Article | Received 6 April 2026; Revised 21 June 2026; Accepted 8 July 2026; Published 22 July 2026
https://doi.org/10.55092/mt20260004

A multi-perception feature-guided network with dynamic sample matching for bearing surface defect detection

Xuan Chen, Yongyi Chen and Dan Zhang*

Department of Automation, Zhejiang University of Technology, Hangzhou, China

* Correspondence author; E-mail: danzhang@zjut.edu.cn.

Highlights:

- A lightweight feature adaptive extraction network is constructed, which not only maintains a low parameter quantity but also enhances the multi-scale feature representation capability.
- A multi-shape adaptive module is designed, which enhances the sensitivity to different types of defects through a dual-axis multi-scale cross-attention mechanism.
- A dynamic sample matching strategy is proposed, which combines classification and location scores to select positive samples, and utilizes a dual-weighting mechanism to suppress the interference of low-quality negative samples.

Abstract: In mechanical systems, bearings play a vital role, yet their surfaces frequently develop defects during the production process. Hence, automated detection of such defects becomes indispensable for ensuring industrial quality control. To enhance the robustness of bearing surface defect detection, a dynamic sample matching detection network is proposed, which is guided by multi-level perceptual features. A lightweight adaptive feature extraction method is introduced to enhance defect representation while maintaining low parameter complexity and computational cost. To address diverse defect patterns, a multi-shape perception module with a dual-axis cross-weighting mechanism across multiple scales is designed to improve sensitivity to different defect types. Furthermore, a dynamic sample selection strategy with dual-label weighting is introduced to improve training sample quality. The proposed approach is evaluated on a self-developed bearing defect platform, where it outperforms mainstream detection models in complex environments and achieves superior performance and robustness.

Keywords: bearings; defect detection; attention mechanism; lightweighting; sample matching

1. Introduction

Bearings are core components in mechanical equipment and serve both supporting and transmission functions. During production and assembly, friction and pressure often cause surface defects. These defects degrade product appearance and may render products unqualified. They also affect the safe

operation of mechanical systems [1]. Therefore, defect detection on production lines, together with the removal of defective products, plays a crucial role in industrial manufacturing and ensures reliable production processes.

Over the past decades, numerous machine vision-based defect detection algorithms have been developed. Traditional methods typically include a feature extractor and a classifier, techniques such as k-nearest neighbors [2], self-organizing maps [3], and support vector machines [4] are extensively adopted as classifiers for surface defect detection. However, these methods usually rely on complex image preprocessing and feature extraction. This reliance reduces robustness and flexibility in practical inspection tasks [5]. To achieve high-precision and high-efficiency quality assurance, more accurate and efficient detection methods are required. With the rapid development of deep learning in recent years, artificial intelligence-based defect detection methods have been widely applied in intelligent manufacturing [6].

As shown in Figure 1, multiple factors affect bearing surface defect detection, including surface markings, complex backgrounds, and diverse defect shapes [7]. To reduce background interference, research attention has been paid on the introducing attention mechanisms and feature enhancement strategies to improve the identification of defect targets. For example, Shi *et al.* [8] incorporated a channel-spatial attention mechanism to enhance the focus on small defects. Uneven illumination and complex textures may cause background patterns to be misidentified as defects. To improve defect representation, Zhao *et al.* [9] proposed self-attention and cross-attention modules to weight intra-layer and inter-layer features. Li *et al.* [10] combined dynamic thresholding with generative adversarial networks for data augmentation and further introduced attention mechanisms and adaptive fusion strategies. Although this method improves detection performance, it introduces significant parameter redundancy and limits deployment efficiency in industrial scenarios.

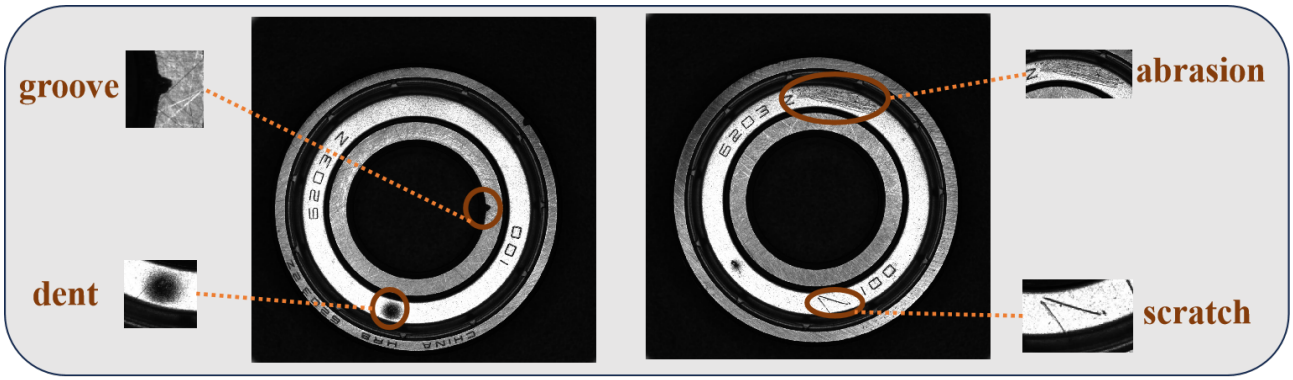

Figure 1. Diagram showing relevant defects on the bearing surface.

To address variations in defect size and morphology, researchers have explored feature integration and semantic alignment methods. Zhao *et al.* [11] proposed a method that integrates global and local features with transfer learning, enhancing small defect detection in data-scarce scenarios. Subsequently, Ma *et al.* [12] employed a hierarchical attention mechanism to integrate texture, semantic, and instance-level features, enabling multi-class detection in complex backgrounds. Gao *et al.* [13] improved the Faster Region-based Convolutional Neural Network (R-CNN) framework for insulated bearing defect detection, with detection accuracy enhanced through optimized k-means clustering and Region of Interest (ROI)

Align. Xu *et al.* [14] further improved multi-scale feature extraction by integrating cross-stage local networks and efficient aggregation layers. In addition, a dynamic task-aligned label matching strategy is introduced to select high-quality samples.

In summary, these methods have advanced defect detection by improving feature representation. However, existing methods still exhibit several limitations, including high model redundancy, limited capability in recognizing deformable defects, and inaccurate sample selection during backpropagation.

Recently, dynamic label assignment strategies have been extensively studied to improve training sample quality in object detection. Task-Aligned Assignment, as presented by Feng *et al.* [15] further aligns classification and localization tasks via a task-aligned head, yielding more consistent selection. The strategy has been integrated into subsequent You Only Look Once (YOLO) versions and it has achieved notable success in general object detection. However, most existing dynamic matching methods primarily rely on loss-based selection or fixed Intersection-over-Union (IoU) thresholds, which are not sufficiently adaptive to the diverse defect morphologies and ambiguous boundaries encountered in bearing surface inspection. In parallel, various defect detection approaches have explored feature enhancement and multi-scale fusion, including turbidity-similarity decoupling for separating surface anomalies, prompt then refine for interactive defect localization, LSINet [16] for rail defect detection, LSNet [17] and WaveNet [18] for salient object segmentation, as well as CCAFNet [19], ECFFNet [20], IRFR-Net [21], and RCNet [22] for road defect analysis. Although these methods have demonstrated promising results, they typically treat feature representation and sample assignment as separate components and fail to incorporate adaptive sample weighting with shape-aware feature perception in an end-to-end unified framework.

To address these issues, a Multi-level Feature-guided Dynamic Matching (MFDM) network is proposed, which improves generalization across multiple defect types. The main contributions are as follows.

(1) A lightweight feature extraction backbone is constructed to balance performance and efficiency. Redundant computation in convolutional layers is reduced through channel separation and recombination. Feature representation is further improved by adaptive channel weighting and channel shuffling.

(2) A multi-shape adaptive detection head and a multi-shape adaptation module are introduced. A dual-axis multi-scale cross-attention mechanism is employed in the module, where both horizontal and vertical directions are considered. Through this design, adaptive weighting of multi-scale defects across categories is achieved. Shape and scale information are integrated in the detection head, and high-quality features are generated. As a result, sample selection during training is improved.

(3) A dynamic sample matching strategy is proposed. Positive samples around defect regions are selected, and both classification and localization scores are taken into account. Additionally, a dual-weighted label assignment mechanism is designed to suppress low-quality negative samples. By this mechanism, the network is guided to focus on high-quality positive samples, while effective supervision for ambiguous samples is maintained.

2. Methods

This section is organized as follows. First, the proposed method is described in detail. Then, the overall network architecture of MFDM is introduced.

2.1. *Lightweight adaptive feature extraction module*

In MFDM, the Down-Sampling Extraction Block (DSEB) and the Feature Enhancement Block (FEB) are designed for feature extraction. As shown in Figure 2, the bearing surface feature maps are sequentially fed into the DSEB and FEB, where parallel processing is performed through the main branch and the shortcut branch, respectively. After down-sampling, the DSEB outputs feature maps, which are subsequently passed to the FEB for further enhancement. To minimize computational overhead, a channel separation strategy is adopted to partition the feature maps before they are fed into the FEB. In addition, to enhance the sparsity of adaptive features and mitigate neuron inactivation, batch normalization is applied after the convolutional layers to accelerate computation, while the Leaky Rectified Linear Unit (ReLU) activation function is used for nonlinear mapping.

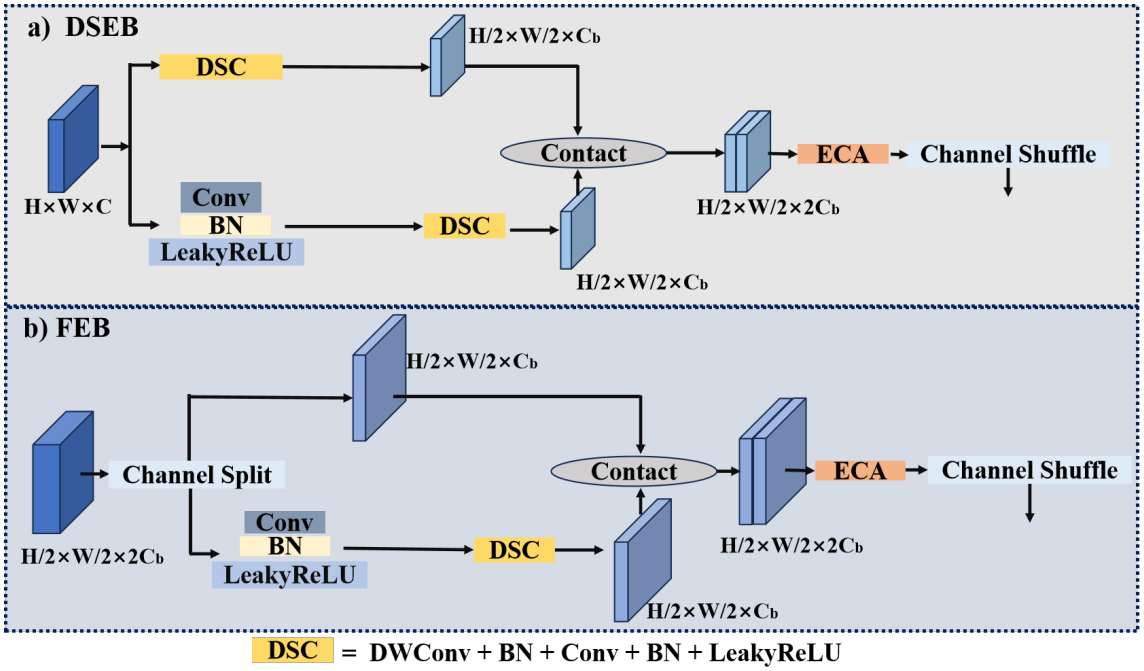

Figure 2. The network structure of the lightweight modules DSEB and FEB.

In terms of feature processing mechanisms, significant differences exist between the DSEB and FEB. A dual-branch down-sampling architecture is adopted in the DSEB. In this design, both the main branch and the shortcut branch perform down-sampling using Depthwise Separable Convolutions (DSC), halving the spatial dimensions of the output feature maps and doubling the number of channels through channel concatenation. In contrast, for the FEB, a channel separation strategy is innovatively introduced, with feature extraction performed only on the main branch. This design effectively reduces operational fragmentation during feature enhancement. As the input features are pre-separated by channel, the output feature maps maintain the same number of channels.

At the outputs of both modules, Efficient Channel Attention (ECA) mechanism and channel shuffle operation are applied to promote feature interaction between different branches. As illustrated in Figure 3, the multi-branch fused feature maps from the module undergo channel attention processing, which adaptively adjusts the weights of different channels. Specifically, global average pooling is first applied to the feature maps, followed by a 1×1 convolution to generate channel weights, which are then

normalized using a Sigmoid function. The size of the convolution kernel varies with the number of channels, calculated as follows:

$$K = \psi(C) = |\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}|_{odd} \tag{1}$$

where *odd* denotes rounding to the nearest odd number. The parameters b and γ are used to dynamically adjust the receptive field of local cross-channel interactions based on the number of input channels C, with values set to 1 and 2 in this work. Finally, the generated weights are multiplied with the original feature maps, enhancing the key channel features while suppressing less important ones.

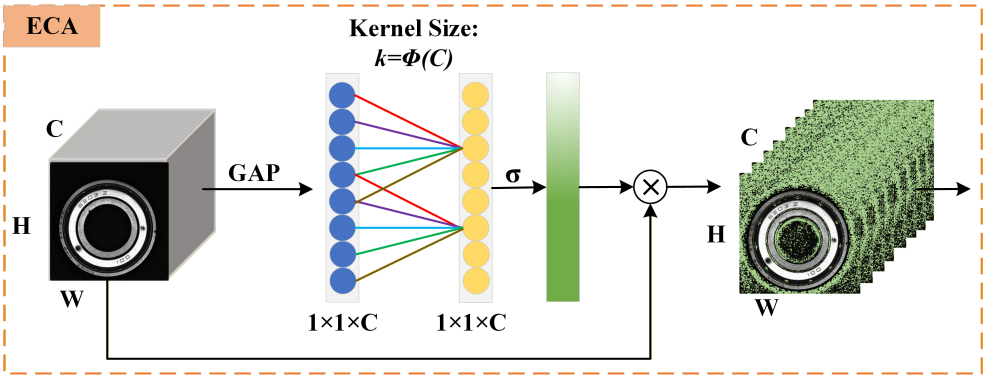

Figure 3. The characteristic processing procedure of the ECA mechanism.

It is worth emphasizing that the adoption of channel splitting together with channel shuffle operations effectively lowers both parameter overhead and computational cost. Furthermore, introducing adaptive channel-wise weighting prior to the shuffle process strengthens the network's capability to capture discriminative features from bearing surface images.

2.2. *Multi-shape adaptive perception module*

Given the diverse types of surface defects in bearings, including elongated scratches with clear directional patterns as well as grooves resembling background textures, traditional convolutional operations are limited in their ability to simultaneously capture such varied forms. To address the issue, a Multi-scale Attention Module (MAM) is introduced at the network's output, through which the multi-level features produced by the model are further enhanced.

To enhance the detection head's capability of modeling geometric variations of defects, deformable convolutions are first employed in the MAM for adaptive sampling of the input feature maps, as illustrated in Figure 4. For each 3 × 3 convolutional kernel, a set of 2D offsets for the nine sampling points is predicted, and the feature map is sampled accordingly. This mechanism enables the network to dynamically adjust its receptive field based on the position and shape of the target, thereby improving its adaptability to intra-class and inter-class variations in defect morphology. The prediction and sampling calculation process of the offset in deformable convolution is as follows:

$$\Delta p = f_{offset}(X, w_{offset}, b_{offset}) \tag{2}$$

$$Y(p) = \sum_{n=1}^{N} w_n X(p + p_n + \Delta p_n) \tag{3}$$

where Δp denotes the predicted offset for each sampling point P on the feature map X, f_{offset} represents the convolution operation used to generate the offsets, w_{offset} and b_{offset} are the learnable weights and biases of the layer, respectively, $Y(p)$ indicates the output at point P, N is the total number of sampling points, and Δp_n corresponds to the offset of the $n-th$ sampling point.

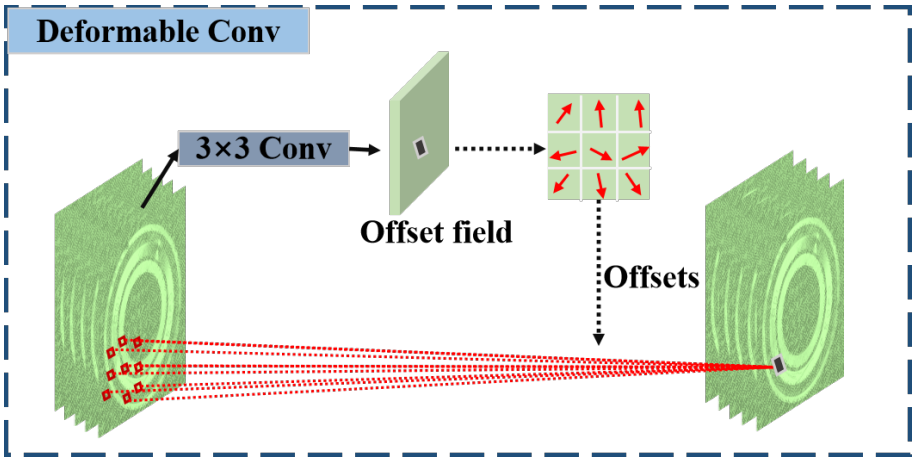

Figure 4. Feature map deformation sampling process.

The feature maps sampled by the deformable convolution are fed into two parallel branches for further processing. In each branch, layer normalization is first applied to maintain stable feature distributions. As shown in Figure 5, multi-scale sampling along the horizontal and vertical directions is performed in the two branches using depth-wise convolutions of different scales, thereby encoding the diversified features after geometric deformation. After multi-dimensional encoding, each branch employs 1 × 1 convolutions to rearrange and smooth cross-channel features, producing feature outputs F_x and F_y, respectively.

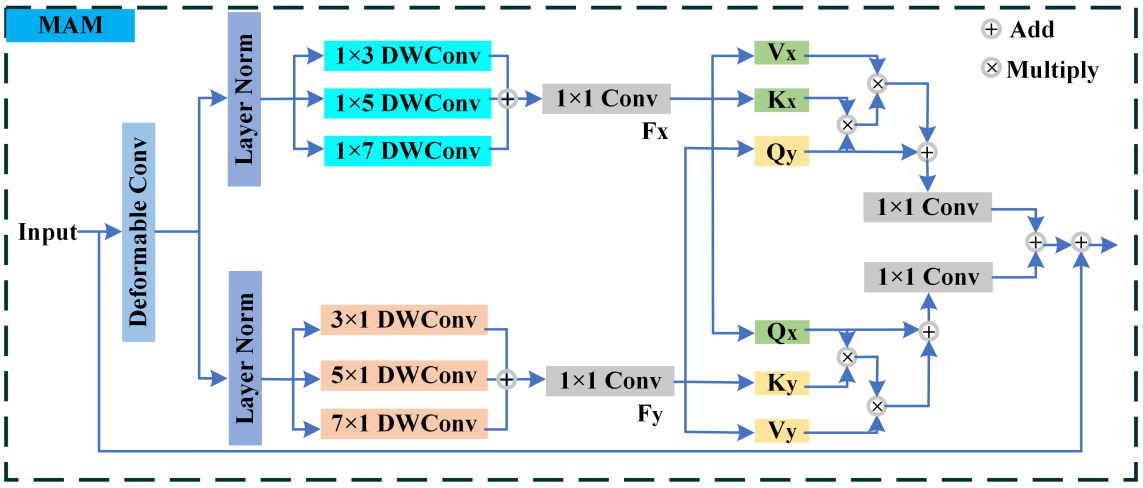

Figure 5. MAM network structure diagram.

To fully integrate contextual information from both branches, a novel dual-axis cross-attention mechanism is proposed. In the horizontal branch, F_x serves as the key (K) and value (V) matrices, while F_y acts as the query (Q) matrix, computing attention responses guided along the vertical direction. In the vertical branch, F_y is used as the key and value matrices, and F_x as the query, to compute attention along the horizontal direction. The computation of the dual-axis cross-spatial attention and the resulting feature outputs are formulated as follows:

$$\begin{cases} \text{Attention}_x = \text{Softmax}\left(\frac{Q_y K_x^\top}{\sqrt{d_{k1}}}\right) \\ O_x = \text{Attention}_x V_x + Q_y \end{cases} \tag{4}$$

$$\begin{cases} \text{Attention}_y = \text{Softmax}\left(\frac{Q_x K_y^\top}{\sqrt{d_{k2}}}\right) \\ O_y = \text{Attention}_y V_y + Q_x \end{cases} \tag{5}$$

where d_{k1} and d_{k2} denote the dimensions of the feature vectors from the horizontal and vertical branches, respectively, while O_x and O_y represent the output features after cross-weighted aggregation. This mechanism, through inter-directional feature interaction and attention re-calibration, significantly enhances the model's perception of complex defect boundaries and geometric shapes on bearing surfaces.

It is important to clarify the essential distinction between the proposed dual-axis cross-attention and existing coordinate attention [23] or axial attention [24] mechanisms, as all three involve operations along spatial axes. Coordinate attention factorizes channel attention into two 1D feature encoding processes along the height and width directions respectively, and then adds the resulting attention maps. This is essentially a variant of single-path self-attention where the two directional features are encoded independently and fused by simple addition. Axial attention performs self-attention sequentially along the height axis and then the width axis, where the two axial attentions are executed in a serial manner without mutual interaction between directions. In contrast, the proposed dual-axis cross-attention employs two parallel branches with multi-scale directional encoding, and introduces a bidirectional cross-attention mechanism where the features from the horizontal branch serve as the key and value matrices while the vertical branch provides the query, and vice versa, as formulated in Equations (4) and (5). This cross-directional interaction, which is absent in both coordinate attention and axial attention, enables the model to jointly reason about horizontal and vertical contextual information, leading to more comprehensive feature aggregation for defects with diverse shapes and orientations. Furthermore, the integration of deformable convolution prior to the dual-axis cross-attention allows adaptive sampling of geometrically varying defect features, which is not considered in existing axial-based methods.

In summary, through the collaborative design of deformable sampling, multi-scale directional encoding, and dual-axis cross-attention, the MAM achieves precise semantic alignment and adaptive enhancement of multi-shaped defect features while maintaining relatively low linear computational complexity.

2.3. Dynamic sample matching mechanism

In this work, the Dynamic Sample Matching (DSM) strategy is developed based on the Simplified Optimal Transport Assignment (SimOTA) framework [25], with several improvements introduced. It is divided into two components, *i.e.*, sample selection and loss function design. By integrating positional and classification information, the evaluation scores of positive and negative samples are dynamically balanced, and dual-label weighting is applied, enabling more effective sample selection and loss formulation.

In the SimOTA strategy, classification and regression losses are first computed for candidate anchors around each ground-truth (GT) box. Subsequently, a set of samples with the lowest losses is dynamically selected as positive samples, while the remaining samples are treated as negative samples. However, during the computation of classification loss, binary cross-entropy is calculated over all classes, incorporating both

matched and unmatched class information. In practice, the number of unmatched samples significantly exceeds that of matched samples, which may lead to the dominance of unmatched samples in the loss. Furthermore, equal loss weighting is assigned to all positive samples, which may result in samples with high IoU but low classification scores being regarded as high-quality, while samples with low IoU but high classification scores are underemphasized. To address these limitations, the classification factor is redefined in the DSM strategy. High-quality positive samples are dynamically selected, and greater emphasis is placed on samples with higher classification scores among those with similar positional scores.

2.3.1. Sample selection

As illustrated in Figure 6, in the DSM framework, the output feature maps are first utilized to perform a preliminary selection, forming a bag of candidate samples. Subsequently, these candidate samples are subjected to a fine-grained evaluation process. In SimOTA, sample selection is conducted based on a combination of classification and regression losses, where samples with the lowest losses are selected as positive samples. In contrast, in DSM, fine-grained selection is achieved by first computing a classification score C, which is then combined with the IoU to calculate the final score P, as follows:

$$C = {C1_i}^{\alpha} \times IoU + (1 - C2_i)^{\beta} \times (1 - IoU) \tag{6}$$

$$C1_i = \sum_{t=1}^{T} P_{i->t} Y_{j->t} \tag{7}$$

$$IoU = \frac{A \cap G}{A \cup G} \tag{8}$$

$$C2_i = \max_{c \neq c_{gt}} P_{i->t} \tag{9}$$

where $C1_i$ denotes the classification confidence of candidate box i for the GT class, $C2_i$ represents the maximum classification confidence of candidate box i for all non-GT classes, $P_{i->t}$ is the predicted probability of candidate box i for class t, Y_j^t indicates whether ground-truth target j belongs to class t (1 if yes, 0 if no), and T represents the set of all target classes. Meanwhile, IoU is the intersection-over-union between the candidate box A and the GT box G, representing the positional information of the sample. After obtaining the confidences of candidate samples across all classes, parameters α and β are used to control the influence of matched and unmatched classes.

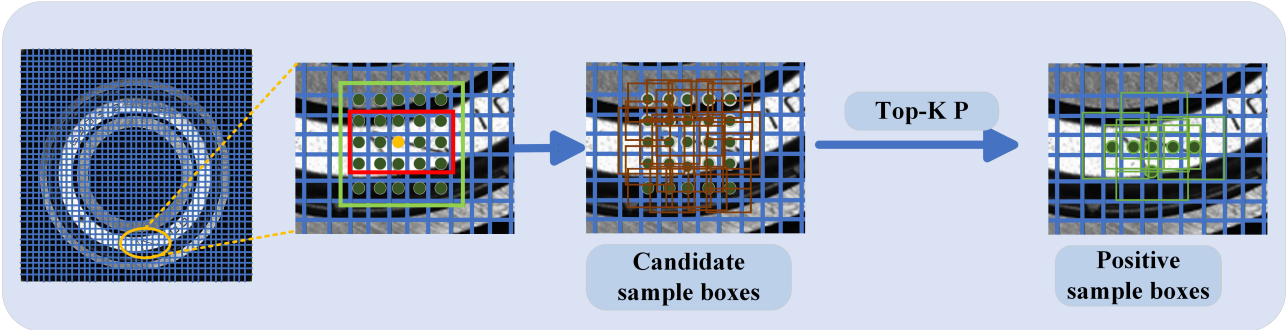

Figure 6. The sample processing procedure of DSM.

During this computation, IoU, $C1_i$, and $C2_i$ are used as weighting factors to adaptively adjust the classification score. When the IoU of different candidate samples is similar, the smaller the confidence of a candidate box for non-GT classes, the higher C will be. Conversely, for unmatched samples, the larger the IoU, the smaller C becomes. This indicates that this computation emphasizes high-IoU, class-matched candidate samples, ensuring they receive greater attention during sample selection.

The term $(1-C2_i)$ in Equation (6) serves as a soft penalty rather than a hard negative constraint. It directly leverages the maximum confidence among all non-ground-truth classes, effectively penalizing candidates that incorrectly assign high probability to any irrelevant category. This design is more flexible than a fixed penalty term, as it adapts to the actual prediction distribution of each candidate. Moreover, this formulation is not sensitive to the number of classes, because $C2_i$ is defined as the maximum over all non-GT classes, which is independent of the class count. Even with a large number of categories, the penalty remains determined solely by the most misleading prediction, preserving the discriminative focus on the most competitive false class.

After obtaining the classification score C, the positional information is incorporated to compute the final score P for evaluating candidate samples. The computation of P is given as follows:

$$P = C \times IoU^{\delta} \tag{10}$$

where the parameter δ serves to balance the relative importance of classification and positional scores. The candidate samples are then sorted according to their IoU values, and the top 10 are summed to obtain the value K. Finally, the positive samples are determined as the top K candidate samples with the highest P scores, while the remaining samples are treated as negative samples.

2.3.2. Loss function

Using the dynamic matching approach, high-quality samples with strong classification and positional scores are selected by the DSM strategy. These samples are subsequently employed for loss computation, with adaptive weighting assigned to each sample. The total loss L is calculated as follows:

$$L = L_{cls} + L_{obj} + \delta L_{reg} \tag{11}$$

where L_{cls}, L_{obj}, and L_{reg} represent the classification loss, objectness loss, and bounding box regression loss, respectively. The parameter δ serves the same purpose as described above. During training, gradients are computed for both positive and negative samples. Consequently, weighting coefficients for positive and negative samples are introduced in the loss calculation of each branch, following a procedure similar to the sample selection strategy. The weighting coefficients are calculated as follows:

$$\omega_{pos} = e^{C} \times C \tag{12}$$

$$\omega_{neg} = C \times (1 - IoU) \tag{13}$$

where ω_{pos} and ω_{neg} denote the weighting coefficients for positive and negative samples, respectively. To prevent gradient explosion, these coefficients are normalized. The losses of different branches are then computed by incorporating these weighting coefficients.

$$L_{\text{cls}} = \sum_{n=1}^{N} \omega_n^{\text{pos}} BCE\left(s_n^{\text{cls}}, 1\right) + \sum_{n=1}^{N} \omega_n^{\text{neg}} BCE\left(s_n^{\text{cls}}, 0\right) \tag{14}$$

$$L_{\text{obj}} = \sum_{n=1}^{N} (\omega_n^{\text{pos}}\, BCE\left(s_n^{\text{obj}}, 1\right) + \omega_n^{\text{neg}}\, BCE\left(s_n^{\text{obj}}, 0\right)) + \sum_{m=1}^{M} \text{FL}(s_m^{\text{obj}}, 0) \tag{15}$$

$$L_{\text{reg}} = \sum_{n=1}^{N} \omega_n^{\text{pos}} \times EIOU_n \tag{16}$$

where N and M denote the numbers of positive and negative samples, respectively, while s^{cls} represents the predicted classification scores and s^{obj} represents the predicted object confidence scores. Binary cross-entropy is used to compute the classification and confidence losses, and ω_{pos} and ω_{neg} are applied to perform dual-label weighted decoupling for positive and negative samples. In addition, Focal Loss (FL) is incorporated into the confidence loss branch to suppress negative samples, enabling the model to provide stronger negative supervision for difficult background samples. Furthermore, the Efficient Intersection Over Union (EIOU) loss is employed in the regression branch to enable more precise localization optimization for positive samples. To justify the choice of hyperparameters α, β, and δ, a grid search is conducted on the validation set. Keeping all other training settings unchanged, each of α, β, and δ is varied over 0.5, 1, 2, resulting in 27 combinations. The mean average precision (mAP) under each combination is evaluated. The experimental results demonstrate that the model achieves the best performance when $\alpha = \beta = \delta = 1$. Deviating any single parameter from 1 leads to a performance drop, with δ showing the highest sensitivity, while variations in α and β cause relatively minor fluctuations. Moreover, fine-tuning within a narrow range around 1 yields stable performance, confirming the robustness of this setting. Therefore, $\alpha = \beta = \delta = 1$ is adopted as the default configuration throughout this work.

2.4. Overall architecture

To address common issues such as missed detections and the loss of fine details in bearing surface defect detection, and to further enhance adaptability to multiple defect types under complex backgrounds, the MFDM model is proposed. The model is designed to improve the network's compatibility with multi-shaped defects. This section presents the overall network architecture, core modules, and specific parameter configurations. The architecture of MFDM is depicted in Figure 7, comprising a feature extraction backbone, a neck network, and a multi-scale detection head.

For practical deployment in industrial environments, a lightweight design is adopted for the feature extraction backbone. Feature extraction modules are employed to capture information at different levels, primarily consisting of DSEB and FEB modules. Through channel separation, reorganization, and shuffle operations, the total number of model parameters is significantly reduced.

The neck network is constructed based on a path-augmented feature pyramid structure. After abstract features are extracted from different receptive paths by the backbone, the Content-Aware ReAssembly of FEatures (CARAFE), a lightweight upsampling operator, is applied in the neck network to fuse positional information from lower-level features with high-level semantic features, thereby enhancing the representational capability of multi-scale features.

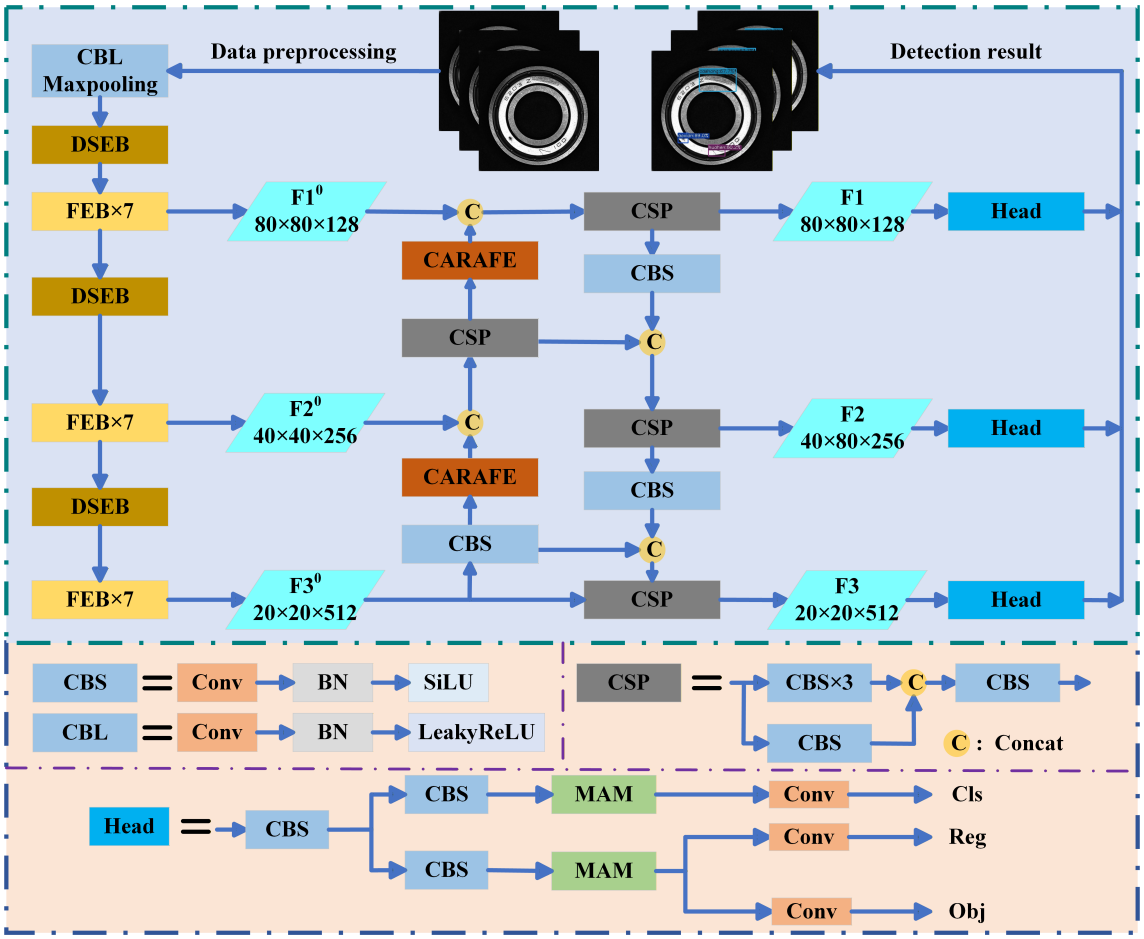

Figure 7. MFDM network architecture diagram.

In the detection head, the three-scale feature maps generated by the neck network are utilized to perform defect classification and bounding box regression. Two independent branches first adjust the channel dimensions of the feature maps via Conv-BatchNormalization-SiLU (CBS) blocks to generate classification and regression features. The multi-scale output features are subsequently refined using the MAM module to improve the discrimination of defects with diverse shapes, and adaptively weighted through a dual-axis cross-spatial attention mechanism. Finally, independent predictions for classification and regression tasks are produced using 1×1 convolutions.

During network training, the DSM mechanism is employed to perform confidence-based sample matching for the outputs of the multi-shape adaptive detection head. High-quality samples are thereby provided for network feedback, enhancing the model's discriminative ability across different defect types and complex backgrounds.

3. Experiment

3.1. Experimental environment

3.1.1. Dataset

In this study, images of defective bearings are acquired using industrial cameras. The image acquisition setup is illustrated in Figure 8a. The captured images are recorded in grayscale to enhance the visibility of surface defects on the bearings. Bearing surface defects are categorized into four types: dents, grooves,

abrasions, and scratches, as shown in Figure 8b. Using this setup, a total of 1000 images containing bearing surface defects are collected and manually annotated to form the dataset. The distribution of defect targets within the bearing surface image dataset is presented in Figure 9, which shows both the number of targets for each defect category and the number of images containing each category.

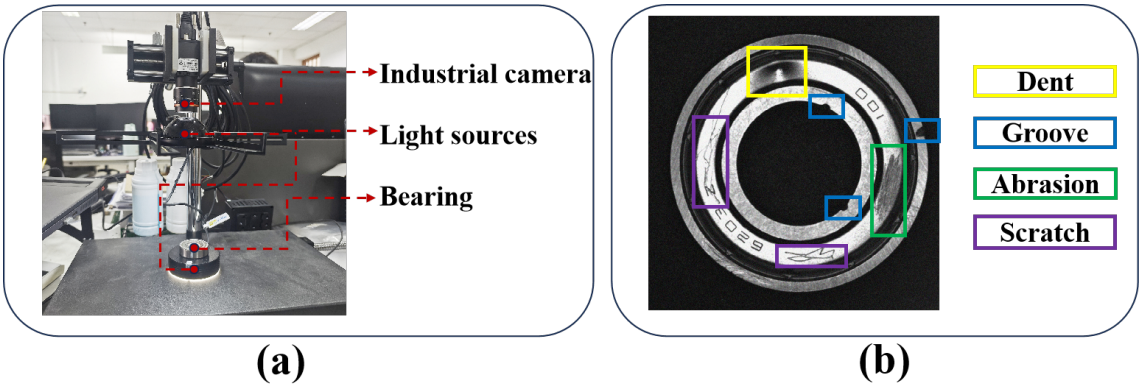

Figure 8. Experimental setup and defect samples. **(a)** Image acquisition platform; **(b)** Picture of bearing surface defects.

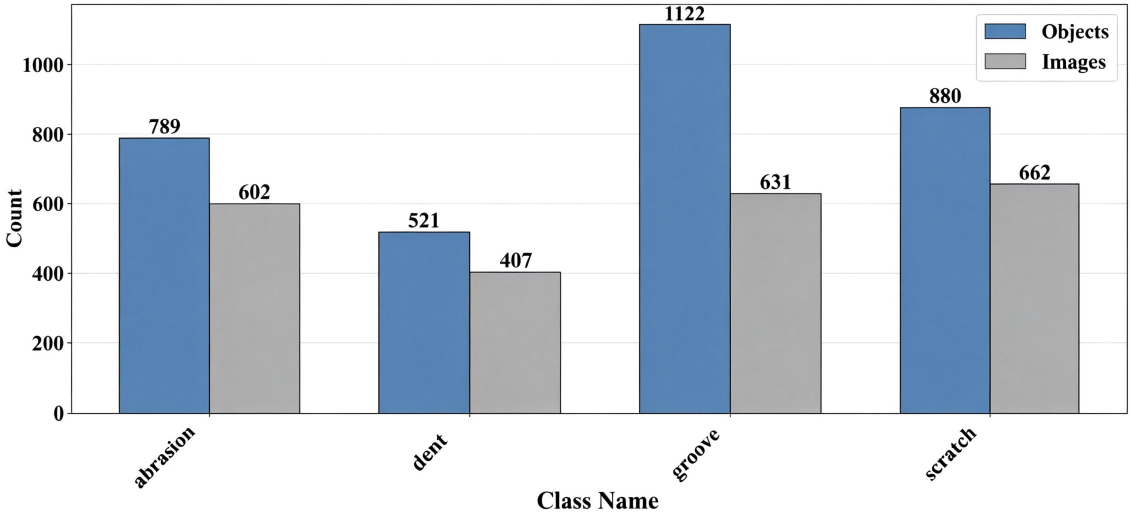

Figure 9. The target quantity for each defect category in the dataset, as well as the number of images that belong to that category.

3.1.2. Experimental settings

The experiments are conducted based on the PyTorch framework and trained using an NVIDIA GeForce RTX 3060 Laptop GPU. The computing platform is equipped with an Intel Core i7-12700H 2.30 GHz processor, with GPU acceleration applied during training. The dataset is split into training and validation sets in an 8:2 ratio. To ensure a fair comparison, all comparative models, including the YOLO series and other detectors, are trained with input images resized to the same resolution of 640 × 640, which is the standard configuration for YOLO models and is also adopted in MFDM network. Input images are resized to 640 × 640, and data augmentation techniques such as random flipping and mosaic are applied. The model is trained for 200 epochs, including a 10-epoch warm-up phase. A cosine annealing learning rate scheduler is adopted, combined with the SGD optimizer. The batch size is set to 4, with an initial learning rate of 0.005. Weight decay and SGD momentum are set to 0.0005 and 0.9, respectively.

3.1.3. Evaluation metrics

In this experiment, multiple evaluation metrics are adopted to comprehensively assess the performance of the MFDM model. Specifically, precision, mAP, recall, and F1-score are calculated to quantitatively evaluate the detection reliability of the model. The definitions and formulations of these metrics are provided below:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{17}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{18}$$

$$\mathrm{AP} = \int_0^1 \mathrm{P}(\mathrm{R})\,\mathrm{dR} \tag{19}$$

$$\mathrm{F1} = 2\,\frac{\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{20}$$

where true positives (TP) refer to the number of correctly predicted positive samples, false positives (FP) denote the number of negative samples incorrectly classified as positive, true negatives (TN) represent the number of correctly predicted negative samples, and false negatives (FN) indicate the number of positive samples incorrectly classified as negative. Precision (P) and recall (R) are calculated to obtain the average precision (AP) for each class [26]. The mAP, computed as the average of AP values across all defect categories, is employed as a key evaluation metric [27]. The F1 score, defined as the harmonic mean of precision and recall, is used to balance these two measures [28].

In addition, the number of model parameters and floating point operations per second (FLOPs) are employed as indicators of model lightweightness, allowing horizontal comparisons with mainstream object detection models. Lower parameter counts and FLOPs correspond to a simpler network structure and reduced inference computation [29]. By quantifying these metrics, the practical performance of the model in bearing surface defect detection can be comprehensively evaluated.

3.2. Experimental results and analysis

3.2.1. MFDM training and detection results

The training curves of MFDM are shown in Figure 10, including both the loss values and learning rate variations. It is observed that during the final 30 epochs, when data augmentation is disabled and the learning rate is reduced to its minimum for fine-tuning, all loss components experience a significant decline. This is attributed to the use of original images during training, which reduces the learning difficulty and accelerates convergence. After training, the MFDM model achieves a detection accuracy of 92.4%. In addition, the model's parameter count and FLOPs are 3.0 M and 13.6 G, respectively, demonstrating a clear advantage in lightweight design.

Given the moderate size of the dataset, which contains 1000 images in total with 800 images for training and the remaining 200 for validation, we carefully examined the potential for overfitting. The loss curves in Figure 10 show that the validation loss decreases steadily in parallel with the training loss throughout the entire optimization process and does not exhibit any increasing trend in the later epochs,

indicating that the model does not suffer from severe overfitting. This robustness can be mainly attributed to the use of data augmentation techniques such as random flipping and mosaic, which effectively increase the diversity of training samples, as well as the compact backbone architecture that inherently limits model capacity and provides structural regularization. Although we have not yet conducted cross-validation on an external bearing defect dataset, the held-out validation set of 200 images covers a wide variety of defect categories, scales, and background complexities, offering a reasonable basis for assessing generalization within the current experimental scope. Moreover, the qualitative detection results on challenging samples presented in Figure 11 demonstrate that MFDM reliably handles defects with low contrast or ambiguous morphology, further supporting its practical robustness. In future work, we plan to collect additional bearing surface data from different production lines and perform cross-dataset or cross-platform evaluations to more comprehensively verify the generalization ability of our method.

After training, bearing surface images that are not included in the training set are selected for detection validation. These images include challenging samples, such as multi-scale defects, defects with low background contrast, and defects with similar morphology. As illustrated in Figure 11, the MFDM model demonstrates excellent detection performance under complex conditions. Its robustness for practical applications is further confirmed.

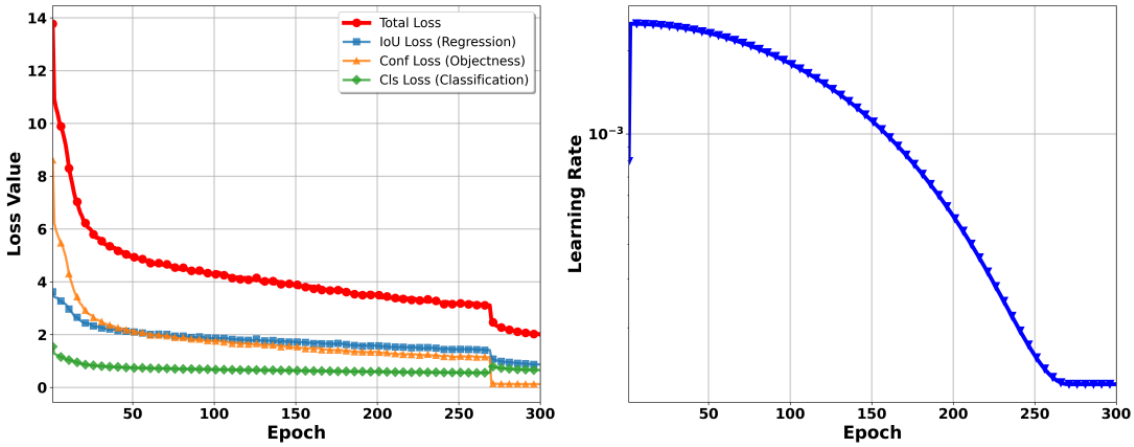

Figure 10. MFDM training curves.

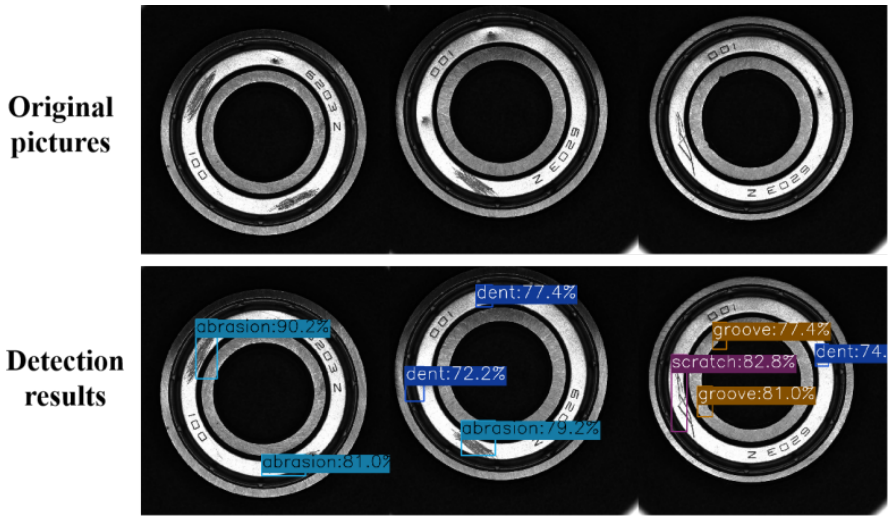

Figure 11. The test results of MFDM.

3.2.2. MFDM ablation experiment analysis

After completing the model training, the network is evaluated, achieving the parameter count of 3.0 M and the mAP of 92.4%. To verify the contribution of each module in the network and analyze their impact on detection performance, an ablation study is conducted. As shown in Table 1, the various innovative components of MFDM are implemented individually and evaluated through experiments.

Net 1 employs the backbone network composed of DSEB and FEB modules. The MAM module is not incorporated, and the SimOTA sample assignment strategy is adopted during training. To isolate the effects of ECA and shuffle within the backbone, we also design three variants with specific components removed, as detailed in Table 1. Net 2 is built upon Net 1, and the MAM module is introduced at the detection head to enable multi-shape adaptive feature perception. Net 3 is also based on Net 1, and the SimOTA strategy is replaced with the DSM training strategy, through which fine-grained sample selection is performed during the feedback process.

Table 1. Ablation networks.

Network	Usage Modules
Net 1	DSEB, FEB
Net 1-NoECA	DSEB, FEB (without ECA)
Net 1-NoShuffle	DSEB, FEB (without Shuffle)
Net 1-NoBoth	DSEB, FEB (without ECA and Shuffle)
Net 2	DSEB, FEB, MAM
Net 2-NoDeform	DSEB, FEB, MAM (without deformable convolution)
Net 3	DSEB, FEB, DSM
MFDM	DSEB, FEB, MAM, DSM

(1) Performance evaluation

After training all ablation models, the AP for each defect category, the overall mean value, and the number of model parameters are calculated. The results are presented in Table 2. It is shown that Net 1 achieves favorable defect detection performance while maintaining an extremely low parameter count. The lightweight nature and strong feature extraction capability of the proposed feature extraction modules are thus demonstrated.

Table 2. Ablation experimental results. Bold indicates the best result in each column.

Network	Groove (AP)	Dent (AP)	Scratch (AP)	Abrasion (AP)	mAP	Parameters (M)
Net 1	81.5%	87.7%	81.3%	78.7%	82.3%	**2.5**
Net 1-NoECA	80.2%	86.4%	79.8%	76.9%	80.8%	**2.5**
Net 1-NoShuffle	79.5%	85.9%	79.0%	76.0%	80.1%	**2.5**
Net 1-NoBoth	77.8%	84.2%	77.5%	74.5%	78.5%	**2.5**
Net 2	83.5%	89.8%	88.0%	85.9%	86.8%	3.0
Net 2-NoDeform	82.5%	88.5%	85.5%	83.5%	84.5%	2.9
Net 3	90.5%	90.9%	**91.9%**	89.8%	90.7%	**2.5**
MFDM	**91.3%**	**93.8%**	91.6%	**90.4%**	**92.0%**	3.0

To further investigate the respective roles of the ECA mechanism and the shuffle operation within the lightweight backbone, we conduct a comparative analysis of the three variants against the complete Net 1 configuration. As quantified in Table 2, the exclusion of ECA results in a 1.5% reduction in mAP, whereas the removal of shuffle leads to a more substantial performance degradation of 2.2%. When both components are simultaneously ablated, the mAP further declines to 78.5%. These observations confirm that both modules positively contribute to feature representation, with the shuffle operation playing a comparatively more critical role in facilitating cross-channel information flow, while the ECA module primarily enhances channel-wise feature discrimination. The complementary nature of their synergistic combination yields superior performance relative to any individual variant, thereby substantiating the architectural rationale of the proposed lightweight backbone design.

By introducing the MAM module for multi-shape feature alignment, Net 2 improves the network's adaptability to various defect types, showing particularly notable improvements in handling scratches and abrasions that are easily confused with the background. To further quantify the individual contribution of deformable convolutions within the MAM module, we design a variant named Net 2-NoDeform, which removes the deformable convolution operation while retaining the dual-axis cross-attention and multi-scale directional encoding. As shown in Table 2, Net 2-NoDeform achieves an mAP of 84.5%, which is 2.3 percentage points lower than that of Net 2. The performance drop is particularly evident for scratch and abrasion defects, whose AP values decrease by 2.5% and 2.4%, respectively. This observation demonstrates that deformable convolutions effectively enhance the model's ability to sample geometrically distorted defect features, especially for elongated or irregularly shaped surface anomalies. Meanwhile, the remaining attention mechanism still provides a clear improvement over Net 1, confirming that the dual-axis cross-attention also contributes positively. Overall, these results validate the synergistic effect of deformable sampling and adaptive attention in the proposed MAM.

Net 3 enhances detection accuracy by selecting high-quality samples to assist training, achieving a clear performance gain while maintaining the same parameter count as Net 1. By combining all proposed components and incorporating the DSM training strategy, the MFDM network achieves notable gains in detection accuracy over the ablation variants. Leveraging multi-level feature perception and fusion together with fine-grained sample selection during training, the model maintains a relatively compact parameter size while significantly improving overall detection performance. These results indicate that MFDM effectively satisfies both accuracy and efficiency requirements in practical applications.

(2) Heatmap visualization

Heatmaps serve as an intuitive visualization tool to reveal the regions of interest in convolutional neural networks. The color intensity is typically positively correlated with the strength of feature activation, and they are widely used to interpret model decision-making and feature focus capability. To systematically compare the roles of different modules in feature aggregation, several representative bearing surface images are selected for detection analysis. These images cover various types and difficulty levels of defects, including challenging scenarios with subtle defects or low contrast between defects and the background. As shown in Figure 12, the trained weights of Net 1, Net 2, Net 3, and MFDM are loaded, and heatmaps are generated at the output of the detection head for the same set of input images. The three variants designed for isolating the ECA and shuffle components are not included in this visualization, as

their performance is inferior to that of Net 1 with only marginal differences among themselves, and their quantitative impacts have been sufficiently characterized in Table 2. Similarly, the variant Net 2-NoDeform, which removes deformable convolutions from the MAM module, is also omitted from the heatmap analysis. The following analysis therefore concentrates on the more visually distinguishable effects contributed by the MAM and DSM modules.

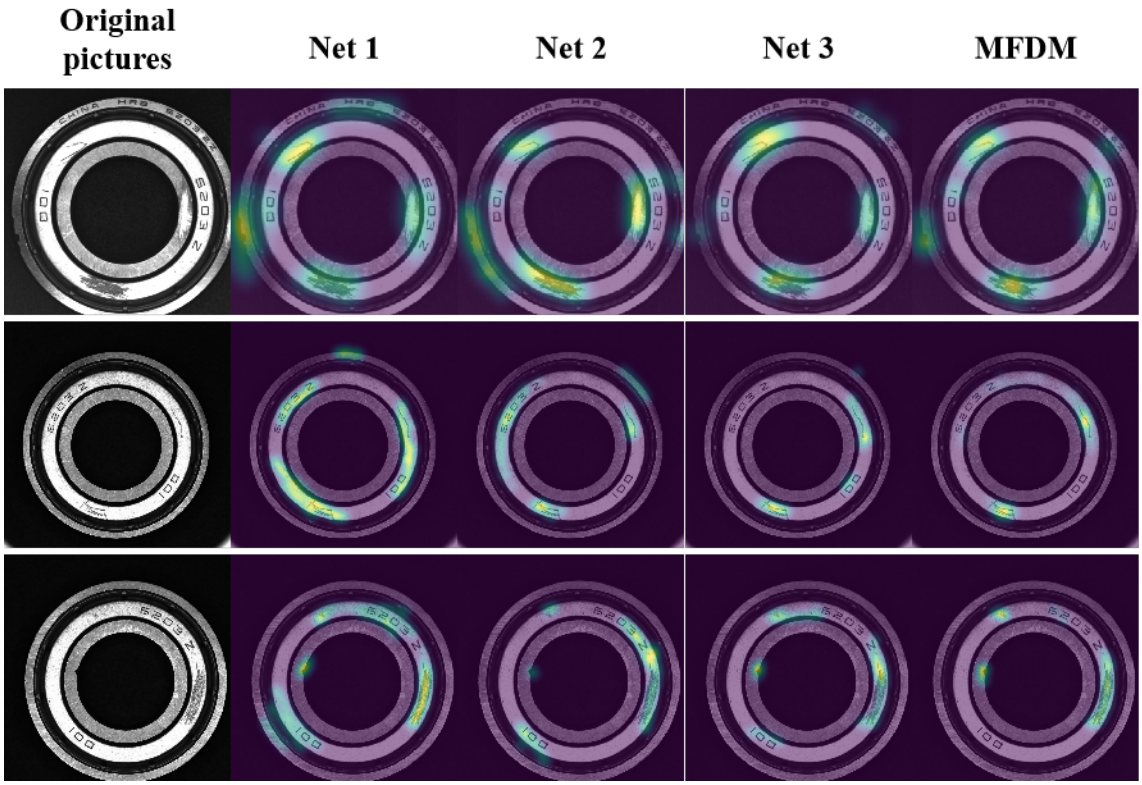

Figure 12. Visualization results of heat maps for different ablation networks.

From the heatmap visualizations of different defect types, it can be observed that there are significant differences in feature aggregation capability among the models. In the absence of the MAM and DSM modules, the model exhibits weak responses to defects with indistinct features, leading to limitations in recognition and an increased risk of missed detections. In contrast, the MFDM model, which incorporates all proposed modules, demonstrates substantially enhanced feature aggregation across various defect types. The corresponding heatmaps show more concentrated response regions that better align with the actual defect locations. This behavior highlights the effectiveness of multi-shape feature perception and dynamic sample adjustment during training. The complementary design of these components jointly improves the model's robustness in complex scenarios.

3.2.3. Comparison with mainstream object detection models

To demonstrate the effectiveness of the proposed method, the MFDM model is compared with several mainstream object detection models. In general, candidate regions are first generated on the bearing surface by two-stage detectors, followed by classification for target recognition. In contrast, classification and localization are directly performed on the input images by one-stage detectors for object detection. Since MFDM is designed as a one-stage end-to-end network, several representative one-stage models

are selected for comparison, including Single Shot MultiBox Detector (SSD) [30], RetinaNet [31], and several variants from the YOLO series, such as YOLOv5s [32], YOLOv8s [33], YOLOv10s [34], YOLOv11s [35], and YOLOv12s [36]. In addition, Faster R-CNN [37] is adopted as a representative two-stage detector for comparison. To ensure fairness, all models are trained under the same experimental settings. The settings include batch size and the number of training epochs.

As shown in Table 3, the comparison of AP values indicates that Faster R-CNN outperforms some one-stage models in the bearing surface defect detection task. This advantage may be attributed to its two-stage architecture, in which region proposal generation and classification are decoupled, thereby enabling stronger feature adaptability. However, a larger number of parameters and higher computational cost are also introduced by this design, which limits its practical applicability to some extent. With the development of dynamic threshold-based sample selection strategies, higher-quality positive and negative samples can be obtained during the training process, thereby effectively improving detection performance. As a result, the YOLO series models are generally observed to outperform traditional methods such as SSD, Faster R-CNN, and RetinaNet in this task. In the proposed MFDM model, detection accuracy is further enhanced by the introduction of multi-shape adaptive feature weighting and a dual-label dynamic sample selection mechanism. Compared with other models, significantly improved performance is achieved by MFDM. As illustrated in Figure 13, superior results across different categories of bearing surface defects are demonstrated by MFDM in terms of F1 score and recall.

To evaluate the practical inference speed for real-time industrial inspection, we measured the per-image inference latency of each model on an NVIDIA RTX 3060 Laptop GPU with a batch size of 1 and an input resolution of 640 $\times$ 640. The average time over 100 runs is reported in Table 4, together with the model parameters and FLOPs. As shown, MFDM achieves the lowest inference time of 10 ms per image, which is substantially faster than the other detectors and confirms its strong potential for deployment in production line inspection systems.

Table 3. AP values of different methods. Bold indicates the best result in each column.

Network	Groove (AP)	Dent (AP)	Scratch (AP)	Abrasion (AP)	mAP
SSD	65.8%	88.5%	79.2%	84.1%	79.4%
Faster-RCNN	84.5%	89.2%	85.7%	87.8%	86.8%
Retinanet	70.3%	89.8%	89.1%	84.2%	83.4%
YOLOv5s	88.1%	90.3%	82.9%	84.5%	86.5%
YOLOv8s	89.9%	90.9%	88.4%	86.4%	88.9%
YOLOv10s	89.3%	89.9%	88.1%	87.7%	88.8%
YOLOv11s	91.1%	92.7%	91.3%	**91.7**%	91.7%
YOLOv12s	89.4%	91.6%	90.9%	88.9%	90.2%
MFDM	**91.3**%	**93.8**%	**91.6**%	90.4%	**92.0**%

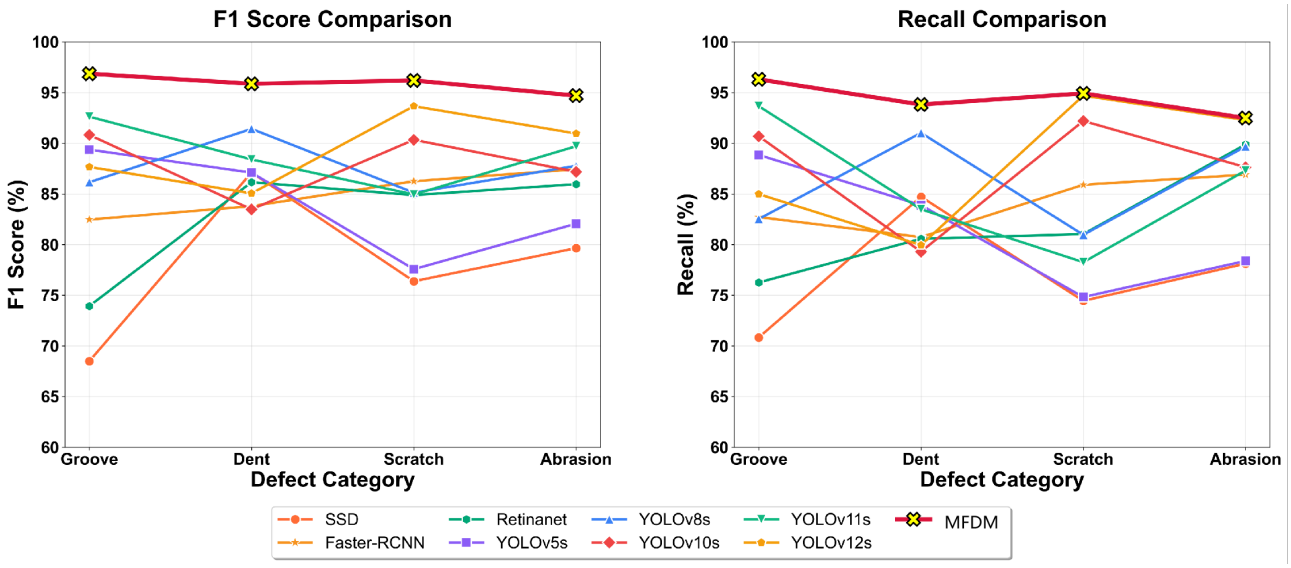

Figure 13. The F1 scores and recall rates of different comparison methods.

Table 4. Parameters, FLOPs and inference time of different methods. Bold indicates the best result in each column.

Network	Params (M)	FLOPs (G)	Time (ms)
SSD	27.2	67.5	30
Faster-RCNN	145.3	210.6	110
Retinanet	41.4	53.9	38
YOLOv5s	8.0	19.3	14
YOLOv8s	11.1	28.7	19
YOLOv10s	8.1	24.8	16
YOLOv11s	9.4	21.6	17
YOLOv12s	9.2	21.5	16
MFDM	**3.0**	**13.6**	**10**

In addition, the model structures of different networks are also analyzed. As shown in Table 4, a simpler architecture is exhibited by MFDM compared to other models. This indicates that better deployability on mobile devices is achieved by the more lightweight MFDM, making it more suitable for practical bearing surface defect detection applications.

4. Conclusion

The challenges of multi-scale defects and diverse defect categories on bearing surfaces are addressed in this work through the proposal of a multi-perception feature-guided dynamic sample matching detection network. A lightweight feature extraction backbone is designed to enhance adaptive feature learning while maintaining a simple network structure. To tackle the difficulty of distinguishing defect features from background information, a multi-shape adaptive perception module is introduced, through which detection maps with enhanced boundary and morphological features are generated. In addition, a dynamic sample selection strategy is employed to filter high-quality training samples, thereby improving the model's feature fitting capability. Through these innovations, the detection performance of multi-scale defects is effectively improved, and robustness in complex environments is enhanced.

Extensive experiments conducted on a self-constructed dataset demonstrate that MFDM outperforms several classical detection models across multiple evaluation metrics. As a novel tool for bearing surface defect detection, significant advantages in practical applications are exhibited by MFDM. Stronger adaptability under complex conditions is achieved through its feature enhancement mechanism and dynamic sample selection strategy, providing an efficient and reliable solution for defect localization and classification. Future work will focus on further optimization of the model structure to facilitate deployment on mobile devices while maintaining strong performance across multiple defect categories, thereby better meeting industrial requirements.

Data availability statement

The data or datasets that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used generative AI tools only to improve language and readability. Specifically, the authors used ChatGPT/Grammarly for language polishing only for drafting in limited sections. The authors take full responsibility for the content of the manuscript.

Acknowledgments

This work was funded by the National Foundation of Science with grant numbers 62322315 and 61873237.

Authors' contribution

Conceptualization, Xuan Chen and Yongyi Chen; methodology, Xuan Chen; software, Xuan Chen; validation, Xuan Chen, Yongyi Chen and Dan Zhang; formal analysis, Yongyi Chen; investigation, Xuan Chen; resources, Dan Zhang; data curation, Xuan Chen; writing—original draft preparation, Xuan Chen; writing—review and editing, Xuan Chen, Yongyi Chen and Dan Zhang; visualization, Xuan Chen; supervision, Yongyi Chen and Dan Zhang; project administration, Dan Zhang; funding acquisition, Dan Zhang. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

Dan Zhang holds the position of Associate Editor for *Mechatronics Technology* and has not peer reviewed or made any editorial decisions for this paper. The authors declare no conflicts of interest.

References

[1] Liu M, Zhang M, Chen X, Zheng C, Wang H. YOLOv8-LMG: an improved bearing defect detection algorithm based on YOLOv8. *Processes* 2024, 12(5):930.

[2] Liu X, Wu L, Guo X, Andriukaitis D, Królczyk G, *et al.* A novel approach for surface defect detection of lithium battery based on improved K-nearest neighbor and Euclidean clustering segmentation. *Int. J. Adv. Manuf. Technol.* 2023, 127(1):971–985.

[3] Tang B, Chen L, Sun W, Lin Z. Review of surface defect detection of steel products based on machine vision. *IET Image Proc.* 2023, 17(2):303–322.

[4] Chen Y, Ding Y, Zhao F, Zhang E, Wu Z, *et al.* Surface defect detection methods for industrial products: a review. *Appl. Sci.* 2021, 11(16):7657.

[5] Ullah W, Khan SU, Kim MJ, Hussain A, Munsif M, *et al.* Industrial defective chips detection using deep convolutional neural network with inverse feature matching mechanism. *J. Comput. Des. Eng.* 2024, 11(3):326–336.

[6] Ma Y, Yin J, Huang F, Li Q. Surface defect inspection of industrial products with object detection deep networks: a systematic review. *Artif. Intell. Rev.* 2024, 57(12):333.

[7] Peng Y, Xia F, Zhang C, Mao J. Deformation feature extraction and double attention feature pyramid network for bearing surface defects detection. *IEEE Trans. Ind. Inf.* 2024, 20(6):9048–9058.

[8] Shi J, Yang J, Zhang Y. Research on steel surface defect detection based on YOLOv5 with attention mechanism. *Electronics* 2022, 11(22):3735.

[9] Zhao Y, Liu Q, Su H, Zhang J, Ma H, *et al.* Attention-based multiscale feature fusion for efficient surface defect detection. *IEEE Trans. Instrum. Meas.* 2024, 73:1–10.

[10] Li L, Liu Z, Zhao H, Xue L, Wu J. The bearing surface defect detection method combining magnetic particle testing and deep learning. *Appl. Sci.* 2024, 14(5):1747.

[11] Zhao Z, Hu B, Feng Y, Zhao B, Yang C, *et al.* Multi-surface defect detection for universal joint bearings via multimodal feature and deep transfer learning. *Int. J. Prod. Res.* 2023, 61(13):4402–4418.

[12] Ma J, Hu S, Fu J, Chen G. A hierarchical attention detector for bearing surface defect detection. *Expert Syst. Appl.* 2024, 239:122365.

[13] Gao L, Jia S, Zhang G, Li Y, Yang M. Detection method for surface defects of insulated bearings based on improved Faster R-CNN. *Bearing* 2023, 4:1–8.

[14] Xu G, Liu C, Hu D, Xiong Y, Gu P, *et al.* Enhanced real-time defect detection for ceramic bearing balls using advanced feature extraction and optimization. *Meas. Sci. Technol.* 2025, 36(5):055014.

[15] Feng C, Zhong Y, Gao Y, Scott MR, Huang W. Tood: task-aligned one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021)*, Montreal, Canada, October 11–17, 2021, pp. 3490–3499.

[16] Zhou W, Wu Y, Qiang F, Yan W. Lightweight scope integration network for rail surface defect detection. *IEEE Trans. Big Data* 2025, 12(2):321–329.

[17] Wang A, Chen H, Lin Z, Han J, Ding G. LSNet: see large, focus small. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025)*, Nashville, USA, June 11–15, 2025, pp. 9718–9729.

[18] Van Den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, *et al.* Wavenet: a generative model for raw audio. *arXiv* 2016, arXiv:1609.03499.

[19] Zhou W, Zhu Y, Lei J, Wan J, Yu L. CCAFNet: crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images. *IEEE Trans. Multimedia* 2021, 24:2192–2204.

[20] Zhou W, Guo Q, Lei J, Yu L, Hwang JN. ECFFNet: effective and consistent feature fusion network for RGB-T salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* 2021, 32(3):1224–1235.

[21] Zhou W, Guo Q, Lei J, Yu L, Hwang JN. IRFR-Net: interactive recursive feature-reshaping network for detecting salient objects in RGB-D images. *IEEE Trans. Neural Networks Learn. Syst.* 2021, 36(3):4132–4144.

[22] Huang X, Liu W, Li M, Nie H. Cross-scale resolution consistent network for salient object detection. *IET Image Proc.* 2024, 18(10):2788–2799.

[23] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*, Online, June 19–25, 2021, pp. 13713–13722.

[24] Ho J, Kalchbrenner N, Weissenborn D, Salimans T. Axial attention in multidimensional transformers. *arXiv* 2019, arXiv:1912.12180.

[25] Ge Z, Liu S, Wang F, Li Z, Sun J. YOLOX: exceeding YOLO series in 2021. *arXiv* 2021, arXiv:2107.08430.

[26] Jung HK. YOLO-Drone: an efficient object detection approach using the ghosthead network for drone images. *arXiv* 2025, arXiv:2511.10905.

[27] Chen W, Li S, Han X. IDD-DETR: insulator defect detection model and low-carbon operation and maintenance application based on bidirectional cross-scale fusion and dynamic histogram attention. *Sensors* 2025, 25(18):5848.

[28] Terven J, Cordova-Esparza DM, Romero-González JA, Ramírez-Pedraza A, Chavez-Urbiola EA. A comprehensive survey of loss functions and metrics in deep learning. *Artif. Intell. Rev.* 2025, 58(7):195.

[29] Shahriar T. Comparative analysis of lightweight deep learning models for memory-constrained devices. *arXiv* 2025, arXiv:2505.03303.

[30] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, *et al.* SSD: single shot multibox detector. In *Proceedings of the 14th European Conference on Computer Vision (ECCV 2016)*, Amsterdam, The Netherlands, October 8–16, 2016, pp. 21–37.

[31] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017)*, Venice, Italy, October 22–29, 2017, pp. 2980–2988.

[32] Jocher G, Chaurasia A, Stoken A, Borovec J, Kwon Y, *et al.* Ultralytics/YOLOv5: v6.1-tensorrt, tensorflow edge tpu and openvino export and inference. 2022. Available: https://ui.adsabs.harvard.edu/abs/2022zndo...6222936J/abstract (accessed on 10 April 2026).

[33] Varghese R, Sambath M. YOLOv8: a novel object detection algorithm with enhanced performance and robustness. In *Proceedings of 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, Chennai, India, April 18–19, 2024, pp. 1–6.

[34] Wang A, Chen H, Liu L, Chen K, Lin Z, *et al.* YOLOv10: real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* 2024, 37:107984–108011.

[35] Lakshmi MCN, Pravallika PV, Sandhya P, Lavanya PV, Prasanth P, *et al.* YOLOv11: a next-generation approach to real-time object detection. *J. Nonlinear Anal. Optim.* 2025, 16(1):748–754.

[36] Alif MAR, Hussain M. YOLOv12: a breakdown of the key architectural features. *arXiv* 2025, arXiv:2502.14740.

[37] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS 2015)*, Montréal, Canada, December 7–12, 2015, pp. 91–99.