Surface Defect Detection Algorithm of Hot-Rolled Strip Based on Improved YOLOv7

Lijia Shen, Wenhua Cui, Ye Tao, Tianwei Shi, and Jinzhen Liao

Abstract—**To enhance the capability of classifying and localizing defects on the surface of hot-rolled strips, this paper proposed an algorithm based on YOLOv7 to improve defect detection. The BI-SPPFCSPC structure was incorporated into the feature pyramid in this algorithm, enabling enhanced extraction of features from small objects and improved accuracy in network model positioning. Additionally, a small object detection layer was introduced to enhance shallow feature capture. Then the CARAFE sampling operator was used for up-sampling to reduce the feature loss problem of small objects. Finally, the WIoU served as the loss function for network model training to expedite convergence. The NEU-DET dataset was utilized for ablation and comparison tests. The findings indicated that the enhanced YOLOv7 model's mAP value had increased to 80.7%. The detection impact was much enhanced in comparison to other traditional models, and the frequency of false and missing detections was also decreased.**

Index Terms—**Hot-rolled strip, YOLOv7, Defect detection, CARAFE, WIoU**

I. Introduction

As a type of steel product, hot-rolled strip steel serves as an indispensable raw material for numerous industrial applications. Within the manufacturing sector, it finds extensive usage in fields such as automotive, construction, pipeline, and shipbuilding. Simultaneously, the production process of hot-rolled strip steel involves heating, rolling, cooling, and other stages that render it susceptible to temperature variations and collisions. Consequently, these elements may result in surface flaws such dents, cracks, and scratches that could endanger public safety in addition to impairing the functionality and longevity of hot-rolled steel strip. Therefore, the detection of surface defects in hot-rolled strip steel holds significant practical significance.

Deep learning has advanced significantly in recent years, and as a result, its industrial intelligent detection application technology has matured. This integration plays a critical role in cost reduction and efficiency enhancement for organizations by improving detection speed and accuracy while also optimizing resource allocation. A common method for deep learning object detection is the YOLO series technique. YOLOv1 [1], initially proposed by Joseph in 2015, stands as the pioneering first-stage deep learning detection algorithm, ensuring a certain level of precision while enhancing object detection speed. Subsequently, a series of YOLO algorithms [2-4] have been proposed, leading to significant advancements in the field of defect detection. In 2021, Cheng et al. [5] proposed an enhancement to YOLOv3 by integrating shallow and deep features within the network architecture, resulting in novel feature layers capable of effectively capturing subtle object characteristics. Additionally, the inclusion of DIoU border regression loss was introduced to expedite model convergence. Kou et al. [6] proposed an end-to-end defect detection model leveraging YOLOv3, lowering computation time by implementing an anchorless feature selection method. Meanwhile, they introduced a dense convolutional block to extract comprehensive feature information, improving the capacity to characterize networks, reuse features, and propagate features. Liu et al. [7] proposed a real-time metal surface defect detection system utilizing an upgraded version of YOLOv4, in which MobileNetv3, a lightweight deep neural network, takes the place of the feature extraction network. To address the issue of imbalanced positive and negative data, they also developed a novel multi-scale adaptive loss function, which greatly raises this model's detection accuracy. An enhanced YOLOX method was suggested by Ge et al. [8]. To increase the algorithm's detection accuracy, a decoupled header and tag assignment approach (SimOTA) is employed. In 2022, Li et al. [9] proposed adding quadruple downsampling into the feature pyramid based on the YOLOv5 detection algorithm. while also integrating the CBAM attention mechanism into the neural network. These changes enhance the neural network's ability to extract information and detect small objects more accurately, as well as the model's capacity to generalize. Guo et al. [10] designed a MSTF-YOLO detection model, by combining the multi-scale feature fusion structure to achieve varied size feature fusion, improved the dynamic adjustment of the detector to different scale objects, and achieved real-time monitoring while improving the detection accuracy. Zhang et al. [11] proposed a lightweight YOLOv3-M3 network, which used K-means++ clustering algorithm to enhance the

Manuscript received September 18, 2023; revised February 8, 2024. This work was supported by Joint Fund Project of the National Natural Science Foundation of China (U1908218), the Natural Science Foundation project of Liaoning Province (2021-KF-12-06), and the Department of Education of Liaoning Province (LJKFZ20220197).

Lijia Shen is a Postgraduate of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, China. (e-mail: slj7708@163.com).

Wenhua Cui is a Professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China. (Corresponding author to provide phone: +86-133-0422-4928; e-mail: taibeijack@126.com).

Ye Tao is a Lecturer of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China. (e-mail: taibeijack@163.com).

Tianwei Shi is an Associate Professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China. (e-mail: tianweiabbcc@163.com).

Jinzhen Liao is a Postgraduate of School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, China. (e-mail: G851792070@163.com).

prediction accuracy. The bounding box regression loss function was further enhanced by adopting the CIoU loss function, thereby optimizing the training efficiency. During the same year, Wang et al. [12] proposed the YOLOv7 series algorithm and proposed the efficient aggregation network (ELAN) and model reparameterization [13] algorithms, which significantly increased the algorithm's detection efficiency and enabled it to perform faster and more accurately than the previous YOLO series. However, despite the high detection efficiency of YOLOv7, small defects on the surface of hot-rolled steel strips can easily be overlooked during model feature learning, resulting in false detections and missed objects. Therefore, there is still considerable scope for enhancing the capture of feature information pertaining to small objects and improving the accuracy of their detection using YOLOv7.

This paper presents an enhanced YOLOv7 algorithm for detecting surface defects in hot-rolled strip steel, with the aim of improving the classification and localization capabilities of small object defects. The main work includes the following four points: (1) In the backbone network, the CCBS module is constructed by introducing coordinate convolution CoordConv, the SPPF structure and the Biformer attention mechanism are used to reconstruct SPPCSPC. This facilitates the model to allocate greater attention towards regions containing small object samples and extract a higher number of features pertaining to small objects. (2) The incorporation of a small object detection layer facilitates the fusion of the high-resolution feature map with the original three scale feature maps, thereby enhancing the extraction of features pertaining to smaller objects. Additionally, a corresponding detection head is incorporated to augment the detection rate and precision of small objects. (3) A new up-sampling operation CARAFE is adopted to augment the receptive field. It can reduce the loss of small object features. (4) The loss function is substituted by the bezel-fitting CIoU incorporating the dynamic non-monotonic focusing WIoU, thereby enhancing the convergence speed of the network model.

II. Related Work

The YOLOv7 algorithm, being an iterative version of the YOLO series, outperforms the majority of existing object detection algorithms in terms of both speed and accuracy. Therefore, the YOLOv7 series model is chosen as the fundamental basis for subsequent research endeavors. The YOLOv7 series comprises three versions: YOLOv7, YOLOv7-Tiny, and YOLOv7-W6. Specifically designed for conventional GPUs, YOLOv7 is optimized to achieve superior performance. On the other hand, YOLOv7-tiny is tailored for embedded edge GPUs while YOLOv7-W6 caters to cloud-based GPUs. The network architecture of the comprehensive YOLOv7 model consists of four key components: Input, Backbone, Neck, and Head. The network architecture of YOLOv7 is illustrated in Fig. 1.

A. Input

The input side that was utilized augmentation of the

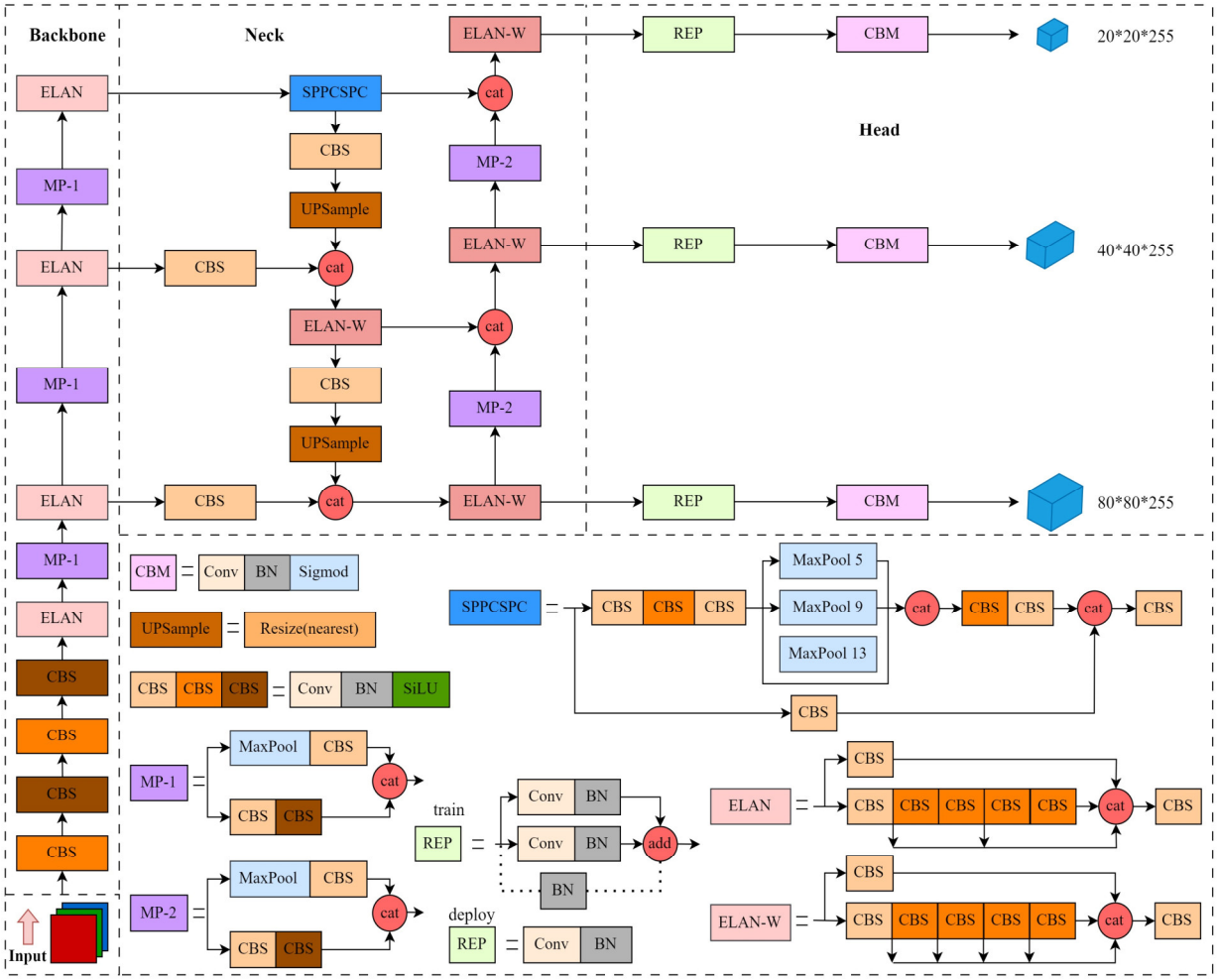

Fig. 1 YOLOv7 network model structure

mosaic data and several more preprocessing steps for the images, thereby optimizing the training effect while reducing graphics card memory consumption. This approach enriched both positive and negative samples for network learning, effectively meeting the requirements of feature extraction.

B. Backbone

The backbone network primarily consists of the CBS convolutional module, ELAN structure of efficient aggregation network and MPConv module. ELAN structure introduces the idea of residual structure, using multiple convolution, normalization and activation function stacking to improve accuracy by increasing depth. At the same time, the residual blocks are internally connected through jump connections, simultaneously addressing the issue of gradient vanishing during deep neural network training. The MPConv module effectively expands the receptive field of the feature graph by leveraging the maximum pooling operation. By integrating this with the characteristic information obtained from conventional convolution processing, it enhances the model's generalization capability.

C. Neck

In the neck, YOLOv7 uses the SPPCSPC structure. Mitigate the distortion induced by image processing and the issue of repeated extraction of image features. Subsequently, the FPN architecture is employed to integrate the three feature layers generated by the backbone, facilitating effective amalgamation of feature information across diverse scales in the model. Simultaneously leveraging the PANET structure facilitates upsampling-based fusion of features at various levels, facilitating information transfer and interaction between low-level and high-level representations, thereby enhancing model detection accuracy.

D. Head

In the head, three detection heads with different object sizes are used, and RepConv modules with different structures are used in the training and reasoning process, which can improve the training accuracy and reasoning speed, and finally output three different scale prediction results.

III. Algorithm Design

A. BI-SPPFCSPC Feature Pyramid Structure

In order to improve enhance small object feature extraction capabilities, the detection accuracy of small objects is further enhanced. In this paper, the SPPCSPC structure was redesigned, and the reconstruction method was to build CCBS module by using coordinate convolution to increase the representation capability of the network, then adopt SPPF structure to speed up the detection speed, and finally incorporate the BiFormer attention mechanism to enhance the detection efficiency of small object defects. After reconstruction, the structure of BI-SPPFCSPC was shown in Fig. 2. The reconstruction process is as follows.

1) CCBS Module

Conventional convolution operations exhibit translation invariance, allowing images to share unified convolution kernel parameters across different locations, thereby facilitating the learning of essential features for tasks like classification. However, when conventional convolution performs local operations in the convolution kernel, the model is limited to perceiving only local information and lacks the ability to perceive the positional information of the current feature within the image. The focus of the proposed method is solely on the pixel value of the input image, while ignoring the information related to the position and coordinate. CoordConv [14] differs from conventional convolution by incorporating two additional channels after the input feature map, representing the i and j coordinates of each original input pixel. These coordinate channels are then connected with the original input feature channels using conventional convolution, enhancing its performance. This approach enables better understanding of spatial relationships and positional information between pixels as depicted in Fig. 3. The CCBS module is constructed by replacing 1×1 conventional convolutions in the original structure with CoordConv layers. During training, perceiving coordinate information improves detection accuracy and enhances precise position perception.

2) SPPF Structure

The SPPF structure initially partitions the features processed by the convolution layer into two segments, one utilizing conventional operations and the other employing maximum pooling processing, as illustrated in Fig. 4. The convolution kernel size within the maximum pooling layer is set to 5×5, followed by sequential feature map input and fusion. Ultimately, the maximum pooling segment is integrated with the conventional processing segment. By virtue of two consecutive 5×5 maximum pooling layers yielding an equivalent feature extraction effect as a single 9×9 maximum pooling layer, three successive 5×5 maximum pooling layers achieve a comparable feature extraction effect to that of a single 13×13 maximum pooling layer. Consequently, while attaining identical operational efficacy as the SPP[15] structure, the SPPF architecture enhances computational efficiency by half and augments network model speed without compromising accuracy.

3) BiFormer Attention Mechanism

BiFormer [16] is a dynamic sparse attention mechanism based on a variant of the Transformer architecture. The approach employs query adaptation to concentrate on a limited set of pertinent tags while disregarding extraneous

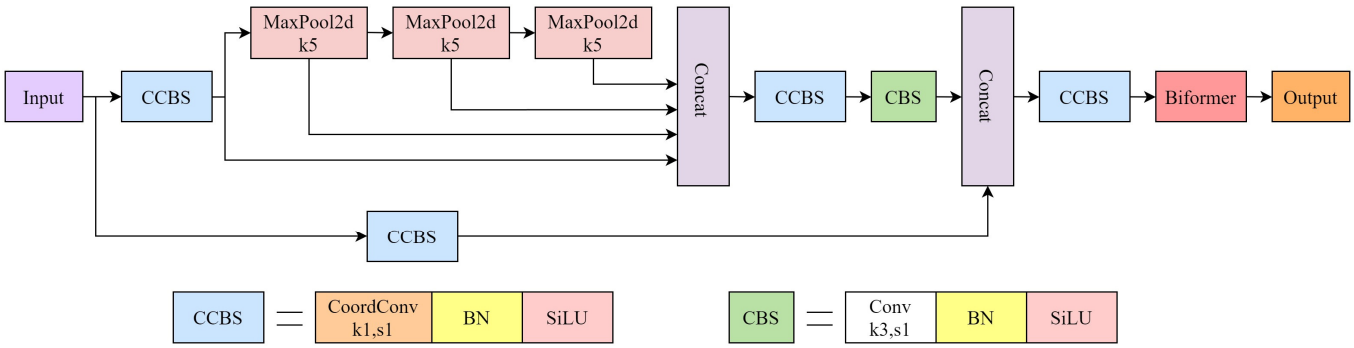

Fig. 2 Structure of the BI-SPPFCSPC

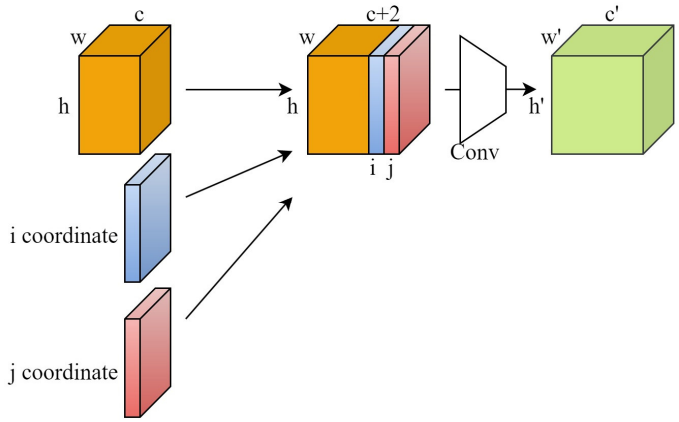

Fig. 3 Structure of CoordConv

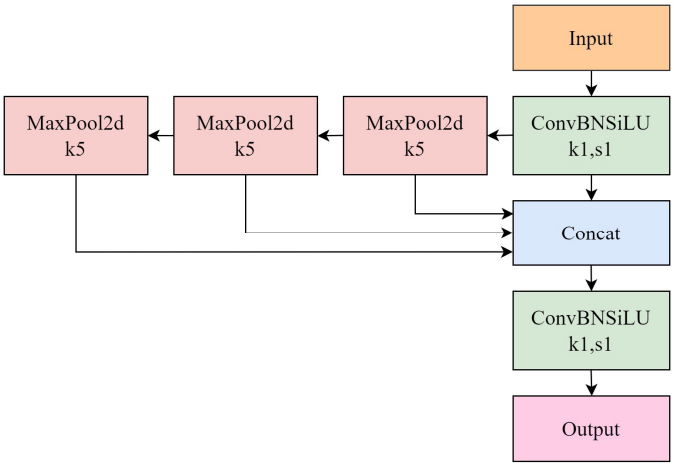

Fig. 4 Structure of SPPF

ones. Its structure is shown in Fig. 5. During the initial stage, BiFormer employs overlapping blocks to embed, and in the second to fourth stage, the Merge module is utilized to decrease the spatial resolution of input data and simultaneously increase channel capacity. Subsequently, consecutive BiFormer blocks are employed for feature transformation. This approach enables more flexible computation allocation and content perception, addressing issues related to high computational complexity and memory consumption. Additionally, it enables the network to prioritize regions that contain small objects, facilitating extraction of precise features and enhancing small object detection performance.

B. Small Object Detection Layer

The original YOLOv7 model incorporates three detection layers with varying scales. When an image of size 640×640 is input, the detection layer outputs scales of 80×80, 40×40, and 20×20 respectively. This network employs receptive field size to differentiate objects, utilizing large-size feature maps for detecting small objects and small-size feature maps for detecting large objects. However, the subsampling factor of YOLOv7 is relatively large, leading to the loss of defect feature information during continuous subsampling. Additionally, deeper feature maps struggle to capture shallow small object features, resulting in suboptimal detection performance for smaller objects. Therefore, a structure of small object detection layer is proposed, and the shallow feature map and deep feature map are fused together, this enables the model to enhance its ability in extracting detailed information from small objects during training, thereby improving the accuracy of object detection.

The network is enhanced by adding a small object detection layer with an input resolution of 160×160 in this study, and the PAFPN structure is used for horizontal connection with the detection layer with 8, 16 and 32 times downsampling. The detection layer of small objects contains more intricate details and features, enough defect feature information can be provided to the deep feature map during the feature fusion procedure to improve the feature capture.

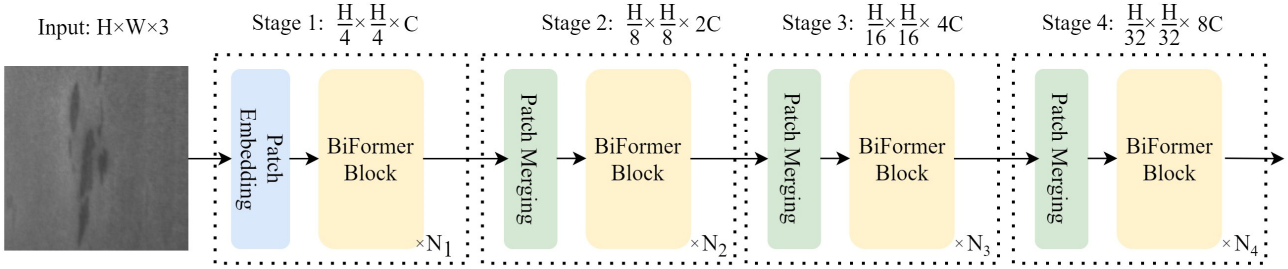

Fig. 5 Structure of BiFormer

At the same time, the detection head corresponding to the small object detection layer is simultaneously incorporated on this basis, thereby enhancing the capability of detecting small objects, which can further reduce the situation of missing and false detection, and enhance the network's capacity for generalization.

C. CARAFE Upsampling Operator

The YOLOv7 model utilizes the nearest neighbor interpolation technique for up-sampling, which provides the benefit of fast processing speed and is suitable for minor image scaling operations. However, due to its simplistic approach of copying values from neighboring pixels, it fails to generate new pixel values and lacks smoothness, resulting in jagged edges in the enlarged images. Additionally, the upsampled kernel of the nearest interpolation is solely determined by the spatial position of the pixel, resulting in a limited perceptual domain that fails to fully capture the content information embedded within the feature map.

This paper proposes a substitution for the nearest neighbor interpolation method by introducing CARAFE[17], a lightweight up-sampling operator. The CARAFE operator incorporates content perception and feature recombination as its fundamental concepts. Fig. 6. illustrates the structure of the CARAFE operator. When an input feature graph with an up-sample rate σ and size H×W×C is provided, the kernel prediction module initially employs a channel compressor using 1×1 convolution to compress the number of channels in the input feature graph to H×W×C_m. This compression helps reduce computational complexity for subsequent steps. Subsequently, a convolution layer with parameter $k_{encoder} \times k_{encoder}$ is utilized to predict an upsampled kernel of size $k_{up} \times k_{up}$, where the number of input channels is C_m and the number of output channels is $\sigma^2 k_{up}^2$. The channel dimensions are expanded within the spatial domain to acquire an upsampled kernel with a size $\sigma H \times \sigma W \times k_{up}^2$. Finally, the weight sum of the convolutional kernel is ensured to be equal to 1 through normalization using the Softmax function. Every position in the output feature map is mapped by the feature recombination module back to its matching location on the input feature map, resulting in the acquisition of a refined feature map $N(\chi_l, k_{up})$. Then, a region $k_{up} \times k_{up}$ centered at each location is extracted for dot product operation with predicted upper sampling kernel $W_{l'}$ at that point. This enables different channels at identical locations to share identical upper sampling kernels and ultimately yields an output feature map with size $\sigma H \times \sigma W \times C$.

After applying the CARAFE operator, the feature fusion network is able to extract more comprehensive contextual information from the input low-resolution feature map, thereby facilitating a better understanding of the global structure of the input image and ultimately enhancing defect detection performance.

D. DIoU Loss Function

For object detection, choosing a loss function is essential since it affects how well objects are localized. In YOLOv7, the CIoU loss function [18] is employed for bounding box regression. This particular loss function incorporates the aspect ratio between predicted and actual frames to effectively address situations where there is no overlap between them, thereby providing accurate motion direction for bounding frames. However, despite its advantages, the CIoU loss function treats all loss variables collectively and fails to consider potential mismatches between actual objects and predicted boxes, resulting in slow convergence and instability.

WIoU (Wise-IoU) [19] is a new bounding box regression loss function with dynamic non-monomonotic focusing mechanism. The evaluation of anchor frame quality in this approach employs "outliers" instead of IoU, while ensuring the high-quality effect through gradient gain and mitigating the influence of detrimental gradients, thereby enhancing the overall algorithm performance. WIoU constructs a two-layer attention mechanism to prevent slow convergence, improve convergence accuracy and enhance model generalization ability. The formula for $WIoU_{v1}$ is in equation (1).

$$L_{WIoU_{v1}} = \exp\left(\frac{(x-x_{gt})^2 + (y-y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) L_{IoU} \tag{1}$$

In order to avoid large harmful gradients for lower quality

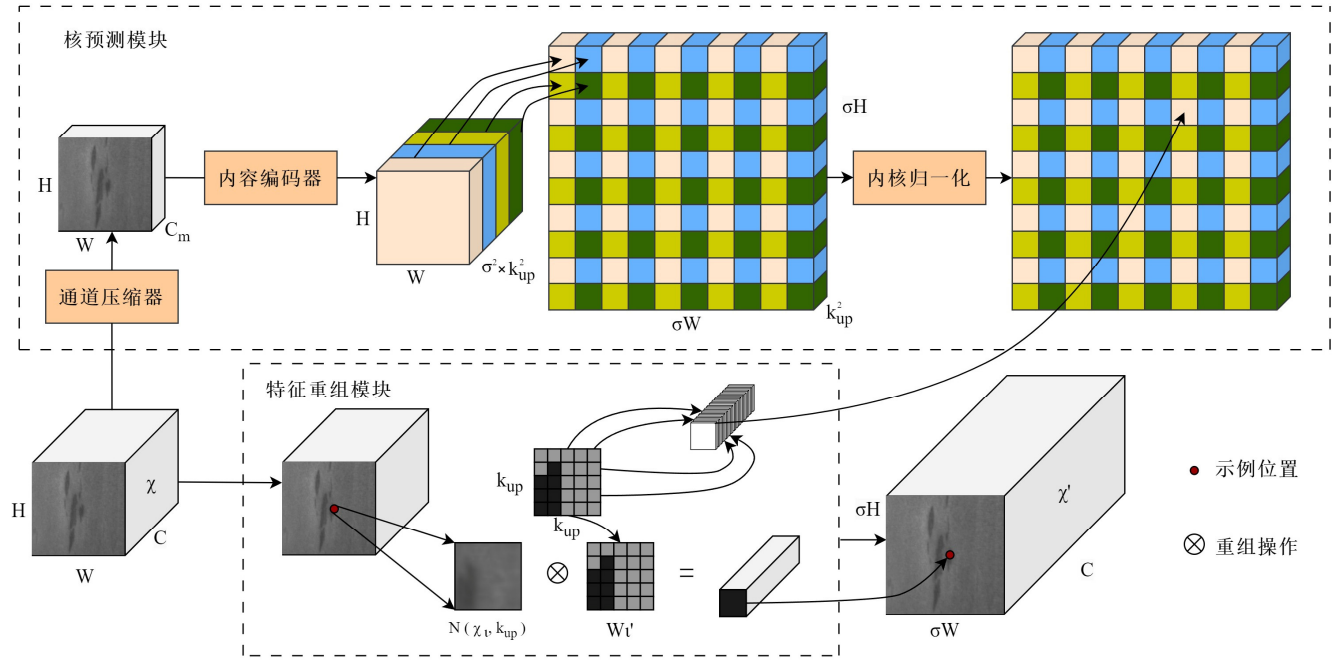

Fig. 6 Structure of CARAFE operator

samples, this paper utilizes β and $WIoU_{v1}$ to construct $WIoU_{v3}$. The formula for $WIoU_{v3}$ is in equation (2).

$$L_{WIoU_{v3}} = \frac{\beta}{\delta \alpha^{\beta-\delta}} L_{WIoU_{v1}} \tag{2}$$

The variables δ and α are represent hyperparameters, while β denotes the outlier factor of the anchor box, exp is an exponential function, the coordinates x and y represent the central point of the predictor box, x_{gt} and y_{gt} are the location of the center point of the real box, W_g and H_g are the width and height of the smallest enclosing box of the predictor box and the real box, * represents the computational separation, L_{IoU} represents the ratio of the crossing regions between the predictor box and the real box.

E. Improved YOLOv7 Network Model Structure

The enhanced model architecture is illustrated in Fig. 7. BI-SPPFCSPC represents a reconstructed feature pyramid structure, which aims to extract more precise features of small objects. Additionally, the inclusion of a small object detection layer enhances the capture of these features, and then the CARAFE upsampling operator is employed as a replacement for the nearest neighbor interpolation method to augment the fusion capability of the feature fusion network. Finally, the network model's frame loss function is swapped out for CIoU with WIoU in order to strengthen the model's capacity for generalization and quicken its convergence.

IV. Experiments

A. Experimental Dataset

The NEU-DET (Northeastern University-Detect) dataset is a collection of neutron data on steel surface defects, published by Northeastern University. It encompasses six distinct categories of common defects observed on the surfaces of hot-rolled strip, including Rolled-in-Scale, Patches, Crazing, Pitted-Surface, Inclusion and Scratches. The collection offers annotations specifying the kind and position of flaws in each image for the purpose of defect detection.

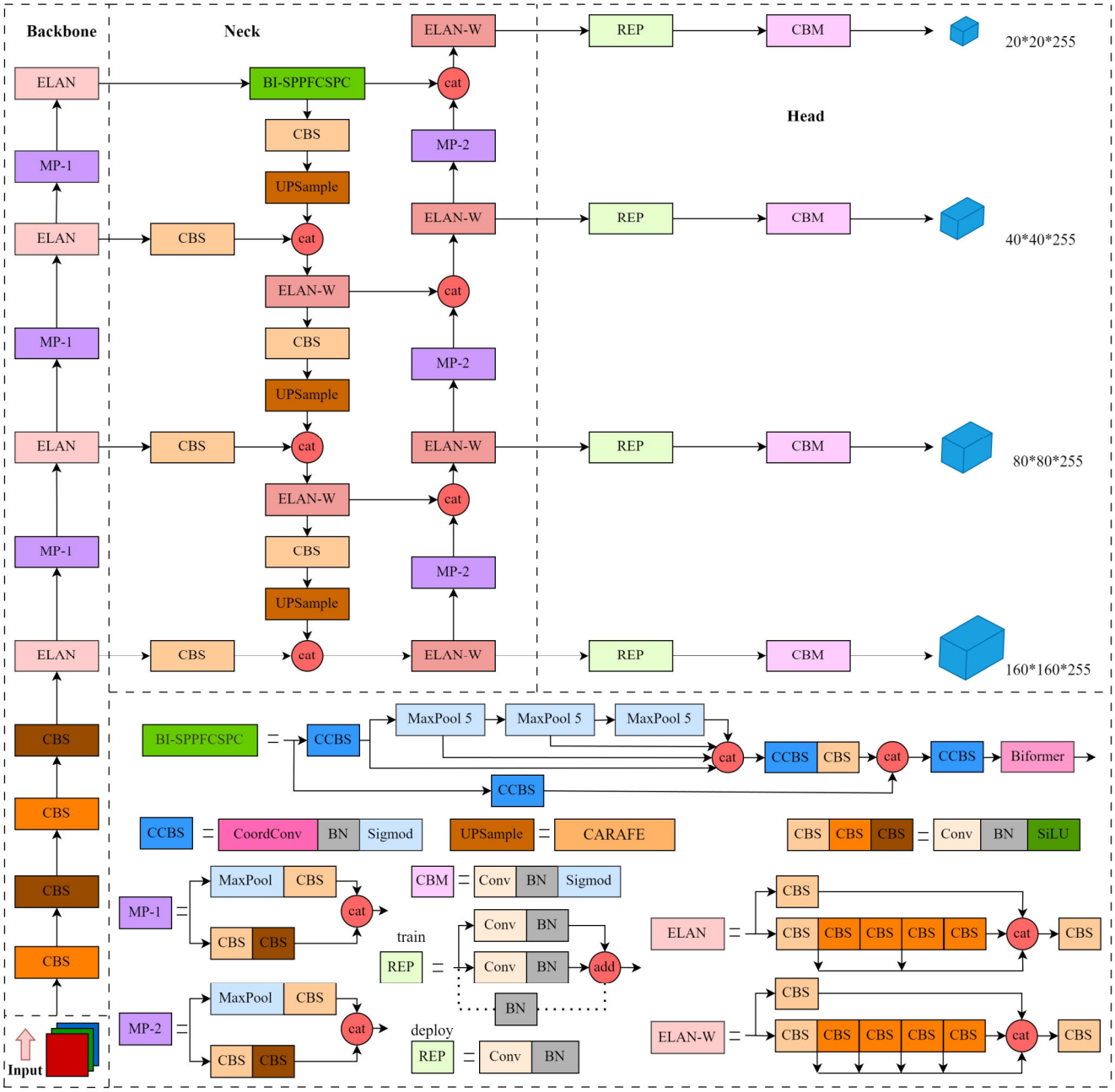

Fig. 7 Improved YOLOv7 network architecture

B. Experimental Environment

The experimental setup employed the Windows 10 operating system, an Intel(R) Core(TM) i7-10700F CPU @ 2.90GHz, and an NVIDIA GeForce RTX 3070 graphics card with 8GB of video memory. The deep learning framework utilized was PyTorch version 1.12.1, implemented in Python 3.9 programming language. The accelerated computing architecture employed CUDA version 11.6 for efficient computations. The algorithm processed input images of size 640×640 and a training batch size of 8 was used for a total of 300 epochs during training process.

C. Evaluation Indicators

This study chooses three widely-used evaluation criteria for object detection tasks in order to thoroughly and impartially assess the trained model's performance: mAP, Params, and GFLOPs to measure the upgraded algorithm's efficacy. The introduction of these metrics is as follows.

1) mAP

The mean average precision (mAP) represents the overall accuracy of object detection across all categories. In this paper, the IoU threshold for judging positive and negative samples is set as 0.5. When the Intersection over IoU value between the detected bounding box and the ground truth box exceeds a predefined threshold, the sample is classified as a positive instance and denoted as TP (true positive). When the IoU value is lower than 0.5, the classification of a sample as negative is indicated by FP (false positive). The ratio of the number of positive samples to the total number of detected objects of this type is denoted as the accuracy P. The formula for P is in equation (3).

$$P = \frac{TP}{TP + FP} \tag{3}$$

FN (false negative) represents the sample that misidentifies the positive sample as the negative sample, and the recall rate is denoted as R. The formula for R is in equation (4).

$$R = \frac{TP}{TP + FN} \tag{4}$$

The two-dimensional P-R curve is plotted with precision (P) on the vertical axis and recall rate (R) on the horizontal axis. The area under the P-R curve, known as average accuracy value AP, represents a measure of performance. The formula for AP is in equation (5).

$$AP = \int_0^1 P(R)dR \tag{5}$$

Compute the average precision (AP) values for each category individually, aggregate all AP values, and then divide by the number of categories to obtain the mean average precision (mAP). The formula for mAP is in equation (6).

$$mAP = \frac{1}{c}\sum_{i=1}^{c} AP_i \tag{6}$$

2) Params

Params is the number of parameters in the network model, denoted as M.

3) FLOPs

FLOPs is the amount of computation that reflects the complexity of the model, denoted as G.

D. Experimental Results and Analysis

The comparison of experimental results between the proposed improved YOLOv7 model and the original YOLOv7 model is presented in Fig. 8, providing compelling evidence to validate the efficacy of our algorithm. The top section displays the detection performance of YOLOv7, while the bottom section shows that of our improved model. In the diagram, each box represents a detected defect region, with the defect category and confidence level annotated above each respective region. The experimental results demonstrate that the enhanced model exhibits superior detection accuracy compared to the original YOLOv7 model across six types of defects. Simultaneously, the improved model exhibits an additional recognition box in detecting cracking and scratch defects compared to the original model, thereby indicating a certain enhancement in addressing the issue of missed detections.

The improved model's detection performance on different types of defects on the surface of hot-rolled steel strips can be analyzed through the utilization of a confusion matrix. Fig.9 presents the confusion matrix generated after training, which consists of six categories. By examining the diagonal elements in the matrix, it becomes evident that a significant majority of defects are accurately predicted and assigned to their respective categories. Furthermore, by analyzing the accuracy distribution within the confusion matrix, it is observed that Patches, Scratches, and Inclusion are predicted with higher precision compared to other categories. This can be attributed to the distinct characteristics exhibited by patch-type defects during training, indicating that their

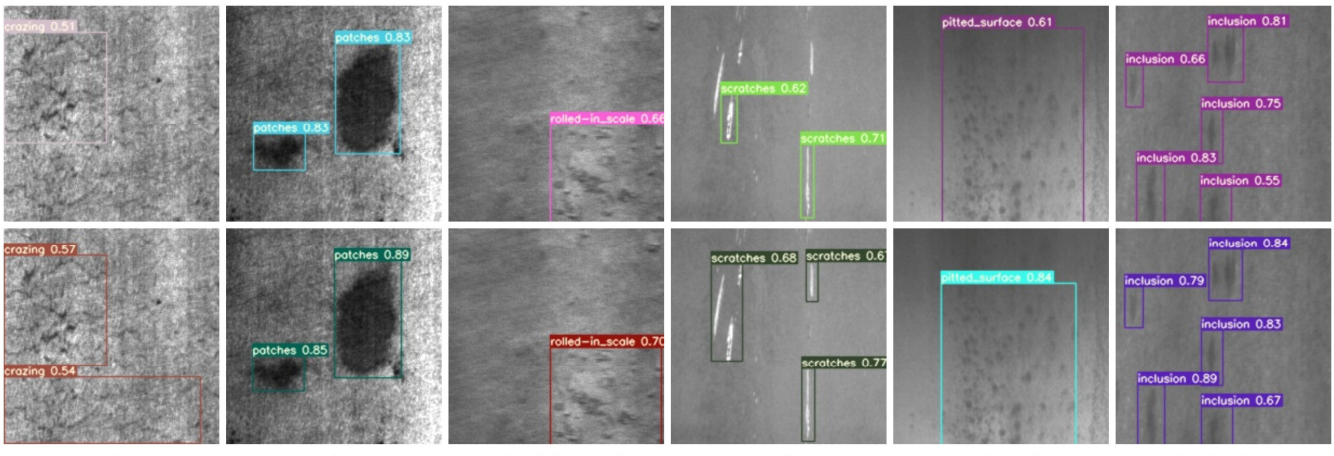

Fig. 8 Comparison of experimental results of hot-rolled steel strips

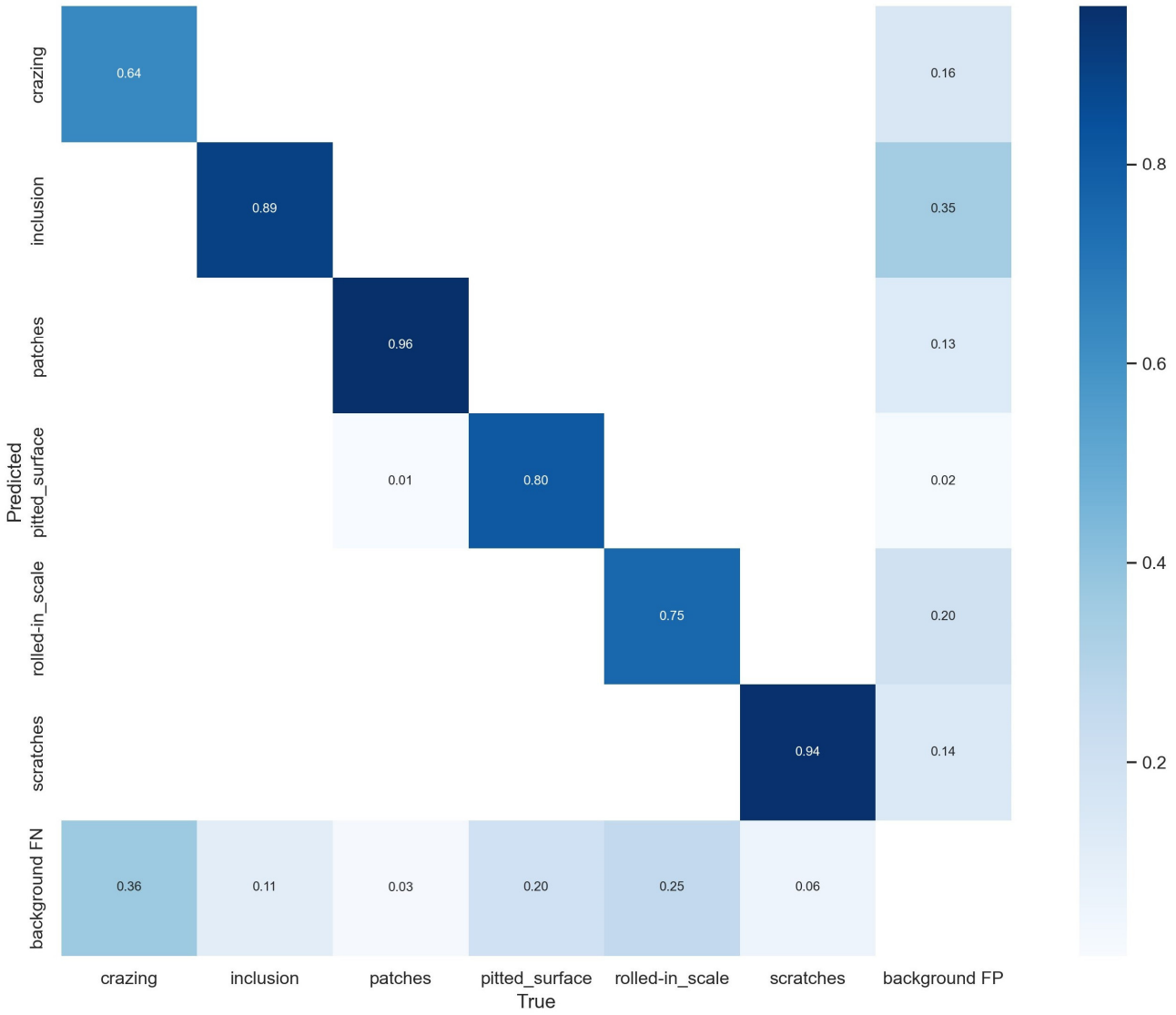

Fig. 9 Confusion matrix for hot-rolled strip experiments

features were more effectively captured and utilized for prediction purposes. Consequently, these features demonstrate superior expressiveness.

The efficacy of the enhanced strategy employed in the proposed algorithm has been validated through a series of ablation experiments. YOLOv7 was used as the baseline model in the experiment, and the corresponding experimental findings were presented in Table I. The effectiveness of each improvement point was verified through a series of eight ablation experiments, denoted as N1-N8. "✓" represents that the improvement point was adopted, and "-" represents that the improvement point was not adopted.

In the table, group N1 directly uses the YOLOv7 model to conduct experiments. The experimental results demonstrate that the mAP value of the baseline model YOLOv7 is 75.5%, the number of model parameters is 37.2M, and the calculation amount is 108.8G. Although it has a great advantage in accuracy, its effect is still not ideal. In the N2 group experiment, the original SPPCSPC was reconstructed into Bi-SPPFCSPC in the backbone network of YOLOv7, and the mAP reached 78.2%. In contrast to the original YOLOv7 model, there was a 2.7% rise in the mAP value, which proved the effectiveness of the reconstructed BI-SPPFCSPC structure. In group N3, the YOLOv7 model incorporates a dedicated layer for small object detection, resulting in an improved mAP value of 78.3%, this improvement of 2.8% compared to the original model effectively enhances the detection performance for small object defects. In the N4 group of experiments, the improved points of N2 and N3 are combined, and the mAP value is increased to 79.5%, thereby substantiating the concurrent enhancement of detection efficacy brought about by the integration of these two structures. In the N5 group of experiments, the upsampling method was replaced in the original YOLOv7 model, and the CARAFE operator was used. The original model achieved a 1.3% increase in mAP value, reaching 76.8%, while exhibiting only minimal augmentation in parameter count and computational complexity. In the N6 group of experiments, the border loss function CIoU was replaced with WIoU, resulting in accelerated convergence of the network model and a 0.6% increase in mAP value. The N7 group of experiments combines the improved points of N5 and N6, and the mAP reaches 77.2%, thereby demonstrating their concurrent enhancement on detection accuracy. The N8 group of experiments is the combination of four improvement points, and the mAP value is increased to 80.7%, the parameter amount is 41.2M, and the calculation amount is 120.7G. Although compared with the original model, the parameter amount and calculation amount are increased, the mAP value

is increased by 5.2%, indicating that the algorithm constructed in this paper can greatly increase the precision with which small object flaws are detected.

The proposed algorithm is compared with object detection algorithms including YOLOv5, YOLOX, YOLOv7, and ST-YOLO[20], and the corresponding results are organized in Table II. The table reveals that the mAP value of the enhanced YOLOv7 model in this study exhibits a 7.5% increase compared to that of the YOLOv5 model, 6.9% higher than that of the YOLOX model, 5.2% higher than that of the YOLOv7 model, and 0.4% higher than that of the ST-YOLO model. Meanwhile, the mAP curve of the improved algorithm in this paper and the original YOLOv7 model in the training process are presented in the form of comparison, as shown in Fig. 10. The proposed algorithm demonstrates a significant enhancement in detection accuracy compared to other mainstream object detection networks, while only a marginal increase in computational complexity and parameters. Moreover, it exhibits superior identification capabilities for detecting small objects.

V. Conclusion

This paper proposes an improved YOLOv7-based model to address the issue of false positives and false negatives in detecting small object defects on hot-rolled strip surfaces. In this model, the BI-SPPFCSPC feature pyramid structure is designed, which uses coordinate convolution to replace the conventional convolution to construct the CCBS module, structure and introduces the BiFormer attention mechanism.

TABLE I
Result of Ablation Experiments

Number	BI-SPPFCSPC	Detection Layer	CARAFE	WIoU	mAP	Params	FLOPs
N1	-	-	-	-	75.5	37.2	108.8
N2	✓	-	-	-	78.2	38.5	115.5
N3	-	✓	-	-	78.3	39.8	112.6
N4	✓	✓	-	-	79.5	40.9	118.7
N5	-	-	✓	-	76.8	37.4	109.1
N6	-	-	-	✓	76.1	37.2	108.8
N7	-	-	✓	✓	77.2	37.5	109.1
N8	✓	✓	✓	✓	**80.7**	41.2	120.7

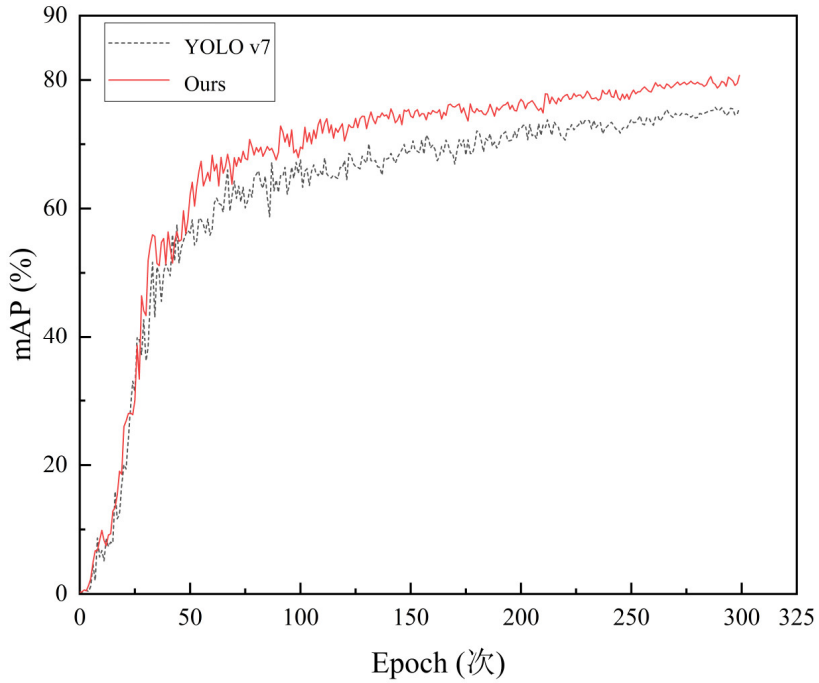

Fig. 10 Comparison chart of experimental results mAP

TABLE II
Performance Comparison of Various Algorithms

Method	mAP	Params	FLOPs
YOLOv5	73.2	46.1	109.2
YOLOX	73.8	54.5	118.6
YOLOv7	75.5	37.2	108.8
ST-YOLO	80.3	55.8	115.5
Ours	**80.7**	41.2	120.7

The structure enhances the model's ability to detect small object defects and facilitates the extraction of more intricate features associated with small objects. Then, a small object detection layer is incorporated to effectively address the issue of excessive feature loss pertaining to smaller objects and enhance feature extraction capabilities. The CARAFE up-sampling operator is subsequently employed to enhance the feature fusion capability. The WIoU loss function is ultimately incorporated into the network model to enhance both convergence speed and accuracy. Utilizing the NEU-DET dataset, the ablation and comparison tests show that the optimized method attains a mAP value of 80.7%, 5.2% greater than the original YOLOv7 algorithm. These results unequivocally demonstrate that revised method improves the detection efficiency of small object defects on hot-rolled steel strip surfaces, while also exhibiting superior detection accuracy in comparison to previous methods. The dataset can be further enriched in future research by collecting a wider range of defect images, thereby improving the model's generalization performance. Additionally, the implementation of pruning techniques can effectively mitigate model complexity and transmission costs, thereby facilitating its lightweight nature.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788, 2016.

[2] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263-7271, 2017.

[3] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv:1804.02767, 2018.

[4] A. Bochkovskiy, C. Y. Wang and H. Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv:2004.10934, 2020.

[5] J. Y. Chen, X. H. Duan and W. Zhu, "Research on Metal Surface Defect Detection by Improved YOLOv3," *Computer Engineering and Applications*, vol. 57, no. 19, pp. 252-258, 2021.

[6] X. P. Kou, S. J. Liu, K. Q. Cheng and Y. Qian, "Development of a YOLO-V3-based model for detecting defects on steel strip surface," *Measurement*, vol. 182, pp.109454, 2021.

[7] Y. Liu, Q. Wang, K. Zhao and Y. Liu, "Real-time defect detection of hot rolling steel bar based on convolution neural network," *Chinese Journal of Scientific Instrument*, vol. 42, no. 12, pp. 211-219, 2021.

[8] Z. Ge, S. Liu, F. Wang, Z. Li and J. Sun, "YOLOX: Exceeding YOLO Series in 2021," arXiv:2107.08430, 2021.

[9] R. Li and Y. P. Wu, "Improved YOLO v5 Wheat Ear Detection Algorithm Based on Attention Mechanism," *Electronics*, vol. 11, no. 11, pp. 1673, 2022.

[10] Z. X. Guo, C. S. Wang, G. Yang, Z. Y. Huang and G. Li, "MSFT-YOLO: Improved YOLOv5 Based on Transformer for Detecting Defects of Steel Surface," *Sensors*, vol. 22, no. 9, pp. 3467, 2022.

[11] B. Zhang, X. Zhang and Z. Li, "An Efficient Face Mask Wearing Detection Algorithm Based on Improved YOLOv3," *Engineering Letters,* vol. 30, no. 4, pp.1493-1503, 2022.

[12] C. Wang, A. Bochkovskiy and H. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7464-7475, 2023.

[13] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding and J. Sun, "RepVGG: Making VGG-style ConvNets Great Again," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13728-13737, 2021.

[14] R. Liu, J. Lehman, P. Molino, F. Petroski Such, E. Frank, A. Sergeev and J. Yosinski, "An intriguing failing of convolutional neural networks and the CoordConv solution," *Advances in Neural Information Processing Systems*, pp. 31, 2018.

[15] K. He, X. Zhang, S. Ren and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904-1916, 2015.

[16] L. Zhu, X. Wang, Z. Ke, W. Zhang and R. W. Lau, "BiFormer: Vision Transformer with Bi-Level Routing Attention," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10323-10333, 2023.

[17] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy and D. Lin, "Carafe: Content-aware reassembly of features," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3007-3016, 2019.

[18] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12993-13000, 2020.

[19] Z. Tong, Y. Chen, Z. Xu and R. Yu, "Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism," arXiv:2301.10051, 2023.

[20] H. Ma, Z. Zhang and J. Zhao, "A Novel ST-YOLO Network for Steel-Surface-Defect Detection," *Sensors*, vol. 23, no. 22, pp. 9152, 2023.