Journal of Intelligent Manufacturing
https://doi.org/10.1007/s10845-025-02613-5

AMFNet: aggregated multi-level feature interaction fusion network for defect detection on steel surfaces

Changyun Wei[1] · Yuhang Bao[1] · Chengwei Zheng[1] · Ze Ji[2]

Received: 9 November 2024 / Accepted: 15 April 2025

Abstract
Current models for detecting defects on steel surfaces struggle to fully utilize potential positional and semantic information. Usually, these models merely combine high-level and low-level features in a straightforward manner, leading to an increase in redundant information. To address this challenge, this study presents an aggregated multi-level feature interaction fusion network (AMFNet). Specifically, the AMFNet incorporates a branch interaction module (BIM) that branches and fuses features channel-wise to facilitate feature interaction. Moreover, it also includes a dilated context module (DCM) that expands the receptive field to capture contextual information across various scales effectively. In addition, we propose a spatial correlation module (SCM) that is designed to recognize spatial dependencies between adjacent feature maps and generate attention weights. Our performance evaluations on the NEU-DET and GC10-DET dataset reveal that our proposed AMFNet significantly outperforms classical object detectors in terms of mean average precision (mAP). Moreover, it also demonstrates a modest improvement over the advanced methods recently introduced by other researchers.

Keywords Surface defect detection · Vision inspection · Deep learning · Multi-scale fusion

Introduction

Surface defect detection plays a pivotal role in object detection, focusing on accurately localizing and identifying defects within images. In automated production lines, the efficient and precise detection of defects is essential, as it directly impacts the safety and quality of manufacturing processes. However, surface defects in real-world production settings are often complex and varied. For example, on steel surfaces, common defects such as cracks, inclusions, patches, pitting, rolled-in scale, scratches, and other imperfections can be particularly challenging to detect. Traditional inspection methods typically rely on manual visual inspection, which is inefficient, time-consuming, and labor-intensive. Additionally, human fatigue and oversight can lead to missed detection of small defects. In contrast, automated visual inspection offers significant advantages over manual approaches by enabling continuous operation while enhancing accuracy and efficiency. Consequently, the demand for efficient and high-quality visual inspection methods tailored to steel surface defect detection has spurred numerous research efforts.

Over the past two decades, advancements in machine vision technology have driven rapid progress in visual inspection. Traditional methods, originally dependent on manually set parameters, have been adapted for detecting surface defects in products. Early approaches typically framed defect detection as a texture analysis problem, using strategies to extract texture information, such as the grayscale covariance matrix (Asha et al., 2011), entropy (Nand et al., 2014), and uniformity. However, these solutions are often highly sensitive to image characteristics, making them less effective in real industrial environments marked by noise and variability. To address these challenges, researchers have proposed alter-

✉ Ze Ji
JiZ1@cardiff.ac.uk

Changyun Wei
c.wei@hhu.edu.cn

Yuhang Bao
yuhang5107@163.com

Chengwei Zheng
zcw721@163.com

[1] College of Mechanical and Electrical Engineering, Hohai University, Changzhou 213200, China

[2] School of Engineering, Cardiff University, Cardiff CF24 3AA, UK

Published online: 28 April 2025

native techniques, including gradient histograms (Halfawy & Hengmeechai, 2014), local binary patterns (Fekri-Ershad & Tajeripour, 2012), and Gabor filter features (Ma et al., 2018). Nonetheless, achieving accurate defect detection typically requires the manual design of features by experienced technicians, highlighting the need for more advanced automated detection methods capable of handling the complexities of industrial environments.

Machine learning techniques have also been applied to surface defect detection, often relying on manually extracted defect features processed by rule-based or learning-based classifiers, such as support vector machines (Zhou et al., 2016), decision trees (Aghdam et al., 2012), random forests (Song et al., 2016), and conditional random fields (Yang & Yang, 2016). The performance of these methods largely depends on the accuracy of feature representation, which must be meticulously designed for each specific application (Usamentiaga et al., 2022). However, modern manufacturing demands detection algorithms that are not only accurate but also timely and robust. For more complex defects, traditional machine learning models often require significant development time and may not generalize well across different defect types.

Recent advancements in computer performance have accelerated the development of deep learning-based methods for defect detection, with multi-scale feature fusion and attention mechanisms emerging as critical techniques for enhancing accuracy. Since defects of varying scales often coexist in defect detection tasks, the extraction and fusion of multi-scale features have become essential. Deep learning models, such as the feature pyramid network (FPN) (Lin et al., 2017a) and bidirectional feature pyramid network (BiFPN) (Tan et al., 2020), have been introduced to improve detection precision by integrating rich semantic and detailed features at different levels, significantly enhancing the detection of small and complex defects. Additionally, attention mechanisms, such as channel attention (CAM) (Hu et al., 2018) and spatial attention (SAM) (Woo et al., 2018), further boost model performance by enabling the network to focus on important regions, thereby improving accuracy, particularly in the presence of complex backgrounds or small defects.

Deep learning algorithms offer significant advantages in addressing these challenges due to their strong generalization abilities, leading to notable advancements in the field. For example, the Surface Defect Detection Network (SDDNet) proposed by (Cui et al., 2021) is a fast and accurate approach that incorporates a feature retention block and a skip dense connection module to enhance detection performance, particularly for small and complex defects. Despite these advancements, existing methods still struggle to balance real-time processing with high precision, especially for defects with diverse textures and sizes. Attention mechanisms have been increasingly integrated into deep learning models to further boost performance. For instance, the work by (Song et al., 2020) introduced attention mechanisms within an encoder-decoder residual network to accelerate model convergence and improve the detection of salient objects. However, challenges remain in preserving fine boundary details, particularly in complex backgrounds. The retinal network proposed by (Luo et al., 2019) utilized differential channel attention and adaptive spatial feature fusion to achieve better segmentation accuracy, yet the trade-off between model size and computational cost still presents challenges for real-time applications. Similarly, the method introduced by (Yu et al., 2021) incorporated channel attention and bi-directional feature fusion in a fully convolutional first-order network to enhance feature representation for surface defect detection, but the handling of large-scale industrial data with high variability in defect types continues to be a challenge. Additionally, the work by (Zhou et al., 2021) proposed an end-to-end dense attention-guided cascade network that integrates multi-scale depth features to improve defect detection accuracy. However, their approach still faces difficulties in detecting low-contrast defects in highly cluttered environments. The work (Xi et al., 2023) introduced YDRSNet, which integrates multi-path fusion with Mask R-CNN, achieving improvements in detection accuracy and efficiency. Yet, a perfect balance between speed and accuracy across varying operational environments remains an ongoing challenge. Similarly, the study (Ren et al., 2024) incorporates deformable convolution (DCNv2) for better detection of irregular defects. It combines the ECA attention mechanism to enhance the weight of important features, and the computational complexity is reduced by replacing the Spatial Pyramid Pooling (SPPF) with SimSPPF, while the bounding box regression is optimized using the SIoU loss function.

Despite significant advancements, current object detection methods for detecting steel surface defects still have limitations and fail to fully leverage available information. One key issue is that they often neglect to prioritize the rich semantic information of defect features during extraction in the backbone network, which is critical for improving recognition accuracy (Peng et al., 2024). Many existing approaches attempt to integrate contextual information for multi-feature fusion by linking high-level and low-level features. However, this simplistic fusion can introduce redundant or noisy information, potentially undermining detection performance. For instance, the study (Zhao et al., 2024) proposed a Multi-Scale Adaptive Fusion (MSAF) method based on YOLOv8n for detecting steel surface defects in complex backgrounds. This approach incorporates multi-scale feature extraction, a lightweight detection head, and enhanced loss functions to improve both detection accuracy and efficiency. However, challenges remain, including false positives and missed detections in complex backgrounds, and the use of MSAF exclusively in the neck does not fully capture the fusion of

high-level semantic information with low-level detailed features. Similarly, the study (Li et al., 2024b) introduced a deep learning-based model for steel surface defect detection with improved feature extraction and fusion capabilities. Specially, it incorporates multi-scale feature extraction (MSFE) and efficient feature fusion (EFF) modules to enhance detection accuracy while reducing the number of parameters. However, the EFF module primarily performs feature fusion between the backbone and neck networks, which may overlook deeper feature interactions, thus limiting the adequacy of feature fusion. Overall, current object detection methods for steel surface defect detection often fall short in handling complex, low-contrast tasks in real-world applications. These challenges highlight the need for continued research and innovation to develop more robust and efficient detection methods capable of addressing the complexities inherent in such tasks.

To address these challenges, this paper proposes AMFNet, an advanced model that incorporates several key components to improve feature extraction and interaction. AMFNet addresses the difficulties in detecting small, low-contrast defects and enhances feature perception through three primary innovations:

- **BIM (Branch Interaction Module)**: The BIM enhances feature fusion by processing and merging multi-scale and multi-level features, thereby improving the model's ability to perceive the input image effectively.
- **DCM (Dilated Context Module)**: This module uses convolutions with different expansion rates to capture contextual information at varying scales, thereby improving feature extraction and the model's robustness to diverse defect types.
- **SCM (Spatial Correlation Module)**: The SCM captures spatial dependencies between neighboring feature maps and generates attention weights to highlight important regions, ensuring that critical information is emphasized during defect detection.

In summary, the main contributions of this paper are as follows: AMFNet effectively detects small and low-contrast defects on steel surfaces, addressing the unique challenges of industrial environments. The BIM module enhances feature perception by seamlessly fusing multi-scale and multi-level information. The DCM module improves the extraction of contextual information across different scales, while the SCM module generates spatial attention, focusing on regions rich in relevant information to enhance detection accuracy. The proposed AMFNet is evaluated on the NEU-DET and GC10-DET datasets. Experimental results demonstrate that AMFNet outperforms state-of-the-art (SOTA) models in both accuracy and computational efficiency.

Related work

Deep learning-based defect detection

Object detection is a major challenge in computer vision, involving both the identification of objects and their precise localization within an image. Recent advancements in deep learning have led to significant progress in object detection, resulting in the development of numerous CNN-based detectors, particularly for defect detection tasks. These detectors are generally classified into two categories: one-stage and two-stage object detectors.

One-stage methods have gained prominence due to their computational efficiency, leveraging CNNs for end-to-end processing to simultaneously predict bounding boxes and classification probabilities. For instance, the Single Shot MultiBox Detector (SSD) (Liu et al., 2016) employs anchors and multi-scale features to enhance small object detection, while YOLOv3 (Redmon et al., 2016) integrates three feature pyramid networks (FPNs) to address objects across various scales. Despite their speed, one-stage methods often compromise on accuracy, limiting their suitability for precision-demanding defect detection tasks. Recent advancements, such as the work by Zhao et al. (2024), introduced the MSAF-YOLOv8n model, which enhances one-stage detection through multi-scale adaptive fusion blocks and innovative loss functions like normalized weighted distance (NWD) and weighted intersection over union (WIoU). This approach improves both accuracy and efficiency, though challenges persist in complex environments with noise and illumination variations.

In contrast, two-stage methods are among the earliest approaches to object detection, involving two steps: generating region proposals and then classifying these proposals. A notable example is the R-CNN family of algorithms. R-CNN (Girshick et al., 2014) first generates candidate regions using techniques like Selective Search, followed by CNN-based feature extraction and classification for each region. Although effective, R-CNN algorithms are slow and inefficient. To overcome these limitations, Faster R-CNN (Ren et al., 2016) introduced RoI Pooling (Region of Interest Pooling) for shared feature extraction, improving both speed and performance. Building on Faster R-CNN, Mask R-CNN (He et al., 2020) adds instance segmentation, allowing it to perform pixel-level segmentation for each detected object. Additionally, Cascade R-CNN (Cai & Vasconcelos, 2019) improves detection accuracy by using multiple cascading detectors. Each stage refines the candidate regions, progressively filtering out less plausible detections, making the detection process more robust. While two-stage object detection algorithms can offer higher accuracy compared to single-stage ones, existing models still have significant room for improvement and hold potential for further enhancement.

In recent years, object detection techniques have continued to evolve, leading to the development of hybrid methods like RetinaNet (Lin et al., 2017b), which combines the strengths of both two-stage and single-stage approaches. Additionally, the Transformer architecture (Vaswani et al., 2017), originally designed for natural language processing tasks, has been adapted for object detection. This adaptation has given rise to Transformer-based models such as DETR (Carion et al., 2020), known for their strong generalization capabilities and faster processing speeds.

Steel surface defect detection

With the rapid development of deep learning in recent years, CNNs have been successfully applied to various steel surface defect detection tasks. To enhance the efficiency and accuracy of steel surface defect detection, (Damacharla et al., 2021) proposed a transfer learning-based U-Net (TLU-Net) framework utilizing ResNet and DenseNet encoders pretrained on the ImageNet dataset. While this framework improves performance, its primary limitation is the reliance on annotated data. Preparing such training data is costly and time-consuming, especially for defect segmentation tasks that require pixel-level annotations. As a result, a novel network architecture, DEA_RetinaNet (Cheng & Yu, 2020), was specifically designed to improve surface defect detection in industrial products. By integrating optimized anchors, a channel attentionA mechanism, and adaptive spatial feature fusion, it achieves remarkable performance gains. Although DEA_RetinaNet improves detection accuracy by addressing class imbalance and enhancing feature fusion, it still suffers from incomplete feature fusion, particularly in deeper layers. The study (Jain et al., 2022) employed Generative Adversarial Networks (GANs) for data augmentation to enhance the performance of steel surface defect detection and classification. By simulating the distribution of training data, additional synthetic data is generated, augmenting the training of real data. To further improve detection capabilities for defects of various sizes, the work (Hao et al., 2021) proposed a steel surface defect inspection model named defect inspection network (DIN) for smart industrial monitoring. This model combines a deformable convolution-enhanced backbone network with a balanced feature pyramid. While the deformable convolutional backbone and balanced feature pyramid improve defect detection, the DIN model can be computationally expensive due to its complex architecture.

Since the methods mentioned above do not fully consider the efficiency of the models or the characterization of defects, some scholars have focused on ensuring model detection performance while accounting for efficiency. The work (Yeung & Lam, 2022) introduced a fused-attention network (FANet) specifically designed for detecting steel surface defects. This framework utilizes a fused-attention mechanism applied to a single balanced feature map, optimizing both accuracy and detection speed. Additionally, it introduces the adaptively balanced feature fusion (ABFF) method, which fuses features with appropriate weights to enhance discriminative power. The fused-attention module (FAM) module is proposed to handle shape variations in defects, improving both localization and classification. To address the challenges posed by steel plate surface defects such as diverse types, complex and irregular shapes, and a wide scale range, the work (Song et al., 2023) proposed a detection method based on deformable convolution and background suppression. This method incorporates an improved Faster RCNN with deformable convolution and region-of-interest (ROI) alignment to enhance detection performance for large-scale defects with complex and irregular shapes. Additionally, the study (Gao et al., 2023) introduced a method that includes feature alignment to map unrecognizable defects so as to recognizable areas. The method employs a hierarchical training strategy to integrate this feature alignment into the training process. While these methods demonstrate strong performance on specific datasets, they may not generalize well to other defect types or surfaces with different characteristics.

Recent studies have further advanced steel surface defect detection through sophisticated feature fusion techniques. For example, He et al. (2019) proposed a multilevel-feature fusion network (MFN), integrating features from multiple CNN layers to improve localization accuracy. This model's key contribution is its integration of multilevel features from various CNN layers, which enhances defect localization accuracy compared to single-level features. By combining multiple feature maps containing both low- and high-level characteristics, this approach improves precision in detecting steel defects. Additionally, Zhang et al. (2023) introduced a cross-scale weighted feature fusion network with Laplace sharpening and an enhanced bidirectional FPN, excelling at small defect detection. In 2024, Li et al. (2024a) enhanced YOLOX with CSPCrossLayer for gradient enrichment, a Squeeze-and-Excitation (SA) module for key feature emphasis, and a PSblock for efficient multi-scale fusion, achieving a strong balance of accuracy and speed. Similarly, (Zhao et al., 2024) proposed MSAF-YOLOv8n, incorporating multi-scale adaptive fusion and advanced loss functions to boost precision in complex backgrounds. Peng et al. (2024) introduced a deformation-aware approach, prioritizing semantic-rich defect features during backbone extraction, addressing limitations in simplistic high- and low-level feature fusion. Despite these advancements, challenges persist in achieving robustness under noise, illumination changes, and low-contrast conditions where defects resemble background features, underscoring the need for continued innovation. Our proposed AMFNet builds on these developments by integrating the Branch Interaction Module (BIM), Dilated Context

Module (DCM), and Spatial Correlation Module (SCM) to enhance feature perception, contextual extraction, and spatial attention, addressing the gaps in existing methods for steel surface defect detection.

The proposed AMFNet

Inspired by Faster R-CNN, this paper introduces the AMFNet, as illustrated in Fig. 1. AMFNet is designed specifically for detecting steel surface defects. This section outlines the backbone network, fusion module, and loss function employed in the proposed method. The architecture of our AMFNet incorporates several key components: BIM, SCM, and DCM. The feature extraction utilizes a feature pyramid network (FPN) with a Faster R-CNN ResNet50 backbone.

Initially, the input image undergoes processing through the FPN, producing a multi-scale feature pyramid denoted as $C_{(i=1,2,3,4,5)}$. Given the C_1 layer has the highest resolution and demands significant computational resources, and considering the necessity to prioritize detailed feature information, we utilize $[C_2, C_3, C_4, C_5]$ as feature inputs for the subsequent processing in the BIM module. Moreover, the SCM module is nested in each BIM module to fuse the adjacent branch features in the BIM module, which can promote the information fusion of adjacent features after the initial fusion.

The feature extracted from C_5 are fed into the DCM, which filters and enhances the extraction of detailed features. Ultimately, the fused and enhanced features $[F_2, F_3, F_4, F_5]$ are inputted into the detection head of the region proposal network (RPN) and RCNN for defect classification and localization.

Branch interaction module

The image of the steel surface is generated by extracting deep features from the backbone of a deep convolutional neural network. These features can capture the complex textures and defects of the steel surface, including small scratches, pits, cracks and other defects. Traditional convolutional neural networks usually have difficulty in capturing both tiny defects (such as scratches and cracks) and large-area distributed defect information. To address the shortcomings of existing methods in dealing with complex textures and multi-scale defects, and to adapt to the diversity of these defects in size, shape and distribution, the BIM module is designed for steel surface defect detection and uses dilated convolutions with different dilation rates to extract multi-scale features. These processed features are then reintegrated with the original input through jump connections, thereby enhancing the network's ability to detect targets of different scales.

As shown in Fig. 2, this paper uses C_5 as the input example. Assume that the input feature map has a size of $C \times H \times W$. The initial feature C_5 is divided into four segments along the channel dimension: $[a_1, a_2, a_3, a_4]$, each with a size of $C/4 \times H \times W$. Each segment undergoes convolution and batch normalization to extract local information, helping to reduce redundancy and optimize the capture of fine defects (such as scratches and cracks), while keeping the feature dimensions at $C/4 \times H \times W$. Subsequently, to select, integrate, and enhance features with different semantic information, the 1×1 convolution and batch-normalized features are input into the SCM, together with the adjacent features $a_{i=1,2,3,4}$ (as shown in the figure). This process generates $S_{i=2,3,4}$, each with a size of $C/4 \times H \times W$. This ensures more accurate handling of small defects while capturing a broader range of surface defects. It further strengthens the module's ability to capture multi-scale and multi-semantic information.

Subsequently, the feature maps S_2, S_3, and S_4 are dilated using dilation rates of 3, 5, and 7, respectively. This allows our network to expand the receptive field while preserving fine details, enabling it to capture defects on the steel surface at different scales. At the same time, this multi-scale configuration balances detail preservation and global information capture in scenarios where defects on the steel surface are distributed diversely, which is difficult to achieve with traditional dilated convolution configurations. Next, these feature maps are upsampled using bilinear interpolation to match the dimensions of the feature map obtained from a_1 after convolution and batch normalization. The upsampled feature tensor is then concatenated with the feature map from a_1 along the channel dimension, forming a new feature map C_5' with dimensions of $C \times H \times W$. This cross-scale feature integration enhances the robustness of the module. To further improve the module's robustness, C_5' is added pixel-wise to the original input C_5, resulting in the final output B_5. This output effectively fuses multi-scale information and enhances the model's ability to adapt to non-defective regions. Additionally, it reduces false positives in non-defective areas, making it particularly suitable for the complex texture environments of steel surfaces. The corresponding calculations are expressed in Eq. (1) as follows:

$$\begin{cases} a_i = chunk(C_5), i = 1, 2, 3, 4 \\ S_i = SCM(a_{i-1}, BN(Conv_1(a_i))), i = 2, 3, 4 \\ C_5' = Cat(BN((US(DilConv_{2i-1}(S_i))), \\ Conv_1(a_1)))), i = 2, 3, 4 \\ B_5 = C_5' \oplus C_5 \end{cases} \tag{1}$$

where *chunk* denotes the partitioning of the tensor based on the channel dimension, and i indicates the layer of divisions by chunks. a_i represents the features after being divided into

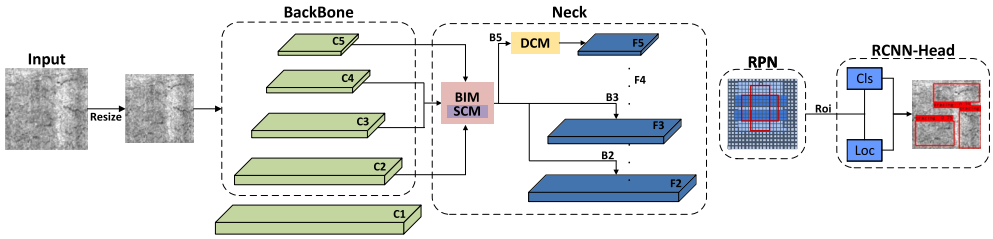
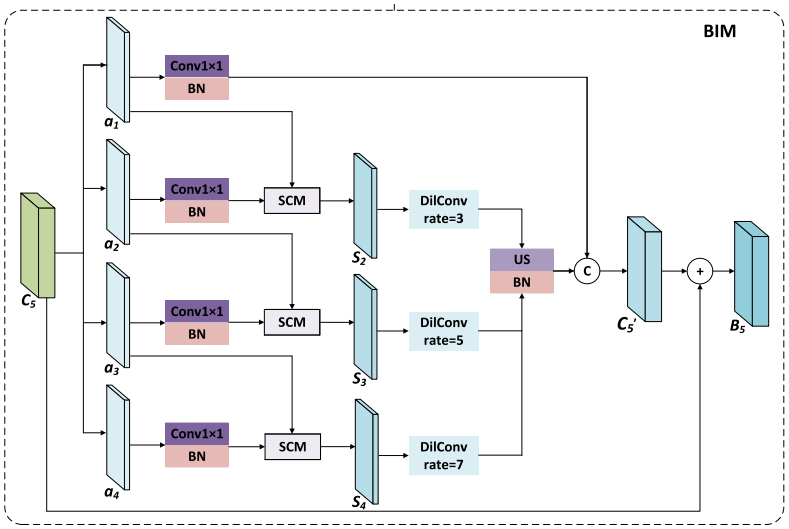


Fig. 1 Overall architecture of the proposed AMFNet

Fig. 2 Detailed illustration of the proposed branch interaction module (BIM)



four layers, and BN stands for batch normalization. $Conv_1$ and $DilConv_{2i-1}$ represent the convolutional layer with a 1×1 convolutional kernel and the dilated convolutional layer with a dilation rate of $2i - 1$, respectively. US represents upsample, and we use bilinear interpolation as the method. Moreover, S_i represents the result of the SCM operation. The operator Cat signifies the concatenation of the feature tensors based on the number of channels. The symbol $\oplus$ signifies element-wise addition.

Spatial correlation module

To improve the network's ability to capture global contextual features and enable more flexible focus on regions of interest, the SCM module is designed to address the complexity of surface textures and the irrelevance of local features on steel surfaces. The innovative design of SCM, as illustrated in Fig. 3, enables more effective identification of important regions while filtering out irrelevant local information.

The SCM retains the original channel number and spatial dimensions ($C \times H \times W$) of the input feature maps x_1 and x_2, ensuring that important spatial information is not lost during feature transformation. This is critical for defect detection tasks, as many defects exhibit distinct spatial patterns.

First, a convolution and ReLU operation is applied to the input tensor x_1, followed by a transpose operation to obtain

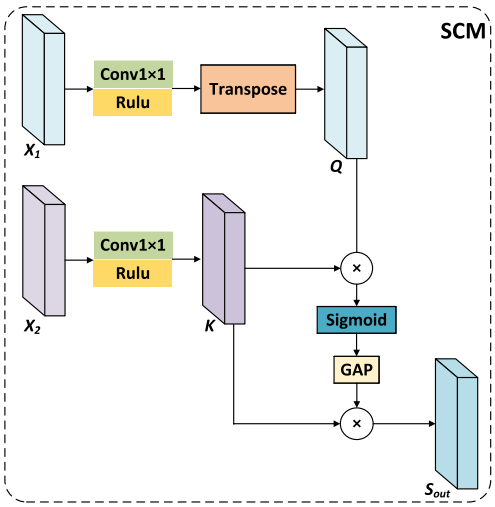


Fig. 3 Detailed illustration of the proposed spatial correlation module (SCM)

Q, which has dimensions of $C \times W \times H$. Then, a direct convolution and ReLU operation are performed on x_1 to obtain K. Next, Q and K are element-wise multiplied. The resulting matrix undergoes a Sigmoid transformation to generate attention weights, allowing the model to automatically select more important features, selectively emphasizing critical features while ignoring irrelevant background or noise. This is espe-

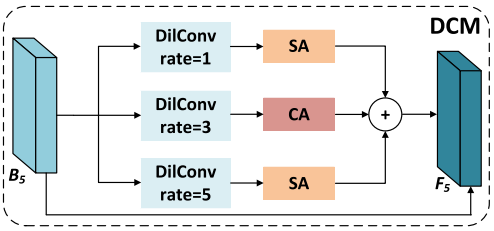


Fig. 4 Detailed illustration of the proposed dilated context module (DCM)

cially important for detecting steel surface defects, where the surface may contain a large amount of irrelevant texture or noise, and the model needs to distinguish the salient features of the defects.

Afterward, global average pooling is applied to the attention weights, producing feature weights of size $C \times 1 \times 1$, which helps to extract comprehensive contextual information and global attention weights. Finally, K is multiplied by the global attention weights to obtain the output tensor, thereby enhancing the attention-weighted region.

Overall, the SCM module improves the model's sensitivity to steel surface defects by integrating convolution, feature mapping, attention mechanisms, and global information, enhancing the model's ability to capture various defect characteristics.

The detailed calculation of this module can be expressed as follows in Eq. (2):

$$\begin{cases} Q = Transpose\left(ReLU\left(Conv_1(x_1)\right)\right), \\ K = ReLU\left(Conv_1(x_2)\right), \\ S_{out} = K \odot GAP\left(\sigma\left(Q \odot K\right)\right). \end{cases} \tag{2}$$

Here, x_1 and x_2 denote the two input features of the SCM module, respectively. Transpose represents the transpose of an input, and $\odot$ indicates the element-wise multiplication operation. Moreover, GAP denotes the global average pooling operation, and σ denotes the Sigmoid function.

Dilated context module

To address the challenge of simultaneously capturing local details and global contextual information in defect detection, this paper introduces the Dilated Context Module (DCM), as shown in Fig. 4. Unlike existing feature fusion methods, DCM offers a novel solution by combining dilated convolutions, multiple attention mechanisms, and feature integration.

First, the input feature B_5 undergoes dilated convolutions with three different dilation rates (1, 3, and 5), producing feature maps with dimensions of $C/3 \times H \times W$. This design significantly expands the receptive field, allowing the model to capture both fine local features and broader contextual information. The multi-scale convolution operation is particularly suited for steel surface defect detection, where defects exhibit highly diverse and irregular shapes and distributions.

Next, spatial attention (Woo et al., 2018) and channel attention (Hu et al., 2018) mechanisms are incorporated into each branch to further refine feature selection. Spatial attention dynamically focuses on regions likely to contain defects, enhancing sensitivity to defect areas, while channel attention adjusts weights based on the importance of specific feature channels, enabling the model to better identify relevant defect characteristics. The combination of these two attention mechanisms is more effective than using a single mechanism, particularly when dealing with complex and varied defect patterns.

Subsequently, the features processed by each attention mechanism are integrated through element-wise addition, resulting in a more semantically rich feature representation. This fusion not only preserves multi-scale contextual information but also optimizes it through the attention mechanisms. Finally, the integrated features are added element-wise to the original input B_5, generating the final output F_5 with dimensions of $C \times H \times W$. This residual connection strategy enhances feature representation while retaining the original information.

Compared to existing methods, the DCM module provides an innovative solution to key challenges in defect detection, such as texture complexity and irregular defect distribution. By combining dilated convolutions with multiple attention mechanisms, DCM enables precise feature capture and representation of complex defects, without incurring additional computational cost. This process is represented in Eq. (3):

$$\begin{aligned} F_5 = {} & SA(DilConv_1(B_5)) \oplus CA(DilConv_3(B_5)) \\ & \oplus SA(DilConv_5(B_5)) \\ & \oplus B_5, \end{aligned} \tag{3}$$

where B_5 denotes the features output from processing C_5 to the BIM, with SA and CA representing spatial attention and channel attention, respectively.

Loss function

We utilize the RPN for region proposals and the RCNN header for defect classification and regression tasks. Both have loss functions consisting of classification and regression. The bounding box regression loss function in the RPN uses $Smooth_{L1}$ loss, calculated as follows:

$$Smooth_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{4}$$

where x represents the discrepancy between the predicted value and the true value.

Table 1 Comparison of NEU-DET and GC10-DET dataset

Dataset	Scale	Type number	Defect types
NEU-DET	1800	6	Cr, In, Pc, Ps, Rs, Sc
GC10-DET	3570	10	Pu, Wl, Cs, Ws, Os, Ss, In, Rp, Cr, Wf

In addition, we use cross-entropy loss as the classification loss function of the RPN. The final RPN loss is expressed as follows:

$$L_{rpn} = \frac{1}{N_{\text{cls}}} \sum_i CE(p_i, p_i^*) + \alpha \frac{1}{N_{\text{reg}}} \sum_i p_i^* Smooth_{L1}(t_i, t_i^*). \quad (5)$$

Here, p_i and p_i^* represent the predicted probability of the i-th anchor being a foreground object and the true label of the i-th anchor, respectively. t_i and t_i^* denote the predicted bounding box parameters for the i-th anchor and the true bounding box parameters corresponding to the i-th anchor, respectively. In addition, α indicates a balancing factor set to 1.

In the NEU-DET dataset utilized in this paper, there exist dense defects such as scratches, patches, among others, which may lead to significant category imbalance issues. To address this concern, we employ Focal Loss as an alternative. Focal Loss allows for adjusting the emphasis on different categories, thereby enabling the model to concentrate more on certain categories. Consequently, for the loss function of the detection head, we utilize the focal loss function for categorization loss, which is calculated as follows:

$$FL(p_t) = -(1 - p_t)^{\gamma} \log(p_t). \quad (6)$$

Here, the term $(1 - p_t)$ acts as a modulating factor to alleviate class imbalance, in which p_t represents the probability that the model predicts a sample as a positive instance.

The final RCNN loss is written as follows:

$$L_{rcnn} = \sum_{i=1}^{N} \big(\quad FL(p_i, p_i^*) + \beta \cdot Smooth_{L1}(t_i, t_i^*) \big), \quad (7)$$

where β is used to judge the importance of balancing classification losses and regression losses. The model's emphasis on the classification and regression tasks can be adjusted by adjusting the value of β, which is usually set to 1.

Experiments setup

Datasets

In this section, we train our network using the NEU-DET (He et al., 2019) dataset. The NEU-DET dataset contains six types of defects: crazing (Cr), inclusions (In), patches (Pc), pitted surfaces (Ps), rolled-in scale (Rs), and scratches (Sc). Each defect type consists of 300 images, all sized at 200×200 pixels. To further verify the generalization of our network, we also employ the GC10-DET (Lv et al., 2020) dataset. GC10-DET has more types of defects (i.e., 10 types), namely punching (Pu), weld line (Wl), crescent gap (Cs), water spot (Ws), oil spot (Os), silk spot (Ss), inclusion (In), rolled pit (Rp), crease (Cr), and waist folding (Wf). It is worth noting that some images may contain multiple types of defects. In our experiments, we randomly split the dataset into training and validation sets with a ratio of 9:1. Table 1 shows the comparison between the NEU-DET and GC10-DET datasets. The details of the defects are as follows:

Implementation details

In this study, we have implemented our model using PyTorch and accelerated the training process using an NVIDIA RTX3070Ti GPU. The parameters of the backbone network are initialized with pre-training weights from ResNet-50 with FPN. Initially, we employed the SGD optimizer; however, to improve convergence and address performance, we transitioned to the Adam optimizer. The initial learning rate is set to 2.5×10^{-5}, with a batch size of 2, and we applied a learning rate decay of 0.05 to dynamically adjust the learning rate during training. The model was trained for 30 epochs to achieve stable convergence and effective optimization of network parameters. To enhance the robustness of the training process and mitigate overfitting, we augmented the dataset using techniques such as horizontal flipping and random flipping.

Evaluation metrics

For object detection tasks, it is conmen to employ precision (P), recall (R), average precision (AP), and mAP as metrics to evaluate detection performance. Each metric is defined as follows. Recall measures the ratio of correctly detected true positives to all actual true positives,

$$Recall = \frac{TP}{TP + FN}. \quad (8)$$

where TP represents the count of instances where the model accurately predicts positive cases as positive, while FN

Table 2 Comparison of various detection models in NEU-DET

Methods	Cr	In	Pc	Ps	Rs	Sc	mAP
Baseline	0.432	0.789	0.885	0.795	0.707	0.893	0.750
SSD	0.382	0.716	0.903	0.754	0.504	0.900	0.693
YoloV3	0.260	0.572	0.815	0.743	0.382	0.703	0.579
YoloX(s)	0.372	0.832	0.924	0.926	0.664	0.918	0.773
YoloX(l)	0.286	0.758	0.928	0.752	0.404	0.872	0.667
YoloX(x)	0.336	0.798	0.918	0.824	0.650	0.852	0.740
Faster R-CNN	0.465	0.795	0.890	0.834	0.652	0.902	0.756
CABF-FCOS	0.554	0.750	0.935	0.889	0.629	0.844	0.767
CANet	0.470	0.836	0.967	0.771	0.597	0.882	0.754
DIN	0.614	0.856	0.930	0.903	0.646	0.883	0.805
AMFNet (SGD)	0.565	0.819	0.952	0.872	0.726	0.933	0.811
AMFNet (Adam)	0.564	0.791	0.942	0.867	0.810	0.943	0.819

denotes the count of instances where the model incorrectly predicts positive cases as negative.

Precision measures the ratio of correctly identified true positives to the total number of examples predicted as positive,

$$Precision = \frac{TP}{TP + FP}. \tag{9}$$

Here, FP represents the count of instances where the model incorrectly predicts negative cases as positive.

AP measures the average of the model's precision at various levels of recall and is typically used to evaluate performance on a single category. In contrast, mAP is the average of the AP values for all categories and is used to evaluate the performance of the entire model on a multi-category task. In this paper, the threshold for AP is set to 0.5, and the calculation for AP and mAP are shown below:

$$\begin{cases} AP &= \int_0^1 P(R)\, dR \\ mAP &= \frac{1}{N} \sum_{i=1}^{N} AP_i, \end{cases} \tag{10}$$

where N denotes the total number of categories, and AP_i represents the first AP value for the i-th category.

Comparative experiment

Detection performance

Table 2 summarizes the detection performance of the proposed AMFNet compared to state-of-the-art methods (i.e., SSD, YOLOv3, variants of YOLOX, CANet (Hou et al., 2023), CABF-FCOS (Yu et al., 2021) and DIN (Hao et al., 2021)) on the NEU-DET dataset, a widely used benchmark for steel surface defect detection. In this paper, we choose the mAP at an IoU threshold of 0.5, a standard metric for object detection tasks.

In general, we can find that AMFNet demonstrates superior overall performance, achieving an mAP of 0.819 with the Adam optimizer, outperforming all compared methods, including SSD (0.693), YOLOv3 (0.579), YOLOX variants (0.667−0.773), Faster R-CNN (0.756), CABF-FCOS (0.767), CANet (0.754), and DIN (0.805). This improvement can be attributed to AMFNet's innovative architecture, which integrates the Branch Interaction Module (BIM), Spatial Correlation Module (SCM), and Dilated Context Module (DCM). These components collectively enhance multi-scale feature extraction, contextual information fusion, and attention-driven defect localization, addressing the challenges of diverse defect sizes, shapes, and distributions on steel surfaces.

Category-specific performance

We can find that AMFNet excels in several defect categories as follows.

- Crazing (Cr): AMFNet (Adam) achieves an AP of 0.564, closely approaching DIN's 0.614 (Hao et al., 2021), the highest among competitors. The BIM's use of dilated convolutions with varying rates enables effective capture of fine, elongated crack features, which are often missed by traditional convolutional networks.
- Scratches (Sc): With an AP of 0.943, AMFNet (Adam) significantly outperforms YOLOX(s) (0.918) and other methods. This can be credited to the BIM's multi-scale feature integration and the DCM's ability to balance local detail preservation with an expanded receptive field, critical for detecting thin, irregular scratches.
- Rolled-in Scale (Rs): AMFNet (Adam) achieves an AP of 0.810, a notable improvement over Faster R-CNN

(0.652) and YOLOX(s) (0.664). The DCM's combination of dilated convolutions and dual attention mechanisms (spatial and channel) enhances sensitivity to defects with varying spatial extents, such as Rs, which often exhibit both localized and distributed patterns.

- Patches (Pc): AMFNet (Adam) records an AP of 0.942, slightly below CANet's 0.967 but surpassing CABF-FCOS (0.935). The SCM's attention-weighted feature fusion contributes to robust contextual understanding, enabling precise detection of patch-like defects amidst complex textures.

While AMFNet does not lead in every category (e.g., Pc), its consistent high performance across all categories underscores its robustness and adaptability to the NEU-DET dataset's diverse defect types.

Impact of optimizer switch: SGD vs. Adam

To investigate the influence of optimization strategies, we trained AMFNet using two optimizers: Stochastic Gradient Descent (SGD) and Adam. The results reveal subtle yet meaningful differences: with regard to the overall mAP, AMFNet (Adam) achieves an mAP of 0.819, slightly higher than AMFNet (SGD)'s 0.811. This 0.008 improvement suggests that Adam provides a marginal but consistent advantage in optimizing the model's weights for defect detection. However, there are category-specific variations with regard to the AP values. For example, AP decreases marginally from 0.565 (SGD) to 0.564 (Adam) in Cr, indicating minimal sensitivity to the optimizer switch for this category; AP increases significantly from 0.726 (SGD) to 0.810 (Adam) in Rs, a 0.084 gain, highlighting Adam's superior adaptability to defects with complex spatial distributions.

The observed differences between SGD and Adam can be explained by their distinct optimization dynamics. As a momentum-based method, SGD updates weights using a fixed learning rate and momentum term, which can lead to stable convergence but may struggle with fine-grained adjustments in deep architectures like AMFNet. Its slightly lower mAP (0.811) suggests that SGD effectively captures general trends but may underfit certain defect-specific features, such as those in Rs (0.726), due to its reliance on a predefined learning schedule. In contrast, by adaptively adjusting learning rates based on first- and second-order moment estimates, Adam excels in navigating the complex loss landscape of AMFNet, which involves multiple modules (BIM, SCM, DCM) and a focal loss function tailored to class imbalance. This adaptability is evident in the improved mAP (0.819) and significant gains in categories like Rs (0.810) and Sc (0.943), where precise weight tuning enhances multi-scale feature representation and attention weighting.

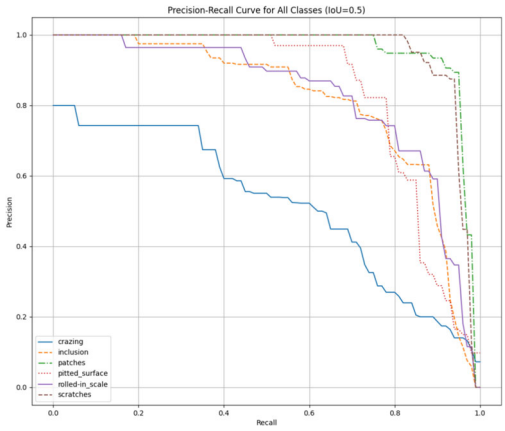


Fig. 5 P–R curves of our AMFNet for each defect category

The NEU-DET dataset's class imbalance, e.g., dense scratches (Sc) versus sparse inclusions (In), further amplifies Adam's advantage. The focal loss, designed to emphasize hard examples, pairs effectively with Adam's adaptive optimization, enabling the model to focus on challenging defects (e.g., Rs) while maintaining performance on easier ones (e.g., Pc). The optimizer switch from SGD to Adam yields a nuanced trade-off: Adam enhances overall performance (mAP: 0.819 vs. 0.811) and excels in categories requiring detailed feature refinement (Rs, Sc), while SGD maintains competitive results with lower computational complexity. The choice of Adam as the preferred optimizer aligns with AMFNet's design goals (i.e., robust multi-scale defect detection). Therefore, we use Adam as the optimizer in the subsequent research of the paper.

Precision-recall curve analysis

Figure 5 illustrates the PR curves for each defect category on the NEU-DET dataset. AMFNet demonstrates superior performance in the Pc (patches) and Sc (scratches) categories, as evidenced by the largest areas under the curve, reflecting high precision and recall. This dominance is attributable to the BIM's multi-scale feature extraction, which effectively captures the distributed patterns of Pc and the fine, elongated structures of Sc. The model also performs robustly in Rs (rolled-in scale), Ps (pits), and In (inclusions), with PR curves indicating competitive precision-recall trade-offs. However, the Cr (crazing) category exhibits relatively lower performance, as indicated by the curve's position in the lower-left region of the plot. Despite this, AMFNet's AP for Cr (0.564 with Adam, Table 2) remains competitive, surpassing most alternatives (e.g., SSD: 0.382, YOLOv3: 0.260) and approaching DIN's leading 0.614. This suggests that while Cr detection remains challenging (likely due to cracks' thin, irregular morphology), AMFNet's design mitigates these difficulties better than most alternatives.

Figure 6 extends this analysis by comparing PR curves across methods at varying confidence thresholds. AMFNet's envelope region encloses the largest area under the curve, outperforming classical methods such as SSD, YOLOv3, and YOLOX variants. This result underscores AMFNet's ability to maintain high precision across a broad range of recall values, a critical advantage in industrial defect detection where both false positives and misses must be minimized. The larger area reflects the synergistic effect of the Spatial Correlation Module (SCM) and Dilated Context Module (DCM), which enhance feature discriminability and contextual awareness, enabling robust detection under diverse operating conditions.

Visual comparison

To complement the quantitative metrics, Fig. 7 provides a qualitative visualization of detection results across methods, with each row representing a defect category and each column corresponding to a model: (a) baseline, (b) SSD, (c) YOLOv3, (d) YOLOX(s), (e) YOLOX(l), (f) YOLOX(x), (g) RetinaNet, and (h) AMFNet. The color-coded bounding boxes highlight detected defects, with confidence scores indicating prediction reliability.

AMFNet (column h) consistently achieves the highest confidence thresholds across all categories, demonstrating its precision in localizing defects of varying sizes and shapes within a single image. For instance, in the Cr category (first row), AMFNet detects all defects with confidence scores exceeding 0.70, whereas competing methods like SSD and YOLOv3 exhibit omissions or lower scores (e.g., <0.50). Similarly, in the In category (second row), AMFNet maintains scores above 0.65, contrasting with the inconsistent detection of baselines. The Sc category (sixth row) further highlights AMFNet's superiority, with precise delineation of elongated scratches, evidenced by tight bounding boxes and scores above 0.87, where other models struggle with fragmented or missed detections.

This visual evidence reinforces AMFNet's ability to mitigate omissions and false positives, a common challenge in steel surface inspection. The BIM's multi-scale feature extraction, bolstered by dilated convolutions, ensures sensitivity to defects across sizes (e.g., fine Sc vs. broad Pc). The SCM's spatial correlation mechanism enhances feature discriminability, reducing false positives in noisy backgrounds, while the DCM's integration of dilated convolutions and dual attention refines feature representations, improving localization precision. These strengths are particularly evident in the high-confidence detections of Cr and In, where competing methods falter, and in the precise boundary delineation of Sc and Rs.

Ablation study

To dissect the contributions of AMFNet's key components, i.e., Branch Interaction Module (BIM), Spatial Correlation Module (SCM), and Dilated Context Module (DCM), we conduct an ablation study on the NEU-DET dataset. This section evaluates their individual and combined impacts on detection performance, computational complexity, and deployment strategies.

Ablation of proposed modules

Table 3 quantifies the effects of integrating BIM, SCM, and DCM into the baseline model, assessing mean mAP at IoU thresholds of 0.5 (mAP50), 0.75 (mAP75), and 0.5:0.95 (mAP50:90), alongside FLOPs and parameter count (millions).

The baseline model yields an mAP50 of 0.744, mAP75 of 0.337, and mAP50:90 of 0.388, with a computational cost of 134.52G FLOPs and 41.38M parameters. The introduction of individual modules results in significant improvements in the model's performance. Specifically, with the inclusion of BIM, the model's accuracy is enhanced through multi-scale feature fusion, raising the mAP50 to 0.786, mAP75 to 0.342, and mAP50:90 to 0.399. Other modules, such as SCM and DCM, also show positive effects on the model's performance, although the improvements are not as pronounced as with BIM. Furthermore, the combination of two modules leads to even higher performance gains, with the best results achieved by the combination of SCM and DCM, which increases the

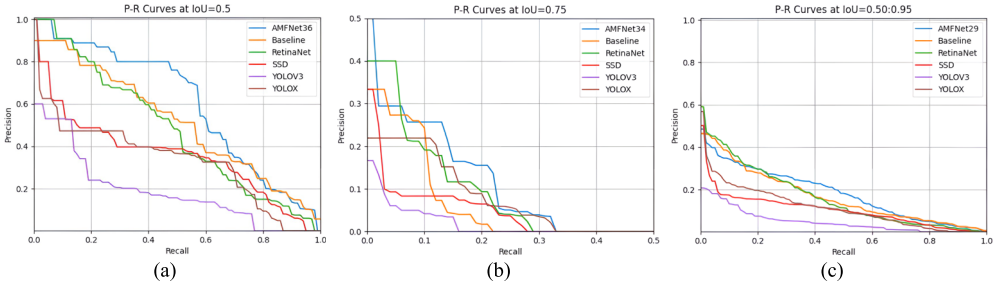


Fig. 6 PR curve comparison with different confidence thresholds on the NEU-DET dataset

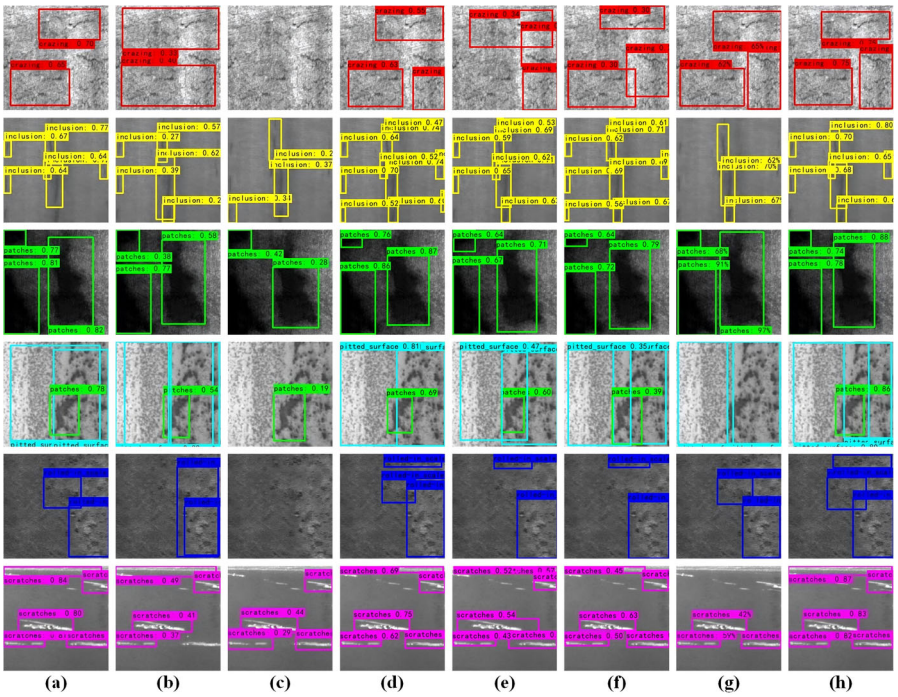


Fig. 7 Visual comparison of our AMFNet test results with state-of-the-art methods on the NEU-DET dataset, with various defect types represented by different colors: **a** baseline, **b** SSD, **c** YoLoV3, **d** YoloX(s), **e** YoloX(l), **f** YoloX(x), **g** RetinaNet, **h** our AMFNet (Color figure online)

Table 3 Ablation Study on the NEU-DET dataset

BIM	SCM	DCM	mAP50	mAP75	mAP50:90	Flops (G)	Params (M)
			0.744	0.337	0.388	134.52	41.38
✓			0.786	0.342	0.399	144.29	42.48
	✓		0.755	0.349	0.384	140.37	44.48
		✓	0.775	0.358	0.393	138.89	43.68
✓	✓		0.802	0.353	0.401	168.68	45.58
✓		✓	0.792	0.367	0.399	158.63	44.78
	✓	✓	0.804	0.371	0.389	175.35	46.78
✓	✓	✓	0.819	0.386	0.422	189.48	47.88

mAP50 and mAP75 to 0.804 and 0.371, respectively. The trade-off is an increase in FLOPS from 134.52G to 175.35G.

The full AMFNet (BIM + SCM + DCM) achieves peak performance: mAP50 of 0.819 (+0.075), mAP75 of 0.386 (+0.049), and mAP50:90 of 0.422 (+0.034). This configuration incurs 189.48G FLOPs and 47.88M parameters, a 41% and 16% increase over the baseline, respectively. The modest computational trade-off relative to the substantial accuracy gains, e.g., a 10% mAP50 improvement, underscores the efficiency of the three-module synergy. Each module contributes uniquely: BIM enhances scale adaptability, SCM improves spatial focus, and DCM refines defect localization, collectively optimizing detection at the standard IoU=0.5 threshold and beyond.

Despite the performance boost, inference speed decreases to ∼15 FPS under our experimental setup (compared to ∼100 FPS for lightweight models like (Li et al., 2024a) with similar hardware). However, steel surface inspection prioritizes accuracy over real-time processing, as defects are typically analyzed offline or in controlled settings. Thus, AMFNet's design prioritizes precision, making the speed-accuracy trade-off justifiable for industrial applications.

Figure 8 visualizes this enhancement via heatmaps from a post-fusion layer, comparing the baseline (first row) and AMFNet (second row) across four steps: original image, heatmap, overlay, and detection result. AMFNet's heatmap (b) exhibits sharper focus on defect regions, with higher activation intensity and clearer boundary delineation than the baseline. The overlay (c) and results (d) show improved localization and confidence scores, attributable to DCM's rich receptive field and attention mechanisms, which enhance sensitivity to defect boundaries and complex textures.

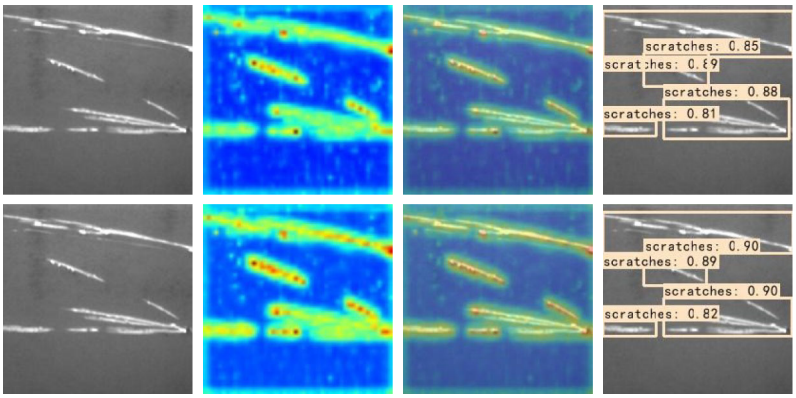


Fig. 8 Heatmaps displayed by specific layers: **a** original image, **b** heatmap, **c** overlay image, **d** result

Table 4 Performance metrics of DCM on the NEU-DET

Models	mAP50	mAP75	mAP50:90	mAP(s)	mAP(m)	mAP(l)
DCM(1)	0.819	0.386	0.422	0.369	0.370	0.564
DCM(2)	0.797	0.362	0.408	0.361	0.359	0.550
DCM(3)	0.788	0.366	0.412	0.357	0.364	0.523
DCM(4)	0.778	0.355	0.405	0.356	0.362	0.517
DCM(1/2/3/4)	0.766	0.331	0.393	0.339	0.355	0.522

Deployment of DCM

Table 4 examines the impact of deploying DCM across different BIM layers, reporting mAP50, mAP75, mAP50:90, and size-specific mAPs (small: mAP(s), medium: mAP(m), large: mAP(l)) on the NEU-DET. The results put particular emphasis on its deployment in the first layer, which demonstrates the best performance. Notably, at IoU = 0.5 (mAP50), it achieves a remarkable accuracy of 0.819, showcasing its efficacy in detecting defects. However, in terms of detecting small defects (mAP(s)), its performance is relatively lower at 0.365. Conversely, it excels in detecting large defects (mAP(l)) and medium-sized defects, achieving 0.556 and 0.366, respectively, underscoring its robust capability in detecting defects of varying sizes when deployed on the first layer of BIM.

With regard to the DCMs deployed at higher layers (DCM2, DCM3, and DCM4), each also exhibits improved performance but slightly lags behind in most evaluation metrics. It should be noted that simultaneous deployment of DCMs across all layers of BIM (DCM1/2/3/4) results in a decrease in all performance metrics, with mAP dropping to 0.393 and mAP50 to 0.766. This suggests that while single-layer deployment can achieve high defect detection performance, simultaneous deployment across multiple layers may lead to performance degradation.

Cross-dataset validation

Generalizability verification

To assess the generalization capability of AMFNet beyond the NEU-DET dataset, we evaluate its performance on the GC10-DET dataset, which encompasses a broader range of steel surface defects with varying morphologies and complexities. Table 5 compares AMFNet against several state-of-the-art methods, including SSD, YOLO variants (YOLOv3, YOLOv5, YOLOv7), Faster R-CNN, EDDN (Lv et al., 2020), DCA_RFCN, EC-YOLO (Cheng et al., 2024), and DCC-CenterNet (Tian & Jia, 2022). In general, AMFNet's mAP of 0.699 exceeds that of Faster R-CNN (0.644), a structurally similar baseline, highlighting the added value of BIM, SCM, and DCM over a standard ResNet50-FPN backbone. Compared to YOLO variants, AMFNet consistently outperforms despite their computational efficiency (e.g., YOLOv7: 0.655), reflecting its superior feature processing for complex defect patterns. Specialized methods like EDDN (0.651) and DCC-CenterNet (0.619) lag behind, likely due to their focus on specific defect types, whereas AMFNet's hierarchical design ensures broader generalization.

Specifically, AMFNet performs well in detecting defect types such as Ws and Os with accuracies of 0.869 and 0.795, respectively, which are the highest scores among the compared models. These defects often exhibit irregular, diffuse shapes with fuzzy boundaries and uneven distributions, posing challenges for traditional detectors. The SCM enhances detection by emphasizing spatial distribution patterns unique

Table 5 Cross-dataset validation on the GC10-DET

Methods	Pu	Wl	Cg	Ws	Os	Ss	In	Rp	Cr	Wf	mAP
SSD	0.852	0.874	0.906	0.612	0.652	0.475	0.268	0.305	0.427	0.592	0.597
Yolov3	0.882	0.896	0.907	0.575	0.754	0.554	0.389	0.354	0.455	0.456	0.619
Yolov5	0.887	0.875	0.931	0.82	0.735	0.523	0.323	0.319	0.355	0.577	0.635
YOLOv7	0.866	0.934	0.952	0.825	0.696	0.489	0.344	0.414	0.402	0.625	0.655
Faster R-CNN	0.895	0.948	0.858	0.825	0.773	0.538	0.455	0.256	0.267	0.627	0.644
EDDN	0.900	0.885	0.848	0.558	0.622	0.65	0.256	0.364	0.521	0.919	0.651
DCA_RFCN	0.893	0.998	0.982	0.705	0.599	0.658	0.264	0.182	0.668	0.67	0.662
EC-YOLO	0.95	0.38	0.96	0.84	0.79	0.62	0.82	0.7	0.38	0.63	0.641
DCC-CenterNet	0.844	0.855	0.962	0.773	0.509	0.848	0.302	0.139	0.499	0.766	0.619
AMFNet	0.899	0.886	0.862	0.869	0.795	0.511	0.515	0.496	0.528	0.633	0.699

to Ws and Os, while suppressing background noise through attention-weighted feature fusion. This capability is critical for distinguishing these defects from complex steel textures. In addition, AMFNet also achieves a high detection accuracy of 0.899 in the Pu category. AMFNet ranks among the top performers (e.g., EDDN: 0.900, EC-YOLO: 0.95), demonstrating balanced adaptability to well-defined, localized defects. The BIM's multi-scale feature extraction ensures robust detection of Pu's distinct geometric characteristics. However, it is most noteworthy that the accuracy of most networks in the three defect categories of In, Rp and Cr is relatively low. These categories are notoriously challenging due to their subtle appearances and low contrast. AMFNet achieves APs of 0.515 (In), 0.496 (Rp), and 0.528 (Cr), surpassing most competitors. For In and Rp, AMFNet trails only EC-YOLO (Cheng et al., 2024) (0.82 and 0.7, respectively), while in Cr, it ranks second to DCA_RFCN (Tian & Jia, 2022) (0.668). The DCM's integration of dilated convolutions and dual attention mechanisms (spatial and channel) enhances sensitivity to these faint, irregular defects by expanding the receptive field and refining feature representations.

The GC10-DET results validate our AMFNet's design principles and also strengthen AMFNet's claim as a robust solution for industrial defect detection, capable of adapting to varied steel surface inspection tasks. Specifically, by leveraging dilated convolutions with varying rates, BIM adapts to the diverse scales of GC10-DET defects (e.g., small Rp vs. diffuse Ws), ensuring comprehensive feature capture. Moreover, The SCM's attention-driven feature fusion enhances discriminability, as seen in Ws and Os, where spatial relationships are critical for accurate detection amidst noise. Lastly, The DCM's combination of dilated convolutions and attention mechanisms enables precise localization of subtle defects (e.g., Cr, In), a key factor in AMFNet's competitive performance across datasets.

5-Fold cross-validation

To mitigate the risk of overfitting and ensure the statistical robustness of AMFNet's performance, we have also conducted 5-fold cross-validation on the GC10-DET dataset. This approach partitions the dataset into five subsets, training and validating the model iteratively to assess consistency across folds. Table 6 reports the AP scores for each defect category and the mAP across all folds for AMFNet and other models.

AMFNet achieves a mean mAP of 0.716 (averaged across folds: 0.719, 0.727, 0.712, 0.693, 0.732), significantly outperforming classical single-stage detectors: SSD (0.586, +0.130), YOLOv3 (0.626, +0.090), YOLOv5 (0.633, +0.083), and YOLOv7 (0.651, +0.065). Compared to Faster R-CNN, a representative two-stage detector, AMFNet improves mAP by 0.075 (0.716 vs. 0.641). These gains highlight AMFNet's robustness against dataset variability, a critical factor in industrial defect detection where overfitting to specific samples can compromise reliability.

With regard to category-specific performance, AMFNet demonstrates superior detection in the Ws and Os categories, with mean APs of 0.844 and 0.801, respectively. These values exceed those of all baselines (e.g., YOLOv7: 0.815 and 0.691 for Ws and Os). The SCM plays a pivotal role here, leveraging spatial attention to emphasize the irregular, diffuse patterns of Ws and Os while suppressing background noise. Additionally, the DCM's expanded receptive field ensures precise localization of these defects, which often exhibit fuzzy boundaries. Across other categories, AMFNet maintains competitive performance without significant degradation. For instance, in Pu, AMFNet's mean AP of 0.925 rivals top performers like YOLOv3 (0.920), while in Cr, its mean AP of 0.563 surpasses most baselines (e.g., SSD: 0.433). This consistency underscores AMFNet's ability to generalize across diverse defect morphologies.

Table 6 Results of 5-fold cross-validation on the GC10-DET dataset

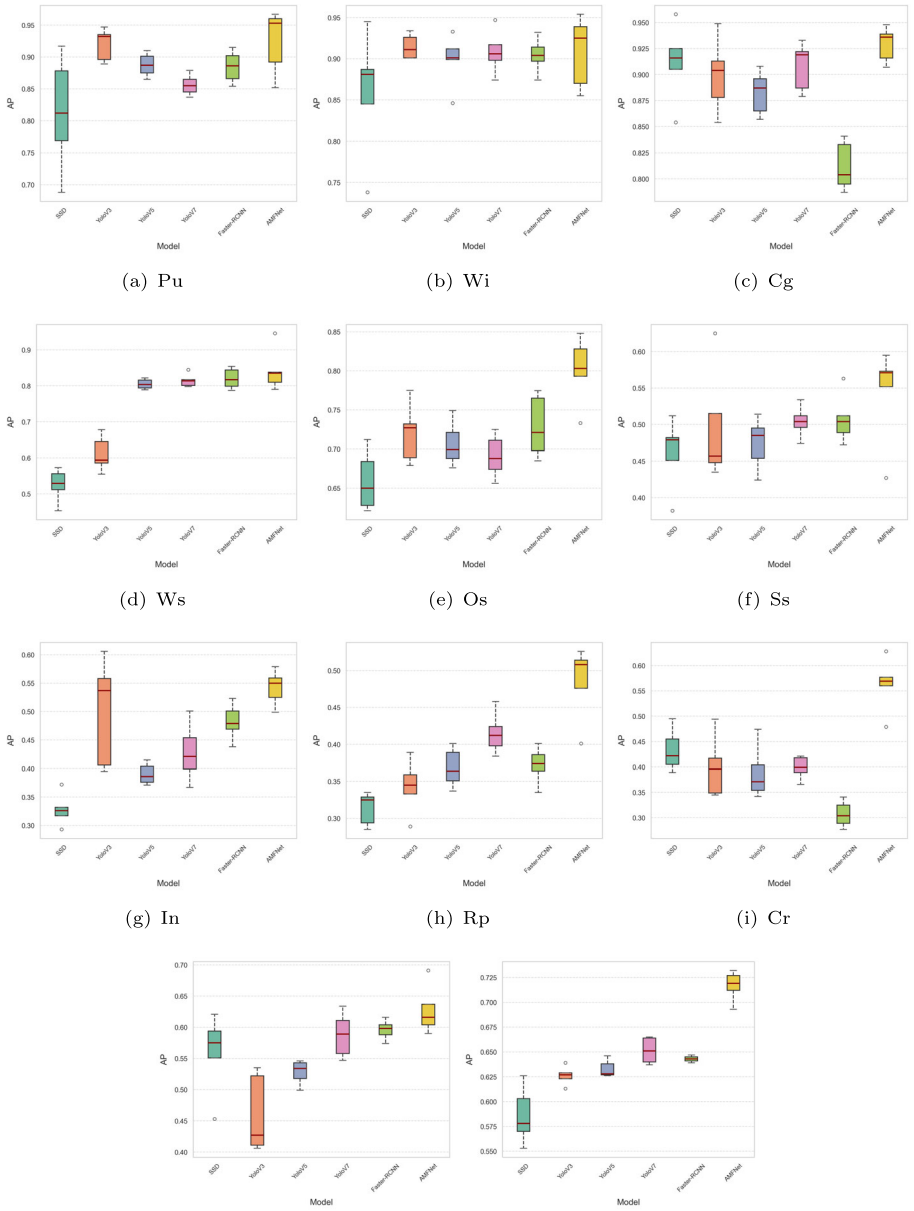
Model	Pu	Wi	Cg	Ws	Os	Ss	In	Rp	Cr	Wf	mAP
SSD	0.812	0.845	0.925	0.453	0.65	0.382	0.293	0.325	0.422	0.594	0.570
	0.688	0.945	0.854	0.512	0.621	0.451	0.326	0.294	0.389	0.453	0.553
	0.769	0.738	0.958	0.529	0.628	0.479	0.317	0.335	0.455	0.575	0.578
	0.917	0.887	0.916	0.573	0.712	0.482	0.332	0.329	0.495	0.621	0.626
	0.878	0.881	0.905	0.556	0.684	0.512	0.372	0.285	0.405	0.551	0.603
YoloV3	0.947	0.926	0.949	0.678	0.689	0.457	0.395	0.345	0.349	0.535	0.627
	0.896	0.901	0.904	0.586	0.732	0.515	0.606	0.333	0.396	0.427	0.629
	0.889	0.934	0.878	0.594	0.727	0.625	0.558	0.289	0.494	0.411	0.639
	0.932	0.911	0.854	0.645	0.679	0.435	0.537	0.389	0.345	0.406	0.613
	0.935	0.901	0.913	0.555	0.775	0.448	0.406	0.359	0.417	0.522	0.623
YoloV5	0.901	0.846	0.908	0.789	0.699	0.495	0.376	0.351	0.354	0.546	0.627
	0.865	0.901	0.887	0.816	0.749	0.424	0.415	0.337	0.371	0.518	0.628
	0.887	0.912	0.896	0.794	0.688	0.454	0.386	0.401	0.342	0.499	0.626
	0.910	0.899	0.865	0.822	0.721	0.514	0.404	0.389	0.404	0.534	0.646
	0.875	0.933	0.857	0.804	0.676	0.485	0.371	0.364	0.474	0.543	0.638
YoloV7	0.855	0.917	0.919	0.845	0.688	0.504	0.421	0.458	0.418	0.611	0.664
	0.879	0.906	0.922	0.814	0.656	0.534	0.399	0.424	0.389	0.589	0.651
	0.845	0.947	0.879	0.798	0.711	0.512	0.501	0.398	0.421	0.634	0.665
	0.865	0.898	0.887	0.801	0.674	0.496	0.454	0.412	0.366	0.547	0.640
	0.837	0.874	0.933	0.817	0.725	0.474	0.367	0.384	0.399	0.558	0.637
Faster-RCNN	0.915	0.904	0.833	0.854	0.685	0.472	0.523	0.386	0.325	0.574	0.647
	0.902	0.932	0.795	0.844	0.7745	0.512	0.501	0.335	0.289	0.598	0.645
	0.866	0.914	0.841	0.817	0.698	0.489	0.479	0.364	0.341	0.616	0.643
	0.854	0.897	0.787	0.799	0.765	0.563	0.438	0.401	0.277	0.604	0.639
	0.886	0.874	0.804	0.787	0.721	0.504	0.469	0.374	0.304	0.588	0.641
AMFNet	0.892	0.855	0.936	0.838	0.828	0.571	0.525	0.514	0.628	0.604	0.719
	0.967	0.939	0.948	0.790	0.733	0.552	0.559	0.526	0.569	0.691	0.727
	0.960	0.925	0.907	0.810	0.848	0.595	0.499	0.508	0.479	0.590	0.712
	0.852	0.954	0.916	0.835	0.793	0.427	0.579	0.401	0.560	0.616	0.693
	0.953	0.87	0.939	0.946	0.803	0.573	0.550	0.476	0.577	0.637	0.732

To visualize performance stability, Fig. 9 presents box plots derived from the 5-fold results in Table 6. AMFNet's mAP box plot exhibits the highest median (∼0.72) and the narrowest interquartile range (IQR) among all models, indicating both superior accuracy and low variance across folds (range: 0.693−0.732). In contrast, SSD's wider IQR (0.553−0.626) and lower median (∼0.58) reflect greater instability. For Ws and Os, AMFNet's boxes show the highest medians (∼0.84 and ∼0.80) and overall heights, confirming its dominance and consistency in these categories. The compact IQR for Ws (0.790−0.946) and Os (0.733−0.848) further validates AMFNet's robustness against data splits, a testament to the BIM's multi-scale feature extraction stabilizing performance across diverse samples.

The 5-fold cross-validation confirms AMFNet's statistical reliability, with an mAP of 0.716 closely aligning with its single-run performance (0.699 in Table 5), suggesting minimal overfitting. The model's stability (narrow IQR) and category-specific strengths (Ws, Os) reinforce its suitability for industrial deployment, where consistent performance across unseen data is paramount.

Conclusion and outlook

In this paper, we propose an aggregated multilevel feature interaction fusion network (AMFNet) for detecting defects on industrial steel surfaces. Unlike previous approaches, our method emphasizes the fusion of features from different paths to achieve effective multi-scale integration. To this end, we introduce the branch interaction module (BIM), which fuses features from multiple paths, fully integrating



(a) Pu (b) Wi (c) Cg

(d) Ws (e) Os (f) Ss

(g) In (h) Rp (i) Cr

(j) Wf (k) mAP

Fig. 9 Box plot performance comparison of different categories and models

shallow fine-grained details with deep semantic information and enhancing both feature extraction and the model's capacity to detect defects at multiple scales. Additionally, we propose a dilated context module (DCM) to expand the receptive field by inflating convolutions, allowing each output to encompass a broader range of information, thus supporting more effective multi-scale fusion. To further improve channel and contextual feature extraction, we introduce the spatial correlation module (SCM), which strengthens the network's ability to recognize and interpret steel surface defects by extracting and emphasizing key regions. Experimental results on the NEU-DET and GC10-DET datasets demonstrate the superior detection performance of AMFNet, showing that these enhancements in feature extraction and defect detection come with only a minimal increase in parameters.

Although the method proposed in this paper improves the performance of steel surface defect detection, it still relies on supervised approaches, which make the manual annotation of datasets both necessary and costly. In future work, we plan to explore semi-supervised and unsupervised methods to address the challenges associated with manual dataset labeling. For instance, we aim to investigate self-supervised learning techniques for feature learning using unlabeled data, as well as the application of Generative Adversarial Networks (GANs) or diffusion models to generate synthetic defect data

for training. Furthermore, the current method has limitations in terms of computational cost and inference speed, which makes it unsuitable for environments with high-speed requirements. In the future, we will explore the application of multi-scale feature fusion techniques for industrial anomaly detection from a more efficient perspective.

Acknowledgements This work is supported in part by National Natural Science Foundation of China [Grant Number 52371275]. We thank all the anonymous reviewers who generously contributed their time and efforts. Their professional recommendations have greatly enhanced the quality of the manuscript.

Data Availability Data will be made available on request.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aghdam, S. R., Amid, E., & Imani, M. F. (2012). A fast method of steel surface defect detection using decision trees applied to lbp based features. In *IEEE conference on industrial electronics and applications (ICIEA)* (pp. 1447–1452). IEEE. https://doi.org/10.1109/ICIEA.2012.6360951.

Asha, V., Bhajantri, N. U., & Nagabhushan, P. N. (2011). Glcm-based chi-square histogram distance for automatic detection of defects on patterned textures. *International Journal of Computer Vision and Robotics, 2*(4), 302–313. https://doi.org/10.1504/IJCVR.2011.045267

Cai, Z., & Vasconcelos, N. (2019). Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(5), 1483–1498. https://doi.org/10.1109/TPAMI.2019.2956516

Carion, N., Massa, F., Synnaeve, G., et al. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229). Springer. https://doi.org/10.1007/978-3-030-58452-8_13.

Cheng, X., & Yu, J. (2020). Retinanet with difference channel attention and adaptively spatial feature fusion for steel surface defect detection. *IEEE Transactions on Instrumentation and Measurement, 70*, 2503911. https://doi.org/10.1109/TIM.2020.3040485

Cheng, Z., Gao, L., Wang, Y., et al. (2024). Ec-yolo: Effectual detection model for steel strip surface defects based on yolo-v5. *IEEE Access, 12*, 62765–62778. https://doi.org/10.1109/ACCESS.2024.3391353

Cui, L., Jiang, X., Xu, M., et al. (2021). Sddnet: A fast and accurate network for surface defect detection. *IEEE Transactions on Instrumentation and Measurement, 70*, 2505713. https://doi.org/10.1109/TIM.2021.3056744

Damacharla, P., Rao, A., Ringenberg, J., et al. (2021). Tlu-net: A deep learning approach for automatic steel surface defect detection. In *International conference on applied artificial intelligence (ICAPAI)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICAPAI49758.2021.9462060.

Fekri-Ershad, S., & Tajeripour, F. (2012). A robust approach for surface defect detection based on one dimensional local binary patterns. *Indian Journal of Science and Technology, 5*(8), 3197–3203. https://doi.org/10.17485/ijst/2012/v5i8.12

Gao, Y., Gao, L., & Li, X. (2023). A hierarchical training-convolutional neural network with feature alignment for steel surface defect recognition. *Robotics and Computer-Integrated Manufacturing, 81*, 102507. https://doi.org/10.1016/j.rcim.2022.102507

Girshick, R., Donahue, J., Darrell, T., et al. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587). https://doi.org/10.1109/CVPR.2014.81.

Halfawy, M. R., & Hengmeechai, J. (2014). Automated defect detection in sewer closed circuit television images using histograms of oriented gradients and support vector machine. *Automation in Construction, 38*, 1–13. https://doi.org/10.1016/j.autcon.2013.10.012

Hao, R., Lu, B., Cheng, Y., et al. (2021). A steel surface defect inspection approach towards smart industrial monitoring. *Journal of Intelligent Manufacturing, 32*, 1833–1843. https://doi.org/10.1007/s10845-020-01670-2

He, K., Gkioxari, G., Dollar, P., et al. (2020). Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(2), 386–397. https://doi.org/10.1109/tpami.2018.2844175

He, Y., Song, K., Meng, Q., et al. (2019). An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. *IEEE Transactions on Instrumentation and Measurement, 69*(4), 1493–1504. https://doi.org/10.1109/TIM.2019.2915404

Hou, X., Liu, M., Zhang, S., et al. (2023). Canet: Contextual information and spatial attention based network for detecting small defects in manufacturing industry. *Pattern Recognition, 140*, 109558. https://doi.org/10.1016/j.patcog.2023.109558

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141). https://doi.org/10.1109/CVPR.2018.00745.

Jain, S., Seth, G., Paruthi, A., et al. (2022). Synthetic data augmentation for surface defect detection and classification using deep learning. *Journal of Intelligent Manufacturing, 33*, 1007–1020. https://doi.org/10.1007/s10845-020-01710-x

Li, C., Xu, A., Zhang, Q., et al. (2024). Steel surface defect detection method based on improved yolox. *IEEE Access, 12*, 37643–37652. https://doi.org/10.1109/ACCESS.2024.3374869

Li, Z., Wei, X., Hassaballah, M., et al. (2024). A deep learning model for steel surface defect detection. *Complex & Intelligent Systems, 10*(1), 885–897. https://doi.org/10.1007/s40747-023-01180-7

Lin, T.Y., Dollár, P., Girshick, R., et al. (2017a). Feature pyramid networks for object detection. In *IEEE conference on computer vision and pattern recognition* (pp. 2117–2125). https://doi.org/10.1109/CVPR.2017.106.

Lin, T. Y., Goyal, P., Girshick, R., et al. (2017b). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988). https://doi.org/10.1109/ICCV.2017.324.

Liu, W., Anguelov, D., Erhan, D., et al. (2016). Ssd: Single shot multibox detector. In *14th European conference on computer vision* (pp. 21–37). Springer. https://doi.org/10.1007/978-3-319-46448-0_2.

Luo, L., Xue, D., & Feng, X. (2019). Hybridnetseg: A compact hybrid network for retinal vessel segmentation. arXiv preprint arXiv:1911.09982 https://doi.org/10.48550/arXiv.1911.09982.

Lv, X., Duan, F., Jj, Jiang, et al. (2020). Deep metallic surface defect detection: The new benchmark and detection network. *Sensors, 20*(6), 1562. https://doi.org/10.3390/s20061562

Ma, J., Wang, Y., Shi, C., et al. (2018). Fast surface defect detection using improved gabor filters. In: *IEEE international conference on image processing (ICIP)* (pp. 1508–1512). IEEE. https://doi.org/10.1109/ICIP.2018.8451351.

Nand, G. K., & Noopur, Neogi N. (2014). Defect detection of steel surface using entropy segmentation. In *Annual IEEE India conference* (pp. 1–6). IEEE. https://doi.org/10.1109/INDICON.2014.7030439.

Peng, Y., Xia, F., Zhang, C., et al. (2024). Deformation feature extraction and double attention feature pyramid network for bearing surface defects detection. *IEEE Transactions on Industrial Informatics, 20*(6), 9048–9058. https://doi.org/10.1109/TII.2024.3370330

Redmon, J., Divvala, S., Girshick, R., et al. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788). https://doi.org/10.48550/arXiv.1506.02640

Ren, F., Fei, J., Li, H., et al. (2024). Steel surface defect detection using improved deep learning algorithm: Eca-simsppf-siou-yolov55. *IEEE Access, 12*, 32545–32553. https://doi.org/10.1109/ACCESS.2024.3371584

Ren, S., He, K., Girshick, R., et al. (2016). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Song, C., Chen, J., Lu, Z., et al. (2023). Steel surface defect detection via deformable convolution and background suppression. *IEEE Transactions on Instrumentation and Measurement, 72*, 1–9. https://doi.org/10.1109/TIM.2023.3277989

Song, G., Song, K., & Yan, Y. (2020). Edrnet: Encoder-decoder residual network for salient object detection of strip steel surface defects. *IEEE Transactions on Instrumentation and Measurement, 69*(12), 9709–9719. https://doi.org/10.1109/TIM.2020.3002277

Song, H., Liu, Z., Du, H., et al. (2016). Depth-aware saliency detection using discriminative saliency fusion. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1626–1630). IEEE. https://doi.org/10.1109/ICASSP.2016.7471952.

Tan, M., Pang, R., & Le, Q.V. (2020). Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10781–10790). https://doi.org/10.1109/CVPR42600.2020.01079.

Tian, R., & Jia, M. (2022). Dcc-centernet: A rapid detection method for steel surface defects. *Measurement, 187*, 110211. https://doi.org/10.1016/j.measurement.2021.110211

Usamentiaga, R., Lema, D. G., Pedrayes, O. D., et al. (2022). Automated surface defect detection in metals: A comparative review of object detection and semantic segmentation using deep learning. *IEEE Transactions on Industry Applications, 58*(3), 4203–4213. https://doi.org/10.1109/TIA.2022.3151560

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 6000–6010). https://doi.org/10.48550/arXiv.1706.03762.

Woo, S., Park, J., Lee, J. Y., et al. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3–19). https://doi.org/10.1007/978-3-030-01234-2_1.

Xi, D., Qin, Y., & Wang, S. (2023). Ydrsnet: An integrated yolov5-deeplabv3+ real-time segmentation network for gear pitting measurement. *Journal of Intelligent Manufacturing, 34*(4), 1585–1599. https://doi.org/10.1007/s10845-021-01876-y

Yang, J., & Yang, M. H. (2016). Top-down visual saliency via joint crf and dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(3), 576–588. https://doi.org/10.1109/CVPR.2012.6247940

Yeung, C. C., & Lam, K. M. (2022). Efficient fused-attention model for steel surface defect detection. *IEEE Transactions on Instrumentation and Measurement, 71*, 2510011. https://doi.org/10.1109/TIM.2022.3176239

Yu, J., Cheng, X., & Li, Q. (2021). Surface defect detection of steel strips based on anchor-free network with channel attention and bidirectional feature fusion. *IEEE Transactions on Instrumentation and Measurement, 71*, 1–10. https://doi.org/10.1109/TIM.2021.3136183

Zhang, Y., Wang, W., Li, Z., et al. (2023). Development of a cross-scale weighted feature fusion network for hot-rolled steel surface defect detection. *Engineering Applications of Artificial Intelligence, 117*, 105628. https://doi.org/10.1016/j.engappai.2022.105628

Zhao, B., Chen, Y., Jia, X., et al. (2024). Steel surface defect detection algorithm in complex background scenarios. *Measurement, 237*, 115189. https://doi.org/10.1016/j.measurement.2024.115189

Zhou, X., Fang, H., Liu, Z., et al. (2021). Dense attention-guided cascaded network for salient object detection of strip steel surface defects. *IEEE Transactions on Instrumentation and Measurement, 71*, 1–14. https://doi.org/10.1109/TIM.2021.3132082

Zhou, X., Liu, Z., Sun, G., et al. (2016). Improving saliency detection via multiple kernel boosting and adaptive fusion. *IEEE Signal Processing Letters, 23*(4), 517–521. https://doi.org/10.1109/LSP.2016.2536743