scientific reports

Check for updates

OPEN

A lightweight algorithm for steel surface defect detection using improved YOLOv8

Shuangbao Ma[1,2], Xin Zhao[2], Li Wan[3], Yapeng Zhang[1,2] & Hongliang Gao[4]✉

In response to the issues of low precision, a large number of parameters and high model complexity in steel surface defect detection, a lightweight algorithm using improved YOLOv8 is proposed. Firstly, GhostNet is utilized as the backbone network in order to reduce the number of model parameters and computational complexity. Secondly, the MPCA (MultiPath Coordinate Attention) attention mechanism is integrated to enhance feature extraction capabilities. Finally, the SIoU (Simplified IoU) is used to replace the traditional CIoU loss function, which can make the anchor frame more fast and accurate in the regression process, to improve the stability and the robustness of detection. The experimental results indicate that these enhancements have led to a reduction of 37% in calculation amount for the improved YOLOv8n algorithm, a decrease of 32% in parameter count, and an increase in average detection accuracy (mAP) by 1.2%. This model achieves a balance between lightweighting and detection accuracy while providing a viable solution for deployment in computationally resource-constrained edge computing environments such as embedded systems and mobile devices.

Keywords YOLOv8, Lightweight, Defect detection, GhostNet, Attention mechanisms, SIoU

As one of the most widely used materials in the industry, steel plays a critical role in manufacturing and construction[1]. With continuous advancements in industrial technology, there has been significant progress in steel production technology, and the industrial market has increasingly focused on product appearance and quality. However, due to various factors in the production process such as raw material quality, production environment, and processing technology, steel surfaces often exhibit different types of defects including cracks, scratches, folding, and ear[2,3]. These defects vary in size and may have subtle or imperceptible features that make them difficult to accurately detect with the naked eye[4–7]. Not only do these defects affect aesthetics but they can also reduce the quality and performance of steel products leading to safety hazards[8]. Therefore, timely and accurate detection of steel surface defects is crucial[9]. Traditional methods for detecting steel surface defects rely on manual visual inspection[10] which suffers from low efficiency, strong subjectivity, and is prone to errors[11,12].

The advent of deep learning has revolutionized the realm of steel surface defect detection. Image - based object detection methods grounded in deep learning are increasingly being integrated into industrial applications. These methods can automatically glean features from steel surface images and detect defects with a certain level of precision. However, the training and testing of deep - learning models necessitate substantial computational resources, such as high - performance graphics processing units (GPUs) and ample memory. This requirement poses a formidable challenge to practical implementations, particularly in scenarios where terminal devices have limited computational capabilities, such as the embedded systems and mobile devices utilized in industrial field operations.

In the context of steel surface defect detection, several critical issues remain to be addressed. Some defects, like micro - cracks and shallow scratches, are minute and inconspicuous, posing difficulties in detection. Occupying merely a few pixels in the image, these defects may elude the detection of traditional deep - learning models. The complex textures, reflections, and noise on steel surfaces further compound the detection challenge. These background factors can disrupt the accurate extraction of defect features, resulting in false positives or false negatives. Furthermore, for industrial production lines, real - time detection is indispensable. Delays in defect detection can lead to the production of defective products, escalating costs and squandering resources.

Therefore, to achieve an optimal balance among model accuracy, computational complexity, and detection real - time performance, this paper presents an enhanced YOLOv8 - based method for the real - time detection

[1]Hubei Key Laboratory of Digital Textile Equipment, Wuhan Textile University, Wuhan 430073, China. [2]School of Mechanical Engineering and Automation, Wuhan Textile University, Wuhan 430073, China. [3]School of Economics, Wuhan Textile University, Wuhan 430073, China. [4]School of Electrical Engineering and Automation, Hubei Normal University, Huangshi, China. ✉email: gaohl2016@hbnu.edu.cn

of steel surface defects. The principal contributions of this research are as follows: Firstly, the convolutional modules and C2f modules within the backbone network are replaced with C3Ghost and GhostConv modules, respectively. This substitution not only enhances the model's representational capabilities but also significantly reduces the number of parameters and mitigates computational complexity. Secondly, MPCA attention mechanism is integrated into the network. This module effectively captures positional relationships across multiple scales, thereby strengthening the feature extraction capability of the backbone network, minimizing background interference, and improving detection accuracy. Finally, the SIoU loss function is adopted to replace the CIoU loss function. This change comprehensively considers the regression direction between the ground-truth box and the predicted box, thereby enhancing the network's bounding box regression performance and optimizing the detection efficacy for steel surface defects.

Distinct from prior studies that concentrated on single - aspect improvements, such as attention mechanisms or lightweight backbone networks, our method innovatively combines GhostNet for parameter reduction, MPCA for multi - scale feature enhancement, and SIoU for precise regression. This integration strikes a balance between detection accuracy and speed, offering a novel solution that has not been previously reported in the literature.

Related work

At present, deep learning - based surface defect detection algorithms are mainly divided into two categories: two - stage and one - stage algorithms. Two - stage algorithms are represented by R - CNN[13], Fast R - CNN[14], Faster R - CNN[15], Mask R - CNN[16], etc. Their detection process usually consists of a region proposal stage and subsequent classification and regression stages. In the region proposal stage, the algorithm uses a region proposal network to generate potential defect regions. In the second stage, the defects in these regions are classified and located. Although this approach can achieve high detection accuracy, it requires multiple forward passes through the network, resulting in high computational costs. This, in turn, leads to slow detection speeds and large model sizes, making it unsuitable for deployment on resource - constrained edge - terminal devices.

One - stage algorithms, such as SSD[17], YOLO[18] (you only look once), etc., have the advantage of directly predicting the bounding boxes and class labels of defects in a single forward pass. Their fast detection speed is beneficial for real - time application scenarios. Among them, the YOLO algorithm can strike a relatively good balance between detection accuracy and speed, meeting the requirements of real - time detection. Therefore, it has been widely used in the field of steel surface defect detection[19].

Many researchers have conducted in - depth studies on steel surface defect detection and proposed a series of improvement methods. Kou et al.[20] proposed an improved YOLOv3 defect detection model. This model optimizes the detection performance at multiple scales through an anchor - free mechanism, effectively shortening the detection time. Li et al.[21]designed an improved steel surface defect detection algorithm based on YOLOv4. By adding a feature alignment module with an attention mechanism and using a decoupled head to independently output classification and regression results, they significantly enhanced the defect detection performance. Guo et al.[22]proposed an improved YOLOv5 defect detection model. By adding a TRANS module designed based on Transformer[23]to the backbone network and detection heads, they improved the accuracy of defect detection. Gao et al.[24]proposed an improved YOLOv7 method for steel surface defect detection. This method incorporates a split residual convolution network (ResNet)[25]to capture gradient features and uses the Normalized Wasserstein Distance and Complete Intersection Over Union (NWD - CIoU) loss function to enhance the detection accuracy of small and fuzzy targets. Li et al.[26]proposed an improved YOLOv8 defect detection model. By introducing an explicit visual center (EVC) into the backbone network, they enhanced the model's feature extraction ability and adaptability, enabling it to better adjust features at different levels and scales. However, the convolutional network classification model proposed by Xing et al.[27]based on YOLOv3 improved the detection accuracy of steel surface defects by replacing the backbone network and introducing a feature pyramid structure[28], but the detection speed decreased. Cheng et al.[29] proposed DEA_RetinaNet based on RetinaNet[30] for steel surface defect detection. This method improved the detection accuracy by introducing channel attention and adaptive spatial feature modules, but a large number of parameters and high computational costs led to extremely low detection speeds.

When dealing with the task of steel surface defect detection, it remains a highly challenging issue to balance detection accuracy, speed, and the number of model parameters. To address these challenges, researchers have explored from multiple aspects.

Some researchers focus on improving the backbone networks of deep - learning models to enhance feature extraction capabilities. For example, lightweight backbone networks such as MobileNetV3 and GhostNet are designed to reduce the number of parameters and computational complexity. To a certain extent, they can effectively reduce the model size. However, in the scenario of steel surface defect detection, when faced with complex defects with irregular shapes and varying textures, these lightweight backbone networks may not be able to detect them accurately, sacrificing detection accuracy.

Attention mechanisms have also been introduced into deep - learning models to improve their performance. Attention mechanisms such as CBAM[31] and SimAM[32]can enhance the model's focus on defect regions. However, they have limitations in steel surface defect detection. They lack the ability to effectively aggregate multi - scale features. Since steel surface defects vary in size, the lack of this ability may lead to missed detections of small - or large - scale defects.

In the research of loss functions, CIoU[33]and DIoU[34]loss functions have been applied to improve the bounding box regression in deep - learning models. However, these loss functions have certain drawbacks. They do not fully consider the directional alignment between the predicted box and the ground - truth box. This oversight may lead to poor model convergence and inaccurate defect localization, especially when the defects have complex orientations.

Recently, some progress has been made in steel surface defect detection methods. DEW - YOLO introduced a lightweight backbone network to reduce model complexity, and GDCP - YOLO integrated a dual - path attention module to enhance feature extraction. However, DEW - YOLO has limited improvement in detecting small defects, and although GDCP - YOLO has improved performance, its computational cost has increased by 18%, which is a major drawback for real - time applications with high requirements for computational resources and detection speed.

In summary, existing steel surface defect detection methods have many shortcomings. They either sacrifice detection speed, reduce detection accuracy, or have defects in multi - scale feature fusion. Therefore, in the steel surface defect detection in resource - constrained environments, there is an urgent need for a method that can achieve a better balance among lightweight design, high detection accuracy, and fast inference speed.

Model and methods

Improved YOLOv8 model

YOLOv8 is the latest version from Ultralytics based on the improvements of YOLOv5,involving optimisations to the backbone network, feature extraction and predictive header. Firstly, it reduces the convolution kernel of the initial convolution to 3×3 and introduces more layer-hopping connections by replacing all the C3 modules with C2f modules, which enhances the fusion of features at different levels and enhances the gradient flow information. In the feature extraction part, YOLOv8 adopts the PANet (Path Aggregation Network)structure with bidirectional pathways, which shorts the information path between the top layer and the bottom layer, and enables the top layer to make better use of the bottom layer features. On the prediction head, YOLOv8 employs a new decoupled head structure that treats classification and regression tasks separately, and shifts from an Anchors-based approach to an Anchors-Free approach to improve the recognition capability of model[35]. As a result, YOLOv8 achieves faster training speed and detection performance.

To enhance the detection performance of the model for steel surface defects and minimize the number of parameters and complexity, an improved YOLOv8n lightweight steel surface defect detection algorithm is proposed. The network structure of improved model is illustrated in Fig. 1.

GhostNet

Steel surface defect detection usually needs to be conducted in real-time scenarios, such as industrial production lines. Therefore, the algorithm must have efficient inference speed. In order to meet the demands of real-time applications, the design of a lightweight network becomes crucial as it can effectively reduce computational complexity and accelerate model reasoning. The YOLOv8 network uses a number of convolutional modules for removing image noise and extracting key features, and as the number of convolutional operations increases, the

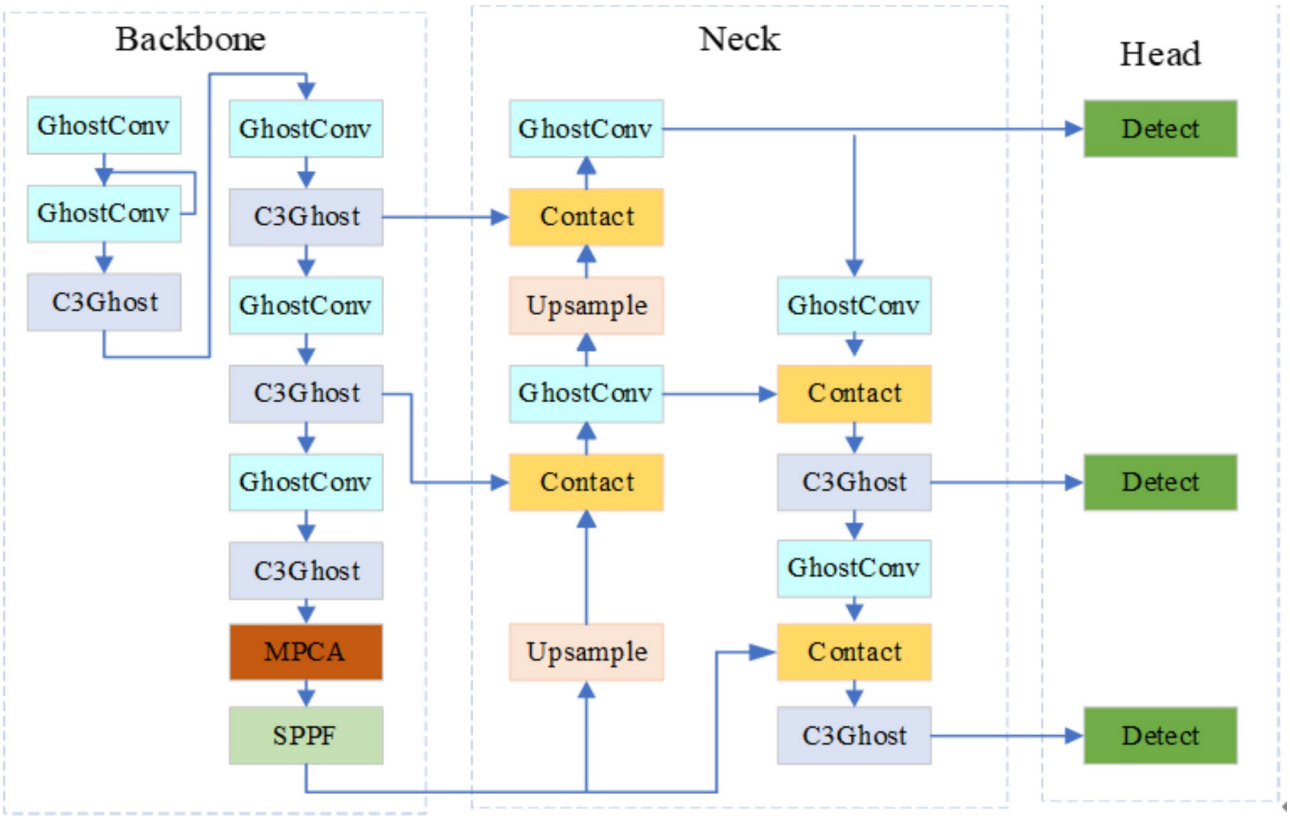

Fig. 1. The Improved Network Structure of YOLOv8n.

computational volume and inference time of the network also increase. In order to optimise the computational efficiency and ensure high-quality feature extraction, GhostNet[36]is introduced to replace the YOLOv8 backbone network, replacing the original Conv layer with GhostConv and the original C2f layer with C3Ghost, which can reduce the computation and inference time while ensuring the accuracy and completeness of the feature representation.

GhostNet is a lightweight neural network architecture designed to enhance computing efficiency and minimize memory consumption by decreasing redundant calculations through efficient feature graph generation. It introduces a new lightweight convolution module (Ghost module), which integrates normal convolutions into two parts. Firstly, it reduces the dimension of an input feature graph using 1×1 convolution operations to decrease channel numbers and extract key information. Secondly, it carries out depth-separable convolutions - a form of layer-by-layer convolutions - utilizing features extracted from previous steps to generate new feature maps. This approach effectively compresses computational requirements while preserving feature extraction efficiency. The structure of the Ghost module is illustrated in Fig. 2.

MultiPath coordinate attention

Steel surface defects usually have different sizes, complex backgrounds and location sensitivity, and may have different characteristics under a variety of environmental and lighting conditions, which is easy to cause problems such as missed and false detection. In order to improve the detection ability of the model for different scales and complex nature, the mechanism of Coordinate Attention with multipath aggregation (MPCA) introduced. Which improves the ability of the model to capture global and local information effectively.

MPCA, as an improved attention mechanism, aims to enhance the performance of deep neural networks in computer - vision tasks. It combines coordinate attention with a multi - path architecture, enabling the acquisition of more comprehensive feature information from different scales and directions.Specifically, MPCA divides the input feature map into four paths[37]. Each path is dedicated to processing information related to a specific scale or direction and is processed by the coordinate attention mechanism. For example, one path may focus on large - scale features, while another path concentrates on small - scale details. Through this initial step, the model can analyze different aspects of the input feature map separately.After the feature - map segmentation is completed, MPCA performs horizontal and vertical global average - pooling operations on each path. The horizontal global average - pooling calculates the average value of each row of the feature map along the horizontal direction, and the vertical global average - pooling calculates the average value of each column. These pooling operations generate two types of coding vectors for each path: horizontal - pooling vectors and vertical - pooling vectors. The horizontal - pooling vectors are used to capture the global information in the horizontal direction, and the vertical - pooling vectors represent the global information in the vertical direction. Both contain important spatial information of the features in the corresponding paths. Subsequently, these coding vectors are spliced and transformed through 1×1 convolutions.Thereafter, MPCA feeds the vectors processed above into a multi - layer perceptron (MLP). The MLP processes these vectors to generate attention weights, and its structure can

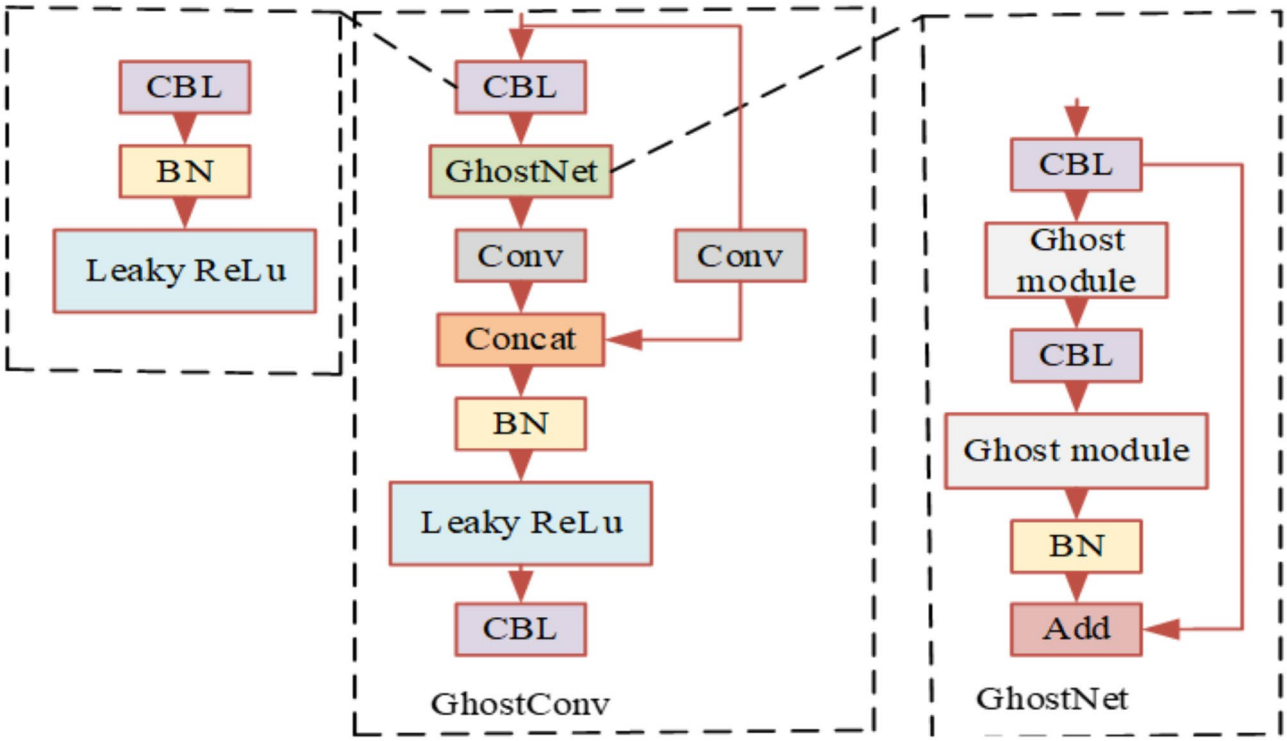

Fig. 2. Structure Diagram of Improved Convolutional.

be adjusted according to the specific requirements of the model. Through a series of linear and non - linear operations, the MLP transforms the input vectors into a set of attention weights, which reflect the importance of different parts of the original feature map.The generated attention weights are then applied to the original feature map on a per - channel basis. That is, each channel of the feature map is multiplied by its corresponding attention weight. This weighting process can highlight the important regions in the feature map while suppressing the less - relevant areas.Finally, the weighted feature maps of each path are combined. According to the design of the MPCA mechanism, this fusion can be achieved through simple addition or concatenation operations. The result of the fusion is the final output of the MPCA mechanism, which contains enhanced feature information and can better represent steel - surface defects. The structure of MPCA is illustrated in Fig. 3.

Simplified IoU

The loss function of YOLOv8 consists of regression loss and classification loss, in which the regression loss part uses CIoU loss function. However, CIoU does not take into account the problem of directional mismatch between the prediction box and the real box, which causes the prediction box to be difficult to align accurately, thus reducing the detection accuracy and training convergence speed of the model. In order to improve the performance of the model, this paper replaced CIoU loss with SIoU loss[38] to better deal with the geometric relationship between the predicted frame and the real frame, enhance the detection accuracy and training efficiency of the model, and improve the detection accuracy and speed of steel surface defects. SIoU redefines the penalty metric by considering the vector Angle between the real box and the predicted box, as shown in Fig. 4.

The SIoU loss function consists of four main parts: Angle lossΛ, distance lossΔ, shape loss $\Omega$and intersection ratio lossIoU, and the formula is as follows:

$$Loss_{SIoU} = 1 - IoU + \frac{\Delta + \Omega}{2},\ IoU = \frac{B \cap B^{GT}}{B \cup B^{GT}},\ \Delta = \sum_{t=x,y} 1 - e^{-(2-\Lambda)\rho t},\ \Omega = \sum_{t=w,h} (1 - e^{w_t})^{\theta} \quad (1)$$

Equation (1) defines the angle loss L_{angle}=1−cos(θ), where θ is the angle between predicted and ground-truth box centers.

$$Loss_{SIoU} = 1 - IoU + \frac{\Delta + \Omega}{2},\ IoU = \frac{B \cap B^{GT}}{B \cup B^{GT}},\ \Delta = \sum_{t=x,y} 1 - e^{-(2-\Lambda)\rho t},\ \Omega = \sum_{t=w,h} (1 - e^{w_t})^{\theta} \quad (2)$$

Where $\rho_x = \left(\frac{b_{cx}^{gt} - b_{cx}}{C_w}\right)^2$, $\rho_y = \left(\frac{b_{cy}^{gt} - b_{cy}}{C_H}\right)$,$w_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})}$, $w_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})}$.

Equation (2) computes shape loss using width (w) and height (h) differences: $L_{shape} = \sum_{t=w,\,h}(1 - e^{-kt})$, with $k = 0.2$.

Angle loss is used to assess the mismatch of direction between the predicted and real frames; Distance loss measures the distance difference between the center point of the predicted frame and the real frame to ensure the accuracy of the frame position; Shape loss deals with the difference in shape between the predicted frame and the real frame to enhance the geometric consistency of the frame; The crossover ratio loss focuses on the proportion of overlapping areas between the predicted and real boxes to reflect the degree of overlap between the two. Through these four loss functions, the target can be located more accurately, and the detection accuracy and robustness can be improved.

Experiment and discussion

Materials

The experiment utilizes the Windows operating system, with a CPU model of Intel i9-9900k@3.60 GHz and a GPU model of NVIDIA GeForce GTX2080Ti graphics card with 11G video memory. The deep learning

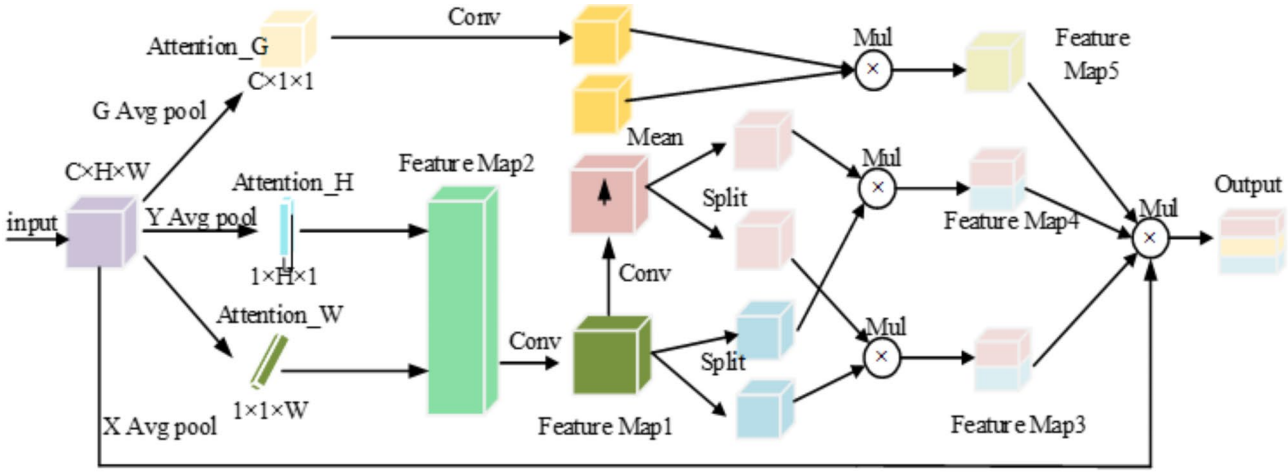

Fig. 3. Structure of MPCA Attention Mechanism.

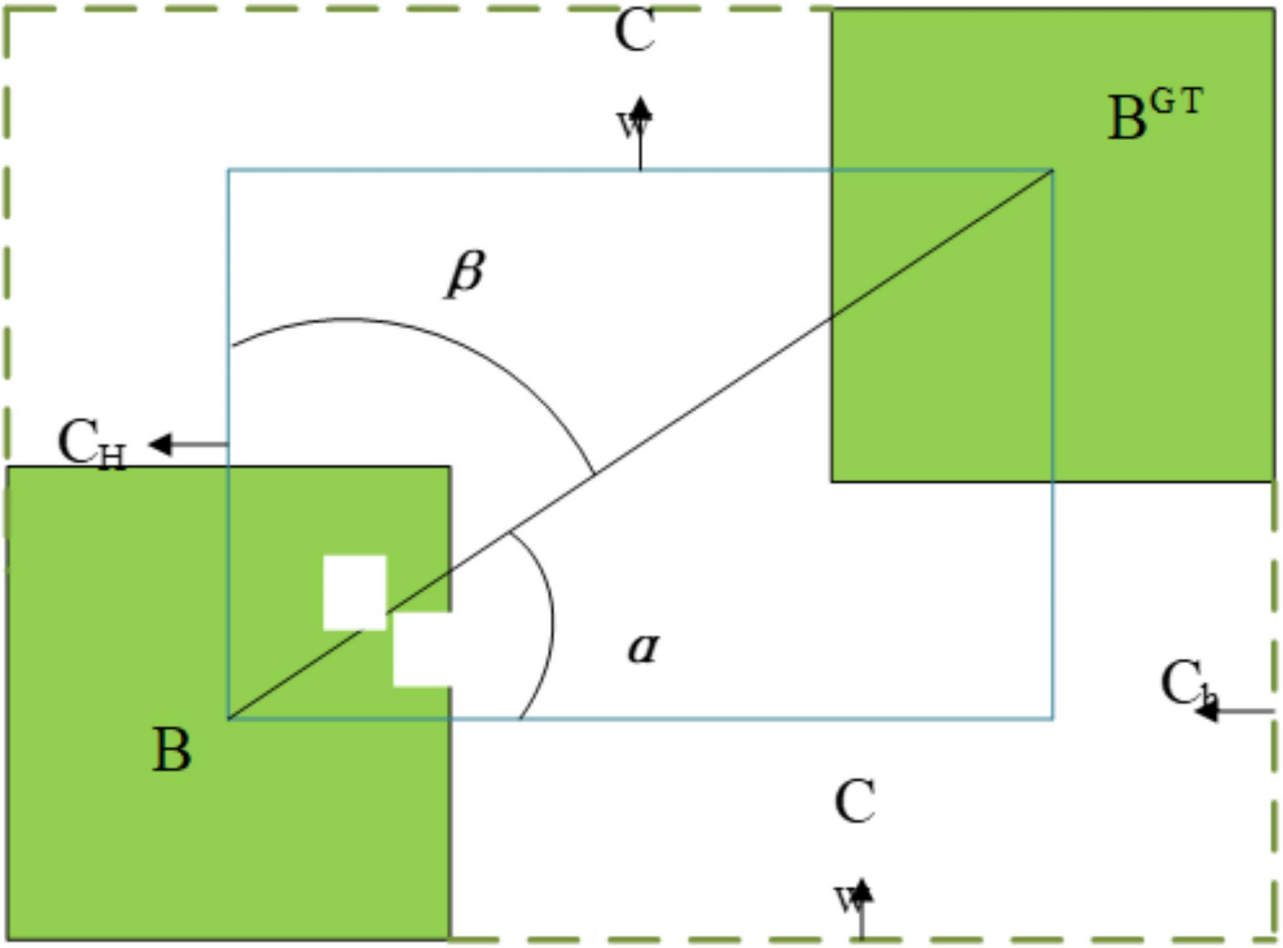

Fig. 4. Schematic diagram of SIoU losses.

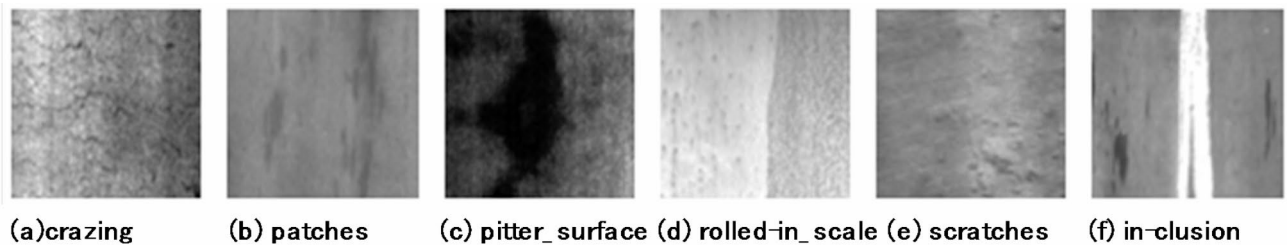

Fig. 5. Defect images of crazing, patches, pitter surface, rolled-in scale, and scratches.

framework used is pytorch-2.1.1, while the programming language is Python-3.11.5 and the CUDA version is 11.7.

In this study, the NEU-DET[39], an open data set from Northeastern University, was used. This dataset is widely recognized as a primary benchmark for steel defect detection[40], and prior studies[20–22,24,26] utilized subsets of similar or smaller scales. The NEU-DET dataset contains a total of six major types of steel surface defects, including crazing, pitter surface, patches, rolled-in scale, scratches and inclusion. With 300 images of each type of defect, for a total of 1,800 images. The dataset was divided into a training set (1440 images), test set (180 images) and verification set (180 images) according to an 8:1:1 ratio. This division method adheres to standard practices and can minimize randomness through stratified sampling across different defect categories. Figure 5 displays different types of Steel Surface defects the six types of steel.

Evaluation metrics

This study employed metrics such as mean average precision (mAP)[40], Parameters (Params), Recall (R), Precision (P) and Billion floating point operations per second (GFLOPS) to evaluate the improved model. The calculation formula is as follows:

$$P = \frac{TP}{TP + FP} \times 100\%,\ R = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} APi,\ AP = \int_{0}^{1} P\mathrm{d}R \tag{4}$$

Lightweight model comparison experiment

In order to prevent an increase in the number of parameters and computational complexity caused by subsequent model improvements, the model is first lightweighted. HGNetV2, MobileNetV3, Ghostnetv3, Mobilenetv4, repvit, and GhostNet are integrated into YOLOv8n by replacing the backbone network, while keeping the detection head and training protocols unchanged. Subsequently, comparative tests are conducted on these improved models, and the results are shown in Table 1.

From the data in Table 1, after introducing GhostNet to replace the original network structure, the model demonstrates unique advantages in many indicators. In terms of the mean average precision (mAP50, at an IoU threshold of 0.5), it is the same as that of the baseline model YOLOv8n, both reaching 0.774. This means that a high level of detection accuracy is maintained, and it can effectively detect steel surface defects. In terms of the number of parameters (Params/M), the GhostNet version has 1.72 M parameters, which is less than the 3.01 M of the baseline model and is between the 1.19 M of MobileNetV3 and the 2.52 M of HGNetV2. While ensuring accuracy, it reduces the storage requirements and computational complexity of the model. The computational amount (GFLOPS) is 5.0, which is also between the 2.4 of MobileNetV3 and the 6.9 of HGNetV2. This indicates that during the inference process, the model will not consume excessive computational resources and has high operational efficiency. When compared with other improved models, the mAP50 of Ghostnetv3 is only 0.714, which is obviously insufficient in detection accuracy. Although Mobilenetv4 has a relatively large number of parameters (5.70 M) and a high computational amount of 22.5, its mAP50 is only 0.755, showing a low cost-performance ratio. Repvit has 6.69 M parameters, a computational amount of 18.5, and an mAP50 of 0.771. Although the accuracy is close, the large number of parameters and computational amount are not conducive to its deployment and operation in practical applications. Taking into account the difficulty of detecting steel surface defects and the performance of the model in terms of accuracy, the number of parameters, and computational complexity, GhostNet ensures a high level of detection accuracy while achieving model lightweighting. Therefore, it is a reasonable and wise decision to select GhostNet for the lightweight operation of the model. In practical applications, it can not only meet the requirements of detection accuracy but also adapt to environments with limited resources and improve the detection efficiency.

Experimental comparison of different attention mechanisms

In order to further verify that the MPCA attention mechanism outperforms other attention mechanisms, we conduct comparative experiments on the reconstructed YOLOv8n model to compare the effects of adding different attention mechanisms at the end of the backbone network, including the channel attention mechanism (CA), the parameter-free attention mechanism (SimAM), and the combined spatial and channel attention mechanism (CBAM). The experiments in Table 2 are based on the benchmark model after lightweighting with GhostNet (Params = 1.72 M, GFLOPS = 5.0), and all attention mechanisms are added on this benchmark.

As illustrated in the Table 2, when detecting steel surface defects, the mAP50 of the MPCA attention mechanism model is 0.784, an increase of 1%, which is better than other mainstream attention mechanisms. The MPCA attention mechanism performs well in steel surface defect detection tasks, significantly improving the detection accuracy without significantly increasing the computational complexity.

IoU comparison experiment

In order to improve the convergence speed and regression accuracy of the model, the SIoU loss function is introduced in this paper and a comparison test with other IoU variants is conducted. The experimental results are shown in Table 3.

Models	mAP50	Params/M	GFLOPS
YOLOv8n(Baseline)	0.774	3.01	8.1
YOLOv8n + HGNetV2[41]	0.767	2.52	6.9
YOLOv8n + MobileNetV3[37]	0.755	1.19	2.4
YOLOv8n + Ghostnetv3	0.714	2.15	6.2
YOLOv8n + Mobilenetv4	0.755	5.70	22.5
YOLOv8n + repvit	0.771	6.69	18.5
YOLOv8n + GhostNet	**0.774**	**1.72**	**5.0**

Table 1. Comparison of lightweighting of different models. Significant values are in bold.

Models	mAP50	Params/M	GFLOPS
YOLOv8n + ghost	0.774	1.71	5.0
YOLOv8n + ghost + CA	0.777	1.72	5.0
YOLOv8n + ghost + CBAM	0.775	1.78	5.0
YOLOv8n + ghost + SimAM	0.775	1.72	5.0
YOLOv8n + ghost + MPCA	**0.784**	**2.04**	**5.1**

Table 2. Experimental comparison of different attention mechanisms. Significant values are in bold.

IoU	Precision	Recall	mAP50	mAP50-95
DIoU	0.74	0.727	0.785	0.454
GIoU	0.747	0.723	0.781	0.441
EIoU	0.717	0.744	0.766	0.441
CIOU	0.768	0.715	0.784	0.444
SIoU	**0.752**	**0.747**	**0.786**	**0.456**

Table 3. Experimental comparison of different loss functions. Significant values are in bold.

Model	mAP50	Params/M	FPS	GFLOPS	模型大小/MB
SSD	0.721	24.4	72.3	62.6	100.3
Faster R-CNN	70.7	100.02	28.9	134.4	159.3
YOLOv3-tiny	0.683	12.13	121.4	18.9	21.4
YOLOv5n	0.744	2.5	139.8	7.1	5.3
YOLOv6	0.77	4.23	121	11.8	8.7
YOLOV8n	0.774	3.01	156.7	8.1	6.3
YOLOv8s	0.779	11.13	106	28.4	22.5
YOLOv9	0.756	60.8	144.3	266.2	122.4
YOLOv10n	0.746	2.3	135.6	6.5	5.8
YOLO11n	0.781	2.6	156.9	6.4	5.5
Ours	**0.786**	**2.04**	**171.5**	**5.1**	**4.5**

Table 4. Result of different detection models. Significant values are in bold.

The model using the SIoU loss function achieves the mAP50 and recall. Although its precision is not the highest, considering all evaluation indicators, SIoU significantly improve the overall performance of the model in the steel surface defect detection task by balancing accuracy and recall rate. Therefore, choosing SIoU loss function as the regression loss function of the model is considered to be best choice.

Comparison experiment of different models

We conducted a comprehensive comparison of the performance of the improved model with other models, such as YOLOv3-tiny, YOLOv5, YOLOv6, and other models, on common datasets. The specific data are shown in Table 4.

Taking into consideration the lightweight implementation and detection effect of the model, it can be concluded that the improved model in this paper exhibits the best performance. The improved model proposed in this study exhibits superior performance in multiple key metrics, achieving an mAP50 of 0.786, which is 0.7% higher than YOLOv8s, which has the highest detection accuracy before. In terms of the number of parameters and model complexity, the improved algorithm in this paper has the lowest number of parameters, only 2.04 M, which is about 18.4% lower than YOLOv5n, which has the lowest number of parameters, and at the same time, the model complexity is the smallest, only 5.1G, which is about 26% lower than YOLOv5n, which has the lowest model complexity. In addition, the FPS of the model is 171.5, which is higher than that of all the models in the table.

The outstanding performance of this model can be attributed to our innovative approach that integrates GhostNet, MPCA, and SIoU. GhostNet, while maintaining a robust feature - extraction capability, leverages its unique architectural design to substantially reduce the number of model parameters and effectively lower the computational complexity. This serves as a fundamental cornerstone for achieving model lightweighting. The MPCA attention mechanism is designed to comprehensively capture feature information across different scales and directions. By doing so, it significantly enhances the model's ability to detect defects of varying scales in complex backgrounds. This ensures that the model can accurately identify target defects under diverse detection scenarios. The SIoU loss function optimizes the model's regression process by incorporating critical factors such

	Category	Crazing	Inclusion	Patches	Pitted Surface	Rolled-in scale	Scratches
mAP50	YOLOv8n	0.456	0.814	0.928	0.791	0.701	0.95
	Ours	0.475	0.829	0.947	0.791	0.702	0.973

Table 5. Comparison results of mAP values for each type of defect detection.

Network	GhostNet	MPCA	SIoU	mAP50	Preci-sion	Recall	Params/M	GFLOPS
YOLOv8n	–	–	–	0.774	0.776	0.704	3.01	8.1
Model 1	√		–	0.774	0.724	0.735	1.72	5.0
Model 2	–	√	-	0.794	0.756	0.737	3.34	8.1
Model 3	–	–	√	0.785	0.76	0.716	3.01	8.1
Model 4	√	√	–	0.784	0.768	0.715	2.04	5.1
Model 5	√	√	√	0.786	0.752	0.747	2.04	5.1

Table 6. Ablation experiments results.

as the vector angle between the predicted bounding box and the ground - truth box. This refinement enables the model to position targets with higher precision, thereby substantially improving the detection accuracy. These three components work in close synergy and complement each other. GhostNet provides a lightweight infrastructure for the model. MPCA is responsible for precise feature extraction and enhancement, enabling the model to better adapt to different defect characteristics. SIoU ensures accurate positioning and regression, guaranteeing the reliability of the detection results. Their collective effect strikes an excellent balance between model lightweighting and detection accuracy. This balance endows the improved model with the ability to maintain a high level of accuracy even under stringent computing resource constraints. As such, the improved model is particularly well - suited for deployment in resource - constrained environments, such as embedded systems or mobile devices. It can efficiently carry out rapid and accurate detection and localization of steel surface defects, thereby meeting the practical requirements of industrial applications.

The Table 5 shows the results of the YOLOv8n model before and after the improvement on the steel surface defects detection, from which it can be seen that the improved model shows different degrees of improvement in most of the defects detection tasks, especially in the detection of cracks, inclusions, patches and scratches.

Ablation experiment

In order to verify the superiority of the algorithm proposed in this paper, we utilized YOLOv8n as the benchmark model and conducted 6 groups of ablation experiments on the NEU-DET dataset. All experiments were carried out under identical network environment and parameters. The symbol "√" indicates that this module was employed in this group of experiments and"-" indicates that this module was not used in this experiment. Refer to Table 6 for detailed experimental results.

The ablation experiments take the original YOLOv8n (Params = 3.01 M) as the benchmark. Each module is added independently or in combination, and the parameters and computational load change dynamically with the addition or subtraction of modules. As shown in the Table 6, compared to YOLOv8n, using the GhostNet module to replace the backbone network, the parameters and FLOPs is reduced by 1.3 M and 3.1G, respectively, while the mAP50value remains unchanged, which is due to the fact that GhostNet drastically reduces the parameter amount and the computation amount through its innovative point-by-point convolution and lightweight design, while maintaining the ability to capture key features. Module 2 introduces the MPCA attention mechanism into the baseline, enhancing the ability of model to capture diverse features and improving the focus on key features, which boosts the mAP50 by 0.794. Model 4 uses SIoU to replace the CIoU loss function of the original YOLOv8n. The SIoU loss function takes into account the shape and orientation information of the target frame more comprehensively by introducing the angle influence factor, which makes the matching between the predicted frames and the real frames more accurate, reduces the leakage and misdetection, and improves the model convergence, and improves the mAP50 by 0.785. Module 3 which combines GhostNet and the MPCA attention mechanism, enhances feature extraction ability while reduceing computational load, improves the mAP50 by 1%, the number of parameters decreases by 32%, and FLOPs decreases by 37% compared to baseline modle. The final improved algorithm integrates the three modules of GhostNet, MPCA, and SIoU, reaching an mAP50 0f 0.786 and a recall rate of 0.747.

Figure 6 shows the finally improved model of F1-Confidence curve, Precision-Confidence curve, Recall-Confidence curve and Precision-Recall curve. Figure 7 shows the confusion matrix of the improved algorithm.

Visual detection results of the improved algorithm

To more intuitively demonstrate the performance of improved model, we applied the model before and after the improvement on the same test set. The improved model adopts a dynamic NMS threshold strategy. For dense small targets, the IoU threshold is increased to 0.6, which effectively reduces the overlapping detection boxes. The detection results are shown in Fig. 8. From the comparison of the detection results in Fig. 8, the original YOLOv8n model has obvious deficiencies. For some small or less - distinct steel surface defects in the figure,

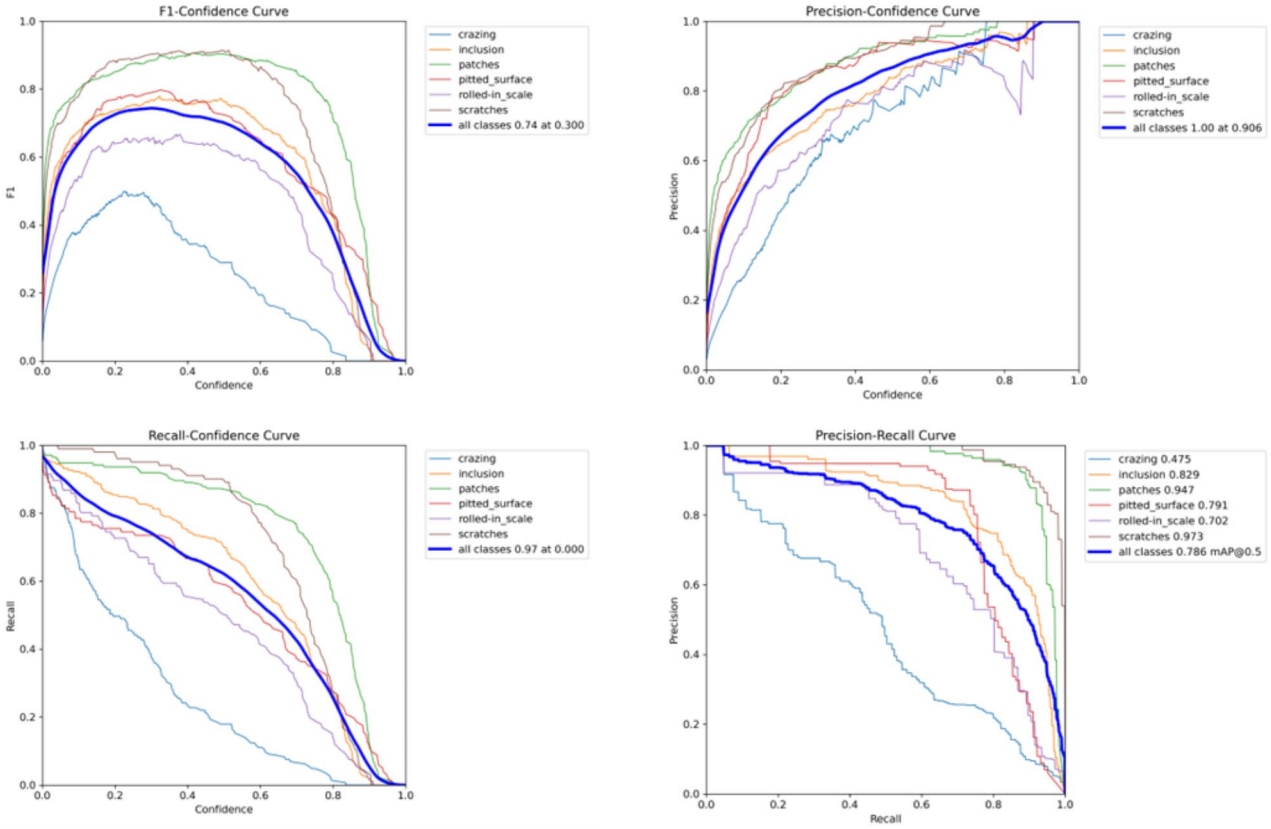

Fig. 6. F1,precision, Recall and PR curves.

such as "inclusion", the original model failed to detect them, failing to identify the corresponding defective areas, resulting in some defects not being boxed and labeled. Meanwhile, in the detection of "patches", the positioning of the detection boxes was inaccurate, and the given confidence levels were not ideal. In contrast, the improved YOLOv8n model has significantly enhanced detection performance. For the "crazing" defect, it can not only locate the defective area more accurately but also increase the confidence level. For defects such as "rolled - in - scale" and "scratches", the improved model can detect them more comprehensively, effectively reducing the missed - detection rate. Moreover, the confidence levels for detecting various defects are more reasonable, and the labeling of detection boxes fits the actual defect shapes better. This indicates that the improved model has a better effect in identifying steel surface defects and can be more reliably applied to actual steel surface defect detection tasks.

Conclusion

In this paper, an efficient and lightweight detection method for steel surface defects based on the improved YOLOv8 is proposed. Initially, the GhostConv and C3Ghost structures are incorporated into the backbone network. This integration serves to not only reduce the number of model parameters but also decrease the computational load, thereby optimizing the model's operational efficiency. Subsequently, the MPCA coordinate attention mechanism module is introduced. By leveraging this module, the method can effectively capture the positional relationships across diverse scales, significantly enhancing the model's feature extraction capabilities. This improvement enables the model to more accurately identify and analyze defect - related features. Finally, the SIoU loss function is adopted to substitute the CIoU loss function. This substitution accelerates the model's convergence speed and enhances the regression accuracy, ensuring that the model can more precisely predict the location and characteristics of steel surface defects.

The experimental results show that the frame rate (FPS) of the proposed method is 171.5 on the NEU-DET dataset, which meets the requirement of real-time detection of steel surface defects. The mean precision (mAP) reached 78.6%, and the number of parameters was only 2.04 M, which achieved a good balance between model lightweight and detection accuracy. It provides a feasible solution for deployment in embedded systems, mobile devices and other edge computing environments with limited computing resources.

In future work, we focus on further improving crack defect detection accuracy while enhancing overall defect detection accuracy to achieve more precise steel surface defect detection results.

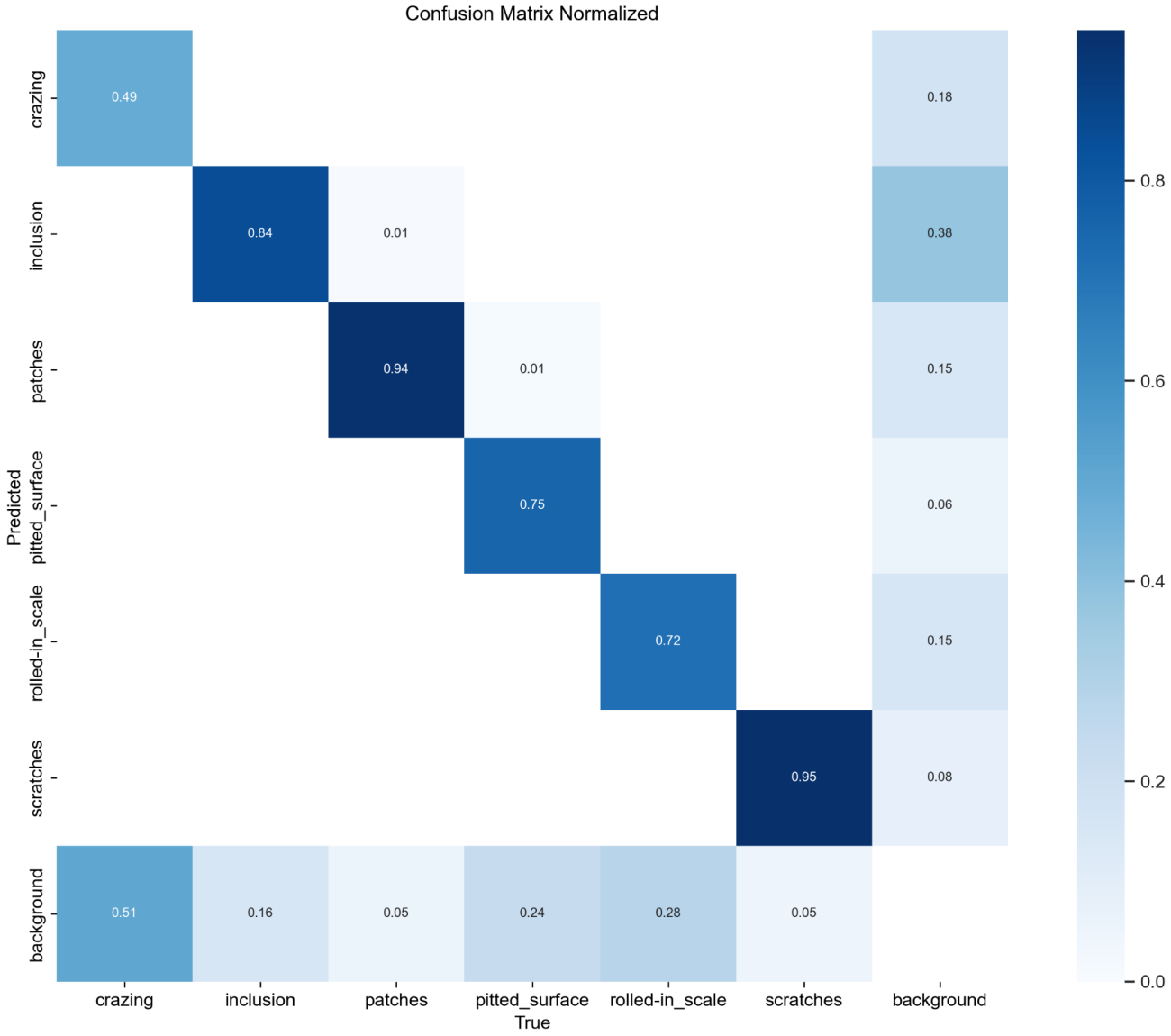

Fig. 7. Confusion matrix.

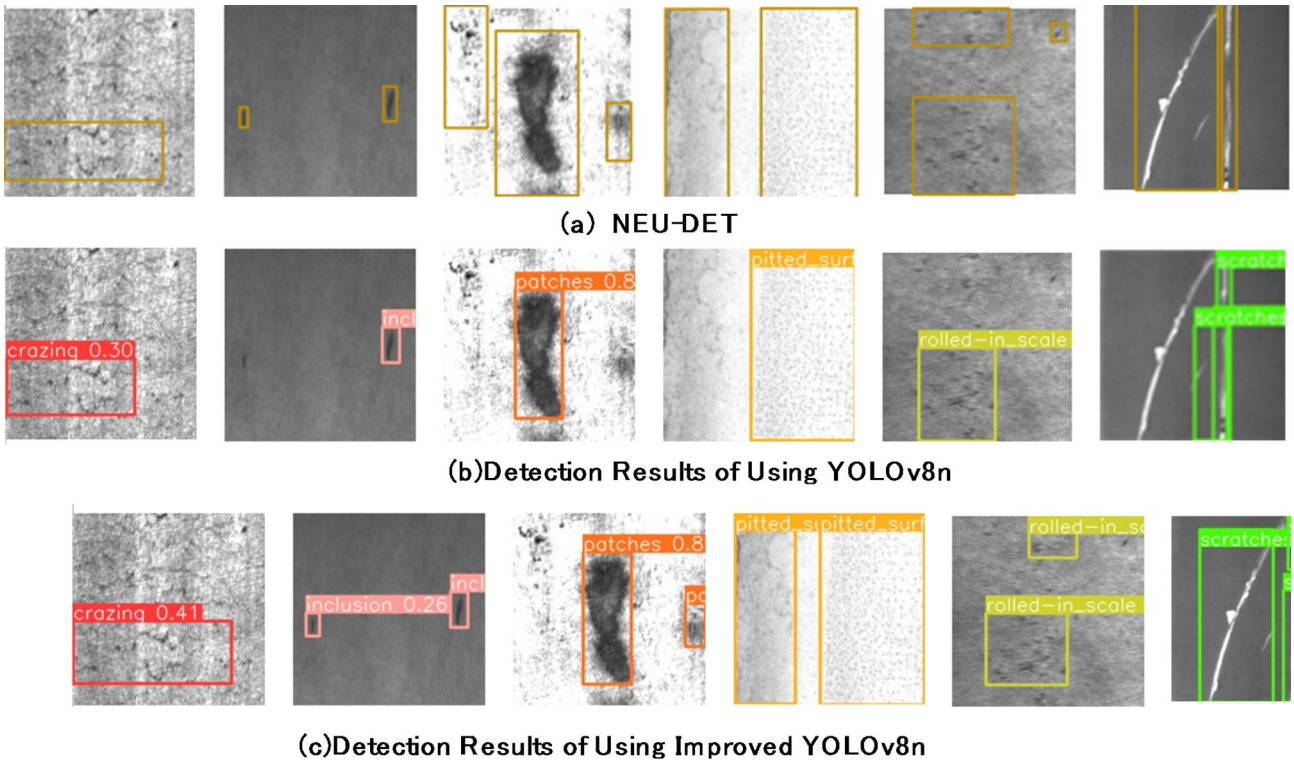

Fig. 8. Steel Surface Defect Detection Results.

Data availability

Data is provided within the manuscript or supplementary information files.

Received: 31 January 2025; Accepted: 6 March 2025
Published online: 15 March 2025

References

1. Zhang, J. et al. Surface defect detection of steel strips based on classification priority YOLOv3-dense network. *Ironmak. Steelmaking.* **48** (5), 547–558 (2021).
2. da da Silva, M. R. et al. A review of gum metal: developments over the years and new perspectives. *J. Mater. Res.* **38** (1), 96–111 (2023).
3. Wang, D. H. et al. Strip surface defect image classification based on double-limited and supervised-connect isomap algorithm. *Acta Autom. Sin.* **40** (5), 883–891 (2014).
4. Li, M., Wang, H. & Wan, Z. Surface defect detection of steel strips based on improved YOLOv4. *Comput. Electr. Eng.* **102**, 108208 (2022).
5. Che, L. et al. Deep learning in alloy material microstructures: application and prospects. *Mater. Today Commun.* 107531. (2023).
6. Lin, K. M., Lin, H. H. & Lin, Y. T. Development of a CNN-based hierarchical inspection system for detecting defects on electroluminescence images of single-crystal silicon photovoltaic modules. *Mater. Today Commun.* **31**, 103796 (2022).
7. Hu, C. et al. Online recognition of magnetic tile defects based on UPM-DenseNet. *Mater. Today Commun.* **30**, 103105 (2022).
8. Huang, Z. C. et al. Effect of repeated impacts on the mechanical properties of nickel foam composite plate/AA5052 self-piercing riveted joints. *J. Mater. Res. Technol.* **23**, 4691–4701 (2023).
9. Zaghdoudi, R., Bouguettaya, A. & Boudiaf, A. Steel surface defect recognition using classifier combination. *Int. J. Adv. Manuf. Technol.* 1–17. (2024).
10. Tian, R. & Jia, M. DCC-CenterNet: A rapid detection method for steel surface defects. *Measurement* **187**, 110211 (2022).
11. Liu, X. et al. Study on the mechanical properties and defect detection of low alloy steel weldments for large cruise ships. *Ocean Eng.* **258**, 111815 (2022).
12. Xing, Z. et al. Rail wheel tread defect detection using improved YOLOv3. *Measurement* **203**, 111959 (2022).
13. Girshick, R. et al. Rich feature hierarchies for accurate object detection and semantic segmentation.Proc. IEEE conference on computer vision and pattern recognition. 80–587. (2014).
14. Girshick, R. Fast r-cnn. Proc. IEEE international conference on computer vision. 1440–1448. (2015).
15. Ren, S. et al. Faster r-cnn: towards real-time object detection with region proposal networks. *Adv. Neural. Inf. Process. Syst.* 28. (2015).
16. He, K. et al. Mask r-cnn.Proc. IEEE international conference on computer vision. 2961–2969. (2017).
17. Liu, W. et al. Ssd: Single shot multibox detector.Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 21–37. (2016).
18. Redmon, J. et al. You only look once: Unified, real-time object detection. Proc. IEEE conference on computer vision and pattern recognition. 779–788. (2016).
19. Yuan, Z. et al. GDCP-YOLO: enhancing steel surface defect detection using lightweight machine learning approach. *Electronics* **13** (7), 1388 (2024).

20. Kou, X. et al. Development of a YOLO-V3-based model for detecting defects on steel strip surface. *Measurement* **182**, 109454 (2021).
21. Li, S. et al. Efd-yolov4: A steel surface defect detection network with encoder-decoder residual block and feature alignment module. *Measurement* **220**, 113359 (2023).
22. Guo, Z. et al. Msft-yolo: improved yolov5 based on transformer for detecting defects of steel surface. *Sensors* **22** (9), 3467 (2022).
23. Vaswani, A. et al. Attention is all you need. *Adv. Neural. Inf. Process. Syst.* 30. (2017).
24. Gao, S., Chu, M. & Zhang, L. A detection network for small defects of steel surface based on YOLOv7. *Digit. Signal Proc.* **149**, 104484 (2024).
25. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 770–778. (2016).
26. Li, J. & Chen, M. DEW-YOLO: an efficient algorithm for steel surface defect detection. *Appl. Sci.* **14** (12), 5171 (2024).
27. Xing, J. & Jia, M. A convolutional neural network-based method for workpiece surface defect detection. *Measurement* **176**, 109185 (2021).
28. Lin, T. Y. et al. Feature pyramid networks for object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2117–2125. (2017).
29. Cheng, X. & Yu, J. Retinanet with difference channel attention and adaptively Spatial feature fusion for steel surface defect detection. *IEEE Trans. Instrum. Meas.* **70**, 1–11 (2020).
30. Lin, T. Y. et al. Focal loss for dense object detection. Proc. IEEE International Conference on Computer Vision, 2980–2988. (2017).
31. Woo, S. et al. Cbam: Convolutional block attention module. Proc. European conference on computer vision (ECCV). 3–19. (2018).
32. Yang, L., Zhang, R., Li, L. & Xie, X. SimAM: A simple, parameter-free attention module for convolutional neural networks. International Conference on Machine Learning, 11863–11874. (2021).
33. Zheng, Z. et al. Distance-IoU loss: faster and better learning for bounding box regression. *Proc. AAAI Conf. Artif. Intell.* **34** (07), 12993–13000 (2020).
34. Rezatofighi, H. et al. Generalized intersection over union: A metric and a loss for bounding box regression. Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 658–666. (2019).
35. Kong, T. et al. Foveabox: Beyound anchor-based object detection. *IEEE Trans. Image Process.* **29**, 7389–7398 (2020).
36. Cao, M. et al. Lightweight tea bud recognition network integrating GhostNet and YOLOv5. *Math. Biosci. Engineering: MBE.* **19** (12), 12897–12914 (2022).
37. Zhang, L. et al. CCDN-DETR: A detection transformer based on constrained contrast denoising for Multi-Class synthetic aperture radar object detection. *Sensors* **24** (6), 1793 (2024).
38. Gevorgyan, Z. SIoU loss: More powerful learning for bounding box regression. Preprint at arXiv:2205.12740 (2022).
39. He, Y., Song, K., Meng, Q. & Yan, Y. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. *IEEE Trans. Instrum. Meas.* **69** (4), 1493–1504 (2019).
40. Liu, K. et al. A new self-reference image decomposition algorithm for strip steel surface defect detection. *IEEE Trans. Instrum. Meas.* **69** (7), 4732–4741 (2019).
41. Guo, A., Sun, K. & Zhang, Z. A lightweight YOLOv8 integrating FasterNet for real-time underwater object detection. *J. Real-Time Image Proc.* **21** (2), 49 (2024).

Author contributions

Shuangbao Ma, Hongliang Gao and Xin Zhao wrote the main manuscript text and Li Wan prepared Figs. 1, 2 and 3,Yapeng Zhang prepared Figs. 4, 5, 6, 7 and 8. All authors reviewed the manuscript.

Funding

This research was funded by the National Natural Science Foundation of China, Project No. 62103309.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-93469-5.

Correspondence and requests for materials should be addressed to H.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.