Measurement Science
and Technology

PAPER

Design of the global-perception focus module and reparameterized dilation-wise for detecting steel surface defects

To cite this article: Jing Liao *et al* 2025 *Meas. Sci. Technol.* **36** 116006

IOP Publishing Measurement Science and Technology
Meas. Sci. Technol. **36** (2025) 116006 (19pp) https://doi.org/10.1088/1361-6501/ae1aa8

Design of the global-perception focus module and reparameterized dilation-wise for detecting steel surface defects

Jing Liao[1,2], **Mingliang Liu**[1], **Lei Jiang**[1,2] and **Kuanching Li**[1,2,*]

[1] School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, People's Republic of China
[2] Sanya Research Institute, Hunan University of Science and Technology, Sanya, People's Republic of China

E-mail: aliric@hnust.edu.cn

Received 26 August 2025, revised 13 October 2025
Accepted for publication 3 November 2025
Published 13 November 2025

Abstract

Surface defect detection is critical in steel industrial quality control. However, because of the diverse nature of defects, the low contrast between defects and the complex background noise, the surface defect detection often results in inaccurate localization, missed detections, and false positives. To overcome these challenges, this work proposes a steel surface defect detection network framework called surface enhanced defect attention network (SEDA-Net), which is based on a global perception focusing module and dilated convolutions. In this framework, the global perception focusing module is designed to eliminate noise ambiguity between defects and the background effectively. Subsequently, the network's receptive field is expanded through reparameterization of the dilated residual module to enhance its ability to model features of low-contrast defects. Finally, a dynamic attention-based intra-scale feature interaction module is introduced to strengthen the interaction between global and local information, thereby enhancing the feature representation of interclass defects. Extensive experiments conducted on two benchmark datasets, NEU-DET and GC10-DET, demonstrate the effectiveness and superiority of SEDA-Net. Experimental results show the proposed method outperformed the state-of-the-art approaches, improving 0.8% to 10.2% in average precision.

Keywords: steel surface defect detection, global-perception focus module, reparameterized dilation-wise, feature extraction network

1. Introduction

The development of steel production technologies in the 21st century has emerged as a critical indicator of industrial progress. A wide variety of steel products, including sheets, coils, strips, and bars, play essential roles in various sectors such as aerospace, automotive engineering, and infrastructure development [1]. Consequently, the identification of surface defects is imperative for improving product quality and fostering intelligent production processes. Nevertheless, the attainment of comprehensive automation in the domain of surface anomaly inspection persists as a significant technical challenge.

Taking hot-rolled steel as an example, the production environment of hot rolling lines is characterized by elevated temperatures, substantial dust, the presence of oil stains, water

* Author to whom any correspondence should be addressed.

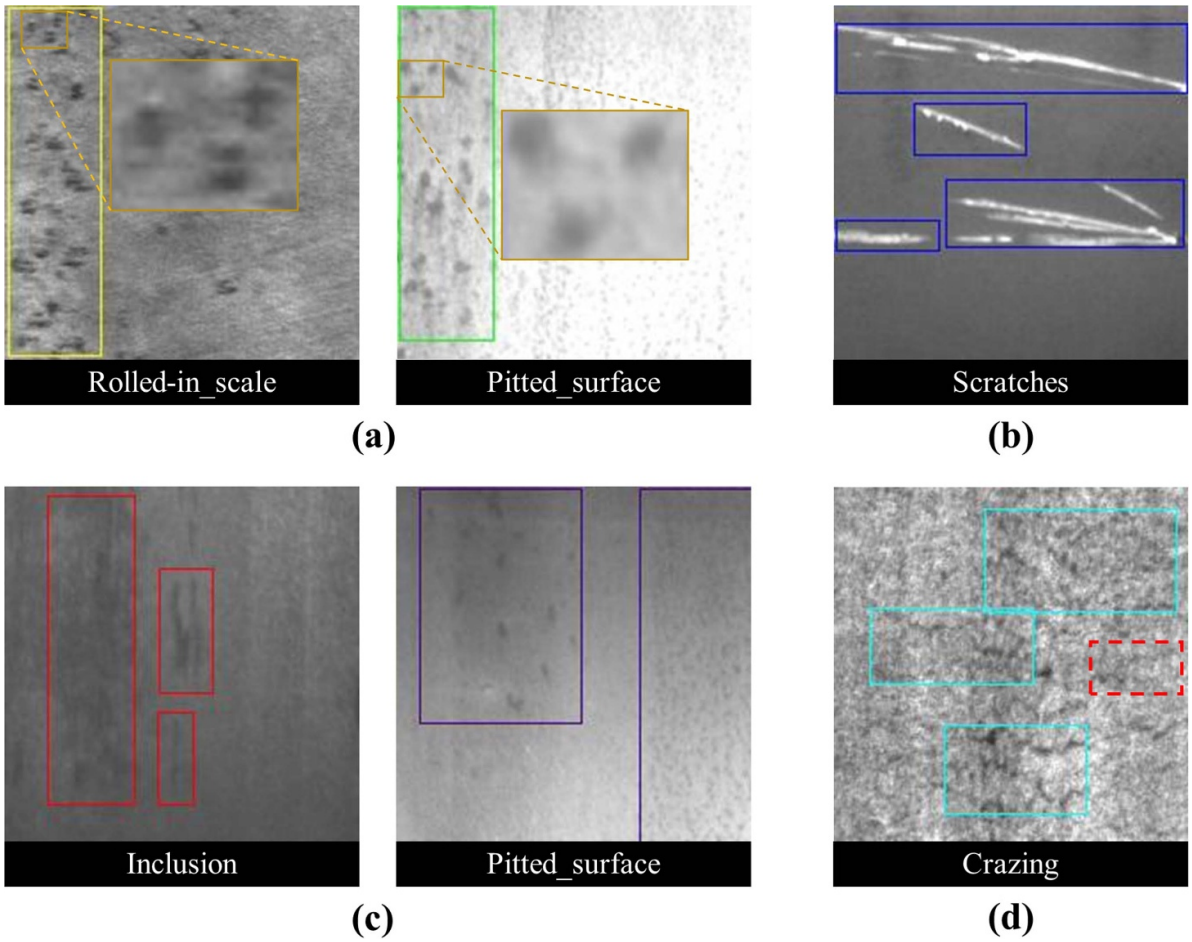


Figure 1. Challenges in defect detection of hot-rolled strip steel. (a) Defects with significant intra-class differences, (b) defects with inter-class similarities, (c) low-contrast defects, and (d) complex background noise.

mist, and elastic vibrations, which collectively introduce various interferences. Consequently, the surface quality of steel plates varies considerably across different production lines and even between different batches. Together, these adverse conditions complicate the reliable identification of surface defects in steel manufacturing.

(1) Inter-class similarity and intra-class diversity: Defects within the same class exhibit significant variations in appearance, as illustrated in figure 1(a). In contrast, defects belonging to different classes share similar characteristics, as demonstrated in figure 1(b), which indicates that the characteristics of defects within a class are diverse, while the characteristics of defects across classes are too similar.
(2) Low contrast: Some defects are similar in color to the material itself, and the low contrast characteristics of industrial images make the contours of defects more blurred and concealed, as shown in figure 1(c).
(3) Noise interference: Due to the complexity of the production environment for hot-rolled strip steel, different texture noises appear on the surface of the steel. Some background noises have similar characteristics to foreground defects, as shown in figure 1(d), where the light blue box indicates a defect and the red box indicates background noise.

Over the past decade, a variety of image processing techniques have been applied extensively to the detection of surface defects, with varying degrees of success. Nevertheless, these methods inherently possess limitations, notably their insufficient capability to represent complex defect patterns. For example, Luo *et al* [2] constructed a real-time hot-rolled flat steel Automated Optical Inspection system. This system leveraged a Dynamic Homogenization Compensation algorithm for image enhancement and introduced an Adaptive Dual-Threshold method for feature extraction and defect detection. Although this approach, through image enhancement, partially alleviates issues of uneven illumination, overexposure, or underexposure, its dependency on simple dual-threshold decision-making for defect classification struggles to handle multi-class and fine-grained defects effectively.

Deep learning, a branch of machine learning, has attracted widespread interest in recent years and has achieved notable advancements in object detection tasks [3]. Several outstanding object detectors have been developed, such as the single-stage YOLO series [4] and the two-stage Faster

R-CNN series [5]. The powerful automated feature extraction capabilities of deep learning methods have made them a popular area of research in defect detection. Due to the industrial demand for real-time defect detection that balances accuracy and speed, single-stage detectors represented by the YOLO algorithm are more favored in industrial scenarios [6]. YOLO11 [7], developed by the authors of YOLOv5 [8] and YOLOv8 [9], introduces conditional channels and positional spatial attention modules, significantly enhancing the model's spatial attention processing capabilities, marking a significant improvement over YOLOv8 [7].

However, due to the inherent complexity and challenges of steel surface defect detection tasks, existing general object detectors often perform suboptimally when directly transferred to this domain. To overcome these limitations, Li *et al* [8] proposed a steel surface defect detection model based on YOLOv5, explicitly addressing the challenges of complex and diverse industrial defect features. Their work introduced an efficient multi-scale feature extraction module, a novel feature fusion scheme, and optimizations to the bottleneck module, thereby improving the model's accuracy in steel surface defect detection. Similarly, to mitigate issues stemming from unstructured features, multi-scale variations, and data scarcity in industrial defects, Liu *et al* [9] introduced a global attention module and a cascaded fusion network based on YOLOv8 for steel surface defect detection. This approach enhanced the model's capability to handle unstructured defects through the global attention module, followed by designing a cascaded fusion network to bolster multi-scale feature interaction. Furthermore, Soft-NMS was employed during the post-processing phase to retain more potentially valid candidate boxes and improve detection performance. While existing methods have made significant progress in detection accuracy, they still struggle to effectively address challenges such as low contrast, noise interference, and diverse defect types in complex industrial scenarios. Consequently, issues like inaccurate defect localization, missed detections, and false positives persist in practical applications.

To tackle these issues, this study redesigns the feature extraction backbone of YOLO11 and proposes surface enhanced defect attention network (SEDA-Net), a steel surface defect detection network that integrates a hybrid attention mechanism with reparameterized dilated convolutions. SEDA-Net consists of three key modules to improve performance in the detection of steel surface defects. First, we propose a globally aware focus module with fine-grained feature enhancement to eliminate ambiguous noise between defects and the background. Second, a re-parameterized dilated residual block with multi-scale receptive field enhancement is designed to strengthen the representation capability of low-contrast defects. Furthermore, a dynamic attention-based intra-scale feature interaction (DyAIFI) module is developed to reinforce global and local information interaction, adapting to intra-class variations and inter-class similarities among defects. Comprehensive experimental analyses were conducted using two distinct steel plate surface defect datasets. These experimental results strongly validate the effectiveness of the proposed method. Summarily, the main contributions of this article are as follows:

(1) To enhance the perception of fine-grained information by low-level features, we have designed a global-perception focus module (GPFM) that suppresses global diffusion background noise through the combined modeling capabilities of deep separable convolutions and cascaded group attention (CGA) modules.
(2) To enhance the receptive field through multi-scale dilation convolution, the reparameterized dilation-wise residual (RepDWR) module has been developed to capture contextual information better and improve the defect areas' distinguishability.
(3) To enhance the dynamic interaction between global and local information, a DyAIFI module has been introduced to address the issues of intra-class differences and inter-class similarities in defects.

The remainder of this article is organized as follows. Section 2 reviews related work, while section 3 provides a detailed description of the proposed SEDA-Net architecture and its key components. Section 4 presents the experimental setup and analyzes the results, and finally, concluding remarks and discussions of potential directions for future research are depicted in section 5.

2. Related work

In industrial inspection, steel surface defect detection has progressed from traditional physical sensor-based methods and signal analysis to vision-based techniques. With the rapid advancement of computational power in recent years, deep learning networks have emerged as a leading solution for detecting defects [6].

2.1. Traditional methods

Early approaches primarily relied on manual inspection, eddy current testing, infrared thermography, magnetic flux leakage, and ultrasonic testing [10]. However, these methods typically suffer from low efficiency, are highly sensitive to environmental and surface conditions, and struggle with real-time performance and precise defect localization. As a result, they fall short of meeting the modern steel industry's demands for automation, accuracy, and high throughput [11]. To address these constraints, methods grounded in computer vision have gained widespread adoption, providing enhanced robustness and flexibility.

Song *et al* [12] proposed a defect detection method for hot-rolled steel strips using the adjacent evaluation completed local binary pattern (LBP), which reduces sensitivity to noise. While LBP is traditionally applied to gray-scale images, Fekri-Erhad *et al* [13] extended its application to color images by introducing a multi-resolution, noise-robust variant that leverages chromatic information. Aminzadeh *et al* [14] employed histogram comparisons

between defect and background regions to determine optimal detection thresholds. Additionally, spectral analysis methods have been explored, transforming images into the frequency or other domains to distinguish defects based on differential responses between defective and non-defective regions.

Although these traditional techniques laid the groundwork for early defect detection systems, they rely heavily on handcrafted features and generally lack the accuracy required for precise localization. Their performance is particularly limited under low-contrast or noisy conditions, where detection reliability significantly degrades.

2.2. Deep learning-based methods

The advent of massive labeled datasets and the rapid development of computing hardware have rendered deep learning a formidable instrument in industrial applications, delivering significant advances in object detection [11]. By constructing multi-layer neural networks, deep learning enables automatic learning of hierarchical feature representations [15, 16]. Compared with traditional vision-based approaches, it eliminates the need for handcrafted features and significantly improves generalization and automation [17]. Consequently, deep learning technology has emerged as the dominant approach to addressing visual inspection problems, particularly in the context of industrial defect detection.

Deep learning-based methods can typically be divided into three categories according to their detection pipelines and network architectures: two-stage, single-stage, and Transformer-based detectors. Two-stage detectors, such as the R-CNN family [5, 18–20], initially produce candidate regions followed by a classification step. In contrast, single-stage detectors like the YOLO series [4, 21] and SSD [22] directly perform classification and localization in one step, offering faster inference. Transformer-based detectors such as DETR [23], Swin Transformer [24], and RT-DETR [25] have further advanced the field by incorporating global attention mechanisms. However, these general-purpose detectors often underperform when applied directly to industrial scenarios due to differences in defect characteristics and environmental conditions. Consequently, recent studies have focused on adapting and enhancing these models for industrial defect detection.

Xun *et al* [26] enhanced RetinaNet by integrating differential channel attention and adaptive feature fusion to address challenges such as overlapping, multiple, and multi-class defects. Cui *et al* [27] improved SSD with feature retention blocks and skip-dense connections, aiming to detect minor defects with significant texture variations. Zhao *et al* [28] proposed an improved Faster R-CNN employing multi-scale feature fusion and deformable convolutions to better handle minor and complex defects. Wang *et al* [29] developed a few-shot detection framework incorporating two domain generalization strategies and a noise regularization mechanism to mitigate issues related to low contrast and high inter-class similarity. Wang *et al* [30] further optimized YOLOv7 by incorporating a BiFPN structure, efficient channel attention, and adopting the SIoU loss to enhance small-target detection. Yeung and Lam [31] introduced an attention-enhanced model with adaptive feature fusion to address defect scale and shape variations. Despite these advances, challenges such as defect diversity, low contrast, and background noise persist.

Regarding the issue of defect diversity, Huang *et al* [32] proposed an aggregation-redistribution network architecture to aggregate and refine features at different hierarchical levels. The introduction of this framework significantly improved the network's recognition accuracy for defects with considerable variations in scale and shape. However, its performance in detecting defects involving geometric complexity remains limited, particularly when identifying defects with irregular shapes, ambiguous boundaries, or substantial background interference.

To address the low-contrast defect problem, Huang *et al* [33] developed a novel lightweight network comprising the location enhanced ghost network (LEG-Net) and the refine grouped spatial network (RGS-Net). The LEG-Net extracts features along both width and height directions, while the RGS-Net emphasizes channel information extraction and leverages its robust nonlinear fitting capability to achieve efficient feature fusion. The combination of these two networks effectively identifies low-contrast defects and small-target defects. Nevertheless, this approach exhibits limitations in detecting defects with high intra-class variation, such as scratch defects.

Concerning the issue of background noise interference in defect detection, Peng *et al* [34] introduced a novel defect detection network that employs a deformable ResNet50 network to extract multi-scale defect features and utilizes a double attention feature pyramid network to mitigate interference from complex backgrounds. This method effectively improved the detection accuracy for multi-scale defects and enhanced localization precision under background noise interference, demonstrating substantial improvement in overall accuracy. However, it similarly exhibits lower accuracy for scratch-type defects characterized by high intra-class variation.

Existing methods struggle to simultaneously address these three critical issues, with many approaches suffering from inaccurate localization, missed detections, and false alarms. Specifically, a critical research gap remains for a model that is both efficient and capable of robustly handling the coexistence of diverse defect types, low contrast, and noise interference. In this study, we propose a novel and computationally efficient detection framework. We introduce targeted, lightweight yet effective modules designed to enhance the model's capability to recognize low-contrast and diverse defect patterns, while suppressing background interference by revisiting the feature aggregation and refinement pipeline based on the intrinsic characteristics of steel surface defects.

3. Proposed SEDA-Net architecture

This section presents the architecture of the proposed SEDA-Net, a novel framework designed to address the key challenges inherent in steel surface defect detection. These challenges commonly include the diversity of defects, low contrast, and

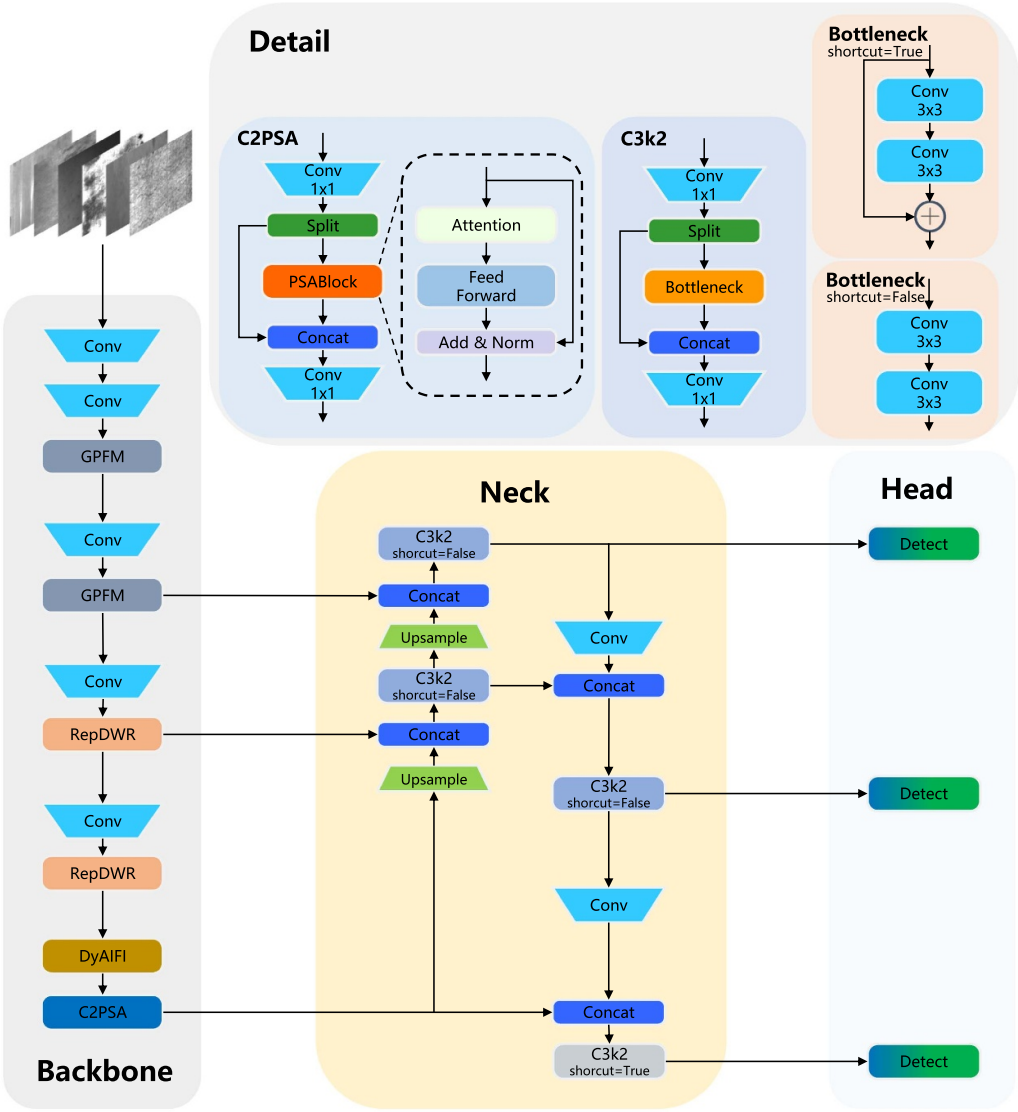


Figure 2. Overall network architecture of SEDA-Net, where Conv, Concat, Split, Upsample, and Detect represent convolution, concatenation, splitting, upsampling, and regression prediction of feature maps, respectively.

background noise, which often lead to missed detections, false positives, and inaccurate localization. The three key modules of SEDA-Net, namely GPFM, RepDWR, and DyAIFI, are subsequently introduced and detailed in turn.

3.1. Overall network architecture

The overall architecture of SEDA-Net is illustrated in figure 2. The network is an improved variant based on YOLO11, which frames object detection as a regression task, simultaneously accomplishing localization and classification within a single inference step. Initially, the input image undergoes normalization and is fed into the custom-designed backbone network for hierarchical feature extraction. Subsequently, these features are fused through the path aggregation feature pyramid network to enhance multi-scale feature propagation. The resulting feature maps are then propagated to the detection head to generate bounding box coordinates and class probabilities.

SEDA-Net establishes a unified framework by cohesively integrating the GPFM, RepDWR, and DyAIFI into a coherent architecture. The proposed network can effectively address the challenges of complex backgrounds and diverse defect patterns, significantly improving both detection accuracy and recall rate.

Specifically, The GPFM integrates depthwise convolution (DWC) with CGA,combining the strengths of convolutional neural networks (CNNs) and Transformers to suppress global background noise andbemphasize salient defect regions. RepDWR begins with a downsampling layer and employs three parallel convolutional branches with varying receptive fields to aggregate multi-scale information, improving sensitivity to blurred or ambiguous defects. DyAIFI, a Transformer-based enhancement module, incorporates 2D positional encoding and multi-head self-attention (MHSA) to model long-range dependencies. It effectively captures intra class variation and inter-class similarity among defects.Each

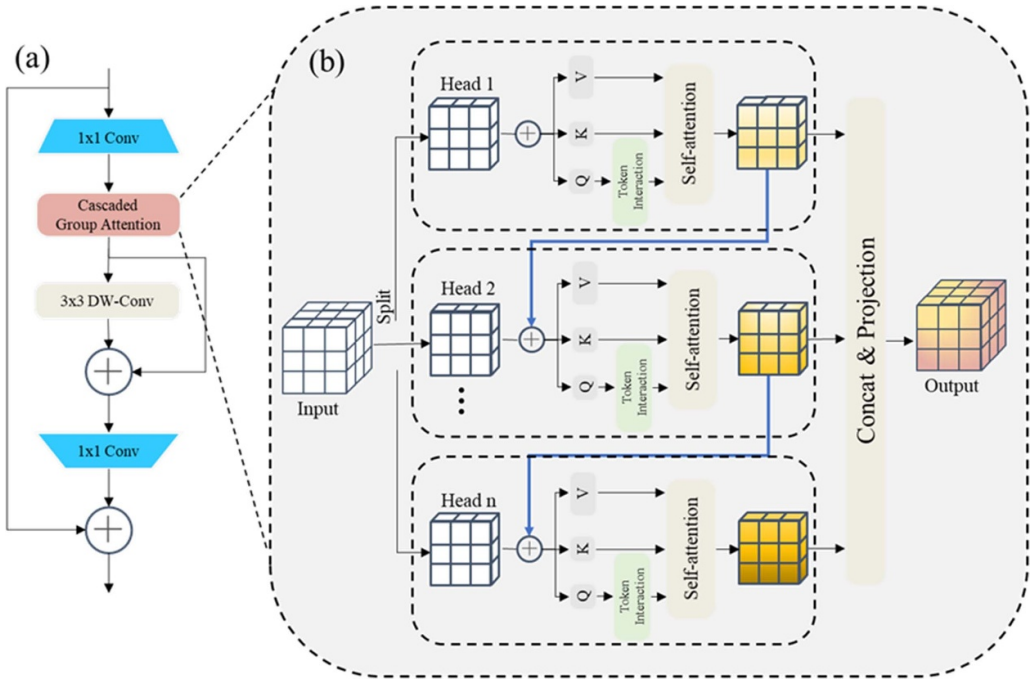


Figure 3. Detailed structure of GPFM.

module will be described in detail in the subsequent sections.

3.2. GPFM

Noise interference is one of the significant challenges in steel surface defect detection tasks, mainly due to the complex production environment of hot-rolled strip steel, where the surface may exhibit a variety of texture noises. Several background noises are highly similar to the characteristics of foreground defects, as depicted in figure 1(d). CNN-based detection algorithms typically utilize convolutional filters to capture localized features from input images. However, depending exclusively on such local cues limits the model's ability to achieve accurate detection.

The MHSA [35, 36] in the Transformer can capture the global relationships between all pixels or regions in an image, overcoming the limitations of the local receptive field in traditional CNNs. It excels in modeling global context and establishing long-range dependencies. However, in the MHSA module, each head uses the same complete input features to calculate the attention map, and multiple heads may learn similar attention patterns, leading to computational redundancy. The CGA [37] module was proposed to address this issue. The CGA module uses different input feature splits for each attention head, reducing computational redundancy, and introduces a cascading mechanism where each head's output serves as the next head's input, which progressively refines the features, enhancing the diversity of attention maps while enabling each head to learn distinct feature representations.

Moreover, Lin *et al* [38] pointed out in their study that the attention mechanism and the convolutional layer should be complementary rather than mutually exclusive. Based on this idea, we designed the GPFM to address the challenge of noise interference. It consists of CGA and DWC, which can be expressed as:

$$\mathcal{F}(\cdot) = (\text{DW-Conv, Skip})\left(\mathbb{CGA}(\cdot)\right). \tag{1}$$

As shown in figure 3, GPFM consists of CGA and DWC connected in series.

The CGA establishes long-range dependencies and effectively captures global information, enhancing the response of key areas while utilizing DWC to extract local features, thereby finely enhancing detailed texture information. Conventional bottleneck modules, which enhance feature representation through stacked convolutions, are hampered by limited local receptive fields, often leading to the confusion of defect regions with noisy textures. While attention-based modules like SE [39] and CBAM [40] refine feature selection by reweighting channels or integrating spatial and channel attention, their effectiveness is compromised in scenarios where background textures closely resemble defect patterns due to their still-limited receptive fields. By contrast, GPFM combines CGA and DWC, allowing it to simultaneously model long-range dependencies and optimize for fine-grained details. This design combines the modeling capabilities of Transformers and CNNs to suppress the diffusion of global background noise and focus on the salient features of foreground defects, thereby effectively addressing the challenge of noise interference in steel surface defect detection.

The Depthwise Separable Convolutions is an efficient convolution method proposed based on extracting spatial features and mixing channel features [41]. Instead of performing standard convolution operations directly, it splits the process into

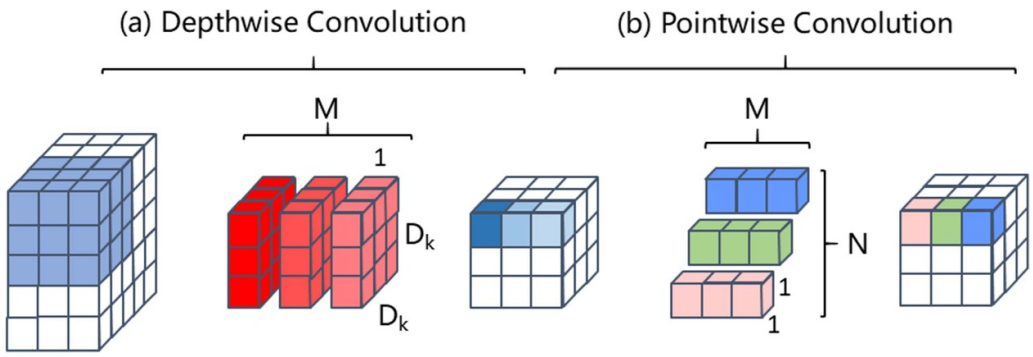


Figure 4. Depthwise Separable Convolution. Constructed from the depth convolution filter in figure (a) and the point convolution filter in figure (b).

two steps: a depthwise operation that filters each input channel independently, followed by a 1×1 convolution (pointwise convolution) that integrates the filtered outputs across channels. This decomposition greatly reduces both parameter count and computational cost. As shown in figure 4 the convolution is factorized into a DWC (figure 4(a)) and a subsequent 1×1 pointwise convolution (figure 4(b)), where the former captures spatial patterns per channel and the latter fuses information across channels.

The DWC for each input channel can be formulated as:

$$F_{h,w,c_{\text{in}}} = \sum_{i,j} W^{\text{dw}}_{i,j,c_{\text{in}}} \cdot X_{h+i-1,w+j-1,c_{\text{in}}} \tag{2}$$

where $X_{h,w,c_{\text{in}}}$ denotes the input feature map at spatial position (h, w) with c_{in} input channels, and $W^{\text{dw}}_{i,j,c_{\text{in}}}$ represents the DWC kernel of size $K \times K$ applied independently to each input channel c_{in}. The output $F_{h,w,c_{\text{in}}}$ is the intermediate feature map obtained by convolving X with the channel-specific kernel at each spatial location.

The computational complexity is:

$$\text{FLOPs}_{\text{depthwise}} = H \cdot W \cdot M \cdot {D_K}^2. \tag{3}$$

DWC is limited to processing individual input channels independently and lacks the capability to integrate information across channels to produce new feature representations. To address this limitation, an additional transformation layer is required to compute a linear combination of the depthwise outputs. This is typically achieved using 1×1 convolutions, also known as pointwise convolutions, which can be formulated as follows:

$$Y_{h,w,c_{\text{out}}} = \sum_{c_{\text{in}}} W^{\text{pw}}_{c_{\text{in}},c_{\text{out}}} \cdot F_{h,w,c_{\text{in}}} \tag{4}$$

where $W^{\text{pw}}_{c_{\text{in}},c_{\text{out}}}$ is the pointwise 1×1 convolution kernel that linearly combines the c_{in} channels of the intermediate feature map $F_{h,w,c_{\text{in}}}$ to produce the final output feature map $Y_{h,w,c_{\text{out}}}$ withc_{out} channels at each spatial position (h, w).

The complete formula is given as:

$$O_{\text{DWS}} = \text{PointwiseConv}\left(\text{DepthwiseConv}\left(I\right)\right). \tag{5}$$

The CGA module is a novel attention module proposed to address the issue of redundant attention heads in MHSA. Inspired by the group convolution mechanism in efficient CNNs, this method introduces a more structured attention modeling approach for visual Transformer architectures. As shown in figure 3(b), the CGA structure explicitly decomposes the attention calculation process of each head by dividing the complete feature representation into different parts and assigning them to individual attention heads. This mechanism effectively enhances the specificity and complementarity among attention heads. It can be represented as:

$$\begin{aligned} \widetilde{F}_{ij} &= \text{Attn}\left(F_{ij}W^{\text{Q}}_{ij}, F_{ij}W^{\text{K}}_{ij}, F_{ij}W^{\text{V}}_{ij}\right), \\ \widetilde{F}_{i+1} &= \text{Concat}\left[\widetilde{F}_{ij}\right]_{j=1:h} W^{\text{P}}_{i}. \end{aligned} \tag{6}$$

where the jth head applies self-attention to the sub-feature F_{ij}, which corresponds to the jth segment of the input feature F_i, that is, $F_i = [F_{i1}, F_{i2}, \ldots, F_{ih}]$, $1 \leqslant j \leqslant h$, with h indicating the total number of attention heads. The projection matrices W^{Q}_{ij}, W^{K}_{ij}, and W^{V}_{ij} are incorporated to transform each sub-feature into query, key, and value representations, respectively. After the attention outputs from all heads are concatenated by $Concat[\cdot]$, they are projected through a linear transformation W^{P}_i to restore the feature dimension to match the original input.

CGA computes attention maps for individual heads in a sequential cascade, as illustrated in figure 3(a). In this design, the output from each attention head is forwarded to the next, enabling progressive refinement of the feature representation.

$$F'_{ij} = F_{ij} + \widetilde{F}_{i(j-1)}, \quad 1 < j \leqslant h \tag{7}$$

where $F'ij$ denotes the sum of the j-th input segment Fij and the output $e_{F_{i(j-1)}}$ from the $(j!-!1)$th attention head, as computed by equation (6). During the self-attention process, $F'ij$ is utilized to update the updated input to the jth head, replacing the original Fij. In addition, a token interaction module is introduced after the W^{Q} projection to enable the attention mechanism to capture both local dependencies and global context, thereby further enhancing the expressiveness of the feature representation.

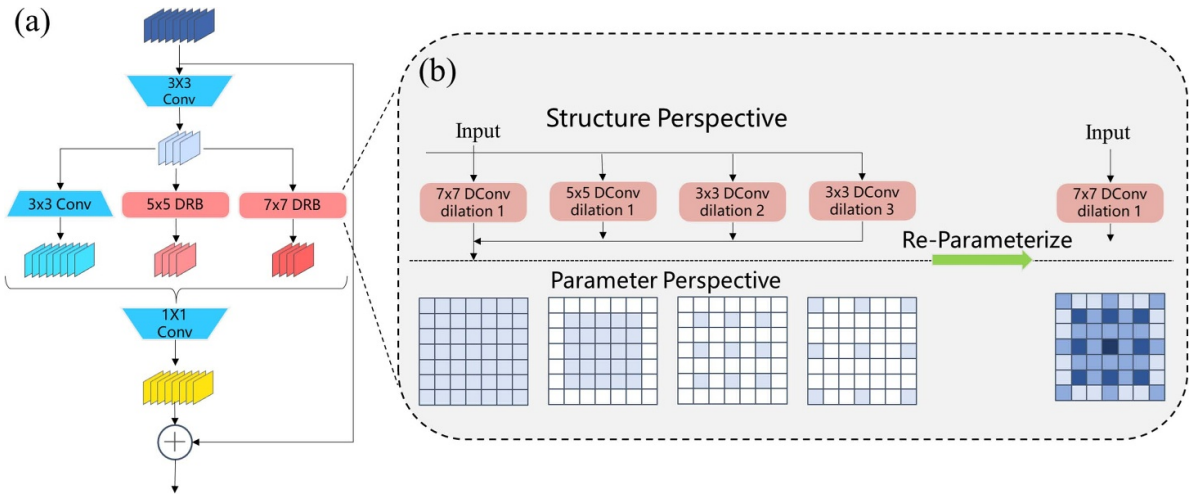


Figure 5. Illustration of the structure of the RepDWR module. (a) and (b) show the RepDWR module and DRB module, respectively. In the example shown in figure (b), $K = 7$. For smaller values of K, the number of dilation layers also decreases accordingly.

3.3. RepDWR module

Low-contrast defect detection is a significant challenge in steel surface defect inspection tasks (as shown in figure 1(c)). To address this issue, we draw inspiration from the design principles of semantic segmentation networks, utilizing dilated convolutions to expand the receptive field while enhancing pixel-level feature capture capabilities [42]. We designed a RepDWR module, whose structure is shown in figure 5. For low-contrast defects, due to the difficulty in capturing edge texture features, the module can effectively locate the defect area by modeling pixel-level details under a large receptive field, thereby adapting to the detection requirements of low-contrast features.

RepDWR uses a two-step method to obtain and integrate multi-scale contextual information effectively. In the first step, regional residualization generates concise feature maps with different regional morphologies, laying the foundation for subsequent morphological filtering. This process consists of 3×3 convolution, batch normalization (BN), and ReLU. In the second step, semantic residualization of regional features is performed using a dilated reparam block. The dilated reparam block employs a multi-branch structure to execute dilated convolutions at different rates, matching specific receptive fields to achieve targeted morphological filtering. The combination of regional residualization and semantic residualization significantly simplifies the task of deep dilated convolution, transforming it from the extraction of complex semantic information to morphological processing of concise features, optimizing the learning process, and enhancing the efficiency of multi-scale contextual information extraction. Finally, a more comprehensive and robust feature representation is generated through feature concatenation, BN processing, pointwise convolution, and residual weighting.

Incorporates a non-dilated small-kernel and several dilated small-kernel convolutions to augment the representational capacity of the large-kernel convolution without dilation.

The architecture of the Dilated Reparam Block (DRB) is illustrated in figure 5(b), which features an enhanced combination of a non-dilated large kernel convolutional layer with a non-dilated small kernel and multiple dilated small kernel convolutional layers. The design parameters encompass the size of the large kernel K, the size of the parallel small kernels K, and the dilution rate r. For example, figure 5(b) shows a structure with three parallel layers, where $K = 7$, $r = (1, 2, 3)$, and $k = (5, 3, 3)$. For smaller K values, the number of layers can be reduced by decreasing the kernel size or dilution rate. For instance, when $K = 5$, the structure uses two layers with $k = (3, 3)$ and $r = (1, 2)$, corresponding to an equivalent kernel size of 5×5. The training-stage structure is shown in equation (8). During the inference stage, the DRB can be converted into a single large kernel convolutional layer, as shown in equation (9). The specific conversion process involves merging the BN layer with the preceding convolutional layer, converting layers with a dilution rate $r > 1$ into sparse convolution kernel functions, and summing all kernel results while adding appropriate zero padding. For example, the layer with $k = 3$ and $r = 2$ in figure 5(b) is ultimately converted into a sparse 5×5 kernel and added to a 7×7 large kernel, with one pixel of zero padding added to each side.

$$Y = \sum_{r \in \mathcal{R}} Y^{(r)} = \sum_{r \in \mathcal{R}} K^{(r)} * X \tag{8}$$

$$\begin{aligned} K^{(\text{merged})} &= \sum_{r \in \mathcal{R}} \text{Align}\left(K^{(r)}, r\right), \\ Y &= K^{(\text{merged})} * X. \end{aligned} \tag{9}$$

Conventional approaches to enlarging the receptive field, such as naively increasing kernel sizes in CNN bottlenecks or employing dilated convolutions, can capture wider context but often suffer from excessive computational overhead and fragmented feature representations. For example, ASPP-based

architectures tend to process dilated branches independently, risking inconsistent cross-scale responses. Moreover, while Inception-style multi-branch designs improve scale diversity, they lack a robust mechanism for consolidating these features into a cohesive representation. These limitations are especially pronounced in steel defect detection, where localizing low-contrast defects like inclusions or blurry scratches demands a synthesis of both global contextual understanding and fine-grained morphological detail.

By contrast, our proposed RepDWR module addresses these challenges by incorporating a two-step residual mechanism with a dilated re-parameterization block. This architecture is designed to progressively refine regional and semantic features, thus enhancing the global context while preserving the integrity of faint boundaries. In the training phase, RepDWR leverages multi-rate, multi-branch dilated convolutions to effectively capture features across a spectrum of receptive fields. Crucially, during inference, these parallel branches are re-parameterized into a single, equivalent large-kernel convolution. This structural transformation retains potent multi-scale contextual expressiveness while drastically reducing computational costs. This unique synthesis of efficiency and performance enables RepDWR to excel at suppressing background interference and accentuating subtle, low-contrast defects, thereby surpassing conventional dilated convolution and large-kernel methods on complex steel surfaces.

3.4. DyAIFI

The intraclass differences and interclass similarities in steel surface defects pose significant challenges for accurate detection, as illustrated in figures 1(a) and (b). To overcome this, the DyAIFI module is introduced to replace the spatial pyramid pooling-fast SPPF module in the YOLO11 baseline.

The SPPF module is an optimized design for feature extraction in the YOLO network. It captures information from different receptive fields through multi-scale pooling operations, as shown in figure 6(a). Specifically, the SPPF module extracts multi-scale features from the feature map through three maximum pooling operations with the same kernel size, concatenates the pooling results with the original feature map along the channel dimension, and then fuses the information through a convolutional layer to generate a more expressive feature representation. The primary function of this module is to capture the different sizes and positions of objects through multi-scale feature fusion while reducing computational complexity through optimized design. However, since SPPF relies on pooling operations with fixed receptive fields, its ability to model global dependencies for complex objects is limited. This limitation may lead to performance bottlenecks when handling tasks with complex feature distributions.

The DyAIFI module is designed based on the Transformer Encoder architecture, aiming to enhance feature interaction and expression capabilities by combining attention mechanisms with positional information. Its structure is shown in figure 6(b). Its core lies in utilizing MHSA mechanisms to capture global dependencies among input features, while

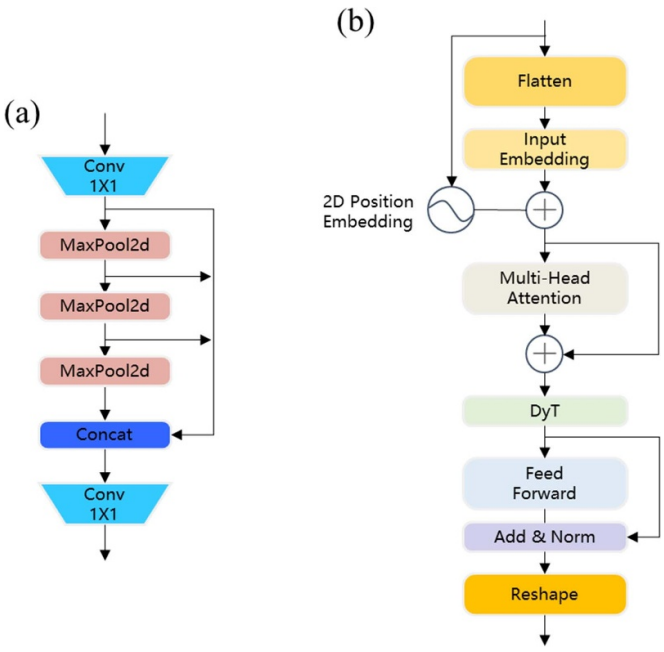


Figure 6. Detailed structure of SPPF module and AIFI module. (a) SPPF, (b) AIFI.

introducing spatial position information through 2D sine-cosine position encoding, thereby improving the ability to model feature spatial relationships. Additionally, the module employs a feedforward network composed of two fully connected layers and nonlinear activation functions to perform nonlinear transformations on features. Then, Dynamic Tanh (DyT) compression is applied to map feature values, supplemented by residual connections and normalization operations to maintain training stability and enhance feature expressiveness. The definition of the DyT layer is given in equation (10). By reshaping the features, this module can ultimately generate high-quality feature representations that retain global semantic information and spatial details.

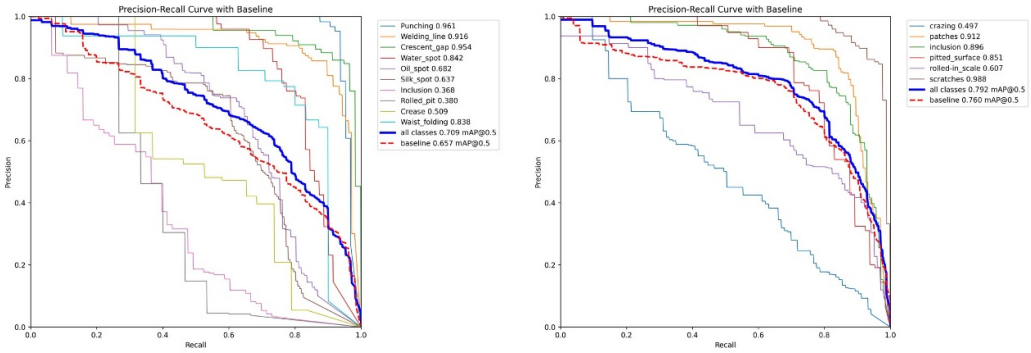
DyAIFI employs a Transformer-based dynamic self-attention mechanism that adaptively models global correlations among features while preserving local structural cues through positional encoding. By integrating the DyT compression layer, this mechanism further enhances adaptability: it prevents feature saturation while enabling fine-grained calibration of defect responses. This design ensures that DyAIFI can effectively distinguish between confusing categories with high inter-class similarity while maintaining robustness against intra-class variations. Compared to the SPPF module, DyAIFI can more flexibly adapt to complex task requirements, balancing global perception and detail retention.

$$\mathrm{DyT}(\boldsymbol{x}) = \boldsymbol{\gamma} * \tanh(\alpha \boldsymbol{x}) + \boldsymbol{\beta} \tag{10}$$

where ω is a learnable scalar parameter that can scale inputs differently based on their range, ω and θ are learnable vector parameters per channel, consistent with the scaling and offset parameters commonly used in normalization layers.

Table 1. The hyperparameters of SEDA-Net.

Hyperparameters	Value	Note
Learning rate	1×10^{-3}	Initial learning rate
Decay strategy	Linear	
Optimizer	AdamW	
Momentum	0.9	
Weight decay	5×10^{-4}	
Total epochs	300	
Batch size	16	
Close mosaic epochs	10	Disable mosaic augmentation for final 10 epochs



(a) (b)

Figure 7. Training loss and mAP curves of SEDA-Net on NEU-DET (a) and GC10-DET (b) datasets.

4. Experimental and results

This section comprehensively evaluates the proposed method on two benchmark steel surface defect datasets (NEU-DET dataset [43] and GC10-DET dataset [44]), and conducted comparative experiments with 14 representative detection networks.

4.1. Experimental environment

The methodology described in this study was implemented using the PyCharm Integrated Development Environment. Model training was conducted on a workstation equipped with one Intel Xeon Silver 4210 R CPU, 32GB of memory, 4TB of storage, and dual NVIDIA A4000 GPUs (16GB of memory each), running Microsoft Windows11. The software environment included Python 3.8.19, PyTorch 2.1.2, CUDA 12.1, and cuDNN 8.8.2. The hyperparameter settings used for model training are summarized in table 1.

4.2. Datasets

NEU-DET Dataset: Provided by Northeastern University, the NEU-DET dataset is a widely used benchmark for evaluating surface defect detection algorithms in steel plates. It comprises six common defect categories encountered in industrial settings: scratches (Sc), spots (Pa), inclusions (In), rolled-in scale (RS), pitted surfaces (PS), and cracks (Cr). This dataset supports consistent benchmarking of detection methods and plays a vital role in advancing the accuracy and efficiency of automated steel quality inspection. Each sample is annotated with corresponding ground truth to facilitate algorithm training and evaluation.

GC10-DET Dataset: The GC10-DET dataset is a comprehensive collection of real-world steel surface defect samples acquired from actual industrial production lines. It covers ten distinct defect types, capturing the complexity and variability of surface anomalies observed in manufacturing environments. The defect categories include punching (Pu), welding line (WL), crescent gap (CG), water spot (WS), oil spots (OS), silk spots (SS), inclusions (In), rolling pits (RP), creases (Cr), and waist folding (WF), each posing unique challenges for detection. All samples are accompanied by corresponding ground truth annotations, enabling rigorous training and evaluation of defect detection algorithms.

4.3. Evaluation metrics

In object detection tasks, Average Precision (AP) and mean Average Precision (mAP) are commonly used as the main evaluation metrics. AP assesses detection accuracy for a particular defect category, while mAP is the average of AP scores across all categories.

Mathematically, the AP for the ith defect category is defined as:

$$\mathrm{AP}(i) = \int_0^1 P(R)\,\mathrm{d}R \tag{11}$$

Table 2. Quantitative comparison of different detection methods on the NEU-DET dataset. The 'bolded' data indicate that a model has the optimal/best-performance.

Model	Recall	mAP@50	mAP@50:95	Params(M)	FLOPs(G)
Faster R-CNN [5]	54.1	74.5	39.6	41.8	177.6
RetinaNet [26]	58.1	74.2	39.9	33.8	340.1
YOLOv5 [7]	69.2	75.7	42.7	9.1	23.8
YOLOv8 [7]	69.6	75.0	43.1	11.2	28.4
YOLOv9 [50]	72.4	75.5	44.7	7.2	26.7
YOLOv10 [51]	69.1	75.1	43.6	8.0	24.5
YOLO11 [7]	72.5	76.1	42.9	9.4	21.3
YOLOv12 [52]	68.2	76.8	42.9	9.2	21.5
DETR [23]	60.5	71.2	40.1	42.2	59.0
RT-DETR(R18) [25]	71.9	73.8	42.1	20.1	61.5
RT-DETRv2 [25]	71.5	77.7	45.4	20	61.1
PMSA-DyTr[a] [49]	71.5	78.7	44.3	59.9	257.4
CAT-YOLO[a] [25]	69.8	77.5	41.0	14.1	31.6
GC-Net[a] [9]	71.2	78.1	42.5	32.6	50.4
LHATA-Net[a] [47]	71.4	78.5	42.8	**3.5**	**18.4**
SEDA-Net	**72.7**	**79.2**	**48.5**	10.5	20.9

[a] indicates that the model is specifically designed for defect detection.

Table 3. Quantitative comparison of different detection methods on the GC10-DET dataset.

Model	Recall	mAP@50	mAP@50:95	Params(M)	FLOPs(G)
Faster R-CNN [5]	44.6	68.3	34.0	41.8	177.6
RetinaNet [26]	48.5	67.7	33.4	33.8	340.1
YOLOv5 [7]	56.0	60.7	32.3	9.1	23.8
YOLOv8 [7]	62.5	66.3	34.7	11.2	28.4
YOLOv9 [50]	65.5	68.6	35.2	7.2	26.7
YOLOv10 [51]	65.2	64.1	33.2	8.0	24.5
YOLO11 [7]	62.6	65.7	33.8	9.4	21.3
YOLOv12 [52]	63.9	68.1	34.8	9.2	21.5
DETR [23]	61.2	62.3	35.1	42.2	59.0
RT-DETR(R18) [25]	63.9	66.3	35.5	20.1	61.5
RT-DETRv2 [25]	58.9	67.4	37.8	20	61.1
PMSA-DyTr[a] [49]	65.7	69.8	36.6	59.9	257.4
CAT-YOLO[a] [25]	63.0	65.9	33.6	14.1	31.6
GC-Net[a] [9]	64.2	68.3	36.0	32.6	50.4
LHATA-Net[a] [47]	63.9	68.4	35.6	**3.5**	**18.4**
SEDA-Net	**66.2**	**70.9**	**38.2**	10.5	10.9

[a] indicates that the model is specifically designed for defect detection.

where i represents the defect type, AP(i) is the detection accuracy of the ith defect type, and n is the total number of categories. P denotes precision and R denotes recall, with the corresponding formulas given as:

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}}, \\ \text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}}. \end{aligned} \tag{12}$$

In this work, mean Average Precision (mAP) is adopted as the primary metric to assess the accuracy of the proposed model. Everingham *et al* demonstrated that mAP serves as a reliable evaluation metric, especially in scenarios involving multiple object categories [45]. Common mAP variants include: (1) map@50, which computes precision based on predicted bounding boxes having an intersection-over-union (IoU) greater than 50%; (2) mAP@50:95, which averages the AP scores calculated at IoU thresholds ranging from 50% to 95% in steps of 5%.

To provide a comprehensive evaluation of the model, this study incorporates not only detection accuracy metrics but also model size and computational complexity (FLOPs) as auxiliary indicators. The number of parameters reflects the model's scale and memory footprint, while FLOPs quantify the computational cost during inference, serving as a proxy for real-time efficiency and energy consumption.

4.4. Training convergence analysis

To comprehensively evaluate the training dynamics and convergence characteristics of SEDA-Net, this study provides the learning curves during the training process. Figure 7 illustrates the variation of the training loss and mAP values across the two

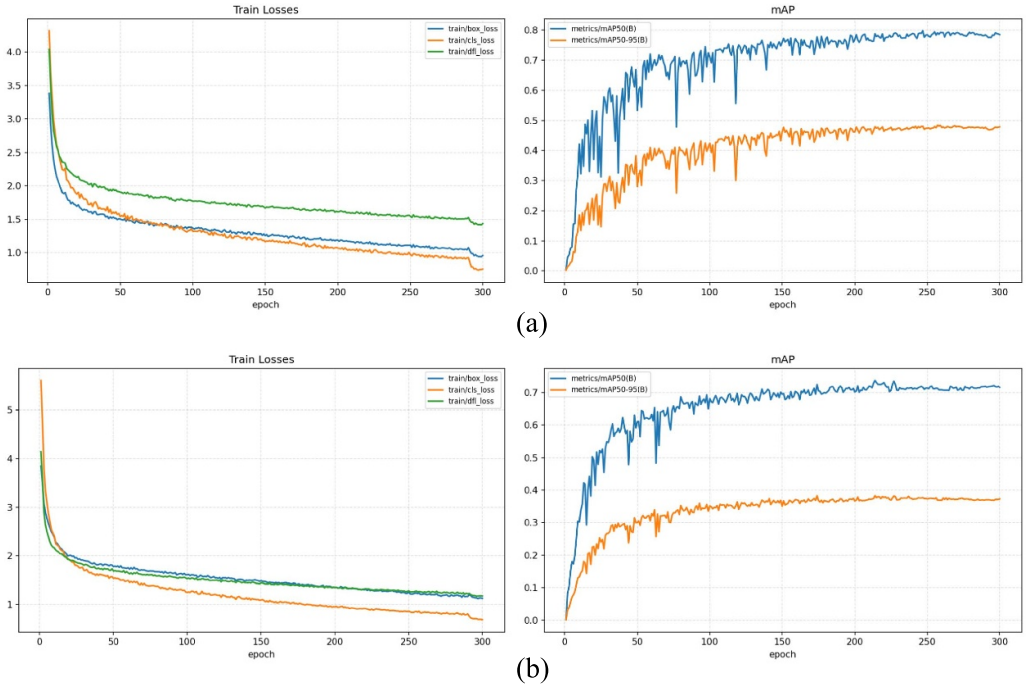


Figure 8. Precision–Recall (PR) curves of SEDA-Net on NEU-DET (a) and GC10-DET (b) datasets.

datasets. The loss curves exhibit a smooth and monotonically decreasing trend, eventually converging without significant fluctuations, which confirms the stability of the optimization process. Concurrently, the mAP curves stabilize after approximately 250 epochs, indicating that the model has reached a state of convergence during training and has not exhibited overfitting. These results confirm the robustness and reliability of the proposed model during the training process.

4.5. Baseline

To comprehensively evaluate the overall performance of the proposed SEDA-Net, this study selected fourteen representative networks, including two major categories: general-purpose object detection networks and networks designed explicitly for steel surface defect detection.

General-purpose detectors include the CNN-based frameworks Faster R-CNN [5], RetinaNet [26], and YOLO series [46], as well as Transformer-based architectures DETR [23] and RT-DETR [25]. Faster R-CNN achieves high-precision end-to-end training through the introduction of a region proposal network; RetinaNet addresses class imbalance via its focal loss; the YOLO family unifies feature extraction, bounding-box regression, and class prediction into a single forward pass, significantly simplifying the pipeline and accelerating inference; DETR removes redundant components by leveraging global self-attention and a bipartite matching loss, yielding accuracy on par with Faster R-CNN in a more streamlined architecture; and RT-DETR augments this design with an efficient hybrid encoder and an uncertainty-driven query selection mechanism to balance end-to-end learning with real-time inference capabilities.

Steel surface defect detectors comprise GC-Net [9], LHATA-Net [47], CAT-YOLO [48], and PMSA-DyTr [49]. GC-Net enhances multi-scale robustness through global attention and cascaded feature fusion; LHATA-Net strengthens the perception of fine cracks and pits via hierarchical attention and adaptive fusion; CAT-YOLO integrates convolutional and attention modules into a lightweight detection head for efficient localization; and PMSA-DyTr employs multi-scale pixel-level self-attention with dynamic transformation modules to improve recognition of low-contrast defects.

All comparison models were tested on the NEU-DET and GC10-DET datasets using the same training environment and evaluation metrics.

4.6. Experimental results and analysis

To assess the practical effectiveness of the proposed method, extensive experiments were conducted on two benchmark datasets: NEU-DET and GC10-DET. These datasets feature diverse defect categories and complex background noise, providing a challenging and reliable basis for performance evaluation.

The quantitative evaluation results presented in table 2 demonstrate that SEDA-Net achieves a mAP@50 of 79.2% on the NEU-DET dataset, outperforming all compared methods. Similarly, on the GC10-DET dataset, the proposed method attains a mAP@50 of 70.9%, as reported in table 3, surpassing existing state-of-the-art approaches.

Figure 8 displays the Precision-Recall (PR) curves of the proposed algorithm across different defect categories on the two datasets. The PR curves of the proposed method consistently encompass the curves of the baseline method (YOLO11)

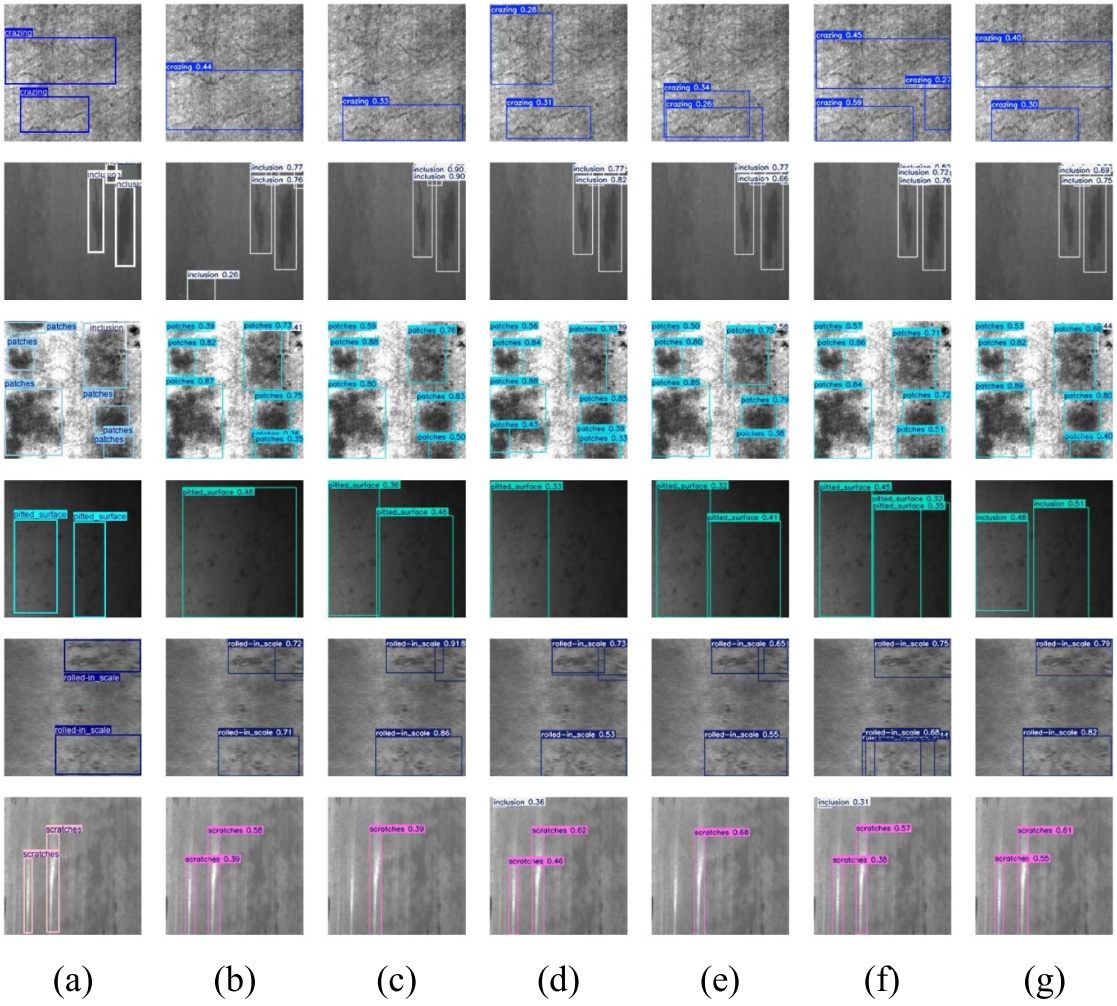


Figure 9. Comparison of detection results for the NEU-DET dataset. From top to bottom, blue, white, sky blue, cyan, dark blue, and pink represent Cr, In, Pa, PS, RS, and Sc defects, respectively. (a) True defect regions. (b) YOLOv11. (c) YOLOv12. (d) LHATA-Net. (e) GC-Net. (f) CAT-YOLO. (g) SEDA-Net.

for all defect categories, which further substantiates the effectiveness of our improvements and confirms the superior detection capability of SEDA-Net.

Figure 9 presents a visual comparison of defect detection results on the NEU-DET dataset. The proposed SEDA-Net outperforms existing methods, particularly under challenging conditions such as background noise and low-contrast defects (rows 1, 5, and 6), where it achieves more precise localization and fewer false positives and false negatives. In scenarios involving high inter-class similarity (row 3), SEDA-Net also demonstrates superior discrimination, yielding higher confidence predictions and improved defect localization. These results highlight the robustness and accuracy of SEDA-Net across various complex detection scenarios.

Figure 10 shows detection results on the GC10-DET dataset. For low-contrast defects (rows 6 and 10), SEDA-Net provides more accurate localization and higher detection confidence. In cases of inter-class similarity (rows 4 and 9), SEDA-Net effectively distinguishes between visually similar defect types.

To provide deeper insights, we visualized the network's output features using heatmaps (figures 11 and 12). These visualizations demonstrate that SEDA-Net accurately focuses on defect core regions while capturing fine-grained texture details, contributing to its strong detection capability across various defect types.

4.7. Ablation study

To evaluate the individual contributions of the GPFM, RepDWR, and DyAIFI modules within the proposed network, a series of ablation experiments were conducted on the NEU-DET and GC10-DET datasets. Starting from the YOLO11 baseline, we tested every possible combination of the GPFM (A), the RepDWR (B), and the DyAIFI (C). The results are reported in tables 4 and 5, which include performance for single-module, dual-module, and full-model configurations.

As presented in tables 4 and 5, integrating any single module yields a significant performance gain over the baseline, validating that each component addresses a distinct limitation.

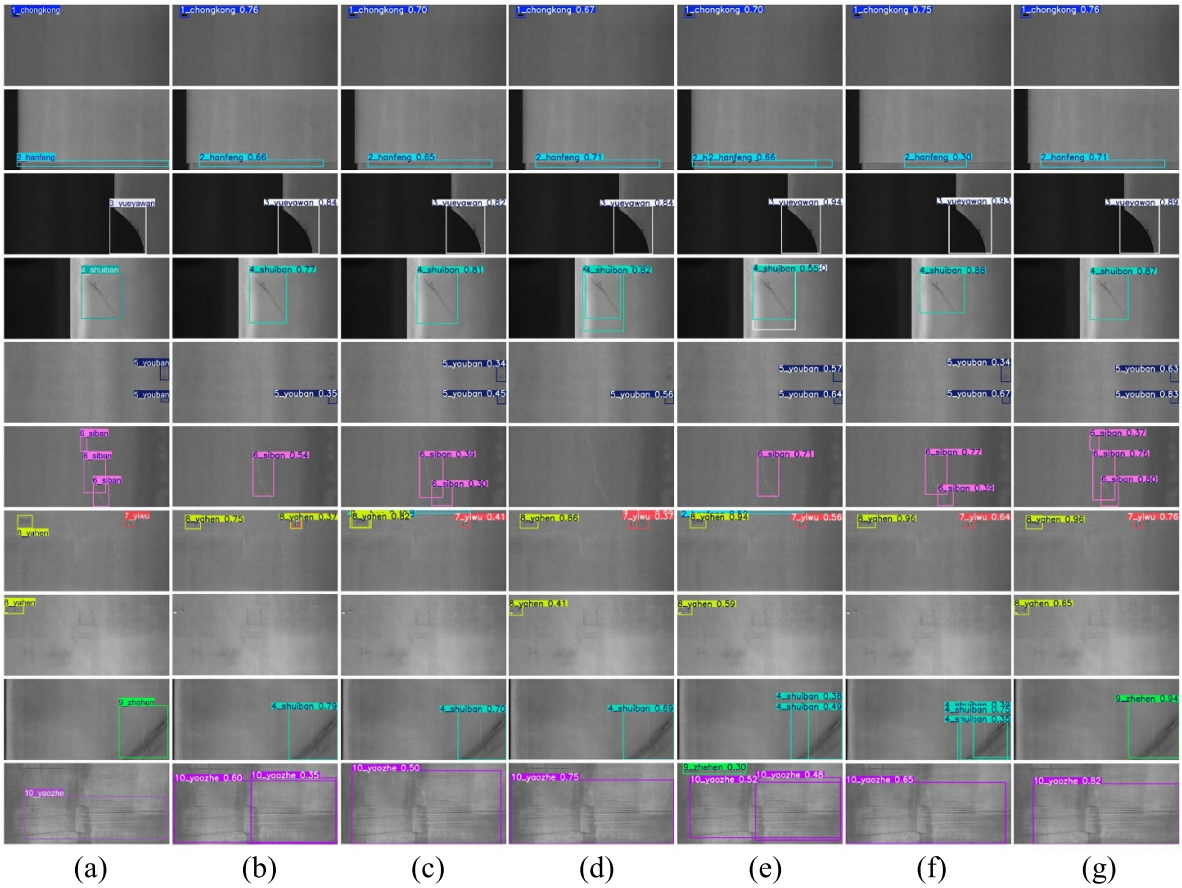


Figure 10. Comparison of detection results for the GC10-DET dataset. From top to bottom, indigo blue, sky blue, light gray, turquoise, navy blue, pinkish purple, coral red, bright yellow, vivid green, and lotus pink represent defects Pu, Wl, Cg, Ws, Os, Ss, In, Rp, Cr, and Wf, respectively. (a) True defect regions. (b) YOLOv11. (c) YOLOv12. (d) LHATA-Net. (e) GC-Net. (f) CAT-YOLO. (g) SEDA-Net.

Table 4. Ablation experiments results for the NEU-DET dataset.

Method	mAP	Cr	Pa	In	PS	RS	Sc	FLOPs	Params
YOLO11	76.0	44.7	88.4	88.3	81.2	56.0	97.6	21.3	9.4
+A	77.4	50.1	89.6	86.2	82.4	59.8	96.1	20.9	9.3
+B	77.4	42.4	90.3	89.4	85.5	61.0	96.0	21.0	9.1
+C	77.5	45.2	87.8	85.3	86.2	62.1	98.3	21.6	10.8
+A+B	78.2	47.1	92.0	88.6	85.6	60.2	95.6	20.6	9.1
+A+C	78.4	48.3	89.4	87.1	86.9	61.2	97.3	21.3	9.4
+B+C	77.7	43.8	89.6	87.6	85.8	62.6	96.8	21.3	9.4
SEDA-Net	79.2	49.7	91.2	89.6	85.1	60.7	98.8	20.9	10.5

Further performance enhancements are observed when combining any two modules, demonstrating their complementary functionalities. The complete SEDA-Net, integrating all three modules, attains the highest mAP on both the NEU-DET (79.2%) and GC10-DET (70.9%) datasets, with only a marginal increase in parameters compared to the baseline. This monotonic improvement across single, dual, and triple module combinations underscores the strong synergy and minimal redundancy among the components. In summary, GPFM, RepDWR, and DyAIFI contribute distinct functions which collectively improve detection robustness on complex steel surface imagery.

5. Disscussion and conclusion

5.1. Disscussion

The experimental results validate that SEDA-Net's superior performance is rooted in its targeted architectural design. The ablation studies (tables 4 and 5) confirm the synergistic and non-redundant contributions of our proposed modules.

The enhanced performance on low-contrast and noisy defects (figures 9 and 10) is primarily attributed to the GPFM, which effectively suppresses background interference while amplifying fine-grained features, a claim substantiated by the focused activation maps in figures 11 and 12 Concurrently, the

Table 5. Ablation experiments results for the GC10-DET dataset.

Method	mAP	Pu	Wl	Cg	Ws	Os	Ss	In	Rp	Cr	Wf	FLOPs	Params
YOLO11	65.7	97.1	93.1	91.5	78.6	65.8	58.8	30.4	33.8	31.2	76.5	21.3	9.4
+A	67.3	96.0	94.1	92.8	79.6	67.4	61.3	33.2	32.7	37.1	78.6	20.9	9.3
+B	67.7	96.4	93.8	92.1	82.3	65.3	61.2	31.9	33.9	43.9	76.4	21.0	9.1
+C	67.3	96.0	91.0	93.5	79.6	67.2	62.3	34.5	34.7	34.6	79.8	21.6	10.8
+A+B	68.5	97.2	94.8	94.3	81.4	68.1	59.5	37.1	34.3	41.9	76.8	20.6	9.1
+A+C	68.7	97.8	90.3	96.4	81.1	65.1	65.4	34.3	32.0	45.7	79.1	21.3	9.4
+B+C	67.8	94.8	92.8	90.7	82.6	63.6	61.2	35.3	36.6	40.7	79.9	21.3	9.4
SEDA-Net	70.9	96.1	91.6	95.4	84.2	68.2	63.7	36.8	38.0	50.9	83.8	20.9	10.5

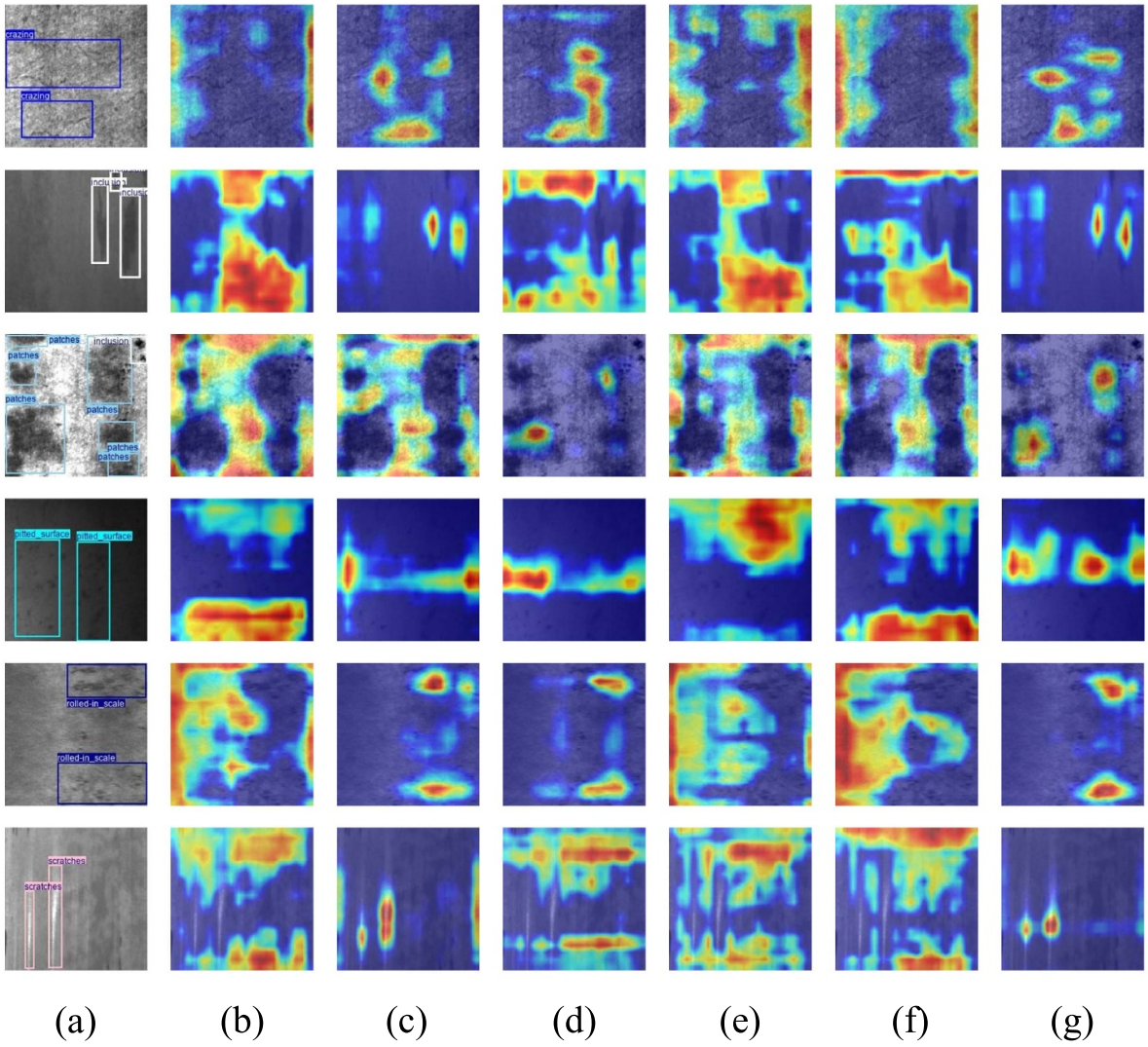


Figure 11. Feature heatmap visualization based on Eigen CAM for different networks on the NEU-DET dataset. (a) True defect regions. (b) YOLOv11. (c) YOLOv12. (d) LHATA-Net. (e) GC-Net. (f) CAT-YOLO. (g) SEDA-Net.

DyAIFI module resolves the critical challenge of inter-class similarity by fostering a more discriminative feature space through dynamic local-global interaction. The efficiency and multi-scale capability of the RepDWR module ensure robust context capture without a significant computational overhead.

SEDA-Net demonstrates high practical feasibility for industrial deployment. On a consumer-grade NVIDIA GeForce RTX 4060 graphics processing unit, SEDA-Net achieves an inference speed of 46.08 FPS. Following optimization and deployment using the TensorRT framework, its performance can be further enhanced to 106.38 FPS. Furthermore, on the NVIDIA Jetson Xavier NX embedded platform, the model maintains a real-time inference speed of 18.9 FPS.

Despite these promising results, this study has certain limitations. Akin to most supervised methods, the performance of SEDA-Net is highly contingent upon the quality and quantity of the training data. However, the acquisition of high-quality defect datasets in industrial settings is often costly and laborious, making data scarcity a prevalent challenge. In our future work, we plan to investigate the use of computer vision techniques for synthetic data augmentation to alleviate this issue and further unlock the full potential of our proposed method.

5.2. Conclusion

This work analyzes the challenges in steel surface defect detection tasks, including inter-class similarity, intra-class variability, low contrast, and noise interference. These characteristics can lead to inaccuracies in defect localization, as well as false negatives and false positives in object detection. To address these challenges, this work proposes a novel steel surface defect detection network called SEDA-Net. We designed a GPFM, which effectively enhances the ability of low-level features to perceive fine-grained information. By combining the joint modeling capabilities of deep separable convolutions and CGA modules, the model effectively suppresses global background noise and enhances its ability to focus on defect features. Additionally, to better capture contextual information, we designed the RepDWR Module, which

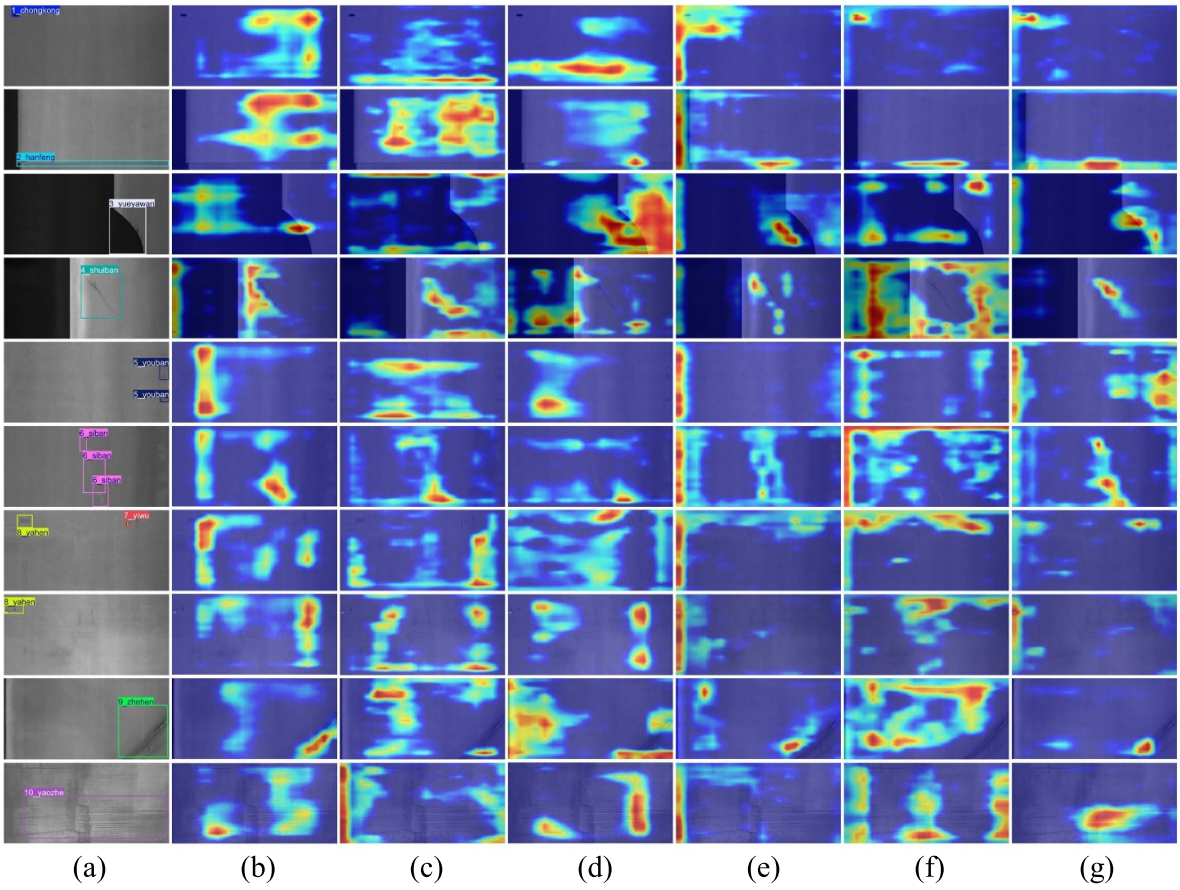


Figure 12. Feature heatmap visualization based on Eigen CAM on different networks for GC10-DET. (a) Actual defect region. (b) YOLOv11. (c) YOLOv12. (d) LHATA-Net. (e) GC-Net. (f) CAT-YOLO. (g) SEDA-Net.

improves the receptive field through multi-scale expansion convolutions to accurately capture the contextual information of surface defects and optimizes inference efficiency through reparameterization techniques. To address the issue of intra-class variability and inter-class similarity in defects, we introduced the DyAIFI Module. This module enhances the dynamic interaction between local features in defect regions and global contextual information, reinforcing the fine-grained discriminative capability.

Compared with other classical detection methods, SEDA-Net outperforms existing state-of-the-art approaches with comparable parameters on both the NEU-DET and GC10-DET datasets, achieving mAP@50 scores of 79.2% and 70.9%, respectively. Concurrently, our algorithm exhibits high real-time performance and is readily deployable. Therefore, the steel surface defect detection method proposed in this paper is efficient, accurate, and practical, making it suitable for industrial applications.

Data availability statement

Data sharing can be available at a reasonable request. The data cannot be made publicly available upon publication due to legal restrictions preventing unrestricted public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

The authors are sincerely grateful to the anonymous reviewers for their careful reading of the manuscript and for providing insightful comments and constructive suggestions that significantly enhanced the quality and clarity of this paper.

Funding

This work is supported by the Natural Science Foundation of Hunan Province under Grant 2025JJ50399.

Author contributions

Jing Liao 0000-0003-3346-7419
Conceptualization (equal), Data curation (equal), Formal analysis (equal), Investigation (equal), Methodology (equal), Software (equal), Supervision (equal), Validation (equal), Writing – original draft (equal)

Mingliang Liu 0009-0009-9273-782X
Formal analysis (equal), Methodology (equal), Resources (equal), Supervision (equal), Validation (equal), Writing – original draft (equal)

Lei Jiang 0000-0002-5654-7748
Methodology (equal), Validation (equal), Visualization (equal), Writing – original draft (equal)

Kuanching Li 0000-0003-1381-4364
Investigation (equal), Methodology (equal), Project administration (equal), Resources (equal), Validation (equal), Visualization (equal), Writing – review & editing (equal)

References

[1] Hou X, Liu M, Zhang S, Wei P and Chen B 2023 Canet: contextual information and spatial attention based network for detecting small defects in manufacturing industry *Pattern Recognit.* **140** 109558
[2] Luo Q and He Y 2016 A cost-effective and automatic surface defect inspection system for hot-rolled flat steel *Robot. Comput. Integr. Manuf.* **38** 16–30
[3] Wang X, Zhou L, Li K C, Zheng S and Fan H 2024 IAtraj: multi-modal trajectory prediction through contextual information spatio-temporal interaction and awareness *Int. J. Interact. Multimedia Artif. Intell.* 1–12
[4] Redmon J, Divvala S, Girshick R and Farhadi A 2016 You only look once: unified, real-time object detection *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 779–88
[5] Ren S, He K, Girshick R and Sun J 2015 Faster R-CNN: towards real-time object detection with region proposal networks *Advances in Neural Information Processing Systems* p 28
[6] Shao R, Zhou M, Li M, Han D and Li G 2024 TD-Net: tiny defect detection network for industrial products *Complex Intell. Syst.* **10** 3943–54
[7] Guiqiang W, Junbao C, Chengzhang L and Shuo L 2025 Edge-YOLO: lightweight multi-scale feature extraction for industrial surface inspection *IEEE Access* **13** 48188–201
[8] Li Z, Wei X, Hassaballah M, Li Y and Jiang X 2024 A deep learning model for steel surface defect detection *Complex Intell. Syst.* **10** 885–97
[9] Liu G, Chu M, Gong R and Zheng Z 2025 Global attention module and cascade fusion network for steel surface defect detection *Pattern Recognit.* **158** 110979
[10] Gao Y, Tian G Y, Li K, Ji J, Wang P and Wang H 2015 Multiple cracks detection and visualization using magnetic flux leakage and eddy current pulsed thermography *Sens. Actuators A* **234** 269–81
[11] Ameri R, Hsu C C and Band S S 2024 A systematic review of deep learning approaches for surface defect detection in industrial applications *Eng. Appl. Artif. Intell.* **130** 107717
[12] Song K and Yan Y 2013 A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects *Appl. Surf. Sci.* **285** 858–64
[13] Fekri-Ershad S and Tajeripour F 2017 Multi-resolution and noise-resistant surface defect detection approach using new version of local binary patterns *Appl. Artif. Intell.* **31** 395–410
[14] Aminzadeh M and Kurfess T 2015 Automatic thresholding for defect detection by background histogram mode extents *J. Manuf. Syst.* **37** 83–92
[15] Cai J, Liang W, Li X, Li K, Gui Z and Khan M K 2023 GTxChain: a secure IoT smart blockchain architecture based on graph neural network *IEEE Internet Things J.* **10** 21502–14
[16] Zhou S, Li K, Chen Y, Yang C, Liang W and Zomaya A Y 2024 Trustbcfl: mitigating data bias in iot through blockchain-enabled federated learning *IEEE Internet Things J.* **11** 25648–62
[17] Liao J, Guo L, Jiang L, Yu C, Liang W, Li K and Pop F 2025 A machine learning-based feature extraction method for image classification using ResNet architecture *Digital Signal Process.* **160** 105036
[18] Girshick R, Donahue J, Darrell T and Malik J 2014 Rich feature hierarchies for accurate object detection and semantic segmentation *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 580–7
[19] Girshick R 2015 Fast R-CNN *Proc. IEEE International Conf. on Computer Vision* pp 1440–8
[20] Dai J, Li Y, He K and Sun J 2016 R-fcn: Object detection via region-based fully convolutional networks *Advances in Neural Information Processing Systems* p 29
[21] Chen Z , Zhang R, Hsieh M-Y, Souri A and Li K-C 2025 Average Sigmoid-Tanh Attention and multi-filter partially decoupled mechanism via YOLOv7 for detecting weld proximity defects *Metall. Mater. Trans. B* **56** 4186–200
[22] Liu W *et al* 2016 Ssd: Single shot multibox detector *Computer Vision–ECCV 2016: 14th European Conf.* (*Amsterdam, The Netherlands, 11–14 October 2016*) (*Proc. Part I vol 14*) (Springer) pp 21–37
[23] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S 2020 End-to-end object detection with transformers *European Conference on Computer Vision* (Springer) pp 213–29
[24] Liu Z *et al* 2021 Swin transformer: hierarchical vision transformer using shifted windows *Proc. of the IEEE/CVF International Conference on Computer Vision* pp 10012–22
[25] Zhao Y *et al* 2024 Detrs beat yolos on real-time object detection *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 16965–74
[26] Cheng X and Yu J 2020 RetinaNet with difference channel attention and adaptively spatial feature fusion for steel surface defect detection *IEEE Trans. Instrum. Meas.* **70** 1–11
[27] Cui L, Jiang X, Xu M, Li W, Lv P and Zhou B 2021 SDDNet: a fast and accurate network for surface defect detection *IEEE Trans. Instrum. Meas.* **70** 1–13
[28] Zhao W, Chen F, Huang H, Li D, Cheng W and Versaci M 2021 A new steel defect detection algorithm based on deep learning *Comput. Intell. Neurosci.* **2021** 5592878
[29] Wang H, Li Z and Wang H 2021 Few-shot steel surface defect detection *IEEE Trans. Instrum. Meas.* **71** 1–12
[30] Wang Y, Wang H and Xin Z 2022 Efficient detection model of steel strip surface defects based on YOLO-V7 *IEEE Access* **10** 133936–44
[31] Yeung C C and Lam K M 2022 Efficient fused-attention model for steel surface defect detection *IEEE Trans. Instrum. Meas.* **71** 1–11
[32] Huang J, Zhang X, Jia L and Zhou Y 2025 An improved you only look once model for the multi-scale steel surface defect detection with multi-level alignment and cross-layer redistribution features *Eng. Appl. Artif. Intell.* **145** 110214
[33] Huang Y, Chen Z, Chen Z, Zhou D and Pan E 2025 Lightweight defect detection network based on steel strip raw images *Eng. Appl. Artif. Intell.* **145** 110179
[34] Peng Y, Xia F, Zhang C and Mao J 2024 Deformation feature extraction and double attention feature pyramid network for

bearing surface defects detection *IEEE Trans. Ind. Inform.* **20** 9048–58
[35] Vaswani A *et al* 2017 Attention is all you need *Advances in Neural Information Processing Systems* p 30
[36] Diao C, Zhang D, Liang W, Jiang M and Li K 2024 A novel attention-based dynamic multi-graph spatial-temporal graph neural network model for traffic prediction *IEEE Trans. Emer. Top. Comput. Intell.* **9** 1910–23
[37] Liu X, Peng H, Zheng N, Yang Y, Hu H and Yuan Y 2023 Efficientvit: memory efficient vision transformer with cascaded group attention *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 14420–30
[38] Lin W, Wu Z, Chen J, Huang J and Jin L 2023 Scale-aware modulation meet transformer *Proc. IEEE/CVF International Conf. Computer Vision* pp 6015–26
[39] Hu J, Shen L and Sun G 2018 Squeeze-and-excitation networks *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 7132–41
[40] Woo S, Park J, Lee J Y and Kweon I S 2018 Cbam: convolutional block attention module *Proc. European Conf. on Computer Vision (ECCV)* pp 3–19
[41] Sandler M, Howard A, Zhu M, Zhmoginov A and Chen L C 2018 Mobilenetv2: inverted residuals and linear bottlenecks *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 4510–20
[42] Chen L, Gu L, Zheng D and Fu Y 2024 Frequency-adaptive dilated convolution for semantic segmentation *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 3414–25
[43] He Y, Song K, Meng Q and Yan Y 2019 An end-to-end steel surface defect detection approach via fusing multiple hierarchical features *IEEE Trans. Instrum. Meas.* **69** 1493–504
[44] Lv X, Duan F, Jiang J-J, Fu X and Gan L 2020 Deep metallic surface defect detection: the new benchmark and detection network *Sensors* **20** 1562
[45] Everingham M, Van Gool L, Williams C K, Winn J and Zisserman A 2010 The pascal visual object classes (voc) challenge *Int. J. Comput. Vision* **88** 303–38
[46] Ali M L and Zhang Z 2024 The YOLO framework: a comprehensive review of evolution, applications and benchmarks in object detection *Computers* **13** 336
[47] Lv S, Liang T, Zhang K, Jiang S, Ouyang B, Li Q and Li X 2025 A lightweight hierarchical aggregation task alignment network for industrial surface defect detection *Expert Syst. Appl.* **263** 125727
[48] Yang J and Liu Z 2024 A novel real-time steel surface defect detection method with enhanced feature extraction and adaptive fusion *Eng. Appl. Artif. Intell.* **138** 109289
[49] Su J, Luo Q, Yang C, Gui W, Silvén O and Liu L 2024 Pmsa-dytr: prior-modulated and semantic-aligned dynamic transformer for strip steel defect detection *IEEE Trans. Ind. Inform.* **20** 6684–95
[50] Wang C Y, Yeh I H and Mark Liao H Y 2024 Yolov9: learning what you want to learn using programmable gradient information *European Conf. Computer Vision* (Springer) pp 1–21
[51] Wang A *et al* 2024 Yolov10: real-time end-to-end object detection *Advances in Neural Information Processing Systems* vol 37 pp 107984–8011
[52] Tian Y, Ye Q and Doermann D 2025 Yolov12: attention-centric real-time object detectors (arXiv:250212524)