A Multi-Scale Steel Defect Detection Method Based on RT-DETR

Zhen Qiang Dai, Shao Chuan Xu*, Xiang Yi Yan and Si Hong Xu

Abstract—Surface defects in steel can significantly affect its mechanical properties, product safety and economic value, making timely detection and repair essential. However, the wide variation in defect sizes and characteristics presents a major challenge for inspection, and existing mainstream models often suffer from missed detections and false positives across different defect scales. To address these limitations, this paper proposes a novel object detection model, VAS-DETR, which integrates RepConv and Efficient Multi-head Attention (EMA) to create a more effective backbone for feature extraction. A Multi-Scale Atrous Fusion (MSAF) module is introduced to enhance multi-scale feature aggregation, thereby improving robustness to defect size variation. Additionally a Multiscale Multi-head Self-Attention (M2SA) mechanism is incorporated into the AIFI module of RT-DETR to better capture fine-grained features. To further improve localization performance, especially when bounding box overlap is limited, the Generalized IoU (GIoU) loss is replaced with a Modified Penalty DIoU (MPDIoU) loss. Experimental results on the NEU-DET dataset demonstrate that VAS-DETR improves mAP_{50} by 3.93% and $mAP_{50:95}$ by 4.58% compared to RT-DETR, while reducing GFLOPs by 12.74% and the number of parameters by 16.85%. The proposed model significantly enhances feature representation and multi-scale fusion capabilities, offering an accurate and efficient solution for industrial steel surface defect detection without increasing model complexity.

Index Terms—Deep Learning, RT-DETR, Parallel Convolution, Defect Detection

I. Introduction

Hot rolled strip refers to strip and plate products manufatured through hot rolling, widely applied in the automotive, electrical machinery, chemical and shipbuilding industries, as well as used as billets for producing cold-rolled and welded pipes. Quality control of hot-rolled strip is crucial, with technical requirements covering dimensional accuracy, flatness, surface quality and mechanical perforance. Among these, surface defects are particularly significant due to their diverse types and unpredictable locations [1]. Defect detection involves both identifying the defect category and localizing its position, making it a highly challenging task. Existing research primarily focuses on magnetic particle inspection, penetrant testing, eddy current inspection, ultrasonic testing, machine vision, and deep learning methods [2-8]. Among these, convolutional neural networks (CNNs), known for their powerful feature extraction capabilities and ability to autonomously learn image representations, have become indispensable in industrial inspection tasks [9]. Currently, mainstream object detection algorithms include Fast R-CNN [10], Faster R-CNN [11], Mask R-CNN [12], SSD [13], the YOLO [14] series, and Transformer-based DETR [15] series. Compared to traditional methods, defect detection techniques based on deep learning are more efficient and accurate in industrial environments. For example, Chen et al. proposed a detection method for ultrasonic images of weld seams by combining Faster R-CNN with a ResNet50 incorporating deformable convolution modules, enhancing small object detection accuracy through K-means clustering and ROI Align [16]. Wang et al. developed a rail defect detection algorithm based on Mask R-CNN, overcoming the limitations of IoU using the CIoU metric [17]. Bo et al. improved the SSD algorithm by integrating channel attention and feature fusion modules to enhance small object detection accuracy in workpiece recognition [18]. Zheng et al. proposed MD-YOLO based on YOLOv5, improving object recognition and localization capabilities by introducing a dynamic head module [19]. Wang et al. developed the improved YOLOv8n-DSDM algorithm, integrating the C2f-DSConv module and a small object layer, and optimized regression loss with MPDIoU to improve detection accuracy [20]. Kong et al. focused on steel surface defect detection by enhancing YOLOv8 with attention-free and SPPF modules to optimize specific region feature extraction and expand the receptive field [21]. Cheng et al. improved the RT-DETR model by incorporating additional prediction layers and the CA mechanism to enhance the accuracy of crack detection in metal components [22].

Although the aforementioned deep learning-based industrial inspection models have largely met the core requirements of modern defect detection, most approaches rely on the integration of shallow features and attention mechanisms to enhance the detection of small targets. While these techniques offer slight improvements in accuracy over baseline models, their effectiveness remains limited. This study investigates the NEU-DET dataset [23], which contains six typical surface defect categories found in hot-rolled steel strips. These defects exhibit a high level of complexity, primarily due to significant dimensional variations not only across different defect types but also within individual categories. This diversity spans large,

Manuscript received April 8, 2025; revised August 17, 2025.

Zhen Qiang Dai is a Postgraduate Student of School of Electronic Information, University of Science and Technology Liaoning, Anshan, 114051 China (e-mail:232085400011@stu.ustl.edu.cn).

Shao Chuan Xu is a Professor of School of Control Science and Engineering, University of Science and Technology Liaoning, Anshan, 114051 China (Corresponding author, phone: 86-0412-5929747; e-mail: shaochuanxu1@163.com).

Xiang Yi Yan is a master's graduate in control science and engineering from University of Science and Technology Liaoning, Anshan, P. R. China (e-mail: 793119513@qq.com).

Si Hong Xu is a master's graduate in control science and engineering from University of Science and Technology Liaoning, Anshan, P. R. China (e-mail: 735911068@qq.com).

medium, and small-scale defect features. Therefore, enhancing small-object detection alone is insufficient to optimize overall accuracy and detection performance. More comprehensive strategies are required to effectively address the challenges posed by multi-scale surface defects.

Given that the RT-DETR model provides real-time, end-to-end detection capabilities and meets the practical requirements of industrial defect inspection, RT-DETR-R18 [25] is selected as the baseline model in this study. To address the challenges posed by defect detection across varying scales, we propose an improved model named VAS-DETR (Variable Scale Detection Transformer), based on RT-DETR-R18. VAS-DETR aims to balance detection accuracy across different defect sizes while ensuring that model complexity remains unchanged, thereby enhancing performance in detecting multi-scale and complex defect features. The main contributions of this study are as follows:

1) In this study, a new target detection model named VAS-DETR is proposed to address the problem of accuracy degradation caused by significant size differences in the detection of surface defects on hot rolled steel. VAS-DETR has significant advantages in the detection of defects with large size differences, and improves the accuracy of the detection of defects on the surface of hot rolled steel under the premise of ensuring that there is no loss in the computational efficiency.

2) Two new modules ERA Block and MSAF are proposed. Among them, ERA Block significantly enhanced the feature extraction capability of the model, while MSAF module improved the performance of the model in capturing global information and local details, further improving the overall detection accuracy.

3) In order to further improve the model performance, the Multi-Scale Feature Integration (MSFI) module is proposed, which integrates the AIFI module of RT-DETR with the M2SA module to enhance the capability of capturing fine features. In addition, Modified Penalized Distance IoU (MPDIoU) is introduced as an improved loss function to make up for the shortcomings of GIoU [24] in small target detection and convergence efficiency.

II. Method Introduction

A. Constructing VAS-DETR

In this paper, we propose VAS-DETR(Variable Scale Detection Transformer), an improved model based on RT-DETR [25], to address the challenges of low detection precision and difficulty in identifying surface defects in hot-rolled steel. The VAS-DETR structure is shown in Figure 1, and its backbone network consists of four stages. Each Efficient RepAttention Block (ERA Block) represents a stage in Figure 1. ERA Block integrates RepConv [27] with Efficient Multi-head Attention (EMA) [28] implementation to ensure the backbone network's ability of extracting image features while improving the inference efficiency. The features of stages 2, 3 and 4 are input into an efficient hybrid encoder consisting of Multi-Scale Feature Integration (MSFI) and Multi Scale Atrous Fusion (MSAF) to convert them into image feature sequences. Then, uncertainty-minimum query selection selects a fixed number of encoder features as initial object queries for the decoder. Finally, the decoder with the auxiliary prediction header iteratively optimizes the object query to generate categories and boxes. In this study, MPDIoU loss function is chosen to replace the original loss function to make up for the shortcomings of the original loss function in small target detection and convergence efficiency. The design and performance of the ERA Block, MSFI and MSAF modules as well as the selection of the loss function will be analyzed in detail in the subsequent part of this paper, and their superiority in industrial defect detection will be verified.

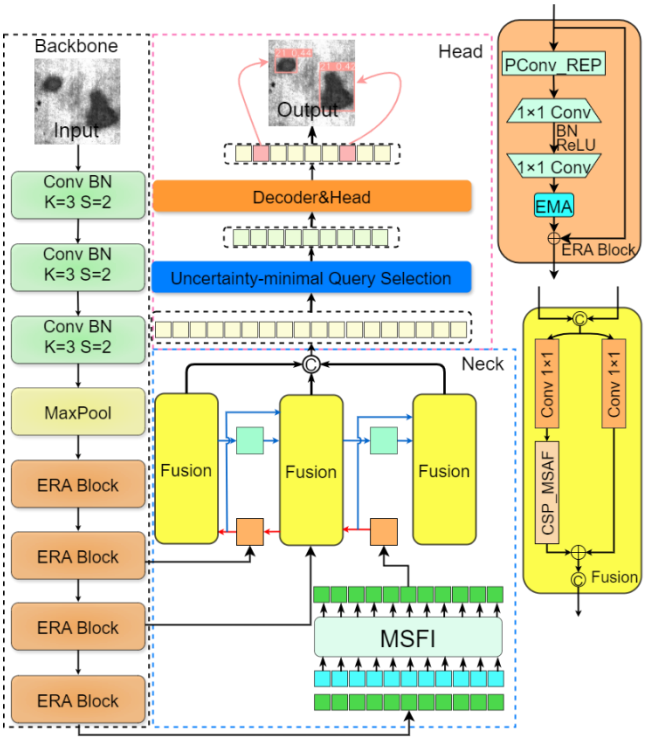


Fig. 1. Network Architecture of VAS-DETR

B. Backbone

To improve the efficiency of model calculation and reduce redundant calculation. In this study, the backbone network of RT-DETR is optimized based on Partial Convolution (PConv) proposed in 2023 and FasterNet [26]. Specifically, this study introduced two key submodules, RepConv [27] and EMA (Efficient Multi-head Attention) [28]. It is named Efficient RepAttention Block (ERA Block).

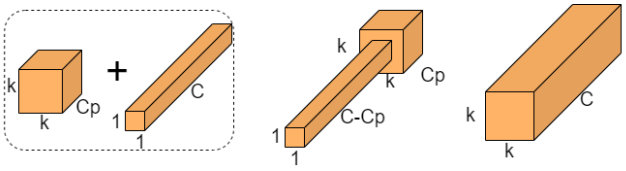
The original FasterNet blocks consist of a PConv layer followed by two PWConv layers. The normalization and activation layers are placed only after the middle layer to preserve the feature diversity and achieve lower latency [26]. PConv divides the input feature channels into processed and unprocessed parts, and only the selected part of the channel performs the convolution operation while keeping the characteristics of the other channels unchanged. PConv achieves flexible feature extraction and fusion in this way. For an input feature map x∈R^(C×H×W), PConv divides the channel dimension C into two parts: the convolved portion x_p∈R^(C×H×W) and the unprocessed portionx_u ∈R^(C×H×W). A regular convolution is applied to x_p as shown in Equation (1), where W is the convolution kernel weight and b is the bias. The final output of PConv is shown in Equation (2).

$$y_p = W * x_p + b \tag{1}$$

$$y = Concat(y_p, x_u) \quad (2)$$

Pointwise convolution (PWConv) aims to make efficient use of information across all channels. On the input feature map, the joint effective receptive field of PConv and PWConv together presents a convolution pattern similar to "T" shape. In this way, the pattern is more centered than regular convolution, as shown in Figure 2.

In Figure 2 (a) A PConv followed by a PWConv. (b) A T-shaped Conv, which spends more computation on the center position compared to a regular Conv (c) [28].



(a) PConv+PWConv (b) T-shaped Conv (C) regular Conv

Fig. 2. Comparison of Different Convolutions.

In this study, the traditional 3×3 convolution in PConv is replaced by RepConv. As shown in Figure 3 RepConv has a convolutional module designed to separate the training and inference phases, which captures rich feature information through a multi-branch structure in the training phase of RepConv and fuses the multi-branches into an equivalent standard convolutional kernel in the inference phase, which drastically reduces the amount of computation and inference latency. Given input features $X \in R^{C\times H\times W}$ and output features $Y \in R^{c'\times H\times W}$, RepConv performs feature extraction during training as shown in Equation (3). During inference, its multi-branch structure is reparameterized into a single convolution operation as shown in Equation (4), where W_{eq} and b_{eq} the equivalent convolution kernel and bias.

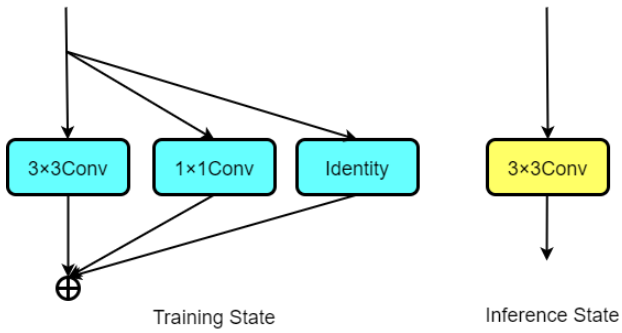


Fig. 3. Training and Inference Stages of RepConv

$$\begin{cases} Y_{3\times3} = W_{3\times3} * X + b_{3\times3} \\ Y_{3\times3} = W_{1\times1} * X + b_{1\times1} \\ Y_{id} = X + b_{id} \\ Y = Y_{3\times3} + Y_{1\times1} + Y_{id} \end{cases} \quad (3)$$

$$\begin{cases} W_{eq} = W_{3\times3} + W_{1\times1} + W_{id} \\ b_{eq} = b_{3\times3} + b_{1\times1} + b_{id} \\ Y = W_{eq} * X + b_{eq} \end{cases} \quad (4)$$

This design offers significant advantages in scenarios with high real-time performance requirements. Therefore, this study incorporates RepConv to optimize the model architecture. Additionally, the Efficient Multi-head Attention (EMA) mechanism is integrated into the Faster_Block module to enhance feature extraction. The EMA module achieves efficient modelling of local and global features through parallel multi-scale branching and cross-space learning methods. Channel dimensionality reduction operations are avoided, reducing information loss while keeping the computational overhead low. Compared to other attention modules, EMA has lower parametric counts and FLOPs while maintaining strong feature modelling capabilities. Its design is well suited to meet the dual requirements of real-time and accuracy for steel defect detection tasks in industrial environments. In summary, this study designs the ERA Block module to replace the Basic Block module in the original model, and its structure is shown in Figure 4.

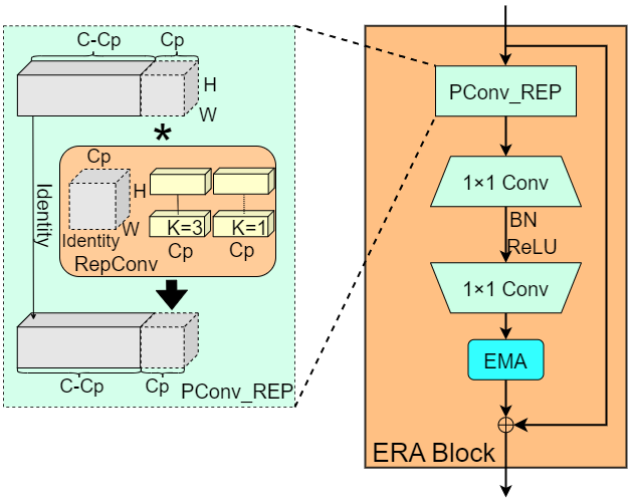


Fig. 4. The Architecture of the ERA Block Module

C. Optimization of the Attention-based Intra-scale Feature Interaction (AIFI) Module

AIFI is a Transformer-based network module that enables global feature modelling through standard Multihead Attention. AIFI embeds spatial information through 2D sine-cosine positional coding to capture long range dependencies. AIFI is able to preserve global features while possessing adaptability to input data, making it suitable for global context modelling tasks. However, due to its bias towards global features resulting in low sensitivity to small targets, there may be a decrease in the detection accuracy of subtle defects on the steel surface, such as cracks and scratches.

To address the issue of low detection accuracy caused by varying defect sizes in steel surface inspection tasks, this study improves the AIFI module by drawing on the M2SA module proposed by Wu et al. A new Multi-Scale Feature Integration (MSFI) module is proposed, with its structure illustrated in Figure 5 [29].

In the MSFI module, the input source features first undergoes a multi-scale multi-head self-attention module and a channel attention module to extract local and global contextual features (denoted as src_2). These features are then updated through residual connections and further enhanced by a feed-forward network, producing src_3. Finally, residual connections are used to fuse the features, generating the output features. The overall process is expressed by the following equations:

$$src_2 = Mutilscal\_MHSA(src) \quad (5)$$

$$src = src + Dropout(src_2) \quad (6)$$

$$src_3 = FC_2(Dropout(ReLU(FC_1(src)))) \quad (7)$$

$$Output = src + Dropout(src_3) \quad (8)$$

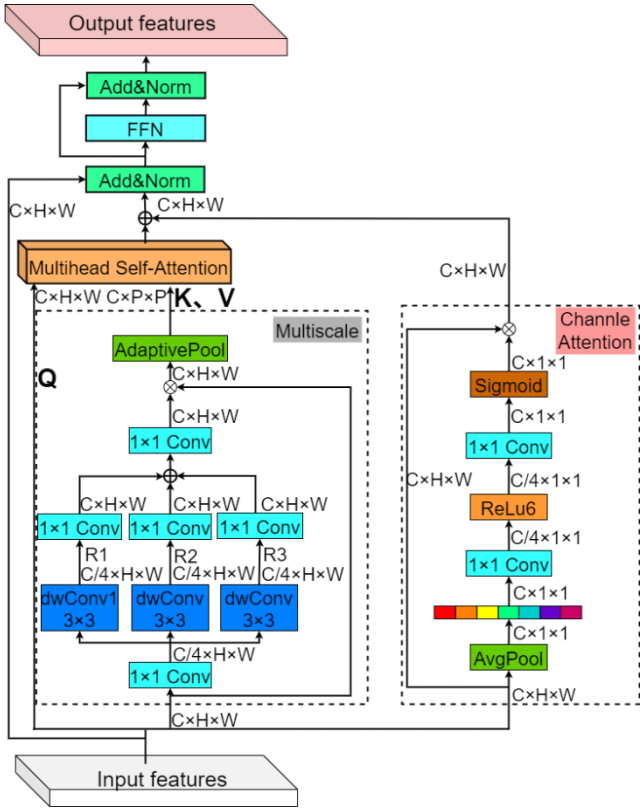


Fig. 5. Structure of the MSFI Module

MSFI differs from AIFI in the calculation of multi-head self-attention. Unlike traditional methods that directly generate query (Q), key (K), and value (V) tensors from input feature X, MSFI generates query tensors from X and key and value tensors from P, where P is a feature with lower resolution but rich multi-scale context information obtained by Multiscale. This process is formally expressed in Equations (9) and (10), where d_k denotes the channel dimension of the key, and $\sqrt{d_k}$ is used for approximate normalization.

$$(Q,K,V) = (XW_q, PW_k, PW_v) \tag{9}$$

$$Attention = Softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V \tag{10}$$

In addition, MSFI introduces a channel attention branch to enhance information extraction along the channel dimension. The features generated by this branch are expressed in Equations (11), (12), and (13).

$$X_p = Avgpool(X) \tag{11}$$

$$X_C = ReLu6\left(Conv_{1\times1}(X_p)\right) \tag{12}$$

$$Attention_c = Sigmoid(Conv_{1\times1}(X_c)) \otimes X) \tag{13}$$

The fusion of Multihead Self-Attention and channel attention features can be expressed as Equation (14).

$$Output = Attention + Attention_c \tag{14}$$

MSFI module combines local features and global features through multi-scale convolution and Multihead Self-Attention to enhance the spatial context awareness ability of the model, and can better capture multi-scale spatial features and context information. Compared to the AIFI module in RT-DETR, it is more suitable for steel surface defect detection tasks.

D.Multi Scale Atrous Fusion (MSAF)

To effectively capture multi-scale information and address the diverse sizes of steel surface defects, this study proposes the Multi-Scale Atrous Fusion (MSAF) module. The MSAF structure is shown in Figure 6. Three parallel convolutional paths are designed in this module, and convolution operations with cavity rates of 1, 2 and 3 are adopted respectively: The first path is an ordinary 3×3 convolution (dilation rate = 1) mainly used to extract local features; The second and third paths use dilation convolution with a dilation rate of 2 and 3, respectively, to gradually expand the sensory field in order to capture medium- and large-scale contextual information. In order to balance the computational complexity and the feature contribution of each path, the number of output channels of the second and third paths is halved. Finally, the three path outputs are spliced in the channel dimension and fused by 1×1 convolution to ensure that the number of output channels is consistent with the input.

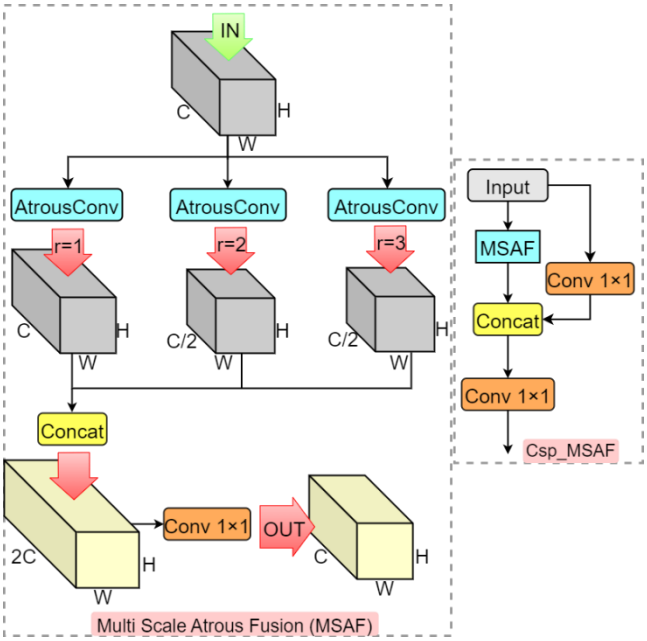


Fig. 6. Structure of the CSP-MSAF Module

To save computational resources and improve inference efficiency, this study combines the MSAF module with the Cross Stage Partial (CSP) structure to construct the CSP-MSAF module, which replaces the Repc3 module in the RT-DETR model. The structure is shown in Figure 6. The CSP-MSAF module achieves efficient feature learning through a branching structure and enhances receptive field representation and detail capturing capability by fusing multi-scale features. With minimal changes to computation and model complexity, this module effectively integrates features from different receptive fields, significantly improving multi-scale steel defect detection capabilities and providing technical support for high-precision detection of diverse defects.

In the MSAF module, the receptive fields of atrous convolution layers with dilation rates of 1, 2, and 3 are illustrated in Figure 7. Red dots represent convolution kernel "pixels," while green areas denote their receptive fields in the original input. (a) For a dilation rate of 1, the receptive field is 3×3, identical to standard convolution. (b) At a dilation rate of 2, the actual 3×3 kernel achieves a receptive field of 7×7. (c) With a dilation rate of 3, the 3×3 kernel expands the receptive field to 11×11 [30].

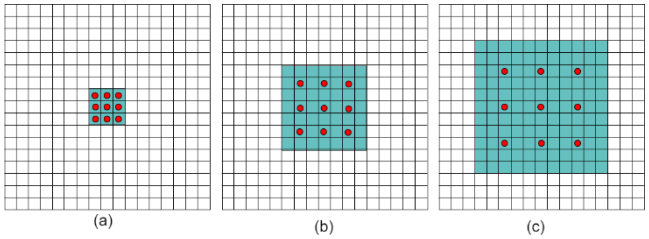


Fig. 7. Schematic Diagram of Receptive Fields in Dilated Convolutions

E. Loss Function

The RT-DETR model uses Generalized Intersection over Union (GIoU) as the loss function, but its optimization relies on the area difference between the predicted box and the minimum enclosing rectangle. This leads to weak optimization performance for small objects and limited gradient changes in the early stages of training, resulting in slow convergence of the model. To address these issues, this study conducts experimental comparisons of GIoU, Inner IoU [31], Focaler IoU [32], MPDIoU [33], Inner-MPDIoU, and Focaler-MPDIoU, and ultimately selects MPDIoU as the loss function for the improved model. MPDIoU introduces a penalty term based on the distance between the vertices of the bounding box in addition to the standard IoU, which measures the position deviation between the predicted and ground-truth boxes. This effectively mitigates the sensitivity issue caused by insufficient bounding box overlap. The method accelerates model convergence and improves small-object detection performance, providing an effective solution for optimizing the RT-DETR model. Its computation is given by formula (15).

$$MPDIoU = \frac{1}{N}\sum_{i=n}^{N}\frac{\max(A_i \cap B_i)+\lambda\cdot\min(A_i \cap B_i)}{A_i \cup B_i} \quad (15)$$

III. Experiment And Results

A. Data and Experimental Setup

The dataset used in this study is the surface defect database released by Northeastern University (NEU) [23], which contains six typical surface defects of hot-rolled steel strips: rolling scale (RS), patch (Pa), crack (Cr), pitting surface (PS), inclusion (In), and scratch (Sc). The database includes 1,800 grayscale images, with 300 samples for each defect type. To address the limited number of samples and reduce overfitting, data augmentation techniques such as rotation, translation, and flipping were applied to the 1,800 images, resulting in 3600 images. The images were split into training, testing, and validation sets at a ratio of 6:2:2, with 2160 images for training, 720 for testing, and 720 for validation, ensuring each defect type was evenly distributed across all sets. This distribution facilitates precise feature learning for each defect type and ensures accurate model evaluation.

The experiments were conducted on a Windows 10 operating system with an Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz, 2.89 GHz (two processors), and an NVIDIA GeForce RTX 4090 GPU. The deep learning framework used was Python version 3.11.9, PyTorch version 2.3.1, and CUDA version 11.2. During the training phase, the initial learning rate was set to 0.0001, momentum to 0.8, input image size was 640×640, batch size was 16, and the number of epochs was 450.

B. Contrast experiment

The GIoU loss function has limitations in gradient variation and sensitivity to small objects, leading to slower convergence and reduced optimization for small targets. To overcome these issues, this study introduces the MPDIoU loss function.

To demonstrate the superiority of the MPDIoU loss function in the multi-scale steel defect detection task. In this study, comparison experiments are conducted on the NEU-DET dataset using a variety of loss functions, including GIoU, Inner IoU [31], Focaler IoU [32], MPDIoU [33], Inner-MPDIoU, and Focaler-MPDIoU. The experimental results are shown in Table 1. From the table it is easy to see that the MPDIoU loss function is most suitable for the steel defect detection task.

TABLE I
Comparison of The Performance of Various Loss Functions in Steel Defect Detection

	GIoU	Inner IoU	Focaler IoU	**MPDIoU**	Inner-MPDIoU	Focaler-MPDIoU
mAP_{50}	0.814	0.817	0.809	**0.818**	0.805	0.819
$mAP_{50:95}$	0.524	0.512	0.509	**0.523**	0.514	0.521

TABLE II
Comparison of Model Accuracy and Computational Performance in Defect Detection Tasks

Model	mAP_{50}	$mAP_{50:95}$	GFLOPs	Parameters
RT-DETR-R18	0.814	0.524	57.3	19974480
YOLOv11-l [34]	0.828	**0.596**	86.6	25283938
YOLOv11-m [34]	0.823	0.594	67.7	20034658
YOLOv10-l [35]	0.826	0.592	127.2	25774580
YOLOv10-m [35]	0.824	0.589	64.0	16491076
YOLOv9-c [36]	0.827	0.588	103.7	25533842
YOLOv9-m [36]	0.821	0.581	76.5	20017330
YOLOv8-l [37]	0.828	0.593	164.8	43611234
YOLOv8-m [37]	0.821	0.591	78.7	25843234
YOLOv5-l [38]	0.816	0.560	109.1	53136034
YOLOv5-m [38]	0.815	0.545	64.4	25068610
VAS-DETR	**0.846**	0.548	**50.0**	**16609608**

In order to prove the performance of the proposed VAS-DETR model in terms of detection accuracy, especially the adaptability of medium to high complexity models in real-time scenarios, this study compared the existing mainstream target detection models. The experimental results are shown in Table 2, from which it can be seen that VAS-DETR achieves a significant improvement between performance and efficiency. In terms of detection accuracy, VAS-DETR's mAP_{50} reaches 0.846, an improvement of 4.44% over the baseline model RT-DETR-R18, outperforming YOLOv11-l by 2.2% and YOLOv8-l by 1.8%. Additionally, VAS-DETR's $mAP_{50:95}$ also increased by 4.58% compared to RT-DETR-R18, indicating a clear advantage in comprehensive detection capability. Compared to other high-precision models like YOLOv11-l and YOLOv8-l, VAS-DETR reduces computational complexity while improving accuracy. Its GFLOPs is 50.0, only 30.4% of YOLOv8-l and 57.7% of YOLOv11-l, with parameter count reduced to 16.6 million, which is 38.1% of YOLOv8-l and 65.7% of YOLOv11-l. This allows VAS-DETR to operate more efficiently in resource-constrained environments.

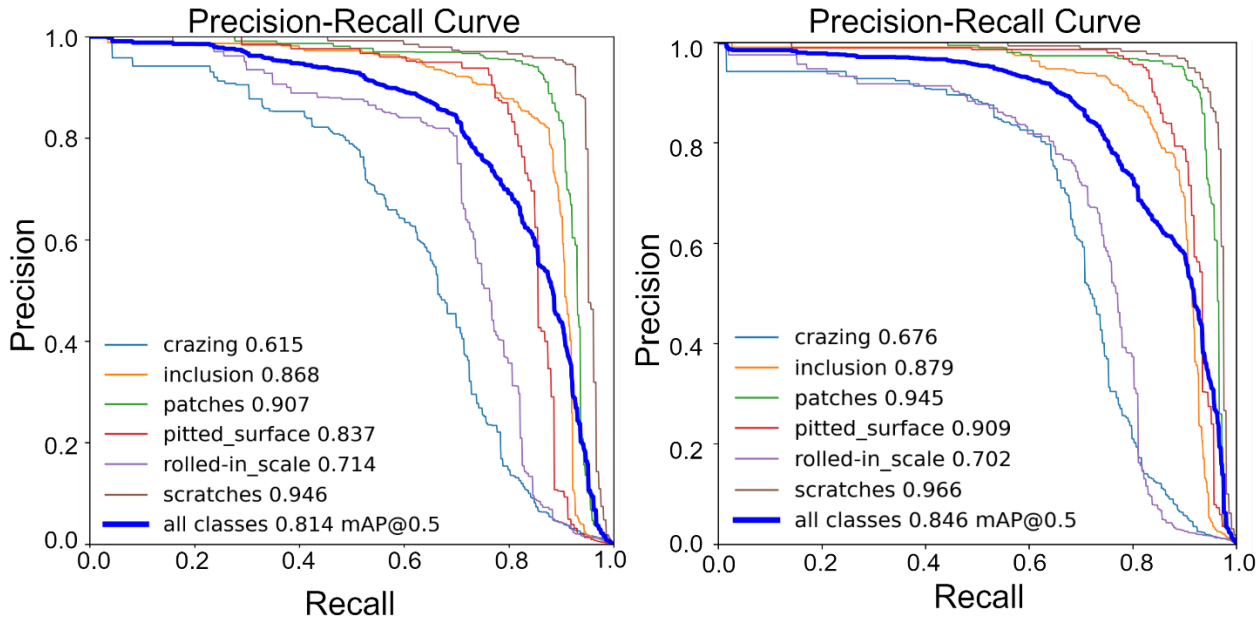


Fig. 8. Comparison of Precision Recall Cure between RT-DETR and VAS-DET

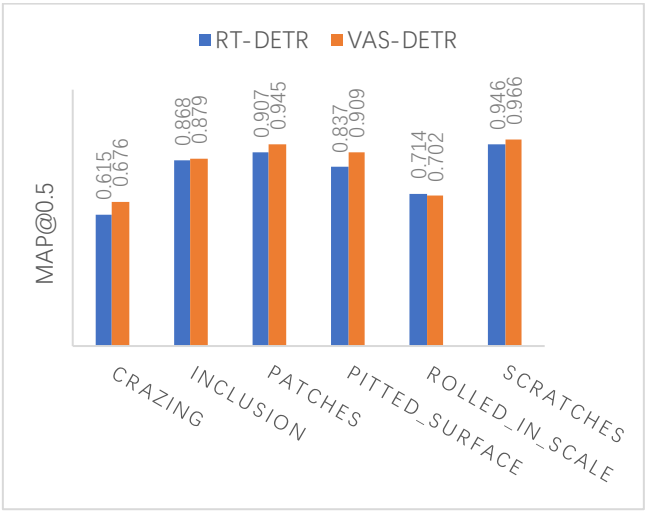


Fig. 9. Comparison of the mAP_{50} performance between the RT-DETR and VAS-DETR models

In order to prove the effectiveness and robustness of the improved strategy, comparative experiments were conducted between RT-DETR and VAS-DETR on the public data set NEU-DET, and the experimental results were visualized by Precision-Recall Curve, as shown in Figure 8. In order to facilitate the comparison of the changes of various types of defects mAP_{50} this study draws the histogram as in Figure 9. By analysing the experimental results through the histograms, it can be intuitively observed that the improved model VAS-DETR significantly outperforms the baseline model RT-DETR in the detection of steel surface defects.

In terms of specific categories, the detection accuracy of crazing and pitted_surface defects increased by 0.061 and 0.072, respectively, indicating that the improved model's ability to capture features of these categories has been significantly enhanced. This performance improvement is mainly due to the Multi Scale Atrous Fusion (MSAF) module designed in this study. The module introduces a parallel cavity convolution mechanism in the process of multi-scale feature fusion, which enhances the adaptability of the model to defects of different sizes and shapes, thus improving the characterization ability of complex surface features. Except that the accuracy of rolled-in_scale is slightly reduced from 0.714 to 0.702, the detection accuracy of the other five types of defects has been improved to varying degrees. This distribution trend indicates that VAS-DETR can improve the detection performance of most defect categories comprehensively after optimizing the model structure. In summary, the improvement of VAS-DETR can not only effectively improve the detection accuracy, but also show stronger adaptability to complex features in specific categories. This shows that the optimization of the VAS-DETR model can achieve a better

balance in the overall detection accuracy, especially for some hard-to-detect defect categories.

C. Ablation experiment

In order to evaluate the necessity of RepConv, EMA, MSAF, MSFI and MPDIoU modules, this paper designed 18 rounds of ablation experiment. The results are shown in Table 3. The ERA Block module and MSAF module can significantly improve the detection accuracy of steel surface defects while reducing the complexity of the model. The ERA Block module integrates RepConv and EMA modules. When only RepConv or EMA were applied to the backbone network independently, the overall detection accuracy of the model decreased. This is due to their respective limitations in feature extraction. RepConv is better at capturing local features efficiently while EMA focuses on modeling the global context. Using one of them alone will lead to the lack of balance between global and local feature modeling, and weaken the overall detection ability. However, when RepConv and EMA are integrated into the backbone network at the same time, they can complement each other and significantly improve the feature expression ability of the model.

In addition, the introduction of MSFI module and MPDIoU loss function further improves the defect detection ability of the model without increasing the complexity of the model. Compared with the baseline model, the performance of VAS-DETR is significantly improved in several indicators, including mAP_{50} from 0.814 to 0.846 and $mAP_{50:95}$ from 0.524 to 0.548. The effectiveness of all modules is proved by ablation experiments.

TABLE III
ABLATION STUDY ON THE VAS-DETR MODEL

Methodology	mAP_{50}	$mAP_{50:95}$	GFLOPs	Parameters
+ERA Block+MSAF+MPDIoU	0.835	0.541	49.8	16574024
+ERA Block +MSAF+MSFI	0.843	0.545	50.0	16609608
+ERA Block+MSFI+MPDIoU	0.826	0.538	51.6	16937288
+MSAF+MSFI+MPDIoU	0.832	0.538	55.5	19682384
+ERA Block +MSFI	0.822	0.524	51.6	16937288
+ERA Block +MSAF	0.834	0.539	49.8	16574024
+ERA Block +MPDIoU	0.823	0.537	51.5	16901704
+MSAF+MSFI	0.831	0.538	55.5	19587368
+MSAF+MPDIoU	0.824	0.534	57.1	19915048
+MSFI+MPDIoU	0.819	0.534	57.1	19915048
+ERA Block	0.820	0.530	51.5	16901704
+MPDIoU	0.818	0.523	57.3	19974480
+MSFI	0.821	0.527	57.1	19915048
+MSAF	0.825	0.523	55.3	19551784
+EMA	0.810	0.518	51.8	16996240
+RepConv	0.810	0.518	49.8	16886960
RT-DETR-R18	0.814	0.524	57.3	19974480
VAS-DETR	**0.846**	**0.548**	**50.0**	**16609608**

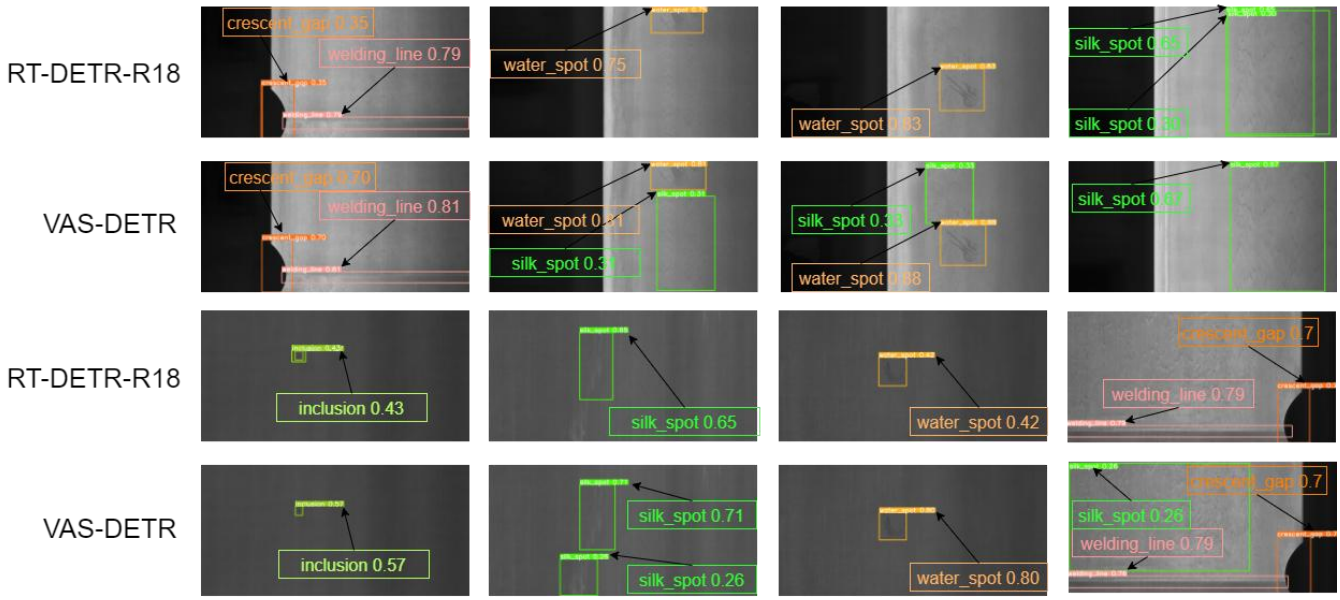


Fig. 10. Visual comparison of detection results on the GC10-DET dataset

D.Evaluation of Model Generalization Performance

To evaluate the generalization capability of the proposed VAS-DETR model across different data distributions, experiments were conducted not only on the NEU-DET dataset but also on the GC10-DET dataset. These experiments provide an effective means to assess the model's performance when facing previously unseen data. The training results of VAS-DETR on the GC10-DET dataset are shown in Figure 11. Experimental results demonstrate that the VAS-DETR model achieves varying degrees of improvement in the mAP_{50}metric across different defect categories compared to the baseline model. The overall average mAP_{50} increased from 0.654 to 0.678, representing a 2.4 percentage point improvement, which reflects a stronger generalization capability. To more intuitively illustrate the advantages of VAS-DETR, visual comparisons between the VAS-DETR and the baseline model were performed, as shown in Figure 10. The comparison clearly indicates that VAS-DETR exhibits fewer issues such as overlapping bounding boxes, low confidence scores, and missed detections. These results strongly support the conclusion that the VAS-DETR model possesses robust generalization ability and reliability when applied to unfamiliar datasets.

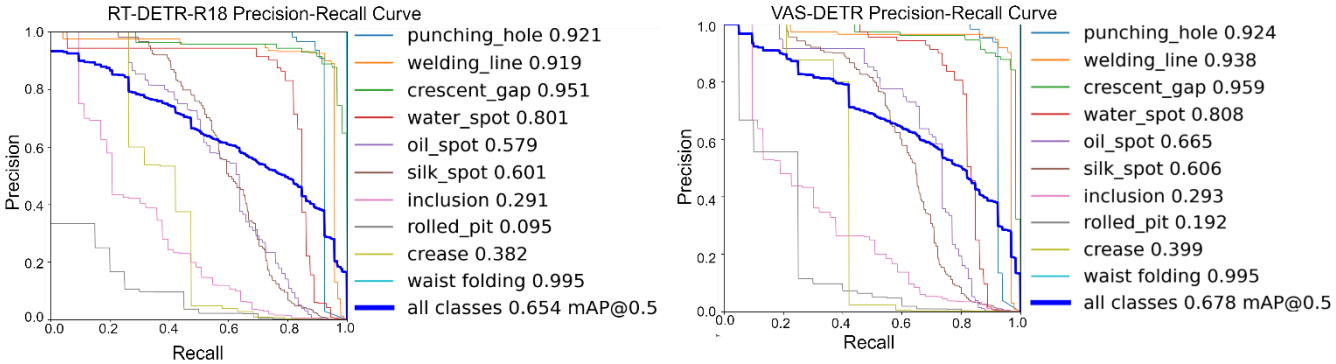


Fig. 11. Comparison of training results on the GC10-DET dataset

E. Visualisation and analysis of test results

In order to show the performance of the VAS-DETR model, this study visualises the detection results of VAS-DETR as well as models of the same magnitude on the NEU-DET dataset. The homogeneous models selected in this study include RT-DETR-R18, YOLOv11-m, YOLOv10-m, YOLOv9-m, YOLOv8-m, and YOLOv5-m.

The visualisation results are shown in Figure 12. A class-by-class defect analysis found that there were omissions in the YOLO series in the crazing class example. The missing detection of YOLO series may be due to the fuzzy characteristics of this type of defect, and the model fails to capture relevant information effectively. Although RT-DETR is able to detect all defects in the image, it lacks in positioning accuracy. The detection frame generated by RT-DETR did not accurately describe the location of the defect, and the defect detection frame in the lower area of the image overlapped. This phenomenon may be due to the failure of the loss function to impose sufficient constraints on the redundant prediction in the process of supervising the decoder to generate the detection box. In contrast, the VAS-DETR model proposed in this study can not only detect two defects in the image, but also accurately depict the minimum external rectangle of the defect. This result shows that the optimization of RT-DETR model on feature extraction capability significantly improves its performance. RT-DETR effectively reduces the redundant prediction by improving the loss function, which makes the detection effect of the model more accurate and robust.

In the inclusion category example, the defect is located in the bottom left edge of the image and is small in size. Compared with the visualization results of the other models, only VAS-DETR successfully detected the defect and had the highest confidence score. In the example of rolled-in_scale and patches categories, the VAS-DETR model was able to detect all defects completely despite the large number of defects and significant size differences. Based on the comparison of visual results of defects of inclusion, roll-in_scale and patches, it can be concluded that the VAS-DETR model is highly sensitive to defects of all dimensions.

In the scratches category example, the scratches varied in size and depth. Not only did VAS-DETR successfully detect the small scratches missed by other models, but the confidence scores of VAS-DETR were higher than those of other models. In the example of pitted_surface category, the performance of the other five models is basically the same, except for one missing detection in YOLOv5m and YOLOv9-m, due to the obvious defect characteristics of this category. To sum up, it can be concluded that VAS-DETR model has excellent performance in the detection of various defects, and solves the problem of low accuracy of individual defects detection.

In addition to displaying the performance of VAS-DETR in the above ways, this study also compared the visual results of the VAS-DETR and RT-DETR-R18 models through heat maps, as shown in Figure 13. The purpose of this analysis is to analyze the feature region of interest and localization performance of the model in defect detection and to verify the effectiveness of the optimization of RT-DETR-R18 in this paper. It is not difficult to see from Figure 13 that VAS-DETR has a more accurate focus area in crazing and inclusion defect detection, and the highlighted part can accurately cover the defect area. Some of the highlighted areas of RT-DETR-R18 deviate from the actual defects and some of the highlighted areas are mis-focused on complex backgrounds. VAS-DETR's heat map focuses on the defect itself and its background suppression and feature focus consistency are significantly better than that of RT-DETR-R18. VAS-DETR achieves optimal results in detecting defects of different sizes.

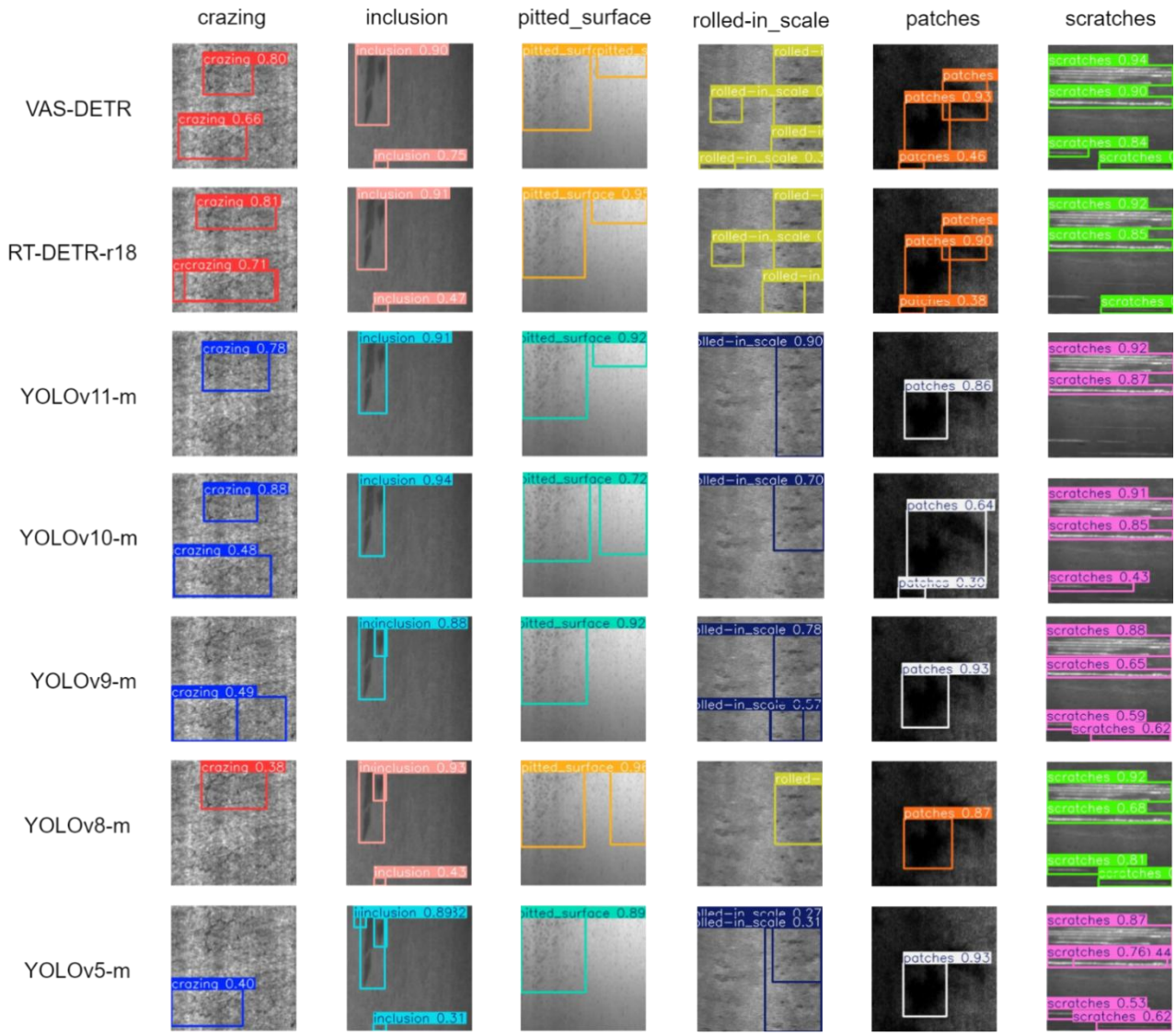


Fig. 12. Visual comparison of detection results

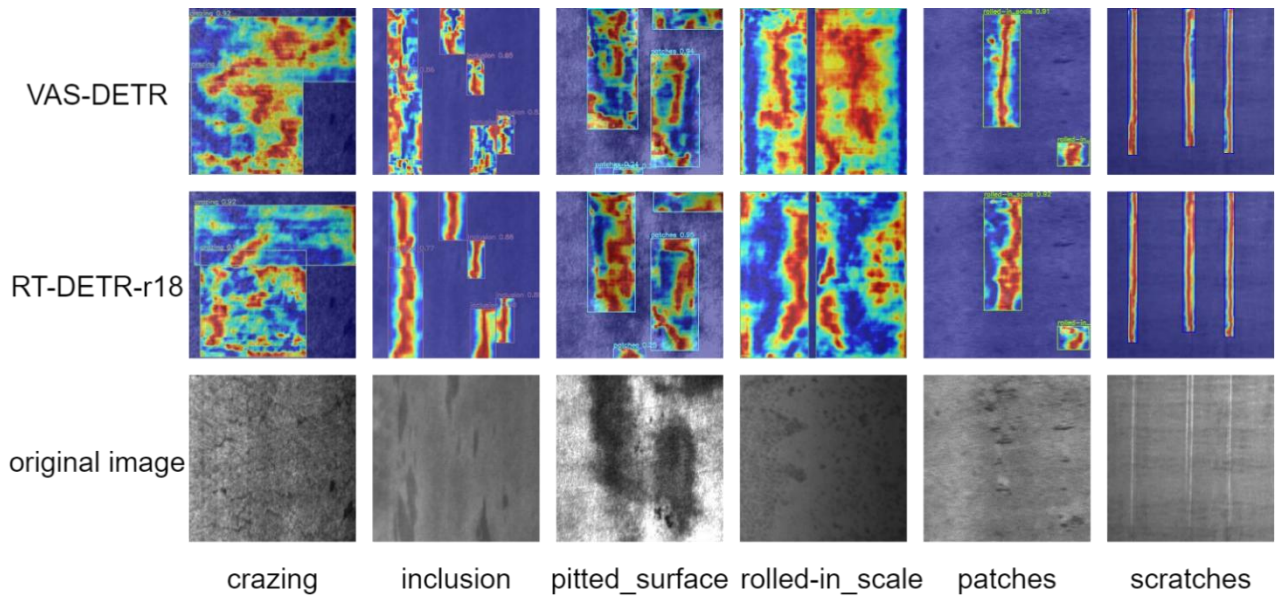


Fig. 13. Comparison of Heatmap Results between VAS-DETR and RT-DETR-R18 Models

IV. Discussion

The experimental results show that the VAS-DETR model proposed in this paper can significantly improve the accuracy of steel defect detection without increasing the amount of calculation. Ablation experiments verify the effectiveness of the loss functions of ERA Block, MSAF, MSFI and MPDIoU. These improvements play an important

role in improving the performance of the model. The comparison of the visual results of various defects proves that the VAS-DETR model has better performance for each kind of defects. The detection accuracy of crazing and pitted_surface defects in VAS-DETR model is significantly higher than that of RT-DETR-R18. This is sufficient to show that the VAS-DETR model proposed in this study reduces the influence of size difference on the detection accuracy.

Although VAS-DETR showed significant advantages in detection performance, it still had some shortcomings. Compared with other models, the missing phenomenon of VAS-DETR has been greatly reduced, but there are some missing phenomena. As shown in Figure 1, scratches and roll-in_scale defects are still missed. This is why VAS-DETR only reduces the impact of size differences on accuracy rather than completely resolving the impact of size differences on accuracy. Future research will further optimize the feature extraction mechanism and focus on reducing the phenomenon of missing and false detection, so as to improve the generalization ability and robustness of the model in complex scenarios, so as to promote the development of the industrial defect detection field.

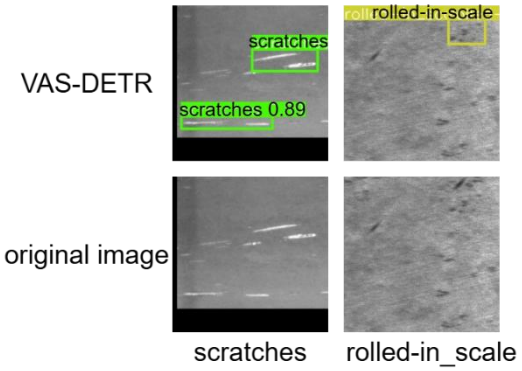


Fig. 14. Examples of poor defect detection results

V. Conclusion

In this paper, an improved target detection algorithm VAS-DETR based on RT-DETR-R18 is proposed to solve the problem of low detection accuracy due to large size difference of steel surface quality defects. Firstly, the ERA Block module is introduced into the backbone network, which uses RepConv to realize the convolution design of separating the training phase from the inference phase. This allows the ERA Block module to improve the feature extraction capability of the overall network while reducing the amount of computation. The MSFI module combined with M2SA module enhances the model's ability to focus on various defects. The MSAF module can effectively expand the receptive field without increasing the amount of computation through the design of parallel cavity convolution, thus improving the multi-scale defect detection ability in the fusion of global and local information. The MPDIoU loss function is optimized for the deviation between the predicted frame and the real frame, which effectively alleviates the problem of insufficient overlap of boundary frames.

In this study, a large number of experiments show that compared with RT-DETR, VAS-DETR improves the mAP_{50} and $mAP_{50:95}$ by 3.93% and 4.58%, respectively, while GFLOPs and Parameters decrease by 12.74% and 16.85%. The effectiveness of each module proposed in this study was further verified by ablation experiments. In comparison with the model of the same magnitude, VAS-DETR can show higher detection accuracy in the face of various defects of different sizes, and effectively reduce missed and false detection. Although the method proposed in this study can significantly improve the accuracy of the model, the deployment of edge devices with limited computing resources still faces certain challenges. The future research direction will focus on the further development of multi-scale defect detection and real-time detection algorithms. At the same time, it should also strengthen the deep integration with the production process to promote the more efficient and widespread detection of steel surface defects, which will have a profound impact on the steel industry and intelligent manufacturing.

References

[1] W. Lin, H. H. Yu, and S. You. “A review of metal surface defect detection based on computer vision,” Acta Autom. Sin, vol. 50, pp. 1261-1283, 2024.

[2] L. Z. Guo, W. X. Mei, M. Hassaballah, L. Y. Hong, and J. X. Song. “A deep learning model for steel surface defect detection,” Complex & Intelligent Systems, vol. 10, no. 1, pp. 885-897, 2024.

[3] L. Hong, Y. Yun, C. Xiang, and K. Y. Hua. “Investigation on the formation mechanism of crack indications and the influences of related parameters in magnetic particle inspection,” Applied Sciences, vol. 10, no. 19, pp. 6805, 2020.

[4] K. Saravanakumar, J. Medoline, S. Ananth, K. M. S. Priyanka, and M. Dharshini. “Defect Identification in Weld Beads of Aluminium 5356 Using Liquid Penetrant Test,” ICAMDMS 2024: Proceedings of the International Conference on Advancements in Materials, Design and Manufacturing for Sustainable Development, ICAMDMS 2024, 23-24 February 2024, Coimbatore, Tamil Nadu, India. European Alliance for Innovation, pp. 341, 2024.

[5] L. Minhhuy, P. H. Pham, L. Q. Trung, S. P. Hoang, L. D. Minh, Q. V. Pham, and V. S. Luong. “Enhancing corrosion detection in pulsed eddy current testing systems through autoencoder-based unsupervised learning,” NDT & E International, vol. 146, pp. 103175, 2024.

[6] H. Z. Long, L. S. Song, C. X. Ming, H. B. Chi, S. Jiao, and Z. Q. Gang. “DFW-YOLO: YOLOv5-based algorithm using phased array ultrasonic testing for weld defect recognition,” Nondestructive Testing and Evaluation, vol. 40, no. 3, pp. 2516-2539, 2025.

[7] R. Z. He, F. F. Zhou, Y. Ning, and W. Yu. “State of the art in defect detection based on machine vision,” International Journal of Precision Engineering and Manufacturing-Green Technology, vol. 9, no. 2, pp. 661-691, 2022.

[8] C. Y. Jun, D. Y. Yuan, Z. Fan, Z. E. H, W. Z. Nan, and S. L. Hao. “Surface defect detection methods for industrial products: A review,” Applied Sciences, vol. 11, no. 16, pp. 7657, 2021.

[9] Z. W. Xuan, S. B. Chao, Z. X. He, L. Ly, and F. S. Wen. “ALdamage-seg: A Lightweight Model for Instance Segmentation of Aluminum Profiles,” Buildings, vol. 14, no. 7, pp. 2036, 2024.

[10] Girshick, Ross. "Fast r-cnn," Proceedings of the IEEE international conference on computer vision. 2015.

[11] C. X. Lei, and A. Gupta. “An implementation of faster rcnn with study for region sampling,” arxiv preprint arxiv:1702.02138, 2017.

[12] H. K. Ming, G. Gkioxari, P. Dollar, and R. Girshick. “Mask r-cnn,” Proceedings of the IEEE international conference on computer vision. 2017.

[13] L. Wei, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, F. C. Yang, and A. C. Berg. “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.

[14] J. Terven, D. C. Esparza, and J. A. R. Gonzalez. “A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS,” Machine learning and knowledge extraction, vol. 5, no. 4, pp. 1680-1716, 2023.

[15] Z. X. Zhou, S. W. Jie, L. L. Wei, W. X. Gang, and D. J. Feng. “Deformable DETR: Deformable Transformers for End-to-End Object Detection,” arxiv preprint arxiv:2010.04159, 2020.

[16] C. C. Hong, W. S. Feng, and H. S. Zhou. “An improved faster RCNN-based weld ultrasonic atlas defect detection method”. Measurement and control, vol. 56, no. 3, pp. 832-843, 2023.

[17] W. Hao, L. M. Jiao, and W. Z. Bo. “Rail surface defect detection based on improved Mask R-CNN,” Computers and Electrical Engineering, vol. 102, pp. 108269, 2022.

[18] B. X. Ning, Z. Z. Yuan, and W. Y. Peng. “An Improved SSD Model for Small Size Work-pieces Recognition in Automatic Production Line,” Journal of Internet Technology, vol. 25, no. 2, pp. 215-222, 2024.
[19] Z. H. Xin, C. X. Xin, C. Hao, D. Y. Xian, and J. Z. Si. “MD-YOLO: Surface Defect Detector for Industrial Complex Environments,” Optics and Lasers in Engineering. vol. 178, pp. 108170, 2024.
[20] W. L. Yang, Z. G. Xue, W. W. Jun, C. J. Yuan, J. X. Yao, Y. Hai, and H. Z. Cheng. “A defect detection method for industrial aluminum sheet surface based on improved YOLOv8 algorithm,” Frontiers in Physics, vol. 12, pp. 1419998, 2024.
[21] H. Kong, and C. You. “Improved steel surface defect detection algorithm based on YOLOv8,”. IEEE Access, vol. 12, pp. 99570-99577, 2024.
[22] C. Zhang, Y. Fan, D. S. Sen, C. F. Lin, and H. J. Zhou. “An Efficient Real-time Metal Crack Detection Model Based on RT-DETR,” 2024 9th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS), Dalian, China, pp. 220-224, 2024.
[23] H. Yu, S. K. Chen, M. Q. Gang, and Y. Y. Hui. “An End-to-End Steel Surface Defect Detection Approach via Fusing Multiple Hierarchical Features,” IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 4, pp. 1493-1504, 2020.
[24] H. Rezatofighi, N. Tsoi, J. Y. Gwak, A. Sadeghian, I. Reid, and S. Savarese. “Generalized intersection over union: A metric and a loss for bounding box regression,” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 658-666, 2019.
[25] Z. Y. An, L. W. Yu, X. S. Liang, W. J. Man, W. G. Zhong, D. Q. Qing, L. Yi, and C, Jie. “Detrs beat yolos on real-time object detection,” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.16965-16974, 2024.
[26] C. J. Run, K. S. Hong, H. Hao, Z. W. Peng, W. Song, L. C. Ho, and S. H. G. Chan. “Run, don't walk: chasing higher FLOPS for faster neural networks,” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.12021-12031, 2023.
[27] M. Soudy, A. Yasmine, and B. Nagwa. "RepConv: A novel architecture for image scene classification on Intel scenes dataset." International Journal of Intelligent Computing and Information Sciences, vol. 22, no. 2, pp. 63-73, 2022.
[28] O. D. Liang, H. Su, Z. G. Zhong, L. M. Zhu, G. H. Yong, and Z. Jian. “Efficient Multi-Scale Attention Module with Cross-Spatial Learning,” ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1-5, 2023.
[29] W. H. Lin, H. Peng, Z.Min, T. W. Long, and Y. X. Yu. “CMTFNet: CNN and Multiscale Transformer Fusion Network for Remote-Sensing Image Semantic Segmentation,” IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-12. 2023.
[30] Y. Fisher, and K. Vladlen. “Multi-scale context aggregation by dilated convolutions,” arxiv preprint arxiv:1511.07122, 2015.
[31] Z. Hao, X. Cong, and Z. S. Jie. “Inner-iou: more effective intersection over union loss with auxiliary bounding box,” arxiv preprint arxiv:2311.02877, 2023.
[32] Z. Hao, and Z. S. Jie. "Focaler-iou: More focused intersection over union loss." arxiv preprint arxiv: 2401.10525, 2024.
[33] S. Ma, and Y. Xu. “Mpdiou: a loss for efficient and accurate bounding box regression,” arXiv preprint arXiv:2307.07662, 2023.
[34] R. Khanam, and M. Hussain. “Yolov11: An overview of the key architectural enhancements,” arXiv preprint arXiv:2410.17725, 2024.
[35] W. Ao, C. Hui, L. L. Hao, C. Kai, L. Z. Jia, H. J. Gong, and D. G. Guang, “Yolov10: Real-time end-to-end object detection,” Advances in Neural Information Processing Systems, vol. 37, pp. 107984-108011, 2024.
[36] W. C. Yao, Y. I. Hau, amd M. L. H. Yuan, “YOLOv9: Learning What You Want to Learn Using Programmable Gradient Infomation,” European conference on computer vision. Cham: Springer Nature Switzerland, pp. 1-21, 2024.
[37] M. Sohan, T. S. Ram, C. V. R. Reddy, “A review on yolov8 and its advancements,” International Conference on Data Intelligence and Cognitive Informatics. Springer, Singapore, pp. 529-545, 2024.
[38] C. X. Han, L, S, Xin, C. F. Kai, L. Chen, and M. Yue. “A review of YOLO object detection algorithms based on deep learning,” Frontiers in Computing and Intelligent Systems, vol. 4, no. 2, pp. 17-20, 2023.

ZHENQIANG DAI was born in Liaoning Province, P. R. China, received the B.S. degree in Communication Engineering from University of Science and Technology Liaoning, Anshan, P. R. China, in 2023.

He is currently pursuing the M.S. degree in Electronic Information with University of Science and Technology Liaoning, Anshan, P. R. China. His research interest is machine vision

SHAOCHUAN XU was born in Liaoning Province, P. R. China, received the B.S. degree in automation from University of Science and Technology Liaoning, Anshan, P. R. China, received the M.S. degree in control science and engineering from University of Science and Technology Liaoning, Anshan, P. R. China, in1995, and 2004.

He is currently a professor in the School of Control Science and Engineering, University of Science and Technology Liaoning, Anshan, P. R. China. He published more than 20 academic papers, more than 20 patents and software copyrights. His research interests include research on industrial intelligent control and machine vision.

XIANGYI YAN was born in Liaoning Province, P. R. China, received the B.S. degree in automation from University of Science and Technology Liaoning, Anshan, P. R. China,received the M.S. degree in control science and engineering from University of Science and Technology Liaoning, Anshan, P. R. China,in2015，and 2019.

He is an Senior Engineer in Tianjin Research Institute of Construction Machinery Co.,Ltd.His research interests include research on artificial intelligence and machine vision.

SIHONG XU was born in Liaoning Province, P. R. China, received the B.S. degree in automation from University of Science and Technology Liaoning, Anshan, P. R. China,received the M.S. degree in control science and engineering from University of Science and Technology Liaoning, Anshan, P. R. China,in2015，and 2019.

She is an intermediate engineer in Tianjin Research Institute of Construction Machinery Co.,Ltd. Her main research direction is intelligent control and machine vision.