Research on Steel Surface Defect Detection Algorithm Based on YOLOv12

Jianing Sun, Yujun Zhang*

Abstract—**Aiming at the problems that exist in steel surface defect detection, such as complex features are susceptible to interference from the background environment and it is difficult to take into account the details and global information, this paper proposes an algorithm for small target detection based on the improved YOLOv12 model. The algorithm replaces the traditional up-sampling method in the head of the model as CARAFE module to reduce the information loss in the feature up-sampling process and enhance the detail recovery ability of small defects. Swin Transformer is introduced to reconstruct the global features and capture the global context information through the self-attention mechanism, which demonstrates stronger multi-scale feature fusion capability when dealing with defects with small size or blurred details. In addition, the adaptive weighting of channel features using SEAttention suppresses the background noise and texture interference, and enhances the ability of the inspection head to focus on critical information. Through the combination of these three key techniques, the model is able to more accurately identify and localize defects that are small and easily obscured by the background. The experimental results validate the effectiveness of the proposed algorithm on the NEU-DET (Northeastern University Detection) dataset with a mean average precision (mAP) of 82.5 %, which is a 3.1 % improvement over the original YOLOv12 model. The experimental data show that the proposed model has significantly improved the detection accuracy on the NEU-DET dataset, and can effectively deal with the problem of small target recognition in complex environments.**

Index Terms—**Steel surface defects, Small target detection, Self attention mechanism, YOLOv12**

I. Introduction

Steel production is susceptible to surface defects such as scratches and cracks due to process fluctuations and environmental factors [1–3]. These defects not only weaken the physical properties of the material, but also pose significant safety risks in petrochemical, shipbuilding, nuclear power and other fields. Steel surface defects are usually characterized by small-scale and low-contrast features, which belong to the category of small target detection, and their fuzzy characterization and lack of contextual information pose a double challenge to the algorithms. Effective steel surface defect detection not only helps to improve product quality and production efficiency, but also has great significance in ensuring production safety [4, 5].

Traditional defect detection methods mainly rely on manual sampling, infrared detection, magnetic leakage detection, and other methods[6]. In complex industrial scenarios, there are sampling imbalances that lead to large errors, limited by the different surfaces of the steel material and lead to inaccurate classification of defects, for the detection of small and narrow crack defects is limited and leads to the defect type is not a complete yank and other problems. The traditional defect detection method seriously restricts the detection accuracy. The deep learning method based on a convolutional neural network shows better detection capability than traditional methods under conventional conditions by automatically learning hierarchical feature representation [7]. However, it still faces the challenge of insufficient feature discrimination when facing complex working conditions such as high background noise interference, tiny defect identification and multi-scale target detection. Overall, the current defect detection technology still faces accuracy and adaptability problems in complex industrial environments.

In the field of steel surface defect detection, improved methods based on the YOLO family provide diverse solutions to address the challenges of multi-scale feature fusion, background noise interference, etc. in complex industrial scenarios. Zhao et al. [8] recognize defects on steel surfaces by combining Generative Adversarial Networks (GANs), Autoencoders (AEs), and LBPs without manual annotation. Based on the YOLOv7 framework, Wang et al. [9] made a breakthrough in improving the accuracy of steel surface defect detection by employing a de-weighted bidirectional feature pyramid network (BiFPN) structure. The method effectively reduces the feature information loss in the convolution process by optimizing the feature information utilization and enhancing the multi-scale feature fusion capability. Zou et al. [10] apply the YOLO framework to steel surface defect detection in infrared images and improve the robustness of the model under complex lighting and noise conditions by means of regularization, but there is a problem that the accuracy of feature extraction may be insufficient under complex background. The DSL-YOLO proposed by Wang et al. [11] designs a multi-scale learning strategy for metal surface defects, aiming to capture fine-grained defect features and improve the detection effect, but it may lead to an increase in the complexity of the model structure, which is prone to overfitting problems. Heliyon [12] provides comprehensive performance analysis and optimization by systematically comparing the performances of different YOLO models in surface defect detection. Directions. However, this work mainly focuses on model evaluation without proposing a new detection architecture. Xie et al. [13] proposed a lightweight heavily parameterized feature pyramid network (DE-FPN), which enhances defect feature characterization by incorporating multi-scale convolutional kernel features, but the feature up-sampling process still involves the risk of loss of local details. Zhang et al. [14] built a lightweight feature pyramid network (FPN) by combining the model

Manuscript received April 1, 2025; revised June 20, 2025. This work was supported by the Key Laboratory of Internet of Things Application Technology on Intelligent Construction, Liaoning Province (2021JH13/10200051)

Jianing Sun is a graduate student of School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (e-mail: 1265612389@qq.com).

Yujun Zhang is a Professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China (Corresponding Author, e-mail: 1997zyj@163.com).

clipping technique with YOLOv7-Tiny and constructed a lightweight FPN with YOLOv7-Tiny, which can be used to detect surface defects. Technique to construct a lightweight and efficient steel defect detection network, which is suitable for real-time application scenarios, but the pruning may lead to insufficient model capacity, which affects the ability to discriminate subtle defects.

A number of researchers mentioned above, from different perspectives, have used a variety of different methods to detect steel surface defects in complex backgrounds and multi-scale scenarios, where feature extraction is not precise enough, leading to difficulties in capturing fine-grained defects. YOLOv12 [15] has achieved higher real-time detection accuracy and speed through improved feature fusion and the introduction of a more efficient attention mechanism in its architecture, but there may still be problems of insufficient local detail capture and inadequate multi-scale feature representation in the steel defect detection of the small target detection task, there may still be problems of insufficient local detail capture and inadequate multi-scale feature representation. In order to improve the model's ability of multi-scale feature fusion and local detail capture for steel surface defects, global information interaction and channel recalibration are enhanced, so as to improve the robustness and accuracy of small target detection. Based on this, this paper improves the YOLOv12 algorithm, aiming to improve the detection performance of steel surface defects by enhancing it. Specifically, the CARAFE module is proposed to replace the traditional up-sampling in the YOLOv12 head network. CARAFE efficiently aggregates multi-scale contextual information through adaptive feature up-sampling. It is useful for improving the recovery of detailed information during convolutional upsampling, which enhances the ability to capture small defects on the steel surface. This fine-grained feature reconstruction improves the discriminative effect of the whole detection network on defective regions in complex contexts. In the head network, the Swin Transformer module and SEAttention module are introduced. The Swin Transformer module is placed after the backward convolutional layers in the detection head and utilizes a hierarchical transformer design to capture global and local information.

The introduction of this module provides richer contextual semantics for steel surface defect detection and helps to recognize small and ambiguous defects. SEAttention module follows the Swin Transformer and dynamically recalibrates the channel features to improve the representation of critical information. It automatically emphasizes the feature channels that are most important for defect detection while suppressing redundant and noisy information. This mechanism improves the feature fusion process and enhances the robustness and accuracy of the model for steel defects in complex industrial environments.

II. Related work

Steel surface defect detection is one of the most important research directions in computer vision, especially in petrochemicals, shipbuilding, nuclear power, and other demanding fields with a wide range of applications. However, steel surface defect images are often characterized by various types of defects, complex morphology, serious background interference, and small and uneven distribution of the target area, which brings significant challenges to traditional detection algorithms. To further improve the detection accuracy, researchers continue to explore new solutions, and gradually introduce the attention mechanism, multi-scale feature extraction methods, transfer learning techniques, and reinforcement learning strategies into the steel surface defects detection task to more effectively capture the weak features and adapt to the complex environment, and to improve the accuracy and robustness of the detection.

In recent years, the YOLO series of models have achieved high detection accuracy, but they still have some limitations in dealing with small target detection of steel surface defects. In order to further improve the detection accuracy and better apply it to more complex cases such as steel surface defect images, a variety of improvement approaches have been proposed. For example, RepVGG [16] adopts a reparameterization strategy to greatly improve the feature expression ability while lightweighting, which can more accurately capture the information of small defects, while ConvNeXt [17] realizes the efficient extraction of detailed features in complex backgrounds through modern convolutional architecture and hierarchical design. In addition, RegNet [18] utilizes automated structure search techniques to build a flexible and efficient network configuration that effectively addresses the challenges of multi-scale features and complex interference. In target detection, HRNet [19] adopts a high-resolution feature retention strategy, which is able to continuously capture fine-grained information throughout the network and effectively discriminate fine defects on steel surfaces. The YOLOX-Nano [20] model adopts a lightweight architecture and a high-efficiency detector head, which is able to ensure real-time detection while capturing tiny targets more accurately, thus improving the accuracy of defect detection on steel surfaces. Scaled-YOLOv4 [21] and other model models in the network structure and loss function, for small targets and complex background defect detection has been specifically optimized, making the positioning and identification of tiny defects more accurate.

Based on these improved approaches, an improved steel surface defect detection algorithm based on YOLOv12 is proposed in this study. The algorithm combines a lightweight up-sampling operator that effectively improves detection accuracy and a fused attention mechanism. The method maintains a high detection performance while significantly improving the detection accuracy, providing an effective solution to the poor detection accuracy of small targets for steel surface defects.

III. Model improvements

A. YOLOv12 Improvement Modules

In order to cope with the problems of steel surface defect detection due to the variety of defect types, complex morphology, serious background interference, small target area, and uneven distribution, the model has made targeted improvements in feature extraction and fusion. Three modules, CARAFE, SwinTransformer, and SEAttention, have been introduced successively, each playing an important role in different stages. The improved YOLOv12 model framework is shown in Figure 1.

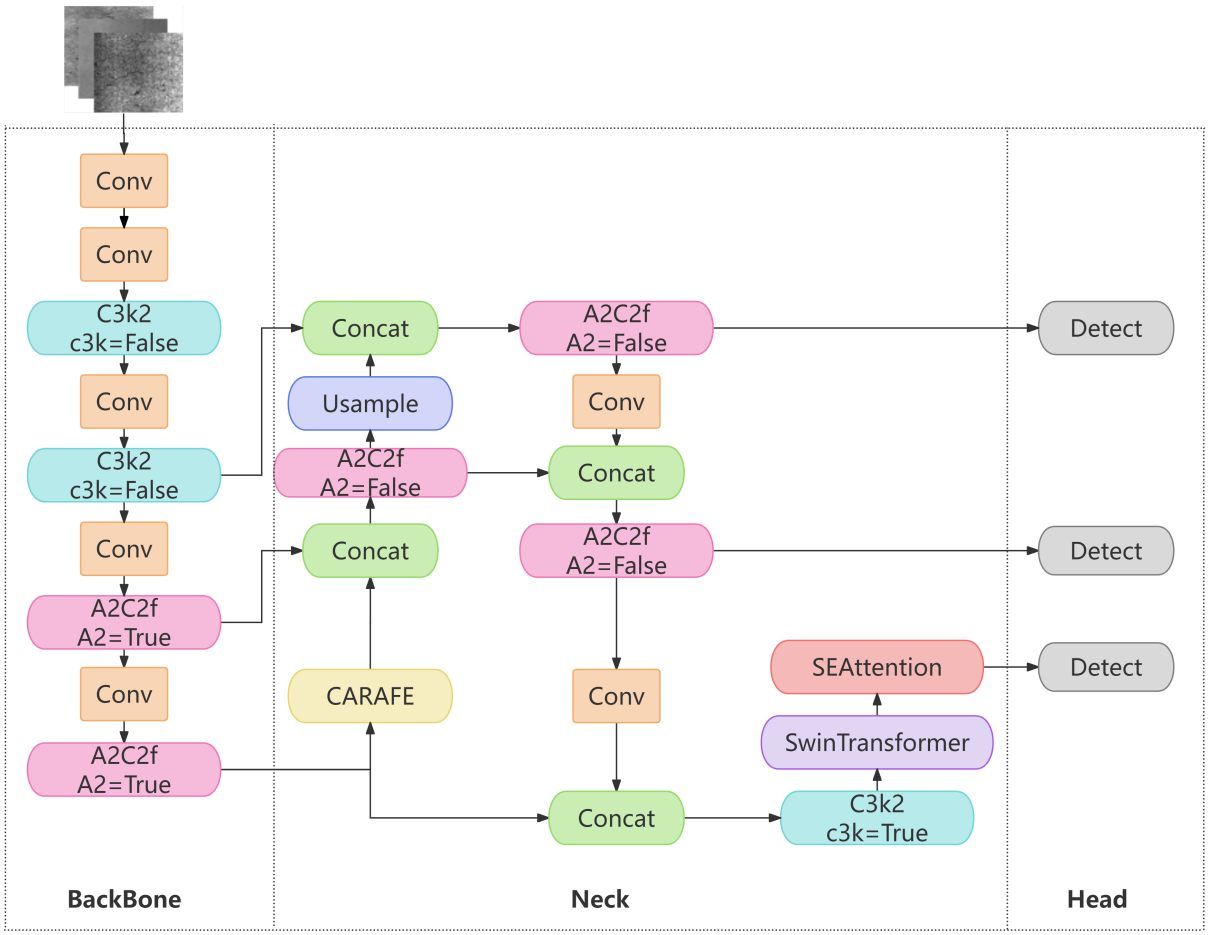

Fig. 1: YOLOv12 improved architecture diagram

In the YOLOv12 neck structure, the CARAFE module is placed in the first up-sampling position, replacing the traditional up-sampling method. During the steel surface defect detection process, defects often have tiny and localized irregular morphology, CARAFE can better capture and restore these fine features, enabling the subsequent detector to obtain clearer and more representative feature information. Swin Transformer module and SEAttention module are added to the front end of the model head in turn. Among them, Swin Transformer utilizes the hierarchical window attention mechanism to model the dependency relationship between local and global, and adaptively captures long-distance and cross-region feature connections to enhance the model's ability to perceive abnormal regions, which improves the recognition accuracy of complex textures and irregular defect patterns on the steel surface. The SEAttention module, added after SwinTransformer, focuses on the channel attention mechanism, which makes it more sensitive to the key texture and edge information in steel defects. This channel-level feature enhancement not only helps suppress redundant background information but also highlights the feature signals in the defect region. This design is useful for steel surface defect detection to further accurately capture and identify abnormal regions, thus significantly improving the detection accuracy.

B. CARAFE

In the traditional up-sampling method, the sampling kernel is determined only based on the spatial location of the pixel point, and its sensing field is only 1×1, which ignores the semantic information of the neighboring regions. The size of defects in steel surface defect detection varies greatly, especially for some small defects, which leads to a degradation in the quality of the feature map after up-sampling, and fails to adequately express the details and contextual information of the input image. CARAFE, as a lightweight universal upsampling operator, effectively solves this problem by introducing an upsampling kernel prediction module and a feature reorganization module. CARAFE helps to better capture and recover the fine defects by adaptively aggregating the surrounding feature information, which improves the ability of multiscale detection. For scenes that require extraction of high-resolution details, CARAFE can better transfer the low-level detail information to the higher level, helping the network to more accurately distinguish tiny texture and defect features. Figure 2 shows the sampling flowchart on the CARAFE operator.

The CARAFE algorithm consists of two parts, the kernel prediction module and the feature reorganization module. It breaks through the limitations of the traditional up-sampling method and enables the model to recover defect details more accurately in the detection task. Let the upsampling

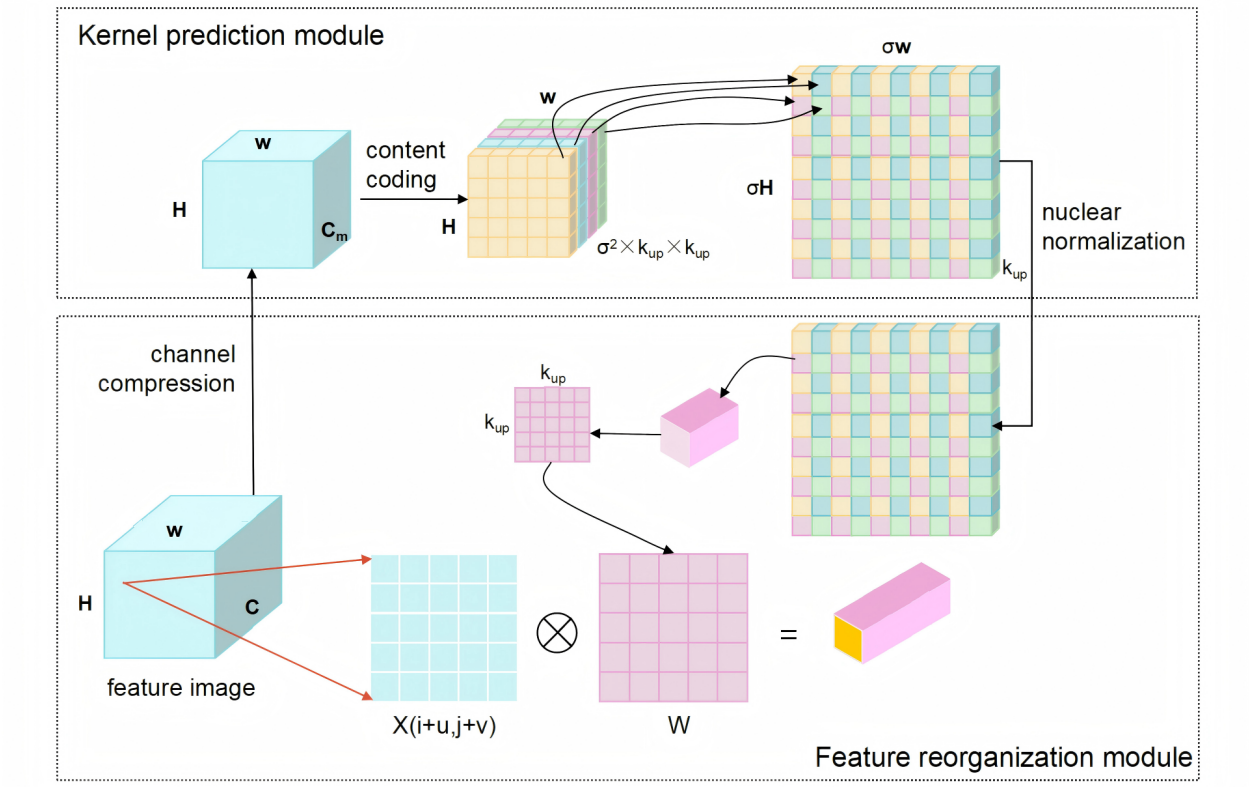

Fig. 2: Upsampling process of CARAFE operator

multiplicity be and the input shape of the previous layer be $X \in R^{B\times C\times H\times M}$, where B is the batch size, C is the number of channels, and H and W are the height and width of the feature map. The specific operation is as follows:

1) The number of channels of the input feature map is compressed using 1×1 convolution to reduce the computation to improve the up-sampling kernel prediction efficiency. The $k_{encoder} \times k_{encoder}$ convolution is then used to predict the up-sampling kernel W_{raw} for each pixel point so that it can be adapted to the features of different regions, allowing the localized a-signatures of steel defects to be better extracted. The up-sampling kernel of dimension $k_{up}^2 \times \sigma H \times \sigma W$ is obtained by unfolding in channel dimension. The up-sampling kernel prediction process is shown in equation (1).

$$W_{raw} = Conv_{k_{encoder}\times k_{encoder}}(Conv_{1\times 1}(X)) \quad (1)$$

2) The up-sampling kernel is normalized using $Softmax$ to avoid certain features from being over-enhanced or ignored. The size of the up-sampling kernel for feature reorganization is extracted in the input feature map X as a $k_{up}\times k_{up}$ local region, which provides rich contextual information for the subsequent feature reorganization and improves the sensitivity of the model to minor defects in steel. The up-sampling kernel normalization and local feature extraction process are shown in equation (2) and (3), respectively. $X(i+u, j+v)$ is a pixel point in the input feature map. $R(i, j)$ is the $k_{up} \times k_{up}$ local region extracted centered on (i, j).

$$W(i,j)(u,v) = Softmax(W_{raw}(i,j))(u,v) \quad (2)$$

$$\mathcal{R}(i,j) = \{X(i+u, j+v) \mid u, v \in \{-\lfloor k_{\text{up}}/2 \rfloor, \ldots, \lfloor k_{\text{up}}/2 \rfloor\}\} \quad (3)$$

3) By means of weighted summation, the normalized up-sampling kernel $W(i, j)(u, v)$ is used for feature reorganization to generate the final high-resolution feature map Y. This can more accurately restore the detailed features of steel defects, reduce misdetections and omissions, and improve the overall detection performance. The computational output process is shown in equation (4).

$$Y(i,j) = \sum_{(u,v)\in\mathcal{R}(i,j)} W(i,j)(u,v) \cdot X(i+u, j+v) \quad (4)$$

C. Swin Transformer

The detection of steel surface defects in the YOLOv12 primitive network head structure undergoes the following process of feature enhancement and reconstruction, multiscale feature fusion, feature convergence and integration in the head. Finally, the high-level features are integrated and the convolution and residual join are used in the C3k2 module for deep extraction and final fusion of the multiscale fused features to form the final feature map required by the inspection head. However, this structure mainly suffers from the problems of limited local receptive field, single information expression, and difficulty in taking both local and global features into account, which affects the accuracy of steel surface defect detection. In this study, after the Swin Transformer module is plugged into the C3k2 module, Swin Transformer, by introducing the self-attention mechanism, is able to provide complementary global context modeling, fine-grained feature enhancement, and improved robustness in terms of global information and long-distance dependence, which are difficult to be covered by convolutional modules. The structure of one of the integrated Swin Transformer

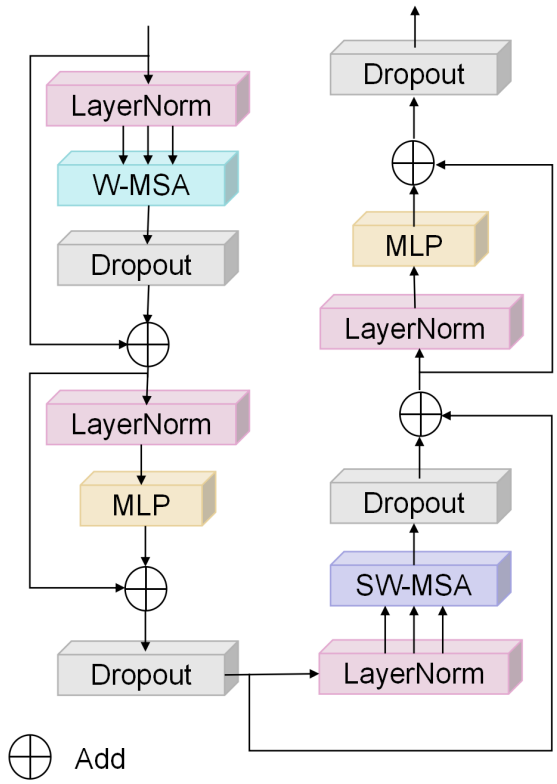

Fig. 3: Swin Transformer architecture

modules is shown in Figure 3. SwinTransformer employs self-attention computation within a local window and realizes cross-window information exchange through sliding window operation, which can effectively capture global semantic information and make up for the problem of limited local receptive field of convolution. In addition, the Windows Multi-head Self-Attention (W-MSA) module used effectively reduces the computation amount and improves the efficiency of image processing compared with the Multi-head Self-Attention (MSA) module of the traditional ViT. The computation amount of MSA is shown in eqation (5), and the computation amount reduced by W-MSA is shown in equation (6).

$$\Omega(\text{MSA}) = 4hwC^2 + 2(hw)^2C \tag{5}$$

$$\Omega(\text{W-MSA}) = 4hwC^2 + 2M^2(hw)^2C \tag{6}$$

Where h, w, C represent the height, width, and number of channels of the feature map, respectively, and M represents the window-size, which is usually 7 by default.

Residual connections are used between sublayers in the module. Assuming that the token composition matrix $X \in R^{M^2 \times d}$ within a certain window is grouped according to a predefined window size M×M, self-attention is computed within each window to capture the detailed relationships within the local region. Where the computation process of W-MSA is shown in equation (7).

$$\text{W-MSA}(X) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + B\right)V \tag{7}$$

Where Q, K and V are vectors of queries, keys, and values, respectively, d_k is the number of channels of the input feature map, and B is the relative positional bias of the processed image. W-MSA transforms localized image blocks into high-dimensional vectors, which retain the steel surface texture and detail information and provide fine-grained inputs for the subsequent attention mechanism. However, W-MSA computes the self-attention only within each fixed window, resulting in the token at the window boundary not being able to fully establish the connection with the token in the neighboring windows, which in turn leads to the incomplete expression of the features in the edge region and affects the detection effect of the subtle defects. In order to solve these problems, the Shifted Windows Multi-Head Self-Attention (SW-MSA) sliding window strategy is introduced after W-MSA.

D. SEAttention

CARAFE improves spatial information recovery by better preserving and fusing low-level details through advanced upsampling methods, while SwinTransformer enhances contextual modeling by capturing global and long-range dependencies using self-attentive mechanisms. While both provide strong support for small target detection, their respective strengths lie in spatial detail and regional information fusion and are not directly optimized for the importance of different channels. It is in this context that the introduction of the SEAttention module is essential, which dynamically adjusts the response weights of each channel through global information aggregation and channel recalibration, allowing the model to automatically focus on those feature channels that are critical for detecting small defects on steel surfaces, further improving detection accuracy. The main operations of SEAttention include Squeeze, Excitation and Scale: the Squeeze operation is to initially process the input feature map through the convolution operator, and then use global average pooling to compress the feature map into a global feature vector of 1×1×C. The Excitation operation processes the global feature vector with two fully-connected layers and nonlinear activation functions to generate the weights of each channel. The Scale operation applies the weight vector obtained from the Excitation operation to the feature map processed by the Squeeze operation to complete the channel re-calibration and obtain the final weighted feature map. The specific operations are as follows:

1)The feature maps$X \in R^{C\times H\times W}$processed by the Swin Transformer module are pooled globally on average to compute global features for each channel. The Squeeze operation aggregates global information to capture the contextual information of the global steel surface defect image and reduce the effect of spatial dimensions. The global average pooling is calculated as shown in equation (8). Where z_c is the global description value of the c channel, $X_c(i,j)$ represents the value of the c channel at position (i,j), and the result z=[z_1, z_2, ... ,z_c] is shaped as R_c.

$$z_c = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_c(i,j) \tag{8}$$

2) In steel surface defect detection, different channels may contain different scale and texture information. Excitation operation learns the relationship between individual channels, i.e., computes the attentional weights between channels

to obtain adaptive weights. First a dimensionality reduction operation is performed using a fully connected layer with ReLU activation, shrinking the number of channels to C/r with the aim of reducing the computational effort. Then the dimensionality is raised to recover the number of channels C and the weights are normalized using Sigmoid activation. The weights obtained by Sigmoid activation can automatically adjust the response strength of each channel, thus highlighting those fine-grained features related to defects. The calculation process is shown in equation (9). Where $W_1 \in R^{(\frac{C}{r})\times C}$is the degenerate matrix, W_2 is the lifting matrix, δ stands for the ReLU activation function, σ stands for Sigmoid normalization.

$$s = \sigma\big(W_2\,\delta(W_1 z)\big) \tag{9}$$

3) The Scale operation is to weight the original feature map to strengthen the feature response of the defect region. After the channel weight s is obtained by Excitation operation, it is reapplied to the original input feature map X to amplify the important channels and suppress the minor channels, so that the model can extract the useful information more effectively and reduce the background interference when detecting the defects on the steel surface. The operation process is shown in equation (10). Where s_c represents the channel weights and X_c is the c channel of the original feature map, $\tilde{X}$ is the weighted output feature map with the same shape as the input feature map.

$$\tilde{X}_c = s_c \cdot X_c \tag{10}$$

IV. EXPERIMENTAL DESIGN AND IMPLEMENTATION

A. Dataset Introduction

In order to verify the superiority of the YOLOv12 algorithm in the task of steel surface defect detection, the NEU-DET dataset is selected in this study and an improved framework adapted to multi-scale defect detection is constructed. The NEU-DET dataset is captured from the surface of steel plates in the actual production and is designed to reflect the common defects in the steel processing. The dataset contains a total of 1800 images covering six typical defect categories, including crazing, inclusion, patches, pitted_surface, rolled-in_scale, and scratches, with about 300 images in each category. In order to ensure the stability of the experimental results, the dataset is divided into training set, validation set, and test set in the ratio of 8:1:1. There are 1440 images in the training set, 180 images in the validation set, 180 images in the test set, and the images are randomly distributed. Since the dataset is balanced and the samples of each category are basically the same, the model fully learns each category during the training process. It avoids category bias, reduces the problem of leakage and misdetection, improves the generalization ability of the model, and provides a good foundation for the improvement process of YOLOv12 algorithm.

B. Experimental environment and parameter configuration

The experiments were conducted on Windows 10 with CUDA 12.1, Pytorch 2.4.0, Python 3.8.10, and NVIDIA GeForce RTX 3090, which has 24G of video memory, providing a solid foundation for efficient training of deep learning models. The core parameters of model training are epoch, batch size, patience, and learning rate, of which the epoch size of this experiment is set to 300, and the batch size is set to 16. For the risk of overfitting during the training process of the model, the patience value of this experiment is set to 50 for the start-stop strategy. If the validation loss is not significantly reduced within 50 consecutive epochs, then the training of the model will be terminated early. In order to balance the training speed and the convergence of the model, the learning rate value size is set to 0.01 in this experiment.The rest of the parameters use the default values.

C. Model evaluation metrics

The experiments use four key metrics to evaluate the performance of the improved YOLOv12 model on the NEU-DET dataset. The four key metrics are Precision, Recall, mean Average Precision (mAP), and F1-score.Precision refers to all precision measures the proportion of all regions labeled as defective by the model that are actually truly defective. The formula is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

where TP is the positive detection rate and FP is the false detection rate. TP is the positive detection rate and FP is the false detection rate.

Recall evaluates what percentage of all actual defects can be detected by the model. The formula is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

where FN is the number of negative samples detected as positive.

mAP is a comprehensive evaluation metric that reflects the overall performance of the model on all detection tasks by calculating the average of precision and recall under different thresholds. The calculation formula is as follows:

$$mAP = \frac{1}{C}\sum_{i=1}^{c} AP_i \tag{13}$$

where c is the total number of image categories, i is the number of detections, and AP is the average accuracy of single category recognition.

$F1 - score$ is a reconciled mean of precision and recall that balances the two. The formula is as follows:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

D. Results and Analysis of the Precision-Recall Curve

The difference in detection performance of the YOLOv12 model before and after improvement on the NEU-DET dataset is clearly contrasted in the curve of the Precision-Recall (P-R) plot. The average mean precision (mAP@0.5) for all categories improves from 79.4 % to 82.5 % in terms of numerical change on the YOLOv12 model before and after the improvement. additionally there is a significant enhancement in terms of both detection precision and recall. Specifically, the AP value of the “crazing” category is improved from 46.1 % to 62.6 %, and the AP value of the “rolled-in_scale” category is improved from 65.0 % to 67.7 %. While the AP value of the “scratches” category is improved from

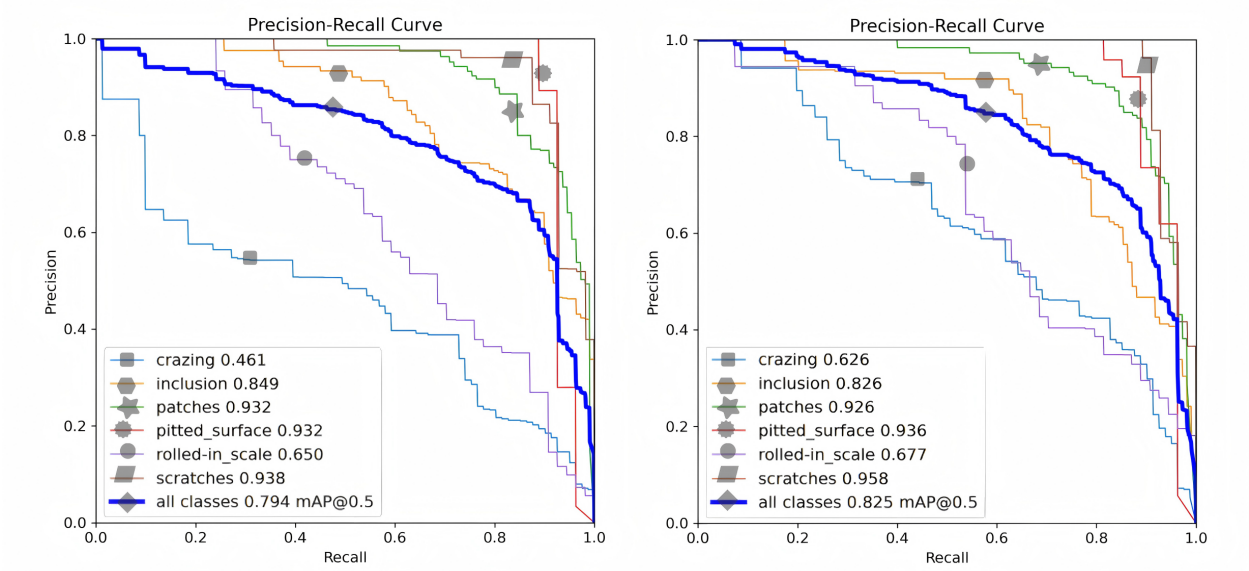

Fig. 4: Results Comparison

93.8 % to 95.8 % on the YOLOv12 model before and after the improvement, there are significant improvements in both detection precision and recall. Although the AP value of the "patches" category decreased from 93.2 % to 92.6 %, the AP values of most of the categories improved overall. In addition, the curve of the P-R diagram is more concentrated and the shape of the curve is smoother, and the balance between precision and recall is enhanced. In conclusion, the performance of the improved YOLOv12 model for steel surface defect detection on the NEU-DET dataset has been effectively improved.

E. Ablation experiments

In order to verify the effectiveness of the improved modules, ablation experiments are conducted on the public dataset NEU-DET to illustrate the impact of different improved parts on the performance of the original YOLOv12 algorithm. We test the CARAFE module, the Swin Transformer module and the SEAttention module respectively, and the results are shown in Table 1. By gradually adding or removing the improved modules, the experimental data can reflect the impact of different modules on the overall detection performance of the model.

According to the data in the table, the mAP of the original YOLOv12 is 79.4 %. Analyzing the data of different module combinations, replacing the traditional up-sampling method of the neck for the CARAFE module improves the ability of capturing and recovering subtle defects, and the mAP is improved by 0.6 %; the Swin Transformer module is introduced in the head to make up for the problem of the limited convolutional local receptive field and to improve the ability of capturing the global semantic information, and the mAP is improved by 1.5 %; and the introduction of the SEAttention module is followed immediately by the introduction of the SEAttention module, which focuses on the feature channels important for detecting steel surface defects, further improves the detection accuracy of the model and obtains 0.1 % mAP improvement. The highest mAP value of 82.5 % is for the improved model retaining the three improvement modules, which is 3.1 % higher than the mAP value of the original YOLOv12 model. This data proves that the combination of the three improvement modules plays a crucial role in improving the accuracy of the model in the detection of steel surface defects.

TABLE I: Ablation experiments

	CARAFE	Swin Transformer	SEAttention	mAP%
YOLOv12	-	-	-	79.4
YOLOv12	√	-	-	80.0
YOLOv12	-	√	-	81.5
YOLOv12	-	-	√	79.5
YOLOv12	√	-	√	81.9
YOLOv12	√	√	-	80.5
YOLOv12	-	√	√	81.2
YOLOv12	√	√	√	82.5

F. Comparison results of different models

In order to comprehensively evaluate the accuracy enhancement effect of the improved target detection model based on YOLOv12 for the detection of NEU-DET dataset, we selected several advanced steel surface defect detection algorithms, including YOLO-LFPD, DCC-CenterNet, YOLOX, YOLOv8s, etc., and the results are shown in Table 2. In comparison, the other models may have slightly higher or lower accuracies in each category, but overall it can be seen that the detection accuracy of the improved model is significantly improved in most of the categories, and in addition, the accuracy of the improved model reaches a satisfactory 82.5 %. In conclusion, the accuracy of our improved model on the NEU-DET dataset is significantly improved, providing a good foundation for subsequent research and applications.

V. Conclusion

In this paper, we propose an improved algorithm model YOLOv12 for steel surface defects detection to solve the

TABLE II: Comparison of different models

Model	crazing	inclusion	patches	pitted_surface	rolled-in_scale	scratches	mAP%
YOLOv7	36.8	85.6	80.7	88.1	58.7	90.4	73.4
YOLOv8s	43.6	82.2	78.1	94.0	66.8	83.3	74.7
YOLOX [22]	46.6	83.1	83.5	88.6	64.8	95.7	77.1
YOLOv5s	46.1	82.2	87.8	91.1	64.9	91.8	77.3
ES-Net [23]	74.1	60.9	82.5	95.8	94.3	67.2	79.1
DCC-CenterNet [23]	45.7	90.6	82.5	85.1	76.8	95.8	79.4
ST-YOLO [22]	54.6	83.0	84.7	89.2	73.2	97.0	80.3
YOLO-LFPD [24]	63.0	82.4	89.8	86.5	71.9	93.9	81.2
Ours	62.6	82.6	92.6	93.6	66.7	95.8	82.5

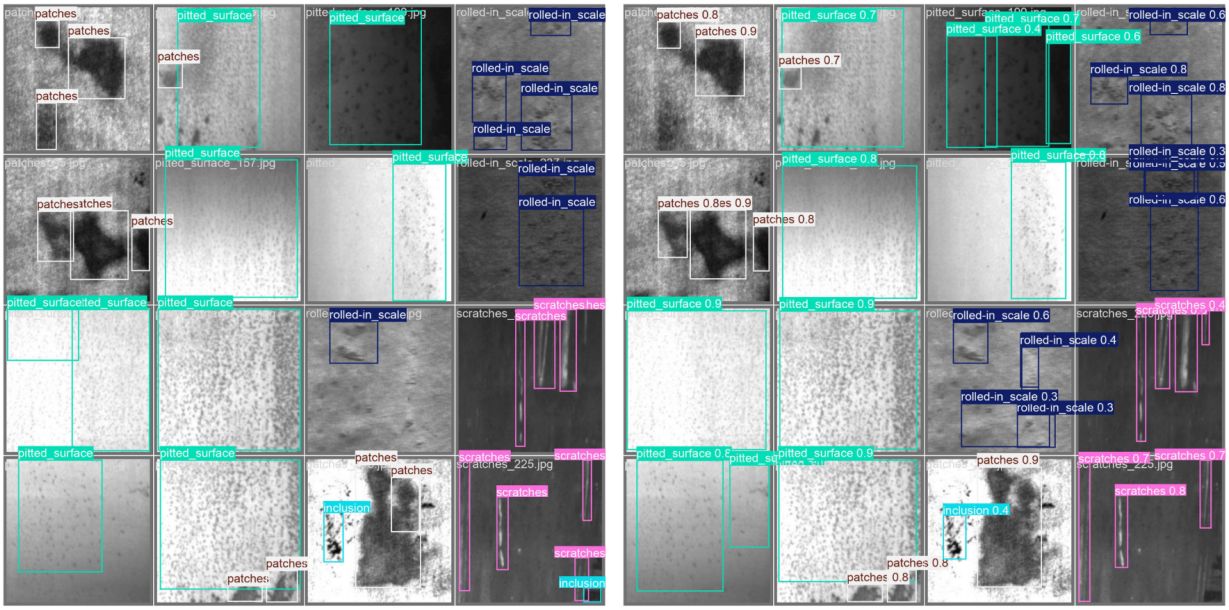

Fig. 5: Results Comparison

problem of low detection accuracy of traditional steel surface defects detection due to the variety of steel surface defects, complex morphology and background. We combine CARAFE module, Swin Transformer module and SEAttention module on the basis of YOLOv12 model. The experimental data proves that the improved model has satisfactorily improved the detection accuracy in dealing with the challenges of complex features that are easily interfered by the background environment and difficult to take into account the details and global information.In the future, we will further reduce the number of parameters of the network model, focus on improving the inference speed of the network, and explore the characterization ability of the model in the case of small samples, so as to further improve the robustness of the model and make it more resistant to interference in complex application scenarios.

REFERENCES

[1] X. Lv, F. Duan, J. Jiang, X. Fu, and L. Gan, "Deep metallic surface defect detection: The new benchmark and detection network," *Sensors*, vol. 20, no. 6, p. 1562, 2020.

[2] R. Mordia and A. K. Verma, "Visual techniques for defects detection in steel products: A comparative study," *Engineering Failure Analysis*, vol. 134, p. 106047, 2022.

[3] M. Duspara, B. Savković, B. Dudic, and A. Stoić, "Effective detection of the machinability of stainless steel from the aspect of the roughness of the machined surface," *Coatings*, vol. 13, no. 2, p. 447, 2023.

[4] T. Xian, H. Wei, and X. De, "A survey of surface defect detection methods based on deep learning [j/ol]," *Acta Automatic Sinica*, pp. 1–19, 2020.

[5] I. S. Ramírez, F. P. G. Márquez, and M. Papaelias, "Review on additive manufacturing and non-destructive testing," *Journal of Manufacturing Systems*, vol. 66, pp. 260–286, 2023.

[6] G. Wang, Q. Xiao, Z. Gao, W. Li, L. Jia, C. Liang, and X. Yu, "Multifrequency ac magnetic flux leakage testing for the detection of surface and backside defects in thick steel plates," *IEEE Magnetics Letters*, vol. 13, pp. 1–5, 2022.

[7] S. Xiong, Y. Tan, G. Wang, P. Yan, and X. Xiang, "Learning feature relationships in cnn model via relational embedding convolution layer," *Neural Networks*, vol. 179, pp. 106510–106510, 2024.

[8] Z. Zhao, B. Li, R. Dong, and P. Zhao, "A surface defect detection method based on positive samples," in *PRICAI 2018: Trends in Artificial Intelligence: 15th Pacific Rim International Conference on Artificial Intelligence, Nanjing, China, August 28–31, 2018, Proceedings, Part II 15*, pp. 473–481, Springer, 2018.

[9] Y. Wang, H. Wang, and Z. Xin, "Efficient detection model of steel strip surface defects based on yolo-v7," *Ieee Access*, vol. 10, pp. 133936–133944, 2022.

[10] Y. Zou and Y. Fan, "An infrared image defect detection method for steel based on regularized yolo," *Sensors*, vol. 24, no. 5, p. 1674, 2024.

[11] Z. Wang, L. Zhao, H. Li, X. Xue, and H. Liu, "Research on a metal surface defect detection algorithm based on dsl-yolo," *Sensors (Basel, Switzerland)*, vol. 24, no. 19, p. 6268, 2024.

[12] G. D. Raj and B. Prabadevi, "Enhancing surface detection: A comprehensive analysis of various yolo models," *Heliyon*, 2025.

[13] W. Xie, W. Ma, and X. Sun, "An efficient re-parameterization feature

pyramid network on yolov8 to the detection of steel surface defect," *Neurocomputing*, vol. 614, p. 128775, 2025.
[14] G. Zhang, S. Liu, S. Nie, and L. Yun, "Yolo-rdp: lightweight steel defect detection through improved yolov7-tiny and model pruning," *Symmetry*, vol. 16, no. 4, p. 458, 2024.
[15] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025.
[16] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13733–13742, 2021.
[17] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.
[18] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10428–10436, 2020.
[19] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.
[20] Z. G. Yolox, "Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, vol. 7, no. 10, 2021.
[21] M. de Jeu and X. Jiang, "Riesz representation theorems for positive linear operators," *Banach Journal of Mathematical Analysis*, vol. 16, no. 3, p. 44, 2022.
[22] H. Ma, Z. Zhang, and J. Zhao, "A novel st-yolo network for steel-surface-defect detection," *Sensors*, vol. 23, no. 22, p. 9152, 2023.
[23] X. Qian, X. Wang, S. Yang, and J. Lei, "Lff-yolo: A yolo algorithm with lightweight feature fusion network for multi-scale defect detection," *IEEE Access*, vol. 10, pp. 130339–130349, 2022.
[24] J. Lu, M. Zhu, K. Qin, and X. Ma, "Yolo-lfpd: A lightweight method for strip surface defect detection," *Biomimetics*, vol. 9, no. 10, p. 607, 2024.