DOI: 10.19884/j.1672-5220.202502011

Steel Surface Defect Detection Based on Improved RT-DETR

HUANG Qiubo*, ZHANG Yakui, WANG Mei
School of Information and Intelligent Science, Donghua University, Shanghai 201620, China

Abstract: To address the challenges of low accuracy in small target detection and high model complexity in steel surface defect detection, a novel defect detection algorithm named FCS-DETR is proposed based on the real-time detection transformer (RT-DETR) framework. The methodology involves three key technical improvements. First, a newly designed DC-RepC3 module replaces the original RepC3 module to expand the model's receptive field for enhanced contextual information capture, which effectively suppresses complex background noise while improving detection accuracy. Second, for selective feature extraction, partial convolution is integrated into the backbone network to significantly reduce the parameter count and computational load, while wavelet pooling is further incorporated to strengthen model robustness. Finally, a Wise-ShapeIoU-NWD joint loss function is proposed, which incorporates both the shape and distribution characteristics of the target, reducing the geometric discrepancies between the predicted and ground truth bounding boxes and thus improving bounding box regression accuracy. Experimental results on the NEU-DET dataset demonstrate that FCS-DETR achieves a mean average precision at an intersection-over-union threshold of 0.50 (mAP_{50}) of 76.6%, representing an improvement of 3.4 percentage points over RT-DETR-R18, while reducing the parameter count and computational complexity of the model by 31.8% and 28.8%, respectively. These advancements establish FCS-DETR as an effective solution for accurate and efficient steel surface defect detection, showing substantial potential for industrial applications.

Keywords: steel surface defect detection; RT-DETR; partial convolution; dilated convolution; wavelet pooling; Wise-IoU
CLC number: TP183 **Document code**: A
Article ID: 1672-5220(2026)05-0080-10

0 Introduction

To enhance the quality of steel, surface defect detection has long been a key area of research in material processing and manufacturing. Traditional methods, such as manual inspection and X-ray detection[1], are capable of identifying defects, but they suffer from limitations including low efficiency, insufficient accuracy, and vulnerability to human error, making them unsuitable for large-scale modern production.

With advancements in computer vision and deep learning algorithms, image-based steel surface defect detection methods have become a prominent research focus. Machine learning, particularly the rapid development of convolutional neural networks (CNNs) and object detection algorithms, has provided new avenues for automating steel surface defect detection. Currently, defect detection methods are primarily categorized into two-stage and single-stage detection algorithms. Two-stage algorithms generate candidate boxes, followed by feature extraction and object regression. While these methods offer high detection accuracy, their large model size and slow detection speed hinder their deployment on edge devices. In contrast, single-stage algorithms (such as the YOLO series[2-4]) bypass the candidate box generation step, directly performing object classification and localization, thus significantly improving the detection speed. The YOLO series, in particular, has garnered significant attention due to its balanced performance in both speed and accuracy. Zhang et al.[5] enhanced YOLOv5s by embedding coordinate attention (CA) and atrous spatial pyramid pooling (ASPP), which broadened the network's receptive field but exhibited limited suppression of complex background noise and might inadvertently amplify irrelevant regions, resulting in only marginal gains in small-defect localization. Li et al.[6] incorporated GhostConv and the SimAM attention mechanism into YOLOv7, which effectively reduced parameter redundancy and sped up inference. However, this architecture still struggles with fusing features across scales, compromising detection consistency between small and large targets. Zhen et al.[7] developed a lightweight YOLOv5 variant using the efficient layer aggregation network (ELAN) structure and depthwise separable convolutions, achieving a smaller model footprint, but lacking dedicated mechanisms to highlight small-scale defect features and underutilizing inter-layer feature interactions, thereby limiting its performance in cluttered backgrounds.

Although these approaches have enhanced detection capabilities, several critical limitations remain

Received date: 2025-02-23
* Correspondence should be addressed to HUANG Qiubo, email: huangturbo@dhu.edu.cn
Citation: Huang Q B, Zhang Y K, Wang M. Steel surface defect detection based on improved RT-DETR[J]. *Journal of Donghua University (English Edition)*, 2026, 43(5): 80-89.

unaddressed in practical steel surface defect detection. First, the detection accuracy for small objects remains insufficient. Surface cracks, rolled-in scale, and other minute defects occupy minimal image area, and their representations are often weakened or lost through successive convolutions and downsampling. Second, steel defects range from large patches to fine pitting and slender scratches or cracks, necessitating robust multi-scale perception and integration strategies. Finally, industrial applications demand efficient inference and low resource consumption, yet high-precision models typically incur substantial computational overhead, hindering deployment on embedded or edge platforms.

To address these issues, this paper proposes three key improvements.

1) Enhanced small-object detection. A novel DC-RepC3 module is introduced to bolster the network's sensitivity and localization precision for small defects; a combined Wise-ShapeIoU-NWD joint loss function integrates geometric shape and spatial distribution cues, improving detection accuracy and robustness.

2) Wavelet-based multi-scale fusion. We replace conventional upsampling and downsampling operations with a wavelet pooling approach, which better preserves high-frequency image details.

3) Lightweight backbone optimization. Partial convolution is employed to streamline the backbone architecture, enhancing computational efficiency, and all accuracy-boosting modules are designed with resource constraints in mind.

1 Related Work

1.1 Real-time detection transformer (RT-DETR)

RT-DETR[8] is an end-to-end real-time object detection framework based on the transformer architecture, designed to address the computational complexity and real-time performance challenges of traditional methods. With its carefully engineered network structure, RT-DETR significantly enhances the computational efficiency while maintaining the detection accuracy, meeting the demands of real-time applications. Its key innovation is the hybrid encoder design, which combines the attention-based intra-scale feature interaction (AIFI) and cross-scale feature fusion (CCFF) modules to effectively process multi-scale features and perform efficient fusion.

1.2 Wavelet pooling

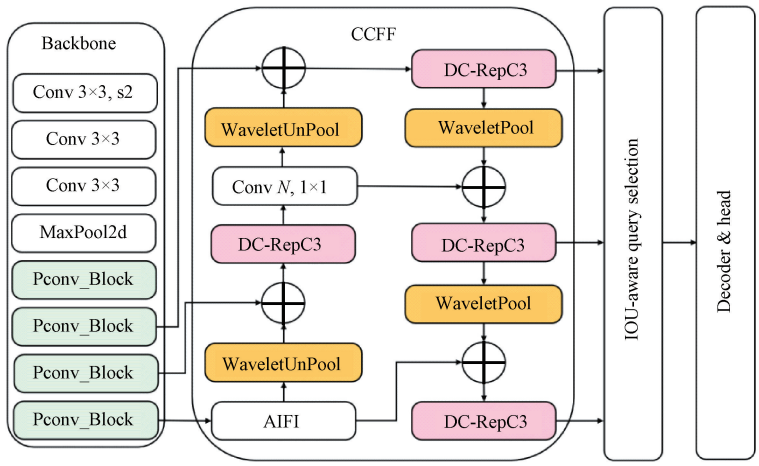
Conventional upsampling and downsampling methods, such as pooling and stride convolutions, enhance efficiency but often overlook the Nyquist sampling theorem, which can lead to aliasing. This aliasing negatively affects the detection of high-frequency features and small targets. To address this issue, the WaveletPool downsampling method based on the wavelet transform[9] has been introduced. By applying the discrete wavelet transform (DWT) and inverse discrete wavelet transform (IDWT), WaveletPool decomposes and reconstructs the feature map, enabling multi-frequency band decomposition and reconstruction during downsampling. This approach effectively preserves low-frequency information while filtering out high-frequency noise. Compared to traditional pooling and stride convolution methods, WaveletPool demonstrates superior performance in retaining small target features, providing more stable and detailed information during multi-level feature fusion, which significantly enhances the accuracy and robustness of small target detection.

2 FCS-DETR

To effectively address the three challenges described in the previous section, namely detecting small objects, identifying multi-scale defects on steel surfaces, and balancing detection accuracy with inference efficiency, we propose FCS-DETR, where FCS stands for focus on small. Built upon the lightweight RT-DETR backbone, our enhanced CCFF significantly enhances robustness and feature representation while remaining well-suited for deployment on edge devices. The following subsections describe our proposed method in detail.

2.1 Overall architecture

The FCS-DETR model introduces three key improvements. The architecture of the FCS-DETR model is shown in Fig. 1. First, in the CCFF module, the DC-RepC3 module replaces the original RepC3 module, expanding the model's receptive field and reducing the impact of background noise. Second, in the backbone network, partial convolution (Pconv) (denoted as Pconv_Block in Fig. 1) is employed to replace traditional convolution, and wavelet pooling methods (WaveletPool and WaveletUnPool in Fig. 1) are used in place of conventional upsampling and downsampling methods. Last, the Wise-ShapeIoU-NWD joint loss function is utilized to refine the bounding box modeling process, minimizing the impact of shape and size discrepancies on the model accuracy.



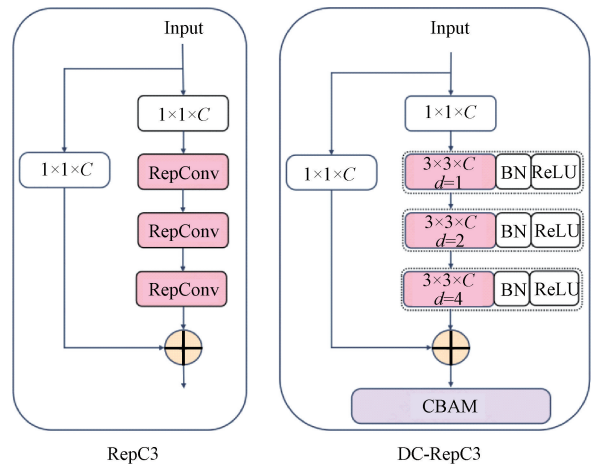
Conv 3×3, s2—3×3 convolution with stride of 2; MaxPool2d—two-dimensional maximum pooling layer; Conv N, 1×1—convolution layer equipped with N output convolution filters.

Fig. 1 Architecture of FCS-DETR model

2.2 Proposed DC-RepC3

The RepC3 module, as shown in Fig. 2, is an efficient feature extraction unit based on the cross-stage partial (CSP) architecture[10]. It optimizes information flow and reduces computational complexity by splitting and merging the feature map. Specifically, the input feature map is divided into two paths, with all convolutional layers set to output tensors of size $1 \times 1 \times C$, where C represents the number of feature channels; one path extracts deep features using three reparameterization convolution (RepConv) layers, while the other retains the original input information. These two feature sets are then merged along the channel dimension. However, the RepC3 module exhibits limitations in small target detection tasks. The fixed receptive field of standard convolution operations hinders feature extraction for targets of varying scales, particularly for small targets that occupy a minor portion of the image and exhibit less distinctive features, which negatively impacts detection accuracy.

To address these challenges, we propose an optimized version of the RepC3 module, namely DC-RepC3, as shown in Fig. 2. First, the original RepConv layers are replaced with convolutions of varying dilation rates d, which expand the receptive field and capture more stable contextual features, particularly when handling complex backgrounds or small target defects. Additionally, we incorporate the convolutional block attention module (CBAM)[11], a lightweight attention mechanism that adaptively adjusts the weights of both the channel and spatial dimensions. This helps suppress background noise and irrelevant features while improving the model's focus on critical features. The channel attention mechanism selects more discriminative feature channels, while the spatial attention mechanism enhances feature representation by focusing on important regions in the spatial domain.



BN—batch normalization; ReLU—rectified linear unit.

Fig. 2 Comparison of RepC3 and DC-RepC3

2.3 Lightweight improvement based on partial convolution

In deep neural network architecture design, conventional convolution operations, while powerful in feature extraction, suffer from high memory access costs and computational redundancy due to their full-channel computation nature. These limitations significantly hinder model deployment in edge computing scenarios.

To overcome these challenges, this study introduces PConv[12], an innovative operator that achieves efficient optimization of computation and storage through a channel-selective computation mechanism. As shown in Fig. 3, PConv adopts a two-stage feature processing strategy: the first stage applies standard convolution to C_1 channels ($C_1 < C$) of the input feature map $X \in \mathbb{R}^{C \times H \times W}$ (where H and W denote the height and width of the feature map, respectively), while the second stage applies an identity mapping to the remaining $C_2 = C - C_1$ channels.

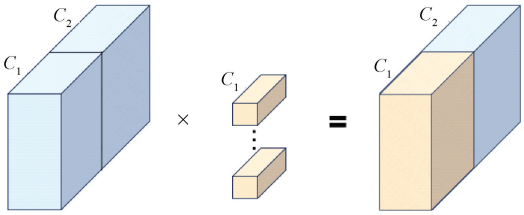


Fig. 3 Schematic diagram of PConv

This architecture can be formally expressed as

$$Y = [\mathrm{Conv}(X_{[1:C_1]}), X_{[C_1+1:C]}], \quad (1)$$

where Y denotes the output feature map of the PConv module; $\mathrm{Conv}(\cdot)$ denotes the standard convolution operation; $[\cdot]$ indicates concatenation along the channel dimension. This design enables PConv to retain the feature extraction capability of traditional convolutions while reducing the computational complexity from $O(k^2C^2)$ to $O(k^2C_1^2 + C_2)$, where k represents the convolution kernel size.

We conduct ablation experiments to verify the effectiveness of the proposed PConv on the NEU-DET dataset, and the quantitative comparison results are presented in Table 1.

Table 1 Experimental results of replacing standard convolution with PConv on NEU-DET dataset

Model	mAP_{50}/%	mAP_{50-95}/%	Params	GFLOPs
Base	73.7	41.7	3.90×10^7	57.0
Base+PConv	74.4	43.2	2.87×10^7	42.9

Notes: mAP_{50} represents mean average precision (mAP) at the intersection-over-union (IoU) threshold of 0.50; mAP_{50-95} represents mAP at IoU thresholds from 0.50 to 0.95; Params represent parameter count; GFLOPs represent total giga floating-point operations.

As shown in Table 1, replacing the standard convolution in the backbone network with PConv ($C_1 = 0.25C$) results in significant improvements on the NEU-DET dataset, achieving an mAP_{50} increase of 0.7 percentage points and an mAP_{50-95} increase of 1.5 percentage points. In addition, the model's parameter count decreases by 26.4% (from 3.90×10^7 to 2.87×10^7), and computational complexity is reduced by 24.7% (from 57.0 GFLOPs to 42.9 GFLOPs). This enhancement is achieved through structural reparameterization, which incurs no loss in inference speed, laying a key technical foundation for the future design of high-precision and low-latency edge computing models.

2.4 Proposed bounding box regression Wise-ShapeIoU-NWD joint loss function

In the original model, the standard IoU loss function emphasizes the IoU between the predicted and ground truth boxes. However, in the case of complex steel surface defects, where the shape and size of the defects vary significantly, traditional IoU loss often fails to distinguish these subtle differences effectively.

To overcome this challenge, this study proposes an improved loss function that is better suited to steel surface defect detection. Specifically, the modified loss function combines WiseIoU v3[13], ShapeIoU[14], and normalized Wasserstein distance (NWD)[15], with the goal of improving the model's ability to adapt to variations in the defect size and shape. First, WiseIoU, as an enhanced IoU metric, is well-suited for object detection tasks involving irregular shapes. Unlike traditional IoU, WiseIoU applies weighting to different regions, allowing the model to focus more on key regions when calculating the IoU. For instance, small localized defects on the steel surface, though having a relatively small area, can significantly affect the overall quality of the steel. Unfortunately, these small defects are often overlooked. WiseIoU's weighted adjustment helps the model focus more on these critical but small defect regions, thereby improving the detection accuracy for irregularly shaped defects. Additionally, ShapeIoU replaces the IoU calculation part in WiseIoU with a combination of shape distance and shape differences, providing a more accurate measure of the geometry of the defect region.

The core formulas for the loss functions are

$$L_{\text{Wise-ShapeIoU}} = r \times R_{\text{WIoU}} \times L_{\text{Shape-IoU}}, \quad (2)$$

$$R_{\text{WIoU}} = \exp\left(\frac{(x - x_{\text{gt}})^2 + (y - y_{\text{gt}})^2}{W_{\text{g}}^2 + H_{\text{g}}^2}\right), \quad (3)$$

$$L_{\text{Shape-IoU}} = 1 - I_{\text{IoU}} + d^{\text{shape}} + 0.5 \times \Omega^{\text{shape}}, \quad (4)$$

where $L_{\text{Wise-ShapeIoU}}$ is the integrated joint loss we proposed; $L_{\text{Shape-IoU}}$ is a shape-aware IoU loss; r is a hyperparameter; R_{WIoU} is the spatially weighted factor that measures the distance between the predicted and ground truth box centers (x, y) and $(x_{\text{gt}}, y_{\text{gt}})$; W_{g} and H_{g} denote the width and height of the ground-truth bounding box, respectively; I_{IoU} denotes the IoU between the predicted and ground truth boxes; d^{shape} measures the shape difference between the predicted and ground truth boxes based on geometric distance; Ω^{shape} acts as a regularization term to limit excessive shape variations, preventing overfitting.

Nevertheless, relying solely on the Wise-ShapeIoU loss still has some limitations. While it improves the matching of defects with different shapes, it still faces challenges when measuring the differences between predicted boxes and ground truth boxes for small defects. NWD, being less sensitive to objects of varying scales, is more suited for measuring the similarity between small objects[15]. Therefore, this study further incorporates the NWD as a metric to quantify the difference between two distributions.

By combining the WiseIoU loss $L_{\text{Wise-ShapeIoU}}$ and NWD loss L_{NWD}, the final joint loss function is obtained through a weighted sum, balancing the contributions of both losses

during model training. Specifically, the modified loss function L_{total} is expressed as

$$L_{\text{total}} = \lambda_1 L_{\text{Wise-ShapeIoU}} + \lambda_2 L_{\text{NWD}}. \quad (5)$$

The coefficients λ_1 and λ_2 control the relative contributions of the WiseIoU loss and the NWD loss. In practice, these coefficients can be tuned based on task requirements to maintain a balance between the two losses during optimization. In the experiments conducted in this study, we set $\lambda_1 = \lambda_2 = 0.5$. This weighted combination enables the model to not only improve localization accuracy but also provide more precise geometric measurements when handling complex and irregular defects, thereby enhancing detection performance.

3 Experiments

3.1 Dataset

The dataset used in this study is the NEU-DET surface defect dataset released by Northeastern University, China[16]. It contains six typical steel surface defect types: crazing (Cr), inclusion (In), patches (Pa), pitted surface (Ps), rolled-in scale (Rs), and scratches (Sc). Each defect type includes 300 grayscale images with a resolution of 200 × 200 pixels, totaling 1 800 images. To ensure adequate model training and reliable testing, the dataset is randomly split into training, validation, and test sets at a ratio of 8∶1∶1, resulting in 1 440 images for training, 180 for validation, and 180 for testing.

3.2 Experimental setup

All experiments were conducted on Ubuntu 20.04.6 LTS with an Intel® Xeon® Platinum 8336C CPU @ 2.30 GHz and an NVIDIA GeForce RTX 4090 GPU (24 GB), implementing FCS-DETR in Python 3.8.19 with PyTorch 2.0.1+cu118, CUDA 11.8, and cuDNN 8.7.0. Input images were resized to 640 × 640 pixels and normalized using ImageNet statistics. The training was performed for 300 epochs with a batch size of 16, using the AdamW optimizer (an initial learning rate of 1×10^{-4} and a weight decay of 0.05) and a cosine annealing learning-rate schedule. Other hyperparameters were set according to the original RT-DETR algorithm.

3.3 Evaluation metrics

The performance of the proposed algorithm was evaluated using five key metrics: precision, recall, mAP, computational complexity (evaluated by GFLOPs), and parameter count. Precision P measures the proportion of true positives among all samples predicted as positive, calculated as

$$P = T_{\text{P}} / (T_{\text{P}} + F_{\text{P}}), \quad (6)$$

where T_{P} is the number of true positives; F_{P} is the number of false positives. Recall R evaluates the model's ability to detect positive samples, calculated as

$$R = T_{\text{P}} / (T_{\text{P}} + F_{\text{N}}), \quad (7)$$

where F_{N} is the number of false negatives. GFLOPs quantifies the model's computational complexity by counting the floating-point operations required during execution, expressed in billions of floating-point operations.

The mAP is a critical measure of model detection performance, with two variants, namely mAP_{50} and mAP_{50-95}, offering a comprehensive view of the model's detection capability across different precision levels.

$$P_{\text{mAP}} = \frac{1}{N} \sum_{n=0}^{N} P_{\text{AP}n}, \quad (8)$$

where P_{mAP} represents the mAP; N is the number of classes; $P_{\text{AP}n}$ represents the AP for the nth class.

Parameter count P_{a} is used to assess the model's storage requirements and complexity, calculated as

$$P_{\text{a}} = \sum_{i=1}^{k} (w_i \times h_i \times c_i), \quad (9)$$

where w_i, h_i, and c_i are the width, height, and number of channels of the ith layer's weight matrix. These metrics provide a comprehensive evaluation of the model's performance in terms of detection accuracy, computational efficiency, and model complexity.

3.4 Heatmap analysis

To assess the impact of the modified loss function on the model's feature learning capacity, this study employs the XGradCAM[17] method for visualizing the defect detection process (Fig. 4). By comparing the activation map distributions of the original model and the improved Wise-ShapeIoU-NWD model, we observe that the latter demonstrates higher accuracy in detecting surface defects with complex geometric features.

The heatmap of the improved model demonstrates far more comprehensive feature coverage. The red and yellow high-activation regions now stretch uniformly from each defect's core out to its morphological edges, effectively eliminating the false positives and missed regions caused by irregular shapes or background noise in the original model. Moreover, this enhancement generalizes across defects of diverse forms and scales. It accurately captures a broader spectrum of geometric features and defect types. In particular, by introducing a Shape-IoU constraint for scratch detection, the model greatly refines the spatial alignment between its activation maps and the true defect boundaries, transforming previously fragmented responses into smooth and continuous contours.

Experimental results reveal that the original model demonstrates a typical localization bias, where high-activation regions are overly concentrated on the defect body, thereby neglecting the edge details. This limitation is effectively mitigated with the introduction of Wise-ShapeIoU-NWD. Specifically, Shape-IoU improves the model's boundary perception by incorporating geometric constraints, enabling it to capture the topological features

of irregular contours. Meanwhile, NWD enhances the model's sensitivity to small deformations and local texture anomalies by reweighting distances in the feature space.

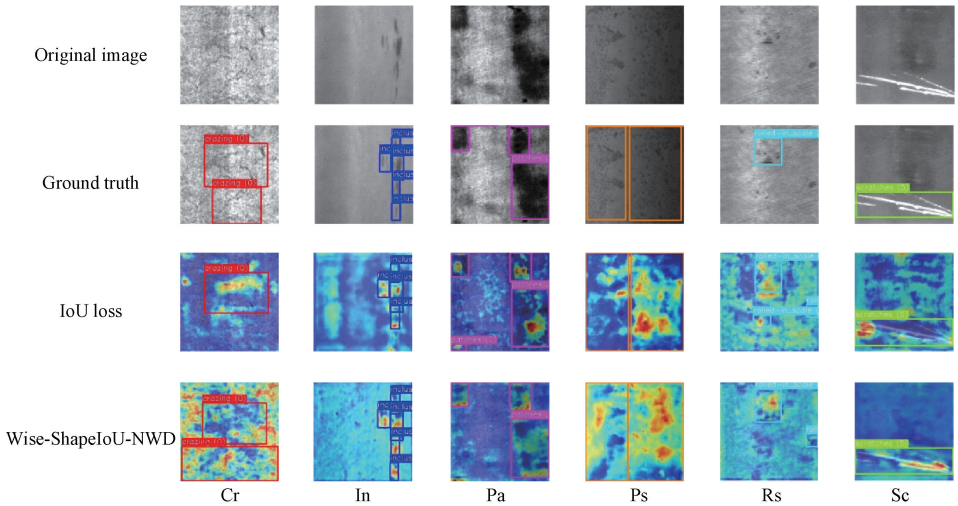


Fig. 4 Comparison of model heatmaps with/without new loss

3.5 Ablation experiments

Before the main ablation studies, we conducted a preliminary experimental design to identify the optimal module parameters and configurations. Each candidate setup was assessed on two criteria: detection performance (mAP_{50} and mAP_{50-95}) and computational complexity (parameter count and GFLOPs), ensuring a robust baseline for all subsequent large-scale evaluations.

3.5.1 DC-RepC3 module hyperparameter tuning

We examined the dilation rate d in combination with different attention mechanisms. Three dilation rates ($\{1, 2, 4\}$, $\{1, 2, 3\}$, and $\{1, 3, 5\}$) were evaluated. The results are shown in Table 2.

Table 2 Results of different dilation rates

Method	mAP_{50}/%	mAP_{50-95}/%	Params	GFLOPs
Base	73.7	41.7	3.90×10^7	57.0
Base+DC-RepC3 ($d=\{1, 2, 4\}$)	74.8	42.9	4.01×10^7	57.0
Base+DC-RepC3 ($d=\{1, 2, 3\}$)	73.5	42.1	4.01×10^7	57.0
Base+DC-RepC3 ($d=\{1, 3, 5\}$)	70.4	38.4	4.01×10^7	57.0

Results indicate that for $d=\{1, 2, 4\}$, the model achieves 74.8% mAP_{50} and 42.9% mAP_{50-95}, increases of 1.1 and 1.2 percentage points over the base configuration, demonstrating an optimal trade-off between receptive field coverage and feature redundancy, and markedly improving small-defect detection. For $d=\{1, 3, 5\}$, mAP_{50} and mAP_{50-95} decline to 70.4% and 38.4%, respectively, revealing that excessively large or unevenly spaced dilation rates can lead to feature sparsification and degrade detection accuracy. Based on these findings, we fixed $d=\{1, 2, 4\}$ and then compared three attention methods (CBAM[11], SE[18], and CPCA[19]). The results are shown in Table 3.

Table 3 Results of different attention methods

Method	AP/%	AP_{50}/%	AP_{75}/%	AP_s/%	AP_m/%	AP_L/%	Params	GFLOPs
Base	41.1	72.7	40.4	28.7	34.0	52.8	3.90×10^7	57.0
Base+DC-RepC3	42.3	73.7	42.3	35.4	35.8	50.7	4.01×10^7	57.0
Base+SE-RepC3	40.3	71.3	39.9	24.0	34.5	50.0	4.01×10^7	57.0
Base+CPCA-RepC3	40.6	72.6	40.3	30.4	35.0	47.7	4.04×10^7	58.2

Notes: AP_{50} and AP_{75} represent the AP calculated at IoU thresholds of 0.50 and 0.75, respectively; AP_s, AP_m, and AP_L represent the AP for small-sized, medium-sized, and large-sized objects, respectively.

As shown in Table 3, in terms of AP_s, DC equipped with the CBAM attention method achieves improvements of 11.4 and 5.0 percentage points over SE and CPCA, respectively. Thus, DC was selected for the final model.

3.5.2 Hyperparameter tuning of Wise-ShapeIoU-NWD loss

Eleven configurations were evaluated with λ_1 ranging from 0 to 1.0 in increments of 0.1 under the constraint $\lambda_1 + \lambda_2 = 1$. As shown in Fig. 5, the model performance in terms of mAP_{50} exhibits a bell-shaped curve, peaking at $\lambda_1 = 0.5$, where the model achieves the highest mAP_{50} of 74.3%. This suggests that a balanced contribution of the Wise-ShapeIoU and NWD losses leads to optimal detection accuracy.

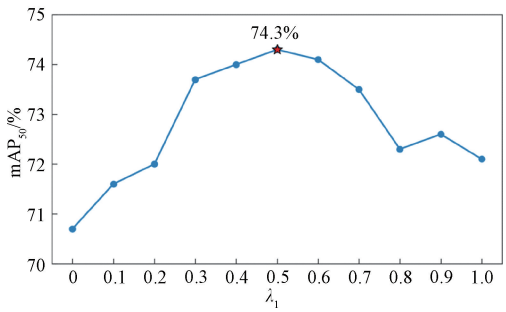


Fig. 5 Comparison of mAP_{50} under varying λ_1 values on NEU-DET dataset

Performance deteriorates at both ends of the curve, with mAP_{50} values dropping to 70.7% and 72.1% at $\lambda_1 = 0$ and $\lambda_1 = 1.0$, respectively. This trend confirms that relying solely on either spatial-aware (NWD) or shape-aware (Wise-ShapeIoU) loss fails to capture the full complexity of object localization and shape representation. Overall, these results demonstrate the importance of jointly leveraging both geometric and distributional cues in loss design, with balanced weighting proving most effective in enhancing model generalization.

3.5.3 Ablation experiments of all modules

To evaluate the impact of each improvement module on the performance of the steel surface defect detection model, this study conducted an ablation study by progressively adding each module to assess the effects of DC-RepC3, PConv, WaveletPool, and the optimized loss function. The experimental results are shown in Table 4, where "√" indicates the addition of a corresponding improvement module, and "−" indicates the absence of a module.

Table 4 Results of ablation experiments

Module				Performance					
DC-RepC3	PConv	WaveletPool	Loss function	*P*/%	*R*/%	mAP_{50}/%	mAP_{50-95}/%	Params	GFLOPs
−	−	−	−	78.0	68.5	73.7	41.7	3.90×10^7	57.0
√	−	−	−	79.7	68.3	74.8	42.9	4.01×10^7	57.0
−	√	−	−	75.1	68.6	74.4	43.2	2.87×10^7	42.9
−	−	√	−	72.3	69.7	74.7	41.5	3.90×10^7	60.9
−	−	−	√	75.9	68.5	74.3	42.4	3.90×10^7	57.0
√	√	−	−	72.4	71.0	74.2	43.0	2.83×10^7	42.9
√	√	√	−	78.1	67.5	75.0	42.6	2.66×10^7	40.6
√	√	√	√	76.4	73.4	76.6	43.1	2.66×10^7	40.6

The ablation experiments presented in Table 4 demonstrate the critical contributions of each proposed module. Specifically, introducing the DC-RepC3 module improves the baseline mAP_{50} from 73.7% to 74.8% and mAP_{50-95} from 41.7% to 42.9%, with only a minor increase in parameter count and negligible computational overhead. This highlights that the DC-RepC3 module enhances the model's representation capabilities, particularly improving its ability to accurately localize defect regions without incurring high computational costs.

Incorporating the PConv module notably reduces the computational complexity and parameter count, lowering the parameter count by approximately 26.4% (from 3.90×10^7 to 2.87×10^7) and GFLOPs by nearly 24.7% (from 57.0 to 42.9). Although the precision metric slightly decreases, the overall mAP performance is largely unaffected. These findings suggest that the PConv module effectively eliminates redundant computations through partial convolutions, optimizing model efficiency while maintaining detection accuracy.

Furthermore, adding the WaveletPool module provides a modest yet meaningful improvement in accuracy, elevating mAP_{50} to 75.0% and mAP_{50-95} to 42.6%, while simultaneously reducing model complexity (parameter count and GFLOPs drop to 2.66×10^7 and 40.6, respectively). This indicates that the WaveletPool module better preserves high-frequency detail crucial for detecting small defects, thus enhancing multi-scale representation without increasing computational burden.

Finally, integrating the Wise-ShapeIoU-NWD joint loss function significantly improves the model's accuracy, achieving the highest performance of 76.6% in mAP_{50}. This substantial improvement illustrates that combining geometric shape and spatial distribution

metrics effectively enhances defect localization precision. The synergy between all proposed modules, including DC-RepC3, PConv, WaveletPool, and the joint loss function, indicates that each component plays a complementary and indispensable role in overall performance optimization.

3.6 Comparative experiments

Table 5 presents the comparative experimental results across several advanced object detection models, including YOLO series (YOLOv5, YOLOv5m, YOLOv6, YOLOv8m, YOLOv9m, and YOLOv10m) and RT-DETR-R18. Among these models, our proposed FCS-DETR achieves the highest recall (73.4%) and mAP_{50} (76.6%), outperforming the best competitors such as YOLOv9m by 1.1 percentage points in mAP_{50} and RT-DETR-R18 by 3.4 percentage points in mAP_{50}.

Table 5 Comparative experimental results

Model	*P*/%	*R*/%	mAP_{50}/%	mAP_{50-95}/%	Params	GFLOPs
YOLOv5	71.1	71.0	73.1	41.2	5.10×10^6	7.1
YOLOv5m	73.2	70.0	75.1	42.7	5.05×10^7	64.0
YOLOv6	75.0	67.7	74.3	42.6	8.40×10^6	11.8
YOLOv8m	70.6	71.7	74.8	42.9	5.21×10^7	78.7
YOLOv9m	76.8	69.8	75.5	44.0	4.09×10^7	76.5
YOLOv10m	71.3	71.9	73.9	42.7	3.35×10^7	63.4
RT-DETR-R18	75.3	66.9	73.2	41.4	3.90×10^7	57.0
FCS-DETR (ours)	76.4	73.4	76.6	43.1	2.66×10^7	40.6

Crucially, our method also demonstrates excellent computational efficiency, using a significantly smaller parameter count (2.66×10^7 vs. 4.09×10^7 in YOLOv9m and 3.90×10^7 in RT-DETR-R18) and lower computational complexity. These results underline FCS-DETR's effectiveness in balancing accuracy and efficiency, making it especially suitable for industrial applications requiring precise defect detection and deployment on resource-constrained devices.

3.7 Generalization experiments

To evaluate the model's generalization performance, we conducted comprehensive experiments on three additional defect detection datasets: GC10, PCB, and RDD2022-China. The GC10 dataset comprises 2 292 grayscale images at a resolution of 2 048 pixel × 1 000 pixel. The dataset was partitioned into training, validation, and test sets in a ratio of 8∶1∶1. For the PCB defect detection task, a dataset of 693 images was utilized, with 552 allocated for training, 66 for validation, and 75 for testing. The RDD2022-China dataset combines two subsets (China_M and China_D). Following preprocessing to exclude five unannotated defect images from China_D, the final dataset contained 4 378 images. These were divided into training, validation, and test sets following a ratio of 8∶1∶1. These experiments utilized the identical configuration settings established for the NEU-DET dataset, ensuring consistent comparisons. The results are represented in Table 6.

Table 6 Results on additional defect detection datasets

Dataset	Model	mAP_{50}/%	mAP_{50-95}/%	Params	GFLOPs
GC10	RT-DETR-R18	74.3	41.5	3.90×10^7	57.0
	FCS-DETR (ours)	74.5	42.1	2.66×10^7	40.6
PCB	RT-DETR-R18	92.8	51.6	3.90×10^7	57.0
	FCS-DETR (ours)	93.6	51.9	2.66×10^7	40.6
RDD2022-China	RT-DETR-R18	87.5	58.0	3.90×10^7	57.0
	FCS-DETR (ours)	88.6	58.3	2.66×10^7	40.6

Specifically, for the GC10 dataset, mAP_{50} increases from 74.3% to 74.5%, and mAP_{50-95} increases from 41.5% to 42.1%. For the PCB dataset, mAP_{50} increases from 92.8% to 93.6%, and mAP_{50-95} increases from 51.6% to 51.9%. Additionally, for the RDD2022-China dataset, mAP_{50} increases from 87.5% to 88.6%, and mAP_{50-95} shows a slight enhancement from 58.0% to 58.3%. In terms of complexity, our proposed FCS-DETR model consistently employs a smaller parameter count (2.66×10^7 vs. 3.90×10^7) and reduces the computational demand (40.6 GFLOPs vs. 57.0 GFLOPs) compared to the RT-DETR-R18 baseline. This represents a reduction of approximately 31.8% in parameter count and about 28.8% in computational demand.

In summary, FCS-DETR achieves notable accuracy improvements across diverse datasets while significantly decreasing computational complexity and model size, underscoring its suitability for practical applications, particularly in resource-constrained environments.

4 Conclusions

We propose an efficient temporal decoding module based on action category prediction queries for video action recognition. By pre-constructing a fixed number of prediction queries and leveraging the powerful global temporal modeling capability of transformer attention mechanisms, this module autonomously learns to classify different actions. The decoder is conceptually and implementation-wise straightforward. It can be readily extended to other action recognition models with relatively low additional cost while enabling efficient temporal modeling, thus improving action classification accuracy. Moreover, incorporating this temporal decoder enhances the model's discriminative ability for subtle actions, which is particularly helpful in scenarios involving similar action sequences. Our experiments demonstrate that the constructed decoder improves the recognition of actions in longer-duration videos, effectively addressing issues related to long-range dependencies. We believe that our work has the potential to contribute to similar tasks in the field of video understanding in the future.

References

[1] Wu G H, Xiong H J. Radiography testing in China [J]. *Chinese Journal of Scientific Instrument*, 2016, 37(8): 1683-1695. (in Chinese)

[2] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 779-788.

[3] Farhadi A, Redmon J. YOLOv3: an incremental improvement[C]//Computer Vision and Pattern Recognition. Heidelberg: Springer, 2018, 1804: 1-6.

[4] Li C Y, Li L L, Jiang H L, et al. YOLOv6: a single-stage object detection framework for industrial applications [PP/OL]. arXiv (2022-09-07) [2025-01-25]. https://arxiv.org/abs/2209.02976. arXiv: 2209.02976.

[5] Zhang S Q, Shi W Y, Zhang S W, et al. Steel surface defect detection based on improved YOLOv5 algorithm [J]. *Science Technology and Engineering*, 2023, 23(35): 15148-15157. (in Chinese)

[6] Li W, Yao X M, Zhang P C, et al. Study on improved YOLO-V7 algorithm of steel surface defect detection [J]. *Mechanical Science and Technology for Aerospace Engineering*. 2025, 44(12): 2191-2200. (in Chinese)

[7] Zhen G Y, Zhao L R, Li W Y, et al. Research on lightweight of steel surface defect detection network based on improved YOLOv5 [J]. *Modular Machine Tool & Automatic Manufacturing Technique*, 2024(3): 58-63. (in Chinese)

[8] Zhao Y A, Lü W Y, Xu S L, et al. DETRs beat YOLOs on real-time object detection [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 16965-16974.

[9] Williams T, Li R. Wavelet pooling for convolutional neural networks[C]//International Conference on Learning Representations. [S.l.]: [s.n.], 2018.

[10] Wang C Y, Liao H Y, Wu Y H, et al. CSPNet: a new backbone that can enhance learning capability of CNN [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2020: 1571-1580.

[11] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module [M]// Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.

[12] Chen J R, Kao S H, He H, et al. Run, don't walk: chasing higher FLOPS for faster neural networks [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 12021-12031.

[13] Tong Z J, Chen Y H, Xu Z W, et al. Wise-IoU: bounding box regression loss with dynamic focusing mechanism[PP/OL]. arXiv (2023-04-08) [2025-01-25]. https://arxiv.org/abs/2301.10051. arXiv: 2301.10051.

[14] Zhang H, Zhang S J. Shape-IoU: more accurate metric considering bounding box shape and scale [PP/OL]. arXiv (2024-01-12) [2024-12-15]. https://arxiv.org/abs/2312.17663. arXiv: 2312.17663.

[15] Wang J W, Xu C, Yang W, et al. A normalized Gaussian Wasserstein distance for tiny object detection[PP/OL]. arXiv (2022-07-14) [2024-12-15]. https://arxiv.org/abs/2110.13389. arXiv: 2110.13389.

[16] He Y, Song K C, Meng Q G, et al. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features [J]. *IEEE Transactions on Instrumentation and Measurement*, 2020, 69(4): 1493-1504.

[17] Fu R G, Hu Q Y, Dong X H, et al. Axiom-based grad-CAM: towards accurate visualization and explanation of CNNs[PP/OL]. arXiv (2020-08-19) [2024-12-15]. https://arxiv.org/abs/2008.02312. arXiv: 2008.02312.

[18] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[19] Huang H, Chen Z, Zou Y, et al. Channel prior convolutional attention for medical image segmentation [J]. *Computers in Biology and Medicine*, 2024, 178: 108784.

基于改进 RT-DETR 的钢材表面缺陷检测

黄秋波*，张亚魁，王　梅
东华大学 信息与智能科学学院，上海 201620

摘　要：为了应对钢材表面缺陷检测中小目标检测精度低和模型复杂度高的问题，提出了一种基于实时检测 transformer（RT-DETR）框架的新型缺陷检测算法 FCS-DETR。该方法包括 3 项关键技术改进：首先，设计了一个新的 DC-RepC3 模块，替代原有的 RepC3 模块，扩展模型的感受野以捕获更丰富的上下文信息，既能有效抑制复杂背景噪声，又能提升缺陷检测精度；其次，通过在主干网络中集成部分卷积实现了选择性特征提取，显著减少了参数量和计算负载，此外结合小波池化以增强模型的稳健性；最后，提出了 Wise-ShapeIoU-NWD 联合损失函数，该函数结合了目标的形状和分布特征，减小了预测框和真实框之间的几何差异，从而提高了边界框回归精度。基于 NEU-DET 数据集的实验结果表明，FCS-DETR 达到了 76.6%的 mAP_{50}（交并比阈值为 0.50 时的平均精度），相较于基线 RT-DETR-R18 提高了 3.4 百分点；同时模型参数量和计算复杂度分别降低了 31.8%和 28.8%。这些进展使 FCS-DETR 成为兼顾高精度与高效率的钢材表面缺陷检测方案，展示了其在工业应用中的潜力。

关键词：钢材表面缺陷检测；RT-DETR；部分卷积；扩张卷积；小波池化；Wise-IoU