Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Steel-CGA: Channel-Spatial-Semantic Attention and Grouped Dilated Fusion for Real-Time Steel Surface Defect Detection

Jianjian Chen
Shandong Youth University of Political Science
Baolai Wang
Shandong Youth University of Political Science
Xinghan Chen
Shandong Youth University of Political Science
Shuai Wang
230013@sdyu.edu.cn

Shandong Youth University of Political Science

Article

Keywords: Steel surface defect detection, Channel-Spatial-Semantic Attention Module, Grouped Dilated Residual Fusion, Adaptive Downsampling

Posted Date: September 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10724970/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Steel-CGA: Channel-Spatial-Semantic Attention and Grouped Dilated Fusion for Real-Time Steel Surface Defect Detection

Jianjian Chen[1,2,3], Baolai Wang[1], Xinghan Chen[1], Shuai Wang[1,2,3*]

[1]School of Information Engineering,Shandong Youth University Of Political Science,Jingshi East Road, Jinan, 250103, Shandong, China.
[2]New Technology Research and Development Center of Intelligent Information Controlling in Universities of Shandong,Shandong Youth University Of Political Science,Jingshi East Road, Jinan, 250103, Shandong, China.
[3]Smart Healthcare Big Data Engineering and Ubiquitous Computing Characteristic Laboratory in Universities of Shandong,Shandong Youth University Of Political Science,Jingshi East Road, Jinan, 250103, Shandong, China.

*Corresponding author(s). E-mail(s): 230013@sdyu.edu.cn;

Abstract

Reliable detection of steel surface defects is essential for ensuring product quality and production safety. However, existing detectors struggle to jointly handle complex-background interference, weak responses to small targets, and spatial information loss under lightweight downsampling. We propose Steel-CGA, a YOLO26s-based detector that addresses these challenges for steel surface defect recognition. Steel-CGA integrates three complementary components: (1) a Channel-Spatial-Semantic Attention Module (CSSAM) appended to the backbone, which fuses channel, spatial, and semantic gated attention through residual connections to suppress complex-background interference; (2) a Grouped Dilated Residual Fusion (GDRF) module embedded in the C3k2 blocks of the neck, which enlarges the receptive field to enrich multi-scale context and strengthen responses to small defects; and (3) an Adaptive Downsampling (ADown) module that replaces the conventional downsampling layers in the backbone, reducing spatial

information loss while preserving inference efficiency. On NEU-DET and GC10-DET, Steel-CGA achieves 76.6% and 62.0% mAP@50, outperforming YOLO26s by 3.5 and 3.6 points, while reducing parameters by 26.0% and FLOPs by 23.7% and running at 134 FPS.

Keywords: Steel surface defect detection, Channel-Spatial-Semantic Attention Module, Grouped Dilated Residual Fusion, Adaptive Downsampling.

1 Introduction

Steel is a fundamental material in modern industry, underpinning key economic sectors such as construction, transportation, aerospace, and energy [1]. However, manufacturing processes such as smelting and rolling readily introduce surface defects, including cracks, scratches, inclusions, and rolled-in scale. These imperfections not only degrade the mechanical properties and service life of the products but also pose serious safety hazards under high-stress service conditions, potentially causing substantial economic losses and even casualties [2]. Rapid and accurate detection of steel surface defects is therefore of critical practical importance for ensuring product quality, maintaining production safety, and advancing intelligent manufacturing [3].

Deep learning has become the mainstream paradigm for steel surface defect detection, as it automatically learns complex visual patterns and substantially improves recognition performance. Convolutional neural networks (CNNs) have outperformed traditional methods across a range of vision tasks, including defect detection [4, 5] and semantic segmentation [6]. Object detectors are commonly grouped into two-stage and one-stage paradigms. Two-stage methods such as Faster R-CNN [7], Mask R-CNN [8], and Cascade R-CNN [9] achieve high accuracy but remain computationally heavy for real-time production lines [10]. One-stage detectors, including SSD [11], RetinaNet [12], and the YOLO series [13–16], offer a better speed–accuracy trade-off and are widely adopted for industrial deployment; recent real-time end-to-end models such as RT-DETR [17] further advance this direction. Recent advances in related vision fields further inform this study: techniques such as feature recalibration and multi-scale context modeling have proven effective in enhancing a model's perception of critical regions [18–20]. These strategies provide feasible pathways for tackling complex-background interference, weak responses to small targets, and spatial information loss under lightweight downsampling in steel surface inspection, laying a solid foundation for the proposed method.

Complex industrial backgrounds remain a primary obstacle in steel surface defect detection. Rolling textures, oxide scale, water stains, and illumination variation often produce responses similar to true defects, leading to frequent false alarms on production lines. To alleviate this issue, attention mechanisms such as SENet and CBAM, together with non-local modeling, have been widely incorporated into steel inspectors to emphasize defect regions and suppress textured backgrounds [20–24]. Dual-attention and coordinate-aware variants further attempt to reweight informative channels and spatial locations under cluttered scenes [19, 25]. Nevertheless,

most of these modules still process channel and spatial cues separately and lack joint channel–spatial–semantic calibration; as a result, discriminative features of low-contrast defects are easily overwhelmed, and false positives remain frequent under uneven illumination [26, 27].

Weak responses to small and multi-scale defects constitute a second persistent challenge. Micro-cracks, fine scratches, and sparsely distributed pits occupy only a few pixels after deep downsampling, so their features are readily diluted by surrounding context. Feature pyramids, large-kernel or dilated convolutions, and multi-scale fusion blocks have therefore been adopted to enlarge the receptive field and improve small-defect perception [28–31]. Lightweight YOLO variants further strengthen cross-scale aggregation for industrial inspection [25, 32]. However, existing multi-scale designs often lack a unified interaction and residual calibration framework across feature levels, which limits complementary use of fine details and global context. Simply stacking fusion modules can also inflate parameters and computation, aggravating the accuracy–efficiency trade-off without guaranteeing stable gains on tiny defects [33–35].

Meanwhile, edge deployment drives aggressive lightweighting through depthwise separable and group convolutions [36–39]. Recent detectors further redesign downsampling operators—for example, space-to-depth style modules and detail-preserving pooling—to mitigate information loss during resolution reduction [31, 32, 40]. Despite these efforts, conventional strided convolution and pooling still discard spatial cues that are critical for tiny defects, and overly aggressive compression often trades away localization accuracy for speed. Reducing such information loss while sustaining real-time inference on resource-constrained devices thus remains an open problem.

In recent years, a number of YOLO-based improvements have been proposed for steel surface defect detection, demonstrating the continued vitality of this research direction. For instance, HDR-YOLO [41] introduces a hierarchical depth-aware refinement strategy to enhance shallow texture preservation and deep semantic modeling simultaneously. YOLO-DCF [42] incorporates dual distillation and context-aware fusion to improve detection robustness under complex backgrounds. SCCI-YOLO [43] integrates SPD-Conv, C2f_EMA, and Inner-IoU loss to achieve lightweight and accurate detection. YOLO-MFD [44] employs multi-scale edge feature enhancement and a global-to-local spatial attention BiFPN to further boost detection performance. Despite these advances, jointly handling complex-background interference, weak small-target responses, and spatial information loss under lightweight downsampling remains a challenging open problem.

To overcome the difficulty of jointly handling complex-background interference, weak small-target responses, and spatial information loss under lightweight designs, we propose **Steel-CGA**, an enhanced defect detector based on the YOLO26s framework and tailored for the accurate recognition of diverse steel surface defects. Our method introduces three core innovations. **First**, to suppress background interference, we design a **Channel-Spatial-Semantic Attention Module (CSSAM)** that fuses channel, spatial, and semantic gated attention via residual connections, markedly improving discrimination under complex textures. **Second**, to mitigate the loss of small-target details, we introduce a **Grouped Dilated Residual Fusion (GDRF)** module that leverages multi-scale dilated convolutions to expand the receptive field

and capture rich context, strengthening responses to tiny defects across scales. **Third**, to optimize the accuracy–efficiency trade-off, we integrate an **Adaptive Downsampling (ADown)** module that minimizes parameter redundancy and information loss, yielding a lightweight architecture without compromising performance.

The main contributions of this study are as follows:

- We design CSSAM, which fuses channel, spatial, and semantic gated attention through residual connections, markedly improving robustness against complex-background interference.
- We propose the GDRF module, which exploits multi-scale receptive fields to strengthen contextual modeling and thereby enhance the detection of small and low-contrast defects.
- We integrate the ADown module to obtain a lightweight architecture that substantially reduces the parameter count while preserving detection accuracy.
- We propose Steel-CGA and validate it on NEU-DET and GC10-DET, where it achieves 76.6% and 62.0% mAP@50, surpassing the YOLO26s baseline by 3.5 and 3.6 points and outperforming current mainstream detectors. Meanwhile, parameters and FLOPs are reduced by 26.0% and 23.7%, respectively, and inference reaches 134 FPS, establishing a new state-of-the-art accuracy–efficiency trade-off for real-time steel surface defect detection.

The remainder of this paper is organized as follows. Section 2 reviews related work on object detection and attention mechanisms. Section 3 details the proposed Steel-CGA framework and its core components. Section 4 presents the experimental setup, comparative results, and ablation studies. Finally, Section 5 concludes the paper and discusses future directions.

2 Related Work

2.1 Two-stage object detection algorithm

Two-stage detectors attain high accuracy by first generating candidate boxes and then performing classification and regression. A representative example is Faster R-CNN [7], which produces high-quality proposals via a Region Proposal Network (RPN) and applies RoI Pooling for subsequent classification and regression. Mask R-CNN [8] extends Faster R-CNN with a segmentation branch, jointly addressing object detection and instance segmentation. Cascade R-CNN [9] adopts a cascade structure to progressively refine box quality, markedly improving performance in complex scenarios. Further enhancements include deformable convolution [45], which adaptively adjusts the kernel shape to better detect irregular targets; RoI Align [8], which replaces RoI Pooling to reduce feature misalignment and improve localization; attention mechanisms [46], which strengthen the representation of key regions; and cascade optimization [9], which dynamically raises the IoU threshold to progressively improve proposal quality. Although such methods excel in complex scenes and small-object detection, their heavy computation and slow inference make them ill-suited to the real-time requirements of industrial settings.

2.2 One-stage object detection algorithm

One-stage detectors predict object locations and categories directly in a single forward pass, greatly improving efficiency. The YOLO series [13–15] attains a strong speed–accuracy balance through continual refinement of network structures and loss functions [47]. SSD [11] exploits multi-scale feature maps to detect objects of different sizes with high efficiency, while RetinaNet [12] addresses class imbalance via Focal Loss and improves small-target detection. Subsequent improvements fall into four main categories. (i) Attention mechanisms [25, 48–50], such as the Convolutional Block Attention Module (CBAM), enhance feature extraction in defect regions by fusing channel and spatial attention in parallel. (ii) Multi-scale feature fusion designs [31, 51], such as the Bidirectional Feature Pyramid Network (BiFPN), aggregate features across scales. (iii) Anchor optimization [52–54] redesigns anchor dimensions (e.g., via K-Means++) to improve recall for small and elongated objects. (iv) Lightweight designs [31, 55–57], such as depthwise separable convolutions, reduce computational cost and boost real-time performance. Despite their fast inference, these methods still leave room for improvement in detecting small objects against complex backgrounds [27].

2.3 Complex-background interference suppression

Complex industrial backgrounds remain a primary obstacle in steel surface defect detection. Rolling textures, oxide scale, water stains, and illumination variation often produce responses similar to true defects, so detectors must selectively emphasize defect cues while suppressing clutter. Classical channel attention, pioneered by SENet [20], recalibrates feature maps with global average pooling and excitation to strengthen informative channels. CBAM [21] further cascades channel and spatial attention, whereas Coordinate Attention [58] embeds directional positional information into channel weights. Dual Attention [19], Non-Local blocks [22], and GCNet [59] capture long-range dependencies and global context beyond local convolutions.

These general-purpose modules have been widely transferred to steel inspection. Zhou et al. [23] insert CBAM into a YOLOX-style backbone and couple it with ASFF, reporting clear mAP gains on NEU-DET by focusing on defect regions amid clutter. Huang et al. [24] replace standard convolutions with deformable kernels and stack CBAM in YOLOv5, improving detection probability by 4.3 points with only about 1 ms extra latency. Liu and Ma [60] design a multi-layer attention network with dilated weighted cross-stage fusion, enlarging attended regions on both NEU-DET and GC10-DET. Wang et al. [61] combine BEMD denoising with SE and Coordinate Attention in YOLOv5s to resist noise-induced false alarms, while Xin and Zhang [62] fuse channel–spatial attention with BiFPN for surface inspection. Lightweight ECA/SimAM injections into YOLO necks further enhance key-feature focus under strip-steel textures [25, 49], and transformer-style channel modeling has been explored for strip defects [63]. Despite these advances, most modules still process channel and spatial cues separately, or couple them only loosely, and rarely introduce an explicit semantic gate that jointly calibrates channel, spatial, and semantic responses [26, 64]. Consequently, low-contrast defects remain easily overwhelmed by textured backgrounds, and false positives persist under uneven illumination [27, 65].

2.4 Small-target and multi-scale defect detection

Weak responses to small and multi-scale defects constitute a second persistent challenge. Micro-cracks, fine scratches, and sparsely distributed pits occupy only a few pixels after deep downsampling, so their features are readily diluted. Feature Pyramid Networks (FPN) [28] propagate high-level semantics top-down to enrich shallow maps; PANet [66] adds a bottom-up path for localization cues; NAS-FPN [67] searches fusion topologies; and BiFPN [39] learns cross-scale fusion weights. Large-kernel and dilated convolutions enlarge the receptive field without aggressively shrinking resolution [29, 68], while adaptive spatial fusion further balances contributions across levels [69].

Steel-oriented detectors build on these ideas with task-specific multi-scale designs. Wang et al. [30] add multi-scale explore blocks and spatial attention to YOLO-v5 for real-time strip inspection. Xie et al. [31] propose LMS-YOLO with lightweight multi-scale mixed convolutions, BiFPN fusion, and NWD loss, improving small-defect mAP on both NEU-DET and GC10-DET while cutting parameters. Yu et al. [70] replace PANet with CoT-enhanced GFPN so that contextual interactions enlarge attended defect regions under clutter. Zhao et al. [71] design a BiFPN-style neck that recovers small-strip defects attenuated by conventional PANet and reduce parameters by over 60% relative to YOLOv5s. Related work further strengthens small-object layers via multi-scale enhancement and dilated re-parameterization [3, 32], or refines fusion strategies for industrial small targets [33, 34]. Nevertheless, many pipelines still lack a unified residual interaction that jointly calibrates fine details and global context inside the neck; simply stacking pyramid or dilated blocks often inflates parameters and FLOPs without guaranteeing stable gains on tiny defects [35, 64]. Balancing multi-scale representation with compact computation therefore remains unresolved for real-time steel inspection.

2.5 Lightweight design and detail-preserving downsampling

Edge deployment further drives aggressive lightweighting. Depthwise separable convolutions in MobileNets [36, 72], channel shuffling in ShuffleNet [37], extreme compression in SqueezeNet [38], Ghost feature generation [73], and Xception-style separable operators [55] substantially cut computation, while EfficientDet demonstrates scalable compound scaling for detection [39]. Dynamic network selection has also been studied to allocate capacity on demand [74].

In steel inspection, these operators are commonly specialized for strip surfaces. Lu et al. [25] replace YOLOv7 backbone blocks with MobileNetv3 and introduce parameter-free attention, reducing FLOPs by about 79% while improving mAP50. Yuan et al. [75] rebuild YOLOv8n with deformable convolutions, Coordinate Attention, Faster blocks, and Ghost convolution for a compact multi-category detector. Chu et al. [76] and Wang et al. [77] similarly couple Ghost/Swin-style light backbones with BiFPN fusion, cutting parameters by roughly one-third while retaining accuracy. LMS- and LSD-style YOLO variants jointly lighten the network and strengthen cross-scale fusion for deployment [31, 71]. Because conventional stride-2 convolutions and pooling discard spatial cues critical to tiny defects, recent work redesigns

downsampling—for example, space-to-depth Ghost convolution in GDM-YOLO to retain information during resolution reduction [32]—or proposes lightweight architectures that explicitly target detail preservation [40]. Even so, aggressive compression still often trades away localization accuracy, and many downsampling replacements either remain information-lossy for micro-defects or introduce overhead that offsets the intended speed gains. Reducing spatial information loss while sustaining real-time inference on resource-constrained devices thus remains an open problem.

To address the three challenges above, we propose Steel-CGA with CSSAM for joint channel–spatial–semantic background suppression, GDRF for grouped dilated multi-scale small-target modeling, and ADown for detail-preserving lightweight downsampling.

3 Proposed Method

3.1 Overview

The baseline You Only Look Once version 26 small (YOLO26s) model comprises a backbone, a neck, and a head. The backbone integrates large-kernel depthwise separable convolutions together with C3k2 and C2PSA modules to enlarge the receptive field and strengthen attention over key regions. The neck adopts a multi-path, multi-scale fusion strategy with optimized cross-layer connections to improve feature representation for small objects. The head uses an end-to-end, NMS-free dual-head parallel design and removes the DFL branch, directly outputting predictions to reduce post-processing latency.

However, steel defect detection is hindered by complex backgrounds, weak small-target responses, and information loss under lightweight downsampling, while the baseline still suffers from feature-map redundancy, a large parameter count, and high computational cost. To address these issues, we build upon YOLO26s and propose a framework tailored specifically for steel defect detection.

Steel-CGA is designed for the accurate identification of complex steel surface defects and integrates three key modules: CSSAM, GDRF, and ADown. CSSAM is placed after the C2PSA module in the backbone, where it aggregates channel, spatial, and semantic attention via residual connections to enhance discriminative feature responses and suppress complex-background interference. GDRF reconstructs the C3k2 structure in the neck, enlarging the multi-scale receptive field to better capture context and strengthen the detection of small-scale defects. Finally, ADown is embedded in the backbone to replace the stride-2 convolutional downsampling after C3k2 or C2PSA; it employs dual-path hybrid pooling with feature reorganization to preserve detail during resolution reduction, thereby mitigating the information loss inherent to conventional downsampling.

3.2 Channel-Spatial-Semantic Attention Module (CSSAM)

We introduce CSSAM into the backbone to fuse three complementary attentions for steel surface defect detection under complex backgrounds. The module adopts a three-branch parallel architecture—channel-group attention, spatial attention, and a

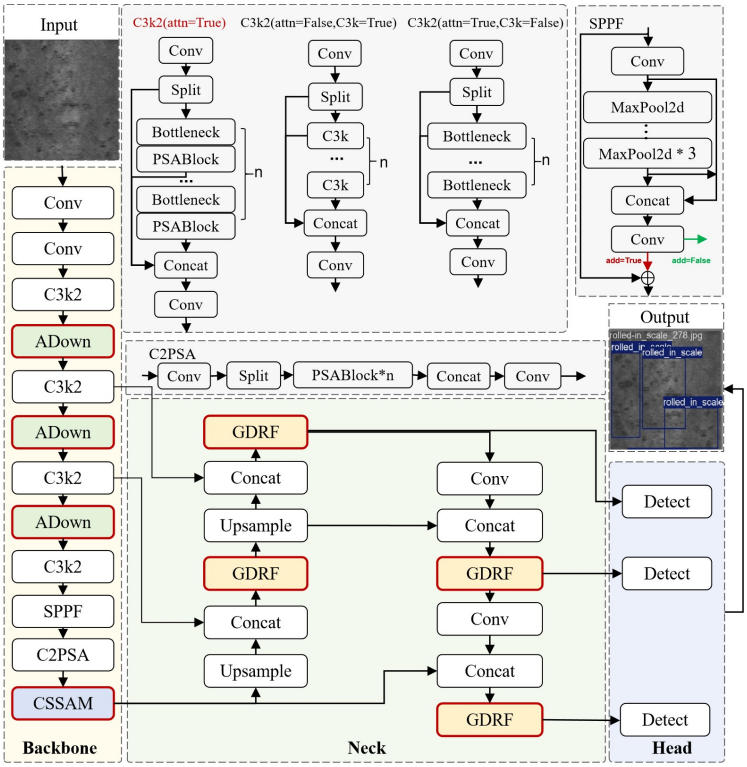


Fig. 1 Overall architecture of the proposed Steel-CGA. CSSAM is incorporated into the backbone, GDRF is embedded in the C3k2 blocks of the neck, and the standard convolutions in the backbone are replaced with ADown.

semantic gate—that captures channel-wise correlations and spatial importance via global average pooling and depthwise separable convolutions. The channel-group and spatial branches adaptively emphasize defect-relevant features and localize defect regions, while the semantic gate and residual connection suppress false positives caused by complex textures, markedly improving background-interference suppression and meeting industrial demands for both high efficiency and high accuracy. The architecture is illustrated in Fig. 2.

CSSAM consists of three parallel, lightweight, and complementary attention branches combined through a residual fusion mechanism, and operates as follows. Given an input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, CSSAM first partitions it into G non-overlapping channel groups $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_G\}$, where each group $\mathbf{X}_g \in \mathbb{R}^{C_g\times H\times W}$ and $C = G \cdot C_g$.

Step 1: Channel-Group Attention. For each group $\mathbf{X}_g$, channel attention is obtained by global average pooling followed by a bottleneck of two 1×1 convolutions and a nonlinear activation:

$$\mathbf{A}_{\text{chan}}^{(g)} = \sigma\left(W_2^{(g)} \cdot \delta\left(W_1^{(g)} \cdot \text{GAP}(\mathbf{X}_g)\right)\right) \tag{1}$$

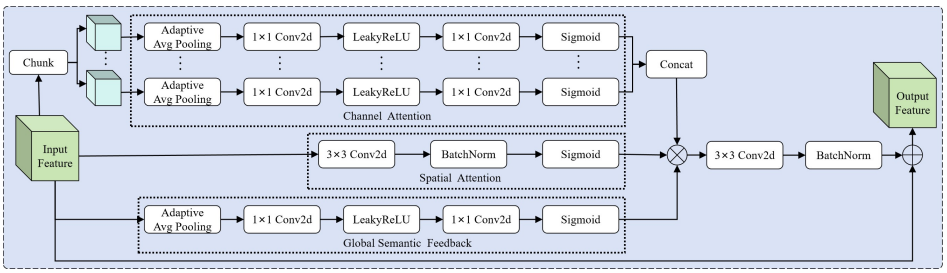


Fig. 2 Architecture of the Channel-Spatial-Semantic Attention Module (CSSAM), comprising channel-group recalibration, spatial attention, and a semantic gate. The three branch outputs are fused to selectively enhance informative features.

where GAP$(\cdot)$ denotes global average pooling, $W_1^{(g)}$ and $W_2^{(g)}$ are learnable 1×1 convolutions, $\delta(\cdot)$ is the LeakyReLU activation function, $\sigma(\cdot)$ is the sigmoid function, $\mathbf{X}_g$ denotes the g-th channel group, and its shape is $\mathbb{R}^{C_g \times H \times W}$.

The channel-refined feature for group g is then:

$$\tilde{\mathbf{X}}_g = \mathbf{X}_g \odot \mathbf{A}_{\text{chan}}^{(g)} \tag{2}$$

where $\odot$ denotes element-wise multiplication. All group-wise outputs are concatenated to reconstruct the full-channel-attended feature map:

$$\tilde{\mathbf{X}} = \text{Concat}(\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_G) \tag{3}$$

where $\tilde{\mathbf{X}}$ denotes the full-channel-attended feature map, and Concat$(\cdot)$ denotes concatenation along the channel dimension.

Step 2: Spatial Attention. To model spatial dependencies, we apply a depthwise separable convolution followed by batch normalization and a sigmoid activation:

$$\mathbf{A}_{\text{spa}} = \sigma\left(\text{BN}(\text{DWConv}(\mathbf{X}))\right) \tag{4}$$

where $\mathbf{A}_{\text{spa}}$ denotes the spatial attention map, $\sigma(\cdot)$ denotes the sigmoid activation, BN$(\cdot)$ denotes batch normalization, and DWConv$(\cdot)$ denotes depthwise separable convolution.

Step 3: Semantic Attention Gate. To capture global semantics, a semantic gate is generated from the original input by a two-layer MLP:

$$\mathbf{A}_{\text{sem}} = \sigma\left(W_4 \cdot \delta\left(W_3 \cdot \text{GAP}(\mathbf{X})\right)\right) \tag{5}$$

where W_3 and W_4 are shared 1×1 convolution layers, $\mathbf{A}_{\text{sem}}$ represents the semantic gating signal, GAP$(\cdot)$ denotes global average pooling.

Step 4: Attention Fusion and Residual Learning. The attention-modulated feature is obtained by combining the three attention maps:

$$\mathbf{X}' = \tilde{\mathbf{X}} \odot \mathbf{A}_{\text{spa}} \odot \mathbf{A}_{\text{sem}} \tag{6}$$

A final projection layer refines the output:

$$\mathbf{Y} = \text{Dropout}\left(\text{BN}\left(\text{Conv}_{1\times1}(\mathbf{X}')\right)\right) \tag{7}$$

where $\mathbf{Y}$ denotes the projected feature map, Dropout$(\cdot)$ is the dropout layer, BN$(\cdot)$ denotes batch normalization, and $\text{Conv}_{1\times1}(\cdot)$ is a 1×1 convolution.

Lastly, a residual connection is applied to preserve the original signal:

$$\hat{\mathbf{Y}} = \mathbf{Y} + \mathbf{X} \tag{8}$$

Where $\hat{\mathbf{Y}}$ denotes the final output of the CSSAM module, whose shape is the same as the input, i.e., $C \times H \times W$, $\mathbf{Y}$ denotes the projected feature map, and $\mathbf{X}$ is the original input.

3.3 The Grouped Dilated Residual Fusion Module(GDRF)

GDRF is obtained by embedding the Dilation-wise Residual (DWR) module from DWRSeg, a multi-scale real-time semantic segmentation model [68], into the C3k2 module of the neck, where it replaces the original bottleneck component. The module enlarges the receptive field via dilated convolutions with varied rates, while residual connections fuse multi-scale information, strengthening the model's capacity to capture small and multi-scale defect features. The architecture is depicted in Fig. 3.

The DWR module aggregates multi-scale context in two stages to enrich the semantic representation of features.

Step 1: Regional Residual. A combination of 3×3 convolutions, batch normalization (BN), and ReLU activations performs preliminary feature extraction, producing feature maps with distinct regional patterns that serve as inputs to the subsequent stage. Let the input feature map be denoted as

$$\mathbf{X} \in \mathbb{R}^{H\times W\times C} \tag{9}$$

The module employs a set of depthwise dilated convolutions with multiple dilation rates for regional feature extraction:

$$\mathbf{F}^{(k)}_{\text{region}} = \text{Conv}_{d_k}(\mathbf{X}) \tag{10}$$

where Conv_{d_k} denotes the convolution with dilation rate d_k, and k indexes branches with different dilation rates.

Step 2: Semantic Residual. The regional feature maps from Step 1 are partitioned into three groups and processed by three parallel dilated convolutions with dilation rates of 1, 3, and 5, respectively, performing semantic feature extraction over predefined receptive fields:

$$\mathbf{F}^{(k)}_{\text{semantic}} = \mathcal{G}(\mathbf{F}^{(k)}_{\text{region}}; \Theta_k) \tag{11}$$

where $\mathcal{G}$ is the dilated convolution function, and Θ_k are the corresponding branch parameters.

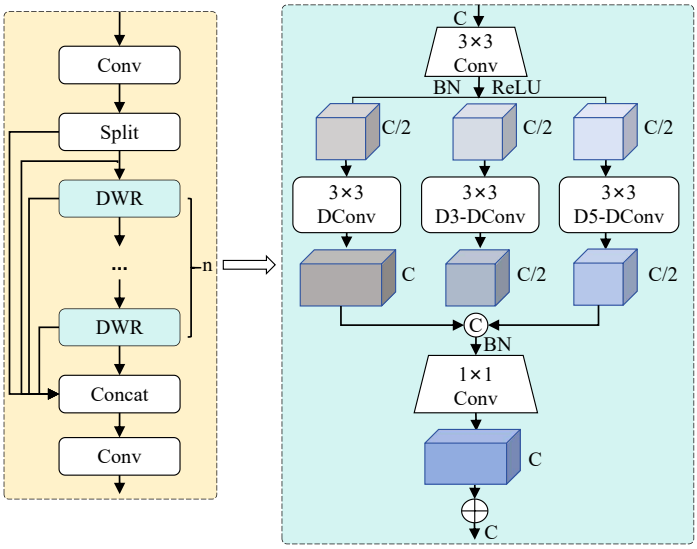


Fig. 3 Illustration of the GDRF module, which enhances small-target and multi-scale feature extraction for steel defects by combining dilated convolutions with residual connections to fuse multi-scale contextual information.

Step 3: Feature Fusion and Residual Connection. We fuse the outputs of the three branches with a 1×1 pointwise convolution. To adaptively balance their multi-scale contributions, we introduce learnable coefficients α_k, normalized by a Softmax so that they sum to one. The unnormalized weights $\hat{\alpha}_k$ are randomly initialized and optimized end-to-end together with the other network parameters via standard backpropagation, requiring no additional attention modules.

$$\alpha_k = \frac{\exp(\hat{\alpha}_k)}{\sum_{j=1}^{K} \exp(\hat{\alpha}_j)}, \quad k = 1, 2, \ldots, K \tag{12}$$

where K denotes the total number of branches, $\hat{\alpha}_k$ are learnable unnormalized parameters, and α_k are the normalized adaptive fusion weights.

Finally, a residual connection adds the original input to the fused feature map, yielding a more comprehensive representation:

$$\mathbf{Y} = \mathbf{X} + \sum_{k=1}^{K} \alpha_k \cdot \mathbf{F}_{\text{semantic}}^{(k)} \tag{13}$$

where α_k is an adaptive weight coefficient.

3.4 Adaptive Downsampling(ADown)

ADown is a convolutional block designed for downsampling in object detection. Through a lightweight two-pathway architecture—a primary convolutional branch and a secondary max-pooling branch—it balances information preservation with computational efficiency. By combining an adaptive strategy with multi-scale fusion, ADown

reduces the parameter count by roughly 20% while improving sensitivity to small targets, thereby outperforming conventional downsampling in parameter count, computational efficiency, and detection accuracy. As illustrated in Fig. 4, the ADown module proceeds as follows.

Average Pooling. The input feature map is denoted as $X \in \mathbb{R}^{C\times H\times W}$, and the pooled output is

$$X_{\text{avg}} = \text{AvgPool2d}(X) \tag{14}$$

where AvgPool2d$(\cdot)$ denotes an average pooling operation with a kernel size of 2×2 and a stride of 2.

Channel Splitting and Convolution. The tensor X_{avg} is split into two parts along the channel dimension:

$$X_{\text{avg1}}, X_{\text{avg2}} = \text{Split}(X_{\text{avg}}) \tag{15}$$

where $X_{\text{avg1}}, X_{\text{avg2}} \in \mathbb{R}^{\frac{C}{2}\times\frac{H}{2}\times\frac{W}{2}}$ and Split$(\cdot)$ denotes an equal split along the channel dimension. The first part X_{avg1} is then convolved:

$$X_{\text{conv1}} = \text{Conv2d}(X_{\text{avg1}}) \tag{16}$$

where Conv2d$(\cdot)$ represents 2D convolution.

Max Pooling and Feature Fusion. The second part X_{avg2} undergoes max pooling (stride 1, with zero padding):

$$X_{\text{max}} = \text{MaxPool}(X_{\text{avg2}}) \tag{17}$$

where MaxPool$(\cdot)$ denotes 2×2 max-pooling with stride 1.

Next, X_{max} is zero-padded to restore its spatial resolution to match the first branch:

$$X_{\text{pad}} = \text{ZeroPad2d}(X_{\text{max}}) \tag{18}$$

where ZeroPad2d$(\cdot)$ denotes the zero-padding operation.

The padded features are then processed by a convolutional layer:

$$X_{\text{conv2}} = \text{Conv2d}(X_{\text{pad}}) \tag{19}$$

Finally, the two convolutional outputs are concatenated along the channel dimension to form the module output:

$$Y = \text{Concat}(X_{\text{conv1}}, X_{\text{conv2}}) \tag{20}$$

where Concat$(\cdot)$ denotes concatenation across channels, and $Y \in \mathbb{R}^{C\times\frac{H}{2}\times\frac{W}{2}}$ denotes the final output of the module.

Parameter analysis. Consider a downsampling layer that maps C input channels to C output channels. Conventional downsampling applies a single 3×3 stride-2 convolution, whose parameter count (ignoring bias) is

$$P_{\text{conv}} = 3 \times 3 \times C \times C = 9C^2. \tag{21}$$

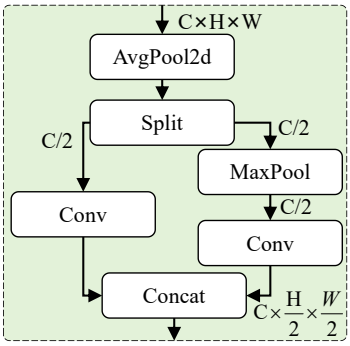


Fig. 4 Illustration of the Adaptive Downsampling (ADown) module, which mitigates information loss during downsampling through a feature-preservation mechanism while substantially reducing computational cost.

In contrast, ADown first splits the feature map into two $C/2$-channel halves and processes them with a 3×3 convolution and a 1×1 convolution, respectively, each producing $C/2$ output channels, while the pooling branches are parameter-free. Its parameter count is thus

$$P_{\text{ADown}} = 3 \times 3 \times \frac{C}{2} \times \frac{C}{2} + 1 \times 1 \times \frac{C}{2} \times \frac{C}{2} = \frac{9}{4}C^2 + \frac{1}{4}C^2 = \frac{5}{2}C^2. \tag{22}$$

Consequently, ADown requires only $P_{\text{ADown}}/P_{\text{conv}} = 5/18 \approx 28\%$ of the parameters of a standard downsampling convolution, i.e., a reduction of about 72% per layer. Aggregated over all downsampling stages of the backbone, this substantially lowers the overall parameter count, while the parallel average- and max-pooling branches retain complementary statistical and salient responses to preserve fine details during resolution reduction.

4 Experiment and Analysis

4.1 Datasets

To comprehensively evaluate the performance and generalization of the proposed model, we conduct systematic experiments on two representative public benchmarks: NEU-DET [78] and GC10-DET [79].

NEU-DET [78] is a standard benchmark for industrial surface defect detection. Developed jointly by Northeastern University (NEU) and several manufacturing partners, it provides genuine defect samples acquired in real production environments with high-resolution imaging systems, ensuring image quality consistent with actual inspection conditions. The dataset covers six fine-grained defect categories—Crazing (Cr), Inclusion (In), Patches (Pa), Pitted Surface (Ps), Rolled-in Scale (Rs), and Scratches (Sc) (Table 1)—with 300 grayscale images of 200×200 pixels per category, totaling 1,800 samples. The taxonomy was validated by quality engineers to comply with industrial standards, facilitating the learning of critical defect patterns. Representative samples are shown in Fig. 5, exhibiting substantial variation in scale, morphology, and

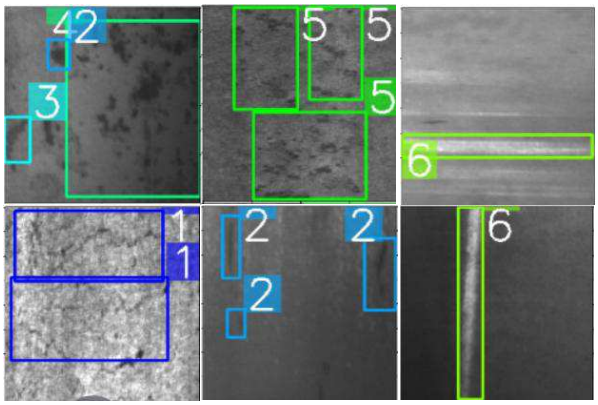


Fig. 5 Representative samples of the six NEU-DET defect categories, with color-coded bounding boxes corresponding to the IDs in Table 1. Each color denotes a specific defect type, and the color–ID mapping is kept consistent throughout the detection pipeline.

Table 1 Instance-level defect distribution of the NEU-DET dataset under the original 1,620/180 image split (training/validation). Counts denote annotated defect instances rather than images.

ID	Color	Defect Category	Train	Validation	Total
1		Crazing (Cr)	550	79	629
2		Inclusion (In)	800	89	889
3		Patches (Pa)	704	99	803
4		Pitted surface (Ps)	347	46	393
5		Rolled-in-scale (Rs)	502	69	571
6		Scratches (Sc)	419	64	483
Total			3,322	446	3,768

spatial distribution. We partition the dataset by stratified random sampling into training and validation sets at a 9:1 ratio (1,620 and 180 images, respectively), preserving balanced class distributions: the training set (90%) is used for parameter optimization and the validation set (10%) for independent performance monitoring. This strategy mitigates class imbalance and, through randomization, prevents data leakage, ensuring reliable evaluation. The dataset also supports multiple annotation formats (e.g., YOLO and COCO), affording compatibility with mainstream deep learning frameworks and flexible experimental design.

GC10-DET [79] was constructed by a research team at Tianjin University from images captured in real industrial production environments for steel surface defect detection. It contains ten common defect types: Waist Folding (Wf), Punching Hole (Pu), Welding Line (Wl), Crescent Gap (Cg), Water Spot (Ws), Oil Spot (Os), Silk Spot (Ss), Inclusion (In), Rolling Pit (Rp), and Crease (Cr). The dataset comprises 2,296 grayscale images of 2048×1000 pixels, split into 1,836 training and 460 validation images. The grayscale modality enhances contrast, suppresses noise, and reinforces

Table 2 Instance-level defect distribution of the GC10-DET dataset under the original 1,836/460 image split (training/validation). Counts denote annotated defect instances rather than images.

ID	Color	Defect Category	Train	Validation	Total
1		Waist Folding (Wf)	10	3	13
2		Punching Hole (Pu)	251	79	330
3		Welding Line (Wl)	404	110	514
4		Crescent Gap (Cg)	214	51	265
5		Water Spot (Ws)	284	71	355
6		Oil Spot (Os)	461	108	569
7		Silk Spot (Ss)	688	196	884
8		Inclusion (In)	265	83	348
9		Rolling Pit(Rp)	71	15	86
10		Crease (Cr)	63	11	74
Total			2,711	727	3,438

detail, mitigating the interference caused by variations in surface reflectance and roughness. The training/validation split is detailed in Table 2.

We train within the YOLO framework using several online augmentation strategies: random horizontal flipping (probability 0.5), random translation (±10%), random scaling (±50%), HSV color jitter (hue ±1.5%, saturation ±70%, value ±40%), mosaic augmentation (probability 1.0), and RandAugment. These augmentations vary the appearance of each original image across epochs and improve generalization; no random augmentation is applied at validation. Tables 1 and 2 report the instance-level defect counts on the original NEU-DET (1,620/180 images) and GC10-DET (1,836/460 images) splits.

4.2 Evaluation Metrics

We report both accuracy and efficiency metrics. Detection accuracy is measured by mean Average Precision at IoU = 0.5 (mAP@50) and the stricter COCO-style mAP@50:95. mAP@50 averages per-category Average Precision (AP) at an Intersection-over-Union (IoU) threshold of 0.5:

$$\text{mAP@50} = \frac{1}{C}\sum_{c=1}^{C}\text{AP@50}(c) \tag{23}$$

where C is the number of categories and AP@50(c) is the AP of category c. mAP@50:95 further averages AP over IoU thresholds from 0.50 to 0.95 with a step of 0.05, and is more sensitive to localization quality.

Efficiency is characterized by the number of parameters (Params, in millions), floating-point operations (FLOPs, in giga), and frames per second (FPS). Unless otherwise stated, FPS is measured on the same NVIDIA GeForce RTX 3060 GPU with batch size 1, including post-processing, under the Ultralytics inference pipeline.

4.3 Implementation Details

All experiments were conducted on dedicated cloud instances to ensure environmental consistency. Each instance was configured with an Intel Xeon Gold 6133 CPU and 30 GB of memory for reliable large-scale data processing, while training and inference were carried out on a single NVIDIA GeForce RTX 3060 GPU with 12 GB of VRAM, which proved sufficient for the modern architectures studied here. The storage subsystem comprised a 20 GB system disk and a 50 GB NVMe SSD data disk for fast access to the training data. GPU acceleration was supported by the CUDA 12.4 toolkit and NVIDIA driver 550.144.0. All experiments were run within the Ultralytics framework, which provides a unified interface for training, validation, and inference. For fairness and reproducibility, all models were trained and evaluated under the same random seed and identical configuration.

The NEU-DET and GC10-DET datasets consist of grayscale industrial defect images that differ substantially from the natural scenes in ImageNet/COCO. Since pretrained features are unlikely to transfer effectively and may even introduce bias, all models are trained from scratch without any external pretrained weights, under a fixed random seed for reproducibility. For YOLO-family models, training follows the YOLO26s recipe for 200 epochs with a batch size of 8 and input images uniformly resized to 640×640 pixels. We use the SGD optimizer with an initial learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005, together with a cosine annealing schedule. To curb overfitting, YOLO's default online augmentations are enabled—random horizontal flipping (probability 0.5), random translation ($\pm$10%), random scaling ($\pm$50%), HSV color jitter (hue $\pm$1.5%, saturation $\pm$70%, value $\pm$40%), mosaic (probability 1.0), and RandAugment—with mosaic disabled during validation. Early stopping is triggered if validation performance does not improve for 20 consecutive epochs. At validation, we enable rectangular inference and multi-scale evaluation and compute all metrics on the validation set. Each model is trained once under the fixed seed, and validation results are reported. Note that the `dfl_loss` field logged by Ultralytics is retained only for legacy compatibility and does not imply an actual DFL module; its value corresponds to the new bounding-box regression loss.

4.4 Comparison with State-of-the-art Methods

As shown in Table 3, we systematically compare Steel-CGA with state-of-the-art defect detection methods on NEU-DET, using mAP@50 as the primary metric.

Steel-CGA achieves 76.6% mAP@50, the best overall result among the compared models, confirming the effectiveness of the integrated architecture in feature extraction and spatial information capture. Relative to the YOLO26s baseline (73.1% mAP@50), our method improves mAP@50 by 3.5 percentage points. Meanwhile, the parameter count falls from 10.0 M to 7.4 M (a reduction of 2.6 M parameters, or 26.0%) and the computational cost from 11.4 G to 8.7 G FLOPs (a 23.7% reduction), and inference reaches 134.0 FPS, evidencing a favorable accuracy–efficiency balance for embedded deployment. These gains stem from three key designs: CSSAM for suppressing complex-background interference via channel–spatial–semantic attention; GDRF

Table 3 Quantitative comparison of steel surface defect detection models on the NEU-DET dataset. Higher mAP@50 and mAP@50:95 indicate better performance; the best results are shown in bold.

Model	Backbone	Params (M)	FLOPs (G)	FPS	mAP@50 (%)	mAP@50:95 (%)
Faster R-CNN[7]	ResNet50	41.4	134	10.1	70.8	35.4
SSD300[11]	vgg16	24.4	31.0	15.6	70.6	33.8
YOLOv7[14]	E-ELAN	7.1	38.5	18.7	64.3	34.1
YOLOv8[80]	CSPDarknet	11.1	14.4	108.8	73.2	39.7
YOLOv9$_s$[15]	CSPDarknet	7.2	13.8	79.4	74.2	38.6
YOLOv10$_s$[81]	CSPDarknet	8.1	12.6	125	64.2	34.6
YOLOv11$_s$[82]	C3k2	9.5	10.9	59.2	73.4	38.9
RTCN$_s$[83]	RDCDLST	-	-	-	67.6	-
GSL-YOLO$_s$[84]	CSPDarknet-Ghost	2.0	5.1	68.8	75.5	-
RT-DETR-I$_s$[85]	ResNet100	32.0	103.5	-	68.7	35.5
SwinYOLO$_s$[77]	GhostNet	4.5	9.9	-	74.9	-
Baseline[16]	C3k2+C2PSA	10.0	11.4	128.0	73.1	38.5
Steel-CGA	C3k2-ADown	7.4	8.7	134.0	**76.6**	**41.4**

for cross-scale context modeling that strengthens small-target responses; and ADown for adaptive downsampling.

Compared with GSL-YOLO$_s$ (75.5% mAP@50), Steel-CGA attains higher accuracy at low parameter complexity. This balance between accuracy and efficiency makes it particularly suited for real-time deployment under limited computational resources.

As shown in Table 4, Steel-CGA achieves the best mAP@50 on three of the six defect categories and remains competitive on the rest. Its core modules—CSSAM, GDRF, and ADown—jointly enhance the detection of subtle and irregular defects through background-aware attention, multi-scale dilated fusion, and adaptive downsampling. The accuracy on the Ps class decreases slightly from 79.0% to 77.8% (a 1.2-point drop): this defect type has highly random textures, low contrast, and weak defect–background boundaries, so the semantic gating in CSSAM, which suppresses background-like responses, may slightly over-attenuate such ambiguous patterns. This indicates that balancing background suppression and fine-grained local cues remains challenging in lightweight architectures. Nevertheless, given the 26.0% reduction in parameters and 23.7% reduction in computation, this minor loss is an acceptable trade-off and points to future optimization for irregular defects. Overall, the balanced architecture delivers consistent improvements across most defect types, demonstrating strong potential for industrial quality-inspection systems.

To further validate the effectiveness and generalization of Steel-CGA, we conduct experiments on the public GC10-DET dataset under identical hardware and hyperparameter settings; the results are reported in Table 5. Steel-CGA again markedly outperforms the YOLO26s baseline in both accuracy and compactness, achieving a 3.6-point gain in mAP@50 (62.0%), a 26.0% reduction in parameters (7.4 M), and a 23.7% reduction in FLOPs (8.7 G). These results show that the proposed model attains higher accuracy on GC10-DET while generalizing robustly across datasets, and

Table 4 Per-category detection performance (mAP@50, %) on the six NEU-DET defect types: Cr (crazing), In (inclusion), P
Ps (pitted surface), Rs (rolled-in scale), and Sc (scratches). The best result per category is shown in bold.

Model	Cr (%)	In (%)	Pa (%)	Ps (%)	Rs(%)	Sc (%)	mAP@50 (%)	mAP@
Faster R-CNN[7]	34.1	84.2	87.6	68.1	55.7	**95.4**	70.8	
SSD300[11]	40.9	69.2	90.6	**91.4**	65.3	66.0	70.6	
YOLOv7[14]	37.9	60.7	75.9	80.7	57.7	80.4	64.3	
YOLOv9$_s$[15]	46.6	83.8	90.8	79.7	67.4	76.9	74.2	
YOLOv10$_s$[81]	39.9	83.4	88.8	61	51.6	60.4	64.2	
YOLOv11$_s$[82]	47.7	**86.0**	92.1	79.5	61.1	73.9	73.4	
RTCN$_s$[83]	41.0	69.7	87.8	72.4	55.1	79.5	67.6	
RT-DETR-I$_s$[85]	29.8	74.9	86.3	77.1	58.8	85.5	68.7	
SwinYOLO$_s$[77]	40.3	82.5	91.2	85.8	60.4	89.1	74.9	
Baseline[16]	50.9	79.4	89.7	79.0	61.3	78.2	73.1	
Steel-CGA	**54.8**	85.7	**92.3**	77.8	**68.5**	80.3	**76.6**	

Table 5 Comparison of steel surface defect detection models on the GC10-DET dataset. Higher mAP@50 and mAP@50:95 in
better performance; the best results are shown in bold.

Model	Backbone	Params (M)	FLOPs (G)	FPS	mAP@50 (%)	mAP@50:95
YOLOv6$_s$[86]	EfficientRep	16.5	22.4	126.1	58.6	30.4
YOLOv7[14]	E-ELAN	7.1	38.5	18.7	58.3	34.3
YOLOv8[80]	CSPDarknet	11.1	14.4	108.8	60.4	31.5
YOLOv10$_s$[81]	CSPDarknet	8.1	12.6	104.2	54.6	28.5
RT-DETR-I$_s$[85]	ResNet100	32.0	103.5	-	41.3	19.4
ELS-YOLO$_s$[87]	C3k2-THK	2.4	5.6	-	54.0	26.7
Baseline[16]	C3k2+C2PSA	10.0	11.4	128.0	58.4	29.9
Steel-CGA	C3k2-ADown	7.4	8.7	134.0	**62.0**	**35.8**

that its balanced optimization of lightweight design and efficiency offers a competitive solution for practical applications.

4.5 Ablation Studies and Analysis

To systematically quantify the contribution of each component, we perform ablation studies under strictly controlled conditions; the results are reported in Table 6. The standalone CSSAM module raises mAP@50 by 0.9 points (73.1%→74.0%) by fusing channel, spatial, and semantic attention to suppress complex-background interference. GDRF alone contributes 0.6 points (73.1%→73.7%), showing that multi-scale context fusion via dilated convolutions strengthens small-target representation. ADown yields a larger gain of 1.1 points through its detail-preserving adaptive downsampling. Notably, combining GDRF and ADown produces a synergistic improvement of 3.2 points (73.1% vs. 76.3%), highlighting their complementary roles for small-defect preservation. Any pairing of CSSAM with the other modules further improves accuracy by reinforcing background suppression. Crucially, integrating all three modules

Table 6 Ablation study of the CSSAM, GDRF, and ADown modules on the NEU-DET dataset, following a controlled-variable protocol. "×" indicates that the corresponding module is excluded from the configuration.

Baseline	CSSAM	GDRF	ADown	mAP@50	mAP@50:95
✓	×	×	×	73.1	38.5
✓	✓	×	×	74.0	39.5
✓	×	✓	×	73.7	40.0
✓	×	×	✓	74.2	37.5
✓	✓	✓	×	74.6	41.5
✓	✓	×	✓	74.4	39.5
✓	×	✓	✓	76.3	41.4
✓	✓	✓	✓	76.6	41.4

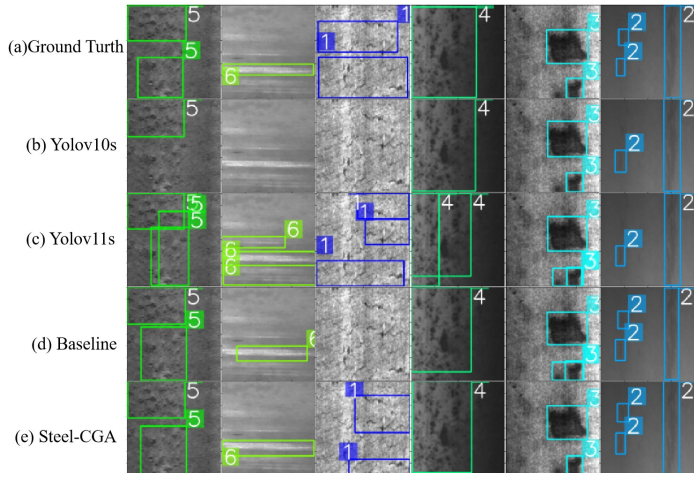


Fig. 6 Qualitative comparison of defect detection on the NEU-DET dataset for the ground truth, YOLOv10s, YOLOv11s, the baseline, and the proposed Steel-CGA.

attains the peak of 76.6% mAP@50, an absolute improvement of 3.5 points, confirming strong inter-module synergy and validating the proposed design.

Beyond these quantitative results, the contributions of individual components are corroborated qualitatively by the more precise defect activations in the heatmaps (Fig. 9) and the reduced misclassifications in the confusion matrices (Fig. 10), together substantiating the rationale and efficacy of our modular design.

4.6 Visualization

Fig. 6 presents a qualitative comparison of detection results on NEU-DET for the ground truth, YOLOv10s, YOLOv11s, the baseline, and Steel-CGA. The comparison reveals clear gains in both accuracy and stability, confirming the leading performance of Steel-CGA on steel surface defect detection. Its strengths manifest along three dimensions. (1) *Superior missed-detection suppression*: as shown in columns 3 and 6,

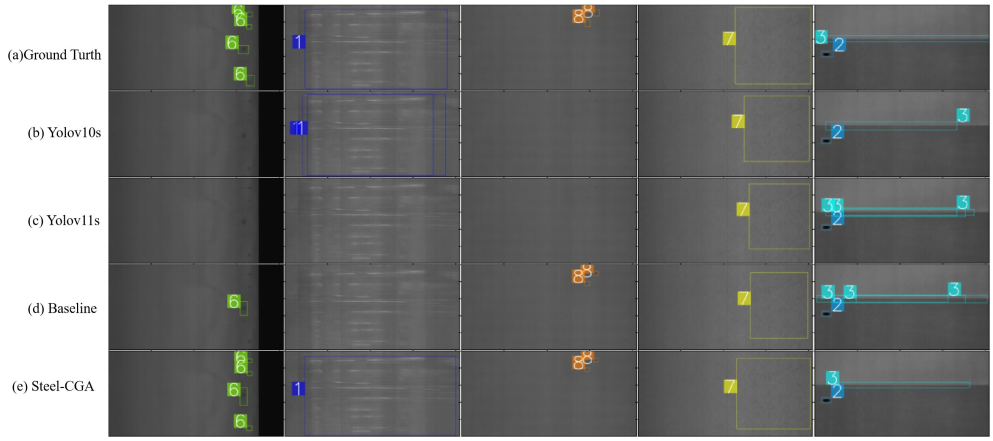


Fig. 7 Qualitative comparison of defect detection on the GC10-DET dataset for the ground truth, YOLOv10s, YOLOv11s, the baseline, and the proposed Steel-CGA. Numeric labels denote the defect categories: 1, Wf; 2, Pu; 3, Wl; 4, Cg; 5, Ws; 6, Os; 7, Ss; 8, In; 9, Rp; 10, Cr.

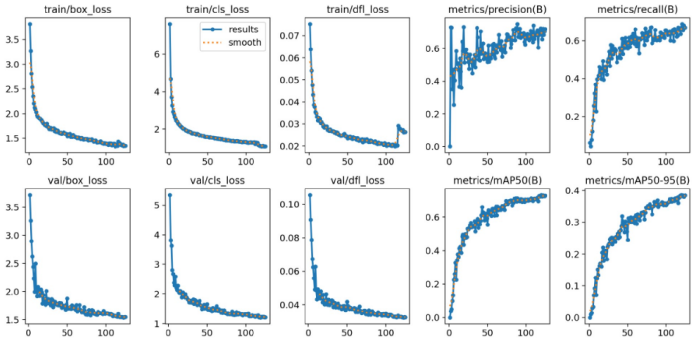
the competing models partially miss Class 1 and Class 2 defects, whereas Steel-CGA identifies them comprehensively through improved feature extraction. (2) *Precise localization*: as shown in columns 1, 2, 4, and 5, our predictions align tightly with the annotations, reducing errors on complex geometries and ensuring reliable region delineation. (3) *Robust feature representation*: by jointly integrating CSSAM, GDRF, and ADown, Steel-CGA extracts discriminative features and maintains stable detection under complex industrial conditions.

To further validate generalization, Fig. 7 shows comparative results on the GC10-DET dataset.

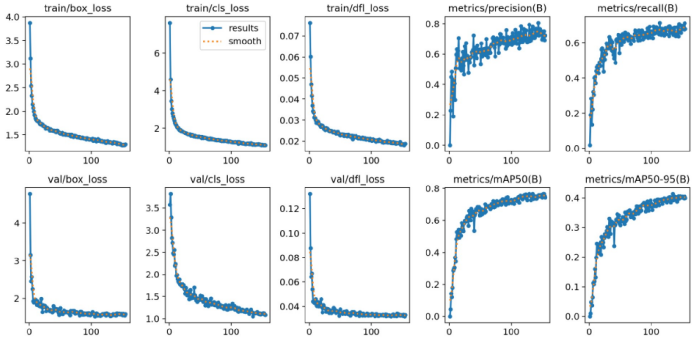
The GC10-DET classes are labeled as 1: Wf, 2: Pu, 3: Wl, 4: Cg, 5: Ws, 6: Os, 7: Ss, 8: In, 9: Rp, 10: Cr. Steel-CGA consistently outperforms the baselines across domains, with three notable advantages. (1) *Strong small-object detection*: as shown in columns 1 and 3 of Fig. 7, our method attains perfect recall on the small Class 6 and Class 8 defects, whereas the baselines miss many instances. (2) *Precise boundary localization*: as shown in column 4, our boxes align markedly better with the annotations than the baselines, especially on geometrically complex defects. (3) *Robustness to cluttered backgrounds*: as shown in columns 2 and 5, our method suppresses repeated false positives and reduces missed instances in complex environments. Across both datasets, Steel-CGA demonstrates consistent reliability in detecting steel surface defects, offering a practical solution for industrial inspection.

Fig. 8 compares Steel-CGA and the baseline through training/validation curves on six key metrics, with the baseline in the upper row and Steel-CGA in the lower row: box loss, classification loss, distribution focal loss, precision, recall, and mAP@50/mAP@50:95.

Steel-CGA outperforms the baseline across these metrics. It attains lower box-regression, classification, and distribution focal losses, leading to more stable and efficient convergence; this benefits from the adaptive channel-group weighting and



(a) Baseline



(b) Steel-CGA

Fig. 8 Training and validation curves of the baseline (top row) and Steel-CGA (bottom row). From left to right: box loss, classification loss, distribution focal loss, precision, recall, mAP@50, and mAP@50:95. Steel-CGA converges faster, attains lower final losses, and consistently achieves superior detection performance across all metrics. Note that the reported dfl_loss corresponds to the new bounding-box regression loss; the name is retained solely for historical compatibility.

semantic gate in CSSAM, which suppress background-induced false positives without added computation. Steel-CGA also achieves higher precision and recall, reflecting stronger suppression of false positives and missed detections, owing to the multi-scale small-target modeling of GDRF and the reduced downsampling information loss of ADown, which jointly preserve global structure and local detail. These improvements culminate in a marked performance gain—76.6% mAP@50 (+3.5 points) and 41.4% mAP@50:95 (+2.9 points)—which, together with the qualitative evidence in Fig. 6, validates the robustness and precision of the proposed framework.

Fig. 9 provides a qualitative comparison of defect localization between the baseline and Steel-CGA. Each example shows the original steel image (left), the baseline activation map (middle), and the Steel-CGA activation map (right).

The heatmap comparison on representative samples highlights three advantages of the proposed approach. First, for irregular defects (e.g., Classes 5 and 6), Steel-CGA exhibits clear focal awareness: as shown in rows 2 and 4, the baseline responses are diffuse with imprecise boundaries, whereas Steel-CGA produces concentrated activations that accurately outline the defect contours. This gain is attributed to the fusion of channel, spatial, and semantic attention in CSSAM, which suppresses background clutter and sharpens defect–background separation, consistent with the reduced false

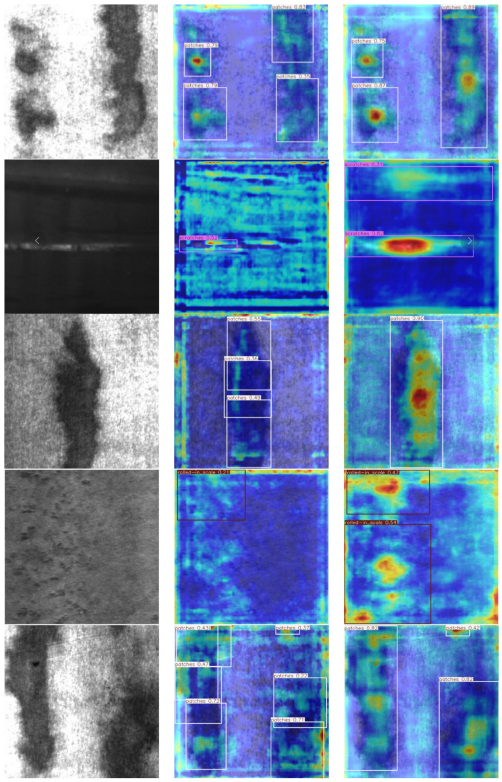


Fig. 9 Visualization of defect-localization heatmaps: (a) original steel image, (b) baseline response, and (c) proposed Steel-CGA response. Steel-CGA produces markedly more precise and concentrated activations within the actual defect regions.

positives in Fig. 6. Second, Steel-CGA localizes small and multi-scale defects more precisely: as shown in rows 3 and 5, the diffuse baseline activations tend to induce repeated and overlapping responses with unnecessary overhead, whereas Steel-CGA yields compact activations tightly aligned with the defect boundaries, owing to the multi-scale receptive-field modeling of GDRF that strengthens small-target responses. Finally, the Steel-CGA heatmaps are clearer and more interpretable, which directly contributes to the higher mAP@50. These qualitative advantages provide strong evidence for the accuracy and robustness of the proposed method.

The confusion-matrix comparison in Fig. 10 shows that Steel-CGA significantly outperforms the baseline on several key metrics, as detailed below.

(1) *Overall classification accuracy.* Under the IoU-0.5 matching criterion (i.e., mAP@50), we define an instance-level detection accuracy to quantify the matching performance of the detection head. Each ground-truth object, each missed detection (false negative), and each false positive is treated as an individual sample; the number of correct matches equals the sum of the diagonal entries of the confusion matrix, and the total equals the sum of all entries. Because correctly rejected background regions (true negatives) vastly outnumber defect regions, including them would dilute the metric and obscure true defect-detection performance; following common practice, we therefore omit true negatives (set to zero) in this matrix. This matrix is used solely

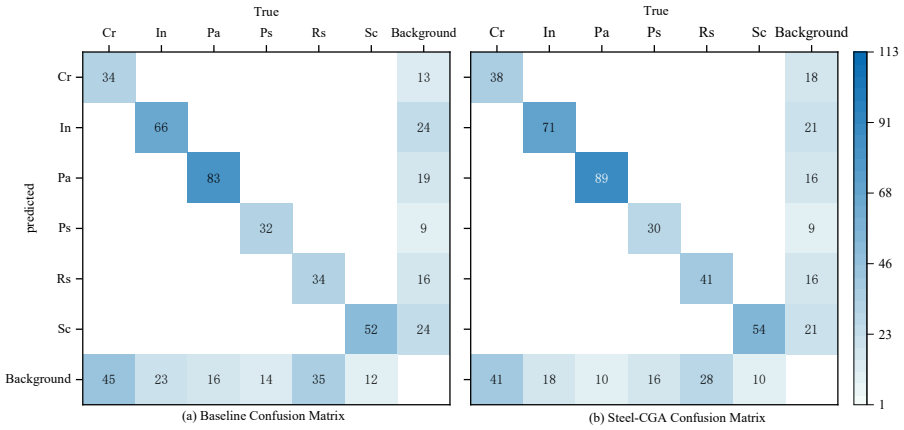


Fig. 10 Confusion-matrix comparison for steel defect detection on the NEU-DET dataset: (a) baseline and (b) proposed Steel-CGA. Steel-CGA increases the correct classifications on the diagonal and reduces off-diagonal errors across all defect types, highlighting its superior discriminative capability.

to compare the matching capability of different modules on defect samples and is not the sole indicator of final performance. Under this scheme, the baseline achieves 301 correct matches (54.63% accuracy), whereas Steel-CGA attains 323 correct matches (59.05%), a gain of 4.42 percentage points. In particular, the precision for In, Pa, Rs, and Sc improves by 3.84, 3.39, 3.93, and 3.58 points, respectively, indicating a higher proportion of true defects among the positive predictions, stronger classification confidence, and substantially fewer false negatives and false positives. We further report the standard mAP@50 and mAP@50:95 metrics as complementary validation, ensuring a comprehensive and consistent evaluation of overall performance.

(2) *Enhanced background-interference suppression.* The baseline misclassifies 145 true defects as background, which Steel-CGA reduces to 123—22 fewer missed detections (a 15.2% relative reduction). This markedly stronger defect-recognition ability arises from the attention mechanism's focus on defect regions, which lowers the chance of false background predictions. Meanwhile, the number of background regions misclassified as defects drops from 105 to 101 (a reduction of 4, or 3.8%); though modest, this further confirms improved background-interference suppression.

(3) *Improved small-object detection.* Cr and Sc are the two smallest defect categories in NEU-DET and are particularly challenging. Steel-CGA raises the recall of Cr from 43.04% to 48.10% (+5.06 points) and that of Sc from 81.25% to 84.38% (+3.13 points), substantially improving detection of small defects, especially the easily missed elongated cracks (Cr). Recall also improves for other categories—In by 5.62 points, Pa by 6.06 points, and Rs by the largest margin of 10.14 points. Overall, Steel-CGA better captures the vast majority of defect types.

Moreover, except for the Ps class, the F1-scores of all five remaining categories improve, with Rs gaining the most (8.0 points), indicating a better precision–recall balance. The weighted-average metrics likewise trend upward, confirming the model's balanced and superior overall performance. Notably, all off-diagonal entries in both confusion matrices are zero, i.e., there are no inter-class misclassifications under the reported matching protocol. This outcome is consistent with our experimental

observations on NEU-DET: the six defect categories differ substantially in texture and morphology, so confusions predominantly occur between defects and background rather than among defect classes. The matrix is constructed by matching predictions to ground-truth instances at IoU = 0.5; unmatched ground truths are counted as background false negatives, and unmatched predictions as background false positives, while class-to-class swaps–if any–would appear off-diagonal. Empirically, neither the baseline nor Steel-CGA produces such swaps on this validation set, so residual errors concentrate on the background row/column. By reducing defects misclassified as background from 145 to 123 and background misclassified as defects from 105 to 101, Steel-CGA strengthens the defect–background decision boundary while preserving zero inter-class confusion. Finally, the most challenging categories with low baseline recall—Rs (49.28%) and Cr (43.04%)—both achieve substantial gains, confirming that CSSAM and GDRF effectively extract features from defects with complex textures and irregular shapes.

In summary, Steel-CGA consistently outperforms the baseline in overall accuracy, missed-detection rate, small-object detection, and per-category recall and F1-score, including on the most challenging categories. This advantage stems from the background-interference suppression of CSSAM together with the multi-scale small-target modeling of GDRF, which jointly enhance defect feature representation. Although the precision for Cr and Ps decreases slightly, Steel-CGA retains a comprehensive overall lead.

5 Conclusion

This paper presented Steel-CGA, a YOLO26s-based detector for steel surface defect recognition. CSSAM fuses channel, spatial, and semantic gated attention to suppress complex-background interference; GDRF enlarges multi-scale receptive fields to strengthen small-target responses; and ADown reduces spatial information loss under lightweight downsampling. On NEU-DET and GC10-DET, Steel-CGA achieves 76.6% and 62.0% mAP@50, outperforming YOLO26s by 3.5 and 3.6 points, while reducing parameters by 26.0% and FLOPs by 23.7% and running at 134 FPS. These results confirm a favorable accuracy–efficiency trade-off for real-time industrial inspection.

Several limitations remain. First, NEU-DET and GC10-DET are relatively small (1,800 and 2,296 images), so broader industrial generalization is not fully established. Second, mAP@50:95 (41.4% / 35.8%) still lags behind mAP@50, indicating room to improve localization under stricter IoU thresholds. Third, all experiments use static images, and robustness to severe illumination changes and noise on production lines requires further study. Future work will enlarge and diversify steel defect data, strengthen boundary regression, and explore compression for edge deployment.

Acknowledgements. This reseach was funded by the Shandong Youth University of Political Science.

Declarations

Competing interests The authors declare no conflicts of interest.

Data availability The raw data supporting the conclusions of this article will be made available by the authors on request.
Author contributions J.J.C. conceived the experiment(s), J.J.C. and X.H.C. conducted the experiment(s), J.J.C., B.L.W. and S.W. analysed the results. All authors reviewed the manuscript.
Funding This study was funded by the Shandong Youth University of Political Science Doctoral Research Startup Fund (XXPY23036).
Additional information Correspondence and requests for materials should be addressed to S.W.

Editorial Policies for:

Springer journals and proceedings: https://www.springer.com/gp/editorial-policies

Nature Portfolio journals: https://www.nature.com/nature-research/editorial-policies

Scientific Reports: https://www.nature.com/srep/journal-policies/editorial-policies

BMC journals: https://www.biomedcentral.com/getpublished/editorial-policies

References

[1] Sun, J., Sun, M., Guo, H., Ji, J., Xu, L., Peng, Y.: Comprehensive prediction of surface quality of hot-rolled strip with multiple and single defects. Iron Steel **56**(1), 51–58 (2021)

[2] Li, Y., Li, Y., Liu, J., Fan, Z., Wang, Q.: Research on segmentation of steel surface defect images based on improved res-unet network. Journal of Electronics & Information Technology **44**(5), 1513–1520 (2022)

[3] Zhang, L., Li, X., Sun, Y., Feng, Y., Guo, H.: Multiscale feature fusion for salient object detection of strip steel surface defects. IEEE Access (2025)

[4] Li, H., Zhang, R., Pan, Y., Ren, J., Shen, F.: Lr-fpn: Enhancing remote sensing object detection with location refined feature pyramid network. In: 2024 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2024). IEEE

[5] Weng, W., Wei, M., Ren, J., Shen, F.: Enhancing aerial object detection with selective frequency interaction network. IEEE Transactions on Artificial Intelligence **5**(12), 6109–6120 (2024)

[6] Liu, Y., Shen, J., Ye, R., Wang, S., Ren, J., Pan, H.: Fp-deeplab: a segmentation model for fabric defect detection. Measurement Science and Technology **35**(10), 106008 (2024)

[7] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 1137–1149 (2015)

[8] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. IEEE transactions on pattern analysis and machine intelligence **42**(2), 386–397 (2018)

[9] Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–6162 (2018)

[10] Zhang, D., Song, K., Xu, J., He, Y., Yan, Y.: Unified detection method of aluminium profile surface defects: Common and rare defect categories. Optics and Lasers in Engineering **126**, 105936 (2020)

[11] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37 (2016). Springer

[12] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)

[13] Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)

[14] Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7464–7475 (2023)

[15] Wang, C.-Y., Yeh, I.-H., Mark Liao, H.-Y.: Yolov9: Learning what you want to learn using programmable gradient information. In: European Conference on Computer Vision, pp. 1–21 (2024). Springer

[16] Jocher, G., Qiu, J., Liu, M., Lyu, S., Akyon, F.C., Kalfaoglu, M.E.: Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models (2026). https://arxiv.org/abs/2606.03748

[17] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16965–16974 (2024)

[18] Bastidas, A.A., Tang, H.: Channel attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)

[19] Nam, H., Ha, J.-W., Kim, J.: Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 299–307 (2017)

[20] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

[21] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)

[22] Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)

[23] Zhou, M., Lu, W., Xia, J., Wang, Y.: Defect detection in steel using a hybrid attention network. Sensors **23**(15), 6982 (2023)

[24] Huang, B., Liu, J., Liu, X., Liu, K., Liao, X., Li, K., Wang, J.: Improved YOLOv5 network for steel surface defect detection. Metals **13**(8), 1439 (2023)

[25] Lu, J., Yu, M., Liu, J.: Lightweight strip steel defect detection algorithm based on improved yolov7. Scientific reports **14**(1), 13267 (2024)

[26] Zhang, C., Zhang, X.: Multi-target domain-based hierarchical dynamic instance segmentation method for steel defects detection. Neural Computing and Applications **35**(10), 7389–7406 (2023)

[27] Tian, R., Jia, M.: Dcc-centernet: A rapid detection method for steel surface defects. Measurement **187**, 110211 (2022)

[28] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)

[29] Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters–improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4353–4361 (2017)

[30] Wang, L., Liu, X., Ma, J., Su, W., Li, H.: Real-time steel surface defect detection with improved multi-scale YOLO-v5. Processes **11**(5), 1357 (2023)

[31] Xie, W., Sun, X., Ma, W.: A light weight multi-scale feature fusion steel surface defect detection model based on yolov8. Measurement Science and Technology **35**(5), 055017 (2024)

[32] Zhang, T., Pang, H., Jiang, C.: GDM-YOLO: A model for steel surface defect detection based on YOLOv8s. IEEE Access **12**, 148817–148825 (2024)

[33] Zeng, Z., Li, L., Zhao, Z., Liu, Q.: Improved fine-grained image classification in few-shot learning based on channel-spatial attention and grouped bilinear convolution. The Visual Computer **41**(6), 4129–4141 (2025)

[34] Zeng, N., Wu, P., Wang, Z., Li, H., Liu, W., Liu, X.: A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection. IEEE Transactions on Instrumentation and Measurement **71**, 1–14 (2022)

[35] Shen, K., Zhou, X., Liu, Z.: Minet: Multiscale interactive network for real-time salient object detection of strip steel surface defects. IEEE Transactions on Industrial Informatics **20**(5), 7842–7852 (2024)

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

[37] Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)

[38] Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016)

[39] Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10781–10790 (2020)

[40] Peng, L., Cao, Y., Sun, Y., Wang, Y.: Lightweight adaptive feature de-drifting for compressed image classification. IEEE Transactions on Multimedia **26**, 6424–6436 (2024)

[41] Qin, Q., Khairuddin, A.S.M., Idros, N., Liu, H., Fan, L., Yang, Z., Li, J., Zhu, C.: Hierarchical depth aware yolo for efficient metal surface defect detection. Scientific Reports **16**(1), 16248 (2026) https://doi.org/10.1038/s41598-026-46074-z

[42] Xing, H., Yang, Z.: Yolo-dcf: dual distillation and context-aware fusion for defect detection. Scientific Reports **16**(1), 15897 (2026) https://doi.org/10.1038/s41598-026-46602-x

[43] Zhou, H., Zou, H., Hu, G.: An efficient and lightweight algorithm for detecting surface defects of steel based on scci-yolo. Scientific Reports **15**(1), 36276 (2025) https://doi.org/10.1038/s41598-025-20154-y

[44] Feng, Y., Jin, L., Yang, H., Liu, S.: An improved steel defect detection model using multi-scale feature fusion based on yolo-mfd. Scientific Reports **16**(1), 313 (2025) https://doi.org/10.1038/s41598-025-29596-w

[45] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 764–773 (2017)

[46] Zhang, W., Liu, J., Yan, Z., Zhao, M., Fu, X., Zhu, H.: Fc-yolo: an aircraft skin defect detection algorithm based on multi-scale collaborative feature fusion. Measurement Science and Technology **35**(11), 115405 (2024)

[47] Jian, T.S., Fauadi, M.H.F.M., Yahaya, S.H., Noor, A.Z.M., Saptari, A.: A deep learning approach for automated pcb defect detection: A comprehensive review. Multidisciplinary Reviews **8**(1), 2025011–2025011 (2025)

[48] Li, C., Xu, A., Zhang, Q., Cai, Y.: Steel surface defect detection method based on improved yolox. IEEE Access **12**, 37643–37652 (2024)

[49] Ren, F., Fei, J., Li, H., Doma, B.T.: Steel surface defect detection using improved deep learning algorithm: Eca-simsppf-siou-yolov5. IEEE Access **12**, 32545–32553 (2024)

[50] Falaschetti, L., Manoni, L., Palma, L., Pierleoni, P., Turchetti, C.: Embedded real-time vehicle and pedestrian detection using a compressed tiny yolo v3 architecture. IEEE Transactions on Intelligent Transportation Systems (2024)

[51] Liu, R., Zhao, L., Ren, Y., Shen, Z., Li, L., Luo, J., Abbas, Z.: A lightweight model based on multi-scale feature fusion for ultrasonic welding surface defect detection. Engineering Applications of Artificial Intelligence **161**, 112208 (2025)

[52] Li, S., Guo, S., Han, Z., Kou, C., Huang, B., Luan, M.: Aluminum surface defect detection method based on a lightweight yolov4 network. Scientific reports **13**(1), 11077 (2023)

[53] Liu, Y., He, J., Liang, J., Wang, W., Wang, Z.: A new improved particle swarm algorithm for optimization of anchor lattice beam support structures. Natural Hazards **121**(1), 225–243 (2025)

[54] Zhou, S., Wang, X., Yang, M., Song, W.: Multi-view clustering with adaptive anchor and bipartite graph learning. Neurocomputing **611**, 128627 (2025)

[55] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)

[56] Wu, Z., Zhang, Y., Wang, X., Li, H., Sun, Y., Wang, G.: Algorithm for detecting

surface defects in wind turbines based on a lightweight yolo model. Scientific Reports **14**(1), 24558 (2024)

[57] Wang, J., Xie, X., Liu, G., Wu, L.: A lightweight pcb defect detection algorithm based on improved yolov8-pcb. Symmetry **17**(2), 309 (2025)

[58] 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, pp. 13713–13722 (2021)

[59] Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0 (2019)

[60] Liu, G., Ma, Q.: Strip steel surface defect detecting method combined with a multi-layer attention mechanism network. Measurement Science and Technology **34**(5), 055403 (2023)

[61] Wang, S., Fan, X., Zhao, Z., Qiao, Z., Wang, N., Qiu, Y., Jia, X.: Research on steel surface defect detection system based on yolov5s-se-ca model and bemd image enhancement. Nondestructive Testing and Evaluation **40**(7), 3080–3099 (2025)

[62] Xin, H., Zhang, K.: Surface defect detection with channel-spatial attention modules and bi-directional feature pyramid. IEEE Access **11**, 88960–88970 (2023)

[63] Liu, H., Chen, C., Hu, R., Bin, J., Dong, H., Liu, Z.: Cgtd-net: Channel-wise global transformer-based dual-branch network for industrial strip steel surface defect detection. IEEE Sensors Journal **24**(4), 4863–4873 (2024)

[64] Mi, Z., Gao, Y., Xu, X., Tang, J.: Steel strip surface defect detection based on multiscale feature sensing and adaptive feature fusion. AIP Advances **14**(4) (2024)

[65] Yang, J., Liu, Z.: A novel real-time steel surface defect detection method with enhanced feature extraction and adaptive fusion. Engineering Applications of Artificial Intelligence **138**, 109289 (2024)

[66] Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768 (2018)

[67] Ghiasi, G., Lin, T.-Y., Le, Q.V.: Nas-fpn: Learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7036–7045 (2019)

[68] Wei, H., Liu, X., Xu, S., Dai, Z., Dai, Y., Xu, X.: Dwrseg: Rethinking efficient acquisition of multi-scale contextual information for real-time semantic

segmentation. arXiv preprint arXiv:2212.01173 (2022)

[69] Liu, S., Huang, D., Wang, Y.: Learning spatial fusion for single-shot object detection. arXiv preprint arXiv:1911.09516 (2019)

[70] Yu, T., Luo, X., Li, Q., Li, L.: CRGF-YOLO: An optimized multi-scale feature fusion model based on YOLOv5 for detection of steel surface defects. International Journal of Computational Intelligence Systems **17**(1), 154 (2024)

[71] Zhao, H., Wan, F., Lei, G., Xiong, Y., Xu, L., Xu, C., Zhou, W.: LSD-YOLOv5: A steel strip surface defect detection algorithm based on lightweight network and enhanced feature fusion mode. Sensors **23**(14), 6558 (2023)

[72] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)

[73] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C.: Ghostnet: More features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1580–1589 (2020)

[74] Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic convolution: Attention over convolution kernels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11030–11039 (2020)

[75] Yuan, Z., Ning, H., Tang, X., Yang, Z.: GDCP-YOLO: Enhancing steel surface defect detection using lightweight machine learning approach. Electronics **13**(7), 1388 (2024)

[76] Chu, Y., Yu, X., Rong, X.: A lightweight strip steel surface defect detection network based on improved yolov8. Sensors (Basel, Switzerland) **24**(19), 6495 (2024)

[77] Wang, Q., Dong, H., Huang, H.: Swin-transformer-yolov5 for lightweight hot-rolled steel strips surface defect detection algorithm. Plos one **19**(1), 0292082 (2024)

[78] Song, K., Yan, Y.: A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. Applied Surface Science **285**, 858–864 (2013)

[79] Lv, X., Duan, F., Jiang, J.-j., Fu, X., Gan, L.: Deep metallic surface defect detection: The new benchmark and detection network. Sensors **20**(6), 1562 (2020)

[80] Varghese, R., Sambath, M.: Yolov8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), pp.

1–6 (2024). IEEE

[81] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., *et al.*: Yolov10: Real-time end-to-end object detection. Advances in Neural Information Processing Systems **37**, 107984–108011 (2024)

[82] Khanam, R., Hussain, M.: Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725 (2024)

[83] Hou, S., He, H., Peng, K., Qiao, S.: Improved swin transformer-based model for hot-rolled strip defect detecting. Computing and Informatics **43**(6), 1352–1371 (2024)

[84] Wang, Z., Jiang, Q., Yan, X.: Steel surface defect detection algorithm based on gsl-yolo. Journal of East China University of Science and Technology (2025)

[85] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16965–16974 (2024)

[86] Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., et al.: Yolov6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976 (2022)

[87] Zhang, Z., Zhong, G., Ding, P., He, J., Zhang, J., Zhu, C.: Els-yolo: efficient lightweight yolo for steel surface defect detection. Electronics **14**(19), 3877 (2025)