2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

SIA-YOLO: A Lightweight Multi-scale Feature Fusion Network For Bearing Surface Defect Detection

Yafei Zhu[1,2], Rangyong Zhang[1,2(✉)], Qijia Ping[1,2] and Jian Li[1,2]

[1] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[2] Shandong Provincial Key Laboratory of Industrial Network and Information System Security, Shandong Fundamental Research Center for Computer Science, Jinan, China
zhangry@sdas.org (R.Z.)

Abstract. Bearing surface defect detection is a key task in manufacturing quality control. However, traditional detection methods often fail to meet the requirements in terms of accuracy and efficiency when faced with defects of small size, diverse shapes and complex backgrounds. To solve this problem, this paper proposes a lightweight multi-scale feature fusion network based on YOLOv11. Firstly, the lightweight New StarNet module is used as the backbone to extract features by stacking multiple star operation blocks, while downsampling is performed using convolutional layers, and nonlinear mapping is achieved through element-wise multiplication. This improves the model's feature extraction capability while reducing inference overhead through lightweight calculation. Secondly, the IRMA attention module is embedded in the neck, so that the model can better extract important features of the bearing surface, while enhancing the small target detection capability and keeping the model lightweight. Finally, the improved AFPN module is used to optimize the detection head, which significantly enhances the model's feature expression capability and effectively improves the model's detection capability for multi-scale defects. Experiments show that the GFLOPs of the SIA-YOLO algorithm on ZC bearing dataset is reduced from 6.4GFLOPs of YOLOv11 to 4.2GFLOPs, a reduction of 34.4%. The mAP@0.5 of the SIA-YOLO algorithm increased by 1.6% from 87.5% to 89.1% . A large number of ablation and comparative experiments have verified the effectiveness and generalization ability of the model in bearing surface defect detection.

Keywords: Bearing Surface Defect Detection, YOLOv11, Multi-scale Feature Fusion, Lightweight, Attention Mechanism.

1 Introduction

Bearings are an indispensable key component in modern industry, and are of great significance for ensuring the normal operation of mechanical equipment and improving production efficiency. Therefore, bearing surface defect detection is of great significance for maintaining the normal operation of mechanical equipment, improving production efficiency, reducing costs and ensuring safety.

In recent years, bearing surface defect detection technology has made positive progress in quality control of industrial applications. However, in complex industrial scenarios, the types of defects are complex and diverse, such as cracks, inclusions, plaques, pitting, rolling scale, scratches, etc., which makes traditional detection methods face severe challenges in accuracy and efficiency. In addition, external factors such as the brightness of the defect surface and the complex background of the defect also affect the detection of bearing surface defects. The development of deep learning technology has provided new ideas for solving these problems. How to use deep learning algorithms to achieve efficient and accurate bearing surface defect detection has become a research hotspot. Among them, the application of YOLO in industrial surface defect detection has steadily increased due to its fast and superior performance. Nevertheless, the direct application of existing methods to industrial bearing surface defect detection still has the following limitations: bearing surface defects usually have different scales and morphologies, and the influence of external factors on bearing surface defect detection makes it difficult for the algorithm to detect subtle defects. For real-time, complex, and real industrial scenarios, an efficient and lightweight bearing surface defect detection algorithm is of great significance.

Some improved methods based on YOLO not only perform well in detection accuracy and robustness, but also significantly improve the efficiency and applicability of the model through lightweight design. For example, Fang et al. [1] proposed YOLOv7-WDD, which improved the mAP by 3.1% compared with YOLOv7 on the NEU-DET dataset by optimizing the feature fusion network and introducing the DECA attention mechanism. Hu et al. [2] proposed a workpiece surface defect recognition method based on improved lightweight YOLOv4. By replacing the original backbone network with MobileNetV2 and introducing deep separable convolution, the model size was reduced by 82.1% and the detection speed was increased by 150% compared with the original YOLOv4 model. Shi et al. [3] proposed a welding robot workpiece surface defect detection method based on machine vision technology, which extracted and classified the workpiece surface defect image through frequency domain feature extraction and nearest neighbor classifier. The CFE-YOLOv8s model [4] proposed by Yang et al. significantly enhances the precision by integrating the CBiF module of CNN and Transformer, the lightweight FC module, and the EFC module with the introduction of the attention mechanism. At the same time, it greatly reduces the model parameters and the amount of calculation, and achieves efficient and lightweight detection effects. The YOLO-DD model [5] proposed by Wang et al. significantly improves the defect detection performance of YOLOv5 by introducing the RDAT, IGFS and SE modules. Zou et al. [6] proposed an industrial scene clothing monitoring method based on improved YOLOv8n and DeepSORT, and significantly improved the small target detection

performance by introducing the FPN-PAN-FPN (FPF) structure, receptive field attention convolution (RFAConv) and focused linear attention (FLatten) mechanism. He et al. [7] introduced DDN, a deep learning-based system for steel surface defect detection which realizes accurate defect classification and positioning by fusing multi-level features, making the model both high-precision and real-time. CAT-EDNet [8] proposed by Luo et al., by introducing the Cross Attention Transformer (CAT) and the Cross Attention Refinement Module (CARM), achieved high-precision defect integrity and boundary detail recognition in the detection of significant defects on the surface of strip steel, with a detection speed of 28 frames per second. Guo et al. [9] proposed an efficient defect detection network EDD-Net, which significantly improved the detection performance of small-scale and low-contrast defects by introducing the improved feature pyramid module GCSA-BiFPN, combined with the global context and spatial attention mechanism for the task of mobile phone surface defect detection. In previous studies, several researchers have enhanced the detection performance of YOLO models by introducing a range of architectural and algorithmic modifications, and improved the efficiency and applicability of the model.

In order to improve the detection accuracy of bearing surface defects while ensuring the lightweight algorithm, this paper studies and develops a lightweight multi-scale feature fusion bearing surface defect detection algorithm. This method improves the defect detection accuracy while reducing the number of parameters, effectively improving the efficiency of bearing surface defect detection. This study makes the following key contributions:

1) First, the lightweight New StarNet module is used as the backbone to extract features by stacking multiple star operation blocks, while using convolutional layers for downsampling and implementing nonlinear mapping through element-wise multiplication, which improves the model's feature extraction capability while reducing inference overhead through lightweight calculation.

2) Secondly, the IRMA attention module is embedded in the neck, which allows the model to more effectively capture the critical features of the bearing surface, while enhancing the small target detection capability and keeping the model lightweight.

3) Finally, the improved AFPN module is used to optimize the detection head, significantly enhance the model's feature expression capabilities, and effectively strengthen the model's performance in detecting defects at different scales.

The rest of this paper is organized as follows. The proposed SIA-YOLO network architecture for bearing surface defect detection is designed in detail in Section 2. Next, experimental analysis is conducted in Section 3 and conclusions are summarized in Section 4.

2 Method

2.1 SIA-YOLO

The dimensions, morphology, and texture of bearing surface defects exhibit significant variability. To more effectively capture the salient features of the bearing surface, we introduce a lightweight multi-scale fusion model named SIA-YOLO. This model has strong feature expression capabilities and can effectively improve the model's detection capabilities for multi-scale defects when computing resources are limited.

First, in order to ensure the model's detection performance while being lightweight to the greatest extent possible, the proposed architecture utilizes the innovative StarNet backbone, which hierarchically extracts visual features through cascaded star-operation modules. Convolutional layers are used for downsampling, and nonlinear mapping is achieved through element-wise multiplication. Secondly, the IRMA attention module is embedded into the neck, so that the model can better extract important features of the bearing surface, enhance the small target detection capability, and keep the model light-weight. Finally, the improved AFPN module is used to optimize the detection head, substantially strengthens the model's feature representation capacity while notably boosting its multi-scale defect detection performance. The overall Network architecture of SIA-YOLO is shown in Fig.1.

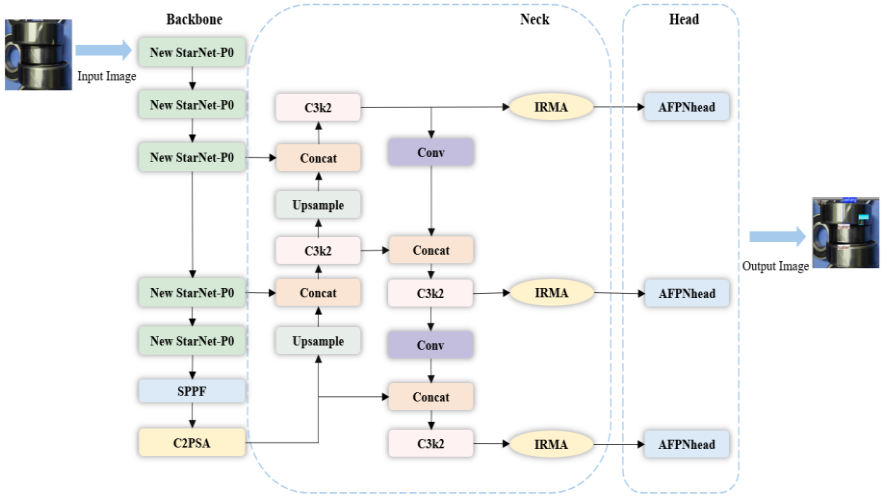


Fig. 1. Network Architecture of SIA-YOLO.

2.2 New StartNet

StarNet [10] is a simple but powerful neural network prototype model that adopts a concise design concept and avoids complex structures and hyperparameter adjustments. The New StartNet proposed in this paper combines the efficient computing power of StarNet with the multi-scale feature extraction of YOLOv11, aiming to ensure that the

model can be lightweight while ensuring detection performance. The New StarNet structure mainly consists of a preliminary feature extraction layer and a deep feature extraction layer. Fig. 2 is the network architecture of New StarNet.

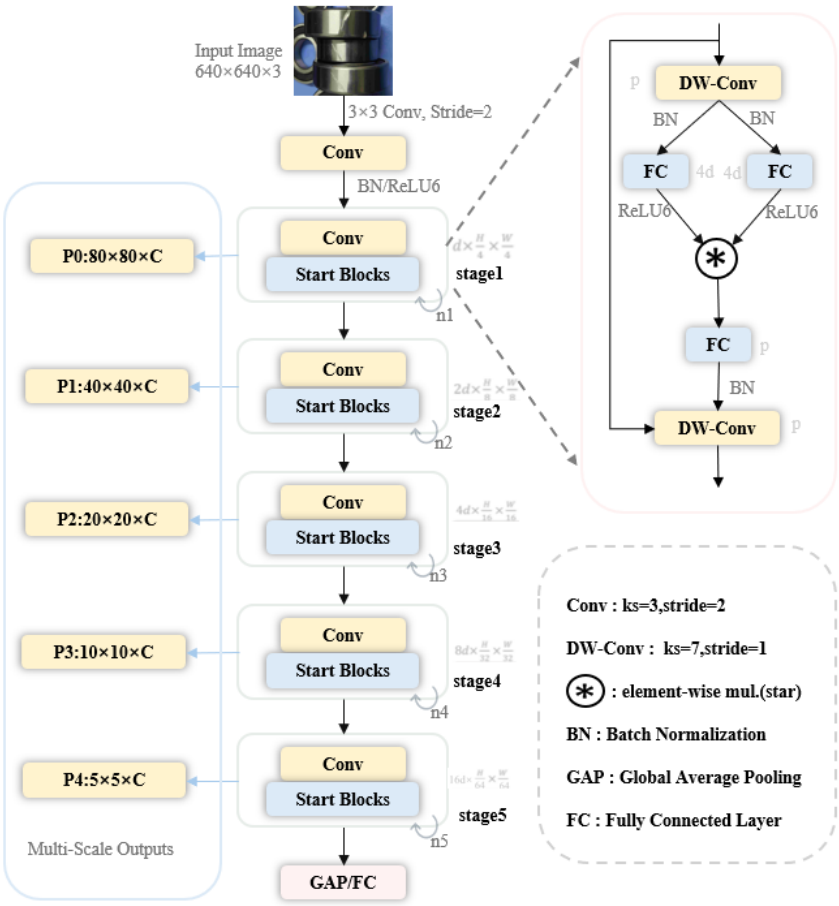


Fig. 2. Network Architecture of New StarNet.

In the initial feature extraction layer, the input image first uses 3×3 standard convolution (stride=1) to expand the channel and combines BN (Batch Normalization) and ReLU6 activation functions to enhance the nonlinear expression ability. This method improves the initial feature representation ability while reducing the resolution and the amount of calculation. Formula (1) is used to calculate the image dimension after the 3×3 convolution transformation.

$$X = ReLU6(BN(Conv_{3\times3}(X))) \quad (1)$$

In the deep feature extraction layer, multiple Star Blocks are used for feature extraction. Each Block consists of Depthwise Separable Convolution, MLP branch, element-wise multiplication, and residual connection. First, local features are extracted through DWConv to reduce the amount of calculation. Then two MLP branches are used for feature conversion; then feature fusion is performed through element-wise multiplication to enhance feature expression capabilities. Finally, residual connection is used to maintain slider flow and improve stability.

This paper uses the New StarNet structure as the backbone for feature extraction, extracts features by stacking multiple star operation blocks, uses convolutional layers for downsampling, and implements nonlinear mapping through element-wise multiplication. It can achieve lightweight while ensuring detection performance, and is suitable for industrial bearing defect detection scenarios.

2.3 IRMA

To more effectively extract critical features from the bearing surface, particularly under constrained computing resources, the IRMA module is incorporated into the neck of the network to bolster feature representation. IRMA extends the EMA attention module [11] by integrating the inverted residual design from iRMB [12], resulting in a compact and effective multi-scale attention module. The IRMA attention mechanism includes Expand Layer, Depthwise Separable Convolution, Compress Layer and EMA attention module, which not only ensures computational efficiency, but also effectively extracts local features, so that the model can still accurately locate the target in a complex background. The module structure diagram of IRMA is shown in Fig. 3.

First, in the Expand Layer, 1x1 convolution is used to expand the number of channels of the input feature from C to C×t (t is the expansion factor, usually 4 or 6), and the feature expression ability is improved by increasing the number of channels. Then 3×3 Depthwise Separable Convolution is used to extract local spatial features to reduce the amount of calculation. Then enter the EMA core attention module to improve the detection accuracy by enhancing the feature representation ability, multi-scale feature representation and global information fusion. Finally, in the Compress Layer, 1×1 convolution is used to compress the number of channels from C×t back to C, restore the number of channels of the feature map, and reduce the amount of calculation.

In the EMA attention module, the input feature map is divided into multiple sub-feature groups, each of which is processed independently, reducing the computational complexity. By grouping, the module can better capture different feature information and reduce the redundancy of the channel dimension. The module uses two parallel sub-networks to capture the channel attention and spatial attention of the feature map. The 1×1 convolution branch encodes the channel information in the horizontal and vertical directions respectively through two global average pooling operations, and then captures the interaction information between channels through 1×1 convolution. This branch avoids the dimension reduction of the channel dimension while retaining the interaction information between channels. The 3×3 convolution branch captures multi-scale feature representation through a 3×3 convolution kernel, increase the receptive

field size. The parallel branch structure optimizes channel weights and spatial weights simultaneously, enhancing the prominence of low-contrast elongated defects. The output features of the two branches are then modulated by the sigmoid function and normalization operation, and finally merged through the cross-dimensional interaction module to capture the pairwise relationship at the pixel level. After the final sigmoid adjustment, the output feature map is used to enhance or weaken the original input features to obtain the final output.

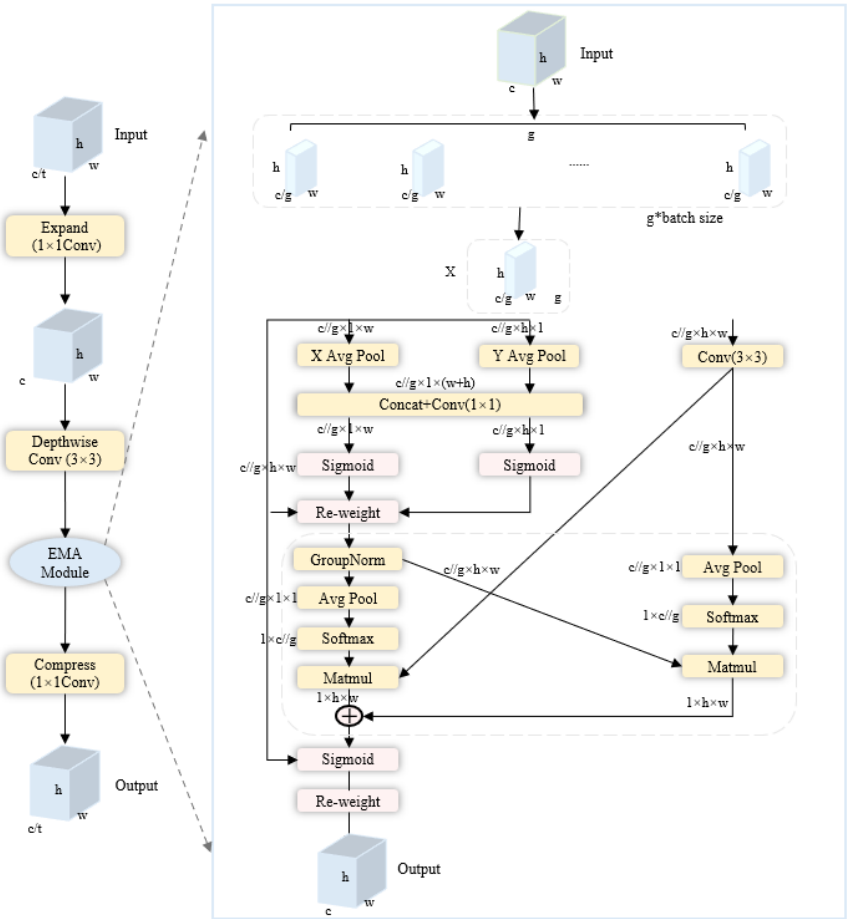


Fig. 3. Network Architecture of IRMA.

This paper integrates the IRMA attention module into Neck to significantly improve the performance of YOLOv11 in industrial defect detection. The IRMA attention

mechanism combines the inverted residual structure with the EMA attention, which enhances the small target detection capability while reducing the inference overhead through lightweight calculation.

2.4 AFPN

This paper improves the AFPN module [13] and introduces it into the model to optimize the detection head, which can effectively improve the model's detection ability for multi-scale defects. Especially in industrial scenarios, the size, shape and texture of defects vary greatly. The AFPN's feature fusion mechanism substantially improves the model's feature representation capacity.

The improved AFPN architecture is shown in Fig. 4. During the bottom-up feature extraction process of the backbone network, AFPN asymptotically fuses the bottom, high, and top features, adopting a progressive fusion strategy to make the features more smoothly merged during the propagation process, rather than direct cascading or simple addition. The goal is to reduce information loss so that shallow features can fully utilize the information of deep features while avoiding conflicts caused by direct fusion. Specifically, AFPN first fuses the bottom-level features, then the deep features, and finally the top-level features, which are the most abstract features. In order to align the dimensions and prepare for feature fusion, this paper uses 1×1 convolution and bilinear interpolation methods to upsample the features. On the other hand, downsampling is performed using different convolution kernels and strides according to the required downsampling rate.

This paper introduces the Adaptive Spatial Fusion Mechanism (ASFF) in the AFPN network [14], which introduces variable spatial weights during the multi-level feature fusion process to enhance the importance of key levels and suppress the influence of conflicting information from different objects. The adaptive spatial fusion mechanism is shown in Fig. 5. By inputting the features into the AFPN for processing, different levels of features can be obtained for fusion, and the results can be input into the detection head for prediction. This improved method aids the model in enhancing detection performance, particularly in handling conflicting information.

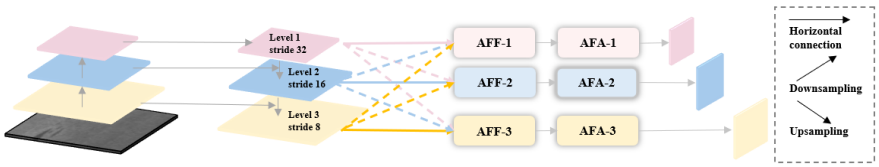


Fig. 4. Network Architecture of AFPN.

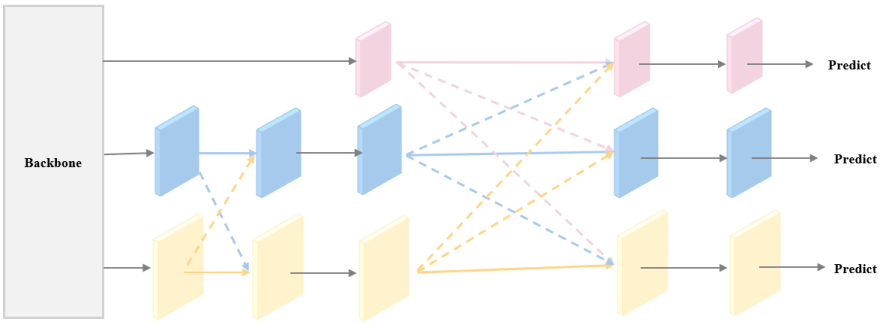


Fig. 5. Adaptive Fusion Mechanism.

In this paper, the AFPN network is improved to enhance the model's perception of fuzzy or unclear edge defects through more efficient information transmission paths. It can accommodate defects of different sizes and shapes, improve generalization capabilities, and achieve faster inference speeds while maintaining high accuracy, making it suitable for industrial bearing inspection.

3 Experiments and Results

3.1 Experimental Environment and Parameter Settings

All the experiments in this paper were conducted on the same computer, and the SIA-YOLOv11 model architecture for bearing surface defect detection was constructed based on the Python3 deep learning framework with Python3 as the main programming language. In this study, YOLOv11n was used as the basic network, and various characteristic ablation experiments and comparative experiments were carried out on YOLOv11n, and the proposed improvements were compared and analyzed. Table 1 outlines the specific software and hardware configurations. Table 2 details the network training parameters employed during model training.

3.2 Dataset

YOLOv11n is used as the base network in this study, and various ablation experiments and comparative experiments are performed on YOLOv11n to compare and analyze the proposed improvements. To assess the detection efficacy of the proposed SIA-YOLO framework, we selected two datasets for experiments: the ZC bearing dataset and the NEU-DET dataset.

The ZC bearing dataset is a self-made bearing defect detection dataset. The dataset was collected and annotated independently in the laboratory. The bearing defect types

include abrasions, scratches, and grooves. The ZC bearing dataset includes 5820 640×640 high-resolution images. Each image has undergone a rigorous annotation process to ensure the accuracy of defect category and location information. Fig. 6 shows the statistical visualization results of the dataset. The dataset is divided into a training set (4074 images), a validation set (1164 images), and a test set (582 images) in a ratio of 7:2:1 to support model training, validation, and testing. The construction of the ZC dataset fully considers the diversity and complexity of actual industrial scenarios, aiming to provide high-quality data support for bearing defect detection tasks.

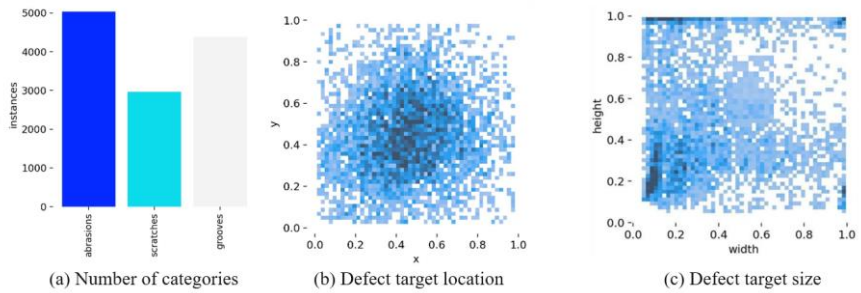
Table 1. Software and hardware configuration details.

Name	parameter
CPU	Intel(R) Core(TM) i9-14900HX@2.20 GHz
GPU	NVIDIA RTX6000
Memory capacity	192GB
System disk	300GB
SSD capacity	500GB
Python version	Python3.9.0
Framework version	Pytorch 1.10.0

Table 2. Details of network training parameters.

Name	parameter
learning rate	0.01
the scale of the input image	640×640
the number of iterations	200
batch size	64
weight attenuation	0.0005

The NEU-DET dataset [15] is publicly released by Northeastern University and is specifically used for hot-rolled strip surface defect detection. The dataset contains a total of 1,800 images, covering six defect categories, with 300 samples for each defect category. The dataset is randomly divided into a training set (1,260 images), a validation set (360 images), and a test set (180 images) in a ratio of 7:2:1.



(a) Number of categories (b) Defect target location (c) Defect target size

Fig. 6. Statistical visualization results of the ZC bearing dataset.

3.3 Evaluation Indicators

In the experiment, the performance of the algorithm is evaluated mainly through a series of indicators. The study employs five key evaluation metrics: Precision, Recall, mAP@0.5, Parameter count, and GFLOPs, to comprehensively assess model performance.

Precision measures the proportion of samples that are actually in the positive class among those determined by the classifier to be in the positive class. The calculation formula is shown in Formula 2:

$$Percision = \frac{TF}{TP+FP} \tag{2}$$

Recall is defined as the proportion of actual positive samples that are correctly identified by the classifier as positive. The corresponding calculation formula is presented in Equation 3:

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

Mean Average Precision (mAP) is the mean of the AP values of all categories. Its calculation formula is shown in Formula 4, where N is the number of categories. The detection performance is assessed by the mAP@0.5 metric, where the Intersection-over-Union threshold is fixed at 0.5 for evaluation.

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{4}$$

Parameters refer to the number of parameters that need to be learned in a neural network, usually including learnable weights and biases such as convolutional layers and fully connected layers. GFLOPs (Giga Floating-point Operations) quantifies the computational cost of a neural network during inference, representing the total floating-point operations (FLOPs) executed in a single forward pass, measured in billions (10^9).

This indicator is essentially a metric for quantifying the load of computing tasks and can be used to effectively reflect the inherent complexity of the model when performing reasoning operations.

3.4 Ablation Experiment

This study conducted 6 sets of ablation experiments on the ZC bearing dataset to test its performance in terms of mAP@0.5, Paramters, Precision, Recall, and GFLOPs. Subsequent experiments gradually added the New StarNet structure, IRMA attention mechanism, AFPN module, and the combination between modules to the modified network to assess their individual and combined contributions. The experimental outcomes are presented in Table 3. This paper abbreviates the New StarNet structure as S, the IRMA attention mechanism as I, and the AFPN module as A.

Table 3. Conducted ablation analyses utilizing the ZC bearing dataset.

Models	mAP@0.5/%	Paramters/10^6	P/%	R/%	GFLOPs/10^9
YOLOv11	87.5	2.6	84.9	76.6	6.4
S-YOLOv11	85.3	1.8	86.1	77.8	4.3
I-YOLOv11	88.5	2.7	87.1	83.7	6.7
A-YOLOv11	87.8	2.3	87.8	81.6	5.8
IA-YOLOv11	90.1	2.5	88.1	82.6	6.2
SIA-YOLOv11	89.1	1.7	88.3	82.6	4.2

Table 3 illustrates that, in comparison to the baseline YOLOv11 model, S-YOLOv11 significantly reduces computational complexity and model size, with GFLOPs and parameter count decreased by 32.8% and 30.8%, respectively, indicating that the use of the New StarNet structure as the backbone network achieves lightweight and improves the detection speed. The I-YOLOv11 model achieves a 1.0% mAP improvement, demonstrating the effectiveness of the IRMA attention mechanism in boosting detection accuracy. The mAP of the A-YOLOv11 model is improved by 0.3%, and the GFLOPs and parameter volume are reduced by 9.4% and 11.5% respectively, indicating that the use of the improved AFPN module to optimize the detection head can reduce the number of parameters while ensuring accuracy. The mAP of the IA-YOLOv11 is improved by 2.6%, which proves that the combination of the IRMA attention module and the improved AFPN module is poised to significantly uplift the accuracy of defect recognition. Finally, In relation to the original YOLOv11 model, SIA-YOLOv11 model achieves a reduction of 34.4% in GFLOPs and 34.6% in the number of parameters, while improving the mAP by 1.6%. These results indicate that the proposed SIA-YOLOv11 model offers a more lightweight architecture with enhanced detection accuracy.

The performance of the SIA-YOLOv11 model is evaluated using the Precision-Recall curve shown in Fig. 7, which is plotted at an IOU threshold of 0.5. The figure shows

the curves of three different categories (abrasions, scratches, and grooves), as well as the average performance of all categories.The closer the curve is to the upper right corner, the better the model performance. As is clear from the figure, the category-wise curves are closer to the upper right corner, and the average precision rate mAP@0.5 of the model is 89.1%, indicating that this model has good precision and recall rates in different categories. In particular, the scratches category performs best.

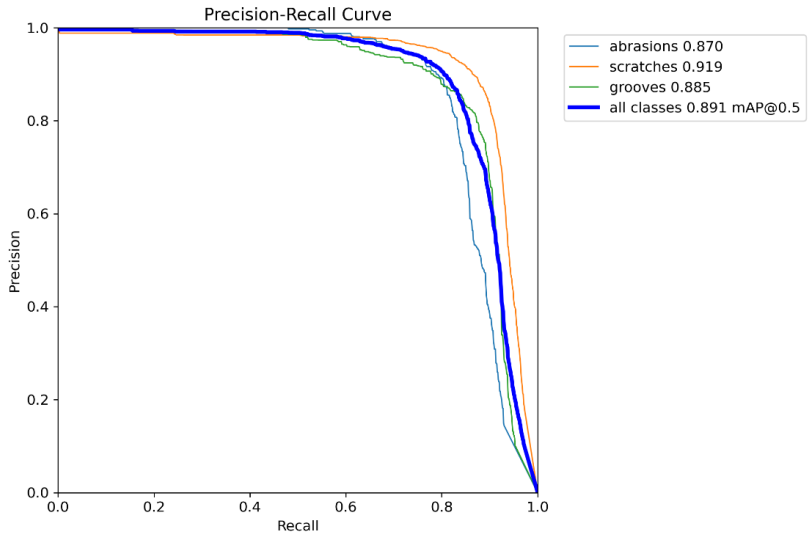


Fig. 7. Precision-Recall analysis conducted at an IoU of 0.5.

In order to investigate the robustness and generalization of SIA-YOLOv11 across datasets, ablation studies were conducted with experiments conducted on the NEU-DET steel surface inspection benchmark. The ablation test results are shown in Table 4. Compared with the YOLOv11 model, the use of the New StarNet structure as the backbone network reduces the GFLOPs and parameters of the model by 32.8% and 30.8% respectively. At the same time, embedding the IRMA attention model and optimizing the detection head with the improved AFPN networks result in more effective gains in relation to detection precision level. SIA-YOLOv11 cuts 34.4% GFLOPs and 34.6% parameters versus YOLOv11, with similar detection performance, while simultaneously improving the mAP by 1.3%. These enhancements demonstrate the enhanced generalization capability of the SIA-YOLOv11 model.

Table 4. Conducted ablation analyses utilizing the NEU-DET dataset.

Models	mAP@0.5/%	Paramters/10^6	P/%	R/%	GFLOPs/10^9
YOLOv11	77.2	2.6	81.4	70.2	6.4
S-YOLOv11	77.3	1.8	70.6	72.8	4.3
I-YOLOv11	77.9	2.7	70.1	71.7	6.7
A-YOLOv11	77.6	2.3	73.5	71.1	5.8
IA-YOLOv11	78.1	2.5	78.6	70.6	6.2
SIA-YOLOv11	78.5	1.7	76.6	71.5	4.2

3.5 Comparison Experiments

This study compares the New StarNet lightweight backbone network with other lightweight networks to verify the impact of the New StarNet lightweight backbone network on model performance. MobileNetv3, ShuffleNetV2, and EfficientNet were selected as comparison references and compared on the ZC bearing dataset. In the comparative experiments, this paper abbreviates the MobileNetv3 module as M, the ShuffleNetv2 module as S, the EfficientNet module as E, and the New StarNet module as SN. The experimental outcomes are presented in Table 5. When juxtaposed with other lightweight network architectures, the New StarNet framework employed in this study as the feature extraction backbone can achieve lightweight while ensuring detection performance, which is suitable for industrial bearing defect detection scenarios.

Table 5. Comparative experiments on various backbones.

Models	mAP@0.5/%	Paramters/10^6	FPS	GFLOPs/10^9
YOLOv11	87.5	2.6	72.6	6.4
M-YOLOv11	85.4	2.2	43.9	5.3
S-YOLOv11	83.9	1.7	40.9	4.1
E-YOLOv11	85.3	3.5	44.1	8.0
SN-YOLOv11	85.3	1.8	40.6	4.3

To assess the effectiveness of the proposed IRMA attention mechanism, we conduct comparative experiments with three established attention approaches: SE, CBAM, and ECA as comparative references. These three attention mechanisms are integrated into Neck respectively, and comparative experiments are carried out on the ZC bearing dataset. The outcomes are detailed in Table 6. This paper denotes the SE module as S, the CBAM module as C, the EMA module as E, and the IRMA module as I. Data analysis demonstrates that this paper integrates the IRMA attention module into Neck, which significantly improves the performance of YOLOv11 in bearing surface defect detection.

Table 6. Comparative experiments on various attention mechanisms.

Models	mAP@0.5/%	Paramters/10^6	P/%	R/%	GFLOPs/10^9
YOLOv11	87.5	2.6	84.9	76.6	6.4
S-YOLOv11	85.1	2.7	84.4	77.4	6.6
C-YOLOv11	86.1	2.5	85.8	76.6	6.4
E-YOLOv11	88.3	3.2	86.8	81.4	7.8
I -YOLOv11	88.5	2.7	87.1	81.7	6.7

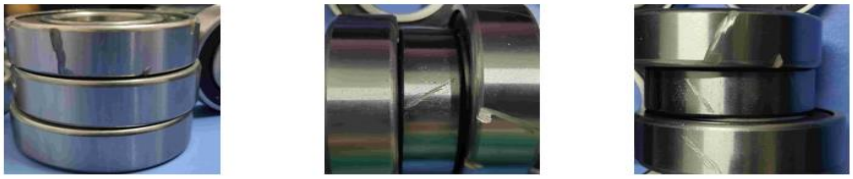
To further evaluate the comprehensive performance of the SIA-YOLOv11 model regarding detection precision and computational intensity on the NEU-DET and ZC datasets, a comparative analysis was carried out against several detection models, including YOLOv5s, YOLOv8s, YOLOv10n, Faster R-CNN, and YOLOv11n, with the corresponding results summarized in Table 7. To highlight the effectiveness of the proposed SIA-YOLOv11 model regarding detection accuracy and efficiency, we compared its performance on the NEU-DET and ZC datasets with that of other mainstream object detectors. The compared models include YOLOv5s, YOLOv8s, YOLOv10n, Faster R-CNN and YOLOv11n. The mAP of the SIA-YOLO model proposed in this paper reached 89.1%, which is higher than all the compared models. Since the SIA-YOLO model uses the New StarNet lightweight network as the model backbone, the number of parameters is only 1.7 and the GFLOPs is only 4.2. In summary, the SIA-YOLO model demonstrates notably high detection accuracy and low computational complexity when applied to the bearing surface defect dataset.

Table 7. Comparative Analysis of Experimental Results.

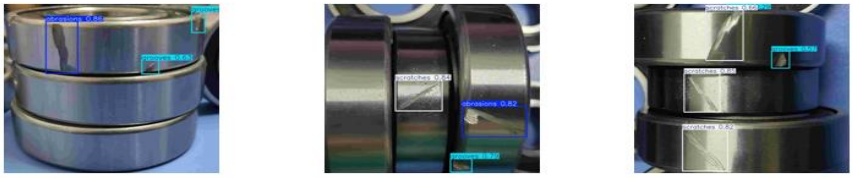
Models	mAP@0.5/%	Paramters/10^6	P/%	R/%	GFLOPs/10^9
YOLOv5s	85.1	7.2	85.4	73.9	17.0
YOLOv8s	86.7	11.2	86.4	71.0	28.6
YOLOv10n	86.9	2.4	84.2	81.3	6.8
Faster R-CNN	83.7	66.0	79.8	78.6	180.3
YOLOv11n	87.5	2.6	84.9	80.6	6.4
SIA-YOLO(ours)	89.1	1.7	88.3	82.6	4.2

In order to better evaluate the generalization ability of SIA-YOLO, the detection results of three bearing defects by YOLOv11n and SIA-YOLO models are shown in Fig.8. As depicted in Figure 8, the SIA-YOLO model introduced in this paper exhibits a more comprehensive capability for detecting defects on the bearing surface, such as abrasions, scratches, and grooves. The confidence scores for detected defect categories

are higher in SIA-YOLO than in the baseline YOLOv11 model, underscoring its enhanced detection capabilities.



a) The origin bearing image



b) The detection results of YOLOv11n

c) The detection results of SIA-YOLO

Fig. 8. Bearing detection results.

4 Conclusions

This study introduces SIA-YOLO, a lightweight approach for multi-scale feature amalgamation designed for industrial bearing surface defect detection. This paper proposes a lightweight multi-scale feature fusion algorithm named SIA-YOLO. This method is applied to industrial bearing surface defect detection. It improves the defect detection accuracy while reducing the number of parameters, effectively improving the bearing surface defect detection efficiency. First, the proposed architecture employs a lightweight New StarNet backbone to enhance feature extraction efficiency, reducing the inference overhead through lightweight calculation, reducing the complexity of the model. Secondly, the IRMA attention module is embedded in the neck, so that the model can better extract important features of the bearing surface, while enhancing the small target detection capability and keeping the model lightweight. Finally, the improved AFPN module is used to optimize the detection head, significantly enhancing the model's feature representation capability, and effectively improving the model's

capability for detecting defects at various scales. The results of ablation experiments and comparative experiments show that the SIA-YOLO model realizes a 34.4% decrease in GFLOPs and a 34.6% cut in parameters, while improving mAP by 1.6% on the bearing dataset. Based on the NEU-DET benchmark dataset for steel defect inspection, the SIA-YOLOv11 model reduces GFLOPs by 34.4% and parameters by 34.6%, while increasing mAP by 1.3%. Compared to other detection algorithms, SIA-YOLOv11 offers higher accuracy, faster speed, and a compact model size.

Acknowledgements. This study was funded by Research and Development of Soil Multi-parameter Composite Sensor and Intelligent Monitoring System (2024CXGC010905), Research and Development and Application of High-end Residential Intelligent Central Air Conditioning and Integrated Internet of Things Configuration Systems (2024TSGC0603), The Construction and Application of a R&D Public Service Platform for Intelligent Innovation Oriented to New-Type Research and Development Institutions (YDZX2023050), Intelligent Manufacturing Empowerment Platform and Its Applications Based on Industrial Internet (YDZX2024121), Research on Key Technologies and Industry Applications of Trustworthy Data Spaces and High-Quality Data Elements (2024ZDZX08), Research on Key Technologies of Sensing for Growth Elements and Vital Signs of Greenhouse Crops (2023TSGC0111), Research on Key Technologies and Applications of Intelligent Management and Control and Informationization Platform for Agricultural Machinery (2023TSGC0587) and Research and Application of Key Technologies for Intelligent Management and Control of Agricultural Machinery and Information Platform in Tai'an City(2023TATSGC042).

References

1. Fang, P., Xu, Z.: YOLOv7-WDD: An Efficient Bi-directional Feature Aggregation Method for Workpiece Defect Detection. In: 2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI), pp. 352–357. IEEE, Changchun (2024)
2. Hu, D.H., Chen, D.F., Yan, K., Cao, Y.: Workpiece surface defects recognition based on improved lightweight YOLOv4. In: 22nd International Conference on Control, Automation and Systems (ICCAS), pp. 1264–1268. IEEE, Busan (2022)
3. Shi, Y., Zhu, Y., Wang, J.: Surface Defect Detection Method for Welding Robot Workpiece Based on Machine Vision Technology. Manufacturing Technology 23(5), 691–699 (2023)
4. Yang, S., Xie, Y., Wu, J., Huang, W., Yan, H., Wang, J., Wang, B., Yu, X., Wu, Q., Xie, F.: CFE-YOLOv8s: Improved YOLOv8s for Steel Surface Defect Detection. Electronics 13, 2771 (2024)
5. Wang, J., Wang, W., Zhang, Z., Lin, X., Zhao, J., Chen, M., Luo, L.: YOLO-DD: Improved YOLOv5 for Defect Detection. Computers, Materials & Continua 78(1), 759–780 (2024)
6. Zou, J., Song, T., Cao, S., Zhou, B., Jiang, Q.: Dress Code Monitoring Method in Industrial Scene Based on Improved YOLOv8n and DeepSORT. Sensors 24(18), 6063 (2024)

7. He, Y., Song, K., Meng, Q., Yan, Y.: An End-to-End Steel Surface Defect Detection Approach via Fusing Multiple Hierarchical Features. IEEE Transactions on Instrumentation and Measurement 69(4), 1493–1504 (2020)
8. Luo, Q., Su, J., Yang, C., Gui, W., Silvén, O., Liu, L.: CAT-EDNet: Cross-Attention Transformer-Based Encoder–Decoder Network for Salient Defect Detection of Strip Steel Surface. IEEE Transactions on Instrumentation and Measurement 71, 1–13 (2022)
9. Guo, T., Zhang, L., Ding, R., Yang, G.: EDD-Net: An Efficient Defect Detection Network. In: 25th International Conference on Pattern Recognition (ICPR), pp. 8899–8905. IEEE, Milan (2021)
10. Ma, X., Dai, X., Bai, Y., Wang, Y., Fu, Y.: Rewrite the Stars. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5694–5703. IEEE, Seattle (2024)
11. Ouyang, D. et al.: Efficient Multi-Scale Attention Module with Cross-Spatial Learning. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE, Rhodes (2023)
12. Zhang, J. et al.: Rethinking Mobile Block for Efficient Attention-based Models. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1389–1400. IEEE, Paris (2023)
13. Yang, G., Lei, J., Zhu, Z., Cheng, S., Feng, Z., Liang, R.: AFPN: Asymptotic Feature Pyramid Network for Object Detection. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2184–2189. IEEE, Honolulu (2023)
14. Liu, S., Huang, D., Wang, Y.: Learning Spatial Fusion for Single-Shot Object Detection. arXiv preprint arXiv:1911.09516 (2019)