基于 YOLOv8 的钢材表面缺陷检测算法

刘 昱,彭 龑
四川轻化工大学 计算机科学与工程学院,四川 宜宾 644002

摘 要:目的 针对现有钢材表面缺陷检测算法检测精度不足和模型复杂度高的问题,提出一种基于 YOLOv8 的改进钢材表面缺陷检测算法,命名为 YOLOv8-RDP。**方法** 首先,引入 RepNCSPELAN4 模块替换 YOLOv8n 模型中的 C2f 模块,通过并行处理不同尺度的特征,并在最终的卷积层中融合这些特征来优化模型的特征提取和融合能力;其次,在骨干网络部位集成 DA(Deformable Attention)注意力机制,通过自适应调整卷积核采样点,增强模型对不同形状和大小物体的特征捕捉能力,从而提高对关键信息的捕捉效率;最后,为减少模型所需的计算资源,结合 PConv(Partial Convolution)模块改进基线模型检测头,利用特征图中的冗余性,根据数据是否缺失动态调整卷积核的作用区域,以减少计算量。**结果** 在 NEU-DET 数据集上的实验结果表明,YOLOv8-RDP 的 mAP 达到了 78.8%,较基线模型提升了 1.8%;参数量减少至 1.87 M,GFLOPs 降至 3.5 G,分别比基线模型降低了 37.9%和 57.0%。改进后的模型在保持高精度的同时,大幅度减少了计算资源的需求。**结论** YOLOv8-RDP 算法在钢材表面缺陷检测中表现出较高的检测精度和较低的模型复杂度,对钢材表面缺陷检测具有一定的应用价值。

关键词:缺陷检测;特征融合;可变形注意力;轻量化检测头
中图分类号:TG115;TP391.41 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2026.0003.004

A Steel Surface Defect Detection Algorithm Based on YOLOv8

LIU Yu, PENG Yan
School of Computer Science and Engineering, Sichuan University of Science & Engineering, Yibin 644002, Sichuan, China

Abstract: Objective To address the problems of insufficient detection accuracy and high model complexity in existing steel surface defect detection algorithms, an improved steel surface defect detection algorithm based on YOLOv8, named YOLOv8-RDP, is proposed. **Methods** First, the RepNCSPELAN4 module was introduced to replace the C2f module in the YOLOv8n model. The feature extraction and integration capability of the model was optimized through parallel processing of features at different scales and their fusion in the final convolution layer. Second, a deformable attention (DA) mechanism was integrated into the backbone network. This mechanism adaptively adjusted the sampling points of convolutional kernels, strengthening the model' s ability to capture features of objects with varying shapes and sizes and thereby improving the efficiency of capturing key information. Finally, to reduce the computational resources required by the model, the detection head of the baseline model was modified using the partial convolution (Pconv) module. By leveraging redundancy in feature maps and dynamically adjusting the active area of convolutional kernels based on data availability, the computational load was reduced. **Results** Experimental results on the NEU-DET dataset show that YOLOv8-RDP achieved an mAP of 78.8%, representing a 1.8% improvement over the baseline model. The number of parameters was reduced to 1.87 M and GFLOPs dropped to 3.5 G, representing decreases of 37.9% and 57.0%

收稿日期:2024-07-12 **修回日期:**2024-11-18 **文章编号:**1672-058X(2026)03-0030-08
基金项目:四川省科技厅项目资助(2019FYG0377).
作者简介:刘昱(1998—),男,四川巴中人,硕士研究生,从事目标检测研究. Email:1477831142@qq.com.
通信作者:彭龑(1967—),男, 湖南武冈人,教授,博士,从事计算机应用研究. Email:2634932795@qq.com.
引用格式:刘昱,彭龑. 基于 YOLOv8 的钢材表面缺陷检测算法[J]. 重庆工商大学学报(自然科学版),2026,43(3):30-37.
Liu Yu, Peng Yan. A steel surface defect detection algorithm based on YOLOv8[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2026, 43(3): 30-37.

compared with the baseline model, respectively. The improved model significantly reduced the demand for computational resources while maintaining high accuracy. **Conclusion** The YOLOv8-RDP algorithm demonstrates high detection accuracy and low model complexity in steel surface defect detection, showing strong potential for practical application in this field.

Keywords: defect detection; feature fusion; deformable attention; lightweight detection head

钢材是现代工业和建筑领域中不可或缺的结构材料,其质量影响到工程的安全性和可靠性。钢材表面的缺陷,如裂纹、夹杂、斑块、麻点表面、轧制鳞片、划痕等,可能导致结构弱点和降低耐久性,进而引发安全隐患。因此,检测钢材表面缺陷成为钢材生产过程中的重要环节。传统的钢材表面缺陷检测算法包括人工检测、无损检测和机器视觉检测[1-3]。人工检测虽然能够进行直观的缺陷识别,但需要大量的检测人员和时间,效率较低,而且容易受到检测人员主观因素的影响,导致检测结果不稳定。无损检测方法包括超声波检测、磁粉检测、渗透检测等,这些技术能够在不损害材料的情况下检测表面的缺陷。然而,这些方法通常需要专业设备和操作人员,检测成本较高。机器视觉检测利用图像处理算法提取图像信息,并根据获得的特征信息来检测缺陷。相比于人工检测,机器视觉检测具有更高的效率;相比于无损检测,机器视觉检测成本更低。

目前,钢材表面缺陷检测算法有 R-CNN(Regions with Convolutional Neural Networks)、SSD(Single Shot MultiBox Detector)、EfficientDet (Efficient Detection)、Swin Transformer 以及 YOLO(You Only Look Once)系列等[4-8]。R-CNN 系列通过提取候选区域并对其进行分类来实现目标检测,但可能会产生一些不准确的候选区域,导致误检或漏检。SSD 能够同时预测图像中的目标位置和类别,但在处理小目标时检测精度不高。EfficientDet 基于 EfficientNet 架构,并引入加权双向特征金字塔网络和复合缩放方法,但需要较长的训练时间来达到较高的准确性。Swin Transformer 引入基于移动窗口的自注意力机制和层级化特征表达方式,有效处理了不同尺度的视觉信息,并降低了计算需求,但训练复杂度高。YOLO 算法则是一种端到端的实时目标检测算法,具备较高的检测速度和准确性。近年来,钢材表面缺陷检测算法得到了快速发展。戴林华等[9]提出一种改进的 YOLOv8 算法 YOLOv8-SSDW,通过在 YOLOv8n 的骨干网络中集成 SKNet(Selective Kernel Network),增强网络的特征提取能力和自适应性。在颈部网络采用 Slim-Neck 结构,减少了模型的参数量,降低了计算成本。再引入可变形卷积,进一步增强网络对不同形状和大小缺陷的识别能力。刘毅等[10]提出一种改进的 YOLOX 钢材表面缺陷检测算法,在 Backbone 部分引入改进的 SE(Squeeze-and-Excitation)注意力机制,并增加最大池化层分支以强化重要特征通道;在 Neck 部分,引入 ASFF(Adaptive Spatial Feature Fusion)模块,更好地融合不同尺度的特征,并通过将 IOU 损失函数替换为 EIOU(Efficient Intersection over Union)损失函数,来改善模型的定位精度。黄硕清等[11]基于 RFB(Receptive Field Block)和 YOLOv5 提出一种新的钢材缺陷检测方法 RFB-YOLOv5-E,通过增加梯度流分支、下采样层和检测头,显著提高了算法的检测精度。宋世奇等[12]基于 YOLOv8 模型,引入新颖的 EP(Expectation Propagation)模块,并融合大分离卷积注意力模块和空间金字塔池化模块,提出了 SPPF-LSKA 模块,该模块降低了计算复杂度,加快了模型运算速度,提高了检测精度。赵倩等[13]提出一种基于 YOLOv8 的 LMS-YOLO 模型,将轻量级多尺度混合卷积模块与 C2f 进行融合,得到 C2f _ LMSMC,从而提取不同尺度的特征进行融合,得到轻量级网络。Li 等[14]在 YOLOv8s 基础上提出 WFRE-YOLOv8s 模型,设计了 CFN(Convolutional Feature Network)模块来代替 C2f 模块,以减少网络的参数量和 FLOPs;引入 RFN(Receptive Field Network)模块作为新的颈部设计,以减少计算开销,同时更好地融合不同尺度的特征;将 EMA(Exponential Moving Average)注意力模块融入主干网络,以增强有价值特征的提取能力,提高模型的检测精度。

综上所述,以上研究方法通过多尺度特征融合、引入注意力机制和优化损失函数等技术,提升了模型的检测精度,但在处理复杂背景和小目标检测方面仍存在不足。同时,通过轻量化设计减少了模型参数和计算资源需求,但在检测小目标样本时的精度有待提高。为了解决上述问题,本文基于 YOLOv8 提出一种改进的钢材表面缺陷检测算法 YOLOv8-RDP。主要贡献如下:

(1) 引入 RepNCSPELAN4 模块,通过识别并解决重复梯度信息,采用高效的梯度传播路径来平衡基线模型性能与计算资源不平衡的问题。

(2) 引入 DA 模块,使得模型可以动态地调整形状和大小,以更好地适应不同任务和输入数据的特点,增强了模型的性能。

(3) 结合 PConv 模块改进检测头,轻量化了模型。

1　改进算法

1.1　基线 YOLOv8n 模型

YOLOv8 是 YOLO 系列的第八代模型,包括 YOLOv8n、YOLOv8s、YOLOv8m、YOLOv8l 和 YOLOv8x5 个模型。该网络模型由 Backbone、Neck、Head 3 部分组

成。YOLOv8 是基于 YOLOv5 进行优化,在 Backbone 和 Neck 部分,两者都使用了 CSP(Cross Stage Partial)梯度分流思想,且都使用了 SPPF(Spatial Pyramid Pooling with Features)模块,不同的是 YOLOv8 使用梯度流更丰富的 C2f 结构,对不同尺度模型调整了不同的通道数。在 Head 部分,将之前的耦合头结构换成了目前主流的解耦头结构,将分类和检测头分离,同时也从 Anchor-Based 换成了 Anchor-Free,使得模型的性能有了进一步提升。YOLOv8n 是 YOLOv8 系列中的轻量级模型,具有较低的计算复杂度,更适合实时应用和资源受限的环境。相比于其他更大规模的模型如 YOLOv8s、YOLOv8m 等,YOLOv8n 在保持良好性能的同时,具有更快的推理速度和更少的计算资源需求。因此,本文选择 YOLOv8n 进行改进。基线 YOLOv8n 网络结构如图 1 所示。

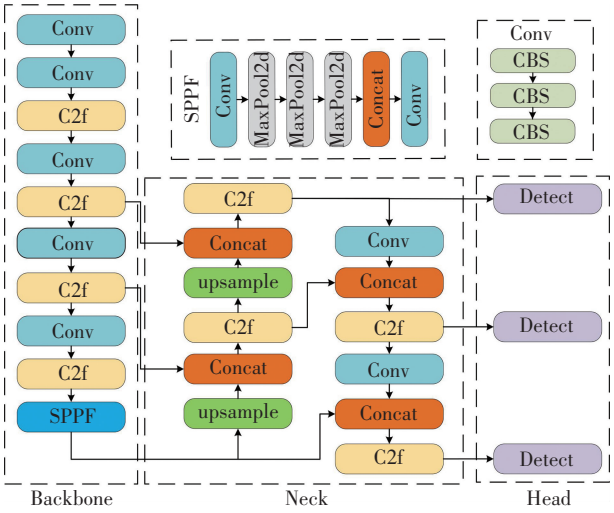


图 1 YOLOv8n 网络结构图

Fig. 1 YOLOv8n network structure diagram

1.2 YOLOv8-RDP 模型

由于基线模型存在检测精度不高,使用计算资源较多等问题,本文提出一种基于 YOLOv8-RDP 的钢材表面缺陷检测算法,主要从 3 个方面对基线模型进行改进。首先,使用 RepNCSPELAN4 替换基线模型的 C2f 模块,强化模型的特征提取能力;其次,为了提高模型的精确度,引入 DA 注意力机制;最后,对基线模型的检测头进行改进。YOLOv8-RDP 的网络结构如图 2 所示。

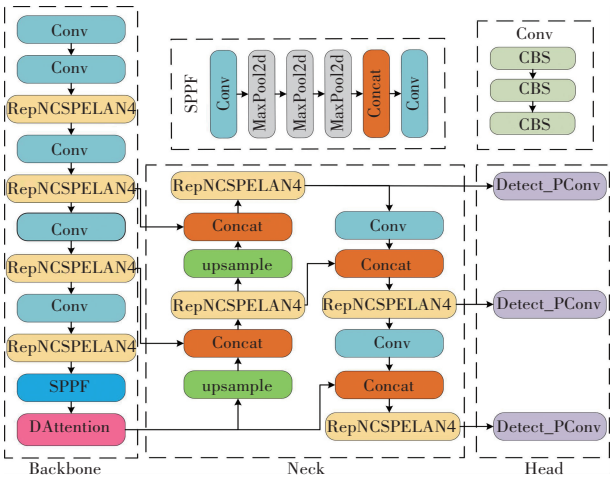
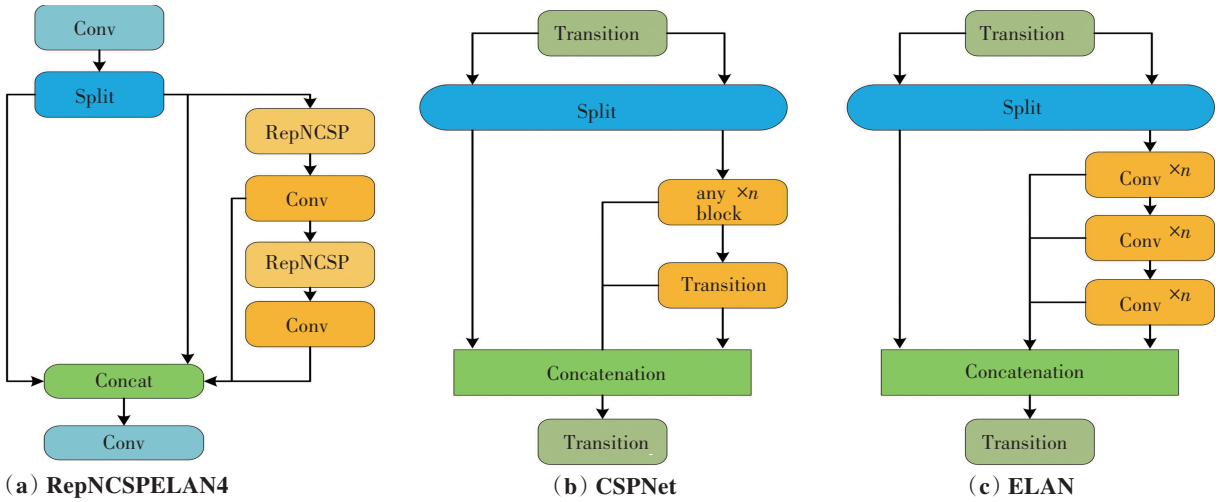


图 2 YOLOv8-RDP 网络结构图

Fig. 2 YOLOv8-RDP network structure diagram

1.3 RepNCSPELAN4 模块

YOLOv8 中使用了 C2f 模块,增强了模型对细节和语义信息的捕捉能力,能够提高检测的准确性和鲁棒性。然而,特征融合操作增加了模型的计算复杂度,使得模型不能很好地平衡性能与计算资源之间的关系。因此,本文将基线模型中的 C2f 模块替换为 RepNCSPELAN4[15] 模块,RepNCSPELAN4 模块是一个特征提取和融合模块,它结合了 CSPNet(Cross Stage Partial Network)和 ELAN(Efficient Layer Aggregation Network)的优点,通过增强神经网络的特征学习能力和梯度路径规划,能在使用较少计算量的情况下提高模型性能。RepNCSPELAN4 的网络结构如图 3(a)所示,CSPNet 和 ELAN 的网络结构如图 3(b)、图 3(c)所示。



(a) RepNCSPELAN4 (b) CSPNet (c) ELAN

图 3 RepNCSPELAN4、CSPNet 和 ELAN 结构图

Fig. 3 Structure diagram of RepNCSPELAN4, CSPNet, and ELAN

CSPNet[16]是一种 CNN(Convolutional Neural Network)骨干网络,它通过创新跨阶段部分网络结构设计来增强 CNN 的学习能力,并减少计算量和内存成本。CSPNet 的核心思想是解决网络优化中的重复梯度信息问题,通过在网络不同阶段整合特征图,以增加梯度信息的多样性,从而提高网络的学习能力 。CSPNet 的结构特点是局部稠密块(Partial Dense Block)和局部过渡层(Partial Transition Layer),这些设计有助于增加梯度路径,平衡每层的计算量,以此减少内存流量。

ELAN[17]是一种高效的神经网络结构,专注于解决深层模型在训练过程中的梯度传播效率问题。它通过细致分析网络中每层的最短和最长梯度路径,采用高效的梯度传播路径进行层级聚合,避免了传统网络因过渡层过多而导致的梯度路径延长问题。ELAN 的设计允许模型灵活地在准确度和计算量之间做出权衡,通过计算块内的堆叠策略,保持了梯度路径的有效性,即便在网络深度增加时也能保证网络的收敛。

1.4　DA 注意力机制

注意力机制通过聚焦输入数据中的关键部分,帮助模型更精确地捕捉完成任务所需的重要信息,从而提高模型在各种任务上的性能。传统注意力机制通常采用固定的权重计算方式,这会导致模型在处理不同任务时缺乏灵活性。然而,小目标图像的特征非常细微,检测起来更加困难。针对这类缺陷的检测,模型更需要具备对细微特征的敏感性。因此,为了使模型更加关注钢材缺陷检测中的细微特征,在基线模型中引入 DA[18]注意力。可变形注意力(DA)是一种用于神经网络中的注意力机制。在传统注意力机制中,权重是通过对位置固定的注意力模型进行计算得到的。而在可变形注意力中,模型可以动态地调整注意力模型的形状和大小,能够更有效地捕捉小目标的特征,从而提高检测性能。DA 注意力包括以下几个部分,如图 4 所示。

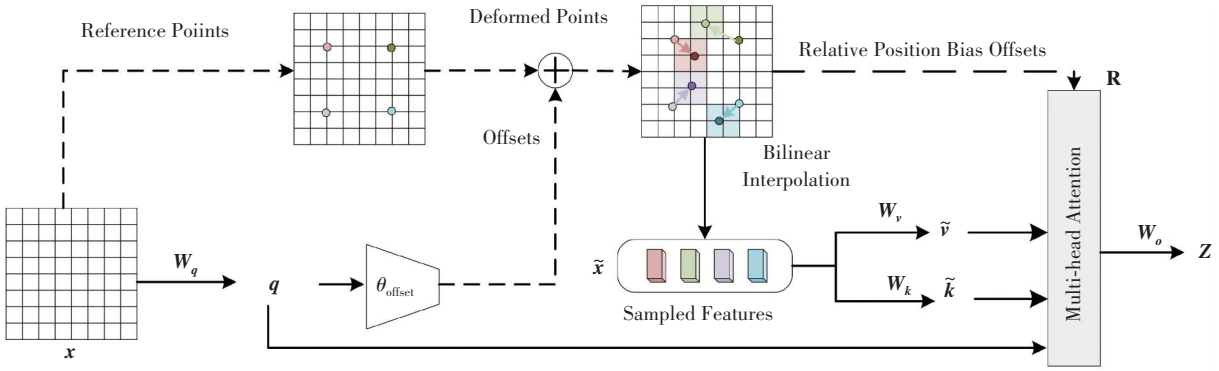


图 4　可变形注意力 DA

Fig. 4　Deformable attention DA

(1) 参考点生成。如图4 所示,基于给定的特征输入图 $\boldsymbol{x}\in\mathbf{R}^{H\times W\times C}$ 生成一个参考网格,其中参考点 $\boldsymbol{p}\in\mathbf{R}^{H_G\times W_G\times 2}$。网格尺寸由输入特征图 $\boldsymbol{x}$ 降采样而来,降采样系数为 r。则查询(Query)、键(Key)、值(Value)的生成方式为

$$\boldsymbol{q}=\boldsymbol{x}\boldsymbol{W}_q \tag{1}$$

$$\tilde{\boldsymbol{k}}=\tilde{\boldsymbol{x}}\boldsymbol{W}_k \tag{2}$$

$$\tilde{\boldsymbol{v}}=\tilde{\boldsymbol{x}}\boldsymbol{W}_v \tag{3}$$

其中,$\boldsymbol{x}$ 是输入的特征矩阵,$\boldsymbol{W}_q$、$\boldsymbol{W}_k$、$\boldsymbol{W}_v$ 是可学习的投影矩阵。

(2) 偏移量计算。为了获得每个参考点的偏移量,将特征图线性投影到 $\boldsymbol{q}=\boldsymbol{x}\boldsymbol{W}_q$,然后送入一个轻量级的子网络 $\theta_{\text{offset}}(\cdot)$ 以生成偏移量 $\Delta\boldsymbol{p}=\theta_{\text{offset}}(\boldsymbol{q})$。输入特征首先通过一个 5×5 的深度卷积层以捕获局部特征;然后,使用 GELU(Gaussian Error Inear Units)[19]激活函数对卷积后的特征进行非线性变换;最后,通过一个 1×1 的卷积层来获取 2D 偏移量。即

$$\Delta\boldsymbol{p}=\theta_{\text{offset}}(\boldsymbol{q}) \tag{4}$$

$$\tilde{\boldsymbol{x}}=\varphi(\boldsymbol{x};\boldsymbol{p}+\Delta\boldsymbol{p}) \tag{5}$$

(3) 可变形采样。使用偏移量 $\Delta\boldsymbol{p}$ 和参考点 $\boldsymbol{p}$ 计算变形点的位置,并将这些位置的采样特征作为键和值与查询一同传入多头注意力机制中,即

$$\varphi(z;(p_x,p_y))=\sum_{(r_x,r_y)}g(p_x,r_x)g(p_y,r_y)z[r_y,r_x,:] \tag{6}$$

其中,φ 表示双线性插值采样函数,g 是定义在 z 上的高斯核函数。

(4) 特征连接。每个头部的特征连接在一起,通过 $\boldsymbol{W}_o$ 投影得到最终输出 $\boldsymbol{z}$。

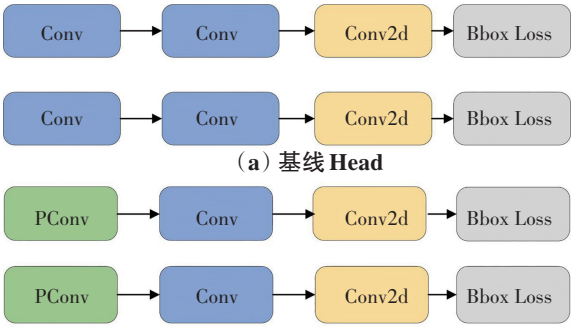
$$\boldsymbol{z}=\text{Concat}(z^{(1)},\cdots,z^{(M)})\boldsymbol{W}_o \tag{7}$$

其中,$z^{(M)}$ 为计算第 m 个注意力头的输出。

1.5　Pconv_Head 检测头

检测头是目标检测模型中的关键组件,负责从特征图中提取并识别目标的位置和类别。然而,检测头需要处理高维特征图,进行多尺度和多任务的复杂预

测,并且常涉及大量的卷积层和后处理步骤,使得检测头所需要的计算资源较多。为了减少模型参数,降低计算复杂度,通过引入 PConv [20]卷积来对检测头进行轻量化改进。基线模型检测头如图 5(a)所示,PConv_Head 检测头如图 5 (b)所示。



(a) 基线 Head

(b) PConv_Head

图 5 基线 Head 和 PConv_Head 结构图

Fig. 5 Structure diagrams of baseline Head and PConv_Head

普通卷积通过卷积核提取输入数据的特征。每个卷积核在输入数据上滑动,计算卷积核和覆盖的输入区域间的点积。普通卷积的 FLOPs 为

$$h\times w\times k^2\times c^2 \tag{8}$$

其中,h 表示特征图的高度,w 表示特征图的宽度,k 表示卷积核的尺寸,c 表示通道数。

部分卷积只对输入通道的一部分应用正则卷积进行空间特征提取,其余通道保持不变,能够减少计算冗余和内存访问。对于连续的内存访问,选取前段或后段连续 c_p 个通道,作为整个特征图的代表进行计算。假设输入端和输出端的通道数一样,则 PConv 的 FLOPs 为

$$h\times w\times k^2\times c_p^2 \tag{9}$$

内存访问量是衡量模型运行时,对内存的读取和写入操作次数的一个指标。内存访问量对于评估模型的效率和性能至关重要。普通卷积的内存访问量为

$$h\times w\times 2c+k^2\times c^2\approx h\times w\times 2c \tag{10}$$

PConv 的内存访问量为

$$h\times w\times 2c_p+k^2\times c_p^2\approx h\times w\times 2c_p \tag{11}$$

2 实验与结果分析

2.1 数据集

本实验以钢材表面缺陷检测的公开数据集 NEU-DET 为实验对象。该数据集涵盖了常见的 6 种钢材表面缺陷,分别为裂纹(Crazing,Cr)、夹杂(Inclusion,In)、斑块(Patches,Pa)、麻点表面(Pitted surface,Ps)、轧制鳞片(Rolled-in scale,Rs)、划痕(Scratches,Sc)。每种缺陷类型都包含 300 张分辨率为 200×200 的灰度样本,共计 1 800 张图像。数据集随机分配为训练集、验证集和测试集,比例为 8∶1∶1。部分图像样本如图 6 所示。

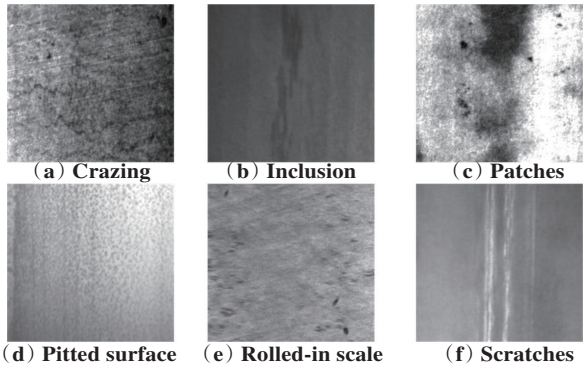


图 6 部分图像样本

Fig. 6 Samples from the dataset

2.2 实验环境

本实验使用 Python-3.8.17 进行编程,torch 版本为 2.1.2,操作系统为 Windows10,训练参数设置为 imgsz 为 640,epochs 为 300,batch 为 16,在最后 10 轮关闭 mosaic 数据增强,使用 SGD 优化器优化网络参数,初始学习率为 0.01,动量参数设为 0.937,权重衰减系数为 0.000 5。实验环境如表 1 所示。

表 1 实验环境表

Table 1 Experimental environment

名 称	配 置
操作系统	Windows 11
CPU	Intel(R) Core(TM) i7-12700H
GPU	RTX 3070 Ti Laptop GPU
Python	3.8.17
内存	16 GB

2.3 评价指标

本文从精确率 P(Precision)、召回率 R(Recall)、平均准确率均值 f_{mAP}(mean Average Precision)、参数量(Params)和计算量(GFLOPs)5 个指标来评估模型的性能。召回率和精确率是二元分类任务中的关键评估指标,它们有助于评估模型的预测准确性。平均准确率均值综合考虑了不同类别的精确度-召回率曲线,反映了模型在不同类别上的检测准确度和检测率。使用平均准确率均值来衡量模型多类别检测的性能。参数量指的是深度学习模型中的可学习参数数量,它是衡量模型复杂度的重要指标。计算量是衡量深度学习模型计算复杂度的指标,表示每秒浮点运算次数的数量。这些指标共同构成了评价模型综合性能的标准。

$$P=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}} \tag{12}$$

$$R=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}} \tag{13}$$

$$f_{\mathrm{mAP}}=\frac{1}{n}\sum_{i=0}^{n}\mathrm{AP}(i) \tag{14}$$

其中,N_{TP}、N_{FP} 和 N_{FN} 分别代表真正例、假正例和假负例。

2.4　消融实验

为了研究模型各个组成部分对整体性能的影响,本文以 YOLOv8n 为基线模型,设计了 7 组消融实验,并将 RepNCSPELAN4 模块记为 R,DA 模块记为 D,Pconv-Head 模块记为 P。包括 YOLOv8-R:只添加 RepNCSPELAN4 模块;YOLOv8-D:只添加 DA 模块;YOLOv8-P:只添加 Pconv-Head 模块;YOLOv8-RD:综合 RepNCSPELAN4、DA 模块;YOLOv8-RP:综合 RepNCSPELAN4、Pconv-Head 模块;YOLOv8-DP:综合 DA、Pconv-Head 模块;YOLOv8-RDP:综合 RepNCSPELAN4、DA、Pconv-Head 模块。消融实验结果如表 2 所示。

表 2　消融实验

Table 2　Ablation study

模　型	R	D	P	*R*/%	*P*/%	mAP/%	Params/MB	GFLOPs
YOLOv8n				84.3	67.7	77.0	3.01	8.1
YOLOv8-R	✓			70.9	73.1	78.2	2.19	5.9
YOLOv8-D		✓		79.3	71.4	78.8	3.27	8.3
YOLOv8-P			✓	76.8	69.6	77.1	2.42	5.5
YOLOv8-RD	✓	✓		71.7	76.1	79.2	2.46	6.2
YOLOv8-RP	✓		✓	76.0	71.1	77.9	1.60	3.3
YOLOv8-DP		✓	✓	79.7	69.0	78.2	2.69	5.8
YOLOv8-RDP	✓	✓	✓	71.8	72.3	78.8	1.87	3.5

由表 2 可知,当添加 RepNCSPELAN4 模块(YOLOv8-R)后,mAP 提升到 78.2%,参数量减少到 2.19 M,GFLOPs 减少到 5.9。这表明 RepNCSPELAN4 模块在提高模型性能的同时还减少了计算资源的需求。添加 DA 模块(YOLOv8-D)后,mAP 提升到 78.8%,但参数量增加到 3.27 MB,GFLOPs 也增加到 8.3。这说明 DA 模块对性能有正面影响,但增加了一些计算资源的使用。引入 Pconv-Head 模块(YOLOv8-P)后,mAP 几乎不变,但参数量显著减少到 2.42 M,GFLOPs 也减少到 5.5。这表明 Pconv-Head 模块在减少模型复杂度方面有积极作用。RepNCSPELAN4 和 DA 的组合有最高的精度,但计算资源相对 RepNCSPELAN4 和 Pconv-Head 的组合使用更多。综合考虑,YOLOv8-RDP 在保持 78.8% mAP 的同时,参数量低至 1.87 MB,GFLOPs 低至 3.5,展现了最优的性能-效率平衡。

2.5　对比实验

2.5.1　与其他注意力改进方法的效果对比

为了探究不同注意力对模型性能的影响,本文将在基线模型的基础上分别添加 DAttention、SimAM 、CPCA(Channel Prior Convolutional Attention)、CAFM(Convolution and Attention Fusion Module)、LocalWindowAttention 注意力机制进行对比实验。实验结果如表 3 所示。

表 3　改进注意力对比实验

Table 3　Comparative experiment of improved attention mechanisms

模　型	Cr	In	Pa	Ps	Rs	Sc	mAP/%	Params/MB	GFLOPs
YOLOv8n	42.7	82.3	91.2	84.2	66.6	95.1	77.0	3.01	8.1
DAttention	44.6	88.2	88.9	88.3	67.3	95.5	**78.8**	3.27	8.3
SimAM	48.7	83.7	89.6	85.1	58.7	94.4	**76.7**	**3.01**	**8.1**
CPCA	45.4	81.2	91.4	88.9	58.1	94.9	**76.7**	**3.13**	**8.3**
CAFM	40.8	85.3	88.9	86.3	65.7	96.0	77.2	3.35	8.4
LocalWindowAttention	38.5	81.2	88.8	82.1	65.1	96.6	75.4	3.10	8.2

由表 3 可知,DAttention 方法在提升模型性能方面表现最为出色,同时保持了较低的参数量和计算资源消耗。LocalWindowAttention 在所有比较中表现最差,无论是在性能还是效率上均未达到预期效果。其中,SimAM 对 Cr 缺陷的 mAP 值提升最大,提升了 6.0%;DAttention 对 In 和 Rs 缺陷的 mAP 值提升最大,分别提升了 5.9% 和 0.7%;CPCA 对 Pa 和 Ps 缺陷的 mAP 值提升最大,分别提升了 0.2% 和 4.7%;LocalWindowAttention 对 Sc 缺陷的 mAP 值提升最大,提升了 1.5%。综合分析实验结果可知,DAttention 对 5 类缺陷的 mAP 值都有提升的效果,

具有更优异的性能。

2.5.2 与其他检测头改进方法的效果对比

为了探究使用不同模块改进检测头对模型性能的影响,本文将分别使用 PConv、RepConv、DBBlock 等模块替换基线模型检测头的 Conv。实验结果如表 4 所示。

表 4 改进检测头对比实验

Table 4 Comparative experiment of improved detection heads

模 型	mAP/%	Params/MB	GFLOPs
YOLOv8n	77.0	3.01	8.1
PConv-Head	77.1	2.42	5.5
RepConv-Head	76.9	4.10	8.1
DBBlock-Head	77.1	3.84	8.1

综合考虑性能和计算资源的平衡时,PConv-Head 使用了最少的计算资源,同时在检测精度上有一定的提升。DBBlock-Head 在保持性能的同时轻微减少了参数量,但计算资源的减少不明显。RepConv-Head 的各项指标都最差。因此,PConv-Head 是最佳的选择。

2.5.3 泛化性验证

为了评估模型的泛化性,使用基线模型与 YOLOv8-RDP 模型在 GC-DET 数据集上进行泛化性实验。工业表面缺陷数据集 GC-DET 包含冲孔、焊缝、新月形缝隙、水斑、油斑、丝斑、夹杂物、轧坑、折痕、腰部折痕等 10 种类型的表面缺陷。在 GC-DET 数据集上收集了 2 294 张图片,数据集随机分配为训练集、验证集和测试集,比例为 8∶1∶1,实验环境与第 3.2 节一致。实验结果如表 5 所示。

表 5 在 GC-DET 数据集上的泛化性实验

Table 5 Generalization experiment on GC-DET dataset

模 型	mAP/%	Params/MB	GFLOPs
YOLOv8n	63.0	3.01	8.1
YOLOv8-RDP	66.1	1.87	3.5

由表 5 可知,在 GC-DET 数据集上 YOLOv8-RDP 的 mAP 值较基线模型提升了 3.1%,参数量较基线模型减少了 1.14 MB,计算量较基线模型减少了 4.6 GFLOPs。实验结果表明,本文提出的改进算法 YOLOv8-RDP 具有良好的泛化性。

2.5.4 与其他经典算法效果对比

将 YOLOv8-RDP 算法与部分经典算法相比较,分别引入 YOLOv5、YOLOv8s 、FCOS 、YOLOv10n 、ATSS (Adaptive Training Sample Selection)、CenterNet、Faster-RCNN、Mask-RCNN 算法在 NEU-DET 数据集进行实验,实验结果如表 6 所示。

表 6 不同模型对比实验

Table 6 Comparative experiments of different models

模 型	mAP/%	Params/MB	GFLOPs
YOLOv5	76.6	7.03	15.8
YOLOv8n	77.0	3.01	8.1
YOLOv8s	77.5	11.13	28.4
YOLOv10n	75.6	2.27	6.5
FCOS	69.7	32.29	128.0
ATSS	69.1	32.1	129.0
CenterNet	67.2	32.1	126.0
Faster-RCNN	76.7	170.0	282.6
Mask-RCNN	78.9	44.4	189.4
YOLOv8-RDP	78.8	1.87	3.5

由表 6 可知,本文所提算法 YOLOv8-RDP 在参数量、计算量的指标中均达到了最优,精确度也仅比 Mask-RCNN 低 0.1%。与基线模型相比,YOLOv8-RDP 的精确度上升了 1.8%,参数量下降了 1.14 MB,计算量下降了 4.6 G。与其他模型相比,具有更优异的表现。相比于 CenterNet,精确度上升了 11.6%;相比于 Faster-RCNN,参数量仅为其 1.1%,计算量仅为其 1.2%;与 Mask-RCNN 相比,虽然精确度下降了 0.1%,但是 YOLOv8-RDP 的参数量和计算量都远低于 Mask-RCNN。与同类型的 YOLOv5、YOLOv8s 和 YOLOv10n 相比,本文所提算法 YOLOv8-RDP 在精确度、参数量和计算量的指标中均达到了最优。综上所述,本文所提算法 YOLOv8-RDP 具有精度高、轻量化的优势,在计算资源有限的设备上具有更佳的部署优势。

3 结 论

根据实际情况中钢材表面缺陷检测中存在的问题,本文基于 YOLOv8n 提出一种改进算法:YOLOv8-RDP。通过引入 RepNCSPELAN4 模块,增强了模型的特征提取和融合能力;通过引入 DA 模块,增强了模型对关键信息的捕捉能力,提升了模型的精确度;通过改进基线模型的检测头,减少了模型的计算冗余。在 NEU-DET 数据集的实验表明:YOLOv8-RDP 的精确度较基线模型提升了 1.8%且参数量和计算量大幅度降低,节省了许多计算资源。与主流算法相比,本研究提出的 YOLOv8-RDP 在保持高检测精度的同时,所需的计算资源更少,速度更快,更有利于部署到计算资源有限的终端设备中,在钢材表面缺陷检测领域有一定的借鉴意义。

但本文仍有许多方面值得进一步研究。尤其是模型对裂纹和轧制鳞片的检测精度相较于其他几类缺陷要低得多,后续考虑使用数据增强技术和更高效的注意力算法提升模型对这两类缺陷的检测精度,进一步提升模型的整体精确度。

参考文献(References):

[1] 盖晨阳, 赵德颖, 李海滨. PP-LCyolov5s: 一种轻量化的钢板目标表面检测算法[J/OL]. 机械科学与技术. https://doi.org/10.13433/j.cnki.1003-8728.20240091.
GAI Chen-yang, ZHAO De-ying, LI Hai-bin. PP-LCyolov5s: a lightweight detection algorithm for steel plate target surface[J/OL]. Mechanical Science and Technology for Aerospace Engineering. https://doi.org/10.13433/j.cnki.1003-8728.20240091.

[2] 单攀蓉, 蒋玉梅, 张才杰, 等. 钢材生产中质量问题的溯源与控制[J]. 冶金与材料, 2023, 43(7): 154-156.
SHAN Pan-rong, JIANG Yu-mei, ZHANG Cai-jie, et al. Traceability and control of quality problems in steel production[J]. Metallurgy and Materials, 2023, 43(7): 154-156.

[3] 宋育斌, 孔维宾, 陈希, 等. 钢材表面缺陷检测研究综述[J]. 软件导刊, 2024, 23(3): 203-211.
SONG Yu-bin, KONG Wei-bin, CHEN Xi, et al. Survey of steel surface defect detection research[J]. Software Guide, 2024, 23(3): 203-211.

[4] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 580-587.

[5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 21-37.

[6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 779-788.

[7] JIANG P, ERGU D, LIU F, et al. A review of YOLO algorithm developments[J]. Procedia Computer Science, 2022, 199: 1066-1073.

[8] WANG C Y, BOCHKOVSKIY A, LIAO H M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 7464-7475.

[9] 戴林华, 黎远松, 石睿. YOLOv8-SSDW: 基于 YOLOv8 的带钢表面缺陷检测算法[J]. 重庆工商大学学报(自然科学版), 2025, 42(4): 44-52.
DAI Lin-hua, LI Yuan-song, SHI Rui. YOLOv8-SSDW: a steel surface defect detection algorithm based on YOLOv8[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(4): 44-52.

[10] 刘毅, 蒋三新. 基于改进 YOLOX 的钢材表面缺陷检测研究[J]. 现代电子技术, 2024, 47(9): 131-138.
LIU Yi, JIANG San-xin. Steel surface defect detection algorithm based on improved YOLOX[J]. Modern Electronics Technique, 2024, 47(9): 131-138.

[11] 黄硕清, 黄金贵. 基于 RFB 和 YOLOv5 特征增强融合改进的钢材表面缺陷检测方法[J]. 计算机工程, 2025, 51(4): 249-260.
HUANG Shuo-qing, HUANG Jin-gui. Improved steel defect detection method based on enhanced fusion of RFB and YOLOv5 features[J]. Computer Engineering, 2025, 51(4): 249-260.

[12] ZHANG X, WANG Y, FANG H. Steel surface defect detection algorithm based on ESI-YOLOv8[J]. Materials Research Express, 2024, 11(5): 056509.

[13] XIE W, SUN X, MA W. A light weight multi-scale feature fusion steel surface defect detection model based on YOLOv8[J]. Measurement Science and Technology, 2024, 35(5): 055017.

[14] HUANG Y, TAN W, LI L, et al. WFRE-YOLOv8s: a new type of defect detector for steel surfaces[J]. Coatings, 2023, 13(12): 2011.

[15] AIBIBU T, LAN J, ZENG Y, et al. Feature-enhanced attention and dual-GELAN net (FEADG-net) for UAV infrared small object detection in traffic surveillance[J]. Drones, 2024, 8(7): 304.

[16] WANG C Y, MARK LIAO H Y, WU Y H, et al. CSPNet: a new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2020: 1571-1580.

[17] LIU W, SHEN Z, XU S. CF-YOLO: A capable forest fire identification algorithm founded on YOLOv7 improvement[J]. Signal, Image and Video Processing, 2024, 18(8-9): 6007-6017.

[18] XIA Z, PAN X, SONG S, et al. Vision Transformer with deformable attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 4784-4793.

[19] NGUYEN A, PHAM K, NGO D, et al. An analysis of state-of-the-art activation functions for supervised deep neural network[C]//Proceedings of the International Conference on System Science and Engineering. Piscataway: IEEE Press, 2021: 215-220.

[20] CHEN J, KAO S H, HE H, et al. Run, don't walk: Chasing higher FLOPS for faster neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2023: 12021-12031.

责任编辑:李翠薇