Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería

SteelGuard-yolo: Steel surface defect detection network based on improved YOLOv5s

Zheng Zhou[1], Min Yuan[1]

1 Lanzhou University, Lanzhou, Gansu 730000, People's Republic of China

Abstract

Steel is playing an increasingly important role in industry, and the detection of defects on its surface is also of great significance. The complex and diverse defects on the steel surface bring great challenges to the detection. In this paper, we propose a SteelGuard-yolo based steel surface defect detection model, whose main role is to improve the existing algorithms for detecting steel surface defects. First, we design the C2f module with weight aggregation and introduce the BiFormer attention mechanism to improve the feature extraction capability of the model. Second, we design a new up-sampling structure and introduce the Multi-Scale Dilated Attention (MSDA) module to effectively improve the feature fusion capability of the model. Finally, we introduce the Simam attention mechanism and use EIoU as a new loss function to improve the robustness and accuracy of the model.SteelGuard-yolo has a powerful multi-scale feature fusion capability and achieves an ideal balance between latency and accuracy. The algorithm proposed in this paper is tested on the NEU-DET dataset and achieves an average accuracy of 69.0%, which compares favorably with most one- and two-stage detection algorithms.

OPEN ACCESS

Published: 10/06/2024

Accepted: 25/05/2024

Submitted: 15/05/2024

DOI: 10.23967/j.rimni.2024.05.011

Abstract: Steel is playing an increasingly important role in industry, and the detection of defects on its surface is also of great significance. The complex and diverse defects on the steel surface bring great challenges to the detection. In this paper, we propose a SteelGuard-yolo based steel surface defect detection model, whose main role is to improve the existing algorithms for detecting steel surface defects. First, we design the C2f module with weight aggregation and introduce the BiFormer attention mechanism to improve the feature extraction capability of the model. Second, we design a new up-sampling structure and introduce the Multi-Scale Dilated Attention (MSDA) module to effectively improve the feature fusion capability of the model. Finally, we introduce the Simam attention mechanism and use EIoU as a new loss function to improve the robustness and accuracy of the model.SteelGuard-yolo has a powerful multi-scale feature fusion capability and achieves an ideal balance between latency and accuracy. The algorithm proposed in this paper is tested on the NEU-DET dataset and achieves an average accuracy of 69.0%, which compares favorably with most one- and two-stage detection algorithms.

1. Introduction:

Steel plays an important role in many fields such as aviation, shipping, real estate and military industry. With the growing maturity of industrial development, the demand for steel in intelligent industrial manufacturing is also increasing. Different types of surface defects are easily produced when steel is processed by machines due to various factors such as external forces, equipment wear and tear, and processing techniques. These defects will adversely affect the appearance, performance and fatigue strength of the whole product [1]. The efficiency and accuracy of traditional inspection methods through manual inspection can no longer meet the inspection needs, and fast and accurate intelligent inspection is conducive to reducing the labor cost, and will also greatly reduce the loss rate of the product. Therefore, it is of great practical value and application significance to study an efficient and accurate detection algorithm for steel surface defects to effectively reduce production costs.

Target detection is a core research area in the field of computer vision, in which various types of machine learning (ML) and deep learning (DL) models are widely used to improve the performance of target detection and its related tasks, and target detection is mainly categorized into two different detection models, one-phase and two-phase.

Two-stage detection methods were the first to appear, and the main element is to generate anchor frames on the input image, followed by detecting the contents of the anchor frames and finally classifying them. Typical algorithms are R-CNN [2], Fast-RCNN [3], Faster-RCNN [4], and Grid-RCNN [5], etc. Ross Girshick et al. proposed R-CNN, which sets a set of anchors at each pixel, and RPN classifies and regresses all of these anchors, and then picks the proposals based on the classification confidence of the proposed top K proposals. Models such as Fast-RCNN and Faster-RCNN, which are improved on its basis, are considered to be the classical models in the second stage. In addition, for target recognition under different viewing angles, lighting and occlusion conditions, Anton Osokin et al. proposed Context-aware CNNs [6], which realized target recognition under different conditions. These models have good detection performance and all of them have achieved excellent results, but their anchor frame localization modules are very similar and there is still room for improvement. To address this problem, lu et al. proposed Grid-RCNN, which effectively utilizes explicit spatial representations to achieve high-quality localization. The two-stage model has higher accuracy but longer detection time, although the EfficientDet series model optimizes the detection time, but it also consumes more computational resources.

One-stage methods no longer need to generate anchor frames, but directly predict the whole image, obtaining an improvement

Correspondence: Zheng Zhou (13733033076@163.com), Min Yuan (yuanm@lzu.edu.cn).

in detection speed. The most typical one-stage target detection algorithms includeYOLO [7],SSD [8],SqueezeDet [9] and DetectNet [10], etc. The OverFeat [11]algorithm proposed by P Sermanet et al. is the basis of the first stage of target detection, which classifies images at different locations in a multi-scale region of the image in the form of a sliding window, as well as trains a regressor on the same convolutional layer to predict the location of the bounding box. On top of this the yolo series of algorithms are also the classic algorithms of the first stage.REDMON J proposed the YOLO (You Only Look Once) model, which has a strong generalization ability as well as adaptability. With the proposal of YOLO, various applications have begun to utilize YOLO for target detection and recognition in various contexts. Aiming at the deficiencies of YOLO family of models in network fusion, Wang et al. proposed gold-yolo [12],which improves on convolution and self-attention mechanisms, and employs Mae-style pre-training to allow the model to gain under unsupervised training. Aiming at the problems of poor performance of yolov2 backbone and underutilization of multi-scale regional features, Huang et al. proposed a DC-SPP-YOLO [13] based on dense connectivity (DC) and spatial pyramid pooling (SPP), which improves the target detection accuracy of YOLOv2 [14]. Aiming at the industrial scenarios where image background interference is large, defect categories are easily confused, defect scales vary greatly, and small defects are poorly detected, Guo et al. proposed MSFT-YOLO [15], which realizes the fusion of features at different scales and enhances the dynamic adjustment of the model to targets at different scales.The first-stage models, such as YOLO, SSD, and RetinaNet [16], are excellent in terms of speed and real-time performance but there is the problem of localization accuracy and relatively low detection accuracy for small targets. One-stage models such as YolovX have much lower accuracy than two-stage models, even though their detection speed is faster than that of two-stage models. Two-stage models such as Faster-Rcnn have higher accuracy than the one-stage detection model, but their computation amount and time are much higher than the one-stage model.

To address the above problems, this paper proposes a new defect detection algorithm SteelGuard-yolo,to realize the improvement of the detection accuracy of the one-stage model with reduced computation, the main contributions are:

(1) Feature extraction: in order to extract different levels of features more accurately, we replace the original C3 module with the improved C2f module with weight aggregation in the backbone part, which fuses the low-level feature map and the high-level feature map, and at the same time, we introduce the BiFormer (BRA) attention mechanism in front of the SPFF module, which enhances the model's perceptual ability and improves the accuracy of target detection and robustness.

(2) Feature fusion at different scales: in order to realize feature fusion at different scales, we design a new up-sampling method in the neck: the mutual fusion of features at different scales is realized by the concatenation and then summation of bilinear interpolation and inverse convolution. We replace the C3 module in the neck with Multi-Scale Dilated Attention (MSDA) module, which suppresses the background sexual information and highlights the perceptual region, while aggregating the semantic information at all scales of the attended perceptual field and reduces the redundancy of the self-attention mechanism.

(3) Detection header: In order to further improve the model accuracy, we redesigned the detection header: we introduced the Simam attention mechanism in front of the original Conv module, which improves the efficiency of the model with the same parameters.

(4) Loss function: in order to solve the problem of imbalance between difficult and easy samples, we replace the original loss function with EIOU, which is easier to obtain more accurate target localization on various scales.

The experimental results showed that the MAP50 of SteelGuard-yolo on the NEU-DET dataset was enhanced by 0.082 compared to the original Yolov5s.

The remainder of this paper is organized as follows.Section 2 Section reviews related work on target detection.Section 3 Section describes the proposed method and the three key modules in detail. Sect.4 Section presents relevant experimental results, which are compared with state-of-the-art algorithms to verify the effectiveness of the proposed method in this paper. Finally, we present in Section 5 Section 5 summarizes the experiments and gives an outlook.

2. Related work

The YOLO (You Only Look Once) family of algorithms is a popular real-time target detection framework that has gained popularity in computer vision for its speed and accuracy. Since the introduction of YOLOv1, the series has gone through several iterations, with each version optimized in terms of performance, speed, and model complexity. The core idea of the YOLO algorithm is to transform the target detection problem into a single regression problem by predicting bounding box and category probabilities directly on the full graph, thus avoiding the region proposal stage in traditional target detection methods. Among many models YOLOv5s is easier to be embedded in resource-constrained platforms while maintaining high target detection accuracy. Based on this, in this paper, YOLOv5s without pre-training weights is chosen as the base model and improved on it to propose SteelGuard-yolo.

2.1 The YOLOv5 model

In the field of target detection, the backbone network of YOLOv5 model plays a crucial role, and its main task is to perform deep feature extraction on the input image. It belongs to the category of one-stage model, which takes into account both speed and accuracy, and is one of the mainstream models at present. Some new techniques are used in YOLOv5, such as adaptive training data enhancement, multi-scale training, multi-scale prediction of the initial detection layer, etc., which make it faster and more accurate.The network results of YOLOv5s are shown in Fig. 1.

The YOLOv5 network structure consists of four parts: input (Input), feature extraction network (Backbone), feature fusion network (Neck), and output (Head).

The input side preprocesses the image, including image flipping and adaptive anchor frame calculation.

The feature extraction network contains a 3-layer structure consisting of the CBS module, the C3 module and the SPPF module. The Conv module is the basic component of the convolutional neural network, which consists of the convolutional layer, the BN layer and the activation function. Among them, the convolutional layer is responsible for extracting the local spatial information of the input features, the BN layer performs the normalization of the feature values after the convolutional layer, and the activation function introduces the nonlinear transformation capability for the neural network. The C3 module is the core component of the YOLOv5 network, which effectively improves the feature extraction capability by increasing the depth and the sensory field of the network. The SPPF module is a pooling module, and its s main role is to realize the spatial invariance and positional invariance of the

input data so as to improve the recognition ability of the neural network.

The feature fusion network adopts FPN and PAN structures, the FPN structure passes the semantic information of the layer feature map to the shallow features from top down, and the PAN structure passes the position information of the shallow feature map to the deep features, thus realizing the multi-scale feature fusion.

The output, on the other hand, contains three detection layers of different sizes to classify and predict the fused features.

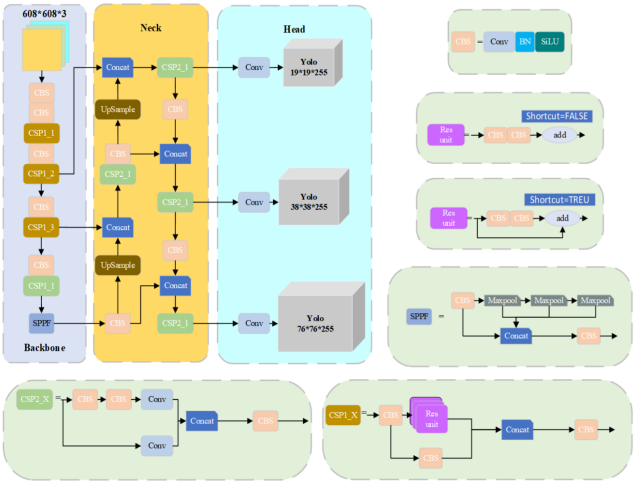

Fig.1. YOLOv5s network architecture.

YOLOv5s is a target detection algorithm based on a lightweight convolutional neural network. Its network architecture consists of three main components; Backbone, Neck and Head.The Backbone network utilizes CSP-Darknet53 for extracting features from the input image. The Neck network integrates the features to improve the detection accuracy. The Head network is used to predict the location and class of the target.YOLOv5 achieves high-performance target detection through key techniques such as multi-scale detection, FPN network, Focal Loss loss function, and non-extremely large value suppression.

2.2. Attention mechanisms

Attention Mechanism is a strategy to simulate human cognitive attention in artificial neural networks, which allows the model to pay more attention to the important parts of the input information Attention Mechanism selectively extracts features of interest in the model in only two steps by fusing the CNN with the attention module. Step 1 adds the attention mechanism branch into the initial network structure for weight learning. Step 2 applies the weights learned in Step 1 to the feature maps output by the CNN. During the training process, the model assigns different weights according to the according to the importance of the feature map in turn. The more important the feature is, the more weights the model assigns to it. The model will focus more on high weighted effective features and low weighted features as well as ineffective features will be suppressed. This process can be described by the following equation:

$$Attention = F(f(x), x)$$

where $f(x)$ denotes the process of learning weights by the attention mechanism, and $F(f(x), x)$ denotes the process of processing the input features according to the weights.

Based on the network structure and region of action, attention mechanisms can be categorized into channel attention mechanisms, spatial attention mechanisms, and hybrid domain attention mechanisms. Hybrid attention mechanisms include Multidimensional collaborative attention (MCA) [17], Bi-Level Routing Attention (biformer) [18], and Multi-Scale Dilated Attention (MSDA) [19].

Multi-Scale Dilated Attention (MSDA) is a typical example of a mixed-domain attention mechanism.

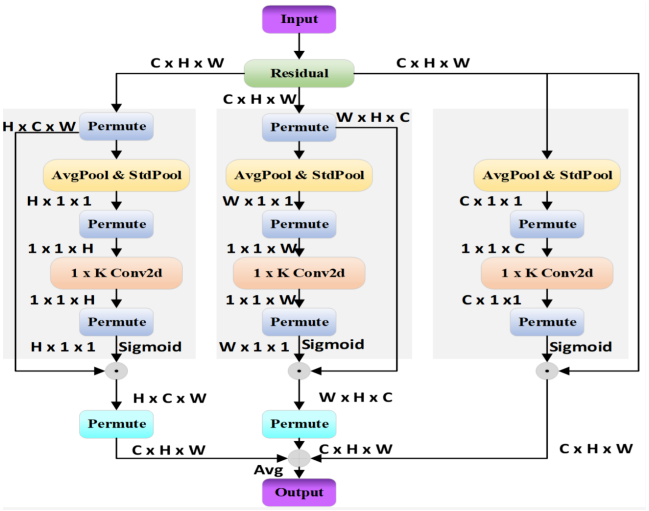

Fig. 2. MCA network structure.

MCA is a statistical moment-based channel attention network. It captures global spatial context through the Ex-tensive Moment Aggregation (EMA) mechanism and efficiently integrates multi-level moment-based information through the Cross Moment Convolution (CMC) module.MCA is lightweight and easy to integrate into various neural network architectures, and it has achieved leading results in image classification, target detection, and instance segmentation tasks, achieving leading results that outperform existing channel attention methods

where C, W and H represent the number, width and height of channels in the feature map, respectively. Here, AvgPool and StdPool represent the global average pooling and theglobal standardized difference pooling, respectively. ⊙ stands for broadcast element-wise multiplication, and ⊕ stands for broadcast element-wise summation. three branches are used to capture the interactions between different dimensions and channels. A substitution operation is used in the first two branches to capture the remote dependence between the channel dimension and any of the spatial dimensions. The final branch aggregates the outputs of all three branches in the integration phase.

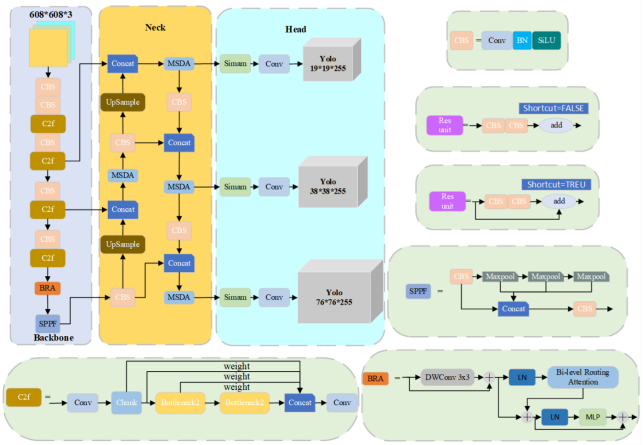

Fig. 3. SteelGuard-yolo network structure.

We made a series of improvements to YOLOv5s: the original C3 convolutional module was replaced with the improved C2f in Backbone, and the Biformer attention mechanism was introduced; a module with both inverse convolution and bilinear interpolation upsampling was designed in Neck, and the original C3 module was replaced with an MSDA module; a convolutional module was added in Head before the Simam attention mechanism was added to improve the detection accuracy and efficiency; the loss function was replaced with EIOU to obtain more accurate target localization.

3. Methodologies

In this section we describe in detail our proposed model SteelGuard-yolo for steel surface defect detection.Figure 3 shows the structure of our proposed SteelGuard-yolo network. The original YOLOv5s model is not effective in detecting multi-scale features due to the large variation of steel surface defects and the different proportions of different defects in the whole image. In addition, the downsampling multiplier of YOLOv5s is large, and it is difficult for the deeper feature maps to receive the feature information of steel surface defects. Therefore, it is especially critical to improve the detection accuracy and robustness of the original model.

First, in Backbone, we replace the original convolutional module C3 with an improved C2f with weight aggregation features, and add an attention mechanism Biformer in front of the Spatial pyramid pooling-based fusion (SPPF) module to enhance the ability of acquiring input features. Second, we design a module with both inverse convolution and bilinear interpolation upsampling in Neck, and we replace the original C3 module in Neck with an MSDA module to enhance the feature aggregation capability of the model. Next we add the attention mechanism Simam in front of the convolution module (conv) in Head to improve the detection accuracy and efficiency. Finally, we replace the loss function with EIOU to obtain more accurate target localization.

3.1. Backbone optimization

For each steel surface defect image, its high-frequency component constitutes the edges and contours of the steel surface defects and represents the fast image localization capability. Meanwhile, the low-frequency components represent regions with smooth changes in their intensity, mainly image patches with similar visual patterns. In steel surface defect images, the low-frequency modules of the images are basically the same, so we replace the original C3 module with the C2f module with weight aggregation, and incorporate an attention mechanism to improve the model's detection attention and accuracy for steel surface defects.

The detailed implementation of C2f with weight aggregation is shown in Fig. 3. When an image with height H, width W, and channels C enters the C2f module, it will be divided into two HW0.5C images. One is left unprocessed and the other is passed into bottleneck2 for feature fusion operation. For feature fusion, we have taken the feature fusion operation with weights as a way to highlight the features that need attention. The operation of Biformer is also shown in detail in Fig. 3. The relative position information is first implicitly encoded using a 3 × 3 convolution. Then, cross-positional relationship modeling and position-by-position embedding are performed sequentially using the BRA module and a 2-layer MLP module with an expansion of e. The BRA module is used to model the relative positional information.

3.2. Neck Optimization

In Neck, we design a module with both inverse convolution and bilinear interpolation up-sampling, and we replace the original C3 module in Neck with MSDA module to enhance the feature aggregation capability of the model. Feature up-sampling is a very important part of deep learning and neural networks. Feature maps of different resolutions are matched based on high-resolution supervision. up-sampling in YOLOv5s uses the nearest neighbor interpolation algorithm by default, which fills new pixel positions by copying the nearest pixel values.This means that there are significant discontinuities in the gray values in the sampled image, resulting in a large loss of image quality. This can manifest itself in the form of noticeable mosaic and jaggedness . This method is overly concerned with the speed of the operation and ignores the accuracy and effectiveness of the up-sampling result. Our design with both inverse convolution and bilinear interpolation up-sampling module effectively remedies this shortcoming. Bilinear interpolation takes into account the weights of the four pixels around a pixel point, which can effectively compensate for mosaic and jaggedness. Inverse convolution enables the network to automatically learn the appropriate upsampling weights for a particular task. The combination of the two effectively improves the smoothness and accuracy of the upsampling results. The original YOLOv5s is not as good as the original YOLOv5s in detecting small targets, so we replace the C3 module in Neck with the MSDA module to enhance the feature aggregation ability of the model.The MSDA module improves the detection accuracy of small targets by generating larger scale feature maps to differentiate the fine features of small targets. Meanwhile, the MSDA module adopts sliding window feature extraction, which effectively reduces the computational requirements and the number of parameters. Finally, the MSDA module introduces a global attention mechanism that combines channel information with global information to create a weighted feature map. This helps to highlight the attributes of the object of interest while effectively ignoring irrelevant details.

3.3. Detection head optimization

The original detection head of YOLOv5 has a large sensing field, and the detection head needs to separate each target accurately, and the detection ability of the model will be reduced when the target is small or dense. The defects on the steel surface are sometimes very dense and small, so YOLOv5 is not good at detecting these defects. Therefore, we introduced the SimAM attention mechanism in front of the original Conv module, which improves the efficiency of the model while keeping the parameters unchanged. SimAM attention mechanism can express the features more finely, so as to better

extract the key target information without introducing too many parameters. In complex contexts, SimAM can better extract target features, thus enhancing the model's ability to perceive the target.

3.4. Loss function optimization

The IoU loss function is the default loss function of YOLOv5. IoU loss can accurately measure the degree of overlap between the predicted and real frames, but there is no specific applicable scenario. It performs well in general, but it is not good enough for steel defect target detection, and the problem of missed detection and false detection often occurs. Therefore, we replace the IoU loss function with the EIoU loss function.The EIoU loss function can describe the target localization objectives more effectively by explicitly measuring the differences in three geometric factors of the bounding box (overlap area, center point, and edge length). This helps the model to converge faster. Also the EIoU loss function takes into account multiple geometric factors of the bounding box, which improves the accuracy of localization.

4. Experimental analysis and results

This section first describes the dataset used, adds more experimental details and evaluation metrics, and finally analyzes and evaluates the experimental results.

4.1. data sets

The use of high-quality and large-scale datasets is particularly important to improve the generalization ability of the model. In this experiment, the NEU surface defect database (NEU-DET) [20-22], which was established by He et al. in 2020, is used, which collects six typical surface defects of hot rolled strip steel, namely Rolled-in scale (RS), Patches (Pa), Crazing (Cr), Pitted surface (PS), Inclusions (In) and Pitted surface (Sc). The database consists of 1,800 grayscale images: 300 samples of each of the six different typical surface defects, while a single image may contain more than one defect, making the dataset suitable for use as a steel defect detection dataset.

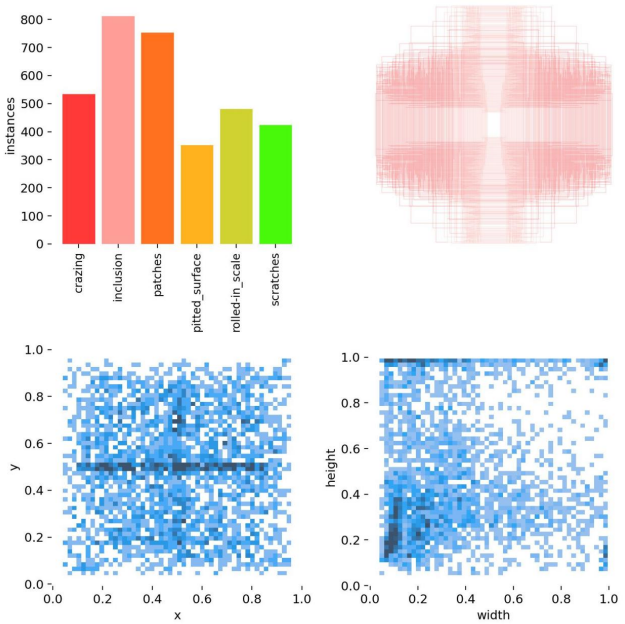

Fig. 4.Data details of NEU-DET dataset.

The NEU-DET dataset is a public dataset for surface defect detection on steel plates and contains a total of 1800 images. The dataset covers six types of steel plate defects, namely 'crazing', 'patches', 'inclusion', 'pitted_surface', 'rolled-in_scale' and 'scratches'. A visual representation and distribution of the various surface defects and the frequency of occurrence are shown.

4.2. Experimental parameters

The system used in this experiment is Windows and the GPU is NVIDIA GeForce GTX 1650. a cross-entropy loss function is used for training. The optimizer is AdaBound, the learning rate is set to 0.01, the weight decay parameter is 0.005, the mean values between the high and low importance groups are 0.249125 and 0.230846 respectively, and the number of training rounds is 50.

4.3. Comparison with mainstream models

In order to verify the superiority of our model, we compare it with current mainstream models in both one-stage and two-stage categories, which include ATSS, CASCADE-RCNN , FASTER-RCNN , SSD300, Retinanet, and YOLOVX. Table 1 lists the comparison of our proposed SteelGuard-yolo with these mainstream algorithms. It can be seen that our algorithm outperforms most of the algorithms for comparison, with Map50 reaching 0.690, which is an improvement of 0.072 over the initial YOLOv5s. We made predictions using a one- and two-stage typical model for target detection and our model, respectively, and Fig. 5 shows the visualization results, where (a) a one-stage typical model of YolovX is used, (b) a two-stage typical model of Faster-Rcnn, (c) is ssdlite, (d) is atss, and (e) is our proposed SteelGuard-yolo.

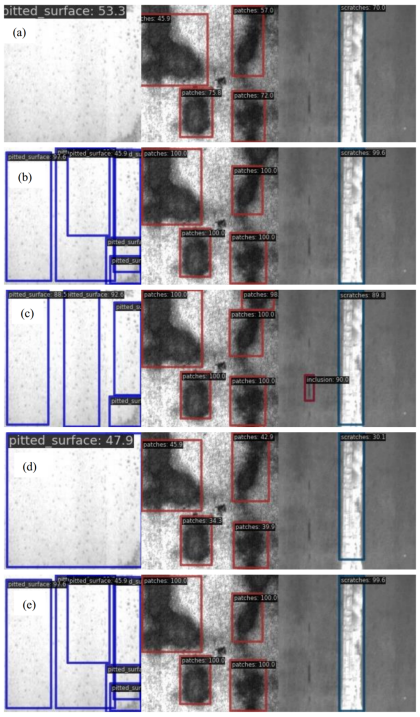

Fig. 5. Comparison of the detection effectiveness of the mainstream algorithms in phase I and II and SteelGuard-yolo on the NEU-DET dataset Example. (a) YolovX (b).Faster-Rcnn (c) ssdlite (d) atss (e) SteelGuard-yolo. the actual detection accuracy of our proposed SteelGuard-yolo in the figure outperforms the other algorithms from (a) to (d).

Table 1. Comparison of SteelGuard-yolo's detection effect with mainstream models on NEU-DET dataset

Models	Map50	Crazing	Inclusions	Patches	Pitted surface	Rolled-in scale	Scratches
ATSS	0.458	0.312	0.256	0.712	0.648	0.563	0.258
FASTER-RCNN	0.745	0.968	0.415	0.909	0.998	0.909	0.273
SSD300	0.973	0.945	0.982	1.000	0.981	1.000	0.930
Retinanet	0.193	0.204	0.029	0.627	0.057	0.221	0.020
YOLOVX	0.397	0.314	0.226	0.611	0.478	0.511	0.242
YOLOV5s	0.618	0.269	0.739	0.839	0.681	0.475	0.704
SteelGuard-yolo	0.690	0.408	0.765	0.847	0.761	0.584	0.772

4.4. Ablation experiments

To further validate the effectiveness of our proposed module, we conduct extensive ablation experiments for NEU-DET. Following the same experimental protocol, we take YOLOv5s without pre-trained weights as the baseline model, and then add components at each position on top of it. As can be seen from Table 2, compared to the original YOLOv5s model, our proposed SteelGuard-yolo improves the APs in all categories, among which Crazing,Pitted surface,Rolled-in scale, and Scratches are significantly improved, with 13.9%, 8%, 10.9%, and 6.8% respectively.

Table 2. Comparison of detection accuracy of different structures.Backbone denotes the improvement done in Backbone part, Neck denotes the improvement done in Neck part, Head denotes the improvement done in Head part and EIOU denotes the replacement of the loss function with EIOU.

Models	Map50	Crazing	Inclusions	Patches	Pitted surface	Rolled-in scale	Scratches
YOLOV5s	0.618	0.269	0.739	0.839	0.681	0.475	0.704
YOLOV5s-Backbone-EIOU	0.669	0.330	0.774	0.882	0.755	0.485	0.784
YOLOV5s-Head-EIOU	0.679	0.335	0.786	0.889	0.783	0.481	0.799
YOLOV5s-Backbone-Neck-Head	0.676	0.322	0.753	0.864	0.787	0.521	0.809
YOLOV5s-Backbone-Neck-EIOU	0.667	0.330	0.713	0.660	0.764	0.516	0.814
YOLOV5s-Backbone-Head-EIOU	0.687	0.358	0.768	0.853	0.817	0.533	0.793
YOLOV5s-Neck-Head-EIOU	0.686	0.361	0.773	0.881	0.790	0.528	0.785
YOLOV5s-Backbone-Neck-Head-EIOU	0.690	0.408	0.765	0.847	0.761	0.584	0.772

5. Conclusion

In this paper, an algorithm for detecting defects on steel surfaces is proposed. By replacing the original C3 module in Backbone with an improved C2f module with weight aggregation, and introducing the BiFormer attention mechanism in front of the SPFF module to enhance the perceptual ability of the model, the accuracy and robustness of target detection are improved. In Neck, the up-sampling method is designed to combine the parallelism of bilinear interpolation and inverse convolution, which realizes the more accurate mutual fusion of different scale features. The C3 module of Neck is replaced by the MSDA module, which suppresses the background information and emphasizes the perceptual region. The Simam attention mechanism is introduced in front of the Conv module of the detection head to better extract the key target information without introducing too many parameters. Finally, the original loss function is replaced with EIOU to improve the localization accuracy. In our future work, we will further develop the lightweight backbone feature extraction network and new feature fusion methods to simplify the network structure architecture and achieve an effective balance between high speed and high accuracy.

REFERENCES

[1] Tang B, Chen L, Sun W, et al. Review of surface defect detection of steel products based on machine vision. IET Image Processing, 2023, 17(2): 303-322.

[2] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[3] Girshick R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[4] Ren S, He K, Girshick R, et al. Faster R-CNNn: Towards real-time object detection with region proposal networks. Ieee Transactions on Pattern Analysis and Machine Intelligence, 2017, 39, 1137-1149.

[5] Lu X, Li B, Yue Y, et al. Grid r-cnn. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 7363-7372.

[6] Shaban M, Awan R, Fraz M M, et al. Context-aware convolutional neural network for grading of colorectal cancer histology images. IEEE transactions on medical imaging, 2020, 39(7): 2395-2405.

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[8] Liu W, Anguelov D, Erhan D, 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, 2016: 21-37.

[9] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. arXiv:1602.07360, 2016.

[10] Li Z, Peng C, Yu G, et al. Detnet: a backbone network for object detection. arXiv:1804.06215, 2018.

[11] Sermanet P, Eigen D, Zhang X, et al. Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv:1312.6229, 2013.

[12] Wang C, He W, Nie Y, et al. Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. Advances in Neural Information Processing Systems, 2024, 36, 1-10.

[13] Huang Z, Wang J, Fu X, et al. DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection. Information Sciences, 2020, 522: 241-258.

[14] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. 7263-7271.

[15] Guo Z, Wang C, Yang G, et al. Msft-yolo: Improved yolov5 based on transformer for detecting defects of steel surface. Sensors, 2022, 22(9): 3467.

[16] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[17] Yu Y, Zhang Y, Cheng Z, et al. MCA: Multidimensional collaborative attention in deep convolutional neural networks for image recognition. Engineering Applications of Artificial Intelligence, 2023, 126: 107079.

[18] Zhu L, Wang X, Ke Z, et al. Biformer: Vision transformer with bi-level routing attention. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 10323-10333.

[19] Jiao J, Tang Y M, Lin K Y, et al. Dilateformer: multi-scale dilated transformer for visual recognition. IEEE Transactions on Multimedia, 2023.

[20] Yanqi Bao, Kechen Song, Jie Liu, Yanyan Wang, Yunhui Yan, Han Yu, Xingjie Li, "Triplet-Graph Reasoning Network for Few-shot Metal Generic Surface Defect Segmentation," IEEE Transactions on Instrumentation and Measurement, 2021, 70, 3083561.

[21] K. Song and Y. Yan, "A Noise Robust Method Based on Completed Local Binary Patterns for Hot-Rolled Steel Strip Surface Defects, " Applied Surface Science, 2013, 285, 858-864.

[22] Yu He, Kechen Song, Qinggang Meng, Yunhui Yan, "An End-to-end Steel Surface Defect Detection Approach via Fusing Multiple Hierarchical Features," IEEE Transactions on Instrumentation and Measurement, 2020, 69(4), 1493-1504.