# DSP-DETR: Towards Efficient Detail–Spatial Pyramid Modeling for Steel Surface Defect Detection

Yihan Chai[1], Qiming Yu[2,*], and Chengming Liu[1,*]
[1]*School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou, China*
[2]*School of Information Science and Technology, Zhengzhou Normal University, Zhengzhou, China*
**Corresponding authors: qmy.09@163.com, cmliu@zzu.edu.cn*

***Abstract*—Surface defect detection in steel is a fundamental task in industrial inspection systems. In complex inspection scenarios, defects tend to be weakened during cross-scale feature modeling. Moreover, spatial structure and defect morphology are not effectively captured, leading to degraded detection accuracy. These issues mainly result from diverse spatial defect distributions and the limited ability of existing frameworks to jointly model local details and spatial hierarchies. To address these challenges, this paper proposes an enhanced RT-DETR framework, termed Detail–Spatial Pyramid DETR (DSP-DETR). In this framework, the backbone integrates a Multi-Granularity Convolution Block (MGCBlock) to extract detailed features at multiple granularities, which enhances the representation of subtle defects while keeping the model lightweight. A Spatial-Aware Pyramid Fusion Network (SAPFPN) is further introduced to refine feature interaction in the pyramid and model spatial structures and defect morphology. Within this network, a spatial–channel attention (SCA) mechanism with directional operations and large receptive fields strengthens the perception of elongated and texture-sensitive defects, enabling more accurate detection of complex surface textures. Specifically, the proposed model achieves a 4.2% improvement in mAP$_{50}$ on the NEU-DET dataset and a 4.5% gain on the GC10-DET dataset compared to the baseline, validating its effectiveness for industrial surface defect detection.**

***Index Terms*—surface defect detection, RT-DETR, feature pyramid, spatial hierarchy modeling**

## I. Introduction

Steel surface defect detection is a fundamental task in industrial quality inspection, as surface imperfections directly affect product reliability and operational safety. Under industrial imaging conditions, defects often show complex appearances with subtle details and irregular spatial layouts, which leads to defect confusion and inaccurate boundary localization in practical inspection. As manual inspection is inefficient and subjective, vision-based automated inspection systems have become a critical component of modern intelligent manufacturing pipelines [1], [2].

In recent years, deep learning–based visual inspection methods have significantly improved detection accuracy for steel surface defects. These approaches enable data-driven feature learning and reduce reliance on handcrafted descriptors. However, steel surface defects exhibit strong diversity in scale, shape, and spatial continuity. Many defects appear as elongated structures with weak contrast or irregular boundaries, which are easily obscured by noise, illumination variation, and surface reflection. These characteristics pose challenges for existing detection frameworks.

Most current detectors rely on hierarchical feature extraction and multi-scale fusion to address defect variability [3]. While such designs improve robustness, subtle textures and fine-grained structures are often weakened during deep feature aggregation. In addition, limited effective receptive fields constrain the modeling of long-range spatial dependencies, which are essential for identifying extended cracks or continuous defect patterns [4]. Although end-to-end detection paradigms introduce global context modeling through attention mechanisms [5], [6], fine local cues may still be insufficiently preserved when global representations dominate feature learning.

Recent studies indicate that effective defect detection requires a balanced integration of local detail perception and long-range structural modeling [7]. In industrial scenarios, small defects demand high-resolution local sensitivity, while elongated or directional defects require awareness of spatial continuity across larger regions. Meanwhile, practical deployment also requires computational efficiency and stable real-time inference [6]. Achieving these objectives within a unified framework remains an open problem.

Motivated by these observations, this paper proposes Detail–Spatial Pyramid DETR (DSP-DETR), an end-to-end detection framework specifically designed for steel surface defect inspection. The main contributions of this work can be summarized as follows:

- A Multi-Granularity Convolution Block (MGCBlock) is proposed to improve local detail extraction under diverse defect patterns.
- A Spatial-Aware Pyramid Fusion Network (SAPFPN) is introduced to enhance multi-scale feature fusion through bidirectional interaction.
- A spatial–channel attention (SCA) mechanism with large-kernel convolution is employed to strengthen directional and long-range feature perception for elongated defects.

## II. Related Work

*1) Fine-Grained Feature Representation for Steel Surface Defect Detection:* Steel surface defect detection requires effective modeling of fine-grained textures and irregular local structures. Many defects, such as cracks, scratches, and inclusions, present weak contrast and blurred boundaries under industrial imaging conditions. Vasan et al. [8] introduced vision transformers for hot-rolled steel surface defect classification and demonstrated that global attention is useful for capturing structural patterns. However, their method mainly focuses on global representations and does not explicitly enhance local defect details, which limits its effectiveness on small or weak defects.

Peng et al. [9] proposed EB-YOLOv8 with weighted feature fusion and BiFPN structures to balance accuracy and efficiency. Its fusion strategy mainly aggregates features across scales and does not explicitly preserve structural continuity. Other studies enhance local feature extraction using multi-scale convolutional designs or lightweight modules [10], [11]. These methods improve efficiency but often suffer from limited receptive fields or insufficient interaction between local details and high-level semantic features.

*2) Multi-Scale Feature Fusion and Hierarchical Semantic Interaction:* Scale variation is a key challenge in steel surface defect detection. Jiang et al. [12] proposed JAFFNet, which applies joint attention to guide cross-scale fusion. This method emphasizes defect regions but performs feature reweighting and does not model hierarchical spatial relationships.

Transformer-based and hybrid architectures have been explored to enhance hierarchical feature interaction. Liu et al. [13] proposed a transformer framework to capture long-range dependencies. However, dense self-attention increases computational cost and reduces sensitivity to local textures. Wang et al. [14] introduced MDT-Net with deformable transformers and query refinement to improve small defect detection, but with higher complexity. Wu et al. [15] proposed SH-DETR by combining CNN backbones with transformer decoders. Although global context modeling is improved, spatial interaction in feature fusion remains implicit. Most approaches focus on scale alignment or attention weighting, while spatially aware hierarchical fusion remains limited.

*3) Structural Continuity, Directional Modeling, and Efficient Detection:* Elongated defects such as cracks require modeling of structural continuity and directional information. CINFormer integrates multi-stage CNN features into transformer layers to balance local texture and global context [16]. However, directional characteristics are learned implicitly, reducing robustness under complex surface conditions. GMBI-Net introduces group-based multiscale bidirectional interaction to enhance cross-scale information flow with low computational cost [17]. This method focuses on scale interaction and provides limited directional sensitivity.

Recent lightweight models emphasize efficiency and real-time deployment [11], [17], [18]. Simplified feature interaction improves speed but often weakens long-range structural modeling. In contrast, the proposed approach enhances directional perception and structural continuity through spatial modeling while maintaining computational efficiency.

## III. Method

In this paper, we propose DSP-DETR, an enhanced version of RT-DETR for steel surface defect detection, as shown in Fig. 1. The backbone is updated to Swift-Backbone by replacing the original basic blocks with MGCBlock. This change improves feature extraction efficiency while keeping the backbone structure unchanged at a high level. The extracted multi-level features are then sent to SAPFPN for fusion and refinement. The SAPFPN, which incorporates SCA to improve multi-scale feature fusion and defect detection. These innovations address key challenges in steel surface defect detection, such as fine-grained feature loss, weak multi-scale interaction, and handling elongated defects.

### A. Swift-Backbone with MGCBlock

Swift-Backbone follows the stage-wise design of the original RT-DETR backbone. The difference is that each stage adopts MGCBlock as the basic unit. In this way, the backbone strengthens feature representation with limited overhead and keeps the overall pipeline compatible with the RT-DETR encoder-decoder design. Steel surfaces often contain small and low-contrast defects, such as micro-cracks, scratches, and inclusions. These patterns are easy to fade after downsampling. To address this issue, we design MGCBlock to preserve discriminative local textures while maintaining stable channel semantics. As shown in Fig. 2, MGCBlock adopts a multi-branch convolutional structure with two parallel paths.

Specifically, we introduce group convolution and pointwise convolution as complementary operations for feature extraction. Group convolution reduces redundant computation and encourages grouped feature specialization, which has been widely adopted in efficient network design [19]. However, grouped processing weakens cross-channel interaction. Pointwise $1 \times 1$ convolution provides a direct and effective way to mix channel information across groups with low computational cost [20]. By combining these two operations, MGCBlock captures local texture patterns while maintaining global channel consistency. This design is especially helpful for steel surface defects with weak contrast and cluttered backgrounds.

Given an input feature map $X \in \mathbb{R}^{C \times H \times W}$, MGCBlock contains two parallel convolutional paths: a $3 \times 3$ group convolution and a $1 \times 1$ pointwise convolution. The group convolution can be written as

$$F_{\text{gc}} = \text{GConv}_{3\times 3}(X) \tag{1}$$

where the operation splits channels into $g$ groups and applies convolution within each group. A more explicit form is

$$F_{\text{gc}} = \text{Concat}\Big(W_{\text{gc}}^{(1)} * X^{(1)}, \dots, W_{\text{gc}}^{(g)} * X^{(g)}\Big) \tag{2}$$

where $X^{(j)}$ is the $j$-th channel group and $*$ denotes convolution.

The pointwise convolution is defined as

$$F_{\text{pwc}} = \text{PWConv}_{1\times 1}(X) \tag{3}$$

The output of MGCBlock is computed as

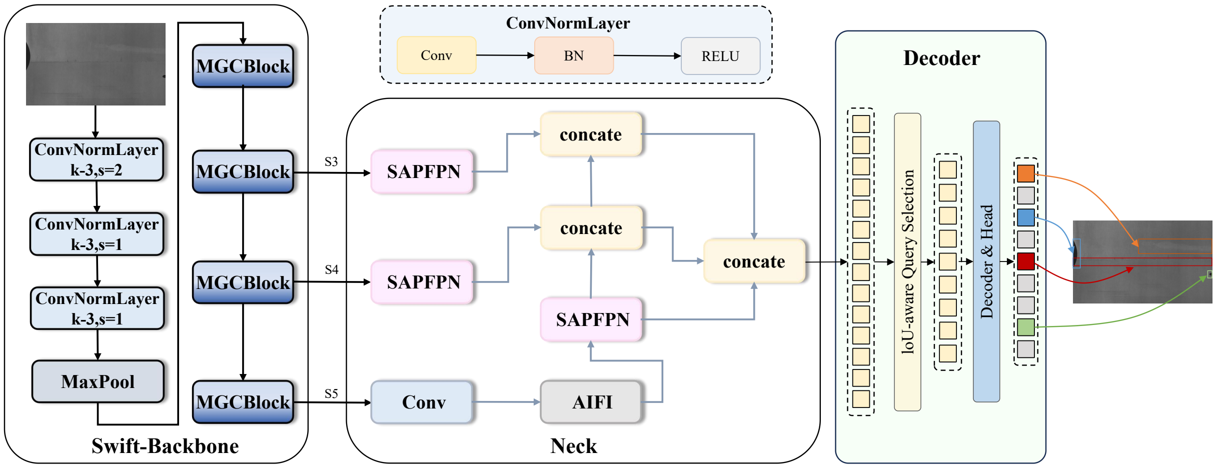


Fig. 1. Overall framework of the proposed DSP-DETR, based on RT-DETR. The framework consists of three main parts: Backbone, Neck, and Decoder. The Backbone enhances feature extraction with the MGCBlock, while the Neck integrates the SAPFPN for multi-scale feature fusion. The decoder processes the fused features based on the RT-DETR architecture and performs defect detection.

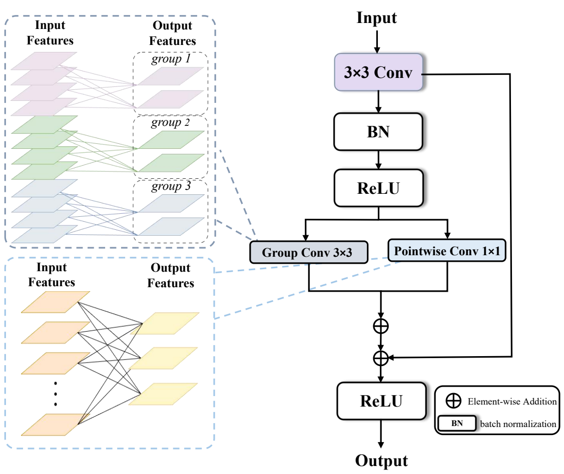


Fig. 2. The structure of MGCBlock with Group Convolution and Pointwise Convolution.

$$Y = \sigma\left(\mathrm{BN}(F_{\mathrm{gc}} + F_{\mathrm{pwc}})\right) \tag{4}$$

where $W_{\mathrm{gc}}$ and $W_{\mathrm{pwc}}$ denote the weights of the group and pointwise convolutions, respectively, and $\sigma$ is the ReLU activation function.

The two convolutional paths are fused by element-wise summation before normalization and activation. This parallel design preserves local texture sensitivity while maintaining effective channel interaction, which improves defect feature separability with limited computational overhead.

### *B. SAPFPN with SCA Module*

To address the strong scale variation and structural diversity of steel surface defects, we design the SAPFPN equipped with direction-sensitive attention. SAPFPN performs feature interaction across multiple pyramid levels and explicitly enhances spatial continuity, which is critical for detecting elongated and low-contrast defects in industrial scenarios. The overall architecture of SAPFPN is illustrated in Fig. 3.

*1) Spatial-Aware Pyramid Fusion:* Steel surface defects exhibit large variations in scale and appearance. Fine-grained defects are mainly captured by shallow features, while elongated and region-level defects rely more on high-level semantics. Conventional FPN structures adopt a unidirectional top-down fusion strategy, which often weakens detail-rich features after repeated upsampling.

SAPFPN adopts a bidirectional pyramid fusion strategy to enable effective interaction between shallow and deep features. High-level semantic features are propagated downward to provide contextual guidance, while low-level features are reinforced upward to preserve local textures. At each fusion node, features from different scales are adaptively combined, allowing spatially consistent information to be retained across pyramid levels.

Formally, the fusion at the $i$-th pyramid level is defined as

$$F_i' = \alpha_i \cdot \mathrm{Up}(F_{i+1}) + (1 - \alpha_i) \cdot F_i \tag{5}$$

where $F_i$ and $F_{i+1}$ denote features from adjacent pyramid levels, $\mathrm{Up}(\cdot)$ represents upsampling, and $\alpha_i$ is a learnable weight that balances semantic and detail information.

This spatial-aware fusion mechanism improves multi-scale consistency and prevents the loss of weak defect cues, which is particularly beneficial for small inclusions and fine surface textures commonly observed in steel products.

*2) SCA with Directional Large-Kernel Convolution:* While pyramid fusion aligns features across scales, it does not explicitly model the structural continuity of defects. Many steel surface defects, such as cracks, scratches, and rolled-in scales, exhibit strong directional patterns and extend over long spatial ranges. To capture such characteristics, we introduce the SCA module with directional large-kernel depthwise convolutions.

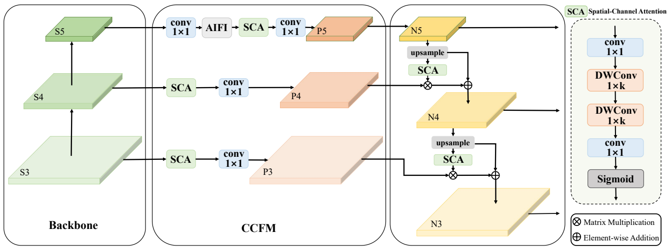


Fig. 3. The structure of SAPFPN for Multi-Scale Feature Fusion with SCA.

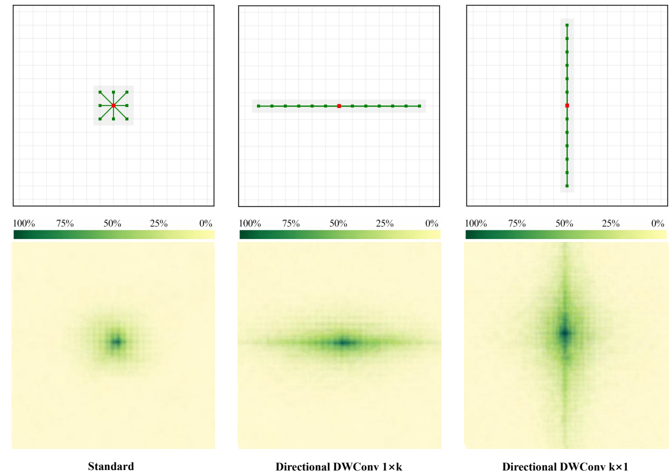


Fig. 4. Visualization of receptive fields for standard convolution and directional depthwise convolutions in the SCA module.

As illustrated in Fig. 4, standard convolution aggregates features within a local neighborhood in an isotropic manner, which limits its ability to model direction-aware structures. This behavior is suboptimal for steel surface defects that exhibit strong continuity along specific orientations.

To address this issue, the proposed SCA module introduces directional depthwise convolutions [21] with large kernels. By separately modeling horizontal and vertical receptive fields, SCA explicitly captures long-range spatial dependencies along principal directions. This design is particularly effective for elongated defects such as cracks, scratches, and weld lines, which are commonly observed in industrial steel inspection.

Given an input feature map $F$, SCA generates a spatial attention factor through sequential directional convolution:

$$A_s = \sigma\big(\mathrm{DWConv}_{k\times 1}(\mathrm{DWConv}_{1\times k}(F))\big) \tag{6}$$

where $\mathrm{DWConv}_{1\times k}$ and $\mathrm{DWConv}_{k\times 1}$ denote horizontal and vertical depthwise convolutions, respectively, and $\sigma$ denotes the sigmoid function. By enlarging the receptive field along horizontal and vertical directions, the proposed design enables effective modeling of long-range structural continuity, which is critical for elongated defects such as cracks and scratches. In this work, the kernel size is set to $k = 13$, which provides an effective receptive field for capturing elongated defect structures without introducing excessive computational cost. The effectiveness of this kernel size is further validated in the ablation experiments.

The final refined feature map is obtained by

$$\tilde{F} = A_s \odot F \tag{7}$$

where $\odot$ denotes element-wise multiplication.

By integrating directional large-kernel attention into the pyramid fusion process, SCA enhances spatial continuity and suppresses background noise. This design is particularly effective for industrial defect detection, where defects often appear as continuous, anisotropic patterns with weak contrast.

Overall, SAPFPN combines bidirectional pyramid fusion with direction-sensitive spatial attention, achieving robust multi-scale representation and improved structural modeling for steel surface defect detection with minimal overhead.

## IV. Experiment and Result Analysis

### A. Datasets

To evaluate the proposed method, experiments are conducted on two public steel surface defect detection datasets: NEU-DET [22] and GC10-DET [23]. These two datasets are widely used in industrial defect inspection and exhibit different defect characteristics.

**NEU-DET Dataset:** The NEU-DET dataset is an open steel surface defect dataset released by Northeastern University. It contains 1,800 grayscale images covering six typical defect categories, including crazing (Cr), Inclusion (In), Patch (Pa), Pitted Surface (Ps), Rolled-in Scale (Rs), and Scratch (Sc). Each category consists of 300 images. Most defects in NEU-DET present thin structures and weak texture variations, which pose challenges for accurate localization and feature extraction.

**GC10-DET Dataset:** GC10-DET is a large-scale steel surface defect dataset collected from real industrial environments. It includes 2,294 high-resolution color images annotated with ten defect categories: Crescent (Cg), Inclusion (In), Rolled-in Scale (Rp), Weld Line (Wl), Crease (Cr), Silk Spot (Ss), Oil Spot (Os), Water Spot (Ws), Punching (Pu), and Waist Fold (Wf). The dataset contains defects with large variations in scale and shape. Some defects appear as long-strip structures, while others are small and difficult to distinguish from the background.

### B. Environment and Experimental Settings

All experiments were implemented on a Linux-based system using the PyTorch framework. Model training and evaluation were carried out on a single NVIDIA RTX 4090 GPU. The software environment was configured with Python 3.11 and CUDA 12.1. The proposed method was trained on two public steel surface defect datasets, GC10-DET and NEU-DET. For optimization, the AdamW optimizer was employed due to its stability in training deep detection models. The initial learning rate was set to $1 \times 10^{-4}$. To avoid potential instability during data loading and training on high-resolution industrial images, no caching strategy was adopted throughout the training process.

To evaluate the effectiveness and efficiency of the proposed method, Average Precision (AP) and mean Average Precision (mAP) were adopted as the primary performance metrics. AP was used to measure the detection accuracy for each defect category, while mAP provides an overall evaluation across all categories.

In addition, model complexity was assessed using the number of parameters (Params) and Floating-Point Operations (FLOPs), which reflect memory consumption and computational cost, respectively. These metrics enable a comprehensive evaluation of the trade-off between detection accuracy and computational efficiency.

### *C. Comparative Experiments*

The proposed DSP-DETR is compared with representative two-stage detectors, single-stage detectors, and recent YOLO-series models on the NEU-DET and GC10-DET datasets, with quantitative results in Table I.

As shown, DSP-DETR has a lightweight architecture with 46.8 GFLOPs and 14.6M parameters, the lowest among all methods. Compared with two-stage detectors like Faster R-CNN and RetinaNet, it reduces computational cost over four times and parameters by more than 60%. Even against lightweight YOLO-based detectors, DSP-DETR maintains a smaller model size, confirming its efficiency.

Despite its low complexity, DSP-DETR achieves the highest detection accuracy on NEU-DET, reaching $mAP_{50}$ of 69.1%. YOLOv5m achieves 65.1% $mAP_{50}$ with 48.0 GFLOPs and 20.89M parameters, while DSP-DETR improves accuracy by 4.0 points and reduces parameters to 14.6M. Compared with YOLOv8m, DSP-DETR gains 5.4 points in $mAP_{50}$ while lowering computational cost by 31.9 GFLOPs, demonstrating superior capture of fine-grained and weak defect patterns.

On the GC10-DET dataset, DSP-DETR achieves an $mAP_{50}$ of 77.7%, which is the highest among all compared methods. Compared with YOLOv11m, DSP-DETR improves the $mAP_{50}$ by 1.2 percentage points while reducing GFLOPs from 67.7 to 46.8 and parameters from 20.04M to 14.6M. Compared with YOLOv10m and YOLOv5m, DSP-DETR achieves accuracy gains of 11.4 and 7.0 percentage points, respectively, demonstrating strong generalization capability on complex and

To further analyze the detection performance, qualitative comparison results are shown in Fig. 5. In the inclusion defect examples, several baseline methods generate redundant bounding boxes or miss weak responses. DSP-DETR produces more consistent localization results with fewer false positives. In challenging cases with low contrast and irregular boundaries, DSP-DETR detects defect regions more completely and suppresses background interference.

TABLE I
COMPARISON WITH EXISTING METHODS ON STEEL SURFACE DEFECT DATASETS

| Model | $mAP_{50}$ ↑ | | FLOPs/G↓ | param/M↓ |
|---|---|---|---|---|
| | NEU-DET | GC10-DET | | |
| RT-DETR [6] | 73.5 | 64.6 | 57.0 | 19.9 |
| Deformable-DETR [5] | 62.8 | 55.5 | 172.6 | 39.6 |
| DAB-DETR [24] | 69.2 | 54.0 | 216.7 | 44.8 |
| Faster R-CNN [25] | 62.8 | 72.6 | 208.0 | 41.4 |
| Cascade-RCNN [26] | 62.8 | 51.8 | 161.7 | 69.8 |
| RetinaNet [27] | 54.1 | 72.5 | 210.0 | 36.5 |
| ATSS [28] | 59.4 | 64.3 | 110.0 | 38.9 |
| TOOD [29] | 55.6 | 62.0 | 199.1 | 32.0 |
| FF-MDE [30] | 70.3 | 57.9 | 55.4 | 28.8 |
| YOLOv5m [31] | 65.1 | 70.7 | 48.0 | 20.8 |
| YOLOv8m [32] | 63.7 | 74.3 | 78.7 | 25.8 |
| YOLOv10m [33] | 55.9 | 66.3 | 58.9 | 15.3 |
| YOLOv11m [34] | 57.8 | 76.5 | 67.7 | 20.0 |
| YOLOv12m [35] | 54.7 | 74.8 | 68.2 | 22.1 |
| **DSP-DETR** | **77.7** | **69.1** | **46.8** | **14.6** |

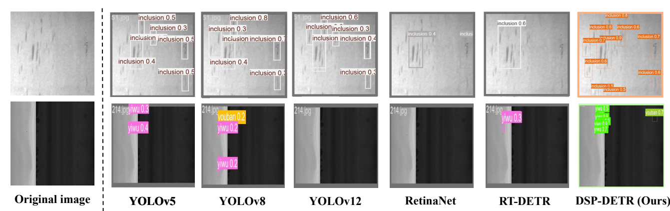


Fig. 5. Qualitative comparison of detection results produced by different methods on steel surface defect images

### *D. Ablation Experiments*

We conduct ablation experiments on the NEU-DET and GC10-DET datasets to analyze the contribution of each component in DSP-DETR. The study focuses on three aspects: the directional large-kernel design in the SCA module, the backbone enhancement with MGCBlock, and the pyramid fusion strategy in SAPFPN with and without SCA. Quantitative results are reported in Tables III, IV, and II, while Fig. 6 provides a visual illustration of the accuracy-efficiency trade-offs among different configurations.

**Effect of directional large-kernel design in SCA.** Table III reports the influence of different directional kernel sizes in the SCA module. On GC10-DET, increasing the kernel size from 3 to 13 improves the $mAP_{50}$ from 66.0% to 67.7%. A similar trend is observed on NEU-DET, where the $mAP_{50}$ increases from 73.9% to 76.3%.

These gains indicate that enlarging the directional receptive field helps capture long-range spatial continuity. This property is important for elongated defects, such as cracks and weld lines, which extend across large regions. Since the kernels are implemented with depthwise convolution, the receptive field is expanded with limited computational overhead.

When the kernel size is further increased to 15, the $mAP_{50}$ drops on both datasets, while FLOPs increase to 60.1G. This suggests that excessively large directional kernels introduce redundant context and weaken local discrimination. Overall, a kernel size of 13 provides the best trade-off between accuracy and efficiency.

**Ablation of MGCBlock and SAPFPN on NEU-DET.**

TABLE II
ABLATION EXPERIMENTS OF MGCBLOCK (A) AND SAPFPN (B) WITH/WITHOUT SCA ON GC10-DET

| A | B | SCA | FLOPs/G↓ | Params/M↓ | $mAP_{50}$/%↑ | $AP_{50}$/%↑ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Cg | In | Rp | Wl | Cr | Ss | Os | Ws | Pu | Wf |
| × | × | × | 57.0 | 19.9 | 64.6 | 93.5 | 33.3 | 22.8 | 86.3 | 24.7 | 57.2 | 68.1 | 87.4 | 91.9 | 80.2 |
| ✓ | × | × | 47.3 | 15.9 | 66.5 | 93.9 | 37.1 | 24.3 | 94.9 | 25.5 | 53.3 | **70.9** | 87.7 | 92.4 | **84.5** |
| × | ✓ | ✓ | 56.7 | 18.6 | 67.7 | 94.3 | 38.6 | 25.6 | 94.3 | 27.6 | **59.9** | 69.6 | 88.6 | 94.2 | 83.9 |
| ✓ | ✓ | × | 46.9 | 14.8 | 68.4 | 94.2 | 39.1 | 29.3 | 93.8 | 30.7 | 57.8 | 70.3 | 89.0 | 95.0 | 84.1 |
| ✓ | ✓ | ✓ | **46.8** | **14.6** | **69.1** | **95.6** | **40.4** | **29.5** | **95.0** | **32.5** | 58.7 | 70.2 | **89.2** | **95.3** | 84.3 |

TABLE III
ABLATION EXPERIMENTS ON DIRECTIONAL KERNEL SIZE IN THE SCA MODULE

| Kernel size | FLOPs/G↓ | Params/M↓ | $mAP_{50}$/%↑ | |
|---|---|---|---|---|
| | | | GC10-DET | NEU-DET |
| 3 | **56.4** | **18.4** | 66.0 | 73.9 |
| 7 | 56.6 | 18.5 | 66.9 | 74.1 |
| 11 | 56.7 | 18.6 | 67.2 | 75.8 |
| 13 | 56.7 | 18.6 | **67.7** | **76.3** |
| 15 | 60.1 | 19.6 | 67.4 | 74.9 |

TABLE IV
ABLATION EXPERIMENTS OF MGCBLOCK (A) AND SAPFPN (B) WITH/WITHOUT SCA ON NEU-DET

| A | B | SCA | FLOPs/G↓ | Params/M↓ | $mAP_{50}$/%↑ | $AP_{50}$/%↑ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Cr | In | Pa | Ps | Rs | Sc |
| × | × | × | 57.0 | 19.9 | 73.5 | 46.3 | 83.8 | 91.4 | 73.9 | 59.1 | 86.2 |
| ✓ | × | × | 47.3 | 15.9 | 74.1 | 44.9 | 81.7 | 91.1 | 77.2 | 59.7 | **90.0** |
| × | ✓ | ✓ | 56.7 | 18.6 | 76.3 | 54.2 | **90.3** | 88.4 | 78.8 | 57.5 | 89.0 |
| ✓ | ✓ | × | 46.9 | 14.8 | 76.6 | 53.7 | 89.2 | 91.9 | 77.1 | 60.4 | 87.4 |
| ✓ | ✓ | ✓ | **46.8** | **14.6** | **77.7** | **55.3** | 88.0 | **92.3** | **80.7** | **61.9** | 88.0 |

The ablation results on NEU-DET are shown in Table IV. The baseline model achieves an $mAP_{50}$ of 73.5% with 57.0 GFLOPs and 19.9M parameters. Introducing MGCBlock reduces the computational cost to 47.3 GFLOPs and the parameter count to 15.9M, while slightly improving the $mAP_{50}$ to 74.1%.

Applying SAPFPN with SCA but without MGCBlock increases the $mAP_{50}$ to 76.3%, indicating that spatially aware pyramid fusion already brings clear performance gains. When SAPFPN is used without SCA, the $mAP_{50}$ reaches 76.6%, which is lower than the configuration with SCA, highlighting the role of directional spatial attention in feature refinement. The full model, which integrates MGCBlock, SAPFPN, and SCA, achieves the best performance, reaching 77.7% $mAP_{50}$ with only 46.8 GFLOPs and 14.6M parameters. Improvements are observed across most defect categories, especially for crack and pitted surface.

**Ablation of MGCBlock and SAPFPN on GC10-DET.** Table II presents the ablation results on GC10-DET. The baseline achieves an $mAP_{50}$ of 64.6%. Adding MGCBlock increases the $mAP_{50}$ to 66.5% while significantly reducing

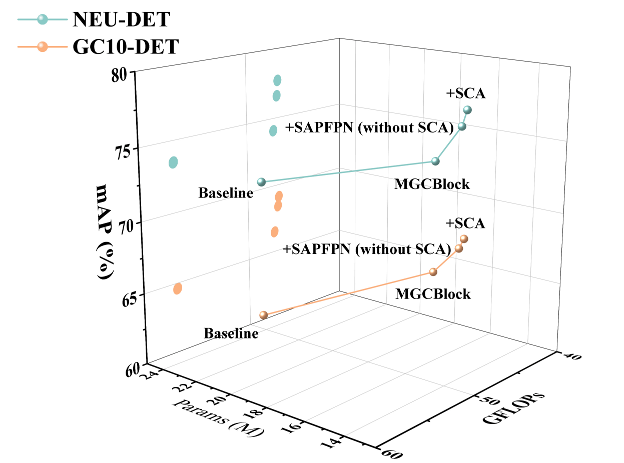


Fig. 6. Visualization of accuracy–efficiency trade-offs in the ablation study on NEU-DET and GC10-DET.

FLOPs and parameters.

Using SAPFPN with SCA alone improves the $mAP_{50}$ to 67.7%, confirming the benefit of spatially aware feature fusion. When SAPFPN is applied without SCA, the $mAP_{50}$ reaches 68.4%, which is lower than the configuration with SCA, indicating that directional spatial attention contributes additional gains beyond pyramid fusion.

The full configuration achieves the highest $mAP_{50}$ of 69.1% with 46.8 GFLOPs and 14.6M parameters. As illustrated in Fig. 6, the progressive introduction of MGCBlock, SAPFPN, and SCA leads to a consistent shift toward higher accuracy and lower complexity on both datasets. Notable improvements are observed for elongated defects such as weld line and crease, which benefit from directional feature modeling.

Overall, the ablation results demonstrate that MGCBlock improves efficiency, SAPFPN enhances multi-scale feature interaction, and the directional large-kernel SCA module strengthens long-range spatial modeling. Their combination achieves consistent accuracy gains with reduced computational cost.

## V. CONCLUSION

This paper presents DSP-DETR, an efficient end-to-end framework for steel surface defect detection in industrial environments. The method targets defects with large scale variation, weak texture, and strong structural continuity while maintaining low computational cost. DSP-DETR enhances feature representation through a multi-granularity backbone and a Spatial-Aware Pyramid Fusion Network. A directional large-kernel spatial attention mechanism is introduced to strengthen long-range and orientation-aware feature modeling, which is effective for elongated defects.

Experimental results on the NEU-DET and GC10-DET datasets show that DSP-DETR achieves higher detection accuracy with fewer parameters and lower computational complexity than existing methods. Ablation studies verify the effectiveness of each component.

DSP-DETR provides a practical solution for real-time steel surface defect inspection and industrial quality control.

## ACKNOWLEDGMENT

This research was supported by the Nature Science Foundation of China (62006210, 62206252), the Key Project of Public Benefit in Henan Province of China (201300210500), and the Key RD Program of Henan (251111212100).

## REFERENCES

[1] X. Wen, J. Shan, Y. He, and K. Song, "Steel surface defect recognition: A survey," *Coatings*, vol. 13, 2023.

[2] B. Tang, L. Chen, W. Sun, and Z.-k. Lin, "Review of surface defect detection of steel products based on machine vision," *IET Image Processing*, vol. 17, no. 2, pp. 303–322, 2023.

[3] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[5] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: deformable transformers for end-to-end object detection," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=gZ9hCDWe6ke

[6] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 16 965–16 974.

[7] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," *Advances in neural information processing systems*, vol. 29, 2016.

[8] V. Vasan, N. V. Sridharan, S. Vaithiyanathan, and M. Aghaei, "Detection and classification of surface defects on hot-rolled steel using vision transformers," *Heliyon*, vol. 10, no. 19, 2024.

[9] M. Peng, S. Bai, and Y. Lu, "Steel surface defect detection algorithm based on improved yolov8 modeling," *Applied Sciences*, vol. 15, no. 15, p. 8759, 2025.

[10] Z. Li, X. Wei, M. Hassaballah, Y. Li, and X. Jiang, "A deep learning model for steel surface defect detection," *Complex & Intelligent Systems*, vol. 10, no. 1, pp. 885–897, 2024.

[11] S. M. Yasir and H. Ahn, "Faster metallic surface defect detection using deep learning with channel shuffling," *arXiv preprint arXiv:2406.14582*, 2024.

[12] X. Jiang, F. Yan, Y. Lu, K. Wang, S. Guo, T. Zhang, Y. Pang, J. Niu, and M. Xu, "Joint attention-guided feature fusion network for saliency detection of surface defects," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[13] G. Liu, Y. Chen, J. Ye, Y. Jiang, H. Yu, J. Tang, and Y. Zhao, "A transformer neural network based framework for steel defect detection under complex scenarios," *Advances in Engineering Software*, vol. 202, p. 103872, 2025.

[14] H. Wang, F. Zhang, and R. Yi, "Multi-scale deformable transformer with iterative query refinement for hot-rolled steel surface defect detection," *Sensors*, vol. 25, no. 22, p. 6890, 2025.

[15] S. Wu, H. Yang, L. Liao, C. Song, Y. Fang, and Y. Yang, "Sh-detr: Enhancing steel surface defect detection and classification with an improved transformer architecture," *PLoS One*, vol. 20, no. 11, p. e0334048, 2025.

[16] X. Jiang, K. Guo, Y. Lu, F. Yan, H. Liu, J. Cao, M. Xu, and D. Tao, "Cinformer: Transformer network with multi-stage cnn feature injection for surface defect segmentation. arxiv 2023," *arXiv preprint arXiv:2309.12639*.

[17] Y. Zhang, C. Chen, Q. Gao, Y. Wang, and B. Fang, "A lightweight group multiscale bidirectional interactive network for real-time steel surface defect detection," *arXiv preprint arXiv:2508.16397*, 2025.

[18] K. Shen, X. Zhou, and Z. Liu, "Minet: Multiscale interactive network for real-time salient object detection of strip steel surface defects," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 5, pp. 7842–7852, 2024.

[19] Y. Ioannou, D. Robertson, R. Cipolla, and A. Criminisi, "Deep roots: Improving cnn efficiency with hierarchical filter groups," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1231–1240.

[20] B.-S. Hua, M.-K. Tran, and S.-K. Yeung, "Pointwise convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 984–993.

[21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[22] K. Song and Y. Yan, "A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects," *Applied Surface Science*, vol. 285, pp. 858–864, 2013.

[23] X. Lv, F. Duan, J.-j. Jiang, X. Fu, and L. Gan, "Deep metallic surface defect detection: The new benchmark and detection network," *Sensors*, vol. 20, no. 6, p. 1562, 2020.

[24] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: dynamic anchor boxes are better queries for DETR," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. [Online]. Available: https://openreview.net/forum?id=oMI9PjOb9Jl

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[26] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[28] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[29] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2021, pp. 3490–3499.

[30] C. Liu, K. Chen, Z. Lian, T. Bai, N. Jia, F. Geng, and N. Wang, "A method for detecting metal surface defects using a dynamic fine-grained multi-branch encoder," *Scientific Reports*, 2025.

[31] R. Khanam and M. Hussain, "What is yolov5: A deep look into the internal features of the popular object detector," 2024. [Online]. Available: https://arxiv.org/abs/2407.20892

[32] M. Yaseen, "What is yolov8: An in-depth exploration of the internal features of the next-generation object detector," 2024. [Online]. Available: https://arxiv.org/abs/2408.15857

[33] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," 2024. [Online]. Available: https://arxiv.org/abs/2405.14458

[34] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," 2024. [Online]. Available: https://arxiv.org/abs/2410.17725

[35] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025.