Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 22, No. 2, 2022

Steel Surface Defect Detection Using an Enhanced YOLOv8s with PPA Mechanism and AFPN

Dong Wang *

Southwest Minzu University, Chengdu 610225, China
* **Corresponding author Email:** 2505315712@qq.com

Abstract: Accurate and real-time detection of steel surface defects is essential for industrial quality control, yet it remains challenging due to extreme scale variations and complex background noise. Conventional object detection models often suffer from feature degradation and spatial information conflicts when processing these microscopic and multi-scale defects. To overcome these limitations, we propose an enhanced YOLOv8s-based detector. First, a Paralleled Patch-Aware Attention (PPA) module is integrated to extract multi-scale defect features, adaptively emphasizing critical textures while suppressing irrelevant industrial background noise. Second, we optimize the neck architecture by introducing an Asymptotic Feature Pyramid Network (AFPN) equipped with an Adaptive Spatial Fusion (ASF) mechanism. This structure progressively fuses non-adjacent hierarchical features, effectively mitigating semantic gaps and spatial information conflicts during multi-level aggregation. Extensive experiments conducted on the NEU-DET dataset demonstrate that the proposed method significantly outperforms the baseline YOLOv8s and other state-of-the-art models. The enhanced model achieves a superior balance between detection precision (mAP@0.5) and real-time inference speed, making it highly suitable for practical industrial inspection tasks.

Keywords: Steel surface defect detection, YOLOv8, Paralleled Patch-Aware Attention (PPA), Asymptotic Feature Pyramid Network (AFPN), Multi-scale feature fusion.

1. Introduction

Steel is a fundamental material in modern manufacturing, widely utilized in industries such as construction, automotive, and aerospace. However, during the manufacturing process, various surface defects—such as crazing, patches, and scratches—inevitably occur due to equipment fatigue or harsh operational environments. These defects not only compromise the structural integrity and durability of the steel products but also cause significant economic losses. Therefore, accurate and real-time steel surface defect detection is a critical step for industrial quality control. Traditionally, defect inspection relies heavily on manual visual examination, which is labor-intensive, time-consuming, and highly subjective, making it difficult to meet the demands of high-speed automated production lines.

In recent years, deep learning-based object detection algorithms, particularly the YOLO (You Only Look Once) [1]series, have dominated the field of industrial defect inspection due to their remarkable balance between precision and inference speed. While baseline models like YOLOv8 achieve satisfactory results on general public datasets, steel surface defect detection presents unique challenges. First, defects exhibit extreme scale variations; for instance, microscopic scratches may occupy only a few pixels, while massive patches can cover significant image areas. During the repeated downsampling operations of standard Convolutional Neural Networks (CNNs), the fine-grained texture information of these minute defects is prone to severe degradation or complete loss. Second, industrial environments often introduce complex background noise, such as uneven illumination or water stains, which easily obscures subtle defect features. Furthermore, traditional multi-scale feature fusion strategies (e.g., PANet) used in baseline models directly fuse non-adjacent hierarchical features, creating massive semantic gaps and spatial information conflicts that restrict the overall detection accuracy.

To fundamentally address the aforementioned challenges, this paper proposes an enhanced real-time detection model based on YOLOv8s, tailored for complex steel surface inspection. Specifically, to mitigate the feature loss of microscopic defects and suppress complex background interference, we introduce a Paralleled Patch-Aware Attention (PPA) module into the network. This module utilizes a multi-branch feature extraction strategy and an adaptive dual-dimensional attention mechanism to accurately capture subtle defect textures. Additionally, to overcome the semantic gaps in multi-scale feature aggregation, we replace the conventional neck with an Asymptotic Feature Pyramid Network (AFPN). By leveraging a progressive fusion paradigm and an Adaptive Spatial Fusion (ASF) mechanism, the optimized neck seamlessly aligns cross-level semantic information and eliminates spatial multi-object conflicts, significantly enhancing the models capability to detect multi-scale defects.

The main contributions of this paper are summarized as follows:

·We propose an enhanced YOLOv8s-based detector specifically optimized for steel surface defect detection, achieving a superior trade-off between detection accuracy and real-time inference speed.

·We integrate the Paralleled Patch-Aware Attention (PPA) module to extract robust multi-scale representations, effectively preventing the degradation of microscopic defect features in deep neural networks and filtering out complex industrial background noise.

·We construct an optimized neck architecture using AFPN equipped with an Adaptive Spatial Fusion (ASF) operation, which progressively aligns non-adjacent hierarchical features and robustly resolves spatial semantic conflicts during multi-scale feature fusion.

·Extensive comparative and ablation experiments on the public NEU-DET[2] dataset validate that the proposed model sig-

nificantly outperforms existing state-of-the-art methods in industrial defect inspection.

2. Proposed Method

2.1. Overview of YOLOv8

YOLOv8 is a state-of-the-art, single-stage object detection model that achieves an optimal balance between detection accuracy and inference speed. In this study, YOLOv8s is selected as the baseline model because its relatively lightweight architecture meets the strict real-time processing requirements of industrial steel surface inspection while maintaining high precision.

The standard YOLOv8 architecture primarily consists of three core components: the Backbone, the Neck, and the Head.

·**Backbone**: The network utilizes a modified CSP-Darknet structure featuring the C2f (Cross Stage Partial bottleneck with 2 convolutions) module. The C2f module is designed to extract rich feature representations from the input steel images. By enriching the gradient flow, it allows the model to capture comprehensive contextual information and distinct defect textures.

·**Neck**: The baseline YOLOv8 employs a Path Aggregation Network (PANet) architecture to fuse multi-scale features extracted by the Backbone. It is responsible for combining semantic information from deep network layers with fine-grained spatial details from shallow layers, which is generally effective for objects of regular sizes.

·**Head**: YOLOv8 adopts an anchor-free, decoupled head structure. By separating the object classification and bounding box regression tasks into distinct branches, it effectively reduces the number of parameters, accelerates the models convergence speed, and improves localization accuracy.

Although the standard YOLOv8s demonstrates strong baseline performance, its conventional feature fusion mechanism and uniform attention distribution can still be further optimized. When facing extreme scale variations (e.g., massive patches versus minute scratches) and complex background noise typical in steel defect images, the baseline model may experience feature loss or misclassification. This limitation necessitates the introduction of the targeted attention mechanism and an optimized Neck structure.

2.2. Paralleled PPA Module

In steel surface defect detection, minute defects such as fine scratches often suffer from severe feature degradation during the multiple down sampling processes of the backbone network. To fundamentally address this issue, we introduce the Paralleled Patch-Aware Attention (PPA) module. As illustrated in Figure 1(a), the PPA module operates through a synergistic combination of a multi-branch feature extraction strategy and a dual-dimensional attention mechanism.

The multi-branch architecture is explicitly designed to capture comprehensive multi-scale representations of steel defects. Given an input feature tensor $F \in R^{H\times W\times C}$, a point-wise convolution is initially applied to map it into a new channel space, yielding $F' \in R^{H\times W\times C}$. The network subsequently processes this tensor through three paralleled branches: a local branch, a global branch, and a serial convolution branch.

In both the local ($p = 2$) and global ($p = 4$) branches, a Patch-Aware mechanism is employed to aggregate nonoverlapping spatial patches. Specifically, computationally efficient Unfold and reshape operations partition F into a set of spatially contiguous patches of size $p \times p, H'/p, W'/p, C'$. After channel-wise averaging, a Feed-Forward Network (FFN) and a $Softmax$ activation function computes the spatial probability distribution to dynamically ad-just patch weights.

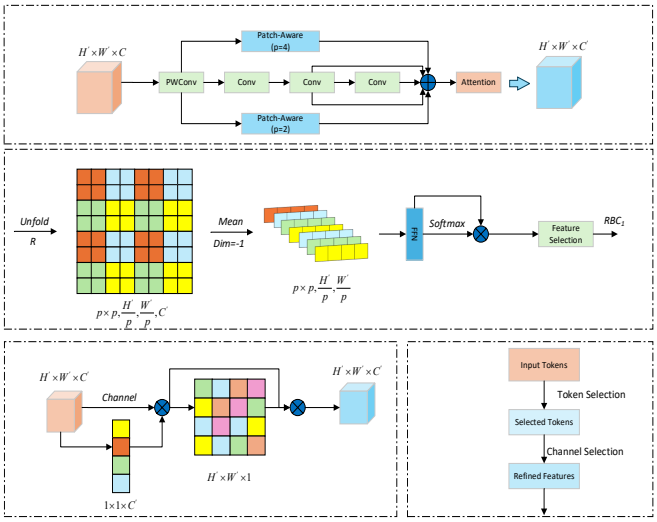

Figure 1. Architecture of the PPA module

During the subsequent feature selection phase (Figure 1(d)), relevant features are chosen from tokens and channels. Let the token dimension be defined as $d = H' \times W'/p \times p$. The weighted tokens are represented as $(t_i)_{i=1}^{C}$, where $t_i \in R^d$. The selective reweighting for each token is formulated as:

$$\dot{t}_i = P \cdot sim(t_i, \varepsilon) \cdot t_i \quad (1)$$

Where $\varepsilon \in R^C$ functions as the task-specific embedding vector determining token relevance, $P \in R^{C\times C}$ is a linear transformation matrix for channel selection, and $sim(\cdot,\cdot)$ denotes the cosine similarity function bounded within [0, 1]. This yields the local and global features $F_{local} \in R^{H\times W\times C}$ and $F_{global} \in R^{H\times W\times C}$.

Concurrently, the serial convolution branch substitutes conventional large-kernel convolutions(7×7,5×5and3×3) with a highly efficient cascade of three convolutional layers, producing $F_{conv} \in R^{H\times W\times C}$,The outputs from all three branches are element-wise summed to generate the fused multi-scale feature$\bar{F} \in R^H \times W' \times C'$.

To further emphasize crucial defect textures and sup-press the pervasive industrial background noise, the fus-ed feature $\bar{F}$ is sequentially fed into an attention block co-nsisting of a 1D channel attention map $M_c \in R^{1\times 1\times C'}$ and a 2D spatial attention map $M_s \in R^{H'\times W'\times 1}$. The sequential refinement is mathematically defined as:

$$F_c = M_c(\tilde{F}) \otimes \tilde{F} \quad (2)$$

$$F_s = M_s(F_c) \otimes F_c \quad (3)$$

$$F'' = \delta(\beta(dropout(F_s))) \quad (4)$$

Here, F_c and F_s denote the intermediary feature representations after channel and spatial selections, respectively. $\delta(\cdot)$ is the Rectified Linear Unit ($ReLU$) activation, $\beta(\cdot)$ refers to Batch Normalization (BN), and F'' represent the final enhanced output of the PPA module.

2.3. Optimized Neck with AFPN

In object detection tasks, extracting and fusing multi-scale

features is of great importance for encoding objects with extreme scale variances. While traditional top-down and bottom-up feature pyramid networks (such as PANet) are widely used, they often suffer from the loss or degradation of detailed feature information during multi-stage transmission, severely impairing the fusion effect of non-adjacent hierarchical levels. For steel surface defect detection, where defect scales vary drastically from microscopic scratches to massive patches, this semantic degradation is particularly detrimental. To address this structural limitation, we construct an optimized Neck based on the Asymptotic Feature Pyramid Network (AFPN)[3].

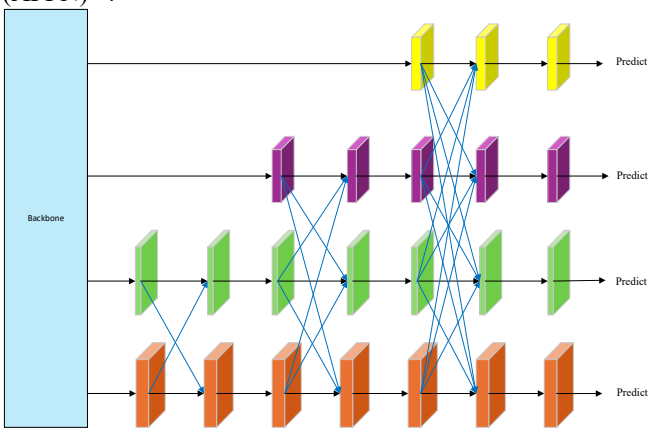

Figure 2. Architecture of AFPN

As illustrated in Figure 2, AFPN discards the traditional sequential fusion paradigm in favor of an asymptotic multi-scale interaction mechanism. Specifically for the YOLO architecture, the backbone extracts three distinct levels of hierarchical features, denoted as $\{C_3, C_4, C_5\}$. Instead of fusing them all at once, AFPN is initiated by fusing the adjacent low-level features (C_3 and C_4). Subsequently, it asymptotically incorporates the higher-level, more abstract semantic feature (C_5) into the fusion process.

To precisely align the spatial dimensions for direct feature interaction, AFPN employs 1×1 convolutions combined with bilinear interpolation for upsampling. Conversely, for downsampling, it utilizes specific convolutional kernels and strides, such as a 2×2 convolution with a stride of 2 for 2-times downsampling. By progressively integrating features through these direct connections, the large semantic gap between non-adjacent levels is effectively mitigated. This ensures that the fine-grained texture information of low-level features is deeply enriched by the high-level semantic context without being washed out during propagation.

Throughout the multi-level feature fusion process, traditional operations like simple element-wise summation or concatenation are inadequate, as severe multi-object information conflicts and semantic contradictions may arise at the same spatial location. To alleviate these spatial inconsistencies, the AFPN incorporates an Adaptive Spatial Fusion (ASF) operation.

As depicted in Figure 3, the ASF module assigns varying, learnable spatial weights to the features from different levels before aggregation. Let $x_{ij}^{n \to l}$ denote the resized feature vector at spatial position (i, j) mapped from level n to the target level l. The resultant fused feature vector, denoted as y_{ij}^{l}, is computed through the linear combination of these spatially weighted features:

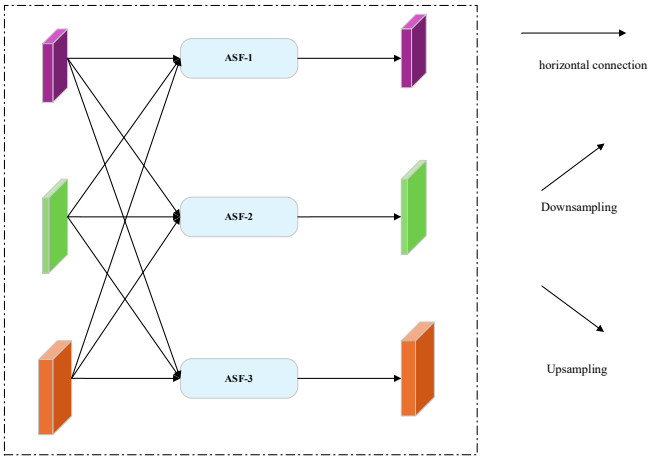

Figure 3. Adaptive Spatial Fusion (ASF) module

$$y_{ij}^{l} = \alpha_{ij}^{l} \cdot x_{ij}^{1 \to l} + \beta_{ij}^{l} \cdot x_{ij}^{2 \to l} + \gamma_{ij}^{l} \cdot x_{ij}^{3 \to l} \quad (5)$$

Where $\alpha_{ij}^{l}, \beta_{ij}^{l}$ and γ_{ij}^{l} represent the dynamic spatial weight matrices corresponding to the three input levels. These weights are adaptively learned by the network and are subject to the strict constraint that their sum at any given spatial location equals exactly 1:

$$\alpha_{ij}^{l} + \beta_{ij}^{l} + \gamma_{ij}^{l} = 1 \quad (6)$$

By dynamically multiplying the multi-scale features with these adaptive weights, the ASF mechanism acts as a robust spatial filter. It forcefully suppresses redundant industrial background noise while actively retaining and emphasizing the most discriminative defect information, thereby significantly boosting the models overall localization and classification accuracy.

3. Experiments and Results

3.1. Dataset and Implementation Details

To evaluate the effectiveness of the proposed enhanced YOLOv8 model, experiments were conducted using the widely recognized Northeastern University Surface Defect Database (NEU-DET). This dataset is specifically designed for industrial steel strip inspection and comprises a total of 1,800 grayscale images with a uniform resolution of 200 × 200 pixels. The dataset meticulously categorizes steel surface defects into six distinct classes: crazing, inclusion, patches, pitted surface, rolled-in scale, and scratches. Each specific defect category contains exactly 300 annotated images, ensuring a balanced data distribution that prevents the network from developing a learning bias during the training phase. For the experimental setup, the dataset was randomly partitioned into training, validation, and testing sets with a ratio of 7:2:1. Specifically, 70% of the images were utilized to optimize the network parameters, 20% were used for validation to tune hyperparameters and prevent overfitting, and the remaining 10% were strictly reserved for the final objective evaluation of the models performance.

The proposed network was implemented using the Pytorch deep learning framework on a Linux operating system. During the training process, the input images were resized to 640 × 640 pixels to maintain consistency with the YOLOv8 architecture. The model was trained from scratch for 100 epochs to ensure full convergence. The network weights were optimized using the AdamW optimizer with an initial learning rate of 0.01. These hyperparameter settings were maintained consistently across all baseline and ablation experiments to

ensure a fair comparison.

3.2. Evaluation Metrics

To quantitatively evaluate the detection performance of the proposed enhanced YOLOv8 model on steel surface defects, several standard evaluation metrics are employed. These core metrics include Precision (P), Recall (R), and mean Average Precision (mAP).

Precision measures the proportion of correctly predicted positive observations out of the total predicted positive observations. Recall measures the proportion of correctly predicted positive observations out of all actual positive observations. They are defined as follows:

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

In these equations, TP(True Positives) represents the number of correctly detected steel defects, FP(False Positives) denotes the number of background regions incorrectly identified as defects, and FN (False Negatives) refers to the number of actual defects that the model failed to detect.

Furthermore, Average Precision (AP) is calculated as the area under the Precision-Recall curve for a single defect class. The mean Average Precision (mAP) is the average of AP values across all N defect classes, providing a comprehensive assessment of the models overall performance. Specifically, $mAP@50$ is calculated at an Intersection over Union (IoU) threshold of 0.5:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{9}$$

Where N is the total number of defect categories (in this study,$N = 6$), and AP_i is the Average Precision for the i-th class. A higher $mAP@50$ score indicates superior accuracy and robustness of the object detection network in industrial environments.

3.3. Comparison with State-of-the-Art Models

To further verify the competitiveness and superiority of the proposed model, we conducted comprehensive comparative experiments with several state-of-the-art (SOTA) object detection models on the NEU-DET dataset. The selected models for comparison encompass a wide range of architectures, including the classic two-stage detector Faster R-CNN, as well as the mainstream and latest one-stage detectors such as YOLOv5s, the original YOLOv8s, YOLOv9, and YOLOv10. To ensure absolute fairness, all models were trained and evaluated under identical experimental configurations.

As presented in Table 1, the proposed model achieves a remarkable improvement in comprehensive detection accuracy compared to other existing methods. While the latest models like YOLOv9[4] and YOLOv10[5] exhibit strong feature extraction capabilities as general-purpose detectors, our proposed model significantly outperforms them in the core metric of mAP@0.5. Compared to the heavy two-stage detector Faster R-CNN, our model not only yields higher precision and recall rates but also maintains a substantially faster inference speed (FPS), which is a critical requirement for real-time industrial defect inspection. These compelling results demonstrate that the synergistic integration of the PPA mechanism and the AFPN architecture effectively resolves the challenges of extreme scale variations and microscopic defect feature loss, allowing our domain-specific network to surpass even the latest generic SOTA models in steel surface defect detection.

Table 1. Comparison with SOTA models

Model	Recall	mAP@0.5	mAP@0.5:0.95	FPS
Faster R-CNN	0.582	0.635	0.312	22
YOLOv5s	0.621	0.688	0.334	145
YOLOv8s	0.644	0.701	0.351	120
YOLOv9	0.658	0.714	0.362	105
YOLOv10	0.665	0.721	0.368	**155**
YOLOv8s+PPA+AFPN	**0.683**	**0.745**	**0.392**	95

3.4. Ablation Study

To thoroughly evaluate the individual contributions and effectiveness of the proposed Paralleled Patch-Aware Attention (PPA) module and the Asymptotic Feature Pyramid Network (AFPN), we conducted a series of ablation experiments on the NEU-DET dataset. The quantitative results of different network configurations are summarized in Table 2.

Table 2. Ablation study results

Model	PPA	AFPN	Recall	mAP@0.5	mAP@0.5:0.95
Baseline			0.644	0.701	0.351
Model A	√		0.662	0.722	0.368
Model B		√	0.669	0.727	0.372
proposed	√	√	**0.683**	**0.745**	**0.392**

As shown in the table, the baseline YOLOv8s achieves a standard mAP@0.5 of 0.701 and a strict mAP@0.5:0.95 of 0.351. While it provides a reasonable starting point, its limited capability to handle complex industrial backgrounds restricts further precision. By independently integrating the PPA module into the backbone (Model A), the network experiences a clear performance boost, with Recall increasing from 0.644 to 0.662 and mAP@0.5 rising to 0.722. This improvement quantitatively demonstrates that the dual-dimensional attention mechanism within the PPA module successfully highlights critical microscopic defect textures while adaptively filtering out irrelevant background noise.

Alternatively, substituting the original feature fusion neck with the AFPN structure (Model B) yields an even higher mAP@0.5 of 0.727. This notable enhancement proves that the asymptotic fusion strategy, combined with the ASF mechanism, effectively bridges the semantic gaps between non-adjacent hierarchical features, providing better localization for defects with drastic scale variations.

Finally, our proposed complete model, which couples both the PPA module and the AFPN architecture, achieves the best overall performance. The synergistic combination dramatically boosts the mAP@0.5 to 0.745 and the highly stringent mAP@0.5:0.95 to 0.392. This compelling outcome confirms that the robust multi-scale feature extraction provided by PPA and the spatial conflict mitigation provided by AFPN are highly complementary. Together, they form a comprehensive and highly accurate solution for complex steel surface defect detection.

3.5. Qualitative Analysis of Detection Results

To intuitively demonstrate the effectiveness and robustness of the proposed model in real-world scenarios, we visualize the detection results on the NEU-DET test set. Figure 4 presents a selection of detection samples, specifically highlighting two challenging defect categories: "scratches" (marked with pink bounding boxes) and "rolled-in scale" (marked with blue bounding boxes).

As shown in the visualization, the proposed model exhibits exceptional localization capabilities under complex industrial backgrounds. For "scratches", which are typically narrow,

elongated, and highly variable in length, the predicted bounding boxes tightly enclose the defective textures without including excessive background noise. This precise localization of fine-grained features vividly demonstrates the superior feature extraction capability of the introduced Paralleled Patch-Aware Attention (PPA) module.

Furthermore, for the "rolled-in scale" defects, which often appear as irregular, mottled patches spanning various scales, the model accurately predicts bounding boxes of corresponding sizes. It successfully detects both large continuous patches and smaller isolated fragments within the same image. This proves the efficacy of the Asymptotic Feature Pyramid Network (AFPN) in handling extreme scale variations and bridging semantic gaps. Overall, the qualitative visual results strongly align with the previous quantitative improvements, confirming that the enhanced YOLOv8s is highly competent for complex steel surface defect inspection.

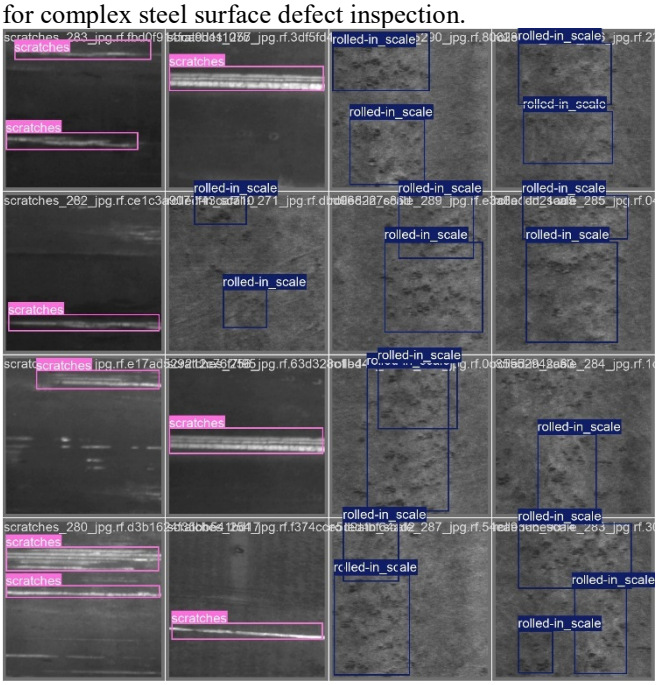

Figure 4. Detection results of the proposed method

4. Conclusion

In this paper, we proposed an enhanced real-time object detection model based on YOLOv8s to address the critical challenges of extreme scale variations and complex background noise in steel surface defect detection. To mitigate the degradation of fine-grained features, we introduced the Paralleled Patch-Aware Attention (PPA) module, which adaptively highlights microscopic defect textures and suppresses irrelevant industrial background interference. Furthermore, we designed an optimized neck architecture utilizing the Asymptotic Feature Pyramid Network (AFPN) equipped with an Adaptive Spatial Fusion (ASF) mechanism. This structure progressively fuses non-adjacent hierarchical features, effectively resolving spatial semantic conflicts during multi-scale aggregation.

Extensive experiments on the NEU-DET dataset demonstrate that our proposed method significantly outperforms the baseline YOLOv8s and other state-of-the-art detectors. It achieves a superior balance between stringent localization accuracy (e.g., mAP@[0.5:0.95]) and real-time inference speed, proving its robust applicability in actual industrial environments. In future work, we plan to further optimize the computational efficiency of the network for deployment on resource-constrained edge devices and explore its generalization capabilities across broader industrial automated inspection scenarios.

Acknowledgment

We thank the researchers who constructed and open-sourced the NEU surface defect database, which provided an essential foundation for the training and evaluation of the models in this study. We also extend our gratitude to the developers of the YOLO and PyTorch open-source communities for their invaluable tools and frameworks.

References

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[2] K. Song and Y. Yan, “A noise robust method based on completed local binary patterns for steel surface defect classification,” IEEE Transactions on Instrumentation and Measurement, vol. 62, no. 11, pp. 2856–2864, 2013.W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[3] G. Yang et al., “AFPN: Asymptotic Feature Pyramid Network for Object Detection,” arXiv preprint arXiv:2306.15988, 2023.

[4] C. Y. Wang, I. H. Yeh, and H. Y. M. Liao, "YOLOv9: Learning Zable gradient information aids network to learn everything," arXiv preprint arXiv:2402.13616, 2024.

[5] C. Li et al., “YOLOv10: Real-Time End-to-End Object Detection,” arXiv preprint arXiv:2405.14458, 2024.