Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 7, No. 2, 2023

Surface Defect Detection of Strip Steel Based on YOLOv5

Dewei Wang, Xiaofang Liu*

School of Computer Science and Engineering, Sichuan University of Computer Science and Engineering, Yibin, 644002, China
* Corresponding author

Abstract: Given the various kinds of surface defects and the lack of obvious features, which lead to detection error and leakage detection, a method of strip surface defects detection with a decoupling head, YOLOv5s, is proposed. Firstly, the K-means + + method was utilized to relocate the anchor frames to produce an optimised coupling between the transcendental and the real frame. SimAM's attention free mechanism was incorporated into the Neck network to improve the performance of the model. Finally, the head of the network is replaced by the decoupling head, which separates classified tasks from regressive tasks, thus enhancing convergence and recognition accuracy. The results indicate that the average precision of the proposed algorithm is 80.7% on the NEU-DET dataset, an increase of 3.2% compared to YOLOv5s, and a transfer frame number of FPS of 50 per second, which balanced detection accuracy and operational efficiency. The increased accuracy of detection as compared to other techniques meets the requirements of precision and timeliness.

Keywords: Surface defect detection, SimAM without attention span, YOLOv5.

1. Introduction

In the field of computer vision, target detection is a challenging task that aims to determine the location of a specific target in a given image and to correctly identify and categorize it for in-depth analysis and understanding of the image content by using specific algorithmic techniques. At present, the dimensional accuracy of strip steel products has basically reached the needs of industrial production, but the issue of strip steel surface quality still needs further improvement. Strip steel in the manufacturing process, subject to the influence of raw material impurities factors, its surface generated cracking, plaque, pitting surface, rolling oxide, scratch, and other defects, will not only affect the appearance of the steel surface but also seriously reduce the corrosion resistance, high-temperature resistance and fatigue strength of steel. Therefore, the detection of strip steel surface defects is important to improve the quality of product production.

In recent years, target detection and image segmentation based on deep learning have gradually emerged. Target detection algorithms can be divided into two categories: one is based on a region-based convolutional neural network (R-CNN), and the other is based on a single stage. R-CNN uses a two-stage approach for target detection. Specifically, the first step is the extraction of candidate frames in the image, which cover all possible objects in the image. The second step is to classify and regress the candidate frames before finally producing results. commonly used models for R-CNN include Faster R-CNN and Mask R-CNN, among others. In contrast, single-stage target detection algorithms process candidate frames directly, without dealing with complex upstream and downstream work, and get the aim frame classification and regression results in a single forward inference, thus, such algorithms are usually simpler and more efficient than R-CNNs. Commonly used models for single-stage target detection algorithms include the YOLO family and SSD (Single Shot multi-box Detector) , among others. In general, different target detection algorithms are suitable for different scenarios and tasks. The single-stage approach has the advantage of faster model inference and easy model deployment, while the two-stage approach simply lies in the relatively high detection accuracy.

2. YOLOv5 Algorithm

The YOLO family of algorithms is a single-stage based target detection algorithm. The algorithm uses a regression model to tackle the target detection problem and aims to achieve accurate and fast determination of the location and class of targets through hierarchical feature extraction and fusion of feature maps. Through continuous innovation and improvement, the algorithm has evolved to the YOLOv5 version, which is the most widely used and has achieved dynamic equilibrium in both detection speed and accuracy.

YOLOv5s is a target detection model whose structure consists of four main parts: the input, Backbone network, Neck feature fusion network, and Head detection head, where the input requires pre-processing of the image and scaling the image according to the size of the network input, before finally performing the normalization operation. the structure of the YOLOv5s network is shown in Figure 1. The algorithm uses Mosaic to enhance the images in the training phase, by cropping and rotating four randomly selected images and stitching them together to a specified resolution size to achieve data enhancement; in addition, before the images are input to the network in bulk, the real frames to which the dataset belongs are clustered, and then the anchor frames are clustered using K-means to achieve backpropagation by adaptively calculating the anchor frame and the real frame difference between the anchor frame and the rear frame to achieve backpropagation.

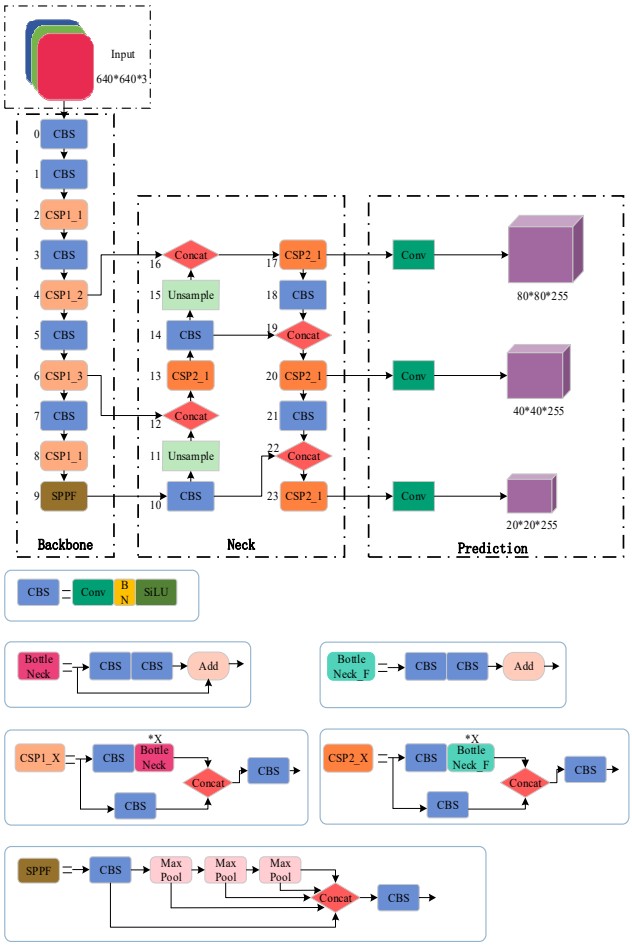

Figure 1. YOLOv5 network structure diagram

The Backbone network consists of a series of convolutional neural networks used to extract image features, mainly consisting of CBS, C3, and SPPF structures. YOLOv5 in version 7.0 replaces the Focus module with a convolutional layer of size 6×6 for the first layer of the network, and is an optimization of the existing algorithm, using a convolutional layer of size The use of a 6x6 convolutional layer is more efficient than using the Focus module.

The Neck section uses a "double tower structure". It consists of a Feature Pyramid Network (FPN) and a Path Aggregation Network (PAN). localization features to a higher level. The combination of the two further enhances the network's ability to fuse features and obtain richer feature information.

The head is the network detection head section of YOLOv5, which contains three Detect detectors. Firstly, for the input image with a resolution of 640×640, the feature extraction network is used to extract features from the input image to obtain three types of feature maps of different sizes, which are used to predict large, medium, and small targets respectively. A diagram of the prediction box is shown in Figure 2. The small black grid (C_x, C_y) is the standard aiming frame, the blue part is the prediction frame, and the dotted line is the a priori frame to be adjusted, whose height and width are p_h and p_w, and the offset of the prediction with respect to the anchor frame is t_w and t_h, from equation (1), we can obtain the prediction frame, b_w, b_h, b_y and b_x.

$$\begin{cases} b_x = 2*\sigma(t_x) - 0.5 \\ b_y = 2*\sigma(t_y) - 0.5 \\ b_w = 4*\sigma(t_w)^2 * p_w \\ b_h = 4*\sigma(t_h)^2 * p_h \end{cases} \quad (1)$$

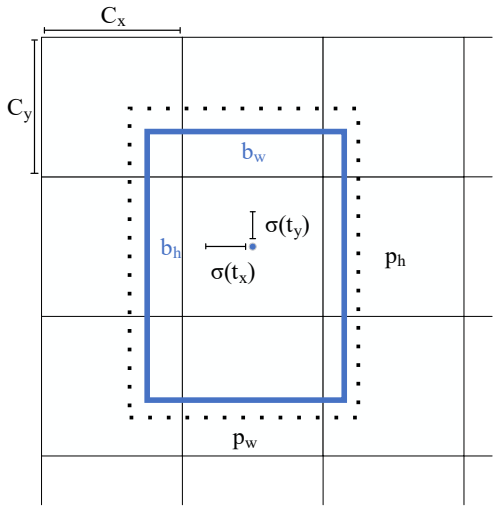

Figure 2. Prediction box diagram

3. Improved to the YOLOv5 Algorithm

(1) Optimal design of the a priori frame

The selection of suitable prior frames plays a key role in improving the training effect of the network. The YOLOv5s network model uses K-means clustering and uses three types of aiming frames to predict three types of detection heads according to three sizes: large, medium, and small. The method is based on the COCO dataset for clustering, and although there is good generalisability in common target detection tasks, there is still some error in using the method for small target identification in complex scenarios, which affects the detection accuracy of the model. Therefore, in order to improve the accuracy of small target detection of strip defects and reduce the error caused by the a priori frame size, the dataset was re-clustered using K-means++, and the clustering results are shown in Table 1.

Table 1. K-means++ generate a priori boxes

特征图	感受野	Achor
8×8	大	(165,85), (91,189), (217,221)
16×16	中	(59,51), (72,95), (209,33)
32×32	小	(23,49), (30,95), (59,51)

The initial points of the K-means++ algorithm are selected randomly from the whole data set, thus jumping out of the initial clusters, which makes the algorithm have a high probability of jumping out of the local optimal solution, thus obtaining the global optimal solution during the iterative process, and its specific computational steps are as follows:

1) Given P candidate frames;

2) Randomly pick a candidate frame as the center of the first cluster;

3) From the remaining P 1 candidate boxes, select the candidate box that is farthest from the first cluster center D

(A, B) is the largest), and this candidate box is used as the second cluster center.

4) Repeat the above steps, and finally select Q cluster centers from P candidate boxes;

5) Replace the initial point of the K-means algorithm with the selected Q cluster cores, and select Q anchor boxes according to the algorithm process.

Where the expression of D(A, B) is :

$$D_{(A,B)} = d(A,B) = 1 - IOU = 1 - \frac{A \cap B}{A \cup B} \tag{2}$$

(2) Introduction of SimAM's non-referential attention mechanism

Since strip defects exist in images with low pixels and are prone to information loss, the SimAM non-parametric attention mechanism is introduced into the YOLOv5s neural network model to enhance the attention of the detected object and improve the target detection accuracy by extracting feature information in the image. SimAM is a novel non-parametric 3D attention module, and its research is based on the human brain attention mechanism characteristics. The module adopts feature maps as an important means to assign uniform weights to 3D attention and aims to enhance the feature extraction capability of the model. Unlike traditional attention modules, SimAM does not introduce additional parameters and has the advantage of being lightweight compared to existing channel and null-field attention modules . The distribution of assigned 3D attention weights is shown in Figure 3. SimAM gives a measure of linear differentiability among neurons to evaluate the importance of each neuron.

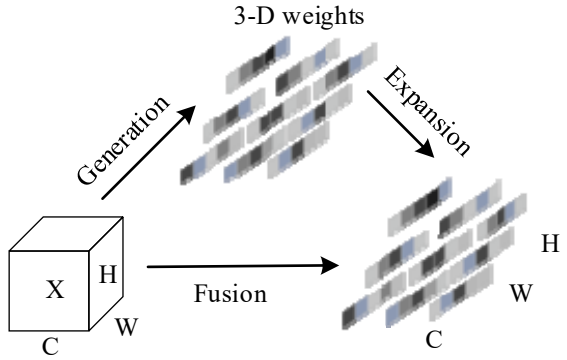

Figure 3. 3D attention weighting

Simulating the neuronal properties, the final energy function is defined as:

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(w_t x_i + b_t))^2 + (1-(w_t t + b_t))^2 + \lambda w_t^2 \tag{3}$$

t and x_i are the target neurons and other neurons of the input features, respectively, w_t and b_t are the weights and deviations of the linear variation of a neuron, and i are indexed in the spatial dimension. $M = H \times W$ is the number of all neurons on a channel. The minimum energy equation can be found by taking the partial derivatives of w_t and b_t substituting the original energy function.

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t-\hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{4}$$

As its energy decreases, the linear separation between the neuron t and other neurons increases, and its importance increases. Therefore, the importance of a neuron can be expressed by $1/e_t^*$ and finally its features are augmented to obtain the processed feature tensor $\hat{X}$.

$$\hat{X} = sigmoid(\frac{1}{E}) \text{ Z } X \tag{5}$$

where E is the set of all channels and spatial dimensions in e_t^*.

(3) Decoupling headers design

From YOLOv6, it is known that the decoupled head structure can take into account the difference in the content of target detection and semantic segmentation, where target detection focuses on the edge information of the target, while semantic segmentation focuses on understanding the pixel content of the object. Therefore, an efficient decoupled detection head with a hybrid channel strategy can further reduce the computational cost, achieve lower inference latency, train and adjust the network model faster in order to make it converge to the optimal state faster and improve the accuracy and precision of the model recognition, which effectively improves the strip defect recognition. The structure diagram of the decoupled head is shown in Figure 4.

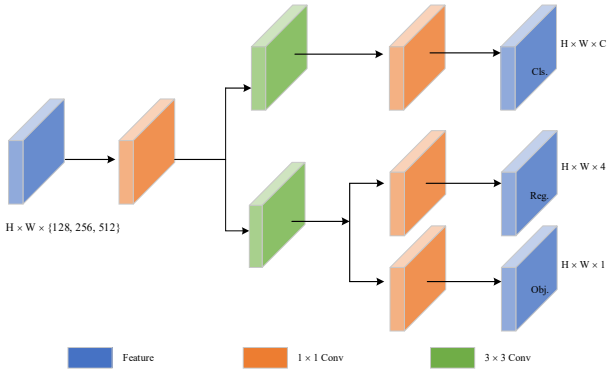

Figure 4. Decoupled head structure diagram

4. Experimental Results and Analysis

(1) Dataset

The NEU-DET dataset used in this study is published by Northeastern University, which is a classical image data set on surface defects of hot-rolled strip steel, containing 1800 images, and the images are divided into 6 broad categories, each corresponding to one type of surface defects, and each broad category includes 300 images. The image size is 200 × 200 to ensure the consistency and comparability of the data. In this study, we used the following six surface defect types: Crazing (Cr), Inclusion (In), Patches (Pa), Pitted Surface (Ps),

Rolled-in Scale (Rs), and Scratches (Sc). The 1800 defect images were randomly divided in the ratio of 8:2, the training set reached 1440, and the remaining 360 were used as the test set. Each class of strip defects is shown in Figure 5.

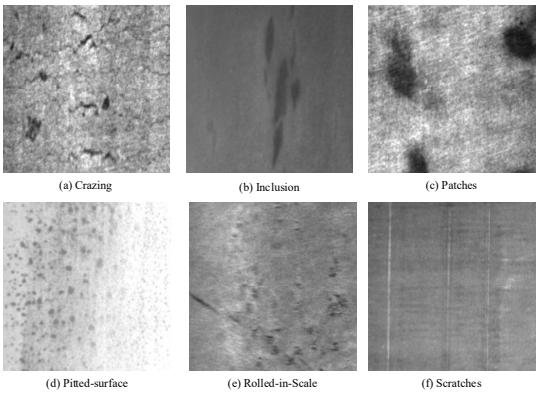

Figure 5. Various strip defects diagrams

(2) Experimental configuration

The experiments use Windows 10 operating system with NVIDIA GeForce RTX3080 GPU and Python 3.8 programming language, based on Pytorch 1.13.1 and CUDA 11.6 deep learning framework. The experimental parameters are configured as shown in Table 2.

Table 2. Experimental parameter configurations

Parameter name	Parameter value
Image size	256x256
Initial learning rate	0.001
Momentum	0.937
Weight decay	0.0005
Batch processing	32
iterations	300

(3) Experimental procedure

To detect the accuracy of strip defects, YOLOv5s is improved. By using the K-means++ algorithm to cluster anchor frames (anchor), the affiliation of the anchor frames clustered by each real frame is regained, and then the category of these real frames is determined to achieve the purpose of clustering aiming frames; in addition, in order to further improve the accuracy of defect detection, for the situation that the target pixels of strip steel defects are low and the information is easily lost, the SimAM non-parametric attention is embedded in the Neck network part mechanism to capture more local information in the image and enhance the extraction of image features without introducing an additional number of parameters to make it more focused on the identification of defects, thus improving the detection accuracy. Finally, in order to overcome the problem of classification and regression conflicts of the defect image output variables, the network head is replaced with a decoupled detection head, which not only increases the model recognition accuracy but also significantly speeds up the convergence of the model. The flow chart of the experimental method is shown in Figure 6.

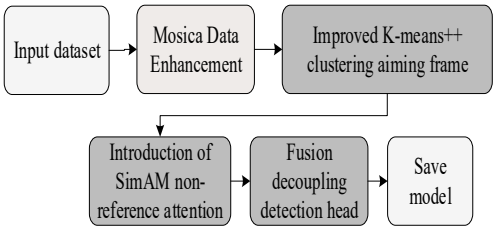

Figure 6. Flow chart of the proposed method

(4) Evaluation index

In order to evaluate and improve the effectiveness of the defect detection model, this paper selects the intersection ratio between the prediction frame and the target frame greater than 0.5 as the criterion for determining the target detection and uses Average Precision (AP), mean Average Precision (mAP) and Frames Per Second (FPS) for a comprehensive and objective evaluation of the model performance. Precision (P) describes the proportion of all detected targets that are correctly predicted by the model, while recall (R) describes the percentage of all targets that are correctly predicted by the model. The expressions are shown in equations (6) and (7), respectively.

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

Where: TP indicates that the target was detected completely correctly; FP indicates that the target was originally the wrong sample but was detected by the model as the correct sample, i.e., error detection; and FN indicates that the positive class of the target sample was predicted as the negative class, i.e., missed detection.

(5) Ablation experiment

Ablation experiment

The aiming frames originally set by YOLO and the aiming frames reacquired after clustering with K-means++ were collated and imported into the model for experiments, and the results were obtained as shown in Table 3. In experiment 1, the K-means algorithm was used for clustering without any modification to it. In Experiment 2, an attempt was made to use the modified clustering algorithm K-means++ for clustering detection of the prior frames. However, the experimental results showed that the mapped value improved by 1.7 percentage points compared with Experiment 1, and the inference speed of the improved algorithm did not decrease significantly. Despite the change in the clustering of the prior frame, the network structure itself did not change. The present algorithm shows superior results in clustering and the resulting prior frames are closer to the actual size of the dataset.

or matching are more fashionable. For example, in some haute couture conferences, many design products will use bead embroidery, plate gold and other three-dimensional embroidery to decorate the whole clothing, which makes the three-dimensional effect of the whole clothing stronger and elegant.

Table 2. K-means++ clustered results

Algorithm	P	R	mAP.5/%	mAP.75/%	FPS
YOLOV5s	70.9	77.0	77.5	42.3	57.47
K-means++	77.0	75.5	79.2	43.8	55.87

The purpose of this paper is to investigate the effect of different improved modules in YOLOv5 on the excellence of strip steel surface defect detection and to analyze it in depth by designing ablation experiments. In this study, the original YOLOv5s were used as the illuminated group, while its backbone, feature fusion network, and detection head parts were fused as the comparison experimental group. The experimental results are shown in Table 3. The detection effect was improved by re-clustering the aiming frame in the YOLOv5s network, and the mAP increased by 1.7%; the mAP was improved by 1.3% by adding the SimAM non-parametric attention mechanism to the Neck feature fusion network; the map was improved by 2.5% by replacing the head of the network with the decoupled detection head, and decoupling the classification task and localization task separately. When all strategies are used simultaneously in the YOLOv5s model, the mAP value of the model is improved by 3.2%, and the final mAP reaches 80.7%, which proves that the improvement of the algorithm is effective.

Table 3. YOLOv5s ablation experiments results

Models	K-means++	Double head	SimAM	Accuracy of each type of defect						mAP.5/%	mAP.75/%
				Cr	In	Pa	Ps	Rs	Sc		
1				42.6	84.0	90.3	84.4	67.6	96.4	77.5	42.3
2	✓			43.3	87.0	89.8	86.9	74.7	93.7	79.2	43.8
3		✓		45.5	88.2	91.9	83.9	75.4	95	80.0	42.9
4			✓	45.9	85.1	89.6	86.8	68.8	96.6	78.8	43.1
5	✓	✓		48.4	85.8	92.2	85.4	74.9	95.9	80.2	43.2
6	✓	✓	✓	53.0	84.9	89.7	85.5	74.0	97.1	80.7	43.3

(6) Comparison Experiment

To verify the superiority of the YOLOv5s model after fusing decoupling heads in this paper for strip surface defect detection, it is compared with current algorithms with good performance, such as the same types of YOLOv3, PP-YOLOE-PLUS and YOLOv7 and SSD and Faster-RCNN, which perform well in all aspects. by training on the NEU-DET dataset and testing, the mAP and inference speed metrics are selected as the measures, and the experimental results are shown in Table 5.

Table 5. Experiments with different algorithms

Model	Backbone	mAP%	FPS
Faster-RCNN	ResNet50	74.45%	10.37
SSD	VGG	73.96%	57.00
YOLOv3	Darknet-53	70.69 %	35.64
PP-YOLOE-PLUS	CSPRepResNet	73.60%	32.07
YOLOv5	CSPDarknet53	77.50%	57.47
YOLOv7	CSPDarknet53	76.90%	44.88
Improve YOLOv5	CSPDarknet53	80.70%	49.87

As can be seen from Table 4, although the Faster R-CNN has slightly higher accuracy than the SSD and YOLOv3 algorithms, it consumes a lot of time and cannot meet the requirement of real-time detection. Although the FPS of SSD algorithm reaches 57 frames/s, its accuracy is lower than YOLOv5s algorithm. YOLOV3 algorithm is inferior to other algorithms in terms of both accuracy and speed. PP-YOLOE-PLUS, although it is a high-precision target detection algorithm model, is not obvious for this dataset crack target, and the effect is not as good as expected. YOLOV7 also uses the CSPDarknet53 structure, which achieves a certain level of detection accuracy, but still suffers from a less-than-ideal state. The improved YOLOv5s algorithm has an overall accuracy improvement of 3.2%, and despite the decrease in speed, there is little loss of accuracy to meet the requirement of real-time detection (FPS higher than 30 fps) to 50 fps.

5. Conclusion

In this paper, we propose an improved YOLOv5s algorithm to address the problems of concentrated strip steel surface defect data, possible small defect targets, and unclear defect features, which may lead to missed detection and low accuracy. Firstly, the a priori frame is re-clustered to locate small targets more precisely; then the lightweight decoupled detection head is replaced so that the classification and localization tasks are performed separately to enhance the information of the network extracted feature map; finally, the inclusion of SimAM non-parametric attention module is added, and the effect is significantly improved without introducing additional parameters. The final experimental results show that the detection method of fusing decoupled detection heads in this paper can effectively improve the problem of inaccurate detection of strip images, and the detection speed also has a good advantage compared with other algorithms, and the final mAP of the model reaches 0.807 with a speed of 50 frames/s. The detection accuracy is higher than that of the generic target detection model, and the method provides useful help for the detection of defects in strips. However, in order to achieve real-time detection, a more lightweight architecture is needed to further optimize the model and improve the model generalization in order to implant the model into the mobile for detecting strip defects in real-time.

References

[1] CAO J, LI Y, SUN H, et al. A survey on deep learning based visual object detection[J]. Image Graph, 2022, 27: 1697-1722.

[2] SHI J, YANG J, ZHANG Y. Research on Steel Surface Defect Detection Based on YOLOv5 with Attention Mechanism[J]. Electronics, 2022, 11(22): 3735.

[3] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 580-587.

[4] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017, 39(6): 1137-1149.

[5] HE K M, KIOXARI G, DOLLAR P, et al. Mask R-CNN[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2961-2969.

[6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[7] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]// Proceedings of the IEEE conference on computer vision and pattern recognition washington D.C. USA: IEEE, Computer Society, 2017: 6517-6525.

[8] REDMON J, FARHADI A. Yolov3: An incremental improvement[C]// Computer vision and pattern recognition. Berlin/Heidelberg, Germany: Springer, 2018, 1804: 1-6.

[9] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv 2022, arXiv: 2207. 02696.

[10] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multiBox detector[C]// Proceedings of the 2016 European Conference on Computer Vision, LNCS 9905. Cham: Springer, 2016: 21-37.

[11] YANG L, ZHANG R Y, LI L, et al. Simam: A simple, parameter-free attention module for convolutional neural networks[C]// International conference on machine learning. PMLR, 2021: 11863-11874.

[12] FANG B, FANG L. Concise feature pyramid region proposal network for multi-scale object detection[J]. The Journal of Supercomputing, 2020, 765(5):3327-3337.

[13] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 8759-8768.

[14] HE Y, SONG K C, MENG Q G, et al. 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.

[15] BAO Y, SONG K, LIU J. Triplet-graph reasoning network for few-shot metal generic surface defect segmentation. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-11.