Structural Durability & Health Monitoring

Tech Science Press

https://doi.org/10.32604/sdhm.2026.079896

ARTICLE

Research on Steel Surface Defect Detection Algorithm Based on YOLOv11-ODF

Zhengxiang Ma[1,2,*], **Xiaofei Ma**[1], **Xiaoliang Liu**[3], **Heng Zhang**[4] **and Weichao Yu**[1]

[1]School of Electronics and Information, Zhengzhou University of Aeronautics, Zhengzhou, China
[2]National and Local Joint Engineering Research Center for Intelligent Building Internet of Things Technology and Application, Zhengzhou, China
[3]Engineering Technology Research Center, Tianzhu Science And Technology Co., Ltd., Zhengzhou, China
[4]Army Artillery and Air Defense Academy, 24 Jianshe East Road, Erqi District, Zhengzhou, China

*Corresponding Author: Zhengxiang Ma. Email: mzx@zua.edu.cn

Received: 30 January 2026; Accepted: 24 March 2026; Published: 30 June 2026

ABSTRACT: Steel surface defect detection is a key technology for ensuring the quality of steel products and the automation of production. However, in actual industrial scenarios, the complex texture background of steel surfaces often leads to low recognition of tiny defect features and easy confusion, and the full extraction and fusion of multi-scale features remain challenging. To address these issues, this paper proposes a lightweight and high-precision detection model based on the improved YOLOv11n, named YOLOv11-ODF. Firstly, in the backbone network, a C3k2_ODConv module integrating full-dimensional convolution (ODConv) is constructed, which enhances the model's ability to capture subtle defect features through multi-dimensional dynamic weights, and combines the C2PSA attention mechanism to optimize the feature representation in both spatial and channel dimensions. Secondly, in the feature fusion network (Head), an OD_WT_Fuse module is designed to replace the traditional fusion method, effectively improving the efficiency of cross-scale information transmission and semantic consistency. In addition, an anisotropic strip spatial pyramid pooling (ASSPPF) module is designed to further expand the receptive field and enhance the robustness of detecting irregular multi-scale defects. Experimental results show that on the NEU-DET dataset, the mAP@0.5 of YOLOv11-ODF reaches 77.1%, significantly improving by 3.2% compared to the original YOLOv11 model; the precision and recall increase by 1.6% and 4.6%, respectively, significantly reducing the missed detection rate of tiny defects. While achieving significant performance improvements, the model parameters only increase by 0.9 M, achieving an excellent balance between detection accuracy and computational efficiency, providing an effective technical solution for high-quality real-time automatic detection in industrial environments.

KEYWORDS: Steel surface defect detection; YOLOv11n; full-dimensional convolution; feature fusion; NEU-DET dataset

1 Introduction

Steel is an indispensable key material in the process of national development and modernization, and its quality directly affects the overall level of infrastructure construction and manufacturing [1]. However, during the production and processing of steel, defects such as cracks, dents, and spots may occur due to factors such as equipment and process flow. These defects may affect the mechanical properties such as strength and toughness of the steel, as well as its service life, and also pose significant safety hazards to its application fields, especially in industries requiring high-precision devices such as aerospace, electronics, and medical care [2]. Therefore, the detection of surface defects in steel is of great significance.

Traditional steel defect detection mainly relies on manual visual inspection or non-destructive testing based on physical sensors [3]. However, in high-intensity production environments, such methods not only have high labor costs and limited detection throughput, but are also easily affected by external factors such as light and vibration, leading to high rates of missed and false detections, and are gradually unable to meet the strict requirements of modern steel industry for high efficiency and digitalization [4].

In recent years, breakthroughs in deep learning technology have provided a new path for the automatic recognition of surface defects in steel. In the field of object detection based on convolutional neural networks (CNN), research routes mainly fall into two-stage methods represented by R-CNN [5] series, such as Mask R-CNN [6], and one-stage methods represented by SSD [7] and YOLO [8] series. For instance, classic two-stage architectures like Faster R-CNN [9] introduced the Region Proposal Network (RPN), which established a strong theoretical advantage in positioning accuracy. However, despite this historical milestone, their complex topological structure and high computational cost often lead to inference delays that struggle to meet the ultra-strict real-time requirements of modern high-speed production lines. In contrast, one-stage detectors integrate defect localization and classification into a single end-to-end regression problem, significantly improving processing speed while maintaining detection performance. Among the many one-stage algorithms, the YOLO (You Only Look Once) series has become the benchmark model in industrial detection due to its excellent detection gain and deployment flexibility. Compared with traditional deep learning algorithms, the YOLO series not only achieves higher parallelism in network structure but also continuously compresses the trade-off space between accuracy and speed through continuous iterations (such as introducing advanced attention mechanisms, multi-scale feature fusion, and decoupled detection heads). Especially in dealing with steel surface tasks with complex backgrounds and highly uneven defect scales, the YOLO series demonstrates stronger global context modeling capabilities and robust real-time inference performance, making it the optimal technical path for achieving high-quality automatic detection of steel surface defects [10–12].

To further improve the efficiency and accuracy of lightweight object detection, researchers have extensively studied various aspects, including network architecture refinement, attention mechanism enhancement, and advanced feature fusion techniques. In the pursuit of architecture simplification and enhanced feature representation, Hou et al. [13] proposed Coordinate Attention (CA) specifically for lightweight mobile networks, which embeds positional information into channel attention to achieve a better balance between detection accuracy and computational cost. Han et al. [14] redesigned the feature extraction backbone using Ghost modules to generate abundant feature maps via cheap operations, effectively streamlining the network for real-time industrial applications. In terms of attention mechanisms and spatial modeling, Wang et al. [15] introduced the Efficient Channel Attention (ECA) module to enhance inter-channel dependencies through local cross-channel interaction without dimensionality reduction; Wu et al. [16] incorporated the parameter-free Simple Attention Module (SimAM) and focal structures into the initial network layers, improving sensitivity to fine-grained defects; Huang and Wang [17] combined the lightweight C2f-Faster block with the Efficient Multi-Scale Attention (EMA) mechanism to enhance detection performance in cluttered environments. In terms of multi-scale feature integration and sampling strategies, Zhu et al. [18] applied deformable convolutions to dynamically adjust spatial sampling locations, significantly enhancing the model's robustness against complex geometric variations of targets; Tan et al. [19] proposed a weighted bi-directional feature pyramid network (BiFPN) to perform easy and fast multi-scale feature fusion, significantly improving the efficiency of cross-scale information flow; Xie et al. [20] developed a lightweight multi-scale feature fusion network based on YOLOv8 to strengthen the extraction and aggregation of steel surface defect features across varying scales; Meng and Wen [21] expanded the receptive field through the Context Anchor Attention (CAA) mechanism and Content-Aware ReAssembly

of FEatures (CARAFE) operator, significantly improving the representation of edge details. Despite these advancements, challenges remain in industrial environments characterized by complex background textures. Traditional convolution and downsampling operations often lead to the loss of edge information and semantic degradation of small defects. Moreover, current feature fusion and pooling methods struggle to provide sufficient representational adaptability and comprehensive global context modeling when dealing with defects with significant scale variations and extreme aspect ratios.

To overcome these limitations, this study proposes YOLOv11 ODF—a lightweight, high-precision detection framework based on the enhanced YOLOv11n architecture. Firstly, the model integrates the ODConv [22] into the backbone network, significantly improving the extraction of fine defect features under complex conditions through a four-dimensional dynamic weight adjustment mechanism. Secondly, in the feature aggregation process, a novel OD_WT_Fuse structure is proposed, which replaces traditional concatenation with dynamic weighted fusion, enabling more effective interaction between deep semantic and shallow spatial features. Finally, to address the frequent occurrence of elongated and irregular defects on steel surfaces, an anisotropic strip spatial pyramid pooling module (ASSPPF) is designed. This module utilizes an adaptive multi-scale strip pooling operator to expand the receptive field, enhancing the detection robustness for different defect geometries while maintaining the model's lightweight nature. The proposed method effectively reduces the omission rate of small and irregular defects in real industrial applications.

2 YOLOv11-ODF Algorithm

YOLOv11, released by the Ultralytics team in 2024, is the latest generation of the YOLO series of models. It inherits the advantages of the single-stage detection algorithm, featuring end-to-end training and real-time inference. Additionally, it optimizes the overall architecture and training strategy, delivering excellent performance in multiple tasks such as detection, segmentation, and tracking. Its network structure comprises a backbone network, a neck feature fusion network, and a decoupled detection head. The backbone incorporates an improved C3k2 module and a C2PSA attention module. The former enhances feature extraction capabilities for complex backgrounds and small targets through larger convolutional kernels and adjustable structures, while the latter integrates channel and spatial attention to boost the response in key areas. Simultaneously, an SPPF layer is introduced at the end to achieve multi-scale feature aggregation. The neck combines FPN and PANet structures for multi-level feature fusion, and the detection head adopts an anchor-free and decoupled design to enhance detection speed and accuracy. Overall, YOLOv11 maintains a good balance between speed and accuracy while preserving lightweight and efficiency, laying a solid foundation for subsequent structural optimization.

To address the detection accuracy limitations in steel production due to complex background textures, large defect scale variations, and the difficulty in capturing irregular defects, this paper proposes a full-dimensional dynamic fusion improved algorithm, YOLOv11-ODF, based on the lightweight model YOLOv11. The improved network structure is shown in Fig. 1. This model, while fully inheriting the efficient residual feature extraction capability of the C3k2 module and the global attention modeling advantages of the C2PSA mechanism in both spatial and channel dimensions, achieves a leap in performance through three core technological innovations. Firstly, it introduces full-dimensional dynamic convolution (ODConv) in the backbone network and neck structure to deeply reconstruct C3k2, forming an enhanced C3k2_ODConv module. This module significantly improves the model's ability to parse extremely fine defect features through multi-dimensional adaptive modulation in the spatial, channel, and input-output dimensions, effectively alleviating the problem of feature loss under complex background interference. Secondly, to solve the semantic consistency problem in multi-scale feature fusion, this paper designs the OD_WT_Fuse fusion structure, which innovatively couples dynamic convolution with wavelet transform convolution

(WTConv) [23] in a residual manner. By leveraging the unique advantages of wavelet transform in spatial-frequency domain analysis, it achieves efficient complementarity and information flow among multi-level features, enhancing the model's perception of edge details. Additionally, for the common strip-shaped or long-proportion irregular defects on steel surfaces, this paper introduces the anisotropic strip spatial pyramid pooling (ASSPPF) module to reconstruct the original SPPF structure. By using an adaptive long-strip pooling operator, it broadens the model's effective receptive field and strengthens the ability to aggregate global context information for non-regular shape defects. Experimental results show that YOLOv11-ODF significantly improves the accuracy and robustness of cross-scale defect detection while maintaining extremely high inference efficiency, providing a more robust technical solution for real-time and automatic detection of high-quality steel surface defects in industrial environments.

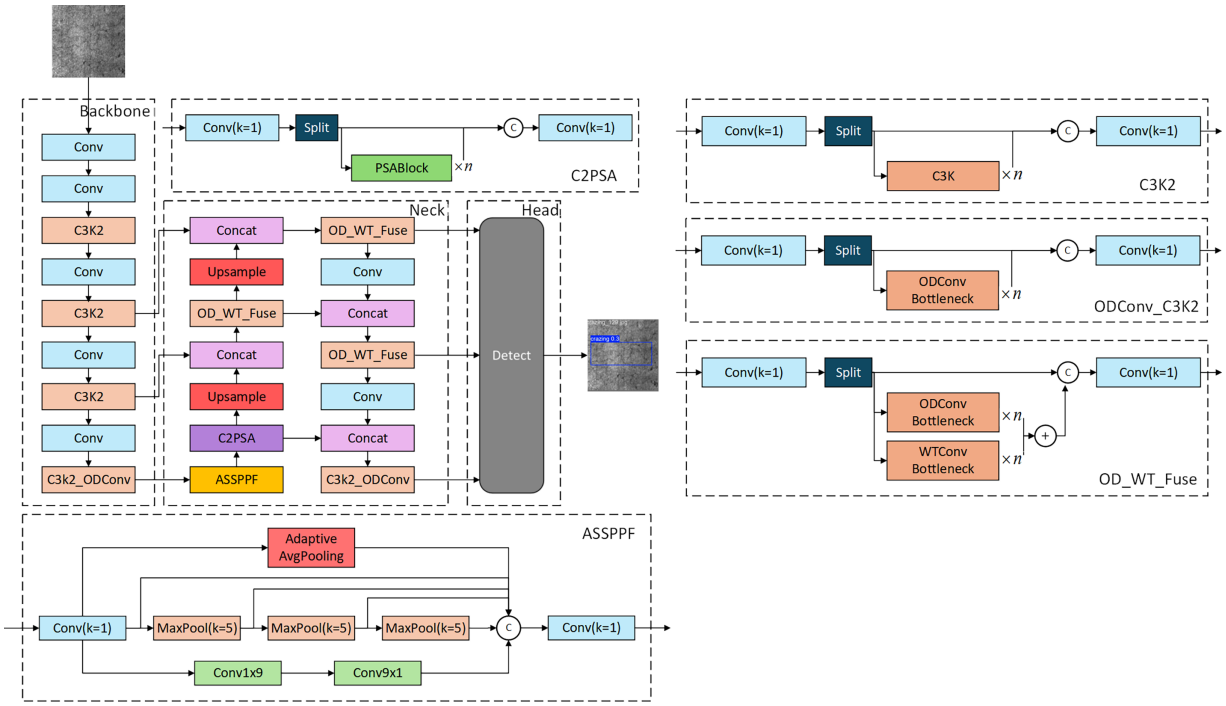

Figure 1: The network architecture of the improved YOLOv11-ODF model.

2.1 ODConv Module

Omni-dimensional Dynamic Convolution (ODConv) is a class of full-dimensional dynamic convolution structures proposed by Li et al., aiming to enhance the adaptive modeling capability of convolution operators for complex scenes without significantly increasing computational complexity. Traditional convolution kernels have fixed weights, and their convolution operation can be formalized as follows:

$$Y = W * X + b \tag{1}$$

where $X \in R^{C_{in} \times H \times W}$ represents the input feature map, $W \in R^{C_{out} \times C_{in} \times k \times k}$ denotes the static convolution kernel, and b signifies the bias term. Since the convolution kernel cannot dynamically adjust with the input, it is often challenging to simultaneously consider local texture and global semantic structure when dealing with steel surface defects that exhibit significant scale differences, complex textures, or low contrast, thus limiting the model's expressive power.

To overcome the aforementioned limitations, ODConv introduces a Multi-dimensional Attention Subnetwork before the convolution operator. This subnetwork generates a dynamic weight set across four dimensions: spatial, channel, convolution kernel, and grouping, through Global Average Pooling (GAP), linear mapping, and Sigmoid activation:

$$A = \{a_8, a_c, a_k, a_9\} \tag{2}$$

These weights are applied to the corresponding dimensions of the convolution kernel, respectively, achieving dynamic re-weighting on a per-dimension basis, thereby constructing a dynamic convolution kernel that adapts to the current input features:

$$\mathrm{W}' = a_8 \odot a_c \odot a_k \odot a_9 \odot \mathrm{W}\ \mathrm{Y} = \mathrm{W}' * \mathrm{X} \tag{3}$$

Here, $\odot$ denotes the element-wise multiplication operation, W' represents the dynamically restructured convolution kernel, and X denotes the input features. Through this mechanism, ODConv is capable of simultaneously modeling feature dependencies across multiple dimensions, enabling dynamic adjustment of the convolution kernel among positions, channels, and filter groups, thereby significantly enhancing the flexibility of feature representation. The differences in dynamic dimensions between traditional dynamic convolution (left) and ODConv (right) are illustrated in Fig. 2:

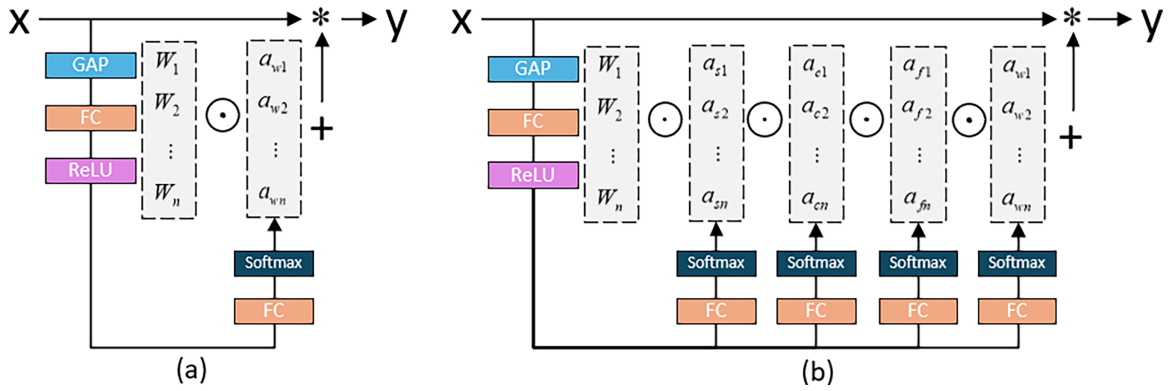

Figure 2: The difference in dynamic dimension between traditional dynamic convolution (**a**) and ODConv (**b**).

To ensure efficiency, ODConv adopts a Parallel Attention Strategy, which simultaneously generates attention weights across four dimensions and performs collaborative modulation. This process can be viewed as a hierarchical dynamic enhancement of convolutional kernels, enabling the network to balance fine-grained texture and high-level semantic information. Fig. 3 illustrates the effective modes of attention mechanisms for four different channels in ODConv, including three stages: attention weight generation, convolutional kernel reconstruction, and dynamic convolution computation.

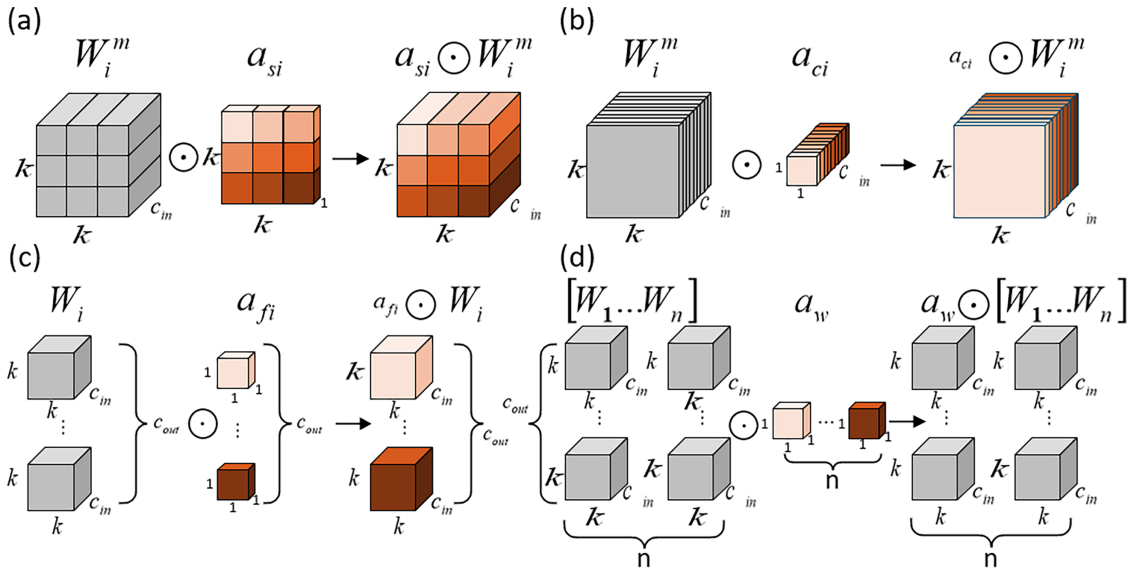

Figure 3: The effective ways of attention mechanisms for four different channels in ODConv.

In the improved YOLOv11 model proposed in this article, based on the balance between feature abstraction and computational complexity, we strategically embed ODConv into the deep semantic stage of Backbone (P5/32) and the multi-scale fusion structure of the detection head (C3k2.ODConv), achieving collaborative optimization at both semantic enhancement and feature interaction levels. Choose a deep architecture instead of a full network replacement, as deep networks have larger receptive fields and higher feature dimensions. ODConv's full dimensional dynamic modulation mechanism can more effectively capture complex global contextual information at this stage and dynamically reconstruct weak features with low contrast; At the same time, introducing this module into the detection of key nodes in the head can significantly improve the dynamic adaptive ability of multi-scale feature fusion, avoiding excessive computational redundancy in the shallow texture extraction stage. The experimental results show that this differentiated layout scheme can significantly improve the accuracy, recall, and generalization performance of the model in NEU-DET steel surface defect detection tasks, especially in the recognition of fine-grained and difficult to identify defects.

2.2 WTConv Module

To enhance the long-range dependency modeling capability of YOLOv11 in steel surface defect detection tasks, this paper introduces the Wavelet Convolution (WTConv) structure proposed by Finder et al. (2024). WTConv utilizes the multi-scale frequency domain decomposition characteristics of two-dimensional Wavelet Transform (WT) to decompose spatial domain convolution into frequency domain convolution and inverse wavelet reconstruction, significantly expanding the effective receptive field of the network while maintaining lightweight.

The receptive field of a traditional convolutional layer expands linearly with the size of the convolution kernel, while the number of parameters grows quadratically with the kernel size. When the convolution kernel is large, the model is prone to issues such as parameter redundancy and performance saturation. For the input feature map X, WTConv first performs a discrete wavelet transform, decomposing it into four frequency bands:

$$[X_{LL}, X_{LH}, X_{HL}, X_{HH}] = WT(X) \tag{4}$$

Among them, the low-frequency sub-band X_{LL} primarily encompasses the overall structure of the target, whereas the high-frequency sub-bands X_{LH}, X_{HL}, X_{HH} retain local details such as texture variations and edges. Compared to traditional convolution, which performs local operations directly in the spatial domain, wavelet domain decomposition enables independent feature modeling at different frequency scales, facilitating the simultaneous capture of contour information and fine-grained texture features of steel defects.

In the wavelet domain, the convolution operation takes the following form:

$$Y = IWT(Conv(W, WT(X))) \tag{5}$$

Due to the downsampling of the spatial resolution of low-frequency components, performing a 3 × 3 small convolution on them can correspond to a larger equivalent receptive field in the input domain, thereby achieving long-range dependency modeling with extremely low computational cost.

Fig. 4 illustrates an example of WTConv performing 3 × 3 convolution on the second-level low-frequency subband $X_{LL}^{(2)}$, which is equivalent to acting on a 12 × 12 receptive field in the input domain, while requiring only 9 parameters.

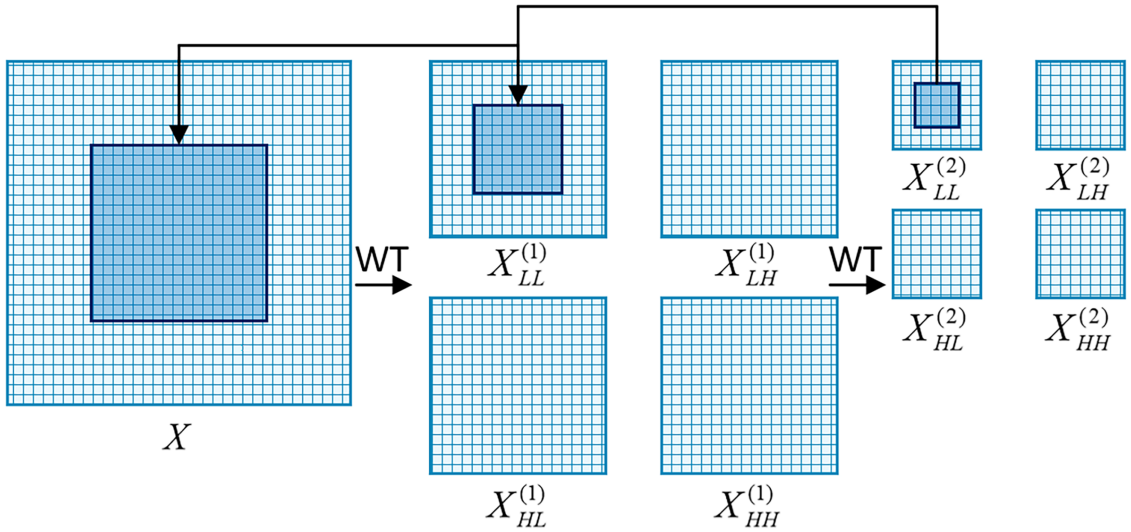

Figure 4: The diagram illustrates the expansion of the receptive field when WTConv performs convolution in the multi-level wavelet domain.

This indicates that WTConv can achieve long-range modeling capabilities that are difficult for traditional convolutions to attain, with extremely low computational costs. More importantly, the multi-frequency expression method in the wavelet domain makes it highly suitable for scenarios in steel defect detection that require both overall shape perception and rely on local texture contrast.

To further demonstrate the execution process of WTConv in real-world networks, Fig. 5 illustrates an example of single-channel convolution in MobileNetV2, encompassing steps such as wavelet decomposition, sub-band convolution, and inverse reconstruction. This provides an intuitive reference for understanding its internal mechanism.

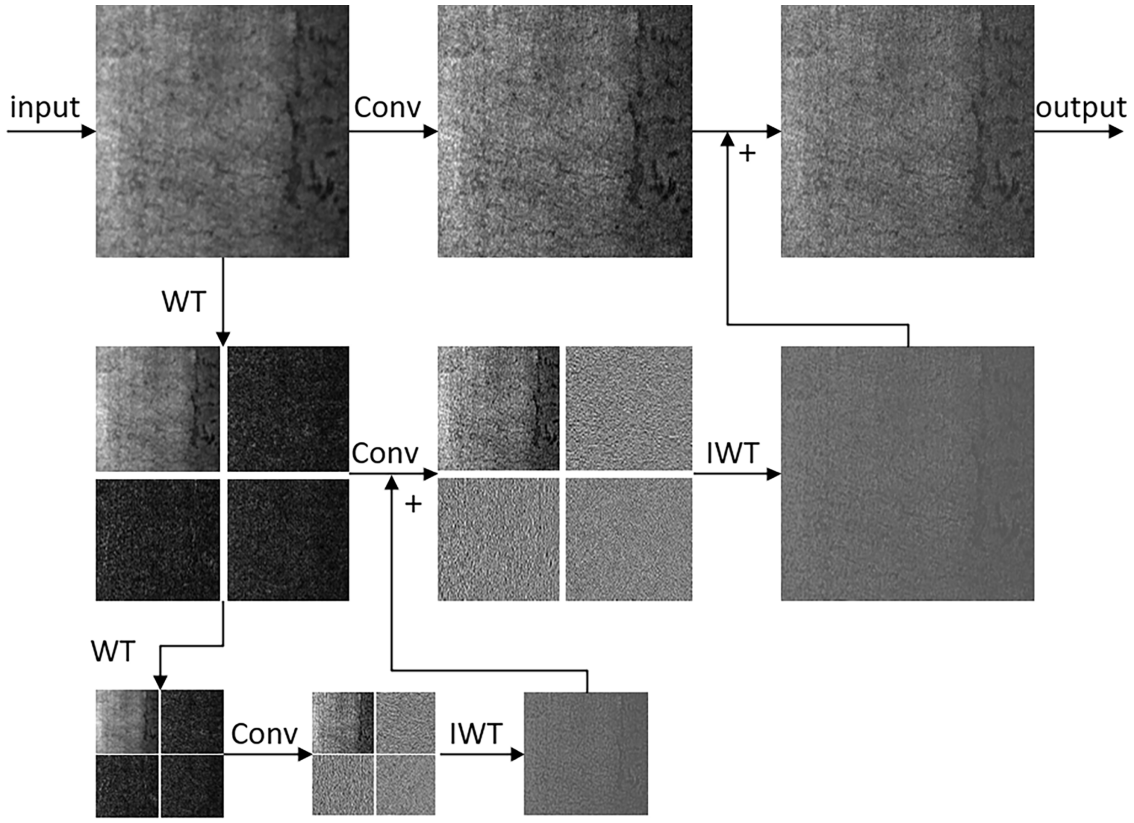

Figure 5: Actual execution example of WTConv in mobile network (decomposition, convolution, reconstruction process).

Compared to traditional convolution, WTConv possesses the following characteristics: (1) Equivalent Large Receptive Field: Multi-level wavelet decomposition provides approximate global context modeling capability while maintaining lightweight. (2) Multi-Frequency Feature Representation: Low-frequency structures and high-frequency details are modeled independently, suitable for multi-type steel defect detection. (3) Shape Bias: Helps reduce false responses caused by noisy background textures. (4) Lower Parameters and Computation: Parameters increase logarithmically with the number of wavelet layers, significantly better than the quadratic growth pattern of large convolution kernels.

In the improved YOLOv11 architecture presented in this paper, WTConv is embedded as a key auxiliary component of the OD_WT_Fuse module into the multi-scale feature fusion stage of the Neck network. Unlike traditional feature extraction methods, WTConv primarily performs frequency domain decomposition and multi-scale reconstruction of cross-layer features through wavelet transform, providing OD_WT_Fuse with auxiliary representations rich in high and low-frequency details. This design enables the model to obtain feature inputs with greater global consistency and frequency differentiation during feature fusion, thereby assisting OD_WT_Fuse in achieving higher-precision cross-scale semantic alignment and background noise suppression.

2.3 Omni-Dimensional Wavelet Fusion Strategy

To further enhance the joint representation ability of the model in the spatial and frequency domains, this paper proposes a novel residual fusion architecture—the OD_WT_Fuse module. This module aims to

construct a more discriminative feature expression paradigm by synergizing the anisotropic spatial modeling capability of the full-dimensional dynamic convolution (ODConv) and the multi-resolution frequency analysis characteristics of the wavelet convolution (WTConv). In the spatial dimension, ODConv achieves adaptive parameter updates in the convolution kernel, channel, and spatial dimensions through a multi-dimensional dynamic weight modulation mechanism, significantly enhancing the model's sensitivity to capturing the details of irregular defects in complex backgrounds. Complementarily, WTConv utilizes the wavelet transform to spectrally decompose the input signal, enabling the parallel extraction of multi-scale low-frequency and high-frequency components, allowing the model to simultaneously consider the global topological structure and local texture evolution features. However, due to the fact that the representation logics of the two focus on dynamic spatial mapping and multi-frequency domain decomposition, respectively, simple linear stacking often fails to effectively bridge the semantic gap and may even introduce feature redundancy. Therefore, OD_WT_Fuse adopts a residual-driven fusion design, by establishing a cross-domain collaborative mechanism, enabling the dynamic adaptability in the spatial domain and the decomposition characteristics in the frequency domain to complement each other within a unified framework, thereby improving the efficiency of information flow while ensuring the integrity and robustness of feature representation [24,25].

The OD_WT_Fuse module consists of four core components: a dual-branch convolutional structure, dynamic gated fusion (Fuse-Gate) [26], a residual balancing mechanism [27], and channel re-calibration [28].

(1) Dual-branch convolutional structure. The input features are first aligned in channels, and then input into the ODConv branch and the WTConv branch, respectively, to extract dynamic convolutional features YOD and multi-frequency convolutional features YWT. This design allows parallel extraction of spatial and frequency domain information, reducing information loss.

(2) Fuse-Gate. To balance the contributions of the two branches, a lightweight gated network is introduced, which generates adaptive weights through global average pooling and two layers of 1 × 1 convolution.

$$(g_{od}, g_{wt}) \tag{6}$$

This weight is used to adjust the importance of the output features of ODConv and WTConv, enabling the module to adaptively select the optimal fusion mode based on the distribution of input features.

(3) Residual balancing mechanism. To enhance the stability of fusion, two learnable parameters α and β are introduced to balance and constrain the fusion results. The core output formula of the module is as follows:

$$Y = (1-\alpha)(g_{od}Y_{OD} + g_{wt}Y_{WT}) + \alpha \cdot \frac{1}{2}(Y_{OD} + Y_{WT}) + \beta X \tag{7}$$

Y_{OD} and Y_{WT} are the outputs of ODConv and WTConv, respectively, X is the input feature, and α, β are learnable weight parameters. This structure enables dynamic weighted fusion and balanced residual paths to work together, which helps to improve gradient stability and feature consistency.

(4) Channel re-calibration. The fused features undergo channel re-weighting through the SE Block to emphasize key channel information and suppress redundant features, thereby enhancing the overall expression quality.

This residual fusion mechanism fully combines the advantages of two types of convolutions: the dynamic nature of ODConv ensures the adaptive ability of feature direction and channel dimension; the multi-frequency decomposition of WTConv enhances the model's ability to capture low-frequency

global information; the fused features achieve complementary enhancement in both spatial and frequency dimensions.

The experimental results show that after embedding the module into the neck network of YOLOv11, it can reduce the parameter count and computational complexity by approximately 10% while maintaining detection accuracy, effectively achieving a balance between model lightweighting and performance. The design of OD_WT_Fuse fully embodies the innovative idea of "spatial domain dynamic modeling + frequency domain multi-scale fusion", providing a new solution for convolutional neural networks in terms of lightweighting and efficient feature representation.

2.4 Anisotropic Striped Spatial Pyramid Pooling Module (ASSPPF)

In the original architecture of YOLOv11, the Spatial Pyramid Pooling-Fast (SPPF) module primarily converts the feature maps output by the Backbone network into fixed-size feature vectors, facilitating the preliminary fusion of multi-scale features. However, in the task of steel surface defect detection, which involves defects with a wide range of scales (such as the coexistence of minor scratches and large patches) and significant industrial background noise interference, relying solely on the local 5 × 5 max pooling SPPF is insufficient to capture global semantic information. Therefore, this paper opts to use the Anisotropic Strip Spatial Pyramid Pooling-Fast (ASSPPF) module as a replacement for the original SPPF [29–32].

The core logic of ASSPPF lies in constructing a dual perception mechanism of "local block-long-range strip". Assuming the input feature map is $X \in R^{C \times H \times W}$, where C, H, and W represent the number of channels, height, and width, respectively. Firstly, the module performs channel compression through a 1 × 1 convolution to obtain relay features X′. The module then branches into two parallel paths:

(1) Cascade Local Extraction Branch: Adopting the serial max-pooling strategy, local scale information is extracted through the cascade of three 5 × 5 pooling layers. The max-pooling operation is defined as MaxPoolk·, and the cascade feature is represented as:

$$y_1 = \mathrm{MaxPool}_5(X') \tag{8}$$

$$y_2 = \mathrm{MaxPool}_5(y_1) \tag{9}$$

$$y_3 = \mathrm{MaxPool}_5(y_2) \tag{10}$$

(2) Anisotropic Strip-Shaped Perception Branch: This represents the core improvement of ASSPPF. To capture scratch information characterized by long-range dependencies, this paper introduces Depth-wise Strip Convolution. This branch comprises a horizontal convolution kernel (1 × 9) and a vertical convolution kernel (9 × 1). The calculation formula for its output feature, ystrip, is as follows:

$$y_{strip} = \mathrm{SiLU}\left(\mathrm{BN}(\mathrm{Conv}_{9\times1}(\mathrm{Conv}_{1\times9}(X')))\right) \tag{11}$$

Among them, SiLU is the activation function of Sigmoid-weighted Linear Unit, and BN stands for Batch Normalization, which is used to accelerate convergence and improve training stability. $\mathrm{Conv}_{9\times1}$ and $\mathrm{Conv}_{1\times9}$ represent horizontal and vertical strip convolutions, respectively. Their function is to capture long-range spatial dependencies in the horizontal and vertical directions without significantly increasing the computational load, thereby enhancing the perception of elongated defects (such as scratches). By extending the receptive field to a cross-shaped strip, this branch can aggregate defective pixels across a larger spatial span, while utilizing grouped convolution (Groups = C') to greatly reduce computational redundancy.

Finally, ASSPPF deeply integrates the original feature $X^{'}$, the three-level local pooling features y_1, y_2, and y_3, as well as the anisotropic strip feature ystrip through a channel concatenation operation. The calculation formula for the output feature Y is:

$$Y = \mathrm{SiLU}(\mathrm{BN}(\mathrm{Conv}_{1\times1}(\mathrm{Concat}(X', y_1, y_2, y_3, y_{strip})))) \quad (12)$$

Among them, Concat $(\cdot)$ denotes feature concatenation in the channel dimension. $\mathrm{Conv}_{1\times1}$ represents pointwise convolution, which is used to perform channel dimensionality reduction and information fusion on the concatenated multi-scale features.

As shown in Fig. 6, ASSPPF effectively compensates for the shortcomings of standard SPPF in handling asymmetric geometric features by introducing a striped convolutional branch. In deep semantic feature maps, this module can accurately capture the structural connectivity of elongated scratches, significantly enhancing the robustness of YOLOv11 in detecting multi-scale and irregular steel defects in complex industrial scenarios.

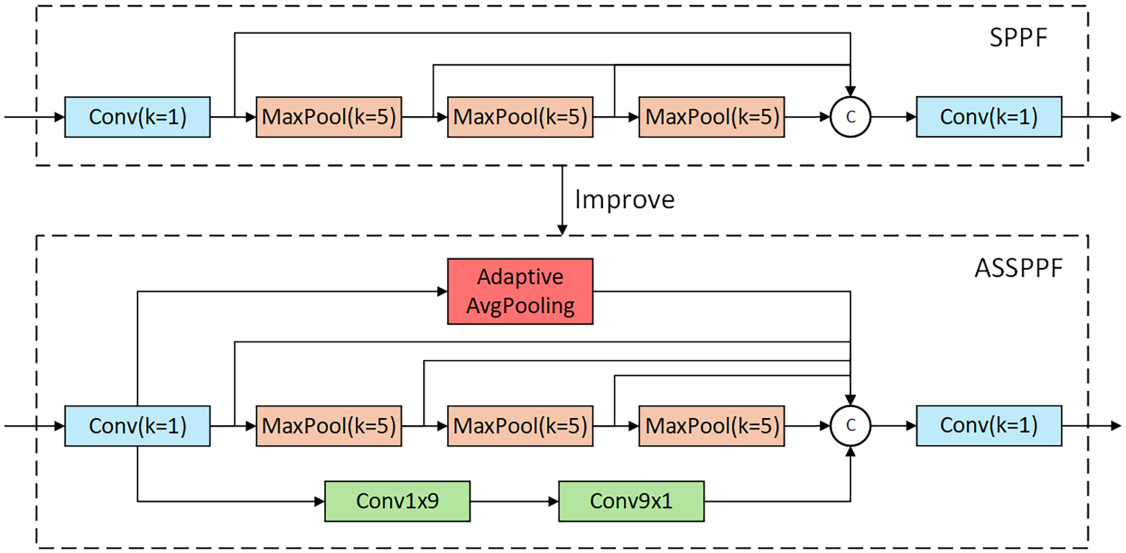

Figure 6: The structure of the improved ASSPPF module.

3 Results Analysis and Discussion

In order to eliminate random fluctuations during the experimental process and ensure the objectivity and robustness of the results, the performance indicators of all models in this paper are the average values obtained after 5 independent repeated training and testing in the same experimental environment. In response to the small size of the NEU-DET dataset, this paper adopts a strategy of taking the average of multiple random experiments: in each experiment, the dataset is shuffled again by changing the random seed, and the training set and validation set are divided according to a fixed ratio. The experiment was conducted continuously for 5 times, and the final result was taken as the arithmetic mean of the 5 independent running indicators. This method effectively validates the robustness and generalization ability of the YOLOv11-ODF algorithm without reducing the training sample size, avoiding the randomness of experimental results.

3.1 Experimental Environment and Parameter Configuration

The experimental platform utilizes an NVIDIA GeForce RTX 3090 GPU, running on the Ubuntu operating system. The deep learning framework employed is Pytorch 2.12, with Python 3.10 and Cuda 11.8. The parameters of the experimental environment are presented in Table 1.

Table 1: Parameters of the experimental environment.

Model	R/%
Epoch	300
Batch size	16
Optimize	AdamW
Image Size	640
Momentum	0.937

3.2 Dataset and Preprocessing

This experiment used the NEU-DET dataset of surface defects on hot-rolled strip steel, publicly released by Professor Song's team at Northeastern University, for model training and evaluation. This dataset contains 1800 grayscale images with a resolution of 200 × 200, covering six typical surface defects commonly found in the production process of hot-rolled steel strips: cracking (Cr), inclusion (In), patches (Pa), pitting surface (Ps), rolled in scale (Rs), and scratches (Sc) [33]. In terms of sample size, each type of defect contains 300 images, totaling 1800 images, demonstrating good category balance. However, due to significant differences in target size, density, and geometric shape among different types of defects, the dataset still presents challenges at the feature distribution level.

To ensure the adequacy of model training and generalization representation ability, the dataset is randomly divided into a training set (1260 images), a validation set (360 images), and a testing set (180 images) in a ratio of 7:2:1. In the preprocessing stage, various data augmentation strategies were applied in this paper to compensate for the insufficient original sample size and alleviate the problem of target scale imbalance: the robustness of the model to geometric distortions was enhanced through random scaling, translation, and flipping; Simulate complex industrial lighting environments using HSV color gamut transformation; And introduce Mosaic and Mixup enhancement techniques to randomly crop and stitch multiple images. The application of these technologies not only enriches the background semantics of the target, but also effectively enhances the model's ability to detect subtle scratches and complex community defects. The visualization effects of various defects are shown in Fig. 7:

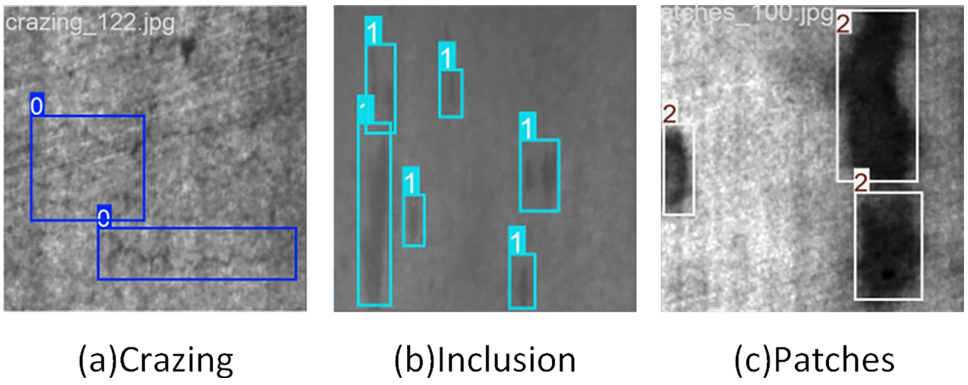

Figure 7: (Continued)

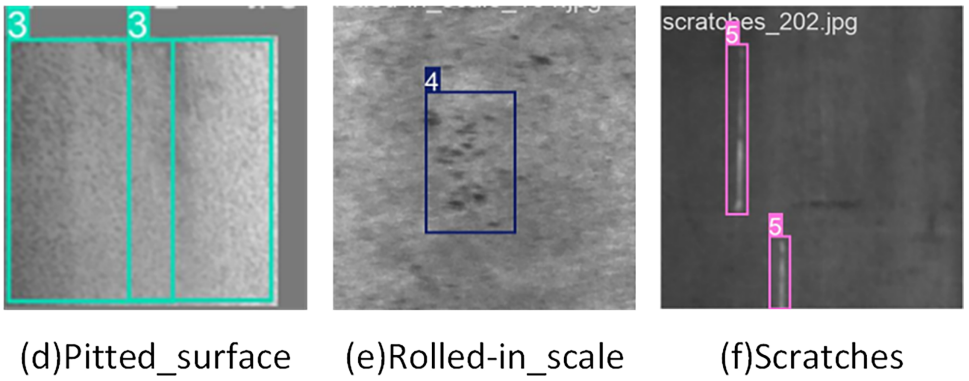

Figure 7: Example of defect images.

3.3 Evaluation Metrics

The experiment utilizes precision P, recall R, average precision AP, mean average precision mAP, and parameter count Params as evaluation metrics. The calculation formulas for each metric are as follows:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{13}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{14}$$

$$AP = \int_0^1 P(R)\, dR \tag{15}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{16}$$

TP is the true case, FP is the false positive case, and FN is the false negative case; AP is the area enclosed by the P-R curve and the coordinate axis, used to measure the model's ability to detect a single category; MAP is the average AP of all categories and is a core indicator for measuring the performance of multi category detection; N is the number of target categories to be tested in the dataset.

3.4 Comparison with Mainstream Object-Detection Algorithms

To verify the superiority of the YOLOv11-ODF algorithm, the improved model was compared with other mainstream algorithms under the same conditions, including YOLOv5n, YOLOv8n, YOLOv9s, YOLOv10n, YOLOv10s, Lgff-YOLO, MHD-YOLO and YOLOv11-EMC [34–37]. The precision, recall, mean average precision, and parameter count of each model were evaluated on the NEU-DET dataset. The comparative experiments are shown in Table 2, with the optimal metrics highlighted in bold.

Table 2: Comparative experimental of different detection networks.

Model	P/%	R/%	mAP@0.5/%	Params/M
SSD	66.9	65.8	65.5	24.5
Faster R-CNN	66.4	62.1	67.8	137.0
YOLOv5n	69.5	65.2	68.3	**1.9**

(Continued)

Table 2 (continued)

Model	P/%	R/%	mAP@0.5/%	Params/M
YOLOv8n	70.2	68.4	71.5	3.2
YOLOv9s	72.1	71.5	74.2	7.2
YOLOv10n	70.1	66.5	72.1	2.7
YOLOv10s	71.0	71.8	74.6	7.2
YOLOv11n	71.6	70.1	73.9	2.58
Lgff-YOLO	72.8	73.1	74.8	3.45
DCD-YOLOv8n	74.1	73.6	75.9	3.12
MHD-YOLO	75.2	74.3	76.4	3.62
MEA-YOLO	74.8	74.5	76.1	2.95
YOLOv11-ODF	73.2	74.7	77.1	3.48

Based on the analysis of the comparative experimental data in Table 2, it can be concluded that the YOLOv11-ODF algorithm proposed in this paper performs the best in the task of steel surface defect detection, achieving an mAP@0.5 of 77.1%, ranking first among all comparative algorithms. When compared with the benchmark model, YOLOv11-ODF demonstrates significant performance gains. Compared to the original YOLOv11n model, the algorithm proposed in this paper improves precision (P), recall (R), and mean average precision (mAP@0.5) by 1.6%, 4.6%, and 3.2%, respectively, while the number of parameters in the algorithm only increases by 0.9 M, still making it a lightweight algorithm model.

As shown in Fig. 8, the improved model outperforms the original YOLOv11n model in detecting most defect categories. Particularly, for the two defect types with complex morphological features, namely Pitted_surface and Rolled_in_scale, the AP values have significantly increased from the baseline of 0.752 and 0.591 to 0.828 and 0.651, respectively. This indicates that by introducing the OD_WT_Fuse module and the ASSPPF mechanism, the model can more effectively capture the minute and highly disruptive background texture features. Additionally, for the extremely difficult-to-detect fine defect Crazing, the improved model has achieved a steady increase in accuracy.

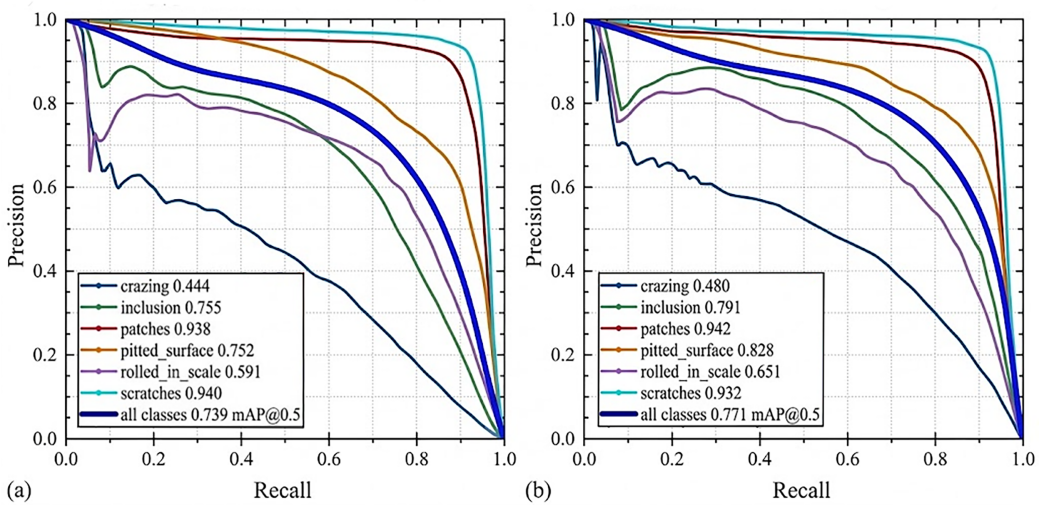

Figure 8: (**a**, **b**) respectively represent the precision-recall (P–R) curves of YOLOv11n (**a**) and YOLOv11-ODF (**b**).

The experimental results show that the ASSPPF module, which aims to enhance the ability of long-range dependency modeling, exhibits slight fluctuations in the detection indicators of scratches category after introducing strip convolution, and its mAP value is adjusted from 0.940 to 0.932. The specific physical mechanism analysis shows that this phenomenon is not due to the failure of the design framework, but reflects the representation trade off between anisotropic feature sampling and defect geometric heterogeneity. Although strip convolution has significant advantages in aggregating axially aligned features, its long path sampling mechanism is prone to introducing axial background noise when dealing with scratches with nonlinear curvature or diagonal distribution features in the NEU-DET dataset. This objectively produces a slight smoothing effect on high-frequency local edge information with clear boundaries, leading to marginal shifts in feature signal-to-noise ratio at extremely fine scales. On the other hand, YOLOv11-ODF significantly enhances the semantic capture ability of the model for irregular and low contrast complex defects (such as iron oxide scales), and exhibits better detection accuracy and generalization robustness than the benchmark model in the vast majority of key defect categories. Even in harsh industrial environments with uneven lighting and drastic changes in defect morphology, the model maintains extremely high performance stability, confirming its practical value in real-time monitoring scenarios on steel assembly lines.

3.5 Confusion Matrix and Class-Wise Performance

To complement the qualitative analysis, we visualized the normalized confusion matrices of the baseline YOLOv11n and the improved YOLOv11-OWT models on the NEU-DET dataset, as shown in Fig. 9.

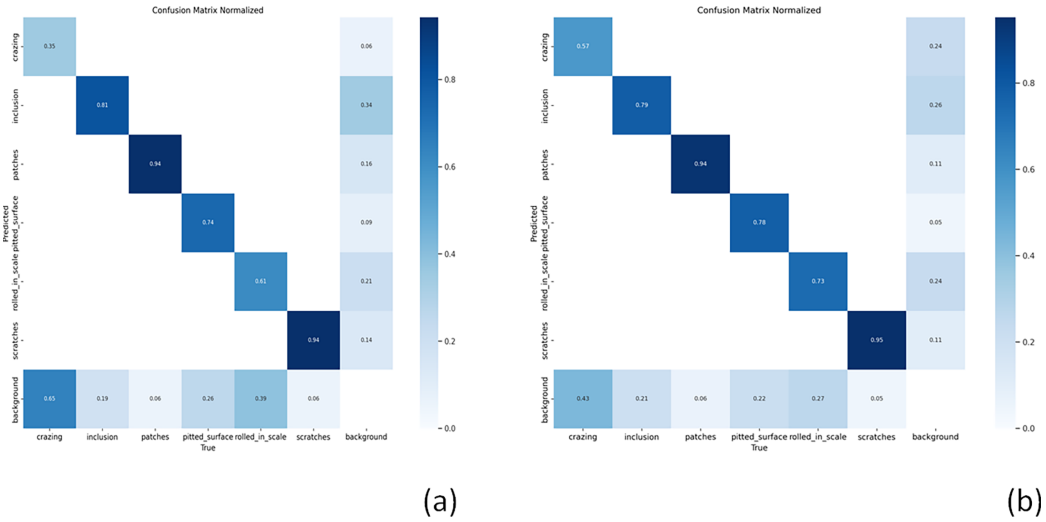

Figure 9: (**a**, **b**) represent the normalized mixing matrices of YOLOv11n (**a**) and the proposed YOLOv11-ODF (**b**).

As illustrated in Fig. 9, a comparison of the normalized confusion matrices between YOLOv11n and YOLOv11-ODF reveals that the proposed model consistently achieves higher diagonal values across most defect categories. This indicates a significant improvement in classification accuracy over the baseline model. The most notable enhancement is observed in the crazing category, where classification accuracy increased substantially from 0.35 to 0.57, reflecting a markedly strengthened ability to capture subtle and diffuse surface texture features. Furthermore, for categories characterized by complex textural patterns, such as pitted_surface and rolled_in_scale, the improved model achieved accuracies of 0.78 and 0.73, respectively, demonstrating more robust feature discrimination. The model also maintains high precision in the patches and scratches categories (at 0.94 and 0.95, respectively). Notably, false positives in the background class have

been significantly reduced—particularly for crazing, where the background misclassification rate dropped from 0.65 to 0.43. these results demonstrate that YOLOv11-ODF not only improves the true positive rate but also effectively enhances the separability between industrial defects and complex backgrounds.

These results confirm that the enhancements integrated into YOLOv11-ODF, particularly the ODF (Omni-Weight Fusion) module, contribute to more precise and robust classification. This is especially evident in defect types that are visually similar or characterized by rich, complex textures.

3.6 Ablation Experiments

This paper uses YOLOv11n as the baseline model and improves upon it, designing a lightweight and efficient YOLOv11-ODF model. To verify the effectiveness of each improved node, this paper conducts four sets of ablation experiments using the NEU-DET dataset. A represents replacing some C3k2 modules with ODConv modules, B represents replacing some C3k2 modules with OD_WT_Fuse modules, and C represents replacing the SPPF module with ASSPPF module.

According to the ablation experiment results shown in Table 3, when introducing the ODConv module alone (Scheme A), the model achieves deep mining of slender scratches and tiny folding features through a four-dimensional dynamic weighting mechanism of spatial, channel, and input-output dimensions, enabling mAP@0.5 Improved from 73.9% to 76.3% and achieved the highest accuracy rate in the entire group (P). This performance gain is attributed to ODConv strengthening the local geometric induction bias of the network during the feature extraction stage, allowing the convolution kernel parameters to adaptively adjust based on the spatial distribution of input content, thereby establishing robust low-level feature representations in complex backgrounds.

Table 3: Ablation experiments of the proposed improvements.

Model	P/%	R/%	mAP@0.5/%	Params/M
YOLOv11n	71.6	70.1	73.9	2.58
YOLOv11n + A	76.3	70.3	76.3	3.09
YOLOv11n + C	71.9	72.2	74.7	2.62
YOLOv11n + A + B	73.9	70.4	75.1	3.45
YOLOv11n + A + C	72.4	72.9	75.5	3.13
YOLOv11n + A + B + C	73.2	74.7	77.1	3.48

Similarly, the independent introduction of the ASSPPF module (Scheme C) yields a highly targeted enhancement, elevating the mAP@0.5 to 74.7% with merely a 0.04 M increase in parameter overhead. Most notably, Scheme C drives the Recall (R) up significantly from 70.1% to 72.2%. From a structural perspective, this emphasizes the efficacy of the anisotropic spatial pyramid pooling mechanism. By adaptively expanding the macroscopic receptive field across different aspect ratios, ASSPPF successfully captures the topological context of variable-scale and highly irregular defects (such as long scratches or scattered patches). This multi-scale contextual aggregation directly reduces the missed detection rate, proving its high parameter-to-performance efficiency.

It was observed in the experiment that when the wavelet transform fusion module (Scheme B) or ASSPPF module (Scheme C) was separately added on the basis of Scheme A, the model performance showed nonlinear regression (75.1% and 75.5%, respectively). From the perspective of Feature Representation Consistency, this fluctuation reflects the mismatch of feature flow distribution caused by single scale enhancement: the high-frequency components introduced by module B are prone to feature aliasing in the

absence of global context constraints, while the macroscopic receptive field provided by module C tends to be feature smooth in the absence of detail support. Both interfere with the distribution fitting of ODConv to the target key area, causing the model to fall into local suboptimal solutions during the optimization process.

When the three collaborate to construct YOLOv11-ODF, the model mAP@0.5 final score was set at 77.1%, an increase of 3.2% compared to the benchmark. This synergistic gain confirms the manifold alignment effect of multidimensional features in the time-frequency and spatial domains: the microscopic high-frequency details provided by module B and the macroscopic topological background captured by module C form a complete feature spectrum, providing a scale consistent input distribution for ODConv. This multi-scale collaborative constraint effectively suppresses the gradient inconsistency introduced by a single module, and achieves high fidelity refining of feature transfer through global and local two-way regularization, thus showing excellent detection performance when dealing with complex and changeable irregular defects on the steel surface.

As illustrated in Fig. 10, the analysis of mAP@0.5 across various defect categories demonstrates that the proposed enhancement modules yield substantial performance gains and synergistic benefits when addressing defects with diverse morphologies. While the baseline YOLOv11n model exhibits balanced but limited performance across categories, the integration of Module A (ODConv) significantly enhances detection accuracy in categories such as Inclusion and Pitted Surface. This validates the superior capability of four-dimensional dynamic convolution in capturing complex local geometric features by adaptively weighting kernel parameters. During further ablation experiments, the incorporation of Module C (ASSPPF) (represented by the blue bars in YOLOv11n + A + C) facilitates a breakthrough optimization for Rolled-in Scale detection. Its accuracy markedly surpasses previous configurations, highlighting the unique advantages of anisotropic receptive fields in processing defects with extreme aspect ratios. Ultimately, the YOLOv11-OWT (Ours) model, which integrates all modules, achieves the highest mAP values across several critical categories, including Patches, Pitted Surface, and Scratches.

This "step-wise" performance escalation provides robust evidence that despite minor feature fluctuations in specific categories (e.g., Crazing) during single-module superposition, the holistic optimization of the YOLOv11-ODF architecture facilitates profound logical complementarity and synergistic effects among the modules. Consequently, the model demonstrates exceptional detection robustness and superior performance across all categories when confronted with morphologically diverse and background-complex industrial steel defects. These findings not only validate the scientific rigor of the proposed scheme in handling extremely irregular steel defects but also underscore its outstanding applicability within complex industrial environments.

4 Conclusions

4.1 Summary of Research Results

Addressing the challenges of strong background interference and difficult capture of subtle features in steel surface defect detection, this paper proposes a lightweight improved model, YOLOv11-ODF, based on the YOLOv11 architecture. This model enhances the perception ability of multi-dimensional spatial features of defects by introducing the Omni-directional Dynamic Convolution (ODConv), and effectively retains key high-frequency details and edge information during downsampling by combining Wavelet Convolution (WTConv). Experimental results show that the improved model achieves a mAP@0.5 of 77.1% on the NEU-DET dataset, an improvement of 3.2% compared to the original model. Meanwhile, the precision and recall rates increase by 1.6% and 4.6%, respectively, significantly reducing the rates of missed and false detections while maintaining the lightweight advantage. This study verifies the effectiveness of dynamic convolution and frequency domain feature fusion in industrial vision tasks.

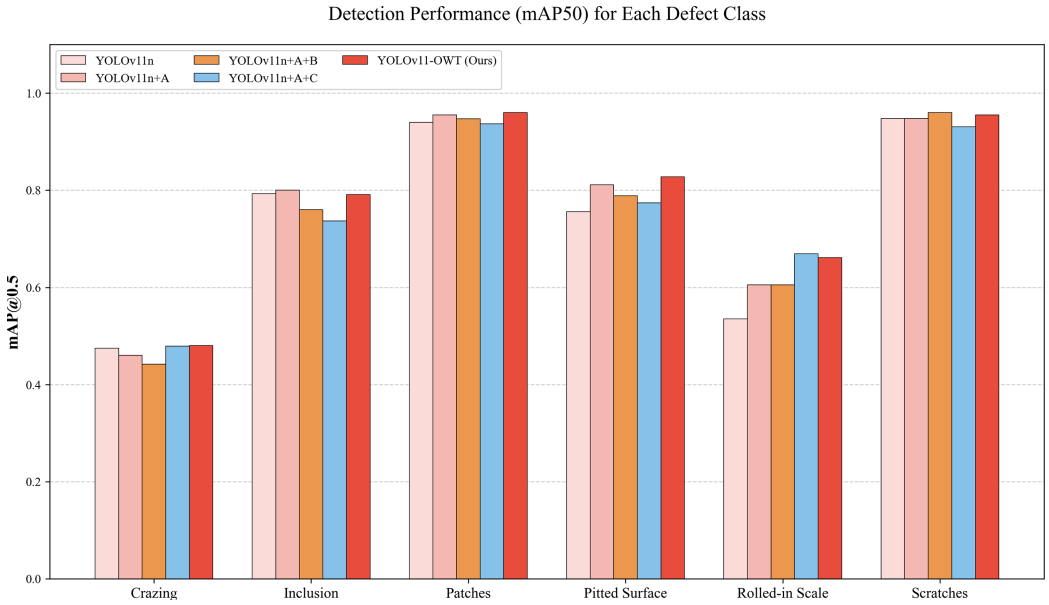

Figure 10: Detection effects of different models for different defect types.

4.2 Limitations and Future Development Directions

The YOLOv11-ODF algorithm proposed in this article performs well on the NEU-DET steel surface defect dataset. Because the model maintains a lightweight parameter count of only 3.48 M and low computational complexity, it demonstrates significant potential for future integration into embedded industrial detection equipment or mobile edge computing platforms. Secondly, the robustness of the algorithm under small sample distributions was verified through multiple random experiments, ensuring its detection stability in practical complex processes. There are still certain limitations to this study: although breakthroughs have been made on a specific dataset of steel defects, the generalization ability of the model still needs to be further validated in the face of extreme changes in lighting or highly complex backgrounds in full scene industrial environments. In addition, this article mainly focuses on two-dimensional image detection, and there are still challenges in identifying small scratches that are sensitive to depth information. Future research directions will focus on the following two points: firstly, introducing semi supervised learning or self supervised learning techniques to further enhance the feature extraction ability of the model using a large amount of unlabeled industrial data; The second is to explore multimodal information fusion strategies, combining deep visual information to meet more complex industrial inspection needs.

Acknowledgement: Thanks to all team members for their work and contributions.

Funding Statement: This research is funded by the Key Research and Development Project of Henan Province (241111223000).

Author Contributions: The contributions of the authors of this article are as follows: conceptualization and methodology: Zhengxiang Ma, Xiaofei Ma, Xiaoliang Liu; data curation: Zhengxiang Ma, Xiaofei Ma; software: Zhengxiang Ma, Xiaofei Ma, Xiaoliang Liu; validation and formal analysis: Zhengxiang Ma, Xiaofei Ma, Heng Zhang; statistical summary: Zhengxiang Ma, Xiaofei Ma, Weichao Yu; writing—original draft: Xiaofei Ma; writing—review & editing: Xiaofei Ma, Heng Zhang; supervision: Zhengxiang Ma. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data will be available on request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang X, Wang Z, Guo C, Han Y, Zhao J, Lu N, et al. Application and prospect of new steel corrugated plate technology in infrastructure fields. IOP Conf Ser Mater Sci Eng. 2020;741(1):012099. doi:10.1088/1757-899x/741/1/012099.
2. Wen X, Shan J, He Y, Song K. Steel surface defect recognition: a survey. Coatings. 2022;13(1):2576. doi:10.3390/coatings13010017.
3. Zhao B, Chen Y, Jia X, Ma T. Steel surface defect detection algorithm in complex background scenarios. Measurement. 2024;237(4):115189. doi:10.1016/j.measurement.2024.115189.
4. Liu Y, Zhang C, Dong X. A survey of real-time surface defect inspection methods based on deep learning. Artif Intell Rev. 2023;56(10):12131–70. doi:10.1007/s10462-023-10475-7.
5. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. p. 580–7. doi:10.1109/CVPR.2014.81.
6. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 2980–8. doi:10.1109/ICCV.2017.322.
7. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot MultiBox detector. In: Computer vision—ECCV 2016. Berlin/Heidelberg, Germany: Springer; 2016. p. 21–37. doi:10.1007/978-3-319-46448-0_2.
8. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 779–88. doi:10.1109/CVPR.2016.91.
9. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. arXiv:1506.01497. 2017. doi:10.1109/tpami.2016.2577031.
10. Ge Z, Liu S, Wang F, Li Z, Sun J. YOLOX: exceeding YOLO series in 2021. arXiv:2107.08430. 2021.
11. Jiao L, Zhang F, Liu F, Yang S, Li L, Feng Z, et al. A survey of deep learning-based object detection. IEEE Access. 2019;7:128837–68. doi:10.1109/ACCESS.2019.2939201.
12. Wang CY, Bochkovskiy A, Liao HM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 7464–75. doi:10.1109/CVPR52729.2023.00721.
13. Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 19–25; Nashville, TN, USA. p. 13713–22. doi:10.1109/CVPR46437.2021.01350.
14. Han K, Wang Y, Tian Q, Wang Y, Chen Z, Luo Y, et al. Ghostnet: more features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 1580–9. doi:10.1109/CVPR42600.2020.00166.
15. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q, et al. ECA-Net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 11534–42. doi:10.1109/CVPR42600.2020.01155.
16. Wu L, Chu YK, Yang HG, Chen YX. Sim-YOLOv8 object detection model for DR image defects in aluminum alloy welds. Chin J Lasers. 2024;51(16):1602103. (In Chinese). doi:10.3788/CJL231485.
17. Huang F, Wang T. Insulator defect detection based on lightweight GCP-YOLOv8s. Prog Laser Optoelectron. 2025;62(2):0212004. doi:10.3788/LOP241147.
18. Zhu X, Hu H, Lin S, Dai J. Deformable convnets v2: more deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 9308–16. doi:10.1109/CVPR.2019.00953.

19. Tan M, Pang R, Le QV. EfficientDet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 10781–90. doi:10.1109/CVPR42600.2020.01079.
20. Xie W, Sun X, Ma W. A light weight multi-scale feature fusion steel surface defect detection model based on YOLOv8. Meas Sci Technol. 2024;35(5):055017. doi:10.1088/1361-6501/ad296d.
21. Meng J, Wen S. Detection of steel surface defects based on improved YOLOv8n algorithm. In: Proceedings of the 2024 2nd International Conference on Algorithm, Image Processing and Machine Vision (AIPMV); 2024 Jul 12–14; Zhenjiang, China. p. 8–12. doi:10.1109/AIPMV62663.2024.10692098.
22. Li C, Zhou A, Yao A. Omni-dimensional dynamic convolution. arXiv:2209.07947. 2022.
23. Finder SE, Amoyal R, Treister E, Freifeld O. Wavelet convolutions for large receptive fields. In: European Conference on Computer Vision. Berlin/Heidelberg, Germany: Springer; 2024. doi:10.48550/arXiv.2407.05848.
24. Li Q, Shen L, Guo S, Lai Z. Wavelet integrated CNNs for noise-robust image classification. arXiv:2005.03337. 2020.
25. Dai Y, Gieseke F, Oehmcke S, Wu Y, Barnard K. Attentional feature fusion. In: Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV); 2021 Jan 3–8; Waikoloa, HI, USA. p. 3559–68. doi:10.1109/wacv48630.2021.00360.
26. Li X, Wang W, Hu X, Yang J. Selective kernel networks. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 510–9. doi:10.1109/CVPR.2019.00060.
27. Bachlechner TC, Majumder BP, Mao HH, Cottrell G, McAuley J. ReZero is all you need: fast convergence at large depth. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence; 2020 Aug 3–6; Online.
28. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 7132–41. doi:10.1109/CVPR.2018.00745.
29. Hou Q, Zhang L, Cheng MM, Feng J. Strip pooling: rethinking spatial pooling for scene parsing. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 4002–11. doi:10.1109/CVPR42600.2020.00406.
30. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2015;37(9):1904–16. doi:10.1109/TPAMI.2015.2389824.
31. Song K, Yan Y. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. Appl Surf Sci. 2013;285:858–64. doi:10.1016/j.apsusc.2013.09.002.
32. Dai X, Gao J, Li C, Yang J. Focal modulation networks. In: Proceedings of the Advances in Neural Information Processing Systems 35; 2022 Nov 28–Dec 9; New Orleans, Louisiana, USA. p. 4203–17. doi:10.52202/068431-0304.
33. He Y, Song K, Meng Q, Yan Y. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. IEEE Trans Instrum Meas. 2020;69(4):1493–504. doi:10.1109/TIM.2019.2915404.
34. Qi X, Dong X. Improved steel surface defect detection algorithm of Yolov7-tiny. Comput Eng Appl. 2023;59(12):176–83. doi:10.3934/mbe.2024016.
35. Ding X, Peng L. HDC-YOLOv8s: an improved YOLOv8s algorithm for steel surface defect detection. In: Proceedings of the 2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT); 2025 Mar 21–23; Xi'an, China. p. 762–7. doi:10.1109/ISCAIT64916.2025.11010361.
36. Guo Y, Tang Y. Detection algorithm of steel surface defects based on MHD-YOLO. Prog Laser Optoelectron. 2025;62(20):132–41. (In Chinese).
37. Zhang Y, Xu B, Yang G. Steel surface defect detection based on multi-scale dynamic convolution and lightweight cross-stage fusion. PeerJ Comput Sci. 2026;12:e3485. doi:10.7717/peerj-cs.3485.