ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

MBD-YOLO: Steel Defect Detection Model Combining Multi-scale Features and Routing Attention

Xvyue Zhang[1, 2, a], Baoping Wang[1, 2, b, *], Qin Sun[1, 2, c], Da Zhao[1, 2, d]

[1]Shandong Jiaotong University, Jinan 250357, China

[2]Shandong Provincial Engineering Research Center for Transportation Construction Equipment Technology and Intelligent Construction, Jinan 250357, China

[a]15753800993@163.com, [b]wangbaoping@sdjtu.edu.cn, [c]sunqin@sdjtu.edu.cn, [d]zhao_da@yeah.net

[*] Corresponding author

Abstract

With the growing demand for high-quality steel, accurate and efficient surface defect detection is becoming increasingly important. Traditional methods often face challenges such as missed detection, inaccurate localization, and poor performance in identifying small defects. This study introduces MBD-YOLO, an enhanced defect detection model based on the YOLOv8 framework. The model incorporates a novel attention module, Bivo, in the feature fusion stage, which captures both global and local features while enabling effective cross-scale information exchange. This significantly improved the accuracy and performance of small-object detection in complex backgrounds. Second, a new DRM detection head was designed. Through its multi-context enhancement and its dual-branch pooling combined with upsampling, the global and local context information is enhanced, and the detection ability of small target, low contrast and direction sensitive defects is significantly improved. The backbone integrates MobileViT, combining CNN and Vision Transformer architectures to improve feature extraction for small targets. Experiments on the NEU-det dataset show that MBD-YOLO achieves an mAP@0.5 of 85.7%, outperforming mainstream models. A large number of experiments not only confirm the combined effectiveness of the proposed modules, provide excellent performance in various defect categories, but also verify the generalization of the model. MBD-YOLO provides a robust and high-precision solution for steel defect detection that meets the needs of modern industrial applications.

Keywords

Deep learning; steel defect detection; YOLO; attention mechanism; Small object detection.

1. INTRODUCTION

With the rapid growth of industrial development in China, steel has become a crucial material in industrial manufacturing because of its excellent surface quality and mechanical properties. It is widely used in various industrial production industries. Concurrently, the demand of enterprises and consumers for higher quality has intensified, leading to increasingly stringent standards for steel. Consequently, the production of high-quality steel has emerged as an urgent issue requiring immediate resolution to meet these high expectations[1].

The steel production process involves a multitude of technical factors, and numerous types of surface defects may arise, including "cracks," "scratches," "patches," "inclusions," "pitting,"

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

and "rolled-in scale," as illustrated in Figure 1. In actual industrial settings, these surface defects on steel not only pose potential safety hazards and risks of production-related accidents, such as conveyor belt failures and crosshead fractures in machinery, but also contribute to severe wear and tear of the rolling equipment. These issues can lead to significant economic losses for manufacturers. Traditional manual inspection methods for detecting steel surface defects are often inadequate because they struggle to accurately distinguish defect types and pinpoint their exact locations[2].

The rapid evolution of neural networks and machine learning algorithms has catalyzed transformative innovations in industrial quality assurance. In steel manufacturing, data-driven defect detection frameworks have demonstrated significant potential for overcoming longstanding technical barriers. These intelligent systems achieve dual optimization of detection precision and operational effectiveness, ultimately contributing to strategic quality enhancement across the production chain.

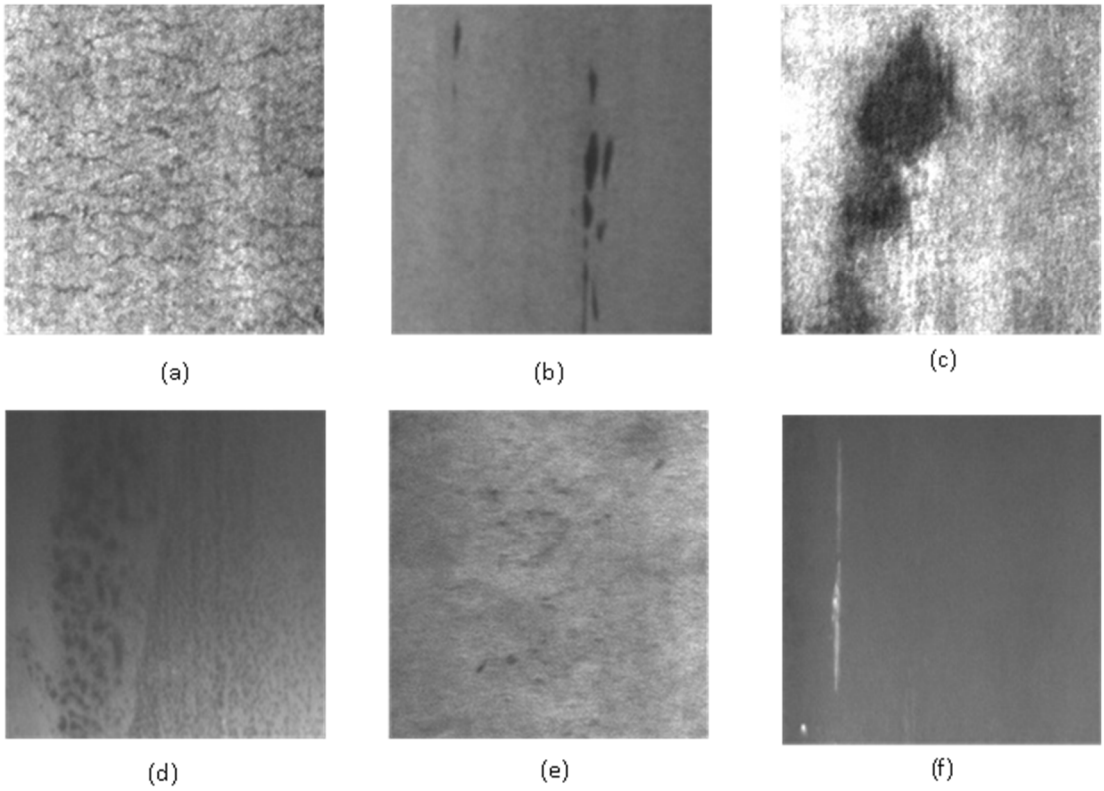

Figure 1. (a) Cracks (b) Inclusions (c) Patches (d) Surface pitting (e) Rolled-in scale (f) Scratches

In the initial stages of research, applying deep learning to defect detection was approached primarily as a pattern recognition problem. Two primary methodologies emerged to tackle this detection task: one based on Support Vector Machines (SVMs) and the other employing Convolutional Neural Networks (CNNs)[3]. Support Vector Machines (SVMs) are supervised learning models designed for classification and regression tasks. The algorithm identifies an optimal separation hyperplane in high-dimensional space by solving a constrained optimization problem. This hyperplane maximizes the geometric margin between the closest samples of distinct classes (support vectors), thereby enhancing generalization performance while minimizing structural risk. As a “shallow learner,” SVMs generally demonstrate strong generalization capabilities compared to other methods. However, traditional SVMs face

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

limitations in real-world detection tasks, such as sensitivity to target variations and challenges in handling complex scenes[4].

The evolution of deep learning has led to the widespread adoption of convolutional neural network (CNN)-based detection fashions in contemporary defect detection systems, which possess the functionality to autonomously analyze function representations. These CNN-based fashions provide end-to-end benefits, resulting in stronger overall performance in defect detection applications. CNN-based detection methods can be classified into two types, primarily based on the inclusion of an area suggestion stage: location notion detection methods and single-stage detection models. Early models, such as those in the R-CNN household (including R-CNN[5], Fast R-CNN[6], and Faster R-CNN), are labeled as place thought detection models, in addition to being referred to as two-stage detection models. Although these two-stage fashions can accomplish excessive accuracy, they are hindered by gradual detection speeds, rendering them insufficient for real-time detection in industrial manufacturing settings. In response to this limitation, subsequent research has led to the improvement of single-stage detection models, which streamline the detection process by framing it as a regression problem with a decreased number of parameters, as depicted in Figure 2.

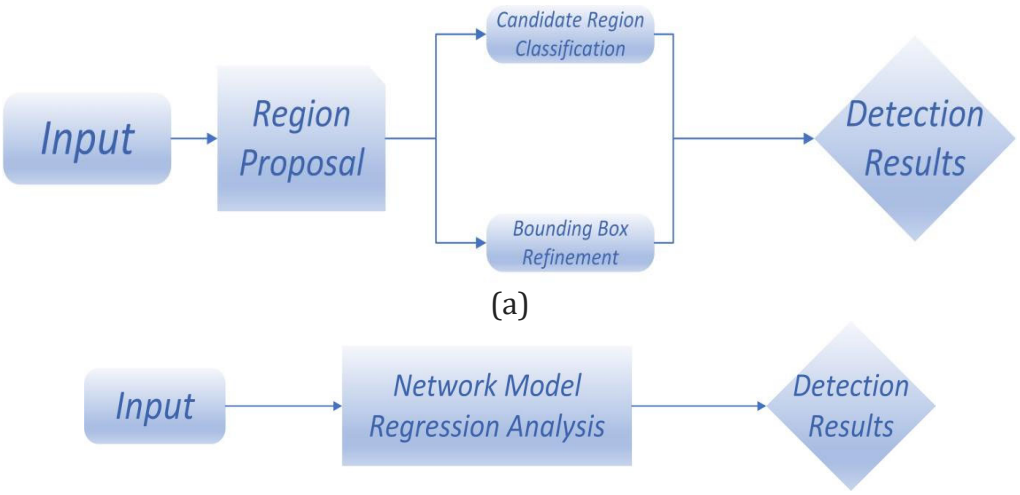

Figure 2. (a)Two-stage process (b)Single-stage process

In 2016, W. Liu et al. delivered the Single Shot MultiBox Detector (SSD)[7] algorithm for object detection. The essential precept of SSD is to simultaneously predict the region and class of aims using a single neural network. This methodology allows the model to execute the detection venture in a single ahead pass, attaining excessive effectivity and accuracy, which renders it specifically appropriate for real-time functions and cellular implementations. Consequently, SSD has been massively adopted in defect detection. In the same year, the inaugural model of the You Only Look Once (YOLO) algorithm was released. YOLO conceptualizes object detection as a regression problem, making predictions throughout a complete photograph using a single neural network[8]. It segments the picture into smaller grid cells, predicting bounding packing containers and class possibilities for each cell. The entire image is processed via a spine community composed of convolutional neural networks, which generates a function map of the entire image. These function maps are built-in through a neck layer to seize multi-scale characteristic information. Since the launch of YOLOv5 in 2020, the YOLO sequence has been utilized in industrial imaginative and prescient inspection, establishing itself as the predominant deep learning model for object detection. In 2023, YOLOv8 was introduced, incorporating greater environment-friendly statistics augmentation and coaching techniques

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

that improved the model's generalization and robustness. This new release improves the overall detection performance for a range of object sizes, specifically for small objects and dense targets, while optimizing the inference velocity and deployment effectiveness through model compression and acceleration techniques. Additionally, YOLOv8 extends its skills to guide multitask learning, permitting it to concurrently tackle duties such as object detection and occasion segmentation[9]. Owing to its excessive efficiency, the YOLO sequence has been broadly applied in visible inspection for industrial production, evolving into the most commonplace deep learning model for object detection. In 2020, Hatab et al. proposed a technique for floor defect detection utilizing the YOLO network, specifically leveraging the YOLOv3 model to obtain excellent effectiveness and accuracy. Their methodology underwent extensive testing, confirming the effectiveness of the YOLO community in floor defect detection. Experimental results indicate that YOLO can successfully discover several types of floor defects at excessive speeds, making it appropriate for realistic applications[10]. In 2023, Zhao et al. developed a better community model based entirely on YOLOv5, termed "RDD-YOLO," which was optimized for the detection of defects in steel. The integration of ResNet as the spine and the implementation of a function pyramid community in the neck of the structure resulted in a more profound community hierarchy. This model attained an accuracy of 81.1% on the publicly available NEU-det dataset[11]. Subsequent research focused on improving YOLOv7 by incorporating the Bi-directional Feature Pyramid Network (BIFPN) shape alongside the ECA interest mechanism, in addition to changes to the loss function. This revised model achieved an accuracy of 81.9% on the NEU-det dataset[12]. Furthermore, the improvement of "WSS-YOLO" represented an development over YOLOv8, introducing dynamic snake convolution and VOV-GSCSP, which contributed to a deeper neck community and accelerated accuracy while maintaining detection efficiency, culminating in an accuracy of 82.3% on the NEU-det dataset[13].

This study provides a more desirable methodology for the detection of floor defects in metals using the YOLOv8 framework. To improve the identification of small targets, a Bivo module was built into the neck of the community architecture. This module employs a bi-level Routing Attention mechanism, which allows Bivo to efficiently seize each world and neighborhood characteristic information, thereby augmenting the traditional interest mechanism and increasing the model's sensitivity to faulty regions. Furthermore, a novel detection head, termed DRM, was developed to amalgamate function maps throughout several scales. Through convolution, upsampling, downsampling, and softmax operations, the DRM module allows the integration and reorganization of points from unique levels, thereby improving detection accuracy.

In the major model, the backbone was updated by replacing the original component with MobileViT, which combines the strengths of convolutional neural networks (CNNs) and visual transformers (ViTs) to enhance detection accuracy while maintaining efficiency. Experimental results on the public NEU-det dataset confirmed the effectiveness of the proposed module and improvement strategies, with a detection accuracy of 85.7%, surpassing the overall performance of the majority of current metal defect detection models.

2. RELATED WORK

2.1. Introduction to the Baseline Model

The YOLOv8 network models are categorized into five types based on model size: YOLOv8n, YOLOv8m, YOLOv8s, YOLOv8l, and YOLOv8x. Among these, YOLOv8n demonstrates superior performance in detecting small targets and offers high detection efficiency, making it the chosen baseline model. The structure of the YOLOv8n network is shown in Figure 3.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

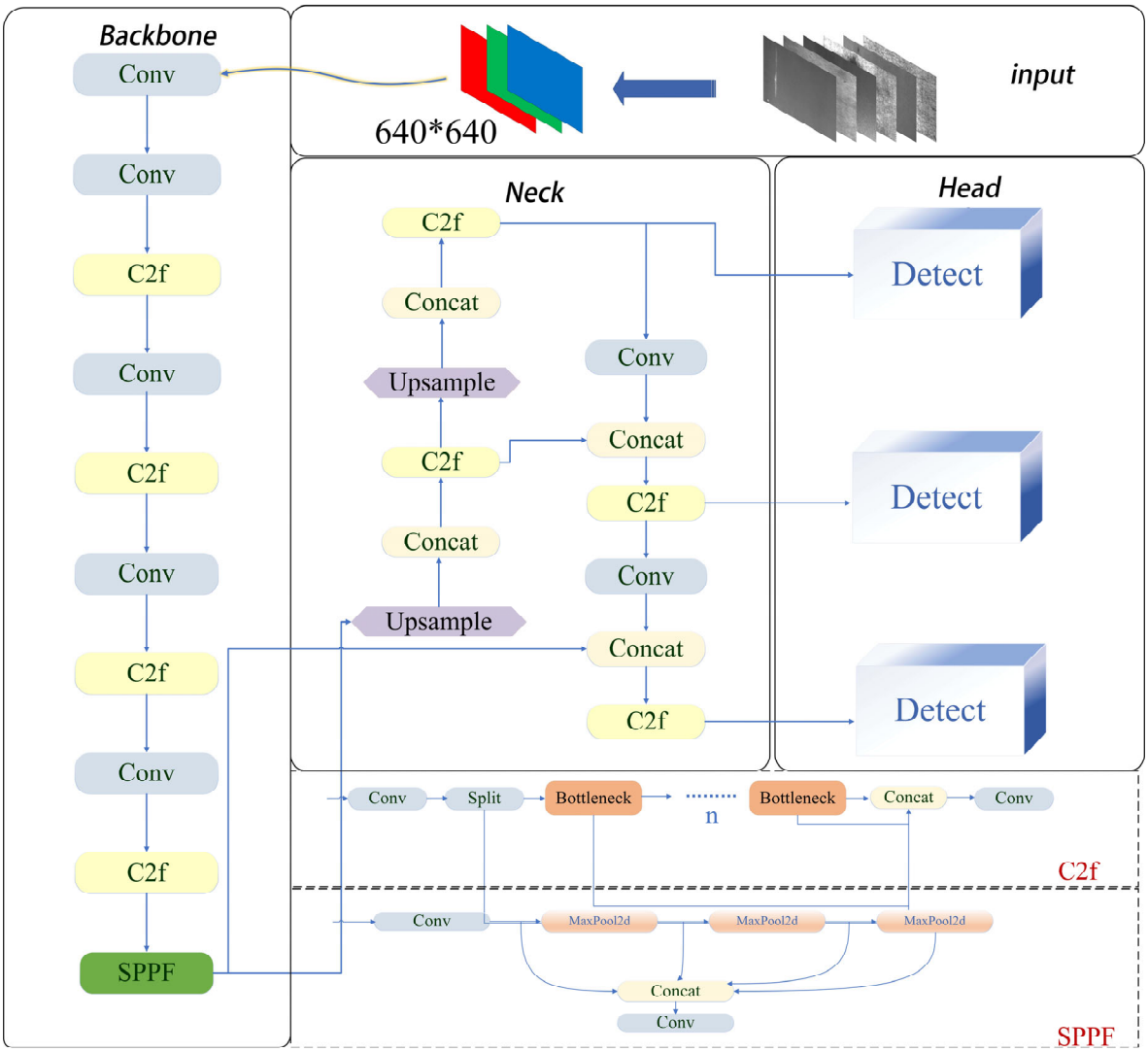

Figure 3. Structure of the YOLOv8 network for the baseline model

YOLOv8 comprises four main components: Input, Backbone, Neck, and Detect. The Backbone component, in particular, includes the basic convolutional module, an enhanced C2f module, and the SPPF module. Together with a downsampling strategy, these modules facilitate the extraction of multi-scale image features at various levels. The most frequently occurring module in the entire model architecture is the convolution module, Conv, whose structure is illustrated in Figure 4.

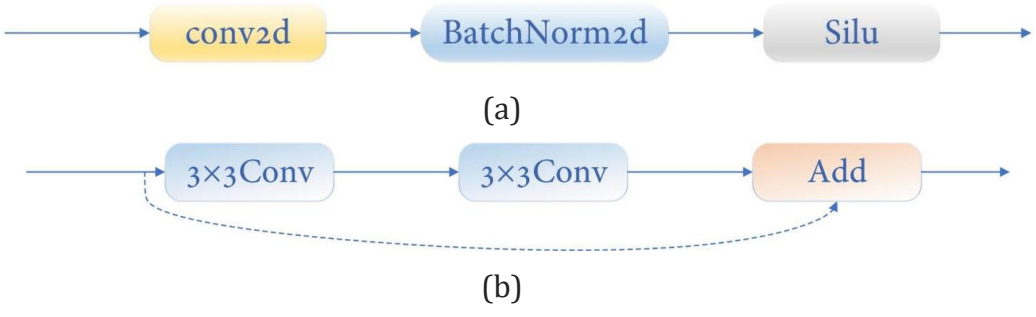

Figure 4. (a)Conv module，the conv2d serves as the convolutional layer.(b)Bottleneck, residual connection module

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

The basic module in the architecture, (a), consists of a convolutional layer (conv2d), a proposed normalization layer (BatchNorm2d), and an activation function (Silu). The residual connection module, Bottleneck, in YOLOv8 is composed of two connected convolutional layers, with a skip connection added between them to address the vanishing gradient problem. This framework improves the capacity for feature extraction while preserving computational efficiency, thereby enabling the model to acquire more resilient features.

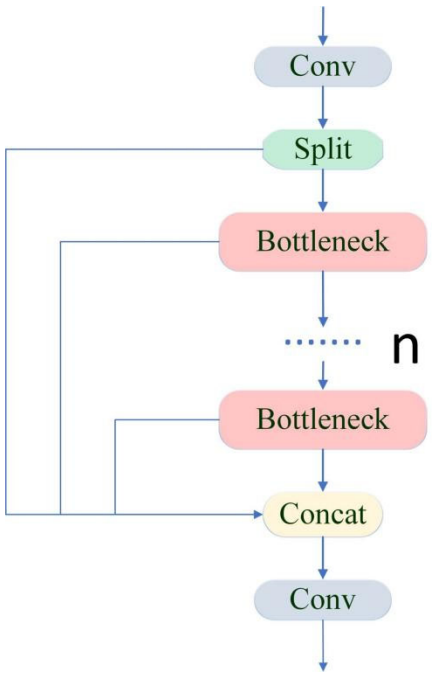

Figure 5. The C2f module structure

The YOLOv8 model introduces an innovative C2f module, consisting of Conv, Bottleneck, and Concat components, as illustrated in Figure 5. Inspired by CSPNet[14], the C2f module begins with an initial convolution (Conv) block that processes the input image and generates intermediate feature maps. The feature maps were divided into two branches: one was directly forwarded to the final concat block, while the other underwent a series of bottleneck operations, including convolution, normalization, and activation. The processed features were then merged with the directly passed features in the concat block and further refined through a convolutional layer to generate the final representation, minimizing the computational and memory costs while maintaining an efficient feature representation through multi-stage fusion. For multi-scale detection, the fusion mechanism facilitates feature exchange across different levels, improving feature reuse and allowing shallower layers to capture deeper-level information.

The backbone network incorporates the enhanced spatial pyramid pooling module (SPPF), an improvement over the original SPP module[15], as depicted in Figure 6. SPPF is designed to expand the receptive field and capture multi-scale features efficiently while keeping computational overhead minimal, thereby boosting object detection performance. Initially, the input feature maps pass through a 1×1 convolutional layer, ensuring the output channel count matches the input while reducing computational costs and facilitating early feature extraction. Next, three Maxpool2d layers process the feature maps to gather information at different scales. Finally, these pooled features are concatenated along the channel dimension within the Concat block and further refined through a convolutional layer, effectively integrating multi-scale features to generate the final output feature map.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

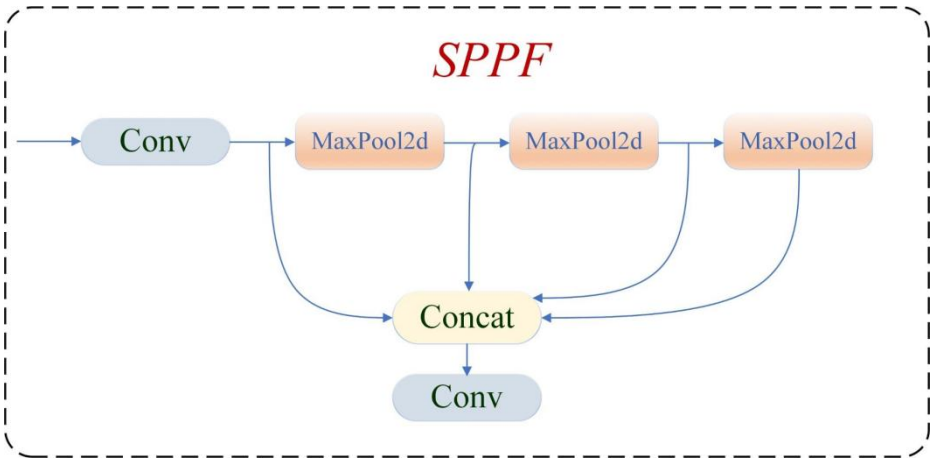

Figure 6. The SPPF module structure

The Neck factor of YOLOv8 was engineered to include the technique and integrate the multi-scale points acquired from the backbone. This element comprises standards from the Feature Pyramid Network (FPN)[16] and Path Aggregation Network (PANet)[17]. The characteristic pyramid structure enhances the semantic richness of high-level characteristic maps through top-down pathways, which are due to this fact mixed with low-level function maps. This integration helps the functionality of the network to efficiently tackle multi-scale object detection challenges. The PANet framework enhances this system by amalgamating neighborhood facts from low-level function maps with world facts from high-level function maps using a bottom-up pathway, thereby improving function illustration and enhancing the flow of records throughout the function maps. This synergistic diagram permits the community to effectively leverage characteristic data at all levels, thereby advertising correct detection outcomes. The facets produced using the neck are then transmitted to the head section, where they undergo additional processing to generate the remaining detection outputs, which consist of bounding container coordinates, class predictions, and self-belief scores. The Head is specially designed to differentiate multi-scale predictions from classification and regression tasks, thereby enhancing detection efficiency. Furthermore, redundant predictions are eradicated through the utility of non-maximal suppression (NMS). In conclusion, the operational workflow of YOLOv8 commences with the extraction of multi-scale facets from the input image using the backbone, observed using function fusion in the neck using the FPN and PANet frameworks. Ultimately, the Head executes object classification and bounding field prediction, yielding the object's class, position, and self-assurance level, while redundant bounding packing containers are suppressed through NMS, resulting in unique object detection.

2.2. Model Improvement Program

2.2.1 Layer 2 Routing Attention Module

Steel defect detection targets are characterized by complex shapes, challenging backgrounds, blurred images, and small sizes, leading to reduced detection accuracy in industrial inspection scenarios. In YOLOv8, the Feature Pyramid Network (FPN) aids in fusing features at various levels; however, its attention mechanism primarily focuses on specific channels or spatial regions, overlooking the interaction of global and local information across feature levels. This limitation hinders the model's ability to capture sufficient global context in complex scenes or in dense backgrounds containing numerous small targets. To address this, the feature fusion component of YOLOv8 incorporates the Bivo module, compared with the traditional attention mechanism of YOLOv8, Bivo adopts the attention mechanism framework in VT model, and uses the mechanism of double-layer routing and adaptive query to realize more effective information

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

interaction between global and local details[18]. On the basis of this framework, the local convolution and routing weight enhancement of the input feature map, the fine-grained optimization of the local attention and the context enhancement of the output feature are carried out respectively.

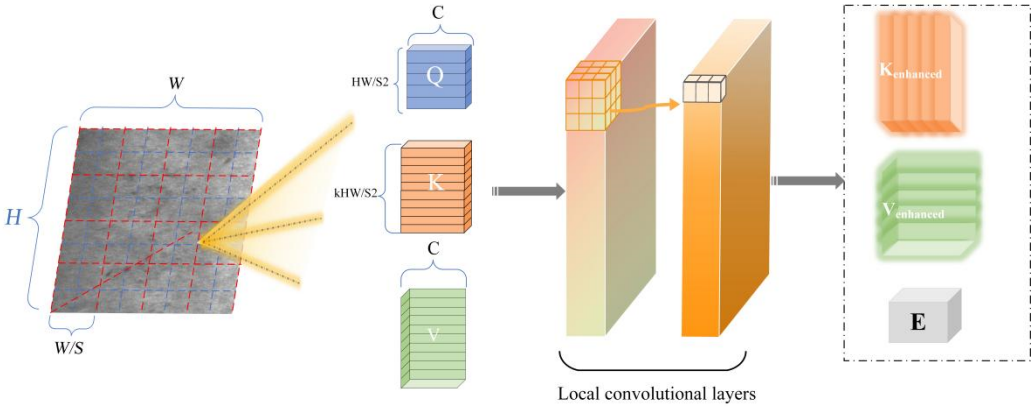

Figure 7. Kenhanced, Venhanced, and local context enhancement E are obtained by local convolution.

As shown in Figure 7,the first input feature map $X \in R^{H\times W\times C}$ will obtain Q (query), K (key), and V (value) through linear mapping, After the input feature map, a local convolution layer is added to enhance the output and features by K (key) and V (value) :

$$Y_{\mathrm{h,w,c}} = \sum_{i=-1}^{1}\sum_{j=-1}^{1} K_{i,j,c}\cdot X_{h+i,w+j,c} + b_c \tag{1}$$

Get Kenhanced, Venhanced, where b is the bias term, and h, w is the output position index. This process effectively captures the local details of the small target (such as texture and edge).

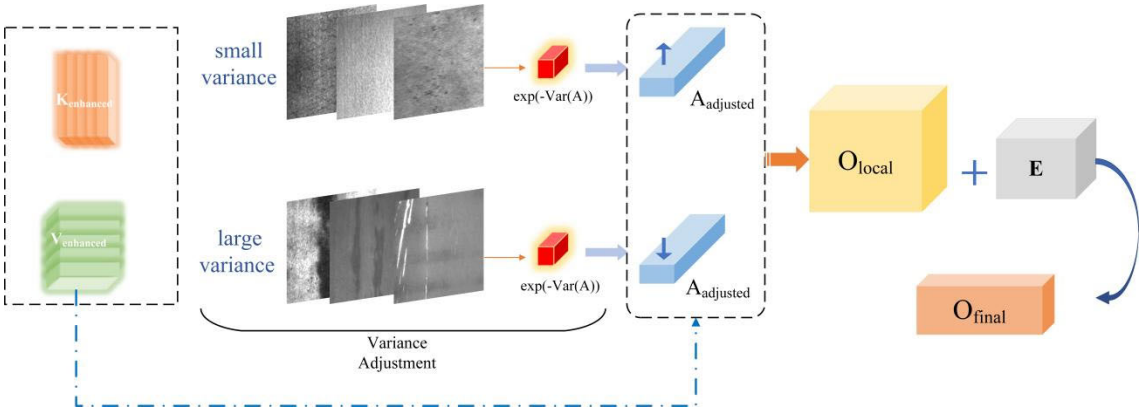

Figure 8. The final attention output process

The Q and K obtained by linear projection and local convolution enhancement are used to calculate the attention weight between regions:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

$$A_{i,j} = \frac{\exp(Q_i \cdot K_j / \sqrt{d_k})}{\sum_{k=1}^{N} \exp(Q_i \cdot K_j / \sqrt{d_k})} \tag{2}$$

Then, the variance adjustment mechanism is introduced to the weight in the top-k module. By adjusting the variance of the weight, the characteristics of which local areas need to be enhanced or suppressed are controlled. First, the variance of each area is calculated to quantify the distribution of the weight :

$$\mathrm{Var}(\mathrm{A}, \mathrm{dim} = \text{-}1) \tag{3}$$

Regions with large variance represent regions with large or unstable changes in features, which are often reflected as edges, noise or rapidly changing regions. This paper introduces an exponential function in the adjustment mechanism to scale the attention weight of each region. The region with large variance has a larger Var(A) value, resulting in a smaller exp(-Var(A)), so the attention weight of the region will be suppressed and the attention to the region will be reduced. For regions with small variance, it is often reflected as background or stable target regions. The information of these regions changes gently and the Var(A) value is small, resulting in a large exp(-Var(A)), thereby increasing the weight of these regions and enhancing their attention. Finally, the regional attention is adjusted to Aadjusted :

$$A_{\mathrm{adjusted}} = A \cdot \exp(\text{-}\mathrm{Var}(\mathrm{A}, \mathrm{dim} = \text{-}1)) \tag{4}$$

the local attention results based on routing weight are obtained :

$$\mathrm{O}_{\mathrm{local}} = A_{adjusted} \cdot V_{enhanced} \tag{5}$$

Ofinal discards the LCE (V) of the traditional local context enhancement term in the final output[19], and uses the local context enhancement term E generated by the depthwise separable convolution :

$$E = \mathrm{D}\mathrm{ep}wiseConv_{3\times3}(X) \tag{6}$$

final output:

$$\mathrm{O}_{\mathrm{final}} = \mathrm{O}_{\mathrm{local}} + E \tag{7}$$

The process is shown in Figure 8.Compared with the traditional attention mechanism, the final output attention can better capture small targets and local details, and under the variance adjustment mechanism, it can better adapt to the rapidly changing area and retain information, which is suitable for small target detection with complex background such as steel defects.

2.2.2 Multi-scale Feature Fusion Detection Heads

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

The traditional detection head in YOLOv8 cannot meet the high-precision requirements of industrial production for detecting steel defects. In this study, a detection head DRM was designed. Compared with the detection head of YOLOv8, multi-scale feature fusion was performed using FPN + PAN. DRM introduces multi-scale feature fusion and multi-context enhancement (PCRC), improves the way of feature fusion, solves the static fusion defects of traditional FPN / PANet, and improves the sensitivity to small targets and sparse defects.

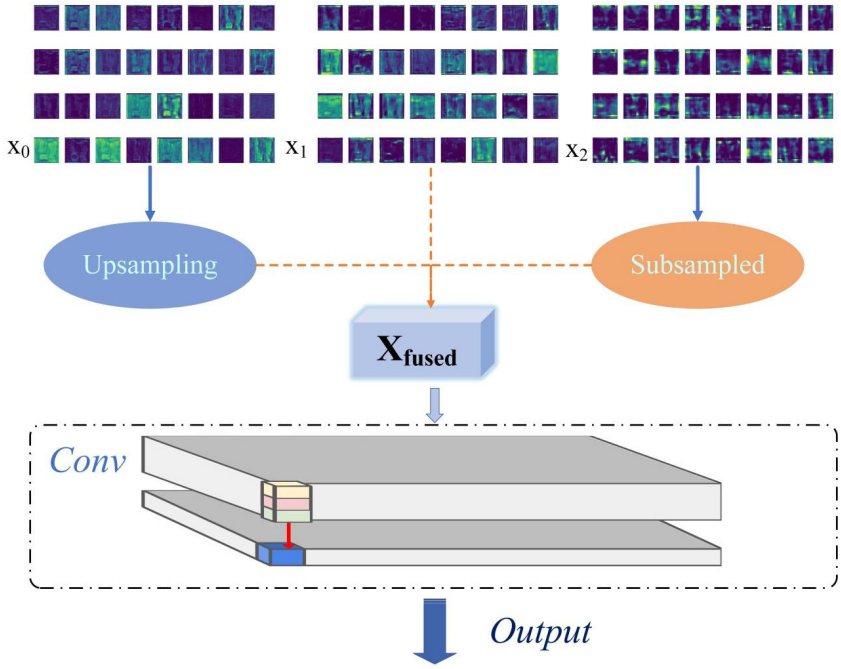

Figure 9. DRM feature extraction process

This mechanism effectively enhances detection efficacy for small-scale targets in complex environments while improving cross-domain generalization across heterogeneous image features. Figure 9 The three probes of the model receive multi-scale feature maps with different resolutions generated from different levels of the model : { x0, x1, x2 }, where:

$$X_0 \in R^{N\times C_0\times H_0\times W_0}, X_1 \in R^{N\times C_1\times H_1\times W_1}, X_2 \in R^{N\times C_2\times H_2\times W_2} \tag{8}$$

x0, x1, and x2 capture small targets, structural information, and global context information, respectively. Then the feature fusion module DRM first upsampled x0 to obtain x'0, and downsampled x2 to obtain x'2. The two are spliced with x1 to obtain :

$$X_{\text{fused}} = [\mathrm{x}_0^{'}, \mathrm{x}_1, \mathrm{x}_2^{'}] \tag{9}$$

the obtained multi-scale feature map is then compressed by 1 × 1 convolution to compress the channel dimension to enhance the expression ability. The process is shown in Figure 9.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

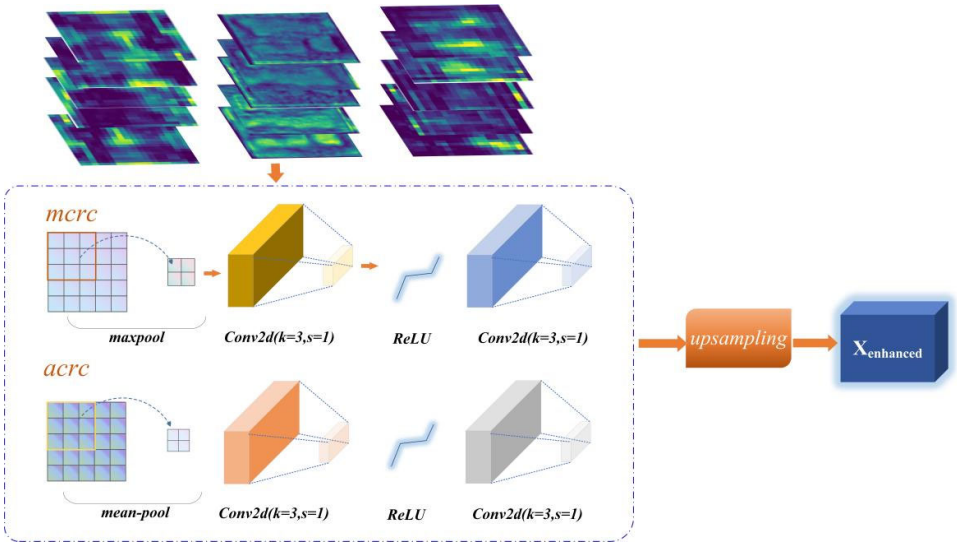

Figure 10. DRM feature fusion process

Subsequently, the feature map is further enhanced by the feature pyramid module (PCRC), and two paths of mcrc and acrc are used to process the fusion features.The process is shown in Figure 10, and the maximum pooling and average pooling are used to obtain :

$$X_{\max} = Conv(MaxPool(X_{fused})), X_{\text{avg}} = Conv(\text{Avg}Pool(X_{fused})) \quad (10)$$

The two pooling methods are used to extract salient region information and smooth context information respectively. The output of the last two paths is unified to the same resolution through upsampling. The results of the two branches are upsampled and added, which preserves the details and enhances the global perception ability. The enhanced feature Xenhanced is generated to achieve multi-scale fusion of features, as shown in Figure 10.

2.2.3 High-Performance Backbone Networks

In real-world steel defect detection, the detection environment is also challenging, particularly when handling multi-scale and complex-shaped steel defects. The YOLOv8 backbone network has limitations in global feature extraction, leading to subpar results, especially for small targets and intricate defects. Unlike the YOLOv8 backbone, MobileViT combines the strengths of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs)[20], leveraging CNN's spatial inductive bias and ViT's global representation capabilities, making it highly effective for small target detection.

The MobileVit network structure consists of three main modules: ordinary convolution Conv, MV2, and MobileVit Block, as shown in Figure 11.

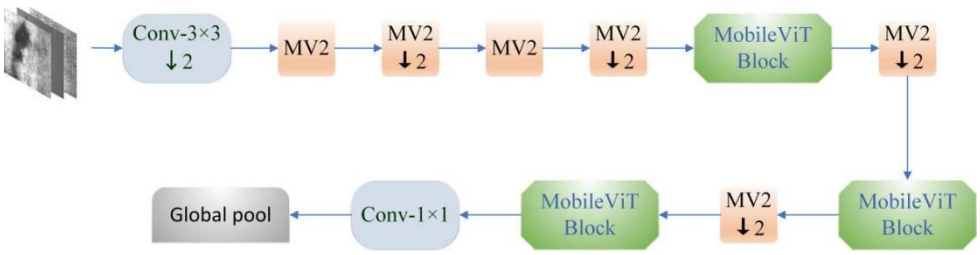

Figure 11. MobileVit network architecture

The external image data were first fed into a convolution module, followed by MV2, where "↓2" denotes downsampling. MV2 is the basic module of MobileNetV2, utilizing depthwise

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

separable convolutions and an inverted residual structure, as shown in Figure 12. When the stride was set to 1, a shortcut connection was included in the structure. In the inverted residual structure of MV2, the process begins by expanding the channel dimension (from low to high dimensions), applying depthwise separable convolutions for feature extraction, and finally reducing it back to a lower dimension through linear projection. This approach minimizes information loss after ReLU activation of high-dimensional data. The downsampled feature map is then processed by the core module in MobileViT, the MobileViT Block, which performs local convolutional feature extraction and global feature fusion.

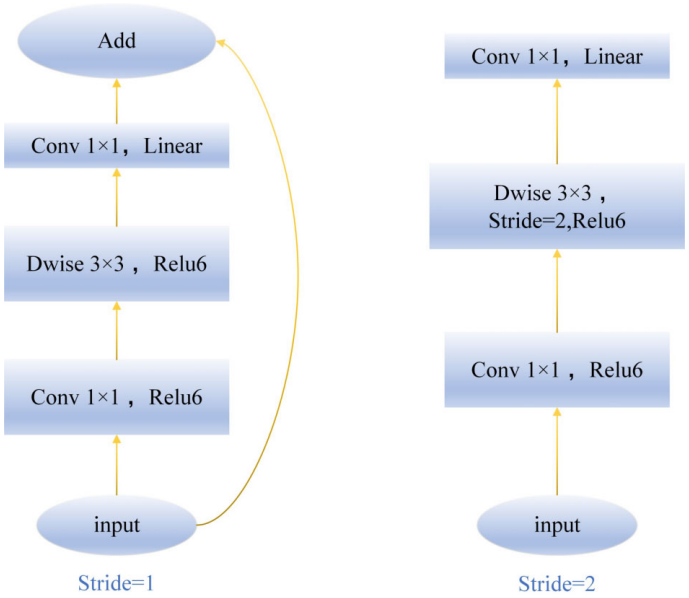

Figure 12. Different stride MV2

The MobileViT block initiates processing by applying a standard n×nconvolutional layer (typically n=3) to extract local spatial features from input tensor X ∈ ℝ^(H×W×C), capturing localized patterns such as edges and textures. Subsequently, a 1×1convolutional layer projects features into a higher-dimensional space X_L ∈ ℝ^(H×W×d)(d > C) to enhance representational capacity.

For global feature modeling, an unfold-transformer-fold mechanism is employed: X_L is partitioned into Nnon-overlapping flattened patches of size P×P(P = w×h, N = HW/P) via unfolding, reorganizing spatial information into sequence X_U ∈ ℝ^(P×N×d). A lightweight Transformer with Llayers then processes pixels at identical spatial positions across patches independently , modeling long-range dependencies while reducing computational complexity from O((HW)^2d)in standard ViTs to O(N^2d). The Transformer output is reconstructed into spatial feature map:

$$X\_F \in R^{\wedge}(H \times W \times D) \tag{13}$$

via folding, preserving original pixel spatial order. A 1×1convolution projects X_Fback to Cdimensions, which is concatenated with input X(early design) or additively fused (residual optimization in later versions). Finally, an n×nconvolutional layer (commonly depthwise separable) fuses local features, global context, and residual information to produce output Y ∈ ℝ^(H×W×C)

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

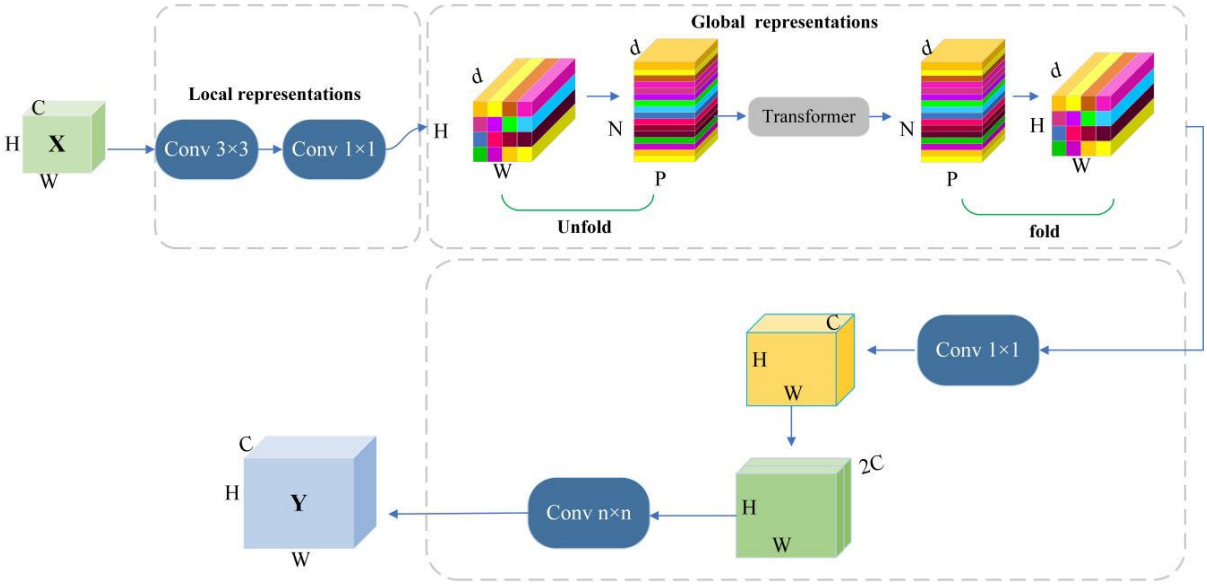

Figure 13. MobileViT Block

2.3. MBD-YOLO

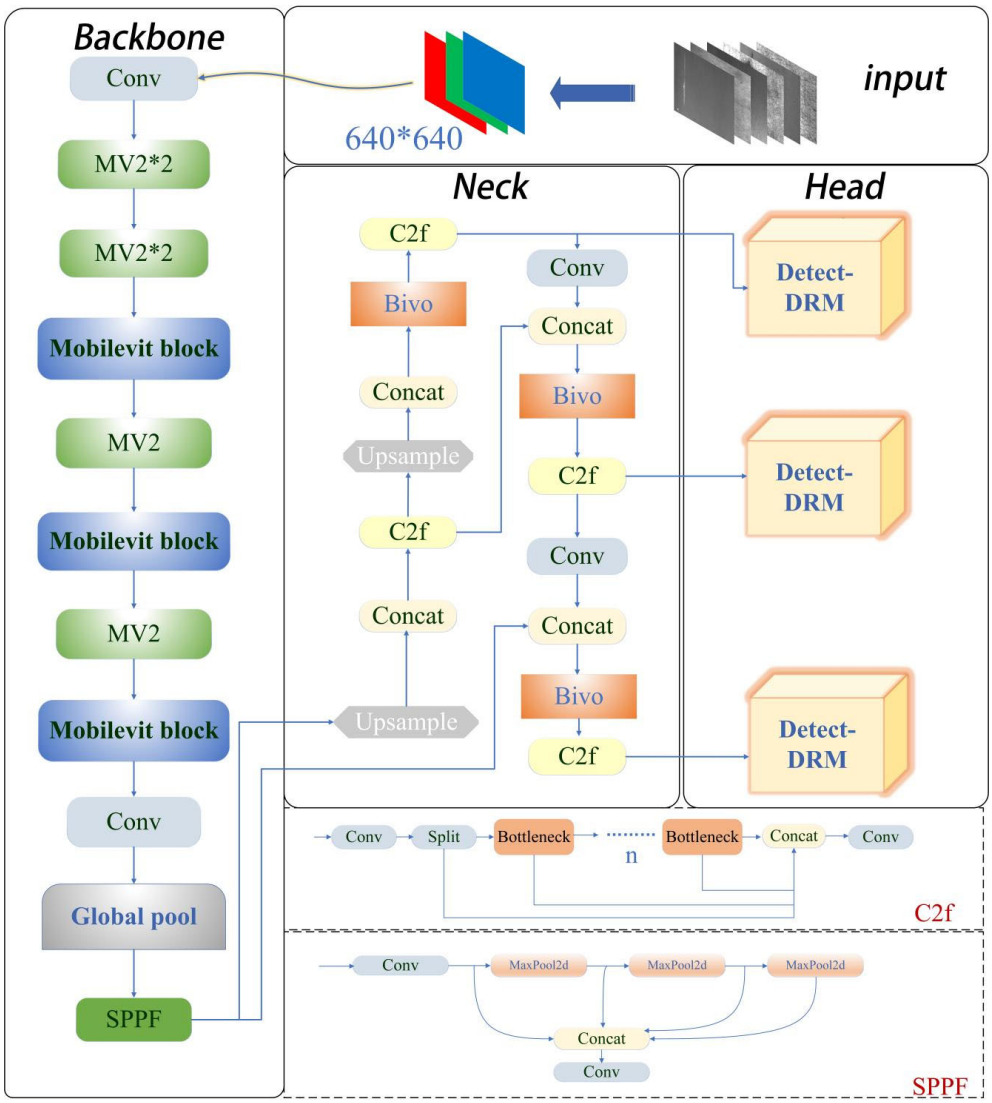

Figure 14. MBD-YOLO

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

Based on YOLOv8n as the baseline model, we developed an improved model, MBD-YOLO, tailored for steel defect detection. The overall structure is illustrated in Figure 14. The backbone uses the MobileViT network, enabling high-performance feature extraction. The integration of the MV2 and MobileViT structures allows the network to effectively capture global contextual information while extracting local details through convolution, thereby enhancing the model's representation capability. A new attention module, Bivo, is added to the neck to improve the information interaction in the feature fusion process through its cross-scale attention mechanism, effectively integrating multi-scale features and increasing robustness in detecting small targets in complex environments. For the detection head, we designed a new small-object detection head, DRM, which adapts to various features and integrates information from different sources, thereby improving the output accuracy and enhancing the detection precision. In summary, MBD-YOLO is a novel model designed specifically for steel defect detection, optimizing feature extraction, information interaction, and detection accuracy through the seamless combination of different modules to satisfy the requirements of this task.

3. EXPERIMENTAL VERIFICATION

3.1. Experimental preparation and evaluation index

3.1.1 Experimental preparation

The experimental implementation was performed in a Linux-based computational environment with the PyTorch framework, accelerated by an NVIDIA GeForce RTX 4090 GPU (24GB VRAM). The software configurations included CUDA 12.4 and Python 3.8.10. The training parameters were set to 300 epochs with a batch size of 8 for model optimization.

The dataset selected for this experiment was the public NEU-det dataset. The dataset contains six common small target steel defects, including ' cracks, scratches, patches, inclusions, pitting, ' and ' rolled-in scales. ' Figure 15 shows some samples from the dataset.

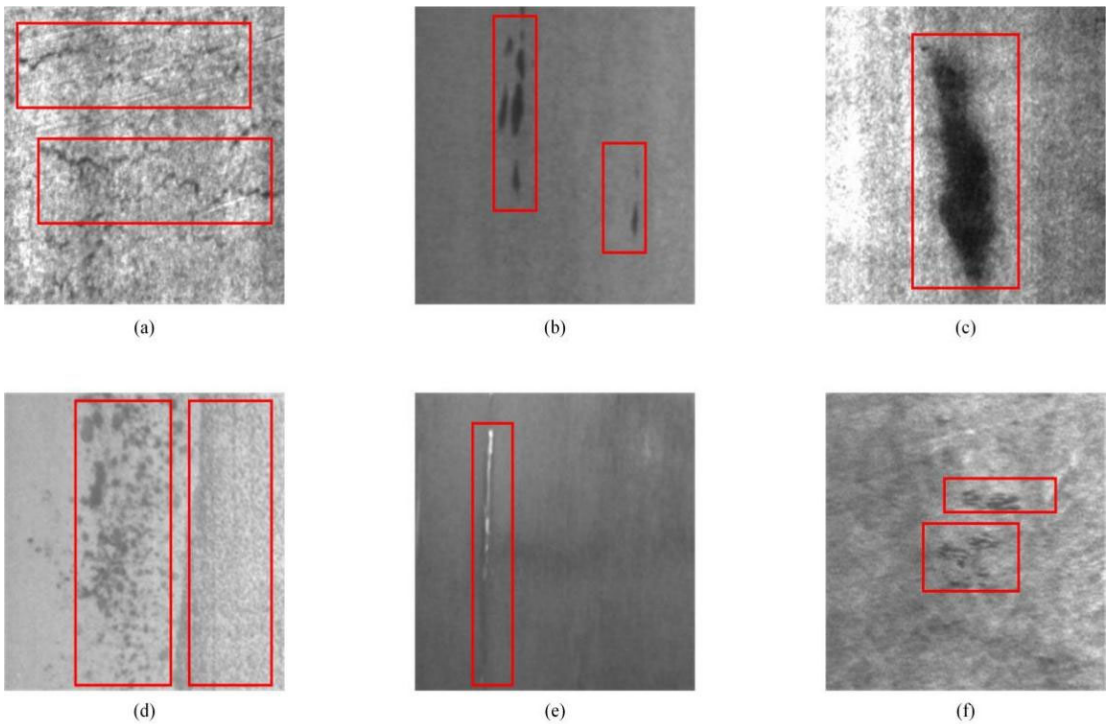

Figure 15. The defect types are a-f, followed by ' crack ', 'inclusion',' patches', 'pitted-surface', 'scratch', ' rolled-in scale '.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

3.1.2 Evaluation indicators

In deep learning-based classification and detection, the confusion matrix is a crucial tool for evaluating the performance of classification models, which is built from the envisioned values and the proper values, encapsulating the classification results via four key components: authentic positives (TP), real negatives (TN), false positives (FP), and false negatives (FN). The comparison metrics of precision, recall, and mean average precision (mAP) at a threshold of 1/2 were derived from these components. The formulation for calculating precision, henceforth referred to as P, is as follows:

$$P = \frac{TP}{TP + FP} \tag{14}$$

TP denotes real positives, signifying situations the place the mannequin precisely recognized an current defect. Conversely, FP refers to false positives, which appear when the mannequin erroneously identifies a defect that is now not present. Precision is a metric that measures the proportion of true positives among all samples predicted as positive by the model, providing an indicator of its predictive reliability.The recall rate, henceforth referred to as R, is computed as follows:

$$R = \frac{TP}{TP + FN} \tag{15}$$

FN denotes false negatives, which pertain to defects that have now not been detected. Recall serves as a metric for evaluating the model's potential to embody applicable instances, in particular measuring the share of real high-quality samples that are precisely recognized. This metric illustrates the model's pass over rate, thereby indicating its effectiveness in figuring out actual fine samples. The metric mAP@0.5 (mean Average Precision at IoU=0.5) quantifies detection accuracy under an Intersection over Union threshold of 50%. For individual object categories, Average Precision (AP) is mathematically defined as the integral of the precision-recall curve, expressed as:

$$AP = \int_0^1 P\left(R\right) \mathrm{d}R \tag{16}$$

$$\mathrm{m}AP = \frac{\sum_{i=1}^{N} AP}{N} \tag{17}$$

Where N is the number of samples.

3.2. Module effectiveness analysis

3.2.1Comparative Experiment of Attention Mechanism

To validate the efficacy of the Bivo attention mechanism in small target steel defect identification tasks, we conducted ablation studies by integrating distinct attention modules into the baseline architecture under identical experimental conditions. Quantitative comparisons of detection performance metrics are detailed in Table 1.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

Table 1. Comparison Experiment Results of Attention Mechanism

Model	P(%)	R(%)	mAP@50(%)	mAP@50:95(%)
Baseline	87.3	99	75	41
+EMA	67.9	79.2	75.8	40.9
+MLCA	77.8	82.4	73.6	41.2
+MSDA	76	75.8	76	40.4
+Triplet attention	75.7	73.2	77.1	41.8
+Irmb attention	67.7	80.5	76.6	40.8
+Bivo	77.8	97	78.5	41.9

The experimental results indicate that incorporating different attention mechanisms leads to a slight improvement in accuracy; however, it comes at the cost of reduced precision and recall, making them less suitable for small-target detection tasks. Among the four key metrics—precision (P), recall (R), mean average precision at 50% intersection over union (IoU) (mAP@50), and mean average precision across IoU thresholds from 50% to 95% (mAP@50:95)—the Bivo attention mechanism demonstrated strong performance, while the Bivo attention mechanism achieved the highest recall (97%), mAP@50 (78.5%), and mAP@50:95 (41.9%).Additionally, the Bivo module demonstrated strong performance in terms of precision (77.8%) and mAP@50 (78.5%). showing a balance between precise positioning and detection accuracy. This superior performance can be attributed to the effectiveness of the Bivo mechanism in feature interaction and multiscale information fusion, particularly for small targets. Therefore, the experiments verified that the Bivo attention mechanism is more suitable for small-target detection tasks and can provide higher accuracy and robustness in target detection at different scales.

3.2.2 ablation experiment

In order to verify the improved scheme, a multi-module ablation experiment was carried out. The experimental results are as follows: Table 2.

Table 2. Ablation experimental results

Baseline model	Bivo	DRM	Mobilevit	P(%)	R(%)	mAP@0.5(%)	mAP@0.5:0.95(%)
YOLOv8n				87.3	99	75	41
	√			77.8	97	78.5	41.9
		√		83.8	98	83	45
			√	81.3	95	75.8	47
	√	√		79.8	97	84.3	50
	√		√	82.5	98	79.9	47.8
		√	√	79.5	95	80.9	51
	√	√	√	89.8	100	85.7	53.7

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

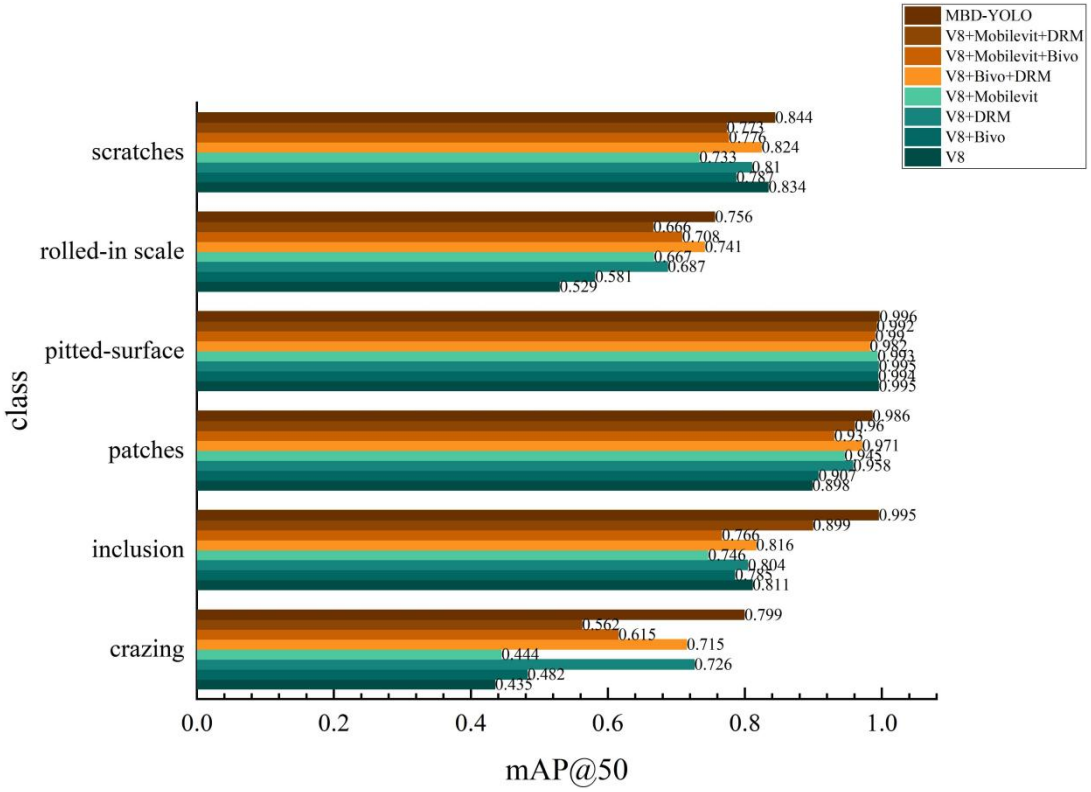

Figure 16. Type-specific defect mAP @ 0.5 ablation experiment

The experimental results revealed significant changes in the model performance with the introduction of various modules. The baseline model (YOLOv8n) achieved a recall rate of 99%, but its mean Average Precision (mAP@0.5) was only 75%. Adding the Bivo module slightly improved the mAP to 78.5%, although the precision (P) and recall (R) decreased, indicating that although Bivo aided feature fusion, its overall impact on performance was limited. When the DRM detection head was introduced, the mAP saw a notable increase to 83%, demonstrating that this module significantly enhanced small-object detection by improving feature fusion and detection accuracy. Although adding MobileViT boosted local feature extraction, its impact on overall performance was less pronounced than that of DRM. The combination of Bivo and DRM further increased the mAP to 84.3%, highlighting their synergistic effect on feature integration and information interaction, thus improving detection in complex scenarios. When DRM and MobileViT were combined, the model achieved a balanced mAP of 79.9%. The best performance was achieved when all three modules were used together, with an mAP of 85.7%, precision of 89.8%, and recall of 100%. This combination also achieved the highest single-class detection accuracy for various defects, indicating that Bivo, DRM, and MobileViT work complementarily to enhance feature extraction and fusion, thereby boosting the accuracy of steel defect detection.

The ablation results in Figure 16 show that for crack and inclusion defects, MBD-YOLO achieved mAP@0.5 of 0.799 and 0.995, respectively, which are substantially higher than those of the other configurations, demonstrating the synergy of Bivo, DRM, and MobileViT in improving small-object detection accuracy. The DRM also performed well in crack detection, reinforcing its effectiveness in detecting small targets. For patch and pitting defects, the mAP@0.5 of MBD-YOLO was close to 1.0, indicating that the improved feature fusion maintained a high precision for larger defects. For rolled-in scale defects, where the targets resembled the background, MBD-YOLO still outperformed the others, with an mAP@0.5 of 0.756. For scratch detection, MBD-YOLO achieved an mAP@0.5 of 0.844, demonstrating robust performance on irregular edges. Overall, the modular synergy of MBD-YOLO provides optimal performance

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

across various defect types, validating its strong multi-scale feature fusion and small-object detection capabilities.

3.3. Comparative experiment

In order to affirm the superiority of the MBD-YOLO model, the scan chosen the oftentimes used goal detection mannequin and in contrast it with the equal facts set. The experimental consequences are proven in Table 3 and Table 4.

Table 3. Experimental esults contrasting

model	P(%)	R(%)	mAP@0.5(%)	mAP@0.5:0.95(%)
YOLOV5	80.4	99	73.2	42.3
YOLOV10	79.8	97	80.1	45.7
SSD	78.9	99	78.9	41.5
DETR	82.4	98	83.3	38.9
Faster R-CNN	81.5	98	81.2	44.8
OURS	89.8	99	85.7	53.7

Table 4. Comparative experiment of mAP @ 0.5 with different types of defects

type	YOLOV5	YOLOV10	SSD	DETR	Faster R-CNN	OURS
Crazing	0.434	0.451	0.52	0.63	0.59	0.799
Inclusion	0.78	0.79	0.68	0.81	0.72	0.995
Patches	0.881	0.823	0.79	0.901	0.912	0.986
Pitted -Surface	0.991	0.994	0.99	0.98	0.995	0.996
Rolled-in Scale	0.534	0.516	0.643	0.656	0.712	0.756
Scratches	0.802	0.75	0.623	0.645	0.783	0.844

The proposed model demonstrates leading performance across comparative benchmarks, achieving 89.8% precision, 99% recall, and mAP scores of 85.7% (@0.5) and 53.7% (@0.5:0.95). These metrics confirm substantial improvements in task-specific detection accuracy, demonstrating the effectiveness of the model's innovative design and training framework. Specifically, MBD-YOLO exhibits superior capability in identifying steel surface micro-defects compared to existing approaches, validating its optimization strategies for sub-scale target detection.

To further assess the performance of MBD-YOLO, we collected iterative data on mAP @ 0.5 and three loss functions throughout the training process, subsequently generating a convergence comparison diagram, as illustrated in Figure 17. In this diagram, box_loss represents the accuracy of target localization, cls_loss indicates the precision of target classification, and dfl_loss reflects the accuracy of bounding box coordinate predictions. The consistently low and stable values of these three loss functions suggest that the model achieved an effective fitting state. Over 300 training iterations, the accuracy of the baseline model converged at 75%, whereas the MBD-YOLO model reached a convergence accuracy of 85.7%. Quantitative evaluations substantiate the proposed model's superiority over the baseline architecture, with three critical observations: Loss Convergence: The training curves exhibit accelerated convergence characteristics, showing 28.6% lower mean loss values than the baseline; Detection Accuracy: Significant absolute gains of 10.7% (mAP@0.5) and 12.7% (mAP@0.5:0.95) are achieved; Robustness: Precision-recall curves demonstrate enhanced stability across varying confidence thresholds.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

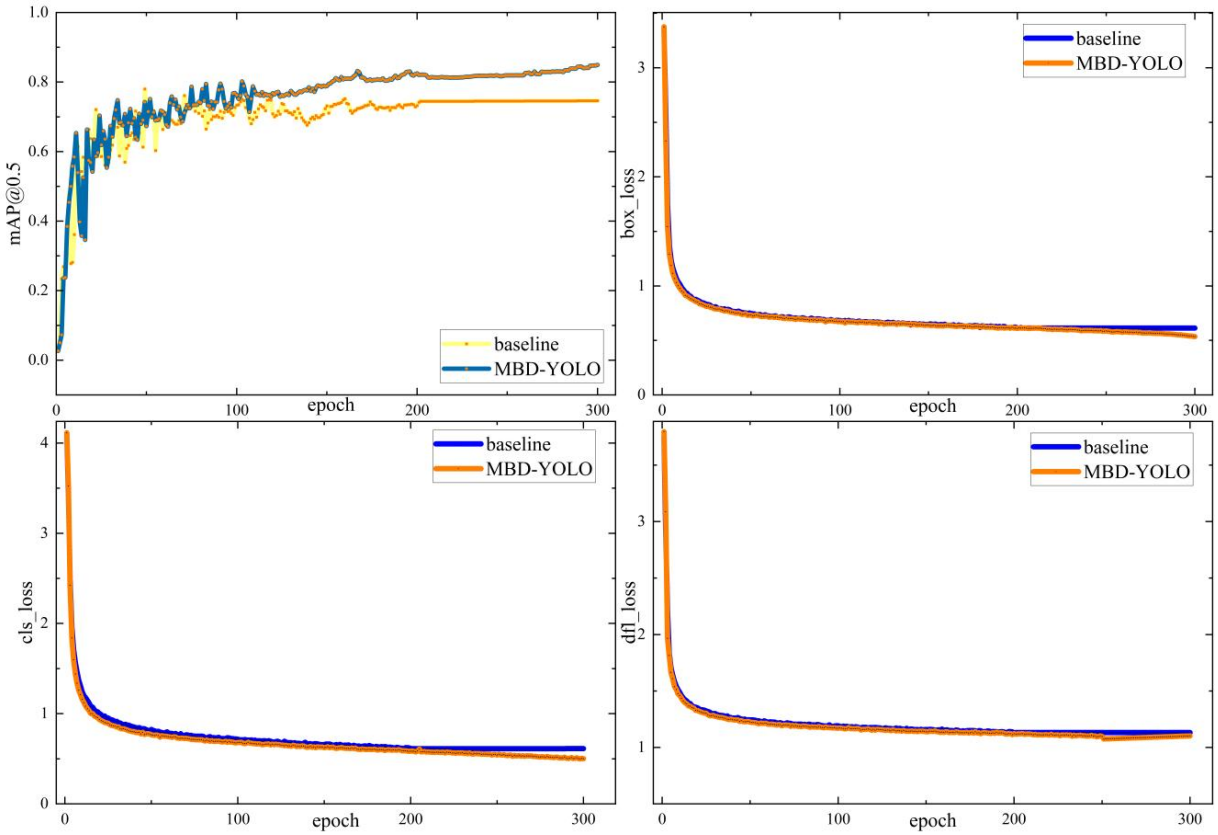

Figure 17. Convergence comparison diagram

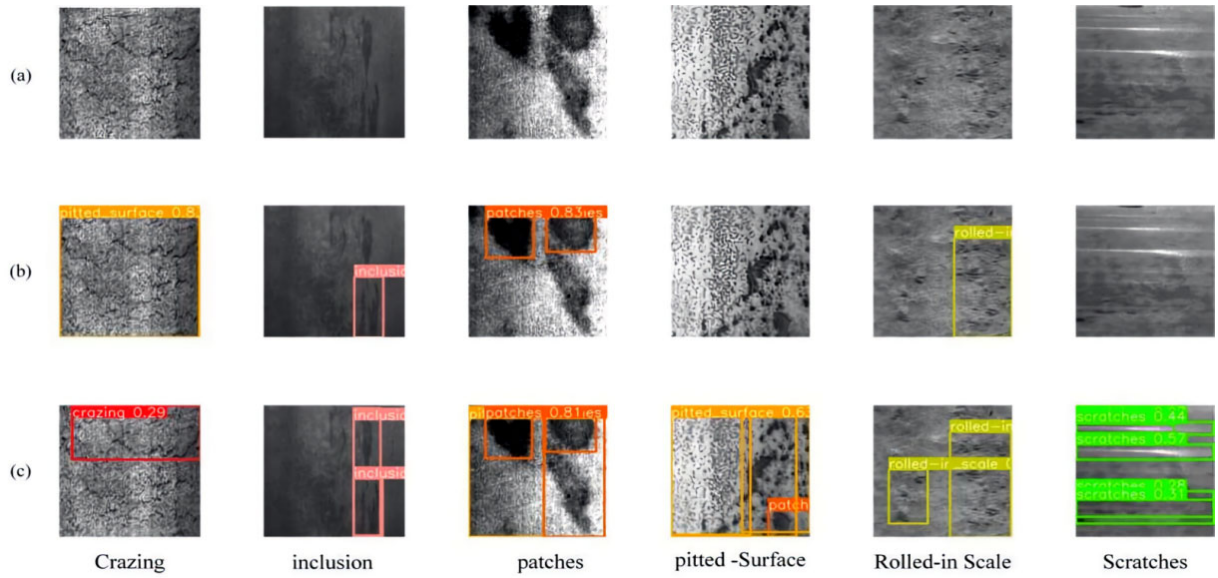

Figure 18. Comparison of test results (a)Original defect image (b)YOLOV8n (c) MBD-YOLO

Figure 18 shows the contrast between the baseline mannequin and some of the detection outcomes of MBD-YOLO. It can be viewed that the extended mannequin can enhance the trouble of false detection and ignored detection of some defects in the baseline model, and in addition illustrates the enchancment of the mannequin characteristic extraction and function fusion stage.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

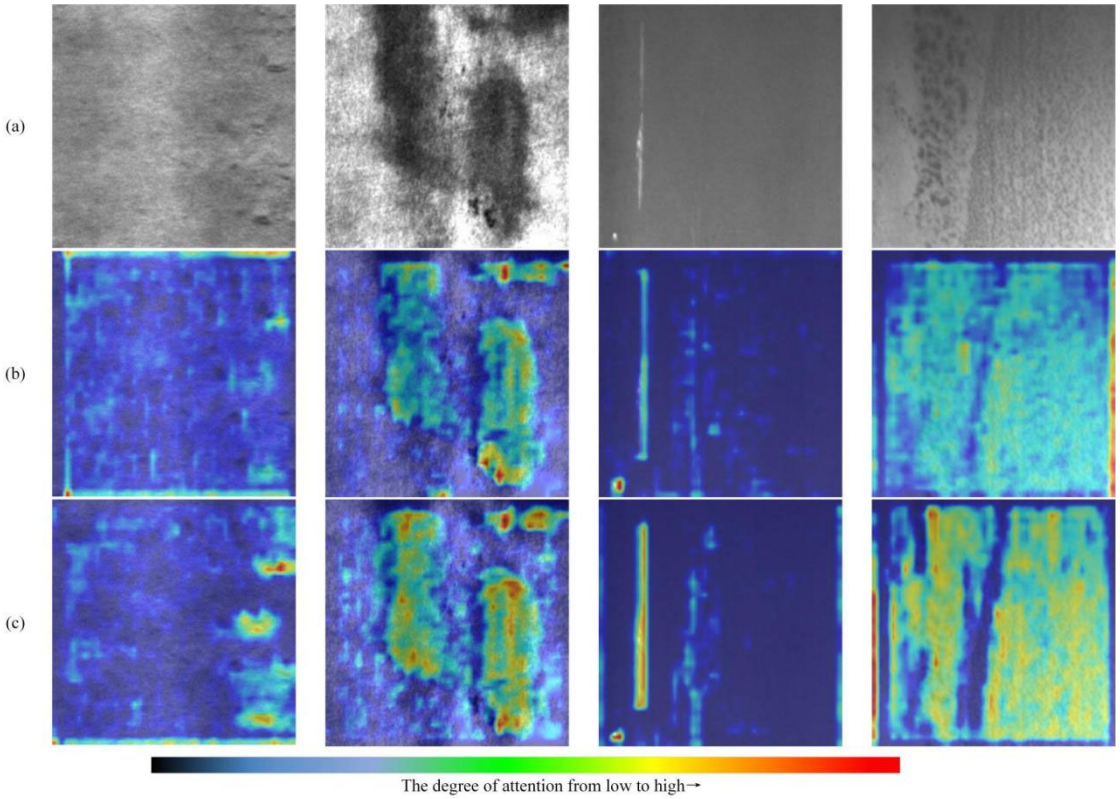

Figure 19. Heat map comparison (degree of interest) (a)Original (b)YOLOV8 (c)MBD-YOLO

The overall performance improvement and the warmth map of the defect in Figure 19, in the discipline of deep learning goal detection, the which means of the warmness map is that the brighter the color (normally crimson or yellow), the greater interest the mannequin gets, the extra satisfied that the area carries the target. In figure (c), the defect area is significantly highlighted (mainly red or yellow), especially in the context of complex surface textures (such as the second and third columns). This shows that MBD-YOLO has a strong focus on the defect area and can accurately locate most defects. In the figure, (b) can also pay attention to the defect area, but compared with MBD-YOLO, some areas (such as the third and fourth columns) show weaker attention, and there may be some missed detection or insensitivity. In Figure (c), the attention of the non-defect area is relatively low (mainly blue and green), indicating that the method has a good ability to suppress background interference. In figure (b), there are also some high attention values (yellow or red) in the background area, which may have some false detection, indicating that the baseline model is less robust to background texture than MBD-YOLO. The experimental results show that the improved model exhibits better generalization ability and robustness. It can accurately identify defect areas and suppress background interference. It has a high degree of attention to the defect area and can complete defect detection tasks well.

3.4. Generalization experiment

Generalization experiments are also needed to further verify the success of MBD-YOLO. Generalization ability refers to the ability of the model to perform well on unseen data or tasks; that is, the rules learned by the model from the training data can be extended to new, different samples or situations. Here, two steel defect datasets GC10-DET, Severstal: Steel Defect Detection, were selected for generalization experiments in the same experimental environment. GC10-DET contains ten common industrial product defects including cracks, inclusions, pitting-

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

surface, and oxide scales. Severstal: Steel Defect Detection includes Edge Cracks, Longitudinal Cracks, Transverse Cracks and Surface Patches. The generalization ability of the model is verified by the comparative experiment of classification defects. The following table is the experimental results to verify the generalization ability of the model, and Table 5 is the experimental results.

Table 5. Generalization experimental results

Dataset	model	Defect type	P(%)	R(%)	mAP@50(%)
Severstal Steel Defect	YOLOV8	Ec	83.7	78.9	81.3
		Lc	79.4	75.6	77.8
		Tc	80.2	76.1	78.5
		Sp	72.3	68.5	70.1
	MBD-YOLO	Ec	87.2	83.5	87.2 (↑5.9)
		Lc	84.1	80.3	84.1 (↑6.3)
		Tc	85.6	81.7	85.6 (↑7.1)
		Sp	78.9	75.2	78.9 (↑8.8)
GC10-det	YOLOV8	Cr	70.3	65.8	68.2
		In	62.4	58.1	60.5
		Pa	66.7	63.2	64.9
		Ps	56.8	53.4	57.6
		Rs	64.5	61.7	63.1
		Sc	59.2	55.9	57.8
		Co	58.1	54.3	59.3
		Pu	68.9	65.2	67.4
		We	63.7	59.8	61.5
		Fo	60.5	57.0	58.9
	MBD-YOLO	Cr	76.1	72.5	75.6 (↑7.4)
		In	68.7	64.3	66.8 (↑6.3)
		Pa	72.9	69.8	71.2 (↑6.3)
		Ps	65.2	62.1	66.3 (↑8.7)
		Rs	70.1	67.5	70.1 (↑7.0)
		Sc	64.8	61.3	65.4 (↑7.6)
		Co	62.4	58.9	64.8 (↑5.5)
		Pu	73.5	70.	73.5 (↑6.1)
		We	68.2	64.5	68.2 (↑6.7)
		Fo	65.7	62.4	65.7 (↑6.8)

The experimental results show that MBD-YOLO also exhibits excellent performance on other steel datasets. In the face of different defects in different datasets, compared with the baseline model, MBD-YOLO can also greatly improve the detection of Severstal sparse defects (such as surface patch Sp) and direction-sensitive defects (such as longitudinal crack Ec) are significantly improved (↑ 5.9–8.8 %), indicating the adaptability of the model to complex industrial scenarios. GC10-DET: The performance was balanced on small targets (Ps) and low-contrast defects, which verified the effectiveness of multi-scale feature fusion. The improvement in all defect types was more than 5.5 %, which proves the robustness of the model in complex scenarios.

In summary, the model can adapt to steel defect detection tasks in different scenarios and diverse datasets and has a strong generalization ability. This means that the model is not only

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

suitable for defect detection in a single scene but also performs well in other scenarios, and its performance can be extended to new data scenarios.

4. CONCLUSION

This study provides a more advantageous model for metal defect detection, termed MBD-YOLO, which is based on the YOLOv8 framework. The mannequin contains several progressive graph factors that drastically improve the accuracy and overall performance of defect detection on metal surfaces in complicated environments, specifically in the context of multi-scale goal detection.

First, the Bivo module, a novel attention mechanism designed in this study, overcomes the limitation of traditional attention mechanisms that focus solely on single-scale features. By effectively combining global and local features, Bivo enables multiscale information interaction. Its attention weight adjustment mechanism enhances the detection of small targets and fine details, whereas the variance adjustment mechanism allows it to adapt to dynamic regions, preserving critical information. Through its routing-weight attention mechanism, MBD-YOLO accurately identifies various small defects, such as cracks and surface pits, even in complex industrial environments. In addition, it significantly improves background suppression, enhancing the overall performance of small-target detection tasks.

Second, a new multi-scale feature fusion detection head DRM was designed. Through its flexible feature processing method, it effectively combines small target features, structural information, and global background information, improves the accuracy of feature fusion, and enhances the accuracy of model detection for small-scale and high-complexity defects.

Additionally, the backbone structure of the model incorporates MobileViT, which combines the advantages of convolutional neural networks (CNN) and vision transformers (ViT). This integration not only maintains efficient feature extraction but also improves recognition performance for small targets. By effectively merging local and global information, the MobileViT module enhances the generalization of the model, allowing it to adapt to diverse and complex industrial detection scenarios.

The experimental results indicate that the MBD-YOLO deep learning model attains a detection accuracy of 85.7% on the NEU-det dataset, with precision and recall reaching 89.8% and 100%, respectively. Ablation studies revealed that the combined integration of the Bivo module, DRM detection head, and MobileViT significantly improved feature extraction and fusion, thereby enhancing the overall detection performance of the model. Comparative analyses further demonstrated that MBD-YOLO surpassed existing object detection models, such as Faster R-CNN, SSD, and DETR, particularly in the detection of small targets and defects within complex backgrounds. The generalization experiments revealed that the model exhibited strong adaptability and performed well in steel defect detection tasks across different scenarios and datasets. This indicates that the model is not only effective in specific defect detection contexts but also maintains high performance in various environments, making it suitable for application in new data scenarios.

REFERENCES

[1] Li, Z., Wei, X., Hassaballah, M., Li, Y., & Jiang, X. (2024). A deep learning model for steel surface defect detection. *Complex & Intelligent Systems*, 10(1), 885-897.

[2] He, Y., Song, K., Meng, Q., & Yan, Y. (2019). An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. *IEEE transactions on instrumentation and measurement*, 69(4), 1493-1504.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

[3] Chorowski, J., Wang, J., & Zurada, J. M. (2014). Review and performance comparison of SVM-and ELM-based classifiers. *Neurocomputing*, 128, 507-516.

[4] Lien, P. C., & Zhao, Q. (2018, August). Product surface defect detection based on deep learning. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)* (pp. 250-255). IEEE.

[5] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).

[6] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).

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

[8] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

[9] Varghese, R., & Sambath, M. (2024, April). Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)* (pp. 1-6). IEEE.

[10] Hatab, M., Malekmohamadi, H., & Amira, A. (2021). Surface defect detection using YOLO network. In *Intelligent Systems and Applications: Proceedings of the 2020 Intelligent Systems Conference (IntelliSys) Volume 1* (pp. 505-515). Springer International Publishing.

[11] Zhao, C., Shu, X., Yan, X., Zuo, X., & Zhu, F. (2023). RDD-YOLO: A modified YOLO for detection of steel surface defects. *Measurement*, 214, 112776.

[12] Wang, Y., Wang, H., & Xin, Z. (2022). Efficient detection model of steel strip surface defects based on YOLO-V7. *Ieee Access*, 10, 133936-133944.

[13] Lu, M., Sheng, W., Zou, Y., Chen, Y., & Chen, Z. (2024). WSS-YOLO: An improved industrial defect detection network for steel surface defects. *Measurement*, 236, 115060.

[14] Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 390-391).

[15] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9), 1904-1916.

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

[17] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8759-8768).

[18] Zhu, L., Wang, X., Ke, Z., Zhang, W., & Lau, R. W. (2023). Biformer: Vision transformer with bi-level routing attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10323-10333).

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0008

[19] Ren, S., Zhou, D., He, S., Feng, J., & Wang, X. (2022). Shunted self-attention via multi-scale token aggregation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10853-10862).

[20] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

AUTHOR CONTRIBUTIONS

Conceptualization, Xvyue Zhang and Baoping Wang.; methodology, Xvyue Zhang; software, Xvyue Zhang; validation, Xvyue Zhang, Qin Sun and Baoping Wang; formal analysis, Xvyue Zhang; investigation, Xvyue Zhang,Baoping Wang; resources, Da Zhao,Lifang Zhao; data curation, Xvyue Zhang; writing—original draft preparation, Xvyue Zhang; writing—review and editing, Baoping Wang,Qin Sun,Da Zhao,Lifang Zhao; visualization, Xvyue Zhang; supervision, Baoping Wang,Lifang Zhao; project administration, Baoping Wang,Lifang Zhao; funding ,Baoping Wang,Lifang Zhao. All authors have read and agreed to the published version of the manuscript.

DECLARATIONS

Competing interests

The authors declare no competing interests.

Ethical approval

This is not applicable in this article.

Funding

This research was supported by Shandong Construction Machinery Intelligent Equipment lnnovation& Entrepreneurship Community (Grant No. GTT20240105). Shandong Province Key R&D Program 2025JMRH0304

Data availability statement

The NEU-det dataset used in this paper has been publicly available for research.It can be accessed through the following URL: http://faculty.neu.edu.cn/yunhyan/NEU_surface_defect_database.html.