Research on Complex Defect Detection Method on Steel Surface based on EBA-YOLO

Wenhui Zhang, Dajian Yi, Zheng Fang, Yi Zhao, Zhangping You

***Abstract*—To address the challenges of detecting complex surface defects in steel materials, an EBA-YOLO object detection model is proposed. The model integrates the efficient channel attention (ECA) module into the backbone network and neck of YOLOv8, enhancing the network's ability to focus on critical features. Additionally, the bidirectional feature pyramid network (BiFPN) is incorporated into the neck to enable the detection model to capture more contextual information across multiple scales, enriching the feature representation. The adaptive spatial feature fusion (ASFF) module is employed to merge features at different scales, and it is added to the multi-scale detection layers in the head to improve the model's ability to detect fine defects in small-sample data. Comparative experimental studies reveal that the mAP-50 (mean Average Precision at an IoU of 0.5) increased by 6.7%, precision across all classes improved by 11.5%, and the recall rate increased by 3.4%. Furthermore, EBA-YOLO maintains a detection speed of 98 frames per second (FPS), ensuring its feasibility for real-time detection applications.**

***Index Terms*—YOLOv8，Steel surface defect detection, Attention mechanism, Multi-feature fusion, Object detection, Visual communication technology**

I. INTRODUCTION

Steel, as the dominant metal in terms of quantity and applications, plays a crucial role in industrial development. Its efficient production processes significantly reduce energy consumption compared to materials like aluminum and offer excellent recyclability, aligning with environmental protection and sustainable development goals. In 2022, global crude steel production reached 1.888 billion tons, with iron, comprising 5.6% of the Earth's crust, providing abundant raw materials. The strength and toughness of steel have driven innovations in modern architecture, such as skyscrapers and bridges, while also playing a pivotal role in transportation, energy, and machinery manufacturing. Steel provides both stability and flexibility to modern society.

Quality issues in flat steel can lead to economic losses and damage to reputation, with surface defects posing a major threat to thin flat steel and wide flat steel. Defects such as cracks and scratches weaken load-bearing capacity, increase the risk of failure, and heighten susceptibility to corrosion. They also reduce aesthetic appeal, disrupt manufacturing processes, and result in higher costs and diminished performance. Consequently, detecting defects is critical to ensuring the safety, reliability, and cost-effectiveness of steel products [1-8]. Traditionally, the detection of surface defects on steel plates has been primarily carried out manually, a process that is both time-consuming and unreliable. To replace manual operations, the rapid development of robotics and vision technologies [9-15] has created an opportunity to leverage computer vision for automating the detection of steel surface defects.

Technologies such as neural networks have also been greatly applied [16-23]. However, traditional methods still face challenges such as low accuracy and high labor intensity. In contrast, the emergence of machine learning represents a significant breakthrough over manual inspection. This approach typically begins with manual feature extraction, followed by inputting these features into a classifier to categorize defects. Yet, as previously mentioned, its reliance on hand-crafted feature extraction rules results in poor flexibility and adaptability, making it difficult to cope with new environments. Additionally, it is susceptible to external factors and noise, which can reduce detection accuracy. Since 2012, convolutional neural networks (CNNs) have become the dominant models in the field of computer vision [24], widely applied to various vision tasks. Object detection techniques are generally divided into single-stage detectors and region-based two-stage detectors. The YOLO family represents single-stage detectors, while the R-CNN family exemplifies two-stage detection algorithms. In recent years, the application of deep learning in industrial fields has grown due to its ability to extract latent features from data without requiring manually designed complex feature extraction rules. For instance, Luo et al. [25] proposed a YOLO-based surface defect detection algorithm that enhances detection speed through feature enhancement, although it still exhibits limitations in accuracy. In another study, Liu et al. [26] integrated attention mechanisms to develop the YOLO-SO model for identifying insulators and detecting defects in aerial images, achieving significant results. By combining these models, they effectively addressed the trade-off between detection speed and accuracy for insulator defects.

Manuscript received January 30, 2025; revised April 18, 2025.

This work was supported by the National Natural Science Foundation of China (61772247), the industry-Academia-Research Cooperation Projects of Jiangsu Province (BY2022651), the Key Foundation projects of Lishui (2023LTH03), Zhejiang Qianlin Sewing Equipment Co., Ltd. Doctoral Innovation Station, Discipline Construction Project of Lishui University (Discipline Fund Name: Mechanical Engineering).

Wenhui Zhang is a professor at Nanjing Xiaozhuang University, Nanjing 211171, P. R. China (e-mail: hit_zwh@126.com).

Dajian Yi is a postgraduate student at Zhejiang Sci-Tech University, Hangzhou 310018, P. R. China (email: ydj1176540684@163.com).

Zheng Fang is an engineer at State Grid Zhejiang Lishui Power Supply Company, Lishui 323000, P. R. China (email: 7899453@qq.com).

Yi Zhao is an engineer at Zhejiang Qianlin Sewing Equipment Co., Ltd and Lishui Key Laboratory of High Power Density Intelligent Drive System, Lishui 323000, P. R. China email: 19870054@qq.com).

Zhangping You is a professor at Lishui University and Lishui Key Laboratory of High Power Density Intelligent Drive System, Lishui 323000, P. R. China (corresponding author to provide phone: +86-15290875562; email: 44536388@qq.com).

Furthermore, Sohan et al. [27] analyzed YOLOv8, highlighting its innovative features, improvements, applicability across different environments, and performance metrics compared to other versions and models, providing a comprehensive evaluation of YOLOv8. Common surface defects on steel, such as cracks and scratches, are characterized by inter-class feature similarities and significant intra-class differences. These challenges, compounded by variations in lighting conditions and material properties, make defect detection more complex. Models often prioritize the identification of intra-class defects, which can lead to reduced classification accuracy.

To address the challenges of poor detection and classification performance in existing steel surface defect detection methods, the EBA-YOLO model is proposed. Specifically, an Efficient Channel Attention Network (ECA-Net) module is integrated into the network's neck to enhance attention and filtering capabilities. This enables the model to focus on relevant steel surface defects while reducing noise and interference. The ECA-Net module allows the YOLOv8 network to effectively filter out irrelevant information and concentrate on identifying valuable target objects, thereby improving its detection performance. In addition to the modifications in the neck architecture, a bi-directional feature pyramid network (BiFPN) is introduced to enable bidirectional information flow. Finally, the adaptive spatial feature fusion (ASFF) module in the prediction head enhances feature fusion across different scales, enabling the model to better learn and identify complex patterns associated with steel failure.

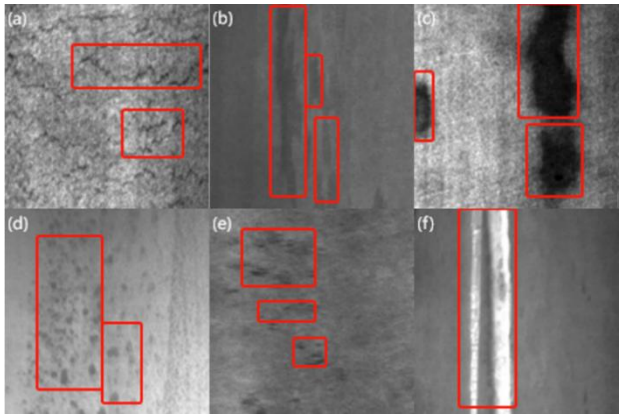


Fig. 1. Defects of different categories: (a) Cracks; (b) Inclusions; (c) Patches; (d) Pitted Surfaces; (e) Rolled-in Scales; (f) Scratches.

In the main network architecture, this paper introduces the ECA attention module and BiFPN feature fusion. BiFPN is a component introduced in the EfficientDet object detection architecture [28], designed to address certain limitations of traditional Feature Pyramid Networks (FPN) in object detection models. By introducing bidirectional connections, BiFPN more flexibly fuses features from different levels, allowing information to flow up and down within the feature pyramid, thereby better capturing contextual information and details across multiple scales. In the prediction head, Adaptive Spatial Feature Fusion (ASFF) is introduced to integrate features at different scales. The NEU-DET dataset was used in the experiments, and the proposed algorithm was compared with the original YOLOv8 algorithm. By incorporating attention mechanisms and multi-scale feature fusion techniques, the model proposed in this paper outperforms the original YOLOv8 architecture. The innovations of the model are listed as follows:

(1) Integration of the ECA Attention Module: The ECA module was incorporated into the backbone and neck of YOLOv8. By selectively enhancing channel features, the ECA module improves the network's ability to focus on critical features, thereby enhancing the detection performance for steel surface defects.

(2) Combined with BiFPN: The bi-directional feature pyramid network (BiFPN) was implemented in the neck, enabling bi-directional feature flow connections and adaptive fusion of multi-level features. This integration allows BiFPN to capture richer contextual information across multiple scales, enhancing the model's fine-grained defect detection capabilities.

(3) Adaptive spatial feature fusion (ASFF): ASFF module was incorporated into the multi-scale detection layers in the head. This module integrates features across different scales, further improving the model's ability to accurately detect fine defects in small-sample datasets.。

The remainder of this paper is organized as follows: Section 2 delves into the existing work related to this study. Section 3 provides a detailed explanation of the methodology of the original YOLOv8 model and the proposed EBA-YOLO detection method. Section 4 discusses in detail the experiments conducted on the NEU-DET dataset and the results obtained, including ablation studies to further demonstrate the superiority of the proposed method. Finally, Section 5 concludes with a summary of the experimental findings and outlines potential directions for future improvements.

II. RELATED WORK

A. Data Augmentation

In the fields of computer vision and object detection, data augmentation is a core technique for improving model generalization and performance. To enhance YOLOv8, a leading real-time object detection architecture, a systematic augmentation strategy was designed. First, basic transformations such as horizontal and vertical flipping were employed to enrich the training data, providing diverse object orientations and perspectives, enabling the model to better adapt to different image features. Second, Mosaic Augmentation was innovatively incorporated, a method that combines four separate images into a single new training image. Moreover, this augmentation strategy increases the diversity of the learning process, strengthening the model's adaptability to various changes in real-world scenarios.

B. Object Detection

With the continuous advancement in the field of object detection, both single-stage and two-stage detectors have achieved significant progress. Among single-stage detectors, the YOLO (You only look once) model stands out for its efficiency, enabling the simultaneous prediction of object classes and bounding box coordinates in a single network inference. This innovative approach made real-time detection possible and laid the foundation for subsequent model iterations, such as YOLOv2 and YOLOv3, which focused on enhancing speed and accuracy. The YOLO series

has evolved further to YOLOv7 and YOLOv8 [29], with YOLOv8 gaining significant attention due to its simplified architecture and notable performance improvements. Meanwhile, SSD (Single shot multiBox detector) [30] introduced a single-stage framework with multi-scale feature layers, improving detection precision. A further advancement came with RetinaNet [31], which incorporated the focal loss mechanism to effectively address the issue of class imbalance, particularly enhancing accuracy in scenarios with abundant background samples.

On the other hand, two-stage detectors, exemplified by Faster R-CNN (Region-based convolutional neural network) [32], adopt a more complex multi-step approach that involves generating region proposals, object classification, and bounding box regression. To achieve higher accuracy, the R-CNN series evolved through various iterations, including R-CNN, Fast R-CNN, Faster R-CNN, and Mask R-CNN, continuously improving object detection performance. This evolution introduced the concept of separating region proposal generation from the detection process, enhancing both flexibility and precision. Cascade R-CNN further advanced this concept by iteratively refining detection quality through a cascade of detectors. Amidst these developments, novel approaches have also had a profound impact on the field of object detection. CornerNet [33] improves detection accuracy by directly predicting key points of objects, while CenterNet [34] focuses on predicting the center and size of objects, achieving excellent performance in real-time detection scenarios. DETR (Detection transformer) leverages the Transformer architecture to reformulate object detection as a set prediction problem, showcasing the potential of attention mechanisms in object detection.

Although the interplay between single-stage and two-stage detectors remains significant, the emergence of YOLOv8 and its various improved versions has introduced new dynamics to the field of object detection. By simultaneously advancing in both speed and accuracy, YOLOv8 and its variants have demonstrated that single-stage detectors can achieve exceptional performance in both aspects, further driving research and discussions on object detection.

C. Attention Mechanism

Attention mechanisms (AMs) represent a revolutionary breakthrough in artificial intelligence and machine learning, enabling models to selectively focus on specific elements within datasets and significantly enhance performance. The pioneering work by Vaswani et al. [35] introduced self-attention through the Transformer architecture, revolutionizing neural machine translation by capturing contextual relationships between words in sequences. Subsequent advancements in attention mechanisms have spurred innovative adaptations, such as CBAM, which harmoniously integrates spatial and channel attention to improve image classification performance. Similarly, the SE module recalibrates feature responses across channels, enhancing model flexibility. The Halo attention method has demonstrated its ability to effectively capture long-range dependencies in images. ECA [36] established an efficient channel attention mechanism, adept at capturing interdependencies among channels with minimal computational overhead. While other attention mechanisms, such as SE or non-local attention, offer similar capabilities, ECA stands out for its computational efficiency, granting it a distinct advantage. These diverse attention strategies collectively underscore the transformative potential of attention mechanisms in advancing AI solutions.

D. Multi-feature fusion

The pioneering paper by Lin et al. [37] introduced the concept of fusing multi-scale features to enhance the representational power of convolutional neural networks (CNNs). This technique aims to combine features extracted from different levels of the network hierarchy to effectively capture fine-grained and high-level contextual information. Multi-feature fusion addresses the limitations of traditional single-scale feature extraction by leveraging the advantages of various features, thereby improving object localization, scale invariance, and semantic contextual awareness.

The field of multi-feature fusion has continuously evolved over the years, giving rise to technologies that have had a significant impact on object detection architectures. The FPN proposed by Lin et al., which was later integrated into YOLOv3 and YOLOv4, paved the way for seamless integration of multi-scale features. This was achieved through a top-down and bottom-up architecture. This innovation greatly enhanced object detection performance, enabling more accurate localization and better semantic understanding. The progressive attention network (PAnet) further optimized multi-feature fusion by dynamically assigning attention weights across different scales, improving the discriminative power of the fused features.

Another important milestone in the field of feature fusion was the introduction of the BiFPN by Tan et al. This network innovatively incorporated a bidirectional information flow mechanism to address the challenges of efficient feature fusion. This mechanism not only improved the efficiency of feature fusion but also ensured high-quality integration of features across multiple scales, significantly enhancing the accuracy and robustness of object detection. Building on this concept, Liu et al. proposed ASFF, which provided a fresh approach to pyramid feature fusion. ASFF introduced a learnable feature fusion strategy that adaptively selects features from different resolutions based on task requirements, effectively enhancing the model's ability to handle diverse object scales and aspect ratios. This flexibility not only improved model performance but also opened new directions and possibilities for future object detection research.

These multi-feature fusion techniques have had a profound impact on the design of object detection architectures, particularly in the widely used YOLO series of models. By integrating FPN, PAnet, BiFPN, and ASFF into the YOLO architecture, these models have not only achieved outstanding detection performance but have also struck the optimal balance between accuracy and efficiency. This integration has not only enhanced the model's ability to understand complex scenes but also improved its real-time performance and adaptability in practical applications, driving continuous advancements and widespread adoption of object detection technology.

III. Proposed methods

A. YOLOv8 Architecture

In recent years, YOLO object detection models have garnered widespread attention for their real-time capability and exceptional performance. YOLOv8, as the latest evolution in the YOLO series, represents a significant technological advancement, achieving remarkable improvements in speed and accuracy compared to its predecessors. The YOLOv8 architecture fully leverages multi-scale features, efficient components, and advanced fusion technologies to deliver state-of-the-art object detection results. By introducing a more lightweight and efficient design, YOLOv8 addresses some of the limitations of previous YOLO versions while maintaining high detection accuracy. YOLOv8 adopts a single-stage processing approach, enabling the entire image to be processed in one pass, making it faster than many other object detection methods. Its architecture utilizes a feature pyramid to capture key multi-scale information, effectively handling detection requirements for objects of various sizes, thus preserving impressive accuracy. The flexibility of YOLOv8 lies in its ability to efficiently handle real-time and large-scale application scenarios, including detection tasks in images and videos. The YOLOv8 architecture is composed of three main components: the backbone, neck, and head. The backbone serves as the main feature extraction network, the neck is responsible for fusing multi-scale features, and the head predicts the object categories and bounding box coordinates. The architectural design of YOLOv8 aims to further optimize performance, allowing it to achieve fast and accurate object detection across a variety of application scenarios. A detailed diagram of the YOLOv8 architecture is shown in Fig. 2.

Backbone: YOLOv8 adopts the improved CSPDarknet53 backbone network, which further optimizes the efficiency and accuracy of the architecture. The CSPDarknet53 architecture introduces Cross-stage partial (CSP) feature fusion, enhancing the network's ability to capture both low-level and high-level features. This design allows the network to not only understand the finer details of objects but also preserve the integrity of contextual information, improving overall detection performance. Neck: YOLOv8's neck utilizes PANet (Path aggregation network) to fuse features from different stages of the backbone network. This feature fusion is crucial for multi-scale feature representation, enabling the model to accurately detect objects of varying sizes and contexts. PANet leverages lateral connections and top-down pathways to aggregate features from different scales, promoting powerful feature fusion and improving the model's ability to detect objects in diverse conditions. Head: YOLOv8's head consists of an anchor-based detection module. For each anchor box, the model predicts the object confidence score, class probabilities, and bounding box coordinates. YOLOv8 incorporates anchor boxes with aspect ratios specifically designed for the target objects, enhancing detection performance by improving the model's ability to handle objects of varying shapes and sizes.

Surface defects in steel are typically characterized by irregular shapes, unpredictable locations, and varying sizes. Additionally, the number of small-scale defects is often large. In this context, the original YOLOv8 model struggles to fully meet the detection requirements. To address this issue, this study enhances the original YOLOv8 network model in several key aspects.

Firstly, improvements were made to the neck of the network by integrating attention mechanisms to emphasize key information while minimizing the influence of irrelevant features. The feature fusion process at the neck was upgraded to enhance the model's ability to capture important features. Additionally, an adaptive spatial feature fusion (ASFF) module was added just before the detection head. These modifications are designed to improve the model's adaptability in recognizing small defects. As a result, these enhancements significantly improve the model's overall performance in defect detection tasks.

C. ECA attention mechanism

Attention mechanisms play a crucial role in contemporary neural network architectures by enhancing the model's information processing capacity and facilitating the connection between different parts of the network. These mechanisms allow the model to selectively focus on relevant features while ignoring irrelevant ones, mimicking the human attention process. Several established attention mechanisms, such as BAM, SE, CBAM, and ECA-Net, have been proven to significantly improve the performance of detection models. ECA-Net, with only a slight increase in complexity, achieves a remarkable performance boost while adding a very limited number of parameters.

Starting from the input feature map containing multiple channels, where each channel represents a different feature, ECA calculates the global context for each channel to indicate its importance relative to other channels. This calculation involves a learnable parameter, typically represented as a 1D convolutional layer, used to compute channel-specific attention coefficients. These coefficients are learned during training, allowing the network to dynamically adjust the importance of each channel. The kernel size (denoted as M) plays a crucial role in this process, as it determines the scope and depth of information integration. By applying a fast 1D convolution with a specific kernel size, ECA effectively captures channel dependencies and interactions in the feature map. The 1D convolution with kernel size M is applied through an adaptive function as shown in equation (1).

$$M = \Psi(C) = \left| \frac{(c)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{1}$$

These attention weights determine the importance of each channel in the overall representation. By integrating the ECA attention module, YOLOv8 can selectively amplify the channels with rich information while suppressing less relevant channels, resulting in a more focused and discriminative feature representation. The integration of the ECA attention module brings several advantages to YOLOv8.

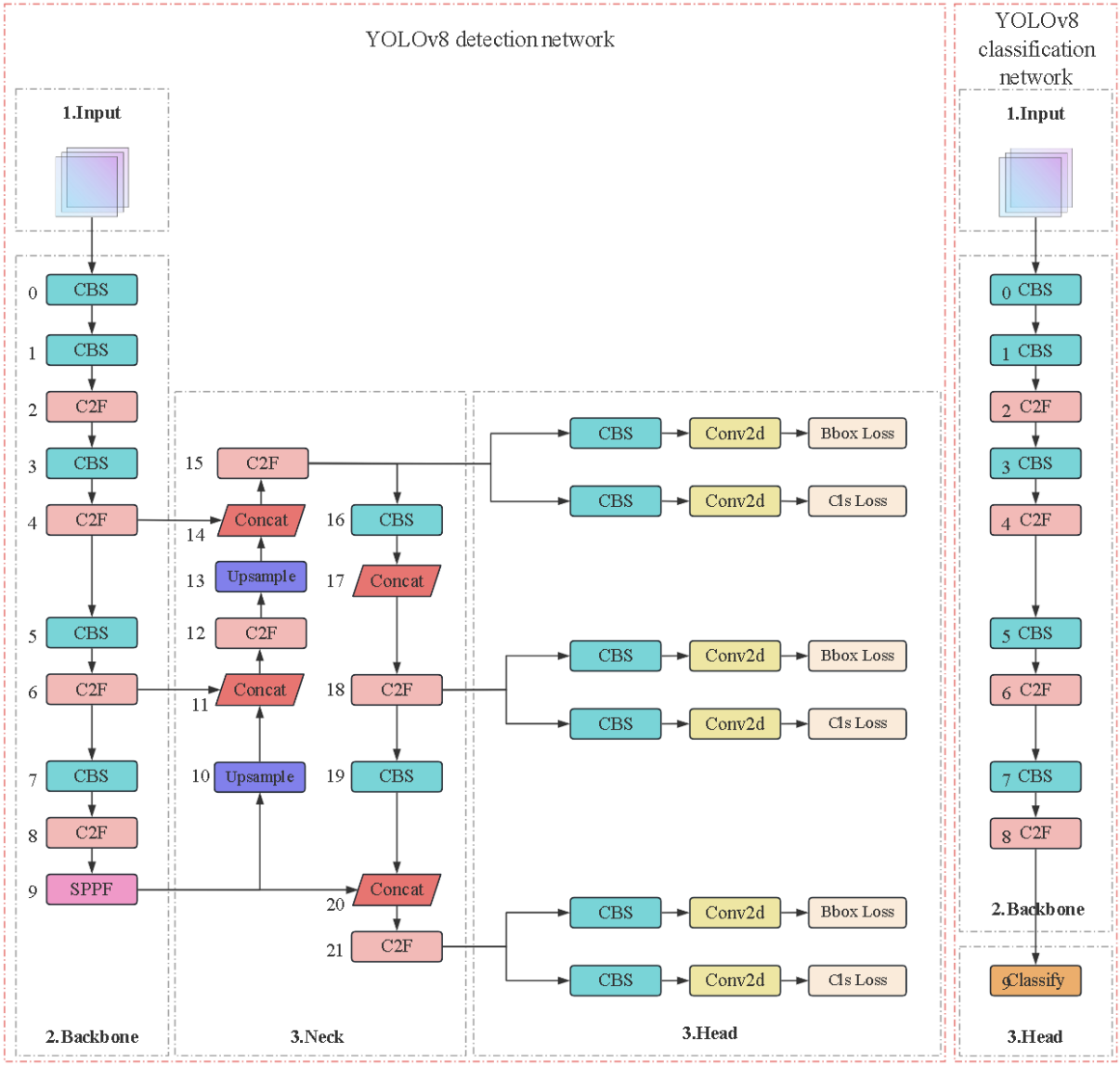


Fig. 2. Original YOLOv8 structure diagram

It improves contextual understanding by capturing long-range dependencies, enhancing the model's ability to recognize objects, especially by emphasizing important features without significantly increasing computational complexity. Fig. 3 illustrates the structural diagram of the ECA module. In summary, the ECA attention module significantly enhances YOLOv8's object detection performance by improving inter-channel dependencies and strengthening feature discrimination. The introduction of this module not only optimizes feature representation but also boosts the model's performance in complex scenarios, enabling YOLOv8 to more effectively identify and locate targets across various applications.

D. BiFPN feature fusion

Replacing the original PANet in the neck component with BiFPN has proven to be highly effective, especially when handling datasets containing small-sized images. BiFPN excels at integrating high-resolution and low-resolution feature data, making it particularly beneficial for detecting defects such as "cracks" and "entangled scales," which typically involve numerous small objects. While FPN effectively generates multi-scale feature maps for object detection, it faces challenges in efficiently handling fine-grained details and maintaining information consistency across different scales. In contrast, BiFPN is an advanced architecture that optimally addresses these shortcomings. It introduces bidirectional connections and lateral connections between adjacent feature maps, enhancing the flow of information both top-down and bottom-up. This bidirectional approach better preserves semantic information and fine object details, ultimately improving the accuracy of object localization and classification. Fig. 4 illustrates the differences between the PANet structure used in the original YOLOv8 and the BiFPN structure integrated into EBA-YOLO. By integrating BiFPN, YOLOv8 achieves more context-aware and fine-grained image understanding, leading to improved object localization and classification accuracy. The bidirectional nature of BiFPN enables it to capture both high-level semantic features and low-level details, providing a comprehensive perspective of the scene. This enhanced feature extraction directly translates to superior object detection performance.

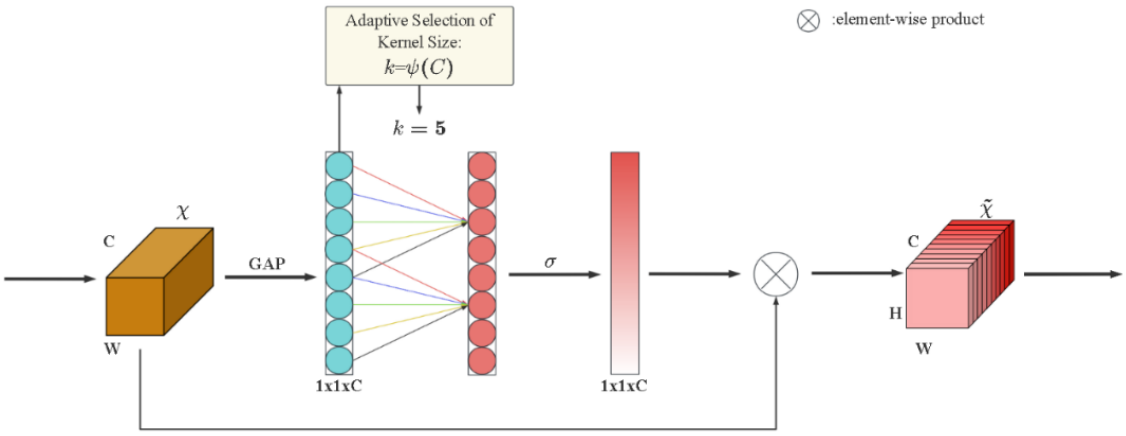


Fig. 3. Schematic diagram of ECA module structure

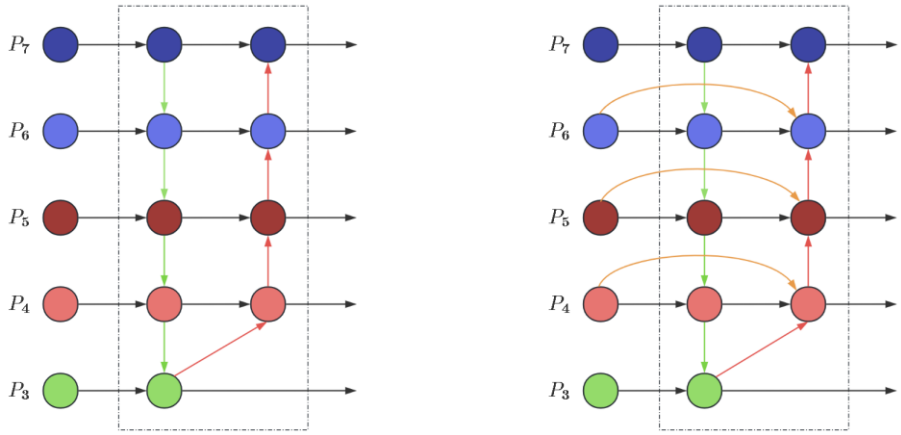


Fig. 4. Comparison of PANet (left) and BiFPN (right)

C. ASFF

Adaptive spatial feature fusion (ASFF) addresses the challenge of effectively integrating multi-scale features from different layers of a convolutional neural network (CNN). Its advantages are manifold [38-41]: first, it significantly improves detection performance by dynamically fusing features from various network layers, thereby achieving more accurate object localization and classification. Second, ASFF excels in handling scale variations, enabling it to adapt to scenarios with objects of different sizes within the same image.

In the process, it maintains computational efficiency, reducing memory usage and computational complexity during training and inference, making it highly suitable for real-time applications. The ASFF module functions by applying feature re-adjustment and adaptive fusion.

Feature Adjustment: $X^{a\to b}$ represents the adjustment of feature maps from a to b, where a and b belong to the set $\{1,2,3\}$. $ASFF-\det ect^l$ is obtained by combining and merging the semantic information from levels 1, 2, and 3, represented by different weights α, β, and γ. Formula (2) gives their definitions:

$$ASFF-\det ect^l = X^{1\to l}\times\alpha^l + X^{2\to l}\times\beta^l + X^{3\to l}\times\gamma^l \quad (2)$$

Adaptive Fusion: After feature resizing, adaptive fusion is applied. $x_{ij}^{a\to l}$ represents the feature vector located in (i, j), where a belongs to the set $\{1,2,3\}$. The feature fusion at a specific level l is expressed by equation (3):

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

y_{ij}^l represents the output feature at position (i, j) within the specific channel level l. Meanwhile, α_{ij}^l, β_{ij}^l, and γ_{ij}^l denote the spatial importance weights of three different feature mapping layers learned by the network prior to level l. These weights can be simple single values applied across all channels. The definition of each weight is provided in equation (4).

$$\alpha_{ij}^l = \frac{e^{\lambda_{\alpha_{ij}}^l}}{e^{\lambda_{\alpha_{ij}}^l} + e^{\lambda_{\beta_{ij}}^l} + e^{\lambda_{\gamma_{ij}}^l}} \quad (4)$$

The values of parameters α_{ij}^l, β_{ij}^l, and γ_{ij}^l are determined by control parameters $\lambda_{\alpha_{ij}}^l$, $\lambda_{\beta_{ij}}^l$, and $\lambda_{\gamma_{ij}}^l$ through the softmax function to compute these weights. The features $X^{1\to l}$, $X^{2\to l}$, and $X^{3\to l}$ assist in the calculation of the weights λ_α^l, λ_β^l, and λ_γ^l through the application of the 1×1-convolution layer. These weights are learned through the standard backpropagation method, similar to other neural networks. In the improved YOLOv8 model, the features from these three layers are dynamically combined at their respective scales, and the fused features are then input into the detection head for classification and detection of steel defects.

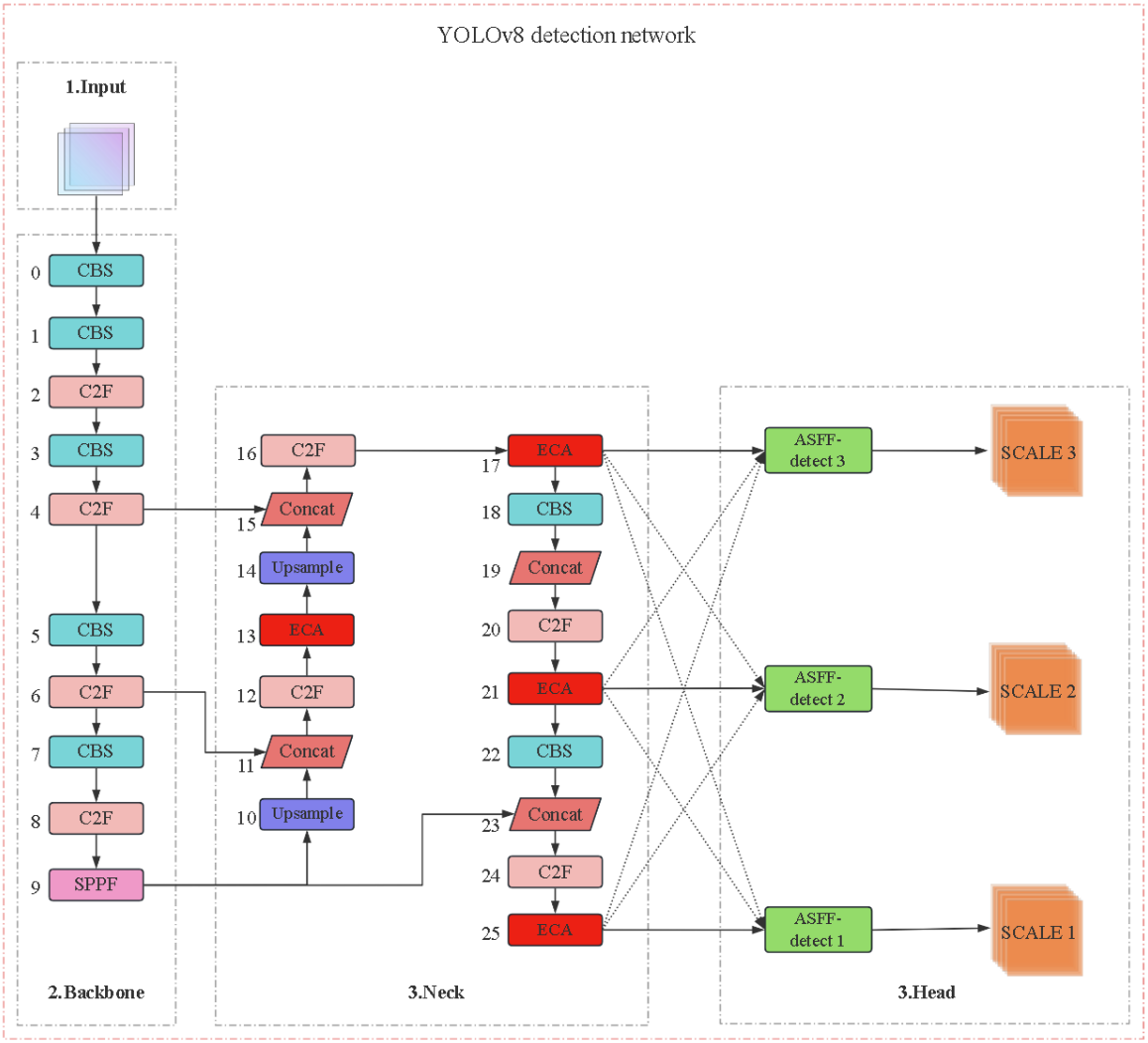


Fig. 5. Structure of EBA-YOLO

This enhancement enables the model to more effectively utilize features from different scales, improving its ability to detect subtle defects. By incorporating several improvements into the YOLOv8 architecture, we propose EBA-YOLO, which stands for ECA and Feature Fusion YOLOv8. The final network structure is shown in Fig. 5.

IV. Experiments and Results

This paper uses the NEU-DET dataset to evaluate the improved YOLOv8 model, and the results show that it achieved 87.4% on the mAP-50 metric.

A. Experimental setup

The code runs in a Windows 11 system environment, using an NVIDIA GeForce RTX 3060 Laptop GPU. During the experimental training, the SGD optimizer was used with an initial learning rate of 0.01 and a weight decay coefficient of 0.0005. The confidence thresholds were set to 0.5 for mAP-50 and 0.95 for mAP-95. The model was trained for 300 epochs per run, with a batch size of 32 and an image input size of 640×640.

The initial dataset used is the NEU-DET steel surface defect dataset, released by Northeastern University. This dataset was first introduced by He et al. in their paper [4], and contains 1,800 images across six defect categories: cracks, inclusions, patches, pitting surfaces, rolled surfaces, and scratches, with 300 images per category. All images in the dataset have a size of 200×200 pixels. After carefully studying the impact of preprocessing and data augmentation techniques on model performance, the dataset was expanded through horizontal flipping, vertical flipping, and Mosaic augmentation. Each image was resized to 640×640 pixels. After augmentation, the total number of images increased to 4,144, and the dataset was divided into approximately 86:7:7 ratios for training, testing, and validation sets, i.e., 3,544 images for training, 300 for testing, and 300 for validation. Figure 1 shows examples of various defects in the baseline dataset. The grayscale images demonstrate that even within the same defect category, significant variations in appearance can exist. For instance, in the scratch defect images, there are both horizontal and vertical scratch patterns.

This paper uses precision, recall, mean average precision (mAP), and frames per second (FPS) as key performance metrics. Precision measures the accuracy of positive sample predictions and reflects the proportion of correct identifications among all positive predictions. Recall, on the other hand, measures the model's ability to capture all relevant instances, indicating the proportion of correct identifications among all actual positive samples. mAP, as a composite metric, evaluates the overall performance of a target detection or recognition system by calculating the

average precision across multiple categories, providing a comprehensive assessment of the model's quality. Lastly, FPS indicates the model's processing speed, which is crucial for real-time applications.

These metrics together provide a comprehensive evaluation of the model's effectiveness, balancing accuracy, efficiency, and comprehensiveness, making them suitable for a wide range of applications from autonomous driving to medical image analysis. The mathematical representations of these metrics are as follows:

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{7}$$

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \tag{8}$$

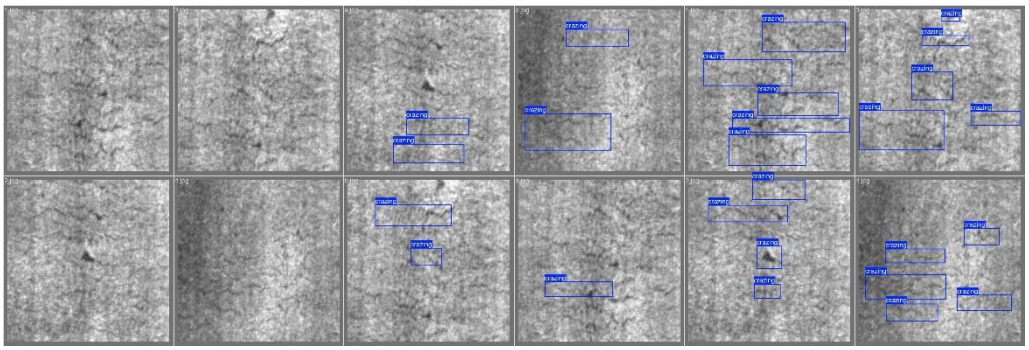
In the text, TP, FP, FN, and TN represent the numbers of true positives, false positives, false negatives, and true negatives, respectively. Precision and Recall are defined by formulas (5) and (6). A detection is considered accurate when the predicted defect category is correct and the Intersection over Union (IoU) exceeds a certain threshold (set to 0.5 in our experiments). Accuracy can be calculated based on TP, TN, FN, and FP, as shown in formula (7). Average Precision (AP) is defined by formula (8), where R_n and P_n represent the recall and precision at the n -th threshold, respectively. The mean Average Precision (mAP) is the average of the AP values for all instances. AP corresponds to the area under the Precision-Recall (P-R) curve.

B. Comparative experiments on NEU-DET dataset

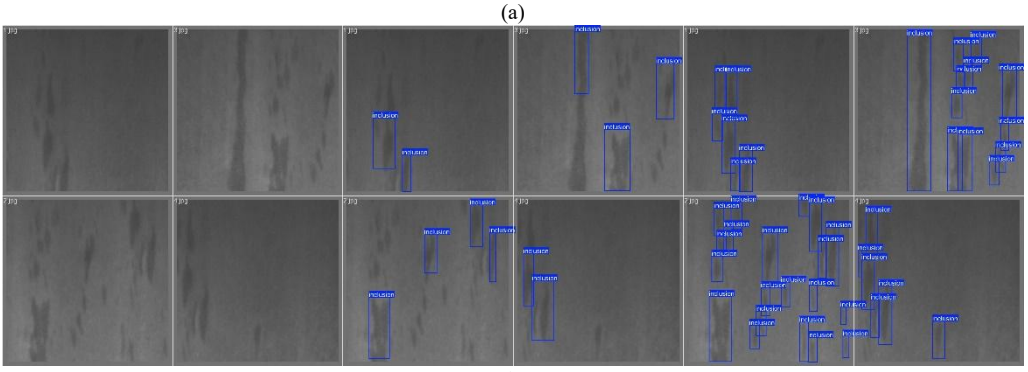
In this study, the EBA-YOLO model was compared with the original YOLOv8s model. The main comparison indicators included precision, recall, and mAP to evaluate the accuracy of the model, and frames per second (FPS) to evaluate the inference speed.

Table I Quantitative comparison between original YOLOv8 and EBA-YOLO

Model	Class	Precision (%)	Recall (%)	Map-50 (%)	Map50-95 (%)
YOLOv8	crazing	50.3	50.2	48.6	21.4
	inclusion	71.9	90.5	86.8	46.3
	patches	85.1	91.3	94.4	60.8
	pitted_surface	75.3	72.4	81.2	47.5
	rolled-in_scale	65.3	63.5	68.3	39.6
	scratches	89.6	92.1	95.4	73.2
EBA-YOLO	crazing	68.2	67.3	67.2	31.6
	inclusion	84.1	91.2	85.1	53.7
	patches	90.3	95.1	97.6	64.2
	pitted_surface	89.5	77.2	88.2	52.9
	rolled-in_scale	83.4	72.1	76.4	50.8
	scratches	94.7	91.4	96.3	77.2



(a)



(b)

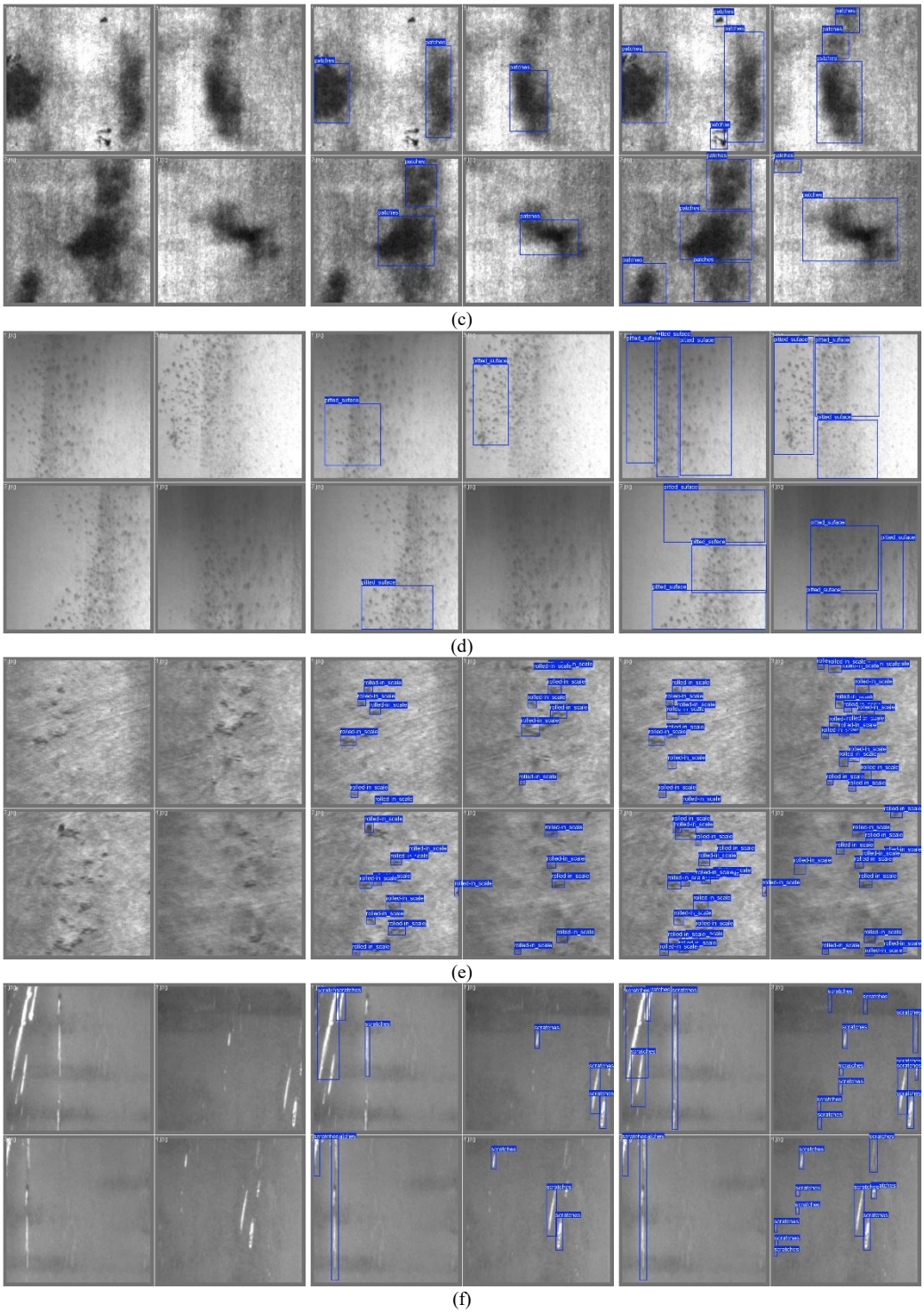


Fig. 6. Performance comparison of the models on the test dataset. (Left) Original dataset, (Middle) YOLOv8 reasoning, (Right) EBA-YOLO reasoning. (a)Crazing; (b) Inclusion; (c) Patches; (d) Pitted_surface; (e) Rolled-in_scale; (f) Scratches

After multiple rounds of rigorous testing and training, the final improved YOLOv8 model was applied to the test image set, and its detection results were compared with those of the original YOLOv8 model. To better illustrate the performance improvements of EBA-YOLO across various categories, the test images were specifically selected from each category.

Fig. 6 presents this visual analysis, with one example chosen per category. By incorporating feature fusion and attention mechanisms into the neck and head components of the EBA-YOLO model, the model demonstrated superior detection capabilities for defects of varying scales (e.g., crazing, patches, rolled-in scale, and scratches in Fig. 6) and different orientations (e.g., inclusion and rolled-in scale in Fig. 6), outperforming the original YOLOv8 model. Moreover, EBA-YOLO also exhibited enhanced overall detection performance (e.g., pitted surface in Fig. 6). Table 1 provides the quantitative analysis results of these two models. As shown in the table, EBA-YOLO outperformed the original YOLOv8 model in terms of mAP for every category. The total mAP-50% across all categories improved by 6.7%, with an 11.5% increase in precision and a 3.4% improvement in recall.

C. Ablation experiment

In computer vision models, when enhancing model performance by introducing new modules, ablation studies

are employed to systematically evaluate the specific contributions of these modules to the overall model performance. In these studies, individual modules or components in the model are gradually removed and their impact on model accuracy, robustness, and efficiency is observed. Through these rigorous experiments, we can gain an in-depth understanding of the relative importance of each module, thereby guiding us to further optimize the model architecture and improve the model's capabilities. Therefore, ablation research is an indispensable tool in the process of iterative model development to ensure that the addition of new modules indeed makes a significant contribution to performance improvement in the field of computer vision. These experiments were also performed on the NEU-DET data set.

This paper uses 7 different variants, each of which represents an improvement on the YOLOv8 architecture, as shown in Table II:

Table II List of changes proposed by ablation studies

Variation	ECA module	BiFPN	ASFF detection
1	☐	☐	☐
2	☑	☐	☐
3	☐	☑	☐
4	☐	☐	☑
5	☑	☐	☑
6	☑	☑	☐
7	☑	☑	☑

As shown in Table II, the ECA module refers to the ECA attention mechanism added to the neck portion of the architecture. BiFPN represents the BiFPN connection module implemented in the neck for feature fusion. ASFF detection indicates the ASFF detection head added prior to the prediction module in the head of the architecture. The improvements are summarized as follows:

(1) Variant 1: The original YOLOv8 model with no modifications.

(2) Variant 2: Improved solely by introducing the ECA module.

(3) Variant 3: Improved solely by incorporating BiFPN connections, implemented in the neck portion of the architecture.

(4) Variant 4: Improved solely by adding an ASFF detection head before the prediction head in the architecture's head section.

(5) Variant 5: Improved through the combination of introducing the ECA module (in the neck section) and the ASFF detection module (in the head section).

(6) Variant 6: Improved through the combination of introducing the ECA module (in the neck section) and the BiFPN connection (also in the neck section).

(7) Variant 7: Further improved by introducing the BiFPN connection into the architecture of Variant 5, achieving the final enhancement.

Table 3 presents the specific improvement results in terms of Precision, Recall, and mean Average Precision (mAP) across all six variants. Each category is represented by the letters a to g, denoting all categories, crazing, inclusion, patches, pitted_surfaces, rolled-in_scale, and scratches, respectively. By analyzing Table 3, several observations can be drawn. Each individual component contributes to better model learning and performance improvement, except for the BiFPN connection module. This module demonstrates its effectiveness when paired with the ECA attention mechanism or the ASFF module, further enhancing their performance. As shown in the table, the addition of the ECA module and the ASFF module increased the mAP of the baseline model by 3.2% and 3.3%, respectively. However, when the BiFPN connection module was integrated into Variants 2 and 4, it further improved the mAP by 2.6% and 1.5%, respectively, compared to the baseline model. By combining all three modules into the baseline model, the overall mAP increased by 7.1%. This demonstrates that the EBA-YOLO model exhibits superior performance and enhanced learning capabilities.

V. Conclusion

Aiming at the problem of complex surface defect detection of steel, an EBA-YOLO target detection model is proposed.

(1) By introducing the ECA mechanism module into the neck part of the network, the dynamic weighting strategy is used to improve the inter-channel feature expression ability and optimize the channel attention mechanism of the detection model. Compared with traditional attention mechanisms, the detection model enhances small target detection and target discrimination capabilities in complex scenes without significantly increasing computational costs.

(2) By leveraging BiFPN's hierarchical information integration capability and ASFF's dynamic adjustment ability, the efficient fusion of BiFPN and ASFF enables complementary optimization of multi-scale features. This enhances the detection model's precision and robustness across various scenarios.

(3) Through comparative experimental studies, it was found that the mAP-50 (mean average precision at an IoU threshold of 0.5) increased by 6.7%, overall precision improved by 11.5%, and recall rose by 3.4%. Meanwhile, EBA-YOLO maintained a detection speed of 98 frames per second (FPS), ensuring its feasibility for real-time detection applications.

Table III Quantitative comparison of different proposed changes in ablation experiments

Variation	Precision (%)							Recall (%)							mAP (all)
	a	b	c	d	e	f	g	a	b	c	d	e	f	g	
1	72.4	50.9	74.1	84.3	76.3	62.7	87.5	77.6	52.1	93.1	91.4	71.2	65.8	94.2	78.5
2	75.6	62.1	76.4	81.9	83.5	69.7	85.9	81.4	62.9	91.7	94.2	77.4	65.8	94.8	82.1
3	73.3	47.8	74.9	84.3	87.1	63.4	87.4	78.2	47.3	91.4	92.4	70.7	70.1	90.6	79.1
4	78.1	60.2	76.2	87.4	89.3	71.2	90.5	81.6	59.1	91.4	95.8	74.5	68.8	96.0	82.5
5	81.1	65.3	79.2	88.5	85.1	75.1	90.3	82.1	71.1	91.8	97.6	73.5	72.6	91.9	84.4
6	77.4	63.2	76.5	85.1	88.6	71.0	87.1	78.9	55.4	90.9	95.2	75.6	68.5	92.1	83.3
7	84.5	68.2	82.6	88.5	89.1	81.0	92.9	80.9	66.7	87.9	95.1	78.2	72.5	91.8	85.5

Therefore, the significant improvements of EBA-YOLO highlight its effectiveness in enhancing detection capabilities while ensuring high efficiency in real-world applications.

REFERENCES

[1] S. Ghorai, A. Mukherjee, M. Gangadaran, & P. K. Dutta, "Automatic defect detection on hot-rolled flat steel products," IEEE Transactions on Instrumentation and Measurement, vol.62, no.3, pp. 612-621, 2012.

[2] W. Zhang, F. Z. Gao, J. C. Huang, et al. "Global Prescribed-Time Stabilization for a Class of Uncertain Feedforward Nonlinear Systems," IEEE Transactions on Circuits and Systems II: Express Briefs, vol.70, no.4, pp. 1450-1454, 2022.

[3] W. Zhang, X. P. Ye, L. H. Jiang, et al. "Output feedback control for free-floating space robotic manipulators base on adaptive fuzzy neural network," Aerospace Science and Technology, vol.29, no.1, pp. 135-143, 2013.

[4] Y. L. L, W. Zhang, and T. Zhou, "Machine Health-Driven Dynamic Scheduling of Hybrid Jobs for Flexible Manufacturing Shop," International Journal of Precision Engineering and Manufacturing, vol.24, no.5, pp. 797-812, 2023.

[5] W. Zhang, J. M. Shen, X. P. Ye, and S. Zhou, "Error model-oriented vibration suppression control of free-floating space robot with flexible joints based on adaptive neural network," Engineering Applications of Artificial Intelligence, no.114, pp. 105028, 2022.

[6] W. Zhang, N. M. Qi, J. Ma, and A. Y. Xiao, "Neural integrated control for free-floating space robot with changing parameters," Science China: Information Science, vol.4, no.10, pp. 2091-2099, 2011.

[7] W Zhang, X Ye and X Ji, "RBF neural network adaptive control for space robots without speed feedback signal," Transactions of the Japan Society for Aeronautical and Space Sciences, vol.56, no.6, pp. 317-322, 2013.

[8] L Jiang, W Zhang, J Shen, et al. "Vibration Suppression of Flexible Joints Space Robot based on Neural Network," IAENG International Journal of Applied Mathematics, vol.52, no.4, pp. 776-783, 2022.

[9] Z You, W Zhang, J Shen, et al. "Adaptive neural network vibration suppression control of flexible joints space manipulator based on H∞ theory," Journal of Vibroengineering, vol.25, no.3, pp. 492-505, 2023.

[10] W Zhang, N Qi and H Yin, "Neural-network tracking control of space robot based on sliding-mode variable structure," Control Theory and Applications, vol.28, no.9, pp. 1141-1144, 2011.

[11] Y Hu and W Zhang, "Modeling framework for analyzing midair encounters in hybrid airspace where manned and unmanned aircraft coexist," Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering, vol.233, no.15, pp. 5492-5506, 2019.

[12] W.Zhang, Z.Wen, Y.Ye, and S.Zhou, "Structural mechanics analysis of bolt joint of rigid flexible coupling manipulator," Journal of Measurements in Engineering, vol.10, No.2, pp. 93-104,2022.

[13] W Zhang, H Chen and Z Wen, "Sliding Mode Control of Flexible Joint Space Robot for Unknown Target,"Aerospace Control and Applications, vol.47, no.3, pp. 49-56, 2021

[14] W Zhang, Y Shang Q Sun, et al, "Finite-Time Stabilization of General Stochastic Nonlinear Systems with Application to a Liquid-Level System," IAENG International Journal of Applied Mathematics, vol.51, no.2, pp. 295-299, 2021.

[15] Z You, D Yi, Z Fang, et.al. "Image Enhancement ANPSO Processing Technology Based on Improved Particle Swarm Optimization Algorithm,"IAENG International Journal of Computer Science, vol.51, no.11, pp. 1781-1792. 2024.

[16] W Zhang, S Ding, D Yi, et.al. "Improved Sliding Mode Variable Structure Control of Robot Manipulators with Flexible Joints based on Singular Perturbation," 2023 International Conference on Telecommunications, Electronics and Informatics (ICTEI), pp. 288-291. 2023.

[17] L Zhu, Xiaochen Huang, X Wu, et.al. "Neural Network Control of Robot under Wheel Slip Conditions based on Observer," Engineering Letters, vol.32, no.10, pp. 2041-2051. 2024.

[18] W.Zhang, Y.Zhu, "Control of free-floating space robotic manipulators base on neural network," International Journal of Computer Science Issues (IJCSI), vol.9, No.6, pp. 322,2012.

[19] Y Fang, W Zhang and X Ye. "Variable Structure Control for Space Robots Based on Neural Networks," International Journal of Advanced Robotic Systems, vol.11, no.3, pp. 35-42, 2014.

[20] W.Zhang, N.Qi, and Y.Li, "Output feedback PD control of robot manipulators dispenses with model base on fuzzy-basis-function-network," Journal of National University of Defense Technology, vol.32, No.6, pp. 163-170,2010.

[21] J. Ma, W. Zhang, and H. P. Zhu, "Adaptive Control for Robotic Manipulators base on RBF Neural Network," TELKOMNIKA, vol.11, no.3, pp. 521-528, 2013.

[22] W.Zhang, Y. Fang, X. Ye, "Adaptive Neural Network Robust Control for Space Robot with Uncertainty," TELKOMNIKA, vol.11, No.3, pp. 513-520, 2013.

[23] W. Zhang, X. Ye, L. Jiang, and Y. Fang, "Robust control for robotic manipulators base on adaptive neural network," The Open Mechanical Engineering Journal，vol.20, no.8, pp. 497-502, 2014.

[24] G. Li, M. Muller, A. Thabet, & B. Ghanem, "Deepgcns: Can gcns go as deep as cnns?" Proceedings of the IEEE/CVF international conference on computer vision pp. 9267-9276, 2019.

[25] H. Luo, P. Wang, H. Chen, & V. P. Kowelo, "Small object detection network based on feature information enhancement," Computational Intelligence and Neuroscience, vol.2022, no.1, pp. 6394823, 2001.

[26] J. Liu, X. Zhu, X. Zhou, S. Qian, & J. Yu, "Defect detection for metal base of TO-Can packaged laser diode based on improved YOLO algorithm," Electronics, vol.11, no.10, pp. 1561, 2022.

[27] M. Sohan, T. Sai Ram, R. Reddy, & C. Venkata, "A review on yolov8 and its advancements," International Conference on Data Intelligence and Cognitive Informatics, pp. 529-545, 2024.

[28] M. Tan, R. Pang, & Q. V. Le, "Efficientdet: Scalable and efficient object detection," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10781-10790, 2020.

[29] J. Terven, D. M. Córdova-Esparza, & J. A. Romero-González, "A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas," Machine Learning and Knowledge Extraction, vol.5, no.4, pp. 1680-1716, 2023.

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, & A. C. Berg, "Ssd: Single shot multibox detector," Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 21-37, 2016.

[31] Y. Wang, C. Wang, H. Zhang, Y. Dong, & S. Wei, "Automatic ship detection based on RetinaNet using multi-resolution Gaofen-3 imagery," Remote Sensing, vol.11, no.5, pp. 531, 2019.

[32] S. Ren, K. He, R. Girshick, & J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE transactions on pattern analysis and machine intelligence, vol.39, no.6, pp. 1137-1149, 2016.

[33] H. Law, & J. Deng, "Cornernet: Detecting objects as paired keypoints," Proceedings of the European conference on computer vision (ECCV), pp. 734-750, 2018.

[34] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, & Q. Tian, "Centernet: Keypoint triplets for object detection," Proceedings of the IEEE/CVF international conference on computer vision, pp. 6569-6578, 2019.

[35] A. Vaswani, "Attention is all you need," Advances in Neural Information Processing Systems, 2017.

[36] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, & Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11534-11542, 2020.

[37] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, & S. Belongie, "Feature pyramid networks for object detection," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117-2125, 2017.

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

[39] X.C. Huang, Z.P. You, W.H. Zhang, et al. "Stabilization control of wheeled mobile robot based on neural sliding mode under wheel slip conditions". International Journal of Dynamics and Control. vol. 13, no.2, pp.1-12, 2025.

[40] D.J. Yi, Z. P. You, W.H. Zhang. "Image Enhancement CHPSO Processing Technology Based on Improved Particle Swarm Optimization Algorithm". IAENG International Journal of Computer Science, vol. 52, no.1, pp.130-142, 2025.

[41] R. Chen, Z. P. You, W.H. Zhang. "Adaptive Fuzzy Sliding Mode Control for Nonlinear Systems with Unknown Dead-zone". IAENG International Journal of Applied Mathematics, vol. 54, no.12, pp.2588-2595, 2025.