electronics

MDPI

Article

FMSD-YOLO12: An Efficient and Lightweight Network for Surface Defect Detection of Ferrite Permanent Magnets

Chuanyu Zhan [1], Haiting Yu [1], Ruize Wu [1] and Junfeng Li [1,2,*]

[1] School of Information Science & Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China; 2023329301054@mails.zstu.edu.cn (C.Z.); 2023312010 72@mails.zstu.edu.cn (H.Y.); 2023331200037@mails.zstu.edu.cn (R.W.)
[2] Changshan Research Institute Co., Ltd. of Zhejiang Sci-Tech University, Quzhou 324299, China
* Correspondence: ljf2003@zstu.edu.cn

Abstract

To address micro-break and edge-chipping defects in ferrite magnetic sheets, as well as the difficulty of balancing detection accuracy and deployment cost under complex grinding-texture interference, this paper proposes FMSD-YOLO12, an efficient and lightweight defect detection model based on YOLOv12. The proposed method follows a task-oriented design for three coupled challenges in ferrite magnetic sheet inspection, namely texture-interfered feature extraction, cross-scale feature inconsistency, and lightweight yet accurate defect localization. Specifically, a Spatially Re-weighted Convolution (SR-Conv) is introduced into the C3k2 backbone module to suppress repetitive grinding-texture noise and enhance the response contrast of subtle defect regions. A Context and Spatial Feature Calibration Network (CSFCN) is further developed to improve semantic consistency and spatial alignment during multi-scale feature fusion. In addition, a Lightweight Shared Detail-Enhanced Convolutional Detection head (LSDECD) is designed to strengthen weak-edge localization while reducing parameter redundancy through re-parameterization. Experimental results show that, with a comparable number of parameters, FMSD-YOLO12 improves mAP@50 by 2.40%, mAP@75 by 3.71%, and mAP@50-95 by 3.03% on the magnetic sheet defect dataset. These results demonstrate that the proposed model achieves a favorable balance between detection accuracy and computational efficiency for irregular defect detection under complex industrial backgrounds.

Keywords: surface defect detection; ferrite permanent magnets; YOLOv12; spatially re-weighted convolution; feature alignment

1. Introduction

Ferrite permanent magnet materials are widely used in critical fields such as electronic communication, automotive manufacturing, aerospace, and specialized micro-motors due to their high coercivity, strong corrosion resistance, and cost-effectiveness [1,2]. However, during complex manufacturing processes such as pressing, sintering, and grinding, surface defects—including micro-breaks, edge wear, dots, and edge chipping—are inevitably introduced due to process variations and environmental factors. These defects not only affect surface quality but may also degrade magnetic field uniformity and mechanical strength, thereby influencing the stability of downstream precision systems. Traditionally, surface quality inspection of magnetic sheets has relied on manual visual inspection. However, with the rapid growth in demand driven by the expansion of the new energy

Check for updates

Received: 26 March 2026
Revised: 26 April 2026
Accepted: 28 April 2026
Published: 30 April 2026

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

vehicle industry, along with increased production scale and speed, manual inspection has become less practical. In particular, strong interference from grinding textures leads to high labor intensity, low efficiency, and a high risk of missed detections, making it difficult to meet the requirements of modern industrial automation.

To address these limitations, extensive efforts have been made in automated defect detection. Early approaches mainly relied on traditional non-destructive testing techniques, such as eddy current testing [3], laser ultrasonic inspection [4], and magnetic particle testing [5]. These methods leverage electromagnetic or acoustic responses of materials and can achieve reliable detection under specific conditions. However, they often require strict inspection setups, incur high system costs, and lack the ability to precisely localize defects on complex surfaces. Subsequently, conventional machine vision-based methods [6], including edge detection (e.g., Sobel), threshold segmentation, and template matching, have been widely studied. Although effective under controlled conditions, these methods often exhibit limited robustness and generalization when dealing with reflective surfaces, complex textures, and large variations in defect scale, such as the coexistence of micrometer-level breaks and centimeter-level edge chipping.

In recent years, deep learning has achieved remarkable success in industrial inspection, anomaly detection, and fault diagnosis. To address critical challenges such as domain shifts, complex feature variations, and data scarcity, a variety of advanced frameworks have been actively explored. For instance, conditional distribution-guided adversarial transfer learning networks have been developed to effectively align multi-source domains and reduce distribution discrepancies [7], while spatial-channel collaborative multi-scale graph interaction deep transfer learning frameworks enhance transferable representations by jointly modeling spatial relationships, channel dependencies, and multi-scale interactions [8]. Additionally, adaptive fused domain-cycling variational generative adversarial networks have shown strong potential for mitigating data scarcity and enhancing robustness under limited sample availability [9].

However, for real-time visual surface defect detection tasks, the direct applicability of the aforementioned methods may still be limited. This is because they primarily focus on cross-domain adaptation and often involve relatively complex training procedures and considerable computational overhead, whereas practical surface inspection systems usually impose strict efficiency constraints. For such tasks, researchers have increasingly turned to dedicated object detection frameworks. Transformer-based methods, such as Real-Time Detection Transformer (RT-DETR) [10], leverage global context modeling to reduce false detections in complex backgrounds, yet their computational cost and model size often restrict deployment on edge devices. In contrast, one-stage detectors represented by the You Only Look Once (YOLO) series [11] utilize convolutional neural networks (CNNs) for hierarchical feature learning and achieve a favorable balance between detection accuracy and inference speed, making them highly suitable for real-time industrial vision applications [12]. Representative recent advances in this line, such as the NMS-free YOLOv10 [13], have further pushed the boundaries of real-time end-to-end detection by minimizing inference latency.

In the recently proposed YOLOv12 [14], Flash Attention and improved compound scaling strategies further enhance feature representation capability. Nevertheless, ferrite magnetic sheet inspection remains challenging due to low image contrast, repetitive grinding-induced textures, and large defect scale variation. Although existing YOLO-based defect detection methods have improved performance through attention enhancement, feature fusion redesign, or lightweight modifications, most of them do not explicitly address these coupled challenges in a unified manner. In addition, some lightweight industrial detection networks achieve lower computational cost at the expense of weakened feature

representation and less accurate localization for subtle or irregular defects. Therefore, a more task-oriented design is still needed to achieve a practical balance between detection accuracy and deployment efficiency in this scenario. To this end, we propose FMSD-YOLO12, an improved detection model for ferrite magnetic sheet surface defects. The proposed method is designed to enhance defect representation under texture interference while maintaining a lightweight architecture for industrial deployment. Its accuracy–efficiency trade-off is further evaluated against both mainstream and lightweight baseline models in Section 4. The main contributions of this work are summarized as follows:

(1) **Structured enhancement of feature extraction.** For texture-interfered industrial inspection, a task-oriented shallow feature enhancement module is introduced. Unlike conventional convolution or generic attention-based enhancement used in existing YOLO-based detectors, the proposed SR-C3k2 incorporates a spatially re-weighted convolution with a center-biased prior to suppress repetitive grinding-texture noise at the early feature extraction stage, thereby improving the signal-to-noise ratio of subtle low-contrast defects.

(2) **Efficient cross-scale feature fusion and alignment.** For cross-scale defect representation, a context and spatial feature calibration strategy is designed. Different from standard feature pyramid fusion or direct concatenation methods, the proposed CS-FCN jointly models contextual semantic refinement and spatial alignment, which improves the consistency between high-level semantics and shallow geometric details for irregular defects with large scale variation.

(3) **Lightweight re-parameterized detection head.** For real-time industrial deployment, a lightweight yet edge-sensitive detection head is developed. Compared with conventional decoupled heads or purely lightweight replacements, the proposed LSDECD combines cross-scale parameter sharing and re-parameterized detail-enhanced convolution, reducing parameter redundancy while strengthening the localization of weak and blurry defect boundaries.

The remainder of this paper is organized as follows. Section 2 reviews related work in object detection and industrial surface defect inspection. Section 3 presents the proposed FMSD-YOLO12 architecture and its key components. Section 4 reports experimental results and ablation studies. Finally, Section 5 concludes the paper and discusses future work.

2. Related Work

2.1. Object Detection Algorithms

Object detection has undergone a transition from handcrafted visual descriptors to deep neural network-based approaches. Traditional methods usually depended on manually engineered features, including Scale-Invariant Feature Transform (SIFT) [15] and Histogram of Oriented Gradients (HOG) [16]. These methods generally require dense window scanning and SVM-based classification, resulting in high computational redundancy and limited generalization ability in complex scenarios. To address these limitations, deep models such as Convolutional Neural Networks (CNNs) were introduced, overcoming this bottleneck through automatic feature learning. Subsequently, two-stage detectors represented by R-CNN [17,18] achieved significant improvements in detection accuracy. Nevertheless, the proposal generation stage introduces considerable latency, making such methods less suitable for real-time industrial inspection. To improve efficiency, the You Only Look Once (YOLO) series reformulates detection as a single regression problem, enabling end-to-end fast inference. Owing to its favorable balance between speed and accuracy, YOLO has become a preferred solution for real-time industrial detection.

The YOLO architecture has continuously evolved to further improve detection performance while maintaining efficiency. Early YOLO models established the one-stage

regression paradigm, and YOLOv2 introduced anchor boxes for improved localization and multi-scale detection [19], but they still suffer from anchor design complexity and limited adaptability. Subsequently, later YOLO variants incorporated CSPNet-style backbones to enhance feature extraction capability [20]. However, their reliance on anchor-based designs still constrains flexibility. To address these issues, YOLOv6 and YOLOv8 [21] transitioned to anchor-free architectures with decoupled heads, significantly improving convergence speed and detection performance while simplifying the model structure. Building upon this, YOLOv9 [22] introduces programmable gradient information and generalized efficient layer aggregation networks, effectively alleviating the problem of deep feature degradation. Furthermore, YOLOv11 [23] optimizes the backbone architecture by incorporating the C3k2 module, improving multi-scale feature fusion efficiency while maintaining lightweight characteristics and achieving better parameter efficiency and detection accuracy, thereby providing a stronger baseline for real-time deployment on edge devices.

In parallel with the evolution of mainstream detection architectures, lightweight object detection has become an important research direction for resource-constrained edge deployment. Representative examples include general-purpose mobile architectures such as MobileNetV4 [24], which improve efficiency through highly optimized lightweight building blocks, and ultra-lightweight detector variants such as YOLOX-Nano from the official YOLOX implementation [25], which further reduce model size and computational cost. However, although these models offer clear advantages in inference efficiency, they often suffer from weakened feature representation capability. In industrial scenarios, this limitation may lead to insufficient sensitivity to subtle edge gradients and tiny defects under complex textured backgrounds.

Nevertheless, many efficient CNN-based detectors still suffer from limited receptive fields, which restrict their ability to capture long-range dependencies. To address this limitation, YOLOv12 introduces a novel "attention-centralized" design paradigm. Unlike previous purely CNN-based architectures, YOLOv12 integrates Flash Attention and region-based attention mechanisms to enhance global context modeling. This hybrid design enables the model to capture long-range dependencies while retaining the efficient inference capability of CNNs, further improving the performance–computation trade-off.

2.2. Industrial Surface Defect Detection

Automated optical inspection based on machine vision has become a standard paradigm for quality control in modern manufacturing [26]. However, due to the complex interaction between intrinsic material properties and imaging conditions, surface defect detection of ferrite magnetic sheets still faces significant challenges. First, defects often exhibit extremely low contrast; for example, micro-breaks and edge chipping are difficult to distinguish from the dark gray background. Second, repetitive textures and granular noise are introduced during sintering and grinding processes, and these "pseudo-defects" are highly similar to real scratches in visual appearance [27]. In addition, the defect scale varies dramatically, ranging from micro-dots occupying only a few pixels to large fractures and edge chipping, which requires the model to simultaneously capture fine-grained local details and global semantic information.

To address these challenges, existing studies have explored various deep learning architectures. Early approaches mainly adopted standard CNNs [28], such as ResNet and VGG, to classify image patches; however, they lack a precise localization capability. More recent studies have shifted toward applying general object detection frameworks to complex industrial defect scenarios. In particular, for magnetic tile and ferrite magnetic sheet defect detection, a number of targeted improvements have been proposed. To handle the highly random distribution of magnetic tile defects and their tendency to be submerged

in complex backgrounds, Zheng et al. [29] proposed an Adaptive Rotation Attention Network (ARA-Net), which introduces Adaptive Rotated Convolution (ARC) to learn multi-view feature representations for capturing subtle defects. Wang et al. [30] developed SAB-YOLOv5 for permanent ferrite rotors, incorporating Atrous Spatial Pyramid Pooling (ASPP) and Bidirectional Feature Pyramid Network (BiFPN) to enhance small defect feature extraction and multi-scale fusion. To address the uneven distribution of defect features caused by curved surfaces, Li et al. [31] proposed YOLO-RDM, which integrates a Residual Parallel Attention (RPA) mechanism and depthwise separable convolution to improve localization accuracy while maintaining model efficiency. Zhang et al. [32] designed a YOLOv8-AHF-based detection network, where hybrid attention modules significantly reduce missed detections and false positives for small magnetic tile defects.

In addition to the above task-specific improvements, recent industrial surface defect detection studies have increasingly emphasized lightweight and deployment-oriented YOLO designs. For example, Ma et al. [33] proposed an improved lightweight YOLOv8 algorithm for steel surface defect detection, balancing detection accuracy and inference efficiency. Wang et al. [34] further developed a lightweight YOLOv11-based framework with an enhanced feature fusion module for small steel defect detection. These studies show that recent industrial defect detection research is not only concerned with accuracy enhancement, but also increasingly focused on compact architecture design for practical deployment. However, most of these methods are not specifically designed for scenarios where low contrast, repetitive texture interference, and irregular defect boundaries occur simultaneously.

Despite these recent advances, several limitations still remain. Standard convolution operators exhibit translation invariance but lack structured sensitivity to subtle spatial variations, making it difficult to effectively capture fine-grained defect features. This issue becomes more severe under strong grinding texture interference, where high-frequency background noise cannot be adequately suppressed, leading to degraded feature representation of low-contrast defects and consequently increased false detections and missed detections in complex scenarios. To address these limitations, we introduce a Spatially Re-weighted Convolution (SR-Conv) operator, which incorporates spatial prior information to suppress background noise and enhance the extraction of subtle defect features on ferrite magnetic sheet surfaces.

2.3. Multi-Scale Feature Fusion

For objects with large scale variations, effective multi-level feature representation is essential. FPN, proposed by Lin et al. [35], builds a top-down fusion pathway to integrate semantic information from deep layers with spatial details from shallow layers. However, the use of a single information propagation path limits the efficiency of feature utilization. To overcome this limitation, Liu et al. [36] introduced the Path Aggregation Network (PANet), which adds a bottom-up augmentation path to shorten the propagation distance of low-level localization information, thereby improving multi-scale detection performance.

Despite these advances, existing architectures typically rely on simple element-wise addition or channel concatenation during feature fusion, while neglecting the inherent semantic gap between features at different levels. As a result, "spatial misalignment" commonly occurs during the fusion process. To address this issue, Ming et al. [37] proposed AlignDet, which mitigates feature discrepancies through a two-stage pre-training framework. Li et al. [38] further introduced a flow-based method to optimize spatial consistency across layers, enabling adaptive feature alignment. In addition, Xiong et al. [39] proposed Deformable Convolution v4 (DCNv4), which dynamically adjusts sampling locations by learning pixel-wise spatial offsets. Related studies have also suggested that explicitly disentangling heterogeneous information can improve robust feature integration under

challenging visual conditions. For example, Tang et al. [40] proposed a divide-and-conquer confluent triple-flow network for RGB-T salient object detection, where modality-specific and modality-complementary information are separately modeled to improve robustness under challenging conditions. These observations together indicate that structured calibration, rather than direct fusion alone, is beneficial when feature inconsistency exists.

Inspired by these advances in feature alignment and structured information modeling, we adopt and extend the idea of Li et al. [41] and introduce a Context and Spatial Feature Calibration Network (CSFCN) for ferrite magnetic sheet defect detection. To address the dual challenges of strong texture interference and irregular defect boundaries, the proposed framework incorporates a Context Feature Calibration (CFC) module to aggregate global semantic information and suppress background noise, and a Spatial Feature Calibration (SFC) module to explicitly learn flow fields and perform grid-based resampling for precise cross-layer alignment. Extensive experiments demonstrate that this design not only enhances global semantic representation but also improves the localization of fine-grained defect contours, leading to superior multi-scale detection performance in complex industrial scenarios.

2.4. Convolutional Operators and Re-Parameterization

In computer vision tasks, standard convolution extracts semantic features by aggregating local pixel intensities. However, this isotropic aggregation mechanism is often insensitive to high-frequency information, such as subtle texture variations and edge gradients, leading to suboptimal performance in detecting low-contrast defects. To address this issue, Yu et al. [42] proposed the Central Difference Convolution (CDC), which explicitly computes the gradient differences between the center pixel and its neighboring pixels within the convolutional kernel, thereby enhancing the model's ability to capture fine-grained textures and local structural variations. This approach has demonstrated strong performance in texture recognition tasks, highlighting the effectiveness of gradient-based operators in extracting intrinsic structural features.

Meanwhile, improving representation capability while maintaining inference efficiency and lightweight design remains a critical challenge. To this end, structural re-parameterization and cross-scale parameter sharing have been widely explored. Ding et al. [43] proposed RepVGG, a re-parameterization technique that adopts a multi-branch structure during training and converts it into a single-branch structure during inference. Based on the linear additivity of convolution, this method merges complex training-time structures into a single equivalent convolution kernel, achieving performance gains without additional inference cost. In addition, RetinaNet adopts shared classification and regression subnets across different feature pyramid levels [44]. This cross-level parameter sharing reduces redundancy in the detection head and encourages the model to learn scale-consistent representations.

Building upon these insights, the proposed method integrates these strategies into a unified design. Specifically, the Lightweight Shared Detail-Enhanced Convolutional Detection (LSDECD) head combines parameter sharing and structural re-parameterization. By leveraging shared convolutional weights, the model reduces parameter redundancy, while the introduction of Detail-Enhanced Convolution (DEConv) further improves the representation of fine defect boundaries. This design effectively alleviates the issue of blurred weak defect contours without introducing additional inference cost.

3. Proposed Methodology

3.1. Base Model: YOLOv12

YOLOv12n is a lightweight variant of the YOLOv12 family, designed to balance detection accuracy and inference speed, making it suitable for real-time industrial defect

detection tasks. The network consists of three main components: a backbone, a neck, and a detection head, as illustrated in Figure 1.

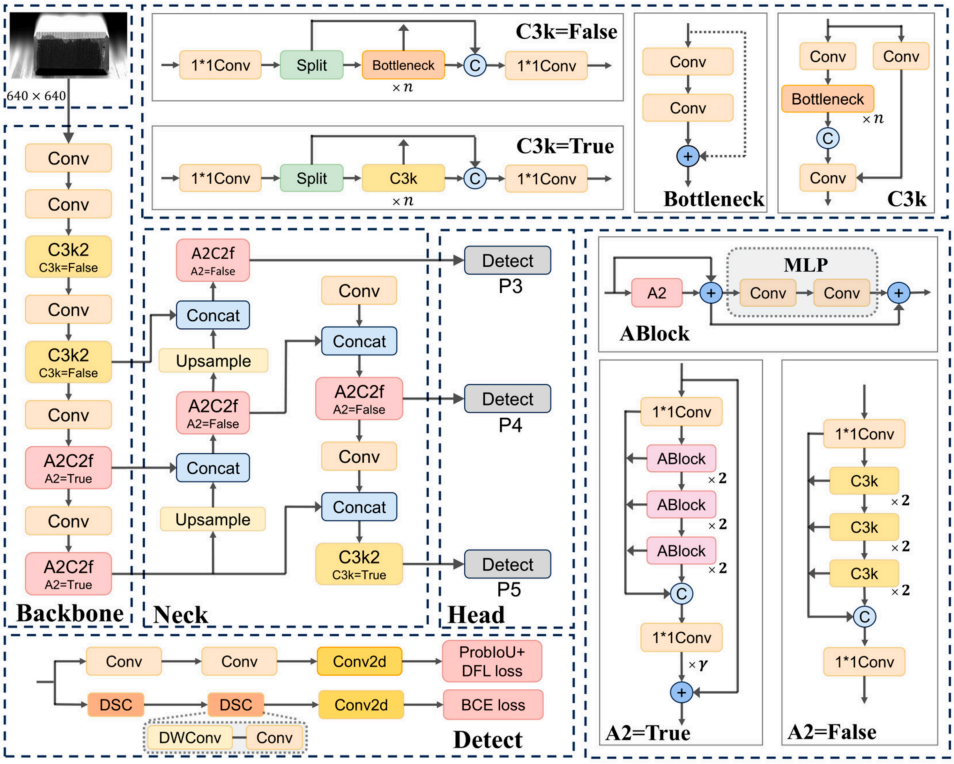


Figure 1. Architecture of the YOLOv12n model.

The backbone is responsible for extracting multi-scale features. To accommodate different feature hierarchies, a hybrid module design is adopted. In the shallow layers (P2, P3), a C3k2 module (C3k = False) composed of stacked Bottleneck blocks is employed, which focuses on capturing high-resolution details such as grinding textures and micro-breaks on the magnetic sheet surface. In the deeper layers (P4, P5), an A2C2f module (A2 = True) based on the R-ELAN architecture is introduced. This module removes the branch-splitting structure in the original ELAN design and instead adopts a chained stacking strategy with learnable residual connections. In addition, an embedded Area-Attention Block (ABlock) enables the modeling of long-range dependencies, which improves the representation of large-scale defects such as edge chipping.

The neck follows the Path Aggregation Network (PANet) architecture, where both top-down and bottom-up pathways are used to facilitate feature fusion across different scales. At this stage, the A2C2f module is configured in a pure convolutional mode (A2 = False), simplifying into a C3k-like stacked structure. This design maintains efficient multi-scale feature aggregation while avoiding the additional computational cost introduced by attention mechanisms, which is important for low-latency industrial applications. The detection head adopts an anchor-free decoupled design. For feature maps at three scales (P3, P4, P5), the classification and regression tasks are handled by separate convolutional branches. The classification branch employs Depthwise Separable Convolution (DSConv) to reduce computational overhead, while the regression branch retains standard convolution to preserve localization accuracy. During training, Distribution Focal Loss (DFL) and Binary Cross-Entropy (BCE) loss are jointly used for optimization.

Overall, benefiting from the R-ELAN architecture and the incorporation of area attention mechanisms, YOLOv12n demonstrates solid real-time performance and feature representation capability in general industrial detection tasks. However, for ferrite mag-

netic sheet surface defect detection, there remains room for improvement. In particular, challenges persist in suppressing high-frequency noise under strong grinding texture interference (affecting low-contrast defect recognition) and in addressing spatial misalignment of irregular defects (e.g., edge chipping and micro-breaks) during cross-scale feature fusion. To address these limitations, an improved model, FMSD-YOLO12, is proposed in this work.

3.2. Proposed Model: FMSD-YOLO12

In this work, an improved model, FMSD-YOLO12, is proposed for surface defect detection of industrial ferrite magnetic sheets. Owing to the dark-gray appearance of magnetic sheets, low image contrast and strong grinding-texture interference often degrade the representation of subtle defects. Meanwhile, large defect scale variation further increases the difficulty of effective feature extraction and localization. Motivated by these challenges, we reconstruct the baseline YOLOv12 framework and develop FMSD-YOLO12, whose overall structure is presented in Figure 2.

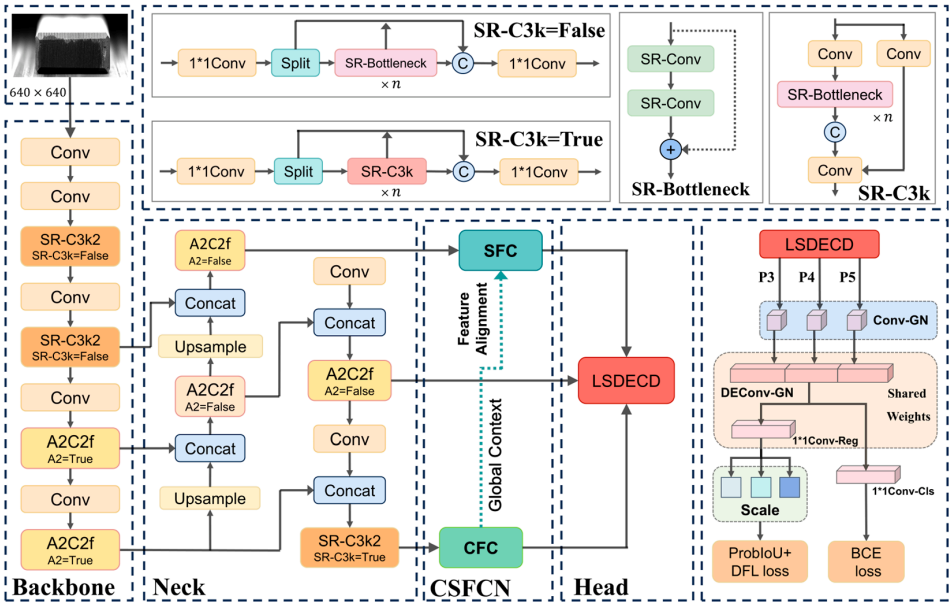


Figure 2. Architecture of the FMSD-YOLO12 model.

Rather than introducing isolated architectural modifications, FMSD-YOLO12 adopts a coordinated three-part design for texture suppression, cross-scale feature calibration, and lightweight edge-aware prediction. Specifically, a Spatially Re-weighted C3k2 module (SR-C3k2) is introduced into the shallow backbone to suppress high-frequency grinding-texture noise, a Context and Spatial Feature Calibration Network (CSFCN) is deployed in the neck to alleviate semantic inconsistency and spatial misalignment during feature fusion, and a Lightweight Shared Detail-Enhanced Convolutional Detection head (LSDECD) is designed to improve weak-edge localization while reducing parameter redundancy.

3.3. Spatial Re-Weighting Module: SR-C3k2

During the manufacturing process of ferrite magnetic sheets, dense high-frequency grinding textures are inevitably introduced on the surface. When detecting low-contrast defects (e.g., micro-breaks), the translation invariance and uniform weighting mechanism of standard convolution tend to cause the feature extraction process to be dominated by surrounding background noise. To address this issue, a Spatial Re-weighting module, termed SR-C3k2, is proposed to improve the feature signal-to-noise ratio at the early stage of feature extraction.

3.3.1. Density-Guided Spatially Re-Weighted Convolution

Inspired by recent studies on optimal weighted convolution in learning models [45,46], the proposed method enhances standard convolution by introducing a Spatial Density Prior to regulate the contribution of neighboring pixels to the central pixel. Based on this concept, a Spatially Re-weighted Convolution (SR-Conv) operator is designed. As illustrated in Figure 3, the SR-Conv operation consists of three stages: Spatial Prior Generation, Weights Re-weighting, and Feature Application.

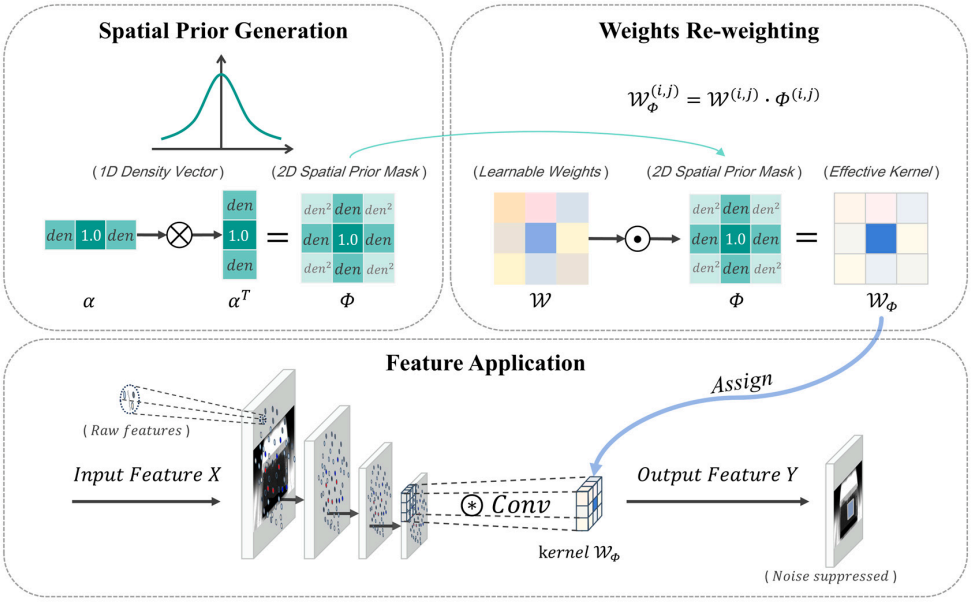


Figure 3. Operational mechanism of the SR-Conv operator. It incorporates a spatial density prior to re-weight learnable filters, acting as a soft-attention mechanism to suppress background noise and enhance defect features. (Note: The $\otimes$ denotes the outer product, $\odot$ denotes element-wise multiplication and $\circledast$ denotes the convolution operation).

The key difference between SR-Conv and standard convolution lies in the explicit incorporation of a spatial prior constraint into the convolution kernel. Given a standard learnable convolution kernel $W \in \mathbb{R}^{C_{out} \times C_{in} \times k \times k}$, a spatial prior mask $\Phi \in \mathbb{R}^{k \times k}$ is constructed based on a distance-decay principle. For the 3×3 case used in this work, the density distribution is controlled by a decay coefficient den ($den < 1.0$). A one-dimensional density vector is first defined as $\alpha = [den, 1.0, den]$, and a two-dimensional symmetric spatial prior matrix Φ is obtained via the outer product of α with itself (denoted as $\alpha\alpha^T$):

$$\Phi = \alpha\alpha^T = \begin{bmatrix} den^2 & den & den^2 \\ den & 1.0 & den \\ den^2 & den & den^2 \end{bmatrix} \quad (1)$$

This formulation yields a symmetric and separable center-biased prior, where the central position is preserved and peripheral responses are progressively attenuated. During forward propagation, the original convolution kernel W is modulated by Φ through element-wise multiplication, yielding a re-weighted kernel W_Φ:

$$W_\Phi^{(i,j)} = W^{(i,j)} \bullet \Phi^{(i,j)} \quad (2)$$

where (i, j) denotes the spatial position within the kernel. This operation introduces a center-biased inductive prior without adding extra learnable parameters.

After obtaining the equivalent convolution kernel, the feature extraction process can then be formulated as:

$$Y = W_{\Phi} \circledast X + B \tag{3}$$

where $\circledast$ denotes the discrete convolution operation, X and Y represent the input and output feature maps, respectively, and B is the bias term.

From a signal processing perspective, the proposed prior matrix can be interpreted as a center-weighted spatial window applied to the convolution kernel. Such a windowing operation reduces the contribution of peripheral positions in the receptive field and introduces a task-oriented low-pass bias, which is beneficial for suppressing repetitive high-frequency grinding textures. Meanwhile, the center-preserving structure helps retain locally coherent weak defect responses, thereby improving the local signal-to-noise ratio during early feature extraction.

From a Bayesian viewpoint, Φ can be interpreted as a location-dependent prior on the effective convolution weights, where positions closer to the kernel center are assigned higher prior importance and peripheral positions are down-weighted due to their greater uncertainty and higher susceptibility to background interference. Under this interpretation, SR-Conv combines learnable convolution weights with lightweight spatial regularization through $W_{\Phi}^{(i,j)} = W^{(i,j)} \bullet \Phi^{(i,j)}$, yielding an interpretable center-biased inductive bias without introducing additional learnable parameters. From a functional perspective (Figure 3, bottom), the resulting re-weighted kernel acts as a lightweight soft-attention mechanism that suppresses peripheral high-frequency activations while preserving defect-related central responses.

3.3.2. Architecture and Application of SR-C3k2

To integrate the advantages of SR-Conv into the detection framework, the original C3k2 module in YOLOv12 is restructured to form the SR-C3k2 module.

Figure 4 presents the structure of SR-C3k2. This module preserves the efficient feature reuse mechanism of the CSP framework, while replacing the standard 3×3 convolutions in the Bottleneck with SR-Conv operators to improve robustness against texture interference. In the forward process, a 1×1 convolution is first used to compress the input channels, after which the feature representation is divided into two paths. The shortcut path retains the original information, whereas the refinement path is processed by n stacked SR-Bottleneck blocks. Finally, the features from the two paths are merged along the channel axis to obtain the output representation.

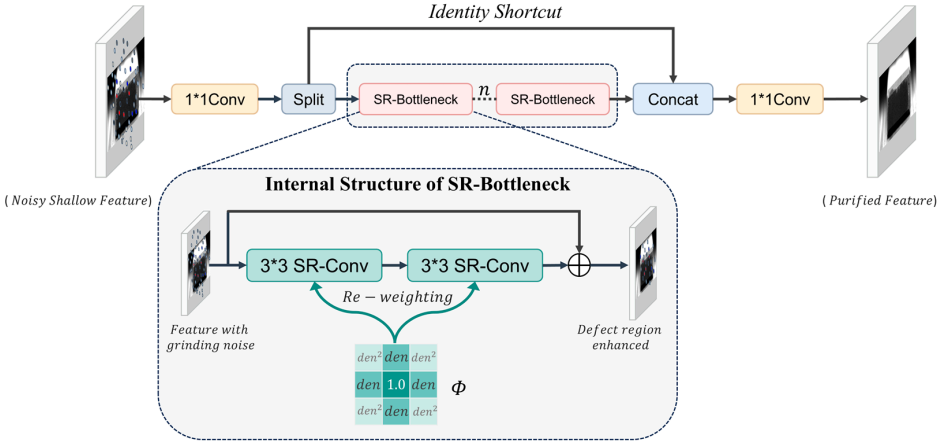


Figure 4. Architecture of the SR-C3k2 module. (Note: $\oplus$ denotes feature addition).

Within the FMSD-YOLO12 framework (Figure 2), SR-C3k2 is primarily deployed in the shallow stages of the backbone (P3 and P4). These layers possess high spatial resolution

and are responsible for extracting low-level features such as edges and contours, but they are also more susceptible to interference from high-frequency grinding textures. Through the introduction of SR-C3k2 at this stage, early feature representations can be moderated through spatial re-weighting. As features propagate through stacked SR-Bottlenecks, the built-in SR-Conv continuously suppresses high-frequency activation responses along edges using its static density mask Φ. Through multiple stacked processing steps, repetitive texture patterns are gradually suppressed, while defect-related structures are retained. This design provides cleaner feature inputs for subsequent modules (e.g., A2C2f and CSFCN), reducing the risk of noise accumulation across layers, while maintaining computational efficiency during inference.

3.4. Context and Spatial Feature Calibration Network: CSFCN

In ferrite magnetic sheet defect detection, defect scales vary significantly, ranging from micro-breaks occupying only a few pixels to large-area edge chipping. Shallow features preserve fine structures but are susceptible to background textures, whereas deep features contain richer semantics but often lose spatial precision. Conventional feature pyramid networks typically rely on bilinear upsampling and channel concatenation for multi-scale feature fusion. However, such rigid fusion strategies neglect both the semantic inconsistency between low-level and high-level features and the spatial displacement introduced during cross-scale aggregation, which may degrade the localization accuracy of irregular defect boundaries. To address this issue, a Context and Spatial Feature Calibration Network (CSFCN) is introduced, as illustrated in Figure 5. CSFCN consists of two complementary components: Context Feature Calibration (CFC), which refines contextual semantics and suppresses background interference, and Spatial Feature Calibration (SFC), which aligns semantic features with shallow geometric details. Through this two-stage design, CSFCN improves semantic consistency and spatial alignment during multi-scale feature fusion. The individual effects of CFC and SFC are further verified in Section 4.4.

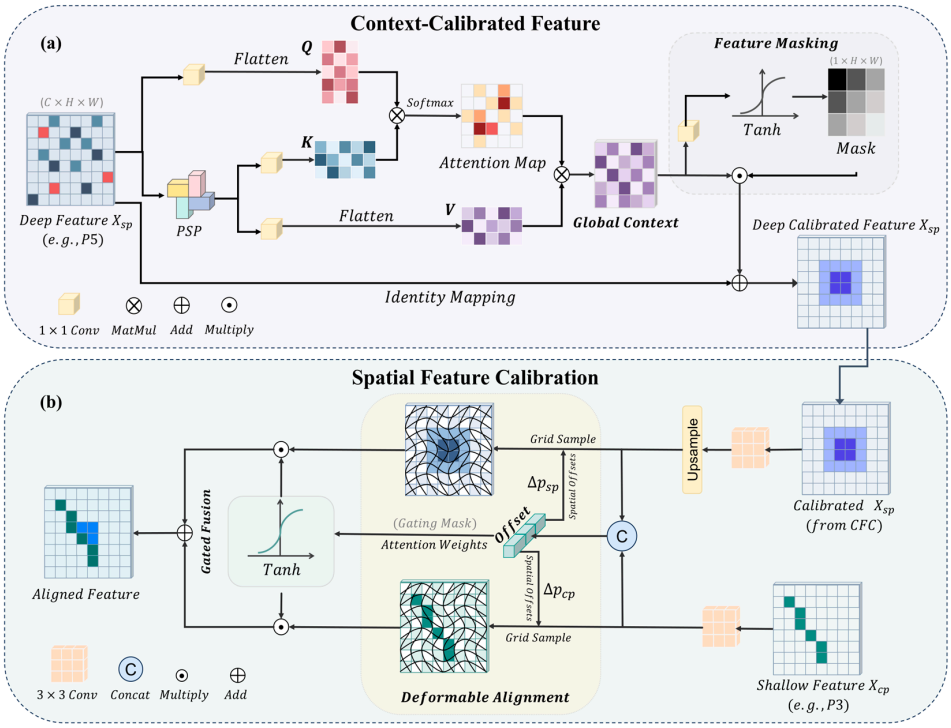


Figure 5. Detailed architecture of the proposed CSFCN. (**a**) The CFC filters out grinding noise and extracts refined global context. (**b**) The SFC dynamically aligns high-level semantic and current structural features via semantic flow-driven grid resampling.

3.4.1. Context Feature Calibration: CFC

Deep features (e.g., from the P5 level) possess large receptive fields and rich semantic information. However, they are also prone to interference from large-scale high-frequency grinding textures (as shown in Figure 5a), which may lead to pixel–context inconsistencies. Directly fusing such features can propagate and amplify noise.

To mitigate this issue, the CFC module is designed to extract refined global contextual representations. Given an input feature $X \in \mathrm{R}^{C\times H\times W}$, a reduced feature X_{red} is first obtained via a 1×1 convolution. The feature is then projected into a dense query matrix Q, while a Pyramid Scene Parsing (PSP) module is employed to aggregate contextual information at multiple spatial scales $S \in \{1\times 1, 2\times 2, 3\times 3, 6\times 6\}$. Based on this, the key matrix K and value matrix V that encapsulate the background prior are constructed as:

$$\begin{cases} Q = \theta_q(X_{red}) \in \mathbb{R}^{N\times d_k} \\ K = \theta_k(PSP(X_{red})) \in \mathbb{R}^{S\times d_k} \\ V = \theta_v(PSP(X_{red})) \in \mathbb{R}^{S\times d_v} \end{cases} \tag{4}$$

where $\theta_q, \theta_k, \theta_v$ denote 1×1 convolutional projections, $N = H\times W$ denotes the flattened pixels, and S represents the pooled sequence length. Next, a global attention map is computed by the inner product of the matrices Q and K:

$$A = \mathrm{Softmax}\left(QK^T\right) \in \mathbb{R}^{N\times S} \tag{5}$$

Then, the global attention map is aggregated with the value matrix V, and the result is reshaped to obtain the global context features:

$$C_{\mathrm{global}} = AV \in \mathbb{R}^{C'\times H\times W} \tag{6}$$

To further refine the contextual representation, a local attention mechanism is introduced. A spatial mask M is generated using the Tanh activation function and applied to re-weight the contextual feature. The calibrated output Y_{CFC} is then obtained via a residual formulation:

$$Y_{\mathrm{CFC}} = X_{\mathrm{red}} + \left(C_{\mathrm{global}} \odot M + C_{\mathrm{global}}\right) \tag{7}$$

After calibration, the output feature suppresses irrelevant background responses while preserving semantically meaningful regions (as shown in Figure 5a), providing more reliable guidance for subsequent feature fusion.

3.4.2. Spatial Feature Calibration: SFC

After obtaining the calibrated deep feature Y_{CFC}, it needs to be fused with a shallow feature (e.g., from the P3 level, denoted as X_{cp}). Shallow features contain precise geometric structures but limited semantic information (the cyan diagonal breaks in Figure 5b), whereas deep features provide coarse semantic localization (the blue halos in Figure 5b). Direct concatenation may therefore result in spatial inconsistency between semantic responses and structural details.

To address this issue, the SFC module explicitly models inter-layer semantic flow instead of relying on rigid fusion. First, the upsampled deep feature X_{sp} is concatenated with the shallow features X_{cp} and passed through an offset prediction layer to estimate spatial offset fields $\Delta p_{cp}, \Delta p_{sp} \in \mathbb{R}^{2\times H\times W}$ and fusion weights W_{att}:

$$[\Delta p_{cp}, \Delta p_{sp}, W_{att}] = Conv_{3\times 3}\left(Concat\left(X_{cp}, X_{sp}\right)\right) \tag{8}$$

Next, grid sampling ($\mathcal{G}$) is applied to perform deformable alignment:

$$\widetilde{X}_{cp}(p) = \mathcal{G}\left(X_{cp}, p + \Delta p_{cp}\right), \widetilde{X}_{sp}(p) = \mathcal{G}\left(X_{sp}, p + \Delta p_{sp}\right) \quad (9)$$

where p denotes spatial coordinates. This operation enables semantic responses from deep features to be adaptively aligned with structural details in shallow features.

Finally, the gated mask activated by Tanh adaptively fuses the aligned dual-layer features:

$$Y_{SFC} = \widetilde{X}_{sp} \odot \left(1 + Tanh\left(W_{att}^{sp}\right)\right) + \widetilde{X}_{cp} \odot \left(1 + Tanh\left(W_{att}^{cp}\right)\right) \quad (10)$$

After spatial calibration, the deep semantic highlights closely align with the shallow geometric breaks (as shown by the blue Aligned Feature region on the far left of Figure 5b). This demonstrates that the SFC mechanism can eliminate spatial misalignment issues in multi-scale fusion, significantly improving the model's accuracy in boundary regression for irregular and faint defects.

3.5. Lightweight Shared Detail-Enhanced Convolutional Detection Head: LSDECD

In the original YOLO architecture, the detection head typically allocates independent convolutional branches for feature maps at different scales (P3, P4, and P5). Although straightforward, this design introduces significant parameter redundancy in deep networks. Moreover, standard 3×3 convolutions tend to smooth local features, making it difficult to capture the high-frequency gradient variations associated with subtle defects such as micro-breaks and edge chipping, which may lead to missed detections. To enhance the perception of weak defect boundaries while maintaining a lightweight design, a Lightweight Shared Detail-Enhanced Convolutional Detection (LSDECD) head is proposed.

3.5.1. Cross-Scale Parameter Sharing and Adaptive Calibration

Inspired by the cross-scale consistency design in RetinaNet, we abandon the redundant structure of multi-branch independent designs and adopt a weight-sharing strategy. Specifically, the P3, P4, and P5 features from the CSFCN are first reduced to 256 channels and then fed into a fully shared feature extraction network. This design leverages the self-similarity of physical defects across different scales, significantly reducing parameter redundancy while improving generalization to multi-scale targets. To address the normalization inconsistency caused by parameter sharing, Group Normalization (GN) is adopted instead of Batch Normalization (BN), a strategy that has been widely validated in recent advanced defect detectors [47]. Since GN is independent of batch size, it provides more stable performance in cross-scale shared architectures. Furthermore, to compensate for the distribution differences in regression magnitudes across scales, a learnable scaling factor (Scale) is introduced after the shared regression branch. This enables adaptive calibration of bounding box predictions at different feature levels.

3.5.2. Re-Parameterized Detail-Enhanced Convolution: DEConv

To better capture the subtle gradient variations at defect boundaries, a Detail-Enhanced Convolution (DEConv) module is incorporated into the detection head. Following the recent work of Chen et al. [48] on detail-enhanced feature modeling, we utilize structural re-parameterization to construct a dynamic architecture featuring "multi-branch training and single-branch inference".

As illustrated in Figure 6, during training, DEConv consists of five parallel branches to explicitly capture edge information from different orientations: the standard convolution branch (Vanilla Conv) for basic low-frequency feature extraction, with weights W_{vanilla}; the Central Difference Convolution (CDC) branch for capturing isolated intensity variations,

with equivalent weights W_{cd}; the Horizontal Difference Convolution (HDC) branch for vertical edge detection, with weights W_{hd}; the Vertical Difference Convolution (VDC) branch for horizontal edge detection, with weights W_{vd}; and the Angular Difference Convolution (ADC) branch for diagonal edge modeling, with weights W_{ad}. Given an input feature map X, the training-time output is formulated as:

$$Y_{train} = \sum_{k \in \{vanilla,cd,hd,vd,ad\}} (W_k \circledast X + B_k) \tag{11}$$

where each branch explicitly models gradient differences between the central pixel and its neighbors, encouraging the network to learn high-frequency edge information and enhancing sensitivity to weak defect contours.

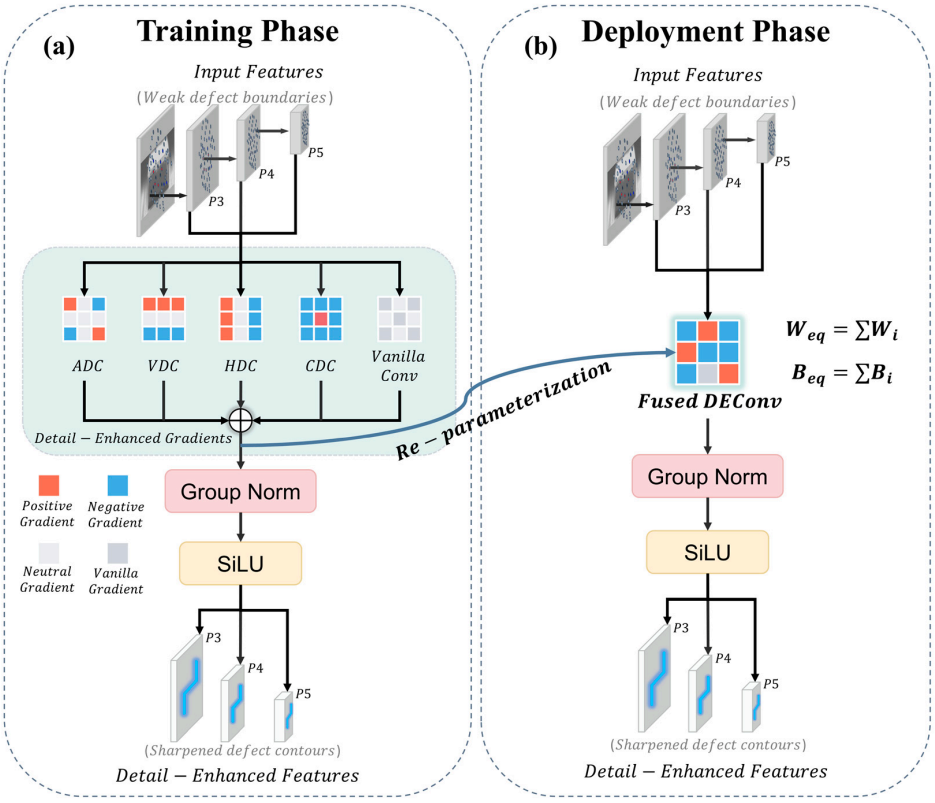


Figure 6. The structural re-parameterization mechanism of DEConv. During training, it employs five parallel branches including central, horizontal, vertical, and angular difference convolutions to capture high-frequency edge gradients. During deployment, it is fused into a single standard convolution without introducing additional inference cost.

Due to the linearity of convolution operations, all parallel branches can be equivalently merged into a single 3×3 convolution during inference, where $W_{eq} = \sum_k W_k$, $B_{eq} = \sum_k B_k$. Finally, the forward propagation during deployment is simplified to:

$$Y_{deploy} = W_{eq} \circledast X + B_{eq} \tag{12}$$

This re-parameterization strategy eliminates the computational overhead of differential branches during deployment, achieving improved edge representation capability without introducing additional inference latency. As a result, the model reduces missed detections caused by blurred or weak defect boundaries while preserving deployment efficiency.

4. Experiments and Analysis

4.1. Experimental Dataset

The experimental dataset was constructed from a real industrial production environment. As illustrated in Figure 7, ferrite magnetic sheets were transported into the imaging area by a conveyor-belt inspection system and captured by high-speed industrial cameras from multiple viewing angles. The imaging setup was relatively fixed, including the camera position, light source, and conveyor path, so that the acquired images could stably reflect practical industrial inspection conditions. However, due to the intrinsic dark-gray appearance of ferrite magnetic sheets and the repetitive grinding textures on their surfaces, the collected images still exhibited low contrast, strong texture interference, and slight appearance variations across samples.

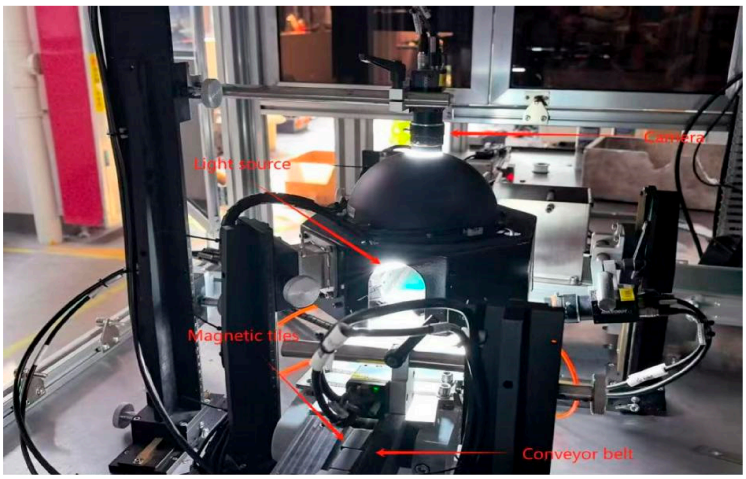
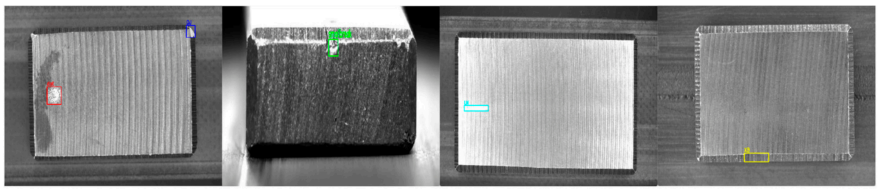


Figure 7. Industrial image acquisition system for ferrite magnets.

Since ferrite magnetic sheets have a relatively high production yield, defective samples are scarce in practice. Therefore, defect samples were manually screened and confirmed by quality inspection personnel before being included in the dataset. After screening, a total of 3592 images were obtained. The dataset was annotated using the LabelImg tool and saved in TXT format, and all labeled defect regions were manually checked to ensure category consistency and bounding-box quality. The dataset was finally divided into training, validation, and test sets at a ratio of 8:1:1.

The dataset contains five defect categories: Dot, DJ, Smallbreak, LM, and XB. Figure 8a–d shows representative annotated samples, and Figure 9 presents the instance distribution of each category. The numbers of instances for Dot, DJ, Smallbreak, LM, and XB are 537, 805, 960, 798, and 816, respectively, indicating a noticeable class imbalance caused by the natural occurrence frequency of different defects. Moreover, the dataset remains challenging. For example, Smallbreak often exhibits large scale variation and irregular defect boundaries, while XB is characterized by subtle low-contrast appearance and strong background interference from grinding textures.



(a) dot and DJ **(b)** smallbreak **(c)** LM **(d) XB**

Figure 8. Production anomaly data and annotations.

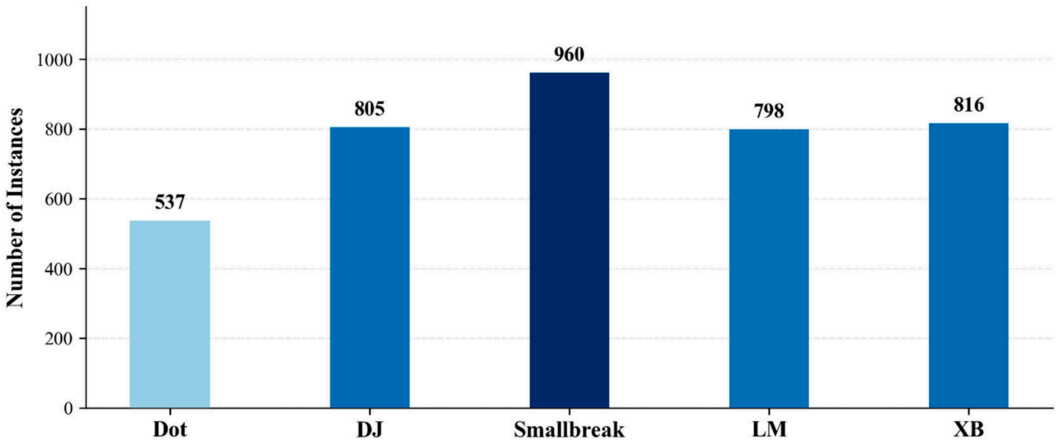


Figure 9. Distribution of defect instances in the custom ferrite magnet dataset.

4.2. Experimental Environment

The experiments were conducted on the Ubuntu 22.04 operating system with the PyTorch 2.3.0 deep learning framework. Detailed information about the experimental environment is provided in Table 1, while the training parameters are listed in Table 2. In addition to these settings, all compared models were trained and evaluated under the same protocol to ensure a fair comparison.

Table 1. Experimental Environment Configuration.

Category	Configuration
GPU	NVIDIA GeForce RTX 3090 24G
System Environment	Ubuntu 22.04
CUDA version	12.1
PyTorch version	2.3.0
Python version	3.12

Table 2. Training hyperparameters and implementation settings.

Parameters	Value
Image Size	640 × 640
Learning Rate (lr0)	0.01
Weight Decay	0.0005
Momentum	0.937
Optimizer	SGD
Batch Size	16
Warmup Epochs	3
Patience	50
Pretrained Initialization	True
Random Seed	0
Deterministic Training	True
Training Epochs	300
DataLoader Workers	4

During training, standard data augmentation in the Ultralytics YOLO pipeline was adopted, including HSV color augmentation, random translation, random scaling, horizontal flipping, and Mosaic augmentation. In this study, Mosaic was enabled throughout the whole training process, whereas MixUp, copy–paste, rotation, shear, perspective transformation, and vertical flipping were disabled.

For model selection, the best-performing checkpoint on the validation set was used for final testing. During inference, all models were evaluated on the test set with an input size of 640 × 640 and a batch size of 16, without test-time augmentation. Non-maximum suppression was applied with an IoU threshold of 0.7 to remove redundant predictions.

4.3. Evaluation Metrics

This study adopts Precision, Recall, F1-Score, mAP@50, and mAP@50-95 as evaluation metrics to assess detection accuracy. To further assess model efficiency, GFLOPs, parameter count, and inference speed (FPS) were also reported. The corresponding calculation formulas for these metrics are provided in Equations (13)–(17).

$$P = \frac{TP}{TP + FP} \tag{13}$$

$$R = \frac{TP}{TP + FN} \tag{14}$$

$$F1\text{-}Score = \frac{2TP}{2TP + FP + FN} \tag{15}$$

$$AP = \int_{r=0}^{1} P(r)dr \tag{16}$$

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n} \tag{17}$$

In these equations, TP, FP, and FN refer to true positives, false positives, and false negatives, respectively. Precision measures the correctness of predicted defect instances, whereas Recall reflects the ability of the model to detect actual defects. The reported F1-Score is averaged over defect categories, reflecting the model's overall balance between prediction accuracy and defect coverage. AP represents the detection accuracy of each individual category, and mAP is obtained by averaging AP over all defect classes to characterize the overall multi-class detection performance. The FPS reported in this paper denotes inference-only speed, excluding preprocessing and postprocessing.

4.4. Ablation Experiments

To examine the effectiveness of the proposed components in ferrite magnetic sheet defect detection, a series of ablation studies were carried out on the custom dataset. All experiments were performed under identical hyperparameter settings, with YOLOv12n adopted as the baseline model. The quantitative comparisons of the three main modules and the internal sub-components of CSFCN are summarized in Tables 3 and 4, respectively.

Table 3. Ablation experiments on FMSD-YOLO12.

SR-C3k2	CSFCN	LSDECD	P (%)	R (%)	F1-Score	mAP@50 (%)	mAP@75 (%)	GFLOPs	Params
×	×	×	85.9	79.0	79.8	84.0	76.1	5.8	2.5 M
✔	×	×	88.6	78.6	80.5	85.8	78.2	5.5	2.5 M
×	✔	×	88.5	78.8	80.7	85.5	76.6	6.7	2.9 M
×	×	✔	83.8	81.0	80.0	84.8	77.6	5.5	**2.2 M**
✔	✔	×	89.8	78.6	81.1	85.9	79.0	6.4	2.9 M
✔	×	✔	**91.2**	76.8	80.0	84.9	78.4	**5.2**	**2.2 M**
×	✔	✔	85.7	**83.4**	**82.6**	86.3	78.7	6.5	2.6 M
✔	✔	✔	85.8	81.9	81.8	**86.4**	**79.8**	6.1	2.6 M

Bold values indicate the best result in each column.

Table 4. Further ablation on the sub-components of CSFCN.

CFC	SFC	P (%)	R (%)	F1-Score	mAP@50 (%)	mAP@75 (%)	GFLOPs	Params
×	×	85.9	79.0	79.8	84.0	76.1	**5.8**	**2.5 M**
✔	×	**90.9**	76.9	80.1	84.9	78.0	6.0	2.8 M
×	✔	86.0	**80.6**	**80.7**	84.9	**78.3**	6.6	2.7 M
✔	✔	88.5	78.8	**80.7**	**85.5**	76.6	6.7	2.9 M

Bold values indicate the best result in each column.

As shown in Table 3, the three proposed modules improve the detector from different aspects. When SR-C3k2 is introduced alone, mAP@75 increases from 76.1% to 78.2%, indicating that the spatially re-weighted convolution can suppress repetitive grinding-texture interference at the shallow stage and improve the localization of subtle defects under a stricter IoU criterion. When CSFCN is added alone, mAP@50 increases from 84.0% to 85.5%, suggesting that the joint modeling of contextual refinement and spatial calibration enhances cross-scale feature consistency, which is particularly beneficial for defects with irregular shapes and large scale variation. The respective contributions of its two sub-components, CFC and SFC, are further analyzed in Table 4. When LSDECD is used alone, the parameter count decreases from 2.5 M to 2.2 M while mAP@75 improves from 76.1% to 77.6%. This shows that the proposed detection head reduces parameter redundancy while improving high-IoU localization by enhancing sensitivity to weak edge gradients via DEConv. Therefore, the three modules contribute to performance improvement in different ways, namely noise suppression, feature alignment, and lightweight edge-aware localization.

To further investigate the internal effectiveness of CSFCN, we separately evaluated Context Feature Calibration (CFC) and Spatial Feature Calibration (SFC) on the baseline YOLOv12n. As shown in Table 4, the two sub-components improve the detector in different aspects. Specifically, when only CFC is introduced, Precision increases markedly from 85.9% to 90.9%, indicating that contextual calibration effectively suppresses misleading background responses and enhances semantic discriminability. However, Recall decreases slightly to 76.9%, suggesting that CFC alone tends to produce more conservative predictions. In contrast, using only SFC improves Recall from 79.0% to 80.6% and increases mAP@75 from 76.1% to 78.3%, demonstrating that spatial calibration is more effective in preserving defect details and improving localization accuracy. When CFC and SFC are combined to form the complete CSFCN, the model achieves the highest mAP@50 of 85.5%, while maintaining a more balanced trade-off between Precision and Recall. Although the SFC-only variant achieves a slightly higher mAP@75, the complete CSFCN provides better overall balance and stronger robustness across evaluation metrics, which is more desirable for practical defect detection. These results indicate that CFC and SFC play complementary roles: CFC mainly enhances semantic discrimination and false-positive suppression, whereas SFC contributes more to spatial alignment and fine-grained localization. Therefore, the complete CSFCN is adopted in the final model due to its more robust and balanced overall performance.

The pairwise combinations in Table 3 further reveal the complementary roles of the proposed modules. The combination of SR-C3k2(A) and LSDECD(C) achieves 91.2% Precision with only 5.2 GFLOPs and 2.2 M parameters, indicating that suppressing shallow texture noise and strengthening edge-sensitive prediction can jointly reduce false positives while preserving a lightweight design. In contrast, the combination of CSFCN(B) and LSDECD(C) achieves the highest Recall (83.4%) and F1-Score (82.6%), together with a higher mAP@50 of 86.3%. This suggests that context-aware feature calibration and detail-enhanced localization are more effective in reducing missed detections, especially for irregular defects whose semantic cues and boundary details need to be jointly preserved. Meanwhile, the combination of SR-C3k2 and CSFCN improves mAP@75 from 76.1% to

79.0%, showing that suppressing grinding-texture interference at the shallow stage can provide cleaner inputs for subsequent cross-scale calibration. In other words, SR-C3k2 mainly improves feature quality before fusion, whereas CSFCN further refines semantic consistency and spatial alignment after fusion. Their sequential effects explain why the A + B configuration yields a clear gain in strict localization accuracy.

Building upon these improvements, the fully integrated model, FMSD-YOLO12, achieves the best overall performance. Compared with the baseline, Recall increases to 81.9% (+2.9%), mAP@50 improves to 86.4% (+2.4%), and the more stringent mAP@75 significantly increases to 79.8% (+3.7%), while maintaining a lightweight model size of only 2.6 M parameters. These results show that jointly optimizing shallow noise suppression, cross-scale feature calibration, and edge-aware prediction leads to more robust overall detection performance without sacrificing lightweight deployment.

It is also worth noting that the fully integrated model does not achieve the best value in every individual configuration. Rather than indicating direct conflict among modules, this phenomenon reflects their non-linear coupling and different optimization emphases. Specifically, SR-C3k2 mainly suppresses grinding-texture interference and reduces false positives at the shallow stage, CSFCN improves semantic consistency and cross-scale spatial alignment during feature fusion, and LSDECD strengthens edge-sensitive localization while reducing parameter redundancy in the detection head. Since these modules optimize different stages of the detection pipeline, their joint effect is manifested as a rebalancing among precision, recall, and localization accuracy, rather than a simple linear superposition. Therefore, the ablation results indicate that the proposed modules are functionally complementary, and the final model achieves the best overall balance by jointly optimizing noise suppression, feature alignment, and edge-aware localization.

To further support this interpretation, we conducted a lightweight feature similarity analysis at the P3 level by computing the cosine similarity between defect-region features and background-region features after ROI average pooling, where a lower cosine similarity indicates better feature separability and less confusion between true defects and background textures. As shown in Table 5, the similarity decreases from 0.3629 in the baseline model to 0.3472 for the A + B configuration, and further to 0.2472, 0.2332, and 0.2186 for the A + C, B + C, and full (A + B + C) models, respectively. This trend suggests that the proposed module combinations progressively reduce feature confusion between defects and grinding-texture background. In particular, the full model achieves the lowest defect–background similarity, which provides feature-level evidence that the three modules are functionally complementary and jointly improve feature discriminability. This interpretation is also consistent with the feature heatmaps shown in Figure 10. Compared with the baseline model, which produces diffuse and blurred activations due to interference from complex grinding textures, FMSD-YOLO12 exhibits more concentrated responses on defect regions and sharper activations around subtle edges, further supporting the complementary effect of the proposed modules.

Table 5. Lightweight P3-level defect–background feature similarity analysis of different module combinations.

Model Variant	P3 Defect–Background Cosine Similarity
Baseline	0.3629 ± 0.1427
A + B	$0.3472 \pm 0.1307 \downarrow$
A + C	$0.2472 \pm 0.1289 \downarrow$
B + C	$0.2332 \pm 0.1449 \downarrow$
A + B + C	$\mathbf{0.2186 \pm 0.1361 \downarrow}$

Bold values indicate the best result in each column. The downward arrow (↓) indicates a decrease compared with the baseline.

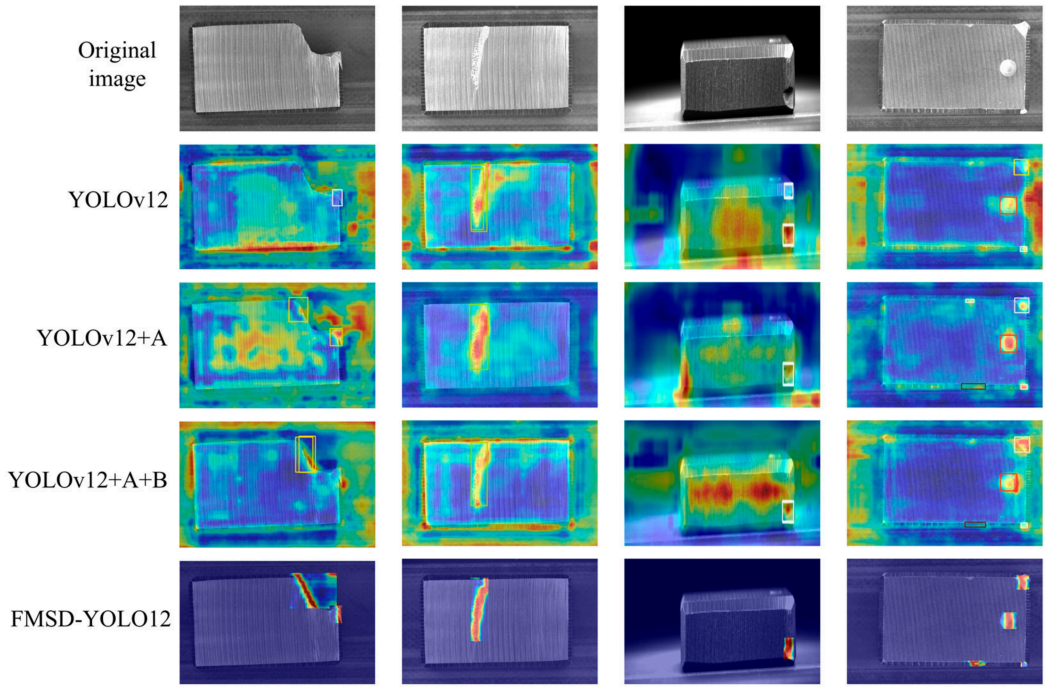


Figure 10. Feature heatmap comparison of different module combinations. A represents the SR-C3k2 module, B represents the CSFCN module, and FMSD-YOLO12 represents the final model incorporating the additional LSDECD module.

4.5. Sensitivity Analysis of the Density Parameter in SR-Conv

In the SR-C3k2 module, inspired by the optimal weighted convolution theory proposed by Cammarasana et al., we reshape the weight distribution of convolution kernels by introducing a two-dimensional spatial prior matrix Φ. According to their study, Φ is constructed via the outer product of a one-dimensional density vector $\alpha = [den, 1.0, den]$. Therefore, the hyperparameter *den* (density attenuation coefficient) directly controls the spatial attention distribution within the convolutional receptive field. To validate the rationality of the parameter selection and explore the optimal configuration, we conducted sensitivity analysis experiments by introducing SR-Conv into three key stages of the network (corresponding to the P3, P4, and P5 feature levels). Under identical training settings, both uniform *den* configurations and representative hierarchical scheduling strategies (as shown in Table 6) were systematically evaluated.

Table 6. Sensitivity analysis of the density parameter (*den*) configuration.

Type	den Schedule (den_1–den_3)	P (%)	R (%)	F1-Score	mAP@50 (%)	mAP@75 (%)	GFLOPs	Params
Std. Conv	(0, 0, 0)	85.9	79.0	79.8	84.0	76.1	5.8	2.5 M
Uniform	(0.9, 0.9, 0.9)	**88.6**	78.6	**80.5**	**85.8**	78.2	5.5	2.5 M
Uniform	(1.0, 1.0, 1.0)	85.6	**79.7**	79.6	84.2	77.5	5.5	2.5 M
Uniform	(1.1, 1.1, 1.1)	80.5	80.8	78.6	82.3	76.0	5.5	2.5 M
Schedule	(0.9, 0.9, 1.0)	87.0	77.1	78.3	82.6	77.0	5.5	2.5 M
Schedule	(0.9, 0.9, 1.1)	87.6	78.6	80.1	84.7	78.1	5.5	2.5 M
Schedule	(0.9, 1.0, 1.0)	85.8	**79.7**	80.3	85.4	77.5	5.5	2.5 M
Schedule	(0.9, 1.0, 1.1)	81.2	81.9	80.0	83.7	77.6	5.5	2.5 M
Schedule	(0.9, 1.1, 1.1)	84.1	77.6	77.9	81.6	73.8	5.5	2.5 M
Schedule	(1.0, 1.0, 1.1)	81.5	80.8	79.6	85.0	**79.6**	5.5	2.5 M
Schedule	(1.0, 1.1, 1.1)	86.5	78.1	80.2	83.0	75.0	5.5	2.5 M

Bold values indicate the best result in each column.

4.5.1. Configuration Strategies of Density Priors

To establish a scientifically grounded and robust baseline for the density parameter *den*, we adopted the empirical configuration space provided in the official implementation [49] of Cammarasana et al., selecting $den \in \{0.9, 1.0, 1.1\}$ as core evaluation values. These values correspond to three distinct spatial priors: center-focused, uniform (standard convolution), and edge-enhanced distributions. Specifically, when $den = 1.0$, the prior matrix Φ degenerates into an all-one matrix, which is equivalent to standard convolution with uniform aggregation. When $den = 0.9$ (<1.0), Φ exhibits a center-focused attenuation pattern, functioning as a low-pass filter that suppresses background noise. Conversely, when $den = 1.1$ (>1.0), the edge weights of the convolution kernel are amplified, leading to high-frequency enhancement effects. Based on these baseline values, we further designed two parameter configuration strategies: a uniform strategy and a hierarchical scheduling strategy. The uniform strategy evaluates the global consistency of spatial priors across all feature levels, while the hierarchical strategy assumes that different layers may require distinct feature extraction behaviors as the receptive field increases with network depth.

4.5.2. Results and Mechanism Discussion

As shown in Table 6, the configuration Uniform (0.9, 0.9, 0.9) achieves the best overall performance, with mAP@50 reaching 85.8%, significantly outperforming both the uniform baseline ($den = 1.0$, 84.2%) and the edge-enhanced strategy ($den = 1.1$, 82.3%). This result demonstrates that, for the dense high-frequency grinding textures on magnetic sheet surfaces, the center-focused soft attention induced by $den = 0.9$ effectively suppresses background interference and enhances sensitivity to subtle defects. In contrast, $den = 1.1$ tends to amplify irrelevant edge textures, leading to performance degradation.

Moreover, all hierarchical scheduling strategies that attempt to assign different filtering behaviors at different network depths fail to outperform the uniform configuration. This can be attributed to the cross-scale homogeneity of background noise in industrial magnetic sheet images. Specifically, relaxing the filtering constraint in deeper layers ($den > 1.0$) may inadvertently reintroduce "pseudo-defects" that were not fully suppressed in earlier stages, thereby interfering with the subsequent semantic alignment process in the CSFCN module. Therefore, maintaining a globally consistent configuration with $den = 0.9$ provides the most effective solution for suppressing background noise and ensuring stable feature representation across scales. Accordingly, in this study, *den* is regarded as a task-specific hyperparameter rather than a universal constant, and its final setting is determined by combining prior empirical guidance with dataset-specific sensitivity analysis.

4.6. Comparative Experiments

To assess the effectiveness and deployment suitability of FMSD-YOLO12, we compared it with several mainstream detection algorithms. In addition to the Transformer-based Real-Time Detection Transformer (RT-DETR) and representative convolutional neural network (CNN)-based YOLO series models, two lightweight baselines, MobileNetV4-YOLO11 and YOLOX-Nano, were also included for a more comprehensive evaluation of the accuracy–efficiency trade-off. All models were trained and evaluated under identical dataset splits and hyperparameter settings. The quantitative results are summarized in Table 7.

Overall, FMSD-YOLO12 achieves a favorable balance between detection accuracy and lightweight deployment on the self-constructed ferrite magnetic sheet dataset. The proposed model attains the best results on several key metrics, with mAP@50 reaching 86.4% and more stringent mAP@75 achieving 79.8%. Notably, compared with the baseline YOLOv12n, FMSD-YOLO12 improves mAP@75 by 3.7% (from 76.1% to 79.8%). This

significant improvement indicates that the proposed model not only detects defects effectively but also generates bounding boxes that more precisely align with defect boundaries, which can be attributed to the CSFCN module and the DEConv operator. Compared with the parameter-heavy RT-DETR (which achieves a high Precision of 90.1% but a lower Recall of 81.5%), the proposed model demonstrates competitive robustness when handling cross-scale irregular defects while only requiring about 13% of its parameters (2.6 M vs. 19.9 M).

Table 7. Performance comparison of FMSD-YOLO12 with mainstream and lightweight detection algorithms.

Algorithm	P (%)	R (%)	F1-Score	mAP@50 (%)	mAP@75 (%)	mAP@50-95 (%)	FPS	Params
RT-DETR	**90.1**	81.5	**84.6**	83.1	79.0	70.5	116.1	19.9 M
MobileNetV4-YOLO11	88.4	77.2	78.8	83.4	76.7	67.6	290.6	2.7 M
YOLOX-Nano	86.2	76.0	77.5	81.5	73.2	64.9	63	**0.9 M**
YOLOv5n	86.4	78.2	79.6	84.3	77.7	69.3	331.5	2.2 M
YOLOv6n	87.3	77.6	79.6	83.8	78.2	69.5	**345.2**	4.2 M
YOLOv8n	84.1	79.4	79.0	83.7	76.8	69.3	341.5	2.7 M
YOLOv9t	87.5	78.0	80.0	85.7	79.2	**71.0**	262.9	1.7 M
YOLOv10n	84.0	77.1	77.3	82.5	76.1	68.1	307.7	2.7 M
YOLOv11n	86.8	79.2	79.8	84.6	78.5	69.9	262.2	2.6 M
YOLOv12n	85.9	79.0	79.8	84.0	76.1	67.9	331.7	2.5 M
Ours	85.8	**81.9**	81.8	**86.4**	**79.8**	**71.0**	197.7	2.6 M

Bold values indicate the best result in each column.

Compared with the lightweight baselines, FMSD-YOLO12 provides a more practical balance between detection accuracy and deployment efficiency for industrial inspection. While YOLOX-Nano possesses the smallest parameter size (0.9 M), its detection accuracy decreases noticeably, with mAP@50 and mAP@75 falling to 81.5% and 73.2%, respectively. Similarly, MobileNetV4-YOLO11 achieves higher FPS (290.6) but exhibits a substantial decrease in Recall (77.2%) and mAP@50-95 (67.6%). These results suggest that extremely lightweight designs can improve compactness, but may also weaken feature representation capability in complex industrial defect scenarios. In contrast, FMSD-YOLO12 maintains competitive efficiency (2.6M parameters) while delivering stronger overall detection performance, indicating that it achieves a favorable balance between computational cost and accuracy within the evaluated real-time industrial defect detection setting.

To further illustrate the detection characteristics of the proposed model in comparison with representative baselines, several representative detection results are visualized in Figure 11.

In terms of stable extraction of subtle features, as shown in the first column (LM defects) and the upper-right region of the third column (Smallbreak defects), all YOLO-based models are able to detect the targets, but the proposed model achieves higher confidence scores (both 0.93), significantly outperforming the baseline YOLOv12n (0.83 and 0.86) and other methods. This demonstrates that, with the assistance of the CFC mechanism, the model effectively prevents subtle defect features from being overwhelmed by complex background semantics, resulting in improved detection stability and accuracy. Regarding robustness against strong texture interference and multi-scale detection (as shown in the second column), the sample is dominated by high-frequency grinding textures and contains both small "Dots" and irregular "Smallbreaks". The baseline YOLOv12n, as well as YOLOv11n and YOLOv8n, exhibit noticeable missed detections, while YOLOv9t produces low-confidence predictions. In contrast, the proposed model accurately detects all targets with high confidence, demonstrating that the SR-C3k2 module effectively suppresses texture noise and enhances multi-scale feature discrimination. For precise regression of

irregular boundaries (as shown in the fourth column), conventional YOLO models tend to produce loose and redundant bounding boxes for highly irregular edge wear (XB). In contrast, driven by the differential edge enhancement of DEConv, the proposed model generates compact and well-aligned bounding boxes that closely match the defect contours. This observation provides intuitive visual evidence for the significant improvement in mAP@75 reported in the quantitative results.

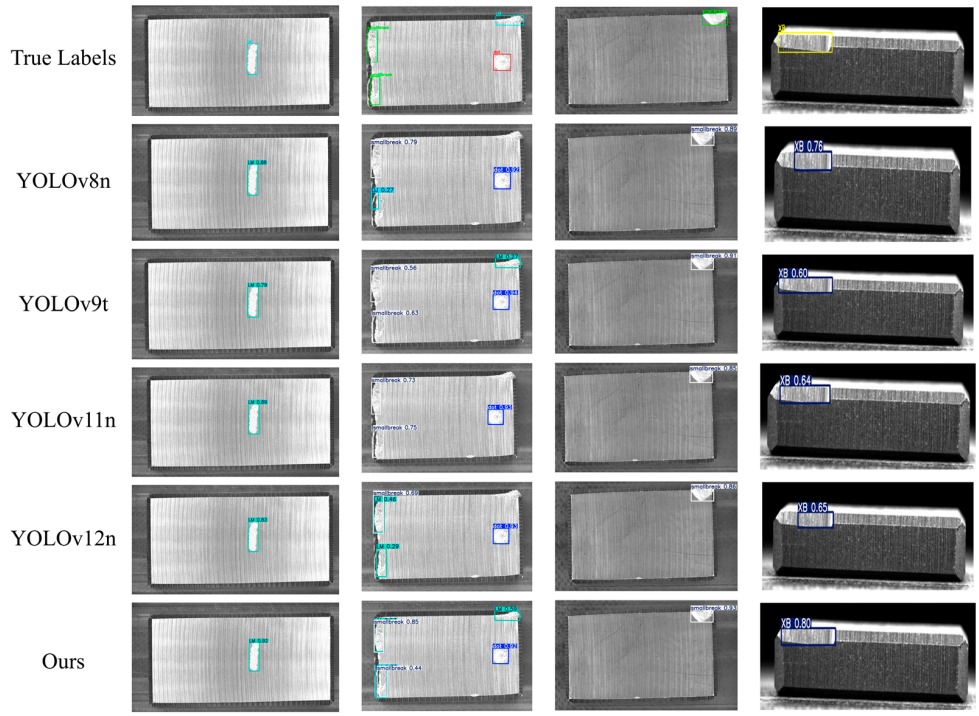


Figure 11. Visual detection examples of FMSD-YOLO12 and representative YOLO-series models. (Note: The colored bounding boxes denote different defect categories, where cyan, light blue, white, and dark blue correspond to LM, dot, smallbreak, and XB, respectively).

Despite the strong performance of FMSD-YOLO12 across various scenarios, some limitations remain. Compared with Transformer-based models such as RT-DETR, which are effective at modeling complex global dependencies, the proposed model may encounter difficulties in certain extreme cases. As illustrated in Figure 12, while FMSD-YOLO12 achieves comparable localization performance in standard single-object scenarios (columns 1 and 2), it struggles with extremely small defects under severe texture interference (e.g., the tiny smallbreak in column 3) and defects with extremely weak edge features (e.g., edge wear in column 4), where RT-DETR can better capture such patterns due to its global attention mechanism.

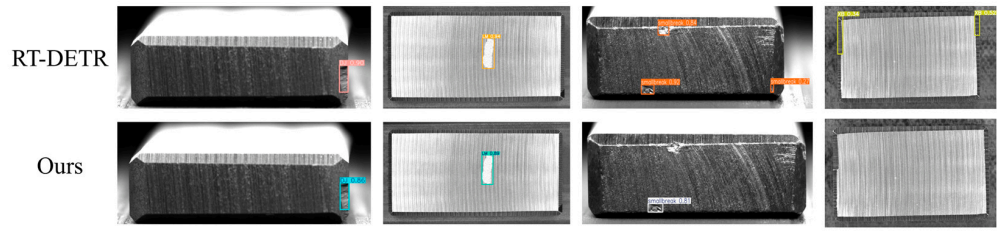


Figure 12. Visual detection examples of FMSD-YOLO12 and RT-DETR. (Note: The colored bounding boxes indicate predicted defect categories. In the results of Ours, cyan, light blue, white, and dark blue boxes correspond to LM, dot, smallbreak, and XB, respectively. For RT-DETR, the box colors follow its own visualization settings, and the specific category is indicated by the label attached to each bounding box).

This observation also explains the higher Precision (90.1%) achieved by RT-DETR in Table 7. However, considering its significantly larger parameter size (19.9 M), FMSD-YOLO12 (2.6 M) achieves a more favorable trade-off by sacrificing only a small portion of performance in extreme cases while maintaining lightweight design and high inference speed. This balance makes it a more practical solution for real-world industrial scenarios that require low-cost and high-throughput deployment.

4.7. Cross-Domain Generalization on NEU-DET

To examine the cross-domain applicability of FMSD-YOLO12 in a different industrial inspection scenario, we evaluated the model on the public NEU-DET steel surface defect dataset [50]. As shown in Figure 13a–f, NEU-DET includes six defect categories from hot-rolled steel strips: crazing, inclusion, patches, pitted surface, rolled-in scale, and scratches. These defects share several challenges with ferrite magnetic sheet defects, such as complex background textures and large scale differences. The experimental results are presented in Table 8.

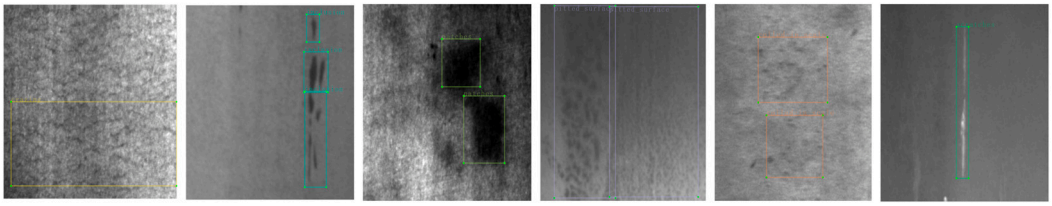


Figure 13. Six typical surface defect categories in the NEU-DET dataset.

Table 8. Performance evaluation of FMSD-YOLO12 against mainstream and lightweight detection algorithms on the NEU-DET dataset.

Algorithm	P (%)	R (%)	F1-Score	mAP@50 (%)	mAP@75 (%)	mAP@50-95 (%)	FPS	Params
RT-DETR	69.3	79.7	73.5	77.4	48.2	44.3	43.9	19.9 M
MobileNetV4-YOLO11	69.3	82.7	74.0	82.1	56.5	49.5	42.2	2.7 M
YOLOX-Nano	83.2	70.6	76.4	77.8	48.3	42.0	59.1	0.9 M
YOLOv5n	69.2	78.5	73.1	80.0	46.0	45.6	50.2	2.2 M
YOLOv6n	62.4	72.6	66.2	76.2	45.0	44.3	42.2	4.2 M
YOLOv8n	79.3	72.9	73.5	80.0	58.2	49.5	42.7	2.7 M
YOLOv9t	84.2	77.7	77.9	85.4	56.5	54.0	43.0	1.7 M
YOLOv10n	62.6	77.2	67.6	74.7	55.5	46.8	43.0	2.7 M
YOLOv11n	72.5	72.9	71.0	78.2	49.3	47.0	41.3	2.6 M
YOLOv12n	81.2	75.9	77.3	83.0	50.9	49.6	39.3	2.5 M
Ours	86.2	73.4	77.1	83.9	49.0	47.8	37.6	2.6 M

The comparison results demonstrate that FMSD-YOLO12 maintains strong competitiveness even without domain-specific prior adaptation to a new industrial material. Specifically, the proposed model achieves the highest Precision (86.2%) among all evaluated methods, outperforming the baseline YOLOv12n (81.2%) and YOLOv9t (84.2%). Furthermore, compared to ultra-lightweight designs, FMSD-YOLO12 shows stronger false-positive suppression capability. For example, MobileNetV4-YOLO11 yields a lower Precision of 69.3%, indicating reduced robustness in distinguishing defects from complex steel textures. This suggests that the combined feature extraction and calibration mechanisms contribute to more robust suppression of "pseudo-defects" under complex industrial backgrounds. In terms of overall detection accuracy, FMSD-YOLO12 achieves an mAP@50 of 83.9%, improving upon the baseline YOLOv12n (83.0%) by 0.9%. These results further validate the robustness and cross-domain applicability of the proposed model.

However, we also observe that FMSD-YOLO12 achieves lower Recall (73.4%) and mAP@75 (49.0%) on the NEU-DET dataset compared with generalization-oriented models such as YOLOv9t. This performance gap may be related to the task-oriented center-focused prior adopted in SR-Conv, whose optimal *den* setting was selected for ferrite magnetic sheet images with dense grinding textures. When directly transferred to a new domain without further re-calibration, the same prior may suppress diffuse or topology-variant defect edges more aggressively than desired, resulting in a slight reduction in Recall and high-IoU localization accuracy. These observations suggest that the task-oriented prior in SR-Conv is still partially scenario-dependent. When extending FMSD-YOLO12 to industrial scenes with substantially different surface textures, illumination conditions, or defect distributions, dataset-specific recalibration or adaptive prior learning may further enhance robustness. Such adaptation is also important under limited-data conditions, where the optimal prior configuration may be difficult to validate reliably.

Overall, although FMSD-YOLO12 is specifically optimized for ferrite magnetic sheet surface characteristics, the generalization experiments demonstrate that its core feature extraction and context-aware architecture still retain strong false-positive suppression. On the steel defect dataset, the proposed model not only maintains competitive real-time performance but also achieves the highest Precision (86.2%) among all compared methods, highlighting its robustness and practical potential for industrial surface defect inspection tasks. In future work, we will further investigate adaptive or learnable density priors in SR-Conv, together with more data-efficient training strategies, to improve robustness under varying texture, illumination, and limited-data conditions.

5. Conclusions

To address the major difficulties in ferrite magnetic sheet inspection, including severe grinding-texture disturbance, large defect-scale variation, and weak boundary perception, this study develops FMSD-YOLO12, an efficient defect inspection framework built upon YOLOv12. The proposed model introduces three targeted modifications to improve defect representation and deployment feasibility in industrial environments. First, the SR-C3k2 module is embedded into the backbone with a spatial density prior, enabling the model to reduce repetitive texture responses and strengthen the representation of subtle low-contrast defects. Second, the Context and Spatial Feature Calibration Network (CSFCN), consisting of Context Feature Calibration (CFC) and semantic-flow-guided Spatial Feature Calibration (SFC), is designed to alleviate spatial misalignment during multi-level feature fusion. Finally, the Lightweight Shared Detail-Enhanced Convolutional Detection (LSDECD) head combines cross-scale weight sharing and re-parameterized Detail-Enhanced Convolution (DEConv), improving the localization of weak and irregular boundaries without increasing deployment-stage computation.

Extensive experiments verify the effectiveness of FMSD-YOLO12. On the self-constructed ferrite magnetic sheet defect dataset, the proposed model achieves 86.4% mAP@50, 79.8% mAP@75, and 81.9% Recall. Additionally, it preserves a compact parameter size and reaches 197.7 FPS in inference-only speed, indicating a favorable trade-off between detection accuracy and processing speed. In addition, the cross-dataset evaluation on NEU-DET suggests a certain degree of cross-domain robustness, particularly in false-positive suppression, achieving the highest Precision of 86.2% among all compared methods.

Overall, within the scope of the evaluated industrial defect detection tasks, FMSD-YOLO12 achieves a favorable compromise among detection accuracy, localization quality, and model efficiency. Compared with extremely lightweight baseline models, it retains stronger recognition capability while keeping a compact model scale and real-time infer-

ence. Therefore, the proposed method offers a feasible approach for automated industrial inspection on edge-side devices, especially in scenarios involving subtle defects, irregular contours, and complex surface textures. For future work, we will explore more lightweight and efficient linear attention mechanisms to further improve robustness under challenging conditions. We also plan to deploy FMSD-YOLO12 on embedded edge devices using optimized inference engines and conduct long-term closed-loop validation in real-world production lines, thereby promoting its practical industrial application.

Author Contributions: Algorithm development and overall project direction were jointly managed by C.Z. and J.L. Software coding and subsequent simulation tests were executed by C.Z., H.Y. and R.W. participated in the establishment of the theoretical model and data interpretation. C.Z. was responsible for drafting the initial manuscript, which was then comprehensively reviewed and refined by J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant No. 62573387, the Key R&D Program of Zhejiang (No. 2023C01062), Basic Public Welfare Research Program of Zhejiang Province (No. LGF22F030001) and the Fundamental Research Funds of Zhejiang Sci-Tech University (No. 24222091-Y).

Data Availability Statement: The dataset used in this study consists of two parts. The first part is a self-constructed dataset collected from real industrial production environments. Due to confidentiality and proprietary restrictions, this dataset is not publicly available. The second part is the publicly available NEU-DET dataset, which can be accessed online. The data supporting the findings of this study, including preprocessing and annotation details, are available from the corresponding author upon reasonable request.

Conflicts of Interest: Author J.L. was employed by Changshan Research Institute Co., Ltd. of Zhejiang Sci-Tech University. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ABlock	Area-Attention Block
ADC	Angular Difference Convolution
AP	Average Precision
ARA-Net	Adaptive Rotation Attention Network
ARC	Adaptive Rotated Convolution
ASPP	Atrous Spatial Pyramid Pooling
BCE	Binary Cross-Entropy
BiFPN	Bidirectional Feature Pyramid Network
BN	Batch Normalization
CDC	Central Difference Convolution
CFC	Context Feature Calibration
CNN	Convolutional Neural Network
CSFCN	Context and Spatial Feature Calibration Network
CSP	Cross-Stage Partial
CSPNet	Cross-Stage Partial Network
DCNv4	Deformable Convolution v4
DEConv	Detail-Enhanced Convolution
DFL	Distribution Focal Loss
DSConv	Depthwise Separable Convolution
F1-Score	F1-Score
FMSD-YOLO12	Ferrite Magnetic Sheet Defect Detection YOLOv12

FN	False Negative
FP	False Positive
FPN	Feature Pyramid Network
FPS	Frames Per Second
GFLOPs	Giga Floating-Point Operations
GN	Group Normalization
HDC	Horizontal Difference Convolution
HOG	Histogram of Oriented Gradients
IoU	Intersection over Union
LSDECD	Lightweight Shared Detail-Enhanced Convolutional Detection
mAP	Mean Average Precision
mAP@50	Mean Average Precision at IoU = 0.50
mAP@50-95	Mean Average Precision averaged over IoU thresholds from 0.50 to 0.95
NMS	Non-Maximum Suppression
P	Precision
PANet	Path Aggregation Network
Params	Number of Parameters in the Model
PSP	Pyramid Scene Parsing
R	Recall
R-CNN	Region-based Convolutional Neural Network
RPA	Residual Parallel Attention
RT-DETR	Real-Time Detection Transformer
SFC	Spatial Feature Calibration
SIFT	Scale-Invariant Feature Transform
SR-Conv	Spatially Re-weighted Convolution
SR-C3k2	Spatially Re-weighted C3k2 Module
SVM	Support Vector Machine
TP	True Positive
VDC	Vertical Difference Convolution
YOLO	You Only Look Once

References

1. Tahanian, H.; Aliahmadi, M.; Faiz, J. Ferrite permanent magnets in electrical machines: Opportunities and challenges of a non-rare-earth alternative. *IEEE Trans. Magn.* **2020**, *56*, 900120. [CrossRef]
2. Du, Z.S.; Lipo, T.A. Cost-effective high torque density bi-magnet machines utilizing rare earth and ferrite permanent magnets. *IEEE Trans. Energy Convers.* **2020**, *35*, 1577–1584. [CrossRef]
3. E. Farag, H.; Khamesee, M.B.; Toyserkani, E. Eddy current sensor probe design for subsurface defect detection in additive manufacturing. *Sensors* **2024**, *24*, 5355. [CrossRef] [PubMed]
4. Zhang, J.; Wan, H.; Sun, F.; Chen, X.; Jia, K.; Fan, L.; Cheng, L.; Xu, X.; Yan, X.; Yuan, P.; et al. Laser ultrasonic inspection of surface and subsurface defects in materials based on deep learning. *Nondestruct. Test. Eval.* **2024**, *40*, 4744–4763. [CrossRef]
5. Li, L.; Liu, Z.; Zhao, H.; Xue, L.; Wu, J. A method for detecting bearing surface defects combining magnetic particle testing and deep learning. *Appl. Sci.* **2024**, *14*, 1747. [CrossRef]
6. Fang, X.; Luo, Q.; Zhou, B.; Li, C.; Tian, L. Research progress of automated visual surface defect detection for industrial metal planar materials. *Sensors* **2020**, *20*, 5136. [CrossRef]
7. Wu, Z.; Jiang, H.; Liu, S.; Liu, Y.; Yang, W. Conditional distribution-guided adversarial transfer learning network with multi-source domains for rolling bearing fault diagnosis. *Adv. Eng. Inform.* **2023**, *56*, 101993. [CrossRef]
8. Wang, X.; Jiang, H.; Dong, Y.; Mu, M. Spatial-channel collaborative multi-scale graph interaction deep transfer learning for unsupervised rotating machinery fault diagnosis. *Eng. Appl. Artif. Intell.* **2026**, *176*, 114691. [CrossRef]
9. Wang, X.; Jiang, H.; Zeng, T.; Dong, Y. An adaptive fused domain-cycling variational generative adversarial network for machine fault diagnosis under data scarcity. *Inf. Fusion* **2025**, *126*, 103616. [CrossRef]
10. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 16965–16974.

11. Ramos, L.T.; Sappa, A.D. A decade of You Only Look Once (YOLO) for object detection: A review. *IEEE Access* **2025**, *13*, 192747–192794. [CrossRef]
12. Hussain, M. YOLO-v1 to YOLO-v8: The rise of YOLO and its complementary role in digital manufacturing and industrial defect detection. *Machines* **2023**, *11*, 677. [CrossRef]
13. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
14. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-centric real-time object detectors. *arXiv* **2025**, arXiv:2502.12524.
15. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
16. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005*; IEEE: Piscataway, NJ, USA, 2005; pp. 886–893.
17. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014*; IEEE: Piscataway, NJ, USA, 2014; pp. 580–587.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9. [CrossRef]
19. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 6517–6525.
20. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. 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 (CVPRW), Seattle, WA, USA, 14–19 June 2020*; IEEE: Piscataway, NJ, USA, 2020.
21. Ultralytics. YOLOv8. Available online: https://docs.ultralytics.com/models/yolov8/ (accessed on 10 March 2025).
22. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning what you want to learn using programmable gradient information. In *Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1–21.
23. Ultralytics. YOLOv11. Available online: https://docs.ultralytics.com/models/yolo11/ (accessed on 10 March 2025).
24. Qin, D.; Leichner, C.; Delakis, M.; Fornoni, M.; Luo, S.; Yang, F.; Wang, W.; Banbury, C.; Ye, C.; Akin, B.; et al. MobileNetV4: Universal models for the mobile ecosystem. In *Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 78–96.
25. Megvii-BaseDetection. YOLOX: High-Performance Anchor-Free YOLO Detector. GitHub Repository. Available online: https://github.com/Megvii-BaseDetection/YOLOX (accessed on 22 April 2026).
26. Ren, Z.; Fang, F.; Yan, N.; Wu, Y. State of the art in defect detection based on machine vision. *Int. J. Precis. Eng. Manuf. Green Technol.* **2022**, *9*, 661–691. [CrossRef]
27. Huang, Y.; Huang, Z.; Jin, T. DEU-Net: A multi-scale fusion staged network for magnetic tile defect detection. *Appl. Sci.* **2024**, *14*, 4724. [CrossRef]
28. Prunella, M.; Scardigno, R.M.; Buongiorno, D.; Brunetti, A.; Longo, N.; Carli, R.; Dotoli, M.; Bevilacqua, V. Deep learning for automatic vision-based recognition of industrial surface defects: A survey. *IEEE Access* **2023**, *11*, 43370–43423. [CrossRef]
29. Luo, F.; Cui, Y.; Wang, X.; Zhang, Z.; Liao, Y. Adaptive rotation attention network for accurate defect detection on magnetic tile surface. *Math. Biosci. Eng.* **2023**, *20*, 18585–18607. [CrossRef] [PubMed]
30. Yu, B.; Li, Q.; Jiao, W.; Zhang, S.; Zhu, Y. SAB-YOLOv5: An improved YOLOv5 model for permanent magnetic ferrite magnet rotor detection. *Mathematics* **2024**, *12*, 957. [CrossRef]
31. Niu, W.; Lv, C.; Zhang, E.; Wei, Z. YOLO-RDM: A high-accuracy and efficient algorithm for magnetic tile surface defect detection with practical applications. *PLoS ONE* **2025**, *20*, e0328815. [CrossRef]
32. Ma, C.; Pan, Y.; Chen, J. Surface defect detection of magnetic tiles based on YOLOv8-AHF. *Electronics* **2025**, *14*, 2857. [CrossRef]
33. Ma, S.; Zhao, X.; Wan, L.; Zhang, Y.; Gao, H. A lightweight algorithm for steel surface defect detection using improved YOLOv8. *Sci. Rep.* **2025**, *15*, 8966. [CrossRef]
34. Wang, Y.; Sun, H.; Luo, K.; Zhu, Q.; Li, H.; Sun, Y.; Wu, Z.; Wang, G. A lightweight YOLOv11-based framework for small steel defect detection with a newly enhanced feature fusion module. *Sci. Rep.* **2025**, *15*, 34322. [CrossRef]
35. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 2117–2125.
36. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 8759–8768.

37. Li, M.; Wu, J.; Wang, X.; Chen, C.; Qin, J.; Xiao, X.; Wang, R.; Zheng, M.; Pan, X. AlignDet: Aligning pre-training and fine-tuning in object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 6866–6876.
38. Li, X.; You, A.; Zhu, Z.; Zhao, H.; Yang, M.; Yang, K.; Tan, S.; Tong, Y. Semantic flow for fast and accurate scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 775–793.
39. Xiong, Y.; Li, Z.; Chen, Y.; Wang, F.; Zhu, X.; Luo, J.; Wang, W.; Lu, T.; Li, H.; Qiao, Y.; et al. Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 5652–5661.
40. Tang, H.; Li, Z.; Zhang, D.; He, S.; Tang, J. Divide-and-Conquer: Confluent Triple-Flow Network for RGB-T Salient Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 1958–1974. [CrossRef] [PubMed]
41. Li, K.; Geng, Q.; Wan, M.; Cao, X.; Zhou, Z. Context and spatial feature calibration for real-time semantic segmentation. *IEEE Trans. Image Process.* **2023**, *32*, 5465–5477. [CrossRef]
42. Yu, Z.; Zhao, C.; Wang, Z.; Qin, Y.; Su, Z.; Li, X.; Zhou, F.; Zhao, G. Searching central difference convolutional networks for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 5295–5305.
43. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. RepVGG: Making VGG-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 13733–13742.
44. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 2980–2988.
45. Cammarasana, S.; Patanè, G. Optimal density functions for weighted convolution in learning models. *arXiv* **2025**, arXiv:2505.24527. [CrossRef]
46. Cammarasana, S.; Patanè, G. Optimal weighted convolution for classification and denoising. *arXiv* **2025**, arXiv:2505.24558.
47. Liu, Y.; Shi, Y. VRU-YOLO: A small object detection algorithm for vulnerable road users in complex scenes. *IEEE Access* **2025**, *13*, 16584–16599. [CrossRef]
48. Chen, Z.; He, Z.; Lu, Z.M. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Trans. Image Process.* **2024**, *33*, 1002–1015. [CrossRef] [PubMed]
49. Cammarasana, S. WeightedConvolution2.0. Available online: https://github.com/cammarasana123/weightedConvolution2.0 (accessed on 25 March 2026).
50. He, Y.; Song, K.; Meng, Q.; Yan, Y. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 1493–1504. [CrossRef]

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