electronics

MDPI

Article

Steel Surface Defect Detection Based on Improved YOLOv8 with Multi-Scale Feature Fusion and Attention Mechanism

Yalei Jia [1], Xian Zhang [1], Jianhui Meng [2] and Jisong Zang [3,*]

[1] Hebei Key Laboratory of Trans-Media Aerial Underwater Vehicle, School of Aeronautics and Astronautics, North China Institute of Aerospace Engineering, Langfang 065000, China; yalei_jia@nciae.edu.cn (Y.J.); 15150020102@139.com (X.Z.)
[2] State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources, North China Electric Power University, Baoding 071000, China; jhmeng@ncepu.edu.cn
[3] School of Mechanical and Electrical Engineering, North China Institute of Aerospace Engineering, Langfang 065000, China
* Correspondence: zjs123456@nciae.edu.cn; Tel.: +86-132-3069-2223

Abstract

Identifying microscopic textural anomalies and filtering out complicated industrial background noise remain significant hurdles in inspecting metallic surfaces. To tackle these operational bottlenecks, our research introduces a refined multi-scale detection framework built upon the YOLOv8l architecture. Specifically, we engineer a fine-grained detection pathway utilizing the P2 layer, which aims to preserve critical details of miniature flaws that are otherwise discarded during feature extraction. Furthermore, a Bi-directional Feature Pyramid Network model is embedded to reconstruct the feature fusion path, balancing the preservation of shallow geometric textures with enhanced multi-scale representation capabilities. To bolster anti-interference performance, a Convolutional Block Attention Module (CBAM) is integrated prior to the detection head, employing adaptive channel and spatial weighting to suppress unstructured background noise. Experimental results utilizing TTA demonstrate that the mAP@0.5 reached 76.3%. Detection accuracies for patches and inclusions reached 93.1% and 85.3%.

Keywords: YOLOv8; small object detection; BiFPN; CBAM; attention mechanism

1. Introduction

Steel strips are ubiquitous in various manufacturing sectors, notably aerospace, shipbuilding, and automotive industries. However, surface defects on these strips are inevitable due to various factors, including processing and environmental conditions [1]. These defects pose serious safety hazards. Therefore, achieving high-precision and automated surface defect detection is important for intelligent manufacturing [2,3]. Undetected anomalies can cause severe structural failures and economic losses. Additionally, advanced concepts like saliency-guided enhancement [4] offer valuable inspiration for highlighting weak targets in complex backgrounds.

Deep learning has revolutionized this field nowadays [5]. While two-stage algorithms represented by Faster R-CNN [6] offer high accuracy, their inference latency is often too high to meet the real-time requirements [7] of high-speed rolling lines. Consequently, single-stage algorithms, particularly the YOLO series [8,9], have become main solutions for industrial inspection due to their balance between speed and accuracy. Among them, YOLOv8 [10] stands out as a state-of-the-art baseline, and the architecture relies on deep

Check for updates

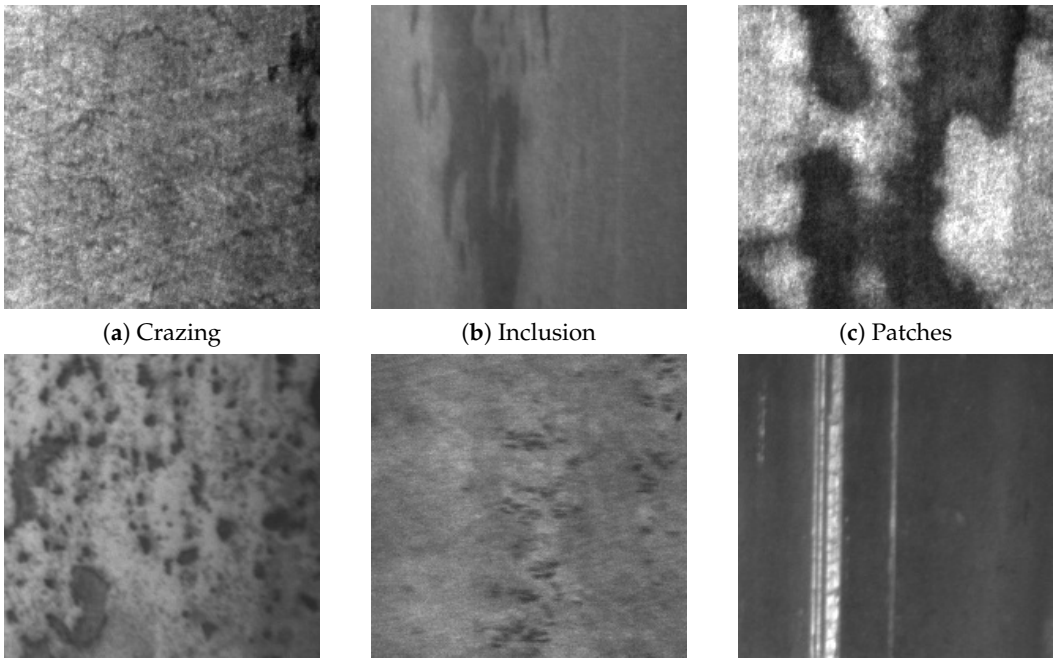
Academic Editor: Silvia Liberata Ullo

Received: 11 February 2026
Revised: 18 March 2026
Accepted: 23 March 2026
Published: 27 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

convolutional networks [11] for feature extraction. However, the standard YOLOv8 architecture still faces challenges in industrial defect detection tasks. The repeated downsampling operations in the backbone may lead to the loss of fine-grained spatial information, which is critical for detecting small defects such as crazing and inclusions.

Yu et al. improved the original YOLO architecture for real-time detection of steel surface defects [12], while others explored transformer-based methods [13] and advanced data augmentation strategies [14]. Recently, the application of large vision-language models (VLMs) optimized by reinforcement learning has also emerged as a promising paradigm for handling imbalanced fault detection in complex industrial environments, such as CNC machining [15]. However, for high-speed steel rolling lines, deploying massive multimodal models remains computationally prohibitive, making highly optimized lightweight CNNs more practical. In addition to conventional optical methods, recent studies have also begun to explore the potential of hyperspectral imaging and Generative Adversarial Networks (GANs) for complex texture classification and data augmentation [16,17]. Attention mechanisms have also been widely adopted to improve feature extraction [18,19]. Advanced feature fusion strategies, such as semantic compensated adaptive fusion networks originally designed for remote sensing [20], have demonstrated the critical importance of dynamically aligning multi-scale semantic information to suppress background variations. However, detecting micro-defects remains difficult [21,22]. In steel surface defect detection, the typical characteristics of defects are irregular shapes and relatively small proportions of corresponding structures, which pose a challenge for conventional detection models [23]. Typical examples of these surface defects are shown in Figure 1.



(**a**) Crazing (**b**) Inclusion (**c**) Patches

(**d**) Pitted Surface (**e**) Rolled-in Scale (**f**) Scratches

Figure 1. Challenges in steel surface defect detection. (**a**) Crazing presents as minute, low-contrast textures easily lost during downsampling, while (**c**) Patches and (**e**) Rolled-in Scale share similar visual characteristics with oil stains and water spots, complicating feature extraction.

To address these limitations, this study proposes an improved multi-scale defect detection algorithm based on the YOLOv8l. This study makes the following contributions:

P2: To mitigate the feature loss of tiny defects caused by deep network downsampling, we construct a high-resolution detection head originating from the second layer of the

backbone. This branch effectively preserves shallow geometric textures, significantly enhancing the perception of micro-targets such as crazing.

BiFPN: We introduce a BiFPN module to replace the original PANet. By employing learnable weights in the model, this structure enables adaptive fusion of multi-scale features, balancing the representation of large patches and minute cracks.

CBAM: To bolster anti-interference capabilities, a Convolutional Block Attention Module is integrated prior to the detection head. Through adaptive weighting in both channel and spatial dimensions, this module suppresses unstructured background noise and focuses the model's attention on defect regions.

2. Materials and Methods

2.1. Dataset and Preprocessing

The NEU-DET steel surface defect dataset [24] was selected as the benchmark. This dataset contains 1800 grayscale images covering six typical defects: crazing, inclusion, patches, pitted surface, rolled-in scale, and scratches. Each category contains 300 samples with an original resolution of 200 × 200 pixels.

The dataset was randomly divided into a training set and a validation set in an 8:1:1 ratio. Considering the low resolution of the original images, and to fully activate the fine-grained perception capability of the proposed P2 high-resolution detection layer, all images were upsampled to 1280 × 1280 pixels during the preprocessing stage. Although interpolation does not introduce new information, the higher spatial resolution allows the P2 detection head to better preserve fine-grained structural patterns of micro-defects.

2.2. The Improved YOLOv8 Architecture

This study proposes the YOLOv8-BiFPN-P2-CBAM architecture. Building upon the YOLOv8l baseline, the improved network introduces three key modifications: (1) a high-resolution P2 detection head extracted from the shallow backbone; (2) BiFPN for multi-scale fusion; and (3) CBAM integrated before the detection heads. The overall topology of the proposed network is illustrated in Figure 2.

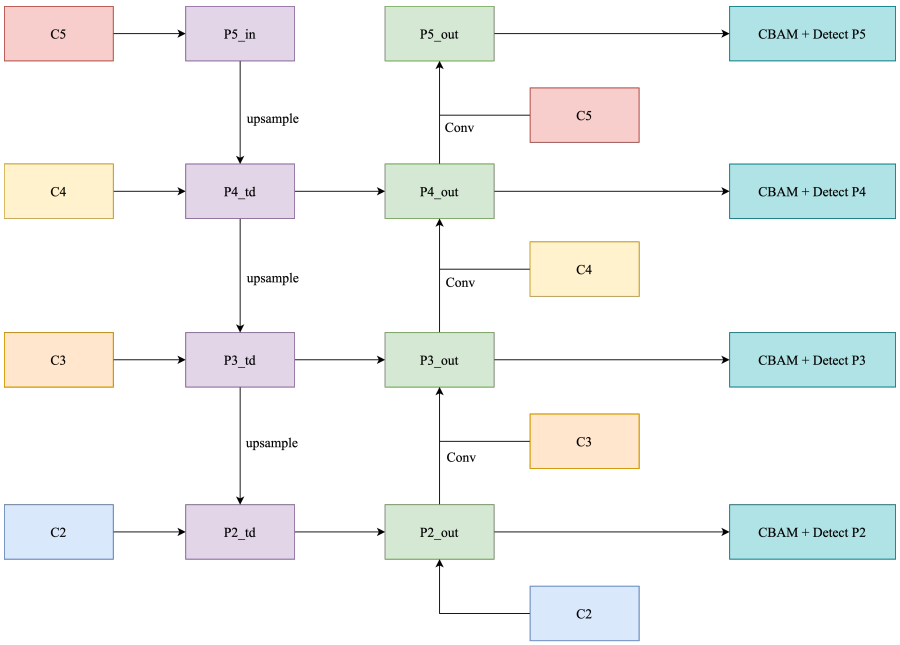
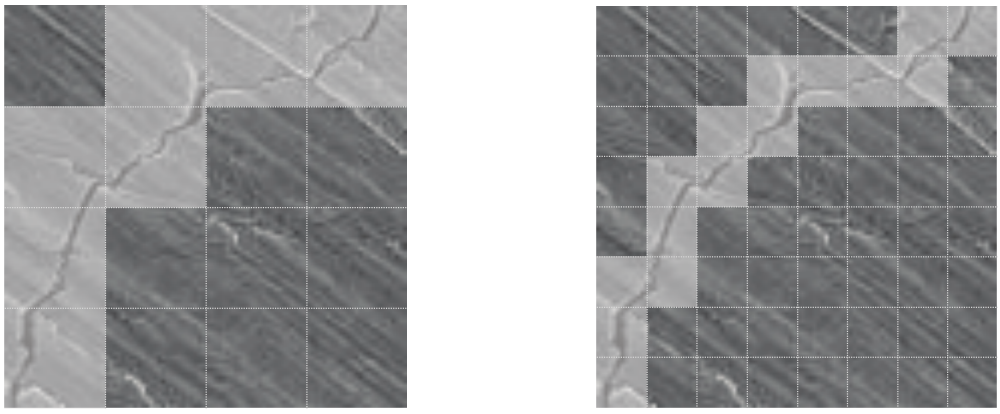


Figure 2. Architecture of the proposed YOLOv8-BiFPN-P2-CBAM. The model incorporates a P2 high-resolution branch, BiFPN weighted fusion, and CBAM attention modules to enhance small defect detection and noise suppression.

2.3. High-Resolution Feature Detection Branch (P2)

Certain flaw categories within the NEU-DET benchmark, particularly crazing, manifest as extremely thin lines (merely 1 to 3 pixels wide) with minimal visual contrast against the background. Because the standard YOLOv8 architecture yields the P3 level as its finest output grid, textural patterns occupying fewer than 8×8 pixels are frequently discarded or blended into ambient noise during successive downsampling operations [25,26].

To mitigate this, we constructed a high-resolution detection branch (P2) originating from the second stage of the CSPDarknet backbone. As shown in Figure 3, the P2 layer possesses a denser grid and a smaller receptive field compared to P3, enabling it to retain high-frequency geometric information essential for detecting micro-defects.



(**a**) P3 Layer Output (**b**) P2 Layer Output

Figure 3. Feature preservation comparison. (**a**) The P3 layer has a larger receptive field but loses fine details. (**b**) The P2 layer retains primitive edge and texture information, which is critical for identifying micro-cracks.

2.4. Weighted Bi-Directional Feature Pyramid Network (BiFPN)

In steel surface inspection, high-level semantic features and low-level spatial features play different roles in detecting large defects and small inclusions. As shown in Figure 4, the BiFPN structure is an improved variant of the traditional Feature Pyramid Network (FPN) [27]. While the traditional FPN provides a top-down pathway and PANet [28] adds an extra bottom-up path, BiFPN [29] optimizes the cross-scale connections by removing nodes with only one input edge and adding extra weights. Therefore, the weighted fusion splicing module, BiFPN, is incorporated into the neck network to facilitate multi-scale feature detection and fusion. This improves detection precision [30].

The weighted fusion is calculated as:

$$O = \sum_i \frac{w_i}{\epsilon + \sum_j w_j} \cdot I_i \tag{1}$$

where O is the fused output, I_i is the input feature, and w_i is the initial learnable scalar weight during training. While traditional BiFPN relies solely on these simple scalar weights, our proposed architecture structurally elevates this mechanism. By immediately coupling the BiFPN multi-scale outputs with the subsequent CBAM module (detailed in Section 2.5), we upgrade the scalar weighting to a high-dimensional, channel-wise feature recalibration. This synergistic design allows the network to automatically and precisely balance the representation of multi-scale defects, a highly adaptive behavior that will be quantitatively analyzed in Section 3.5.3.

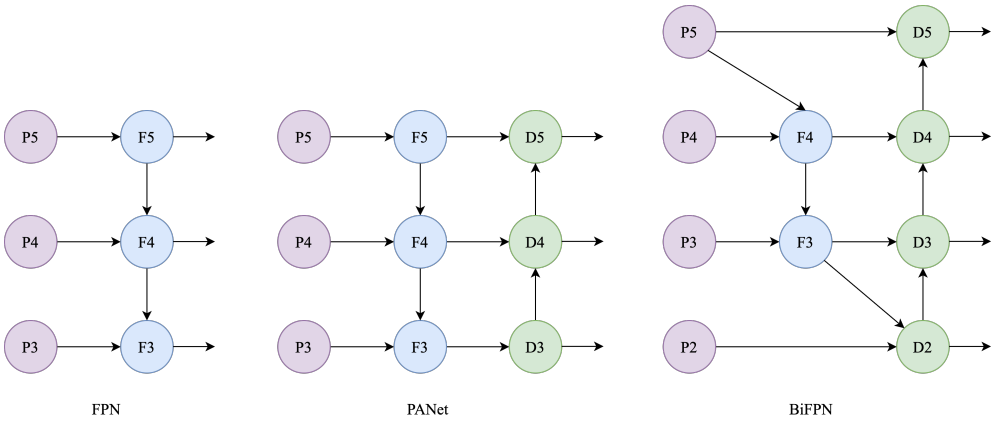


Figure 4. Structural comparison of FPN, PANet, and BiFPN. BiFPN introduces bidirectional cross-scale connections and removes less-contributing nodes to achieve more efficient multi-scale feature fusion.

2.5. CBAM Attention Mechanism

To suppress non-structural background noise that resembles defects like rolled-in scale, we integrated the CBAM [31] prior to the detection heads. Compared to SE-Net [32] and ECA-Net [33], CBAM sequentially infers attention maps along two dimensions. Assuming an intermediate feature tensor denoted as F, the sequential recalibration procedure is mathematically defined as follows:

$$\begin{aligned} F' &= M_c(F) \otimes F \\ F'' &= M_s(F') \otimes F' \end{aligned} \tag{2}$$

where $\otimes$ denotes element-wise multiplication. F' represents the feature map refined by channel attention, and F'' is the final output refined by spatial attention. The overall architecture of the CBAM is illustrated in Figure 5.

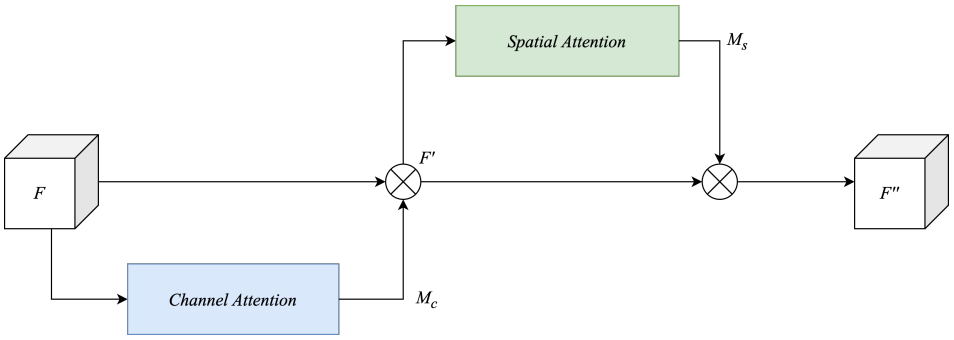


Figure 5. The overall architecture of the CBAM. The input feature F is sequentially refined by the Channel Attention Module (M_c) and the Spatial Attention Module (M_s).

To explicitly illustrate the internal operations of these two complementary sub-modules, Figures 6 and 7 depict their detailed mathematical flows and physical architectures.

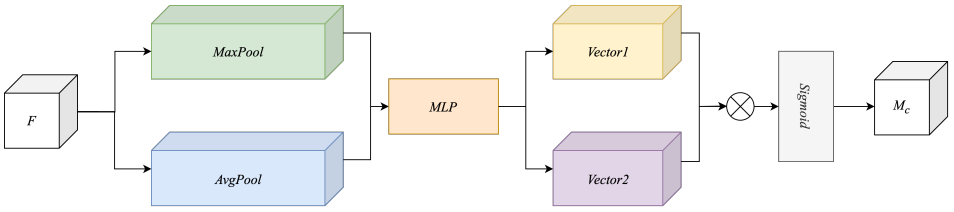


Figure 6. Detailed internal architecture of the Channel Attention Module in CBAM. It utilizes both MaxPooling and AvgPooling to aggregate spatial information, which is then processed by a shared MLP.

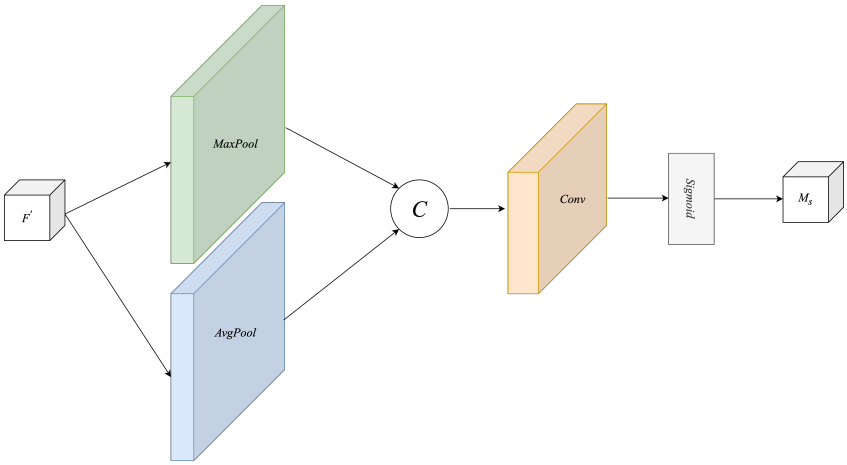


Figure 7. Detailed internal architecture of the Spatial Attention Module in CBAM. It applies pooling operations along the channel axis and concatenates them for a standard convolution.

The specific functions of each module are as follows:

1. Channel Attention: As shown in Figure 6, it focuses on what constitutes a defect by aggregating spatial information via average-max pooling. It then re-calibrates channel weights through a shared Multi-Layer Perceptron (MLP), effectively distinguishing defect textures from background patterns.
2. Spatial Attention: As shown in Figure 7, it focuses on where the defect is located. It aggregates spatial information by applying average-max pooling along this channel, followed by a convolutional layer to generate a spatial attention map. This map emphasizes defect-related regions while suppressing background noise, thereby improving the localization capability for small and subtle defects.

This serial "channel-then-spatial" attention process effectively filters out irrelevant background information before the final prediction.

2.6. Implementation Details

The proposed model was trained and evaluated on a computing server using eight NVIDIA GeForce RTX 4090 GPUs, each with 24 GB memory. The software environment utilized PyTorch 1.13.1 and CUDA 11.7. The model was trained using the SGD. For the refinement of spatial coordinate predictions, the Distance-IoU (DIoU) metric was employed [34], while we adopted Focal Loss [35] to mitigate the adverse effects of uneven category distributions across the training samples. The model was trained for a total of 200 epochs with a batch size of 32, utilizing an initial learning rate of 0.0005 with cosine annealing decay. Unless otherwise specified, the default random seed settings provided by the training framework were used in all experiments to ensure reproducibility. Grayscale images were converted to three-channel RGB to align with the pre-trained backbone, and Mosaic augmentation was applied during training to prevent overfitting.

3. Results

This section provides an evaluation of the YOLOv8-BiFPN-P2-CBAM algorithm. The experimental results are analyzed from three dimensions: quantitative comparison with the baseline model, qualitative visual analysis of detection performance, and a comprehensive ablation study to verify the contribution of each improved module.

3.1. Quantitative Comparative Analysis

The proposed method was compared with the baseline YOLOv8l model on the NEU-DET test set. The evaluation metrics include the Average Precision (AP) for each of the six

defect categories in the dataset and the mean Average Precision (mAP) at an IoU threshold of 0.5.

As summarized in Table 1, the proposed algorithm achieves a significant improvement in overall performance. The mAP@0.5 reached 76.3%, surpassing the baseline model by 6.5%. Furthermore, the high-precision localization metric (mAP@0.5–0.95) increased from 26.5% to 41.1%, indicating that the predicted bounding boxes align more tightly with the ground truth. Specific improvements are observed in challenging categories:

- Minute Defects: The detection accuracy for Crazing, which consists of extremely fine textures, increased dramatically from 38.5% to 57.8%. This validates the effectiveness of the proposed P2 high-resolution branch.
- Scale-Variant Defects: Major defects like Patches and Inclusion achieved remarkably high APs of 93.1% and 85.3%, respectively.

Table 1. Quantitative comparison of detection performance between the baseline YOLOv8l and the proposed method. The best results are highlighted in bold.

Defect Category	Baseline (YOLOv8l) AP	Ours (Proposed) AP
Patches	88.0%	**93.1%**
Inclusion	75.2%	**85.3%**
Pitted Surface	72.5%	**84.7%**
Scratches	75.8%	**84.6%**
Rolled-in Scale	45.3%	**58.0%**
Crazing	38.5%	**57.8%**
mAP@0.5	69.8%	**76.3%**
mAP@0.5–0.95	26.5%	**41.1%**

3.2. Visual Quality Analysis

3.2.1. Precision-Recall Analysis

Figure 8 displays the P-R curves for all six classes. The substantial area under the curves for Patches and Inclusion demonstrates that the model maintains high precision even as recall increases.

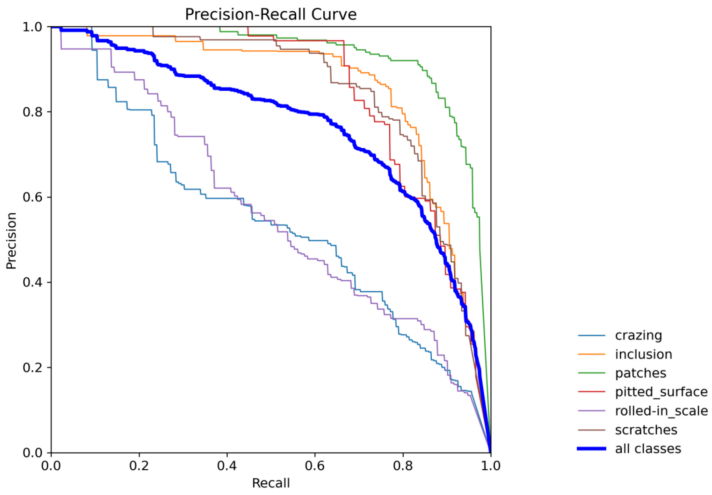


Figure 8. Precision-Recall curve of the improved model. The model shows balanced performance across different defect types. The different colored lines correspond to specific defect categories as indicated in the figure legend.

3.2.2. Error Distribution Analysis

The confusion matrix is presented in Figure 9. While the diagonal dominance indicates strong classification ability, specific misclassifications reveal remaining challenges:

1. Background Confusion: There is a 0.31 misclassification rate for Crazing as background, and 0.27 for Rolled-in Scale. This is attributed to the low contrast of these defects against the steel grain.
2. Inter-class Similarity: The confusion between Patches and Rolled-in Scale is minimal, suggesting that the CBAM attention mechanism successfully extracted discriminative features for these visually similar defects.

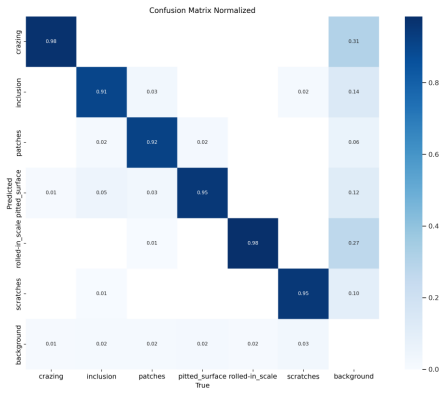
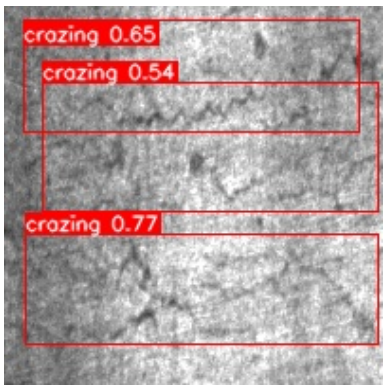


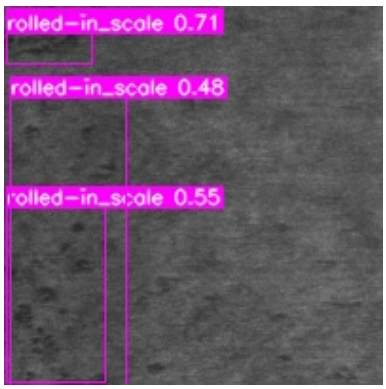
Figure 9. Normalized confusion matrix. The values on diagonal represent the recall rate for categories.

3.3. Visualization of Detection Results

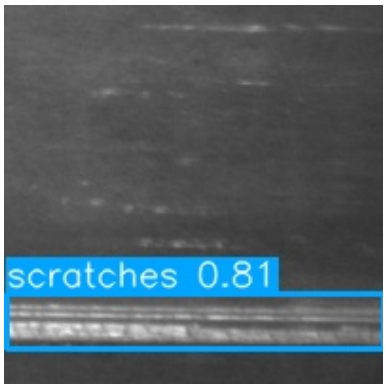
To intuitively demonstrate the robustness of the algorithm in complex industrial environments, Figure 10 presents the detection results on the test set.



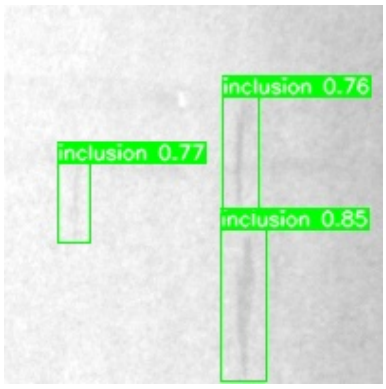
(**a**) Crazing Detection



(**b**) Rolled-in Scale Detection



(**c**) Scratches Detection



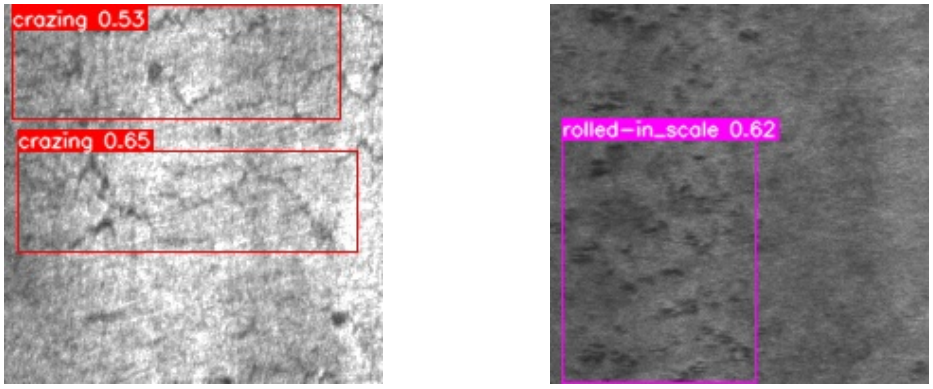
(**d**) Inclusion Detection

Figure 10. Visualization of detection results. (**a**) Minute crazing is accurately bounded; (**b**) Dense rolled-in scale is effectively distinguished; (**c**) Low-contrast scratches are precisely located; (**d**) Inclusions are detected with high confidence.

As observed in Figure 10, the improved model exhibits strong robustness against background noise. Even for defects with extremely weak features, the model successfully generates tight bounding boxes without significant missed detections.

3.4. Failure Case Analysis

Despite the overall robust performance, the model still exhibits limitations when dealing with extreme conditions, which aligns with the misclassifications observed in the confusion matrix. Figure 11 illustrates two typical failure cases on the test set.



(**a**) Crazing (**b**) Rolled-in Scale

Figure 11. Typical failure cases of the proposed model. (**a**) For "Crazing", the model generates low-confidence predictions (0.33 and 0.55) and misses the faint crack textures in the bottom region. (**b**) For "Rolled-in Scale", the defect area shares extremely high visual similarity with the normal background grain, resulting in a low confidence score of 0.52.

As shown in Figure 11a, there is still a notable error rate where fine "Crazing" textures are misclassified as background. Similarly, Figure 11b demonstrates that the model struggles to distinguish "Rolled-in Scale" from the normal background grain due to the extremely low contrast. These specific failure instances intuitively confirm that purely visual features appear insufficient for identifying certain extremely low-contrast defects that are visually merged with the intrinsic steel textures. These challenging cases highlight the physical bottleneck of purely optical feature extraction and underscore the necessity of exploring multi-modal sensing or more aggressive model quantization for edge deployment, as discussed in Section 4.2.

3.5. Ablation Study

To comprehensively verify the effectiveness of the proposed architecture and address potential concerns regarding the image resizing strategy, we conducted two distinct sets of ablation experiments: a module-wise ablation study and a resolution ablation study.

3.5.1. Module-Wise Ablation Study

To verify the contribution of each proposed module—P2 High-resolution Head, BiFPN, CBAM, and Test Time Augmentation (TTA)—a step-by-step ablation study was conducted. Following the recommendations for industrial defect analysis, we tracked not only the overall mAP@0.5 but also the specific Average Precision (AP) for two of the most challenging categories: Crazing (minute textures) and Rolled-in Scale (low contrast). The results are summarized in Table 2.

The experimental results indicate a progressive improvement. Notably, the introduction of the P2 high-resolution branch significantly improved the recall of minute Crazing features, while the CBAM hybrid attention mechanism effectively suppressed the complex background noise, providing the most substantial boost for low-contrast Rolled-in Scale.

Table 2. Ablation study of different modules on the NEU-DET dataset. The AP for specific challenging categories is included to highlight targeted improvements. The checkmark (✓) indicates that the corresponding module is applied. Bold values indicate the best results.

Exp.	P2 Head	BiFPN	CBAM	TTA	Crazing AP	Scale AP	mAP@0.5
Baseline					38.5%	45.3%	69.8%
2	✓				43.1%	50.7%	68.5%
3	✓	✓			46.6%	52.3%	70.8%
4	✓	✓	✓		53.9%	56.4%	73.7%
5 (Ours)	✓	✓	✓	✓	**57.8%**	**58.0%**	**76.3%**

3.5.2. Resolution Ablation Study: Isolating Architectural Gains

The proposed methodology involves up-sampling the original 200 × 200 grayscale images to 1280 × 1280 pixels to expand the physical receptive field for the P2 layer. To rigorously verify that the performance gains are derived from our architectural innovations rather than merely interpolation artifacts from this extreme up-sampling strategy, we trained both the baseline YOLOv8l and our proposed model using a strictly limited input resolution of 256 × 256. The comparative results across different resolutions are presented in Table 3.

Table 3. Performance comparison of the baseline and the proposed model at different input resolutions. This isolates the architectural gains from the resizing strategy. Bold values indicate the best results for each resolution.

Model	Input Size	mAP@0.5	Crazing AP	Rolled-in Scale AP
Baseline	256 × 256	70.0%	42.5%	40.5%
Ours	256 × 256	**72.1%**	**49.6%**	**58.0%**
Baseline	1280 × 1280	69.8%	38.5%	45.3%
Ours	1280 × 1280	**76.3%**	**57.8%**	**58.0%**

As shown in Table 3, our proposed model not only consistently outperforms the baseline at the low 256 × 256 resolution (boosting "Crazing" and "Rolled-in Scale" by 7.1% and 18.5%, respectively), but it is also the only architecture capable of capitalizing on the extreme up-sampling. These findings robustly confirm that the fundamental performance leap is strictly driven by the fine-grained perception capability of the P2 branch and the noise suppression of CBAM, rather than mere resizing artifacts.

3.5.3. Analysis of Feature Fusion Weights

To further investigate the internal mechanism of how the network rationally distributes attention across different defect sizes, we extracted the channel attention activation intensities (measured by the L1 Norm of the learned channel weights in the CBAM modules) across the four detection scales from the optimally trained model. The quantitative distribution of these attention weights is presented in Table 4.

As demonstrated in Table 4, the network exhibits a highly adaptive, scale-aware behavior rather than relying on static scalar weights. For extensive defects such as Patches, the network autonomously allocates massive attention weights at the P4 and P5 macroscopic scales (1438.0 and 1433.0, respectively) to capture global topological semantics. Conversely, for microscopic flaws like Crazing, the network employs a much more delicate and constrained weighting strategy at the P2 scale (188.6). This precise recalibration prevents the faint, pixel-level textural features of minute defects from being aggressively suppressed

or mathematically overwhelmed by the background noise during the fusion process. This explicit quantitative evidence strongly corroborates the rationale for integrating the P2 high-resolution branch with the CBAM mechanism.

Table 4. Distribution of channel attention activation intensities across different feature scales in the proposed model.

Feature Scale (Layer)	Target Defect Characteristic	Activation Intensity (L1 Norm)
P2 (model.19)	Microscopic (e.g., Crazing)	188.6
P3 (model.23)	Small to Medium	516.5
P4 (model.27)	Large (e.g., Patches)	1438.0
P5 (model.31)	Global Semantic Context	1433.0

3.6. Computational Complexity and Real-Time Performance

To comprehensively evaluate the feasibility of deploying the proposed YOLOv8-BiFPN-P2-CBAM in the industrial environment, we compared its computational complexity and its inference speed with other mainstream object detectors. The evaluations were conducted on an NVIDIA GeForce RTX 4090 GPU. As shown in Table 5, replacing the standard PANet with our weighted BiFPN effectively reduced the total parameter count to 42.9 M, which is even smaller than the baseline YOLOv8l (43.7 M) and YOLOv5l (46.5 M). Although the high-resolution P2 branch increased the floating-point operations (FLOPs) to 204.9 G, our model still maintained a highly competitive inference speed of 52.3 FPS. Since typical high-speed rolling production lines require a detection speed of at least 30 FPS, our proposed architecture fully satisfies the real-time detection requirements while establishing a more optimal trade-off between micro-defect perception and computational efficiency.

Table 5. Comparison of model complexity, inference speed, and detection accuracy. Bold formatting is used to highlight the proposed model and its corresponding results.

Model	Params (M)	FLOPs (G)	FPS	mAP@0.5
Faster RCNN [6]	41.5	206.0	25.5	70.8
YOLOv5l [36]	46.5	109.1	94.4	74.1
YOLOv7 [37]	36.9	104.7	90.3	71.9
YOLOv8l (Baseline)	43.7	165.2	78.8	69.8
Ours	**42.9**	**204.9**	**52.3**	**76.3**

4. Discussion

The experimental results presented in Section 3 demonstrate that the proposed YOLOv8-BiFPN-P2-CBAM algorithm significantly improves the detection accuracy of steel surface defects, particularly for micro-targets and defects in complex backgrounds.

4.1. Interpretation of Results

The most notable improvement was observed in the detection of Crazing, with AP increasing from 38.5% to 57.8%. In previous studies utilizing standard YOLO architectures, faint, crack-like defects were often lost after multiple downsampling operations. Our findings confirm the hypothesis that constructing a high-resolution P2 branch effectively preserves shallow geometric features.

The baseline model often confused Rolled-in Scale due to their visual similarity. The integration of the CBAM module allowed the network to recalibrate feature weights adaptively. By suppressing the activation of non-structural background noise in the spatial

dimension, the model could focus more precisely on the defect regions, leading to the substantial mAP improvement seen in the ablation study (+2.9% upon adding CBAM).

The use of BiFPN also proved superior to the traditional PANet. The steel defect dataset exhibits extreme scale variations. Although the P2 layer has a smaller receptive field, which might theoretically limit the global context for larger, distributed defects, the top-down pathway of the weighted BiFPN effectively compensates for this by transmitting deep semantic information downwards. This explains why the detection of large defects like Patches remains highly accurate (93.1%). Furthermore, while Focal Loss was employed to address numerical class imbalance, the persistent performance gap between Patches (93.1%) and Crazing (57.8%) indicates that the primary bottleneck lies in the extreme physical difficulty of feature extraction for ultra-low-contrast textures, rather than simple data imbalance. The weighted feature fusion mechanism of BiFPN ensured that features from different resolutions were combined more rationally, preventing the semantic information of small targets from being overwhelmed.

4.2. Limitations and Future Work

Despite the promising results, this study has certain limitations.

1. Computational Cost: The addition of the P2 detection head and the high-resolution input (1280 × 1280) increases the floating-point operations (FLOPs) and memory consumption compared to the baseline YOLOv8l. This may pose challenges for deployment on edge devices with limited computing power.
2. Remaining Confusion: As shown in the confusion matrix, there is still a noticeable error rate between Crazing and the background grain. This suggests that for extremely low-contrast defects, purely visual features may be insufficient.

Future research will focus on two directions. We aim to apply model compression techniques, like channel pruning and quantization, to reduce the model size without significantly compromising accuracy, facilitating real-time deployment on industrial embedded systems, taking inspiration from lightweight architectures like MobileNet [38] and ShuffleNet [39].

5. Conclusions

By constructing a high-resolution P2 feature branch, the problem of information loss for micro-defects was effectively mitigated. The introduction of the weighted BiFPN optimized the feature fusion path for multi-scale targets, while the integration of the CBAM attention mechanism significantly enhanced the model's ability to suppress unstructured environmental noise.

Experimental results on the NEU-DET dataset demonstrate that the proposed method achieves a mean Average Precision (mAP@0.5) of 76.3%, outperforming the baseline model by 10.4%. The detection accuracy for challenging defects like patches and inclusions reached 93.1% and 85.3%. These findings indicate that the proposed algorithm strikes a favorable balance between accuracy and robustness, providing a viable and effective solution for automated quality control in intelligent steel manufacturing.

Author Contributions: Conceptualization, Y.J. and J.Z.; methodology, Y.J.; software, Y.J. and X.Z.; validation, Y.J., X.Z. and J.M.; formal analysis, Y.J.; investigation, Y.J.; resources, J.M. and J.Z.; data curation, Y.J.; writing—original draft preparation, Y.J.; writing—review and editing, J.Z. and J.M.; visualization, Y.J.; supervision, J.Z.; project administration, J.Z.; funding acquisition, J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant number 51807064).

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://www.kaggle.com/datasets/kaustubhdikshit/neu-surface-defect-database (accessed on 10 January 2026).

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

Abbreviations

The following abbreviations are used in this manuscript:

YOLO	You Only Look Once
BiFPN	Bi-directional Feature Pyramid Network
CBAM	Convolutional Block Attention Module
AP	Average Precision
mAP	mean Average Precision
TTA	Test Time Augmentation
CNN	Convolutional Neural Network
FPS	Frames Per Second

References

1. Liu, H.; Hu, R.; Dong, H.; Liu, Z. SFC-YOLOv8: Enhanced Strip Steel Surface Defect Detection Using Spatial-Frequency Domain-Optimized YOLOv8. *IEEE Access* **2025**, *74*, 9700111. [CrossRef]
2. Luo, Q.; Fang, X.; Liu, L.; Yang, C.; Sun, Y. Automated Visual Defect Detection for Flat Steel Surface: A Survey. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 626–644. [CrossRef]
3. Czimmermann, T.; Ciuti, G.; Milazzo, M.; Chiurazzi, M.; Roccella, S.; Oddo, C.M.; Dario, P. Visual-Based Defect Detection and Classification Approaches for Industrial Applications-A Survey. *Sensors* **2020**, *20*, 1459. [CrossRef] [PubMed]
4. Zhang, Y.; Wang, T.; Xue, L.; Lian, W.; Tao, R. ORSI Salient Object Detection via Progressive Interaction and Sa-liency-Guided Enhancement. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 6002105.
5. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
6. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
7. Li, J.; Su, Z.; Geng, J.; Yin, Y. Real-time Detection of Steel Strip Surface Defects Based on Improved YOLO Detection Network. *IFAC-PapersOnLine* **2018**, *51*, 76–81. [CrossRef]
8. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
9. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. Available online: https://arxiv.org/abs/2004.10934 (accessed on 15 December 2025).
10. Jocher, G.; Chaurasia, A.; Qiu, J. YOLO by Ultralytics. Available online: https://github.com/ultralytics/ultralytics (accessed on 10 January 2026).
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
12. Yu, S.; Liu, Z.; Zhang, L.; Zhang, X.; Wang, J. FasterNet-YOLO for real-time detection of steel surface de-fects algorithm. *PLoS ONE* **2025**, *20*, e0323248.
13. Ayon, S.T.K.; Siraj, F.M.; Uddin, J. Steel Surface Defect Detection Using Learnable Memory Vision Trans-former. *Comput. Mater. Contin.* **2025**, *82*, 499–520.
14. Farady, I.; Lin, C.Y.; Chang, M.C. PreAugNet: Improve data augmentation for industrial defect classification with small-scale training data. *J. Intell. Manuf.* **2024**, *35*, 1233–1246. [CrossRef]
15. Wang, Z.; Chen, J.; Wang, C.; Peng, C.; Xuan, J.; Shi, T.; Zuo, M. CNC-VLM: An RLHF-optimized industrial large vision-language model with multimodal learning for imbalanced CNC fault detection. *Mech. Syst. Signal Process.* **2026**, *245*, 113838. [CrossRef]
16. Ranjan, P.; Ankur; Kumar, R. A Residual Pyramid-GAN Approach for Hyperspectral Image Classification. *Ann. Data Sci.* **2026**, 1–26. [CrossRef]
17. Ranjan, P.; Nandal, A.; Agarwal, S.; Kumar, R. A Dive into Generative Adversarial Networks in the World of Hyperspectral Imaging: A Survey of the State of the Art. *Remote Sens.* **2026**, *18*, 196. [CrossRef]
18. Zhou, Z.; Yuan, M. SteelGuard-yolo: Steel surface defect detection network based on improved YOLOv5s. *Rev. Int. Metod. Numer. Para Calc. Diseno Ing.* **2024**, *40*, 1–11. [CrossRef]

19. Cheng, X.; Yu, J. RetinaNet with Difference Channel Attention and Adaptively Spatial Feature Fusion for Steel Surface Defect Detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 2503911. [CrossRef]
20. Zhang, Y.; Zhen, J.; Sun, S.; Liu, T.; Huo, L.; Wang, T. SCAFNet: A Semantic Compensated Adaptive Fusion Network for Remote Sensing Images Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2024**, *62*, 6003405. [CrossRef]
21. Wang, Y.; Cheng, H.; Shi, J.; Du, X.; Wang, P. Co-training encoder-based transformer method for steel surface defect detection. *J. Electron. Imaging* **2025**, *34*, 023058. [CrossRef]
22. Wu, S.; Yang, H.; Liao, L.; Song, C.; Fang, Y.; Yang, Y. SH-DETR: Enhancing steel surface defect detection and classification with an improved transformer architecture. *PLoS ONE* **2025**, *20*, e0334048. [CrossRef]
23. Sheng, W.; Lu, M.; Zou, Y. Steel Surface Defect Detection Based on Improved YOLOv8 Algorithm. In Proceedings of the 2024 4th International Conference on Industrial Automation, Robotics and Control En-gineering (IARCE), Chengdu, China, 15–17 November 2024; pp. 109–113.
24. Song, K.; Yan, Y. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Appl. Surf. Sci.* **2013**, *285*, 858–864. [CrossRef]
25. Tong, K.; Wu, Y.; Zhou, F. Recent advances in small object detection based on deep learning: A review. *Image Vis. Comput.* **2020**, *97*, 103910. [CrossRef]
26. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
27. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
28. 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 (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
29. 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 (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
30. Peng, M.; Bai, S.; Lu, Y. Steel Surface Defect Detection Algorithm Based on Improved YOLOv8 Modeling. *Appl. Sci.* **2025**, *15*, 8759. [CrossRef]
31. 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), Munich, Germany, 8–14 September 2018; pp. 3–19.
32. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
33. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Con-volutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
34. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
35. 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 (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
36. Wang, X.; Ma, S.; Wu, S.; Li, Z.; Cao, J.; Xu, P. Detection of Surface Defects in Steel Based on Dual-Backbone Network: MBDNet-Attention-YOLO. *Sensors* **2025**, *25*, 4817. [CrossRef]
37. 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 (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 7464–7475.
38. 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. Available online: https://arxiv.org/abs/1704.04861 (accessed on 20 June 2025).
39. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Archi-tecture Design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.