OEE | 光电工程 Opto-Electronic Engineering

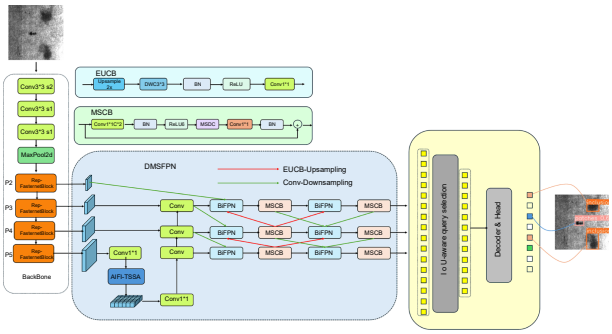
Article

2026 年 1 月, 第 53 卷, 第 1 期

RMD-DETR:一种改进的钢材表面缺陷检测算法

蒋宏捷[1,2], 肖小玲[1,2*], 颜　昕[3]

[1]长江大学计算机科学学院, 湖北 荆州 434023;
[2]长江大学人工智能科研平台, 湖北 荆州 434023;
[3]武汉理工大学计算机与人工智能学院, 湖北 武汉 430070



摘要: 针对钢材表面缺陷检测中存在的检测精度低、跨尺度适应性差等问题, 本文提出一种基于改进 RT-DETR 的检测模型 RMD-DETR。首先, 引入轻量级 FasterNet 作为模型骨干网络, 同时对 FasterNetBlock 中的 PConv 进行重参数化改进, 在保持轻量化的同时增强特征表达能力, 降低计算复杂度; 其次, 设计了动态多尺度特征融合网络 (DMSFPN), 结合多尺度卷积块 (MSCB) 与高效上采样模块 (EUCB), 增强跨分辨率特征的语义一致性; 最后引入基于变分率降噪的 TSSA 机制, 进一步提出 AIFI-TSSA 模块, 提升复杂背景下的检测稳定性。实验结果表明, 在 NEU-DET 数据集上, 改进模型的 mAP@0.5 达到 76.7%, 较基准 RT-DETR 提升了 2.6%。在跨尺度检测中, 模型漏检率从基准的 55.4% 降至 50.9%, 降低了 4.5%, 同时参数量与计算量分别减少了 39.9% 和 39.3%, 帧率 (FPS) 提升了 30.4% (78.9 f/s), 满足工业产线大于 30 f/s 的实时检测需求。在天池铝型材数据集和 GC10-DET 数据集上的泛化实验中, mAP@0.5 分别提升了 2% 和 3%, 验证了模型在工业场景中的高效性与鲁棒性。

关键词: 表面缺陷检测; 重参数化; DMSFPN; AIFI-TSSA; 轻量化模型

The English title and abstract appear at the end of the paper.

DOI: 10.12086/oee.2026.250181 | CSTR: 32245.14.oee.2026.250181 | 中图分类号: TP391.4 | 文献标识码: A

蒋宏捷, 肖小玲, 颜昕. RMD-DETR:一种改进的钢材表面缺陷检测算法 [J]. 光电工程, 2026, **53**(1): 250181

Jiang H J, Xiao X L, Yan X. RMD-DETR: an improved steel surface defect detection algorithm[J]. *Opto-Electron Eng*, 2026, **53**(1): 250181

1　引　言

钢材作为国家基建的核心材料, 其质量直接影响建筑结构安全。受生产工艺与技术限制, 钢材生产过程中易出现轧制氧化皮、开裂、划痕及点蚀等表面缺陷, 导致抗腐蚀性、耐磨性和抗拉伸性能显著降低, 威胁工程可靠性。发展高效准确的表面缺陷检测方法, 对于确保产品质量、提高生产效率和降低成本具有重要意义。

近年来, 基于深度学习的检测方法凭借其高精度与强泛化能力, 逐步替代传统算法成为钢材缺陷检测的主流技术。当前主流模型主要分为两类: 以 Fast R-CNN[1]、Faster R-CNN[2] 为代表的二阶段算法, 通过选择性搜索生成缺陷候选区域进行定位分类; 以及以 SSD[3]、YOLO[4] 系列为核心的一阶段算法, 通过端到端特征提取实现实时检测。尽管上述目标检测模型在诸多任务中表现出色, 但在针对钢材表面微小缺陷 (如微小划痕、细微开裂等) 的检测上仍存在明显局限。这是由于这类缺陷通常具有较小尺寸、低对比度以及特征模糊等特性, 其检测难度远高于宏观缺陷。若此类微小缺陷未被有效识别, 可能在后续加工过程中快速扩展, 进而引发

收稿日期: 2025-06-22; 修回日期: 2025-09-23; 录用日期: 2025-09-29; 网络出版日期: 2026-01-25
基金项目: 国家自然科学基金资助项目 (61771354)
*通信作者: 肖小玲, xxl@yangtzeu.edu.cn
版权所有©2026 中国科学院光电技术研究所

结构失效。二阶段算法如 Faster R-CNN 在 NEU-DET 数据集上仅达 71.9% mAP@0.5，且参数量高达 41.37 M，难以满足工业实时性需求。一阶段算法中性能最优的 YOLOv8m 的 mAP@0.5 为 75.1%，但其 69% 的召回率表明对小尺度缺陷漏检率较高，且 78.7 GFLOPs 的计算成本仍制约边缘部署。针对工业场景，研究者通过优化网络结构及特征融合机制，有效克服钢材表面复杂背景干扰，显著提升微小缺陷识别与多尺度检测稳定性。Xia 等[5] 采用可变形卷积增强骨干网络，结合平衡特征金字塔融合策略，在提升多形状、多尺寸缺陷检测精度的同时优化了推理效率。

随着单阶段检测算法性能发展，其较低的计算成本成为应用的重要考量因素，同时单阶段算法可以一次性完成目标检测和定位，只需要一次反向传播即可得到结果。梁礼明等[6] 在 YOLOv7-tiny 中融合 GAC-FPN 与 BRA 机制，通过构建渐进式特征融合路径和动态稀疏感知的注意力筛选，显著增强了钢材表面微小缺陷的检测精度与实时性。马玉洁等[7] 在 YOLOv8 中融合特征空间压缩与双域注意力机制，通过构建 FSCConv 特征压缩模块和 CSP-OK 多尺度优化架构，显著增强了钢材表面微小缺陷的跨尺度检测能力。现有方法虽在尺度不变的缺陷检测中表现良好，但对跨尺度缺陷的适应性不足，在尺度变化大于 10 倍时漏检率超 50%。此外，基于 CNN (convolutional neural networks) 的模型依赖 NMS (non-maximum supression)[8] 剔除冗余预测框，这种后处理机制会制约检测性能。

为实现端到端的检测效果，同时提高模型的检测性能，Zhao 等[9] 研发出实时目标检测模型——RT-DETR，和传统基于 CNN 架构的检测模型有所差异，相较 YOLO 系列，RT-DETR 的端到端架构具有双重优势：其一，摒弃 NMS 后处理可避免重叠缺陷框的误抑制，在密集缺陷场景显著降低漏检率；其二，Transformer 的全局注意力机制能建模长程依赖关系，有效解决钢材表面纹理干扰导致的局部特征混淆问题，而 YOLO 的 CNN 局部感受野在此类任务中存在固有局限。RT-DETR 基于 Transformer 架构，凭借强大的全局特征学习能力，可并行预测多目标类别、位置及边框，显著提升了收敛速度与推理效率，但是该模型仍具有参数量和计算量较大等不足。

针对钢材表面缺陷检测中存在的检测精度低、多尺度目标漏检率高以及模型计算复杂度大等问题，本文基于 RT-DETR 目标检测算法，提出了 RMD-DETR 改进模型，以提高检测精度、降低误判率，以及降低模型的参数量和计算量。本文的主要贡献如下:

1) 本文在骨干网络中引入 RepConv 重参数化技术对 FasterNetBlock 模块进行结构优化，在保持模型轻量级优势的同时，有效降低了计算复杂度并提升推理效率。通过训练阶段的多分支学习与推理阶段的结构简化，实现精度与速度的均衡优化。

2) 设计了 DMSFPN (dynamic multi-scale feature fusion network) 作为 RT-DETR 的颈部网络，通过引入 MSCB (multi-scale convolution block) 和 EUCB(efficient up-convolution block)，融合不同层次和分辨率的深浅层特征，提升目标检测性能。同时，以 Add 代替 Concat 操作并结合 RT-DETR 的高效注意力机制，减少参数和计算量，实现自适应加权融合，平衡了检测精度和模型复杂度。

3) 在颈部网络设计中，首先将基于变分率降噪的 TSSA(token statistics self-attention) 机制引入 AIFI 模块，构建 AIFI-TSSA 增强模块。该模块以线性复杂度实现全局上下文建模，在保留传统注意力全局依赖捕获能力的同时，显著降低计算成本，增强特征表达的多样性，进而提升复杂场景下的检测稳定性。

2 RT-DETR 模型介绍

RT-DETR 是基于 Transformer 架构的实时检测模型，通过将特征处理解耦为尺度内高维特征优化与跨尺度交互模块，显著降低计算复杂度并兼顾工业实时需求。其架构由主干网络、多尺度融合混合编码器及配备辅助检测头的解码器组成，训练中采用“IoU-aware”查询选择机制，通过联合优化初始参数与监督策略提升定位精度，结构如图 1 所示。

模型主干网络采用 ResNet-18 进行多尺度特征提取，将最后三层特征输入混合编码器。其中，AIFI 模块通过自注意力机制强化 S5 高层语义特征，CCFM 模块基于卷积实现跨尺度特征交互，最终输出融合多粒度信息的特征序列。解码器引入 IoU 作为分类约束，联合优化定位与分类任务，使检测结果的类别判断与位置匹配显著提升。

3 改进的 RMD-DETR

RMD-DETR 基于 RT-DETR 框架，面向钢材缺陷检测提出多维度优化方案：主干网络通过重参数化技术融合多分支训练与单分支推理，平衡轻量化与特征表达；

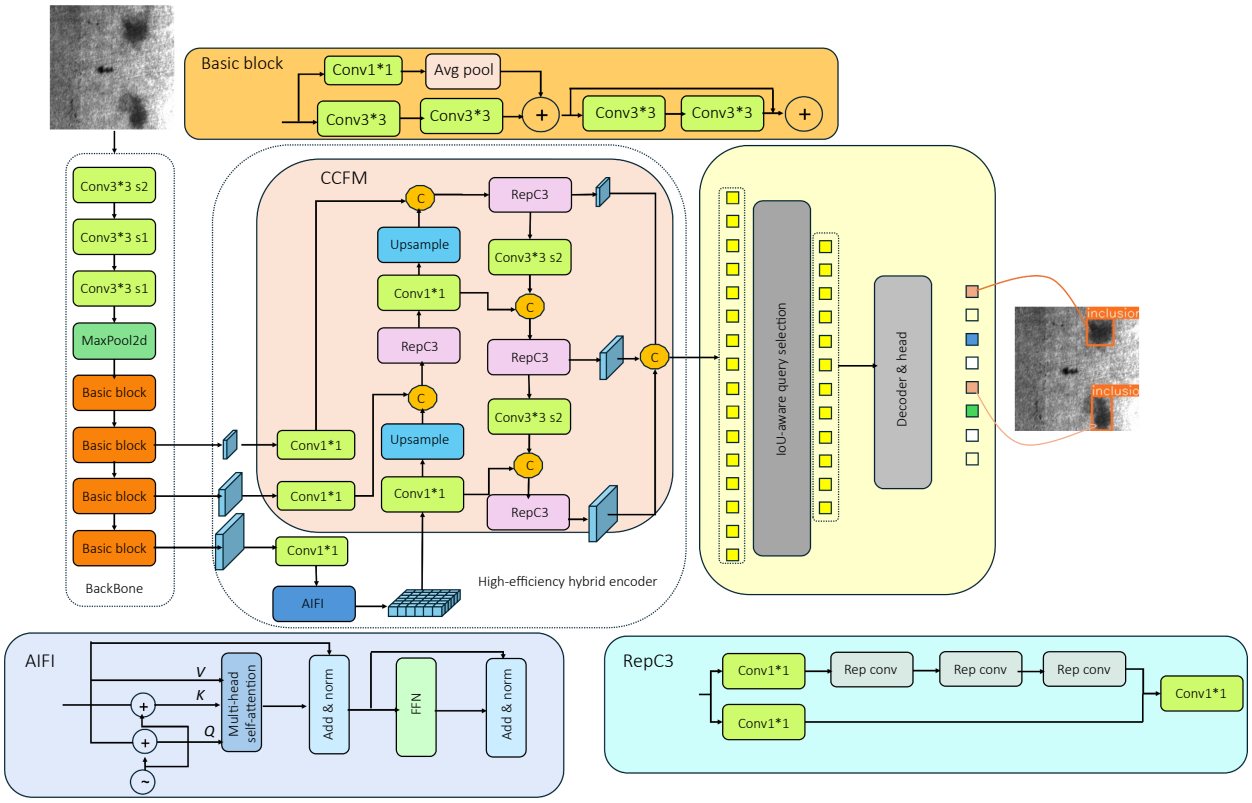


图 1 RT-DETR 网络结构图

Fig. 1 RT-DETR network architecture diagram

动态多尺度融合网络结合双向路径与可变形卷积，实现跨分辨率特征自适应对齐，提升微小缺陷定位精度；全局注意力模块嵌入统计降噪机制，以线性复杂度筛选关键语义区域，抑制背景噪声。模型采用端到端解码器直接生成检测结果，摒弃手工锚框设计，在降低计算冗余的同时，显著增强复杂工业场景下的检测稳定性与泛化能力，改进结构图如图 2 所示。

3.1 主干网络改进

RT-DETR 在进行目标检测时，使用 ResNet-18 作为特征提取网络，但是 ResNet-18 受限于浅层网络架构与固定卷积模式，在复杂目标检测任务中难以实现多尺度特征的有效融合与动态适应性表达。为了提取到更加丰富的目标特征，本文基于 FasterNet[10] 的轻量化思想，采用融合 PConv (partial convolution) 与重参数化改进 RT-DETR 骨干网络。通过多阶段堆叠设计，在 Blocks 模块中同步融合 PConv 的稀疏特征提取能力和 Rep 的动态表达能力，相较于 ResNet-18 实现了更轻量化的多尺度特征融合。

PConv 通过常规卷积操作对输入通道的子集执行空间特征提取，同时保持剩余通道的原始信息完整性，输入的特征图为$X \in \mathbb{R}^{c\times h\times w}$，其中 c 表示通道数，h 表示特征图的高，w 表示特征图的宽。

PConv 将连续的c_p个通道作为整个特征图用于计算代表。其 FLOPs 可以由式 (1) 得到。

$$f_{\mathrm{FLOPS}} = h\times w\times k^2\times c_p^2, \tag{1}$$

式中：k 代表卷积核大小。当c/c_p=1/4 时，局部卷积的 FLOPs 为普通卷积的 1/16，因此可大幅减少参数量和计算量，进一步提高模型性能。

3.1.1 基于重参数化的 Rep-FasterNetBlock

尽管 FasterNetBlock 通过 PConv 显著提升了计算效率，但其结构单一的操作符设计制约了模型在训练阶段对多尺度特征的学习能力。针对这一表征瓶颈，RepConv[11] 的结构重参数化机制在训练阶段通过多分支特征融合增强空间多样性，同时保持推理阶段的高效性，从而可以与 PConv 的局部感受野特性形成层次化特征互补。因此本文引入 RepConv (reparameterizated convolution) 对 PConv 模块进行优化构建 RPConv 模

块，进而构成重参数化模块 Rep-FasterNetBlock 如图 3 所示。

Rep-FasterNetBlock 其核心通过多分支的 3×3 标准卷积、1×1 标准卷积与批量归一化的并行设计，在训练阶段利用多路径特征融合增强模型表征能力，同时借助重新参数化技术，在推理时将多分支结构等效映射为单一卷积操作，显著降低计算复杂度，这种动态适应机制既保证了训练阶段对复杂特征的捕捉精度，又确保了实际使用时的高效推理速度。

其中 BN 层特征图第 i 个通道的计算公式如下：

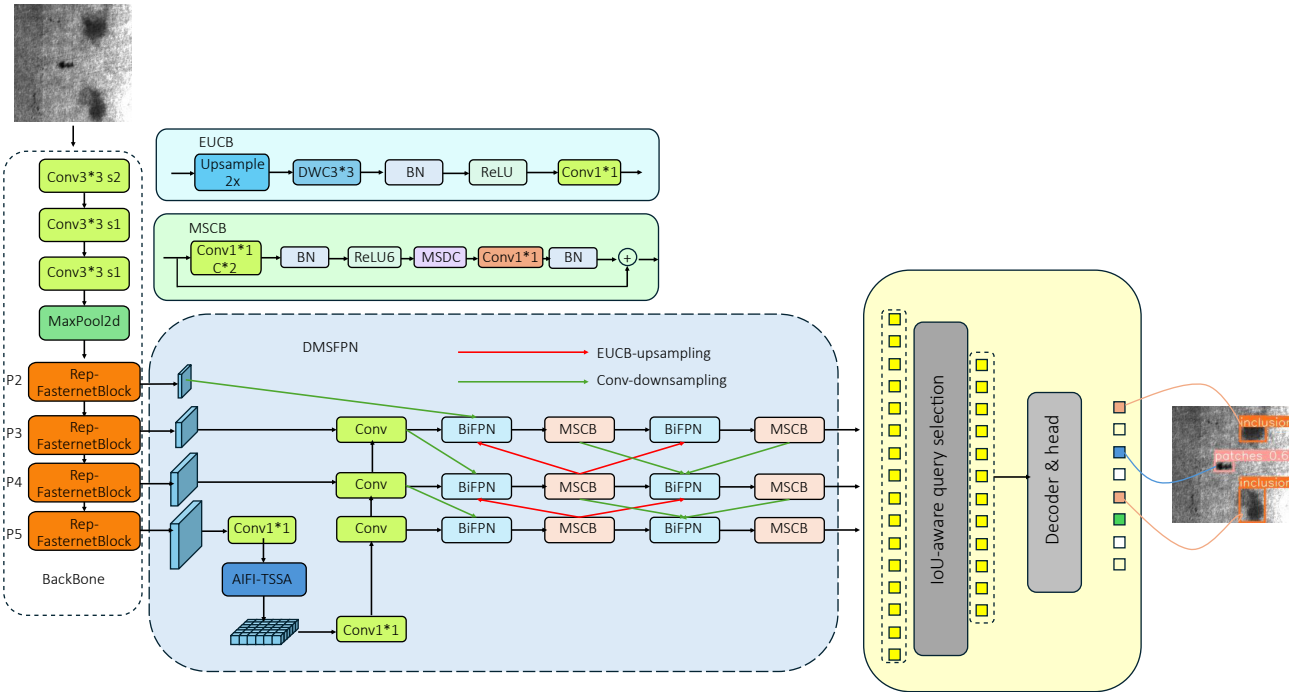


图 2 RMD-DETR 结构图

Fig. 2 RMD-DETR architecture diagram

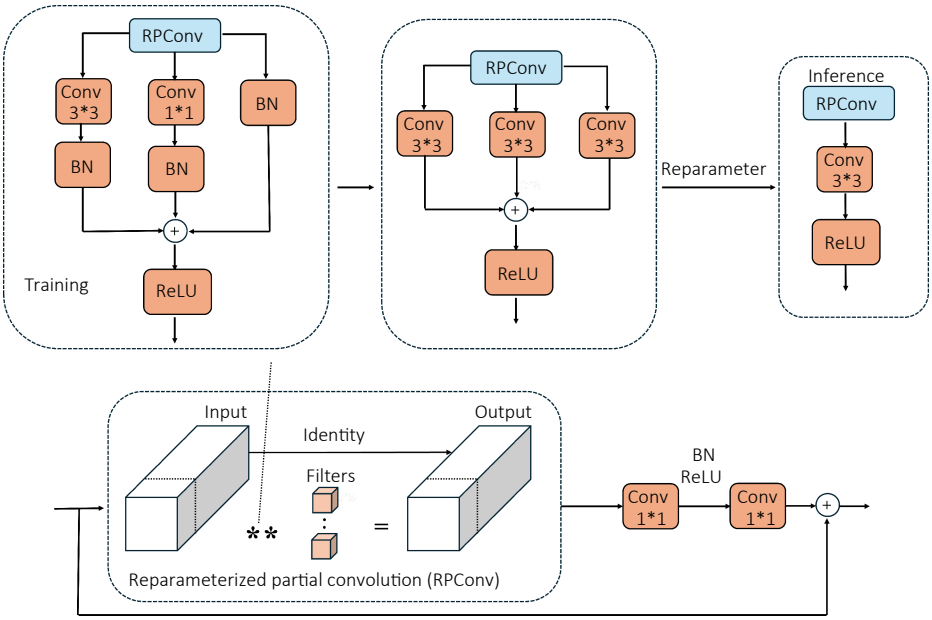


图 3 Rep-FasterNetBlock 结构图

Fig. 3 Rep-FasterNetBlock architecture diagram

$$BN(M*W,\mu,\sigma,\gamma,\beta)_{:,i,:,:}=\gamma_i\cdot\frac{(M*W)_{:,i,:,:}-\mu_i}{\sqrt{\sigma_i^2+\varepsilon}}+\beta_i\,, \tag{2}$$

进一步整理得:

$$BN(M*W,\mu,\sigma,\gamma,\beta)_{:,i,:,:}=\frac{\gamma_i}{\sqrt{\sigma_i^2+\varepsilon}}\cdot(M*W)_{:,i,:,:}-\frac{\gamma_i\cdot\mu_i}{\sqrt{\sigma_i^2+\varepsilon}}+\beta_i\,, \tag{3}$$

式中：M表示输入特征图；W表示卷积核大小；$(M*W)_{:,i,:,:}$表示该张量在第 i 个通道上的二维特征图，即对所有批次样本及空间位置进行索引，其中“:”表示对该维度上的全部元素取值；μ_i表示第 i 个通道的均值；σ_i表示第 i 个通道的标准差；γ_i表示第 i 个通道的缩放因子；β_i表示第 i 个通道的偏移因子；ε表示极小的常数，防止分母为 0。

将每个 BN 及其前面的 Conv 层转换为带有偏置向量的 Conv。设$\{K',b'\}$为从$\{K,\mu,\sigma,\gamma,\beta\}$转换的核和偏差则有

$$K'_{i,:,:,:}=\frac{\gamma_i}{\sqrt{\sigma_i^2+\varepsilon}}\cdot K_{i,:,:,:}\,, \tag{4}$$

$$b'_i=\beta_i-\frac{\mu_i\gamma_i}{\sqrt{\sigma_i^2+\varepsilon}}\,, \tag{5}$$

式中：K 为原始卷积核权重；K'为重参数化后的卷积核权重；b'为重参数化后的偏置项。

将式 (4) 和式 (5) 代入式 (3) 得到推导式 (6)，该公式通过结构化的参数融合机制，将卷积与批量归一化的多项参数高效整合，从而显著降低推理时的计算复杂度提升了推理速度。

$$BN(M*K,\mu,\sigma,\gamma,\beta)_{:,i,:,:}=(M*K')_{:,i,:,:}+b'_i\,. \tag{6}$$

3.2 颈部网络改进

RT-DETR 原模型在多尺度特征交互中存在固定尺寸卷积核难以动态适配多尺度缺陷的几何特性差异，导致高层语义特征与底层细节特征在融合过程中出现语义割裂，尤其对微小缺陷的定位精度显著不足，同时层级间重复卷积操作产生大量无效计算负载，制约实时检测效率。为此本文设计 DMSFPN 结构替代原有跨尺度交互模块。该改进通过精细化特征融合粒度和优化计算路径，在保持全局语义捕捉能力的同时提升特征表达效率。在跨分辨率融合上原始结构的静态卷积与单向特征传递易导致多尺度缺陷语义错位及计算冗余，而 DMSFPN 通过动态核自适应机制、基于可变形卷积的空间连续性重建策略及双向交互闭环设计，以动态特征学习替代静态特征映射，以双向信息交互突破单向特征传递的局限，从而实现计算效率与特征一致性的协同提升。

本文基于图 4 的 DMSFPN 架构，通过层级特征金字塔融合多尺度特征优化与轻量计算。颈部网络采用尺度自适应多尺度卷积模块 (multi-scale convolution block, MSCB)[12]，在高层特征中部署大卷积核增强大目标识别能力，同时在底层特征中密集堆叠小核提升小目标定位精度，通过动态核选择机制平衡语义与空间特征表达。跨层特征交互采用 BiFPN[13] 策略，基于双向递归路径实现跨尺度特征交互：自上而下传递高层语义增强浅层全局感知，自底向上融合底层细节优化深层定位精度，减少冗余参数并增强跨分辨率语义一致性；高效上采样模块 (efficient up-convolution block, EUCB)[12] 引入可变形

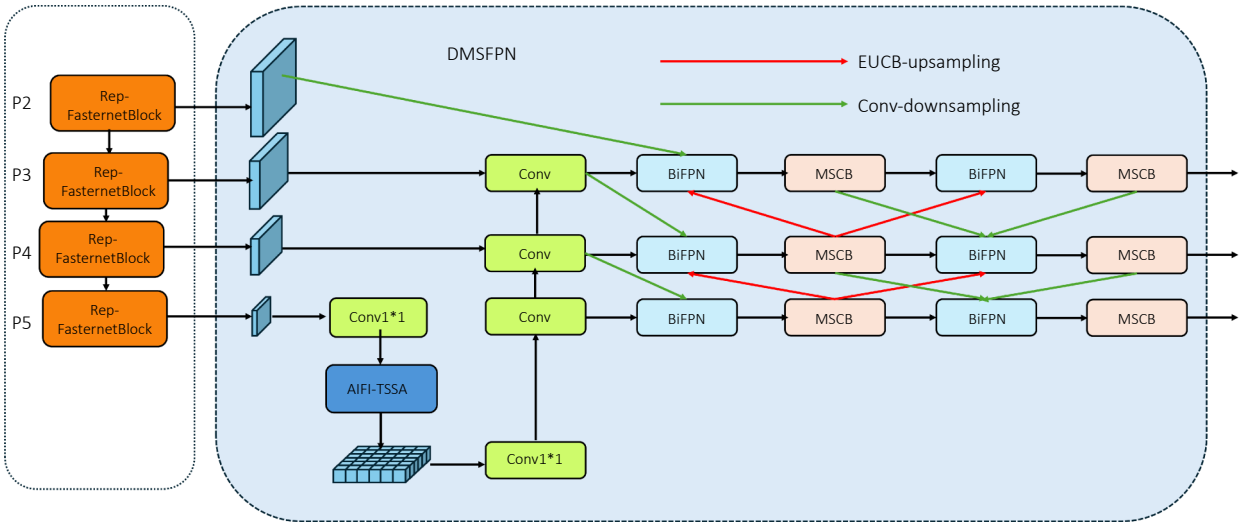


图 4 DMSFPN 网络结构

Fig. 4 DMSFPN network architecture diagram

卷积与通道交互机制，通过空间自适应采样增强特征图的空间连续性，解决常规上采样导致的细节模糊问题。各层级经 MSCB 压缩通道维度后，通过 EUCB 进行分辨率恢复与特征精修，最终由 BiFPN 模块完成多路径特征聚合，输入检测头实现端到端预测。

3.2.1 多尺度卷积块 (MSCB)

MSCB 优化级联扩展路径的特征生成能力。该模块基于 MobileNetV2 的反向残差结构 (IRB)[14] 进行重构，具体包含四阶段特征变换：首先通过 1×1 卷积将通道维度扩展两倍 (扩展后通道数：原始通道数=2：1)，经批量归一化 (BN) 与 ReLU6 激活函数实现非线性映射。随后引入多分支深度卷积 (multi-scale depth-wise convolution, MSDC) 并行处理不同感受野的特征，其中各分支采用差异化的卷积核尺寸以捕捉多尺度空间上下文信息。针对深度卷积忽略通道关联性的局限，通过跨组通道混洗[15] 操作强化特征通道间的信息交互。最终采用 1×1 卷积进行通道维度压缩，配合 BN 层恢复原始通道维度。相较于传统逆残差块，该模块通过多尺度卷积核的协同计算与动态通道交互机制，在保持轻量化特性的同时显著增强了多分辨率特征的表达能力，MSCB 表达式为

$$MSCB(x) = BN(PWC_2(CS(MSDC(R6(BN(PWC_1(x))))))), \quad (7)$$

式中：PWC_1和PWC_2分别是负责扩展和压缩通道数量的逐点卷积层；$MSDC$是多尺度扩张卷积，用于捕捉输入特征的多尺度上下文信息；CS是增强特征交互的通道混洗操作；$R6$是ReLU6激活函数。MSCB 结构图如图 5 所示。

3.2.2 多分支深度卷积 (MSDC)

借鉴 YOLO-MS[16] 的异构核选择 (HKS) 策略，MSDC[12] 模块通过集成差异化的深度可分离卷积核 (DWC) 实现跨层级特征融合如图 6 所示。该模块集成 1×1、3×3、5×5 三种差异化的深度可分离卷积核实现跨层级特征融合。其中，1×1 卷积核专注捕捉微小缺陷的局部纹理特征，3×3 卷积核聚焦中等尺度缺陷的边缘与形态，5×5 卷积核则用于覆盖大面积缺陷的全局分布信息。各支路经 BN-ReLU6 优化后，通过动态加权筛选关键特征融合。因各卷积核对应的通道特征易形成不同尺度特征在通道维度独立分布，因此通道混洗操作通过跨组打乱通道索引，使 1×1 核的细节特征、3×3 核的中等尺度特征与 5×5 核的全局语义特征在通道维度全面交叉融合，缓解多尺度特征的冗余与冲突，增强特征多样性。

计算公式如下：

$$MSDC(x) = \sum_{ks \in KS} DWCB_{ks}(x), \quad (8)$$

式中：$DWCB_{ks}(x)$=$R6(BN(DWC_{ks}(x)))$，其中$DWC_{ks}(x)$是与核大小ks的深度卷积。此外， MSDC 使用递归更新的输入 x，其中输入 x 剩余的连接到先前的$DWCB_{ks}$，为了更好地正则化，如以下公式所示。

$$x = x + DWCB_{ks}(x). \quad (9)$$

3.2.3 高效上卷积块 (EUCB)

EUCB 是一种分层特征增强结构，通过多阶段轻量化操作平衡效率与性能。其首先通过 UP (upsampling) 进行 2 倍上采样提升空间分辨率后，使用深度可分离卷积解耦空间与通道运算以降低参数量，结合 BN-ReLU 增强梯度稳定性和非线性表达能力，最终通过 1×1 卷积动态调整通道维度，结构图如图 7 所示。该设计通过结构优化在减少计算复杂度的同时提升特征重建精度，实现效率与多尺度表征能力的协同优化。计算公式如下：

$$EUCB(x) = C_{1*1}(ReLU(BN(DWC(UP(x))))). \quad (10)$$

3.3 AIFI-TSSA 模块

基于 Transformer 目标检测模型中的自适应特征交互模块存在双重局限：其全局自注意力机制导致计算复杂度与图像块数量成平方关系，高分辨率图像处理时面临显存与算力瓶颈；同时，局部相似性度量易受冗余纹理或遮挡干扰，降低特征表征鲁棒性。

针对上述问题，本文采用基于统计特征压缩的 TSSA (token statistics self-attention) 注意力机制[17] 进行改进。该方法摒弃传统逐对相似性计算，通过分析特征通

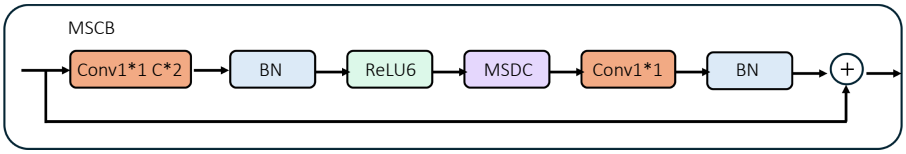


图 5 MSCB 结构图

Fig. 5 MSCB architecture diagram

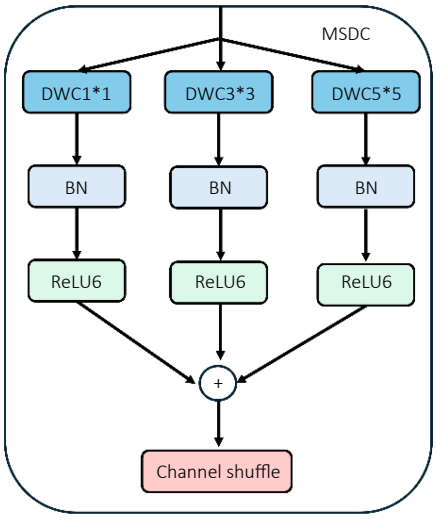


图 6 MSDC 结构图

Fig. 6 MSDC architecture diagram

道间的全局统计相关性构建低秩投影空间，利用图像块二阶统计量建立跨通道协方差结构，结合变分编码率最大化准则优化特征空间压缩，引导注意力聚焦于高层语义一致的特征维度。该机制在实现计算复杂度呈线性的同时，通过最大化特征子空间编码率差异有效提升抗噪声干扰能力。

本文引入的 TSSA 机制基于最大编码率压缩准则构建，给定输入特征矩阵$\boldsymbol{Z}\in\mathbb{R}^{d*n}$，TSSA 通过变分上界构建压缩目标函数其公式如下：

$$\mathcal{R}_{c,f}^{\mathrm{var}}\left(\boldsymbol{Z},\Pi\,|\,\{\boldsymbol{U}_k\}_{k=1}^{K}\right)=\frac{1}{2}\sum_{k=1}^{K}\frac{n_k}{n}\sum_{i=1}^{d}f\left(\frac{1}{n_k}\left(\boldsymbol{U}_k^{\mathrm{T}}\boldsymbol{Z}\mathrm{Diag}(\pi_k)\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{U}_k\right)_{ii}\right). \quad (11)$$

通过优化式 (11) 对特征矩阵 $\boldsymbol{Z}$ 求梯度可得更新规则：

$$\nabla_{\boldsymbol{Z}}R_{c,f}^{\mathrm{var}}(\boldsymbol{Z},\Pi\,|\,\{\boldsymbol{U}_k\}_{k=1}^{K})=\frac{1}{n}\sum_{k=1}^{K}\boldsymbol{U}_k\cdot\underbrace{\mathrm{Diag}\left(\nabla f\left[(\boldsymbol{U}_k^{\mathrm{T}}\boldsymbol{Z})^{\circledcirc 2}\frac{\pi_k}{\langle\pi_k,1\rangle}\right]\right)}_{\doteq\boldsymbol{D}(\boldsymbol{Z},\pi_k|\boldsymbol{U}_k)}\boldsymbol{U}_k^{\mathrm{T}}\boldsymbol{Z}\mathrm{Diag}(\pi_k), \quad (12)$$

其中对角激活矩阵 $\boldsymbol{D}$ 定义为

$$\boldsymbol{D}(\boldsymbol{Z},\pi_k\,|\,\boldsymbol{U}_k)=\mathrm{Diag}\left(\nabla f\left[\left(\boldsymbol{U}_k^{\mathrm{T}}\boldsymbol{Z}\right)^{\circledcirc 2}\frac{\pi_k}{\langle\pi_k,1\rangle}\right]\right). \quad (13)$$

通过最大后验估计推导软聚类机制：

$$\Pi=softmax\left(\frac{1}{2\eta}\begin{bmatrix}\left\|\boldsymbol{U}_1^{\mathrm{T}}z_1\right\|_2^2 & \cdots & \left\|\boldsymbol{U}_K^{\mathrm{T}}z_1\right\|_2^2\\ \vdots & \ddots & \vdots\\ \left\|\boldsymbol{U}_1^{\mathrm{T}}z_n\right\|_2^2 & \cdots & \left\|\boldsymbol{U}_K^{\mathrm{T}}z_n\right\|_2^2\end{bmatrix}\right), \quad (14)$$

式中：η为可学习温度参数，通过特征在投影空间的能量分布实现语义一致性，U_k表示第k个投影矩阵，n_k为分配到第k个簇的 token 数量，$f(\cdot)$为非线性压缩函数，$Diag(\pi_k)$表示由簇分配概率向量π_k构成的对角加权矩阵，∇_Z表示对输入特征矩阵Z求梯度，z_i表示输入特征矩阵Z中的第i个 token 特征向量。在 TSSA 的软聚类中，可学习温度参数通过调控特征在投影空间的能量分布，动态调整聚类粒度。在温度较低时，概率分布集中，特征被聚合为数量较少的簇，利于区分特征差异显著的缺陷；温度较高时，分布更平滑，特征簇数量相应增加，可捕捉相似缺陷的关联性，增强复杂背景下的检测鲁棒性。

最终注意力机制形式化为

$$TSSA\left(\boldsymbol{Z}\,|\,\{\boldsymbol{U}_k\}_{k=1}^{K}\right)\doteq-\frac{\tau}{n}\sum_{k=1}^{K}\boldsymbol{U}_k\boldsymbol{D}(\boldsymbol{Z},\pi_k\,|\,\boldsymbol{U}_k)\boldsymbol{U}_k^{\mathrm{T}}\boldsymbol{Z}\mathrm{Diag}(\pi_k) \quad (15)$$

因此可以看出 TSSA 通过统计特征压缩替代传统的成对相似性计算，显著降低复杂度，使模型处理高分辨率图像更高效。其利用全局统计关联和动态子空间聚焦，在保留关键语义信息的同时减少冗余计算，提升对噪声和复杂背景的鲁棒性，TSSA 原理图如图 8 所示。

在 AIFI-TSSA 模块中噪声率采用“初始自适应设定与动态反馈调整”的双阶段策略，初始值基于输入特征的统计特征自适应确定，以平衡噪声注入强度与有效特征保留；在迭代过程中，噪声率会根据模型任务损失进行动态优化，通过反馈机制适配不同场景下的噪声需求。算法伪代码如算法 1 所示。

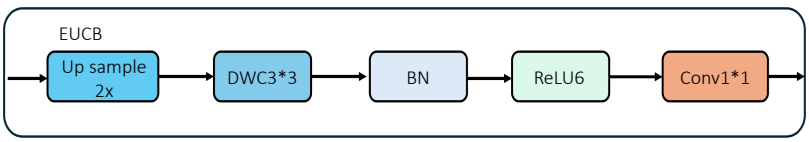


图 7 EUCB 结构图

Fig. 7 EUCB architecture diagram

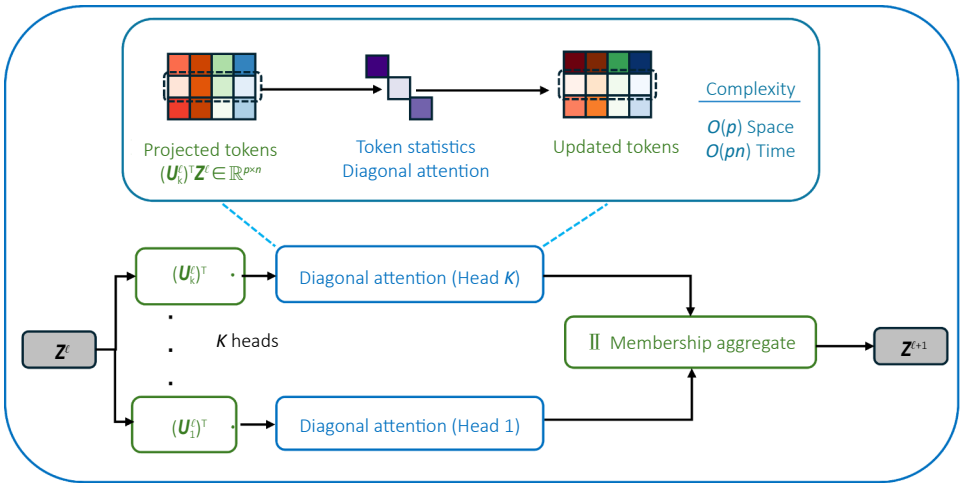


图 8 Token Statistics Self-Attention 结构图

Fig. 8 Token statistics self-attention architecture diagram

算法 1：TSSA 变分降噪与重参数化采样流程

Require: 输入特征集(X)，初始噪声率$(\epsilon_{\text{init}} = 0.1 \cdot std(X))$参数范围$([\mu_{\text{low}}, \mu_{\text{high}}], [\sigma_{\text{low}}, \sigma_{\text{high}}])$

Ensure: 处理后特征集 $(X_{\text{processed}})$

1) $(\epsilon \leftarrow \epsilon_{\text{init}})$// 加载初始噪声率，$(X_{\text{processed}} \leftarrow \varnothing)$

2) for each $(x \in X)$ do

3) $(x_{\text{noisy}} \leftarrow x + \epsilon \cdot N(0, I))$// 高斯噪声模拟真实场景干扰

4) $(\mu \leftarrow \frac{1}{n}\sum_{i=1}^{n} proj\,(x_{\text{noisy}}, i))$// 投影特征均值

5) $(\sigma^2 \leftarrow \frac{1}{n}\sum_{i=1}^{n} (proj\,(x_{\text{noisy}}, i) - \mu)^2)$// 投影特征方差

6) $(\mu \leftarrow clamp(\mu, \mu_{\text{low}}, \mu_{\text{high}}))$// 参数截断避免数值溢出

7) $(\sigma \leftarrow clamp(\sqrt{\sigma^2}, \sigma_{\text{low}}, \sigma_{\text{high}}))$

8) $(x_{\text{reparam}} \leftarrow \mu + \sigma \cdot \mathcal{N}(0, I))$ // 重参数化采样

9) $(x_{\text{final}} \leftarrow LayerNorm(x_{\text{reparam}}))$// 层归一化后处理

10) $(X_{\text{processed}} \leftarrow X_{\text{processed}} \cup \{x_{\text{final}}\})$

end for

// 噪声率自适应调整

11) $(\epsilon \leftarrow \epsilon \cdot \exp(-\alpha \cdot computer_{\text{loss}}\,(X_{\text{processed}})))$ //(α)为学习率系数

12) return $(X_{\text{processed}})$

4 实验结果与分析

4.1 数据集介绍

本文采用东北大学 NEU-DET[18] 钢材缺陷数据集与天池铝型材表面瑕疵数据集和 GC10-DET 钢材缺陷数据集作为实验数据源。NEU-DET 数据集包含图像 1800 张，涵盖开裂、点蚀表面、内含物、轧制氧化皮、斑块、和划痕 6 类典型钢材表面缺陷，各类别样本均衡分布各 300 幅。在数据划分策略上，按照 8∶1∶1 比例划分训练集、验证集、测试集，数据集部分示例图如图 9 所示。

天池铝型材表面瑕疵数据集由工业场景真实采集，包含 1885 张铝型材图像，涵盖 10 类缺陷，分别为不导电 (Bdd)、凸粉 (Tf)、擦花 (Ch)、脏点 (Zd)、碰伤 (Ps)、涂层开裂 (Tckl)、起坑 (Qk)、桔皮 (Jp)、凹陷 (Ax)、漏底 (Ld)。数据按 8∶1∶1 划分，训练集 1508 张、验证集 188 张、测试集 189 张，数据集部分示例图如图 10 所示。

GC10-DET 是在真实工业场景中收集的表面缺陷数据集共有 2294 张图像，它包含 10 种类型钢板表面缺陷，分别为：冲孔 (punching-hole, Ph)、焊缝 (welding-line, Wl)、月牙弯 (crescent-gap, Cg)、水斑 (water-spot, Ws)、油斑 (oil-spot, Os)、丝斑 (silk-spot, Ss)、异物 (inclusion, In)、压痕 (rolled-pit, Rp)、折痕 (crease, Cr)、 腰部折痕 (waist-folding, Wf)。实验中将数据集按照 8∶1∶1 划分成训练集、验证集、测试集，数据集部分示意图如图 11 所示。

4.2 实验环境

本文的实验训练环境配置了 Windows11 操作系统、Intel Core i5-14600KF 和 NVIDIA GeForce RTX 4060TI，

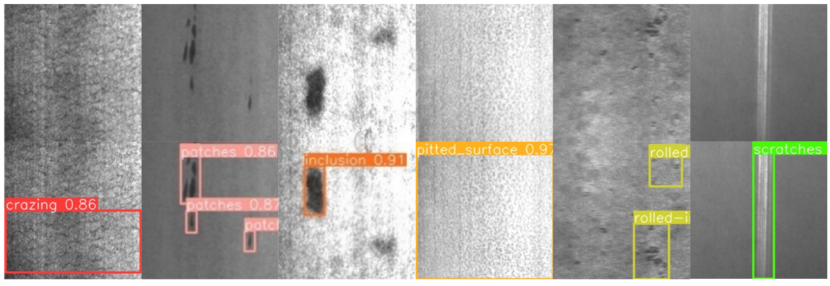


图 9 NEU-DET 数据集部分示例图

Fig. 9 NEU-DET dataset sample image

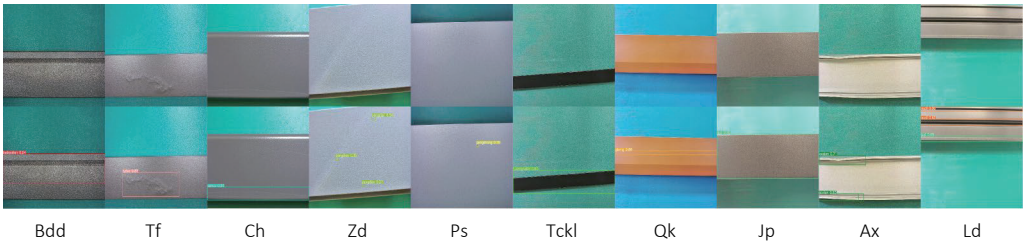


图 10 天池铝型材表面瑕疵数据集部分示例图

Fig. 10 Tianchi aluminum profiles surface defects dataset sample images

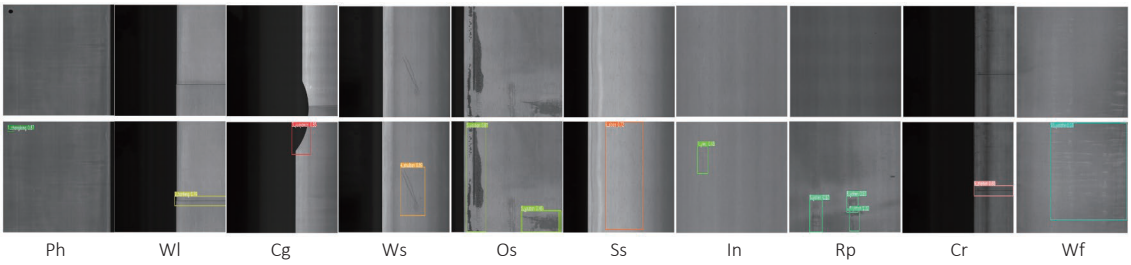


图 11 GC10-DET 数据集部分示例图

Fig. 11 Sample images from the GC10-DET dataset

显存为 16 G，编译环境为 Python 3.8.16+torch 1.13.1+CUDA 11.7。训练参数设置：epochs 为 300，batch size 为 8，进程数 workers 为 2。采用 AdamW 优化器，基础学习率设置为 0.0001，输入图像大小为 640 pixel×640 pixel。

4.3 评价指标

本文所采用的评价指标有召回率 (recall, R)、精确率 (precision, P)、平均精度 ($mAP@0.5$)、F1 因子 (F_1)、参数量 (Params)、浮点运算次数 (GFLOPs)。

精准度和召回率的表达式如下：

$$P=\frac{TP}{(TP+FP)}, \tag{16}$$

$$R=\frac{TP}{(TP+FN)}, \tag{17}$$

式中：TP表示模型正确预测的正类样本数；FP表示模型错误预测的正类样本数；FN表示模型漏检的真实正类样本数。

F_1的表达式为

$$F_1=\frac{2(P\times R)}{(P+R)}. \tag{18}$$

mAP的表达式为

$$mAP=\frac{1}{n}\sum_{i=1}^{n}\int_0^1 P(R)\mathrm{d}(R). \tag{19}$$

4.4 消融实验

为了验证各个模块的有效性，本文以 RT-DETR 为基准方法，采用 NEU-DET 数据集在测试集上设计消融

实验，实验结果如表 1 所示，同时将基准模型 RT-DETR 简称为 BASE，将主干网络 Rep-FasternetBlock 模块改进称为 A，将 DMSFPN 改进称为 B，将 AIFI-TSSA 模块改进称为 C。

由表 1 对比方法 1 和方法 2 可知，引入 Rep-FasternetBlock 对主干网络进行轻量化操作后模型的计算量和参数量相对于原模型分别减少 24.7% 和 29.5%，同时平均精度和帧率 (frame per second, FPS) 分别增加了 1.3% 和 34.0%，从而证明了该模块能够显著减少参数量和计算量提升模型的性能。对比方法 1 和方法 3 可知，当模型加入 DMSFPN 模块，融合不同层次和分辨率的深浅层特征后在各类平均精度增加约 1.0%，召回率提升约 2.3%。对比方法 1 和方法 4，引入 AIFI-TSSA 模块在增强特征表达的多样性后模型能够有效地提升精确率和召回率。最后将三个改进点同时添加到 RT-DETR 网络中，所提模型的精确率为 77.8%，召回率为 72.6%，mAP@0.5 为 76.7%，帧率为 78.9 f/s，相较于原模型提升了 1.7%、5.2%、2.6% 和 30.4%，同时模型的参数量、计算量和模型的大小分别减少了 39.9%、39.3% 和 39.4%。在 mAP@0.5:0.95 这个更严格的指标上，改进后的模型的精度提升了 1.8%，验证了改进后模型的有效性。

实验数据表明，本文提出的模型在缺陷检测任务中展现出显著优势。相较于基准模型，网络结构的改进不仅提升了检测精度，同时实现了轻量化目标，充分验证了该模型在工业场景中的实用性与高效性。

4.5 对比实验

4.5.1 不同模型对比实验

为了验证本文模型的性能，选取 SSD、Faster-RCNN 以及不同版本的 YOLO 模型在东北大学 NEU-DET 数据集上进行对比实验，实验结果如表 2 所示。

结果显示本文的模型相较于其他系列模型表现出更卓越的性能。本文模型不仅 mAP@0.5达到了 76.7%，而且精确率和召回率分别达到了 77.8% 和 72.6%，同时模型在参数量、计算量和存储空间占用等关键效率指标上均实现有效压缩。在与其规模较为相近的模型对比中虽然参数量较大，但其检测精度和精确率召回率均较高，使得算法在工业质检设备部署场景中展现出更高的工程实用价值。

4.5.2 AIFI 模块改进对比实验

为了验证在 AIFI 模块中使用 TSSA 注意力机制在钢材表面缺陷检测中的优越性，本文在模型改进的基础上修改了 AIFI 模块的不同注意力机制作为对比实验，其中不同注意力机制包括：PA (polar attention)[19]、DPBA (dynamic pos bias attention)[20]、DHSA (dynamic-rangehistogram self-attention)[21]、EAA (efficient additive attention)[22]、CGA (cascaded group attention)[23]、DA(deformable attention)[24]。结果如表 3 所示，引入 TSSA 注意力机制后其平均精度达到 76.7%，精确率以 77.8% 为对比实验最优；同时，模型计算量与体积较基线方法分别缩减至 34.6 G 和 23.4 M，实现同类型较小规模。尽管参数量与召回率略低于最优模型，但差异均较小，在精度与效率的协同优化中验证了该机制的有效性。

4.5.3 跨尺度对比实验

为验证模型在跨尺度检测任务中的鲁棒性，本文将 NEU-DET、天池铝材和 GC10-DET 数据集的输入图像尺寸缩小至原始尺寸的 1/10，进行对比实验。如表 4 所示，改进模型的召回率在三个数据集上分别达到 49.0%、

表 1 消融实验结果

Table 1 Ablation experiment results

Method	Model	*P*/%	*R*/%	F1/%	mAP@0.5/%	mAP@0.5:0.95/%	FPS	Params/M	GFLOPs	Model size/M
1	BASE	76.1	67.4	70.0	74.1	41.4	60.5	19.88	57.0	38.6
2	BASE+A	75.3	71.4	72.4	75.3	42.5	**81.1**	14.0	42.9	27.5
3	BASE+B	75.9	69.7	71.8	75.1	42.2	57.2	17.95	48.6	34.7
4	BASE+C	81.6	68.0	72.4	74.6	41.8	64.1	19.75	57.1	38.3
5	BASE+A+B	73.5	70.3	71.2	75.8	43.0	70.4	12.07	**34.5**	23.6
6	BASE+A+C	**83.8**	68.4	73.4	76.0	42.4	75.5	13.87	43.0	27.3
7	BASE+B+C	77.4	71.8	73.1	75.9	43.0	61.7	17.81	48.7	34.5
8	BASE+A+B+C	77.8	**72.6**	**74.7**	**76.7**	**43.2**	78.9	**11.94**	34.6	**23.4**

42.9% 和 52.7%，显著超过原模型对应指标；同时在 mAP@0.5 和 mAP@0.5:0.95 两项指标上均优于基准模型。该系列实验结果充分证实，在尺度变化条件下，改进模型仍能有效降低漏检率并提升检测精度，具备显著的跨尺度适应优势。

4.6 模型泛化实验

为了验证模型泛化能力，本文在天池铝型材数据集进行消融实验和对比实验，将 DMSFPN 改进称为 A，将主干网络 Rep-FasternetBlock 模块改进称为 B，将 AIFI-TSSA 模块改进称为 C。如表 5、表 6 所示，改进

表 2　不同模型对比实验

Table 2　Comparison of different models

Model	*P*/%	*R*/%	mAP@0.5/%	Params/M	GFLOPs	Model size/M
SSD	38.0	41.8	37.7	24.41	141.3	187.0
Faster-RCNN	61.9	69.9	71.9	41.37	134.0	315.0
YOLOv3	70.6	69.6	73.8	61.52	154.6	315.0
YOLOv7	67.1	59.3	67.2	37.22	105.2	71.3
YOLOv5m	76.1	67.9	73.9	25.05	64.0	48.2
YOLOv8m	74.6	69.0	75.1	25.84	78.7	49.6
YOLOv11m	65.9	**75.4**	74.9	20.03	67.7	38.7
YOLOv11s	64.8	72.1	74.0	9.42	21.3	18.3
YOLOv12m	63.9	73.4	74.0	19.58	59.5	37.9
YOLOv12s	67.2	64.3	73.2	**9.08**	**19.3**	**17.8**
RT-DETR	76.1	67.4	74.1	19.88	57.0	38.6
RMD-DETR	**77.8**	72.6	**76.7**	11.94	34.6	23.4

表 3　不同注意力机制对比

Table 3　Comparison of different attention mechanisms

Attention	*P*/%	*R*/%	mAP@0.5/%	Params/M	GFLOPs	Model size/M
DHSA	77.6	66.5	72.6	12.21	34.9	23.9
DPBA	71.7	69.1	73.3	12.07	34.8	24.8
PA	69.5	68.4	73.7	12.24	34.8	23.9
DA	77.1	69.3	75.9	12.08	34.7	23.7
CGA	75.3	**72.7**	76.5	**11.90**	34.6	23.4
TSSA	**77.8**	72.6	**76.7**	11.94	**34.6**	**23.4**

表 4　跨尺度对比实验

Table 4　Cross-scale performance comparison experiment

Dataset	Model	*P*/%	*R*/%	mAP@0.5/%	mAP@0.5:0.95/%	Miss rate/%
NEU-DET	RT-DETR	56.5	44.6	48.5	25.1	55.4
	RMD-DETR	**57.4**	**49.1**	**53.0**	**26.8**	**50.9**
Tianchi Aluminum	RT-DETR	**48.1**	40.9	37.0	21.3	59.1
	RMD-DETR	42.9	**42.9**	**38.5**	**23.2**	**57.1**
GC10-DET	RT-DETR	61.4	50.8	51.1	24.6	49.2
	RMD-DETR	**64.5**	**52.7**	**54.4**	**25.7**	**47.3**

后模型 mAP@0.5 达到 53.2%，较原模型提升 2%，精确率与召回率分别提升至 68.8% 和 57.0%，且减少了模型参数量与计算量。在不同模型对比实验中，本文改进模型指标均取得较好的效果。实验结果证实该方案在不同数据集场景下仍保持较高检测精度体现了良好的鲁棒性和适用性。

为了进一步验证本文改进模型的泛化能力，在 GC10-DET 数据集上也进行了泛化实验。如表 7、表 8 所示，将 DMSFPN 改进称为 A，将主干网络 Rep-FasternetBlock 模块改进称为 B，将 AIFI-TSSA 模块改进称为 C。消融实验结果表明：改进模型的 mAP@0.5 达到 70.0%，较原始模型提升了 3%；同时，在更严格指标 mAP@0.5:0.95 上也取得了提升。尽管模型精确度略有下降，但其召回率提升至 71%，能更有效改善模型漏检情况。此外，模型的参数量和计算量均显著降低。上述结果充分验证了模型改进的有效性。

4.7 不同缺陷检测性能对比

为了更进一步的验证改进后的模型对于小缺陷的识别能力，本文将原模型与改进后的模型分别在三个数据集上统计两类模型针对不同缺陷类别的检测精度，实验结果如表 9 所示。实验结果表明在三个数据集上，本文模型相较于原模型，不仅对大部分缺陷类别的检测精度更高，且其平均检测精度均高于原模型，进一步验证了所提模型的鲁棒性和对小缺陷检测能力的提升。

4.8 可视化分析

为更直观观察改进模型的检测效果，在 NEU-DET 测试集上添加高斯噪声进行实验，进一步验证模型的鲁棒性，如图 12 所示。实验表明：a 组、c 组中原模型和

表 5 天池铝材数据集消融实验

Table 5 Ablation study on the Tianchi aluminum profile dataset

Model	*P*/%	*R*/%	mAP@0.5/%	mAP@0.5:0.95/%	Params/M	GFLOPs
BASE	67.6	51.2	51.2	29.7	19.88	57.0
BASE+A	66.8	52.8	52.0	29.7	17.95	48.7
BASE+A+B	56.7	54.0	52.3	30.5	12.08	**34.5**
BASE+A+B+C	**68.8**	**57.0**	**53.2**	**31.1**	**11.94**	34.6

表 6 不同模型对比实验

Table 6 Comparison of different models

Model	*P*/%	*R*/%	mAP@0.5/%	mAP@0.5:0.95/%	Params/M	GFLOPs
RT-DETR	67.6	51.2	51.2	29.7	19.88	57.0
YOLOV8m	49.7	53.6	52.8	29.2	25.84	78.7
YOLOV11m	55.8	49.0	51.9	29.7	20.03	67.7
YOLOV12m	63.6	52.0	51.3	31.1	19.58	59.5
RMD-DETR	**68.8**	**57.0**	**53.2**	**31.1**	**11.94**	**34.6**

表 7 GC10-DET 数据集消融实验

Table 7 Ablation study on the GC10-DET dataset

Model	*P*/%	*R*/%	mAP@0.5/%	mAP@0.5:0.95/%	Params/M	GFLOPs
BASE	**72.4**	66.8	67.0	32.4	19.88	57.0
BASE+A	72.7	69.3	68.1	32.1	17.95	48.6
BASE+A+B	71.2	67.2	68.7	33.5	12.07	**34.5**
BASE+A+B+C	71.0	**71.0**	**70.0**	**33.7**	**11.94**	34.6

表 8 不同模型对比实验

Table 8 Comparison of different models

Model	*P*/%	*R*/%	mAP@0.5/%	mAP@0.5:0.95/%	Params/M	GFLOPs
RT-DETR	**72.4**	66.8	67.0	32.4	19.88	57.0
YOLOV8m	69.2	62.7	64.5	32.4	25.84	78.7
YOLOV11m	62.3	69.6	64.8	33.3	20.03	67.7
YOLOV12m	63.6	67.2	65.4	33.2	19.58	59.5
RMD-DETR	71.0	**71.0**	**70.0**	**33.7**	**11.94**	**34.6**

表 9 不同检测模型的多类别缺陷检测对比实验

Table 9 Comparison of multi-class defect detection performance for different detection models

Dataset	Model	AP@0.5:0.95/%									
NEU-DET		Cr	Pa	In	Ps	RS	Sc				
	RT-DETR	**15.7**	41.2	58.3	48.5	29.2	55.6				
	RMD-DETR	15.5	**43.6**	**60.8**	**50.2**	**32.2**	**56.7**				
天池铝材		Bdd	Tf	Ld	Qk	Ps	Tckl	Zd	Jp	Ax	Ch
	RT-DETR	43.7	27.1	**24.9**	39.7	**2.5**	**64.5**	**9.1**	47.5	24.8	13.2
	RMD-DETR	**46.1**	**35.9**	24.0	**42.7**	1.0	64.0	7.3	**49.2**	**26.4**	**14.1**
GC10-DET		Cg	Cr	Ss	Ws	Wl	In	Os	Rp	Ph	Wf
	RT-DETR	54.9	21.1	**22.0**	**46.0**	47.9	**14.8**	31.1	2.3	44.5	**39.8**
	RMD-DETR	**58.3**	**27.8**	21.0	45.1	**48.7**	13.5	**34.8**	**3.8**	**47.0**	37.3

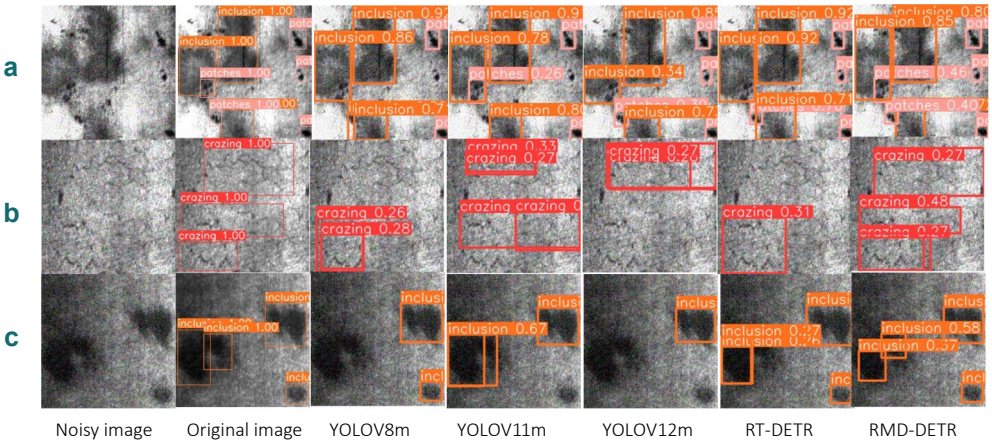


图 12 检测结果图

Fig. 12 Detection results diagram

YOLO 系列存在缺陷漏检，改进模型精准检测出细微缺陷；b 组中原模型置信度低，改进模型检测置信度显著提升并且 YOLO 系列模型边界框重合问题突出，同时存在明显的漏检情况，改进模型能较准确地识别目标位置与大小。结果显示，改进方案在保持检测效率的同时，显著降低复杂工业场景下的误报率和漏检率，有效提升细微缺陷识别的鲁棒性。

为更准确地评估所提模型的检测能力，对原数据集添加高斯噪声，在此基础上采用热力图可视化技术标识图像中的显著区域，直观展示模型预测过程中关注的目标定位及其空间分布特征。相关可视化结果如图 13 所示。实验对比表明，相较于基准模型，所提模型在缺陷检测中展现出更高的精度与噪声抗干扰性，证实了模型鲁棒性的显著提升。

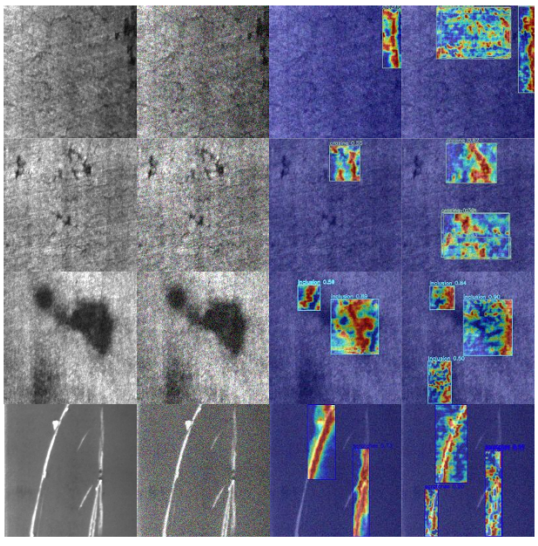


图 13 热力图对比结果图

Fig. 13 Heatmap comparison results diagram

5 结 论

本文提出的 RMD-DETR 通过引入重参数化的 Rep-FasterNetBlock 优化骨干网络、设计 DMSFPN 增强跨尺度特征融合、构建 AIFI-TSSA 模块提升复杂背景鲁棒性，在 NEU-DET 与天池铝型材数据集和 GC10-DET 数据集上均实现了检测精度与效率的协同提升，验证了其在钢材表面缺陷检测任务中的有效性。但模型仍存在局限：尽管相较于 RT-DETR，RMD-DETR 的参数量已显著减少，但在资源极度受限的边缘设备部署场景中，仍有进一步压缩的空间，可通过模型剪枝等策略进一步降低参数规模，在精度损失可控范围内提升部署灵活性。在工业部署方面，依托模型轻量化特性，可采用 TensorRT 工具进行量化优化，适配边缘计算设备并支持 ONNX 格式导出，便于集成至工业质检流水线，进一步发挥其在实际生产场景中的应用价值。

作者贡献

蒋宏捷：研究思路设计与资料的收集与整理、实验实施与数据分析、论文撰写； 肖小玲：论文选题与研究方案指导、论文修改与完善；颜昕：研究 方案指导、论文内容审核与修改建议。

利益冲突

所有作者声明无利益冲突

参考文献

1. Girshick R. Fast R-CNN[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*, Santiago, Chile, 2015: 1440–1448. https://doi.org/10.1109/ICCV.2015.169.
2. Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. *IEEE Trans Pattern Anal Mach Intell*, 2017, **39**(6): 1137−1149.
3. Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[C]//*Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016: 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.
4. Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016: 779–788. https://doi.org/10.1109/CVPR.2016.91.
5. Xia B Z, Luo H, Shi S G. Improved faster R - CNN based surface defect detection algorithm for plates[J]. *Comput. Intell. Neurosci.*, 2022, **2022**: 3248722.
6. Liang L M, Long P W, Lu B H, et al. Improvement of GBS-YOLOv7t for steel surface defect detection[J]. *Opto-Electron Eng*, 2024, **51**(5): 240044.
梁礼明, 龙鹏威, 卢宝贺, 等. 改进 GBS-YOLOv7t 的钢材表面缺陷检测[J]. 光电工程, 2024, **51**(5): 240044.
7. Ma Y J, Cao C Q, Zhang J. Steel surface defect detection based on YOLOv8-SOE[J]. *Opto-Electron Eng*, 2025, **52**(5): 250032.
马玉洁, 曹雏清, 张靖. 基于 YOLOv8-SOE 的钢材表面缺陷检测[J]. 光电工程, 2025, **52**(5): 250032.
8. Bodla N, Singh B, Chellappa R, et al. Soft-NMS—improving object detection with one line of code[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*, Venice, Italy, 2017: 5562–5570. https://doi.org/10.1109/ICCV.2017.593.
9. Zhao Y A, Lv W Y, Xu S L, et al. DETRs beat YOLOs on real-time object detection[C]//*Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2024: 16965–16974. https://doi.org/10.1109/CVPR52733.2024.01605.
10. Chen J R, Kao S H, He H, et al. Run, don't walk: chasing higher FLOPS for faster neural networks[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 2023: 12021–12031. https://doi.org/10.1109/CVPR52729.2023.01157.
11. Ding X H, Zhang X Y, Ma N N, et al. RepVGG: Making VGG-style ConvNets great again[C]//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021: 13728–13737. https://doi.org/10.1109/CVPR46437.2021.01352.
12. Rahman M, Munir M, Marculescu R. EMCAD: efficient multi-scale convolutional attention decoding for medical image segmentation[C]//*Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2024: 11769–11779. https://doi.org/10.1109/CVPR52733.2024.01118.
13. Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020: 10778–10787. https://doi.org/10.1109/CVPR42600.2020.01079.
14. Sandler M, Howard A, Zhu M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018: 4510–4520. https://doi.org/10.1109/CVPR.2018.00474.
15. Zhang X Y, Zhou X Y, Lin M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018: 6848–6856. https://doi.org/10.1109/CVPR.2018.00716.
16. Yang Z Q, Guan Q, Zhao K E, et al. Multi-branch auxiliary fusion YOLO with re-parameterization heterogeneous convolutional for accurate object detection[C]//*Proceedings of the 7th Chinese Conference on Pattern Recognition and Computer Vision*, Singapore, 2024: 492–505. https://doi.org/10.1007/978-981-97-8858-3_34.
17. Wu Z Y, Ding T J, Lu Y F, et al. Token statistics transformer: linear-time attention via variational rate reduction[C]//*Proceedings of the 13th International Conference on Learning Representations*, Singapore, 2025.
18. He Y, Song K C, Meng Q G, et al. An end-to-end steel surface defect detection approach via fusing multiple hierarchical

features[J]. *IEEE Trans Instrum Meas*, 2020, **69**(4): 1493−1504.
19. Meng W K, Luo Y D, Li X, et al. PolaFormer: polarity-aware linear attention for vision transformers[C]//*Proceedings of the 13th International Conference on Learning Representations*, Singapore, 2025.
20. Wang W X, Chen W, Qiu Q B, et al. CrossFormer++: a versatile vision transformer hinging on cross-scale attention[J]. *IEEE Trans Pattern Anal Mach Intell*, 2024, **46**(5): 3123−3136.
21. Sun S Q, Ren W Q, Gao X W, et al. Restoring images in adverse weather conditions via histogram transformer[C]//*Proceedings of the 18th European Conference on Computer Vision*, Milan, Italy, 2024: 111–129. https://doi.org/10.1007/978-3-031-72670-5_7.
22. Shaker A, Maaz M, Rasheed H, et al. SwiftFormer: efficient additive attention for transformer-based real-time mobile vision applications[C]//*Proceedings of 2023 IEEE/CVF International Conference on Computer Vision*, Paris, France, 2023: 17379–17390. https://doi.org/10.1109/ICCV51070.2023.01598.
23. Liu X Y, Peng H W, Zheng N X, et al. EfficientViT: memory efficient vision transformer with cascaded group attention[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 2023: 14420–14430. https://doi.org/10.1109/CVPR52729.2023.01386.
24. Xia Z F, Pan X R, Song S J, et al. Vision transformer with deformable attention[C]//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022: 4784–4793. https://doi.org/10.1109/CVPR52688.2022.00475.

作者简介



蒋宏捷 (2003-)，男，硕士研究生，主要研究方向为计算机视觉与图像处理。

E-mail：2025710773@yangtzeu.edu.cn



颜昕 (1970-)，男，副教授，博士，主要研究方向：新型智能通信网络与系统、云计算与移动计算、大语言模型应用。

E-mail：yanxin@whut.edu.cn



【通信作者】肖小玲 (1973-)，女，教授，博士，主要研究方向为智能信息处理与网络安全。

E-mail：xxl@yangtzeu.edu.cn

RMD-DETR: an improved steel surface defect detection algorithm

Jiang Hongjie[1,2], Xiao Xiaoling[1,2*], Yan Xin[3]

[1] School of Computer Science, Yangtze University, Jingzhou, Hubei 434023, China;
[2] Artificial Intelligence Research Platform, Yangtze University, Jingzhou, Hubei 434023, China;
[3] School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, Hubei, China 430070

Abstract

Objective Steel surface defects such as cracks, scratches, rolled-in scale, pitting, patches, and inclusions are common in industrial manufacturing and may cause severe degradation in corrosion resistance, wear resistance, and mechanical reliability, which makes accurate and efficient surface defect inspection essential for quality control. Although deep learning based detectors have achieved promising results, reliable detection of tiny defects remains difficult because these defects often present small size, low contrast, blurred boundaries, and large scale variations, while industrial backgrounds contain complex textures and noise that easily interfere with feature extraction. Two-stage detectors usually require heavy computation and fail to meet real-time requirements, and many one-stage detectors still suffer from high miss rates on small defects and limited cross-scale adaptability. RT-DETR introduces an end-to-end real-time detection framework by eliminating non-maximum suppression and leveraging Transformer-based global modeling, which improves dense target prediction, yet its parameter redundancy and insufficient cross-resolution feature alignment still restrict its efficiency and robustness in

Foundation item: National Natural Science Foundation of China (61771354)
*Correspondence: Xiao X L, E-mail: xxl@yangtzeu.edu.cn

industrial scenarios.

Methods To address these limitations, this study proposes RMD-DETR, an improved RT-DETR based detection algorithm that enhances detection accuracy, cross-scale stability, and inference efficiency through three coordinated modifications. First, a lightweight FasterNet backbone is introduced to replace the original feature extractor, and the partial convolution in FasterNetBlock is further optimized with a reparameterization strategy to strengthen feature representation under limited computation. During training, multi-branch convolutional paths are used to enrich spatial diversity and improve learning capacity, while during inference these branches are equivalently fused into a single convolution operation, which simplifies the computational graph and accelerates deployment without sacrificing accuracy. Second, a dynamic multi-scale feature fusion network (DMSFPN) is designed as the neck to alleviate semantic inconsistency and misalignment between high-level semantic features and low-level detailed features, which are critical for localizing micro-defects. DMSFPN integrates a multi-scale convolution block (MSCB) to capture context information at different receptive fields, where heterogeneous kernel branches focus on tiny, medium, and large defect patterns, and a channel shuffle mechanism is employed to enhance cross-channel interaction and reduce feature redundancy. To improve detail recovery in cross-scale aggregation, an efficient upsampling convolution block (EUCB) is introduced, which combines lightweight convolution operations with deformable convolution based adaptive sampling, thereby mitigating the feature blurring problem caused by standard interpolation and preserving spatial continuity for small targets. A bidirectional feature fusion strategy is adopted to enable both top-down semantic propagation and bottom-up detail enhancement, improving cross-resolution consistency while avoiding redundant repeated computation. Third, to enhance feature interaction under cluttered backgrounds, a token statistics self-attention (TSSA) mechanism based on statistical feature compression and variational rate reduction is incorporated into the adaptive feature interaction module, forming an AIFI-TSSA enhancement block. Unlike conventional self-attention that computes pairwise token similarity with quadratic complexity and high memory cost, TSSA performs global context modeling with linear complexity through token statistics driven projection and denoising, which reduces computation overhead and improves robustness against redundant textures, occlusion, and noise. The proposed RMD-DETR is evaluated on three industrial datasets, including NEU-DET, the Tianchi Aluminum profile surface defect dataset, and GC10-DET, and comprehensive experiments are conducted to verify the effectiveness of each proposed component, including ablation studies, comparative experiments with representative detectors, cross-scale evaluation by resizing inputs to a smaller scale, and generalization tests across different industrial scenes.

Results and Discussions Experimental results on NEU-DET demonstrate that RMD-DETR achieves 76.7% mAP@0.5, outperforming the baseline RT-DETR by 2.6%, while precision and recall reach 77.8% and 72.6%, indicating improved localization reliability and reduced missed detections. The model simultaneously reduces parameters and computational cost by 39.9% and 39.3%, and the inference speed increases to 78.9 f/s, satisfying the real-time inspection requirement of industrial production lines. In cross-scale detection experiments, the missed detection rate decreases from 55.4% to 50.9%, showing stronger adaptability under significant scale variation. Generalization experiments further confirm the transferability of the proposed design, where mAP@0.5 improves by 2% on the Tianchi dataset and by 3% on GC10-DET compared with the baseline model. Visualization analysis under Gaussian noise demonstrates that the proposed method maintains higher confidence and more complete defect coverage, especially for micro-scratches and fine cracks that are easily ignored in complex textures. These results indicate that the three improvements are complementary: the reparameterized lightweight backbone improves efficiency while retaining feature extraction capability, DMSFPN enhances multi-scale semantic alignment and detail preservation, and AIFI-TSSA provides stable global modeling with reduced complexity, enabling the detector to focus on discriminative defect regions rather than background patterns.

Conclusions In conclusion, RMD-DETR provides an effective balance between accuracy and speed for steel surface defect detection, achieving enhanced robustness and cross-scale performance across multiple industrial datasets, and it is suitable for practical quality inspection systems requiring high throughput and reliable detection. Future work will explore further compression for extremely resource-constrained edge devices and deployment optimization through quantization and acceleration tools to improve engineering applicability.

Keywords: surface defect detection; reparameterization; DMSFPN; AIFI-TSSA; lightweight model

Jiang H J, Xiao X L, Yan X. RMD-DETR: an improved steel surface defect detection algorithm[J]. *Opto-Electron Eng*, 2026, **53**(1): 250181; DOI: 10.12086/oee.2026.250181

Opto-Electronic Journals Group
www.oejournal.org

扫描二维码，获取 PDF 全文