2025 年 第 25 卷 第 16 期
2025, 25(16): 06831-10

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2404413
引用格式:胡名琪, 陈辉明, 徐伟, 等. 融合 MobileNetv3 的轻量级 YOLOv8 钢材表面缺陷检测[J]. 科学技术与工程, 2025, 25(16): 6831-6840.
Hu Mingqi, Chen Huiming, Xu Wei, et al. Surface defect detection on lightweight YOLOv8 steel incorporating MobileNetv3[J]. Science Technology and Engineering, 2025, 25(16): 6831-6840.

融合 MobileNetv3 的轻量级 YOLOv8 钢材表面缺陷检测

胡名琪[1], 陈辉明[1], 徐伟[2], 郭诚君[1], 刘秋明[2*]
(1. 江西理工大学先进铜产业学院, 鹰潭 335000; 2. 江西理工大学软件工程学院, 南昌 330013)

摘 要 针对钢材表面缺陷人工检测成本高昂、检测精度不高,以及传统的目标检测方法模型复杂,导致对终端检测设备的计算资源要求较高等问题,融合 MobileNetv3 轻量化 YOLOv8 算法提出一种轻量级缺陷检测算法 YOLOv8n-MDC。首先,以 YOLOv8n 为基础,将 YOLOv8n 的自带 IoU(intersection over union)候选框损失函数替换成 WIoU(weighted IoU)函数,通过增添非单调聚焦机制,提高模型的鲁棒性。其次,使用 MobileNetv3 网络替换 YOLOv8n 的骨干特征提取网络模块,将轻量级网络用于特征提取端降低网络复杂度,减少冗余开销。最后,在特征融合阶段使用 DW 卷积和 C3Ghost 模块对原网络的相应模块进行替换,使改进后的网络减少模型参数,进一步提升检测速度。使用钢材表面缺陷数据集 NEU-DET 进行模型验证,YOLOv8n-MDC 模型 mAP 达 81.3%,较 YOLOv8n 模型提升 5%;参数量与计算量分别为 1.02 M 和 2.1 GFLOPs,仅为原模型的 33.9% 和 25.9%,达到工业要求。提出的轻量级算法在保证检测精度提升的同时大大降低了算法的复杂度和计算资源的开销,为钢材表面缺陷检测提供了一个优化思路。

关键词 钢材表面缺陷; 缺陷检测; 轻量级网络; YOLOv8; MobileNetv3
中图法分类号 TP391.4; **文献标志码** A

Surface Defect Detection on Lightweight YOLOv8 Steel Incorporating MobileNetv3

HU Ming-qi[1], CHEN Hui-ming[1], XU Wei[2], GUO Cheng-jun[1], LIU Qiu-ming[2*]
(1. Advanced Copper Industry College, Jiangxi University of Science and Technology, Yingtan 335000, China;
2. School of Software Engineering, Jiangxi University of Science and Technology, Nanchang 330013, China)

[**Abstract**] To address the high cost and low accuracy of manual inspection for steel surface defects, as well as the excessive computational resource requirements caused by complex traditional target detection models, a lightweight defect detection algorithm named YOLOv8n-MDC was proposed by integrating MobileNetv3 with YOLOv8. Firstly, based on YOLOv8n, the original intersection over union(IoU)-based bounding box loss function was replaced with weighted IoU(WIoU), enhancing model robustness through a non-monotonic focusing mechanism. Secondly, the backbone feature extraction network of YOLOv8n was substituted with MobileNetv3, utilizing its lightweight architecture to reduce network complexity and redundant computational overhead. Finally, during the feature fusion stage, depthwise separable convolution (DWConv) and C3Ghost modules replaced the original components, further minimizing model parameters and accelerating detection speed. Evaluated on the NEU-DET steel surface defect dataset, the YOLOv8n-MDC achieves an mAP of 81.3%, representing a 5% improvement over the baseline YOLOv8n, while its parameter count and computational complexity are reduced to 1.02 M and 2.1 GFLOPs (33.9% and 25.9% of the original model, respectively), meeting industrial requirements. This lightweight algorithm significantly reduces computational complexity and resource consumption while enhancing detection accuracy, offering an optimized solution for industrial steel surface defect inspection.

[**Keywords**] steel surface defects; defect detection; lightweight networking; YOLOv8; MobileNetv3

钢材表面缺陷检测是钢铁行业中重要的研究领域[1]。随着钢铁产品在现代工业中的广泛应用,对其质量、安全性和成本效益的要求也越来越高[2]。钢材表面缺陷不仅会影响产品的质量和性

收稿日期: 2024-06-13; **修订日期**: 2025-03-04
基金项目: 国家自然科学基金(52361008);江西省省级引导市县科技发展专项资金项目(2021SYD001);江西先进铜产业研究院自立研究课题(ZL 陈辉明-202002);江西省自然科学基金(20242BAB25073)
第一作者: 胡名琪(2001—),女,汉族,江西吉安人,硕士研究生。研究方向:材料工程人工智能。E-mail:1456667454@qq.com。
*通信作者: 刘秋明(1985—),男,汉族,江西南昌人,博士,副教授。研究方向:物联网与智能计算 。E-mail:liuqiuming@jxust.edu.cn。

能,还可能导致安全隐患,因此对钢材缺陷的检测至关重要。然而,传统的钢材表面缺陷检测方法[3]多依赖于昂贵复杂的设备和繁琐的人工操作,效率低下且成本高昂。因此,研究基于轻量级算法[4]的钢材表面缺陷检测具有重要意义,它能够实现高效、精准的自动化检测,提高生产效率,降低成本,同时为工业生产的智能化和数字化发展提供可行的解决方案。

目前主流的目标检测算法包括 Faster R-CNN[5]、YOLO(you only look once)[6]、SSD(single shot multi-box detector)[7]等。首先,Faster R-CNN 包含两个阶段的处理过程[8],需要更多的计算资源和内存空间,相比于 YOLO 更为复杂且在速度和效率上较慢[9];其次,SSD 通过在不同层级上检测目标来实现多尺度的目标检测[10],虽在检测速度方面表现出色,但这种方法无法充分利用图像中的语义信息且在检测精度方面低于 YOLO,导致对小目标的检测效果不佳;最后,虽然 Transformer 模型已经被成功地应用于图像领域[11],如图像分类和图像生成,但在目标检测任务中的表现相对较少,其性能和效率不如专门针对目标检测设计的算法。综上,YOLO 算法具有快速地检测速度,可以实时地对图像进行目标检测,适用于工业生产线上的高效自动化检测需求[12]。然而,由于目标检测算法对终端检测设备的计算资源要求较高,这使得模型部署在终端存在较大的困难,这给钢材表面缺陷检测应用领域带来了巨大挑战。目前,对 YOLO 模型进行轻量化处理、减少其目标检测对终端需要的计算资源、便于钢材表面缺陷检测的实际应用,逐渐成为热点研究方法。

为了降低深度学习算法所需的计算资源同时提高检测精度,以便于模型在目标检测应用终端的部署,王春梅等[13]在 YOLOv8 的基础上,将轻量级的 VanillaNet 模型替换普通卷积,降低了特征提取端的复杂度;然后加入 SPD(space-to-depth)模块,加快算法的计算性能,在钢材表面缺陷检测数据集(Northeastern University detection dataset,NEU-DET)数据集上的平均精度均值(mean average precision,mAP)达到 80.8%,网络的参数量为 1.96 M,计算量为 6.0 GFLOPs(GFLOPs 指每秒可以执行的十亿次浮点运算次数)。崔克彬等[14]为了更好地确定密集目标,改进了 CBAM(convolutional block attention module)注意力机制,同时将 YOLO 的 FPN(feature pyramid network)模块换成 BiFPN(bidirectional feature pyramid network),提出了一种 MCB-FAH-YOLOv8 算法,在 NEU-DET 数据集上的 mAP 达到 81.8%,但牺牲了模型的参数量和计算量。Hao 等[15]首先利用形变卷积增强骨干网络提取特征,然后采用平衡特征金字塔提高了特征融合端的分辨质量,最后通过检测端实现钢材缺陷的定位和分类,提出了一种 DIN(defect inspection network)目标检测模型,在 NEU-DET 数据集上的 mAP 达到 80.5%。Zhou 等[16]提出一种轻量化轧制钢带表面缺陷检测模型 YOLOv5s-GCE,原文使用 Ghost 模块用于替换原 YOLOv5s 模型中一部分的 CBS(convolution-batch-norm-SiLU)结构;新增 EIoU(enhanced intersection over union)函数,提高预测帧回归的准确率,加速其收敛;并且采用 CA(coordinate attention)注意力方法强化关键特征通道及其位置信息,使模型在降低参数量和计算量的同时能够正确识别和找到目标。以上研究阐述了轻量化处理对钢材表面缺陷检测的意义,说明了钢材表面缺陷小目标检测的优化方向。然而在追求轻量化的同时难免会牺牲一些精度,但是过于追求精度又会使轻量化工作难以展开。在钢材表面缺陷检测工业部署中,终端的计算能力是有限的,通常只能处理几 GFLOPS 级别的图像,并且内存极其有限。虽然 YOLO 原型有着部署简单、检测速度快等特点,但在原模型上计算量达 8 ~ 150 GFLOPS,此计算量远超于工业预期部署检测能力,缺陷检测设备难以负担庞大的计算量。

鉴于此,针对提高检测速度、精度和降低部署难度的问题,在 YOLOv8n 网络模型基础上提出一种面向钢材表面缺陷检测的轻量级算法 YOLOv8n-MDC。在 NEU-DET 数据集上进行大量实验,与主流的目标检测算法相比,该算法在提高检测精度和速度的同时所需的计算资源更少。

1 相关工作

1.1 YOLOv8n

YOLOv8 是在 YOLOv5 的基础上改进的,是 YOLO[17]系列的最新版本。该模型采用一系列创新技术和优化,旨在平衡目标检测的准确性和速度。是一种高效的实时目标检测算法。其采用了类似于 Darknet 的网络结构,主要由骨干特征提取端(Backbone)、颈部特征融合端(Neck)和头部检测端(Head)3 个部分组成。

(1)Backbone。负责从输入图像中提取特征,是 YOLOv8 的骨干模块,用来学习检测目标的语义信息和上下文信息。通常包括 Darknet-53 或者其他一些轻量级网络。使用 C2f 模块替换了 YOLOv5 中的 C3 模块,C3 融合了 ResNet 结构和区域跨阶段网络分流的思想,而 C2f 模块利用 ELAN(efficient layer aggregation network)结构并行更多的梯度分流来获取更多的信息,轻量化的同时提高了模型精度。

(2)Neck。负责融合不同层次的特征,提高目标检测的准确性。相比于 YOLOv5,为了压缩模型的同时提高模型的性能,删除了采样层前的卷积层。

(3)Head。负责预测目标的位置和类别信息。YOLOv8 的 Head 部分通常由多个卷积层和全连接层组成,最后输出目标的坐标、类别和置信度等信息。

YOLOv8 模型根据参数量和网络层数的不同,由小到大可以分为 YOLOv8n、YOLOv8s、YOLOv8 m 等,为了降低对钢材表面缺陷检测终端设备的计算资源,使得模型更容易部署在终端,选择 YOLOv8n 作为轻量化研究对象。

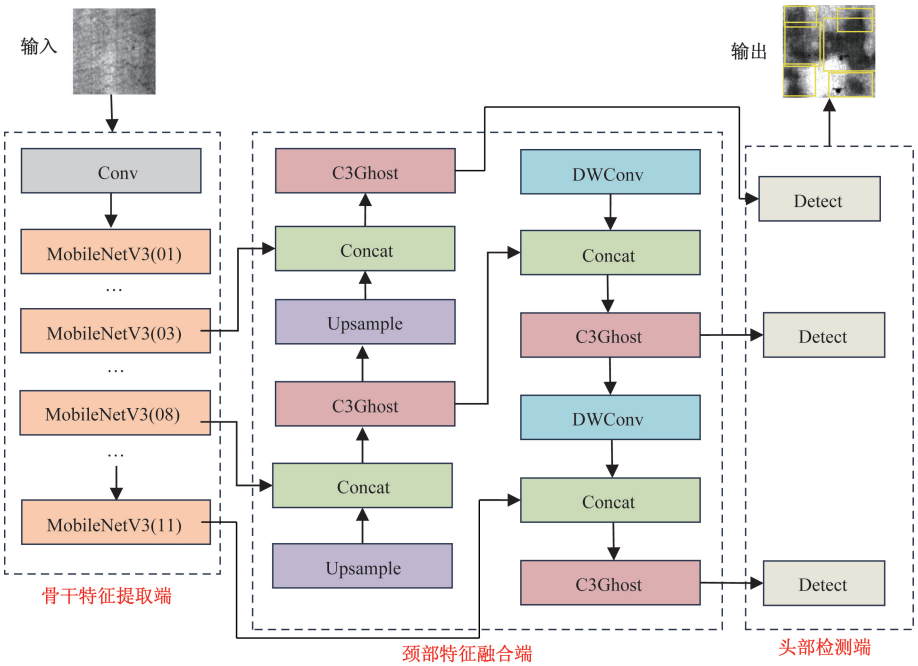
1.2　轻量级模型

目前高效的机器学习网络往往有着巨大的计算量和网络复杂度,导致对计算资源有着较高的要求,因此在实际工业生产应用中,由于终端检测设备的计算能力有限,算法在部署时通常面临困难。不同的工业生产问题对应着不同的模型使用以及不同的轻量化模型方法。Suryarasmi 等[18]为了解决织物制造业计算机计算能力有限使得卷积神经网络(convolutional neural network,CNN)难以部署的问题,在 CNN 模型的基础上提出了一种轻量型 FN-Net 算法用于织物缺陷检测,与主流架构相比,计算量要求低的同时检测精度也更高。Ding 等[19]为了实现对焊接质量进行快速准确的现场检测,以 MobileNetV2 为网络骨干,嵌入卷积块注意力模块,提出了一种基于改进轻量级 MobileNetV2 算法的原位焊缝表面缺陷识别方法,降低了模型的计算量和参数的同时提高了检测精度。Shi 等[20]为了解决由于深度学习模型计算量大难以实现电力行业对嵌入式设备需要实时反馈的需求,基于 YOLOv5 网络引入轻量级坐标注意力模块的同时,对空间和通道同时进行大核卷积网络解耦,降低了卷积计算开销。Ma 等[21]为了降低深度学习对计算机设备的要求,同时平衡铝带表面缺陷检测和速度,在 YOLOv4 的基础上,利用 Ghost 卷积构建 GMANet 骨干网络同时将 union 注意力模块嵌入其中,使得模型计算量减少了 80.41%,检测速度提高了 3 倍,实现了模型的轻量化。

2　基于 YOLOv8 的改进算法(YOLOv8-MDC)

2.1　YOLOv8-MDC 网络结构

为了解决 YOLOv8n 网络计算资源要求较高,网络结构复杂,使得在工业缺陷检测应用领域终端难以部署的问题,提出了面向钢材表面缺陷检测的轻量级 YOLOv8n-MDC 算法。首先将 YOLOv8n 的自带 IoU(intersection over union)函数替换成了 WIoU(weighted IoU),此目标检测损失函数可以在不影响泛化能力的同时,增添非单调聚焦机制。本文算法的整体构架如图 1 所示,为了去除骨干网络中其余



Concat 操作是把多个张量沿着指定维度拼接起来,形成一个尺寸更大的张量;Upsample 为上采样操作,用于增大数据的尺寸;Detect 为检测层,负责完成目标检测的相关任务

图 1　YOLOv8-MDC 网络结构图

Fig. 1　Network structure diagram of YOLOv8-MDC

的分支结构,利用轻量级 MobileNetV3 的 block 模块替换 C2f 模块,构建新的特征提取端,轻量化模型复杂度的同时提高了检测精度。此外,在特征融合网络中将 YOLOv8n 原网络中的 C2f 模块改进为 C3Ghost 模块,将普通卷积使用 DW 卷积代替,使得网络在提取特征的过程中进一步提升模型的检测性能。

2.2 WIoU 模块

IoU 损失函数是 YOLOv8 系列中用于计算检测框(Bounding box)重叠度的一种标准方法。在目标检测任务中,IoU 通常用于衡量两个检测框之间的相似度,即两个检测框之间的交集与并集之间的比值。传统 IoU 只考虑了预测框和真实框的重叠部分,没有考虑两者之间的区域,导致在评估结果时可能存在偏差。为了提高 YOLOv8 的检测精度,引入具有动态非单调 FM 的 Wise-IoU[22](WIoU)损失函数。

在钢材表面缺陷数据集中,难免会出现低质量样本,这会导致样本类别不平衡问题,使网络精度降低。与传统的 IoU 相比,WIoU 考虑了目标的交集和并集,通过对交并比进行加权处理,能够有效地处理目标检测中存在的类别不平衡问题。这意味着模型会更关注那些类别不平衡的目标,提高对少见类别的检测准确性。在出现低质量样本的情况下,WIoU 可以更好地指导模型进行训练,提高对类别的检测准确性。

WIoU 中的权重参数可以根据具体任务和数据集进行调节,可以平衡不同目标类别之间的重要性,使模型可以更灵活地适应不同的场景和需求。WIoU 不仅考虑了预测框和真实框之间的重叠程度,还考虑了两者之间的交集和并集,这有助于模型更好地捕捉目标的准确位置,提高模型对边界框预测的准确性。综上,WIoU 通过加权交并比处理类别不平衡,提高了模型对少见类别的检测能力;同时,WIoU 可使模型生成更精确的边界框,更准确地定位目标。使用的 WIoU v1 损失函数的表达式为

$$L_{\mathrm{WIoUv1}} = R_{\mathrm{WIoU}} L_{\mathrm{IoU}} \tag{1}$$

$$R_{\mathrm{WIoU}} = \exp\left[\frac{(x - x_{\mathrm{gt}})^2 + (y - y_{\mathrm{gt}})^2}{W_{\mathrm{g}}^2 + H_{\mathrm{g}}^2}\right] \tag{2}$$

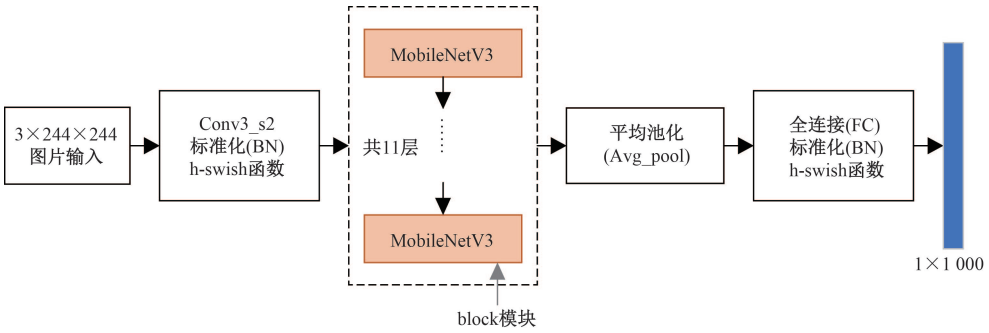
式中:$R_{\mathrm{WIoU}} \in [1,\mathrm{e})$ 能放大普通预测边界框的 L_{IoU};$L_{\mathrm{IoU}} \in [0,1]$,能降低高质量边界框的 R_{WIoU};W_{g}、H_{g} 为最小检测框的尺寸大小;x、y 和 x_{gt}、y_{gt} 分别为预测框和目标框在样本中的位置。

因此,通引入 WIoU 损失函数替换 YOLOv8n 中传统的 IoU 损失函数,更好地平衡了预测框与目标框之间几何度量的惩罚因子。避免了因为训练数据中低质量样本的存在,导致几何度量增加对低质量样本的惩罚因子从而使模型的泛化性能下降的问题。实现了在不增加额外计算量和参数量的基础上,提高模型的泛化能力。

2.3 利用 MobilenetV3 模块构建新的特征提取端

原始的 YOLOv8 系列中,特征提取端使用了复杂度较高的 C2f 模块,这使得模型计算量大大增加,导致检测速度变慢,在工业生产缺陷检测终端难以被应用,难以达到实时检测的目的。为了解决这个问题,通过采用 MobileNetV3 轻量化主干网络作为特征提取端的基础结构[23],从而在保证检测性能的同时,将网络结构精简到最小,减小了模型的参数量和计算量。

MobilenetV3 是基于 MobileNetV2 基础上改进的,体现在改网络结构、修改激活函数和增添注意力机制上。首先,为了具有丰富的预测特征,MobileNetV2 采用倒瓶颈结构和变体的模型,使用 1 × 1 卷积作为最后一层,以便扩展到更高维度的特征空间。然而,这会产生额外的计算开销,使网络的性能降低。如图 2 所示,MobileNetV3 将该层放在 average pooling 层之后,降低了计算资源开销并有利于高维特征的提取。最后一组特征现在以 1 × 1 的



Conv3_s2 是步长为 2 的 3 个卷积层组成结构

图 2 MobileNetv3 的网络结构图

Fig. 2 Network architecture diagram of MobileNetv3

空间分辨率代替 7 × 7 的空间分辨率计算。通过改良之后,降低了特征计算的开销,并且不会产生额外的参数,这使得 MobileNetV3 在计算量和速度方面有了大幅的改善。

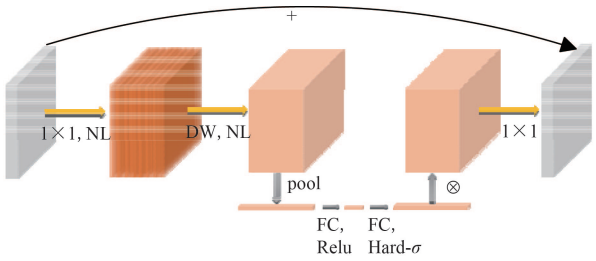
其次,MobilenetV3 的 Block 中加入了 SE 注意力机制模块,具体改动如图 3 所示。在经过了 DW 卷积操作处理后,对数据进行池化操作,再经过新建全连接层并且加上 ReLU 函数处理。最后,MobilenetV3将 MobilenetV2 中原有的激活函数 swish 修改成 h-swish,计算公式为

$$\text{swish}x = x\sigma(x) \tag{3}$$

$$\text{h-swish}(x) = x\frac{\text{ReLU6}(x+3)}{6} \tag{4}$$

$$\text{ReLU6}(x) = \min[\max(x,0),6] \tag{5}$$

式中:x 为输入值;$\sigma(x)$ 为 sigmoid 函数,$\sigma(x) = \frac{1}{1+e^{-x}}$;min(·) 表示取最小值;max(·) 表示取最大值。



NL 为非线性激活函数;DW 为深度可分离卷积;pool 为池化操作;FC 表示全连接层; Hard-σ 为 Hard-sigmoid 函数(Sigmoid 函数的一种近似替代函数)

图 3　MobileNetv3 的 Block 模块

Fig. 3　Block module of MobileNetv3

ReLU6 的优化可以在几乎所有的软件和硬件框架上使用。其次,在量子化模式下,它消除了由于近似 S 形曲线的不同而可能造成的数值精度损失。最后,h-swish 能够用分段函数来表示,用于降低内存访问频率,进而有效减少计算的开销。因此,改进后的激活函数在简化了计算步骤的同时,也与原函数的曲线接近,可以实现有效的替换,保证精度的同时提高了其在终端的部署能力。

2.4　利用 DWConv 与 C3Ghost 模块构建新的特征融合端

深度卷积(depthwise convolution,DWConv)是卷积神经网络中的一种操作,与普通卷积相比,DWConv在卷积层中对每个输入通道都使用一个独立的卷积核进行卷积操作。对于每个输入通道,不需要在每个卷积核中使用同样的权重参数处理所有输入通道。这种单独处理每个通道的方式,使得DWConv能够大幅减少模型的参数量,需要的计算量也更少,从而降低了模型复杂度,同时保持模型性能。由于钢材表面终端检测设备具有计算和存储资源有限的特点,这种轻量级模型显得尤为重要,因此使用 DWConv 构建 C3Ghost 模块和特征融合端。

Ghost[24] 模块是一种高效且轻量级的卷积结构,旨在减少模型参数量,常用于特征融合等任务。图 4(a)和图 4(b)分别为 YOLOv8n 使用的普通卷积和 Ghost 卷积结构。其核心思想是通过两步操作生成特征图。首先,对输入特征图进行少量的普通卷积,生成 m 个通道的本征特征图。接着,利用 $d \times d$ 大小的深度卷积对每个通道执行轻量线性运算(d 为深度卷积操作中卷积核的尺寸大小),得到 m 个新通道的映射特征图。最终,将本征特征图与映射特征图叠加,形成最终的输出特征图。

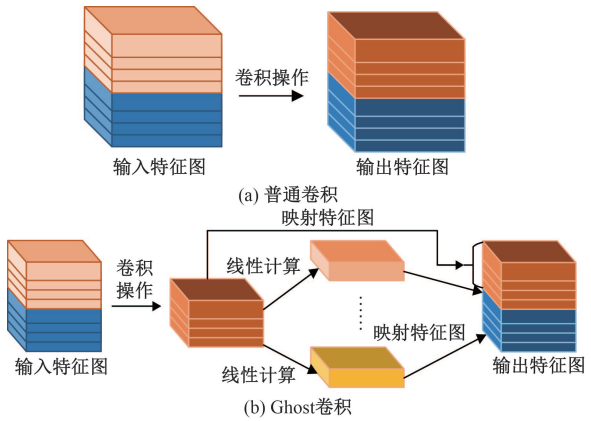


图 4　普通卷积与 Ghost 模块对比图

Fig. 4　Comparison between ordinary convolution and Ghost modules

Bottleneck 模块是 YOLOv8n 中 C2f 结构的一部分,其结构如图 5(a)所示。由普通卷积(Conv)、BN(batch normalization)以及 SiLU 激活函数组成 CBS 模块。提出一种轻量化的 Ghost Bottleneck 模块,通过用 Ghost 卷积替换原有 Bottleneck 中的 CBS 卷积,并在两个 Ghost 卷积之间嵌入 DW 卷积,再引入 DW + CBS卷积分支模块,形成新的 Ghost Bottleneck 结构,如图 5(b)所示。接着,使用该模块替换 YOLOv8n 中的 Bottleneck 模块,构建了轻量化 C3Ghost 结构,如图 5(c)所示。最后,将 C3Ghost 结构应用于 YOLOv8n 的特征融合端,同时用 DWConv 深度卷积替换 CBS 卷积模块,得到新的特征融合端设计,如图 6 所示。这一改进在保持识别精度的同时,显著减少了模型的计算量和参数量。

3　实验与分析

3.1　实验数据集

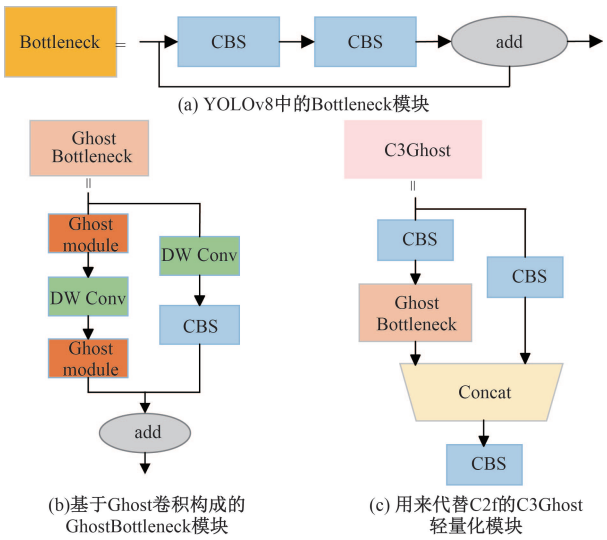
使用的数据集为 NEU-DET 钢材缺陷数据集,

其中包含 1 800 张灰度图像。数据集包含 6 类常见钢材表面缺陷,分别为:裂缝(crazing, Cr)、夹杂物(inclusion, In)、斑块(patches, Pa)、点蚀表面(pitted-surface, Ps)、氧化皮(rolled-in-scale, Rs)和划痕(scratches, Sc)。图 7 为表面缺陷示例图。训练集、验证集和测试集的分配比例为 8∶1∶1,样本量分别为 1 440、180、180 张。

为了保证各类别标签数量分布均匀,防止由于数据的问题影响模型的训练,因此使用旋转、对比度变化、翻转、裁剪、缩放等方式对数据进行了清洗和增强,有利于提高模型的鲁棒性和泛化能力,防止模型过拟合。图 8 为数据标签分布情况。其中,图 8(a)为每种缺陷目标框的数量,裂缝、夹杂物、斑块、点蚀表面、氧化皮和划痕大约分别为 560、790、690、310、460 和 410 个,检测类别标签数量较为平均;图 8(b)~图 8(d)分别为所有缺陷目标框的形状、中心位置坐标和长宽,可以看出,缺陷目标框在样本数据的位置分布和尺寸大小都较为均匀。

3.2 实验设置

本实验采用的 GPU 为 RTX 3090,显存为 24 GB,CPU 为 15 vCPU Intel(R) Xeon(R) Platinum 8358P CPU @ 2.60 GH。采用 Pytorch1.12.1 框架,编程语言为 Python3.8,同时安装了 CUDA11.0 以支持 GPU 的使用。实验阶段图片的输入尺寸为 640×640×3,数据增强方式为 Mosaic 方法。



Concat 操作是把多个张量沿着指定维度拼接起来,形成一个尺寸更大的张量;add 为元素相加操作

图 5 Bottleneck 模块与 C3Ghost 轻量化模块结构图

Fig. 5 Structural diagram of the Bottleneck module and the C3Ghost lightweight module

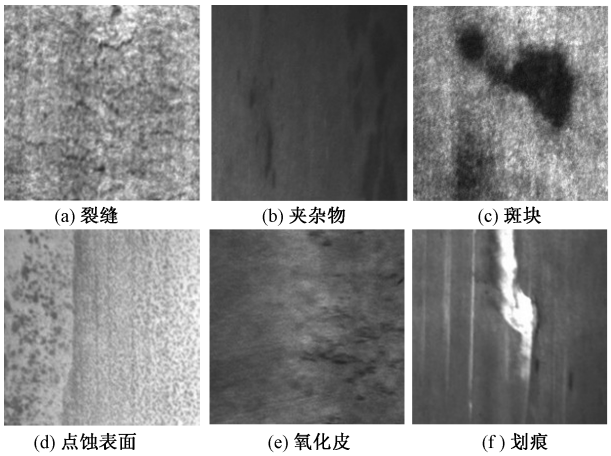
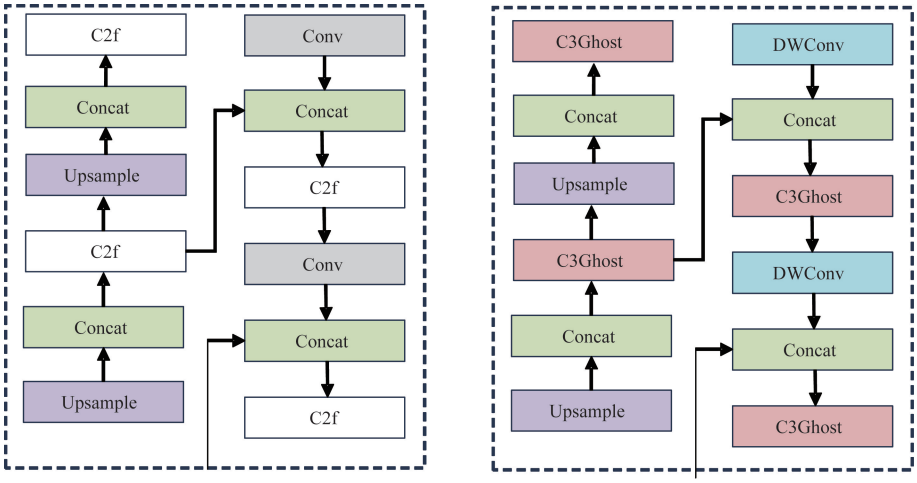


图 7 表面缺陷示例图

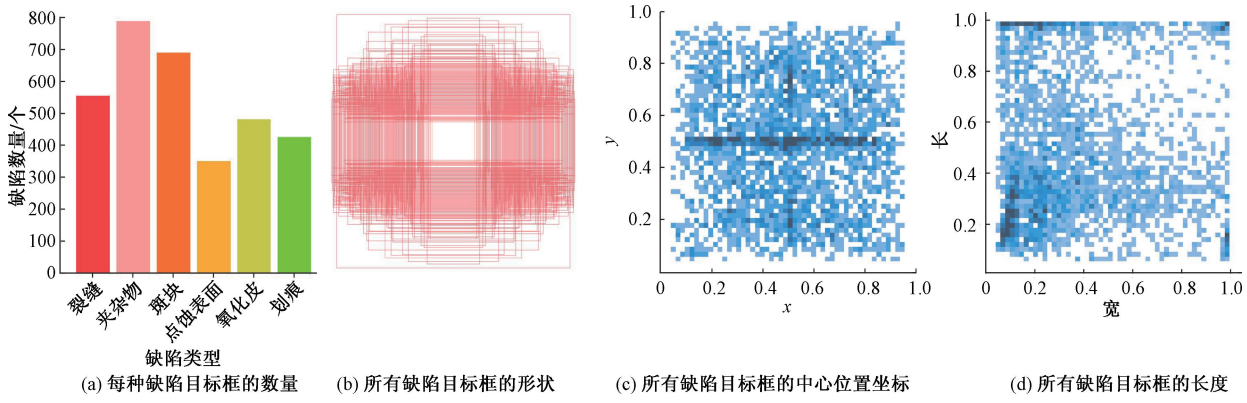
Fig. 7 Example diagram of surface detects



Concat 操作是把多个张量沿着指定维度拼接起来,形成一个尺寸更大的张量;Upsample 为上采样操作,用于增大数据的尺寸;Conv 为普通卷积层;C2f 为对 CSP(cross stage partial)模块的改进和优化版本

图 6 原特征融合端与改进后的特征融合端对比图

Fig. 6 Comparison between the original feature fusion end and the improved feature fusion end



(a) 每种缺陷目标框的数量　(b) 所有缺陷目标框的形状　(c) 所有缺陷目标框的中心位置坐标　(d) 所有缺陷目标框的长度

图 8　数据样本分布

Fig. 8　Distribution of data sample

3.3　评价指标

为了全面评价本文改进模型的性能,采用以下评价指标。

(1)召回率(R)表示预测正确的样本占所有目标样本的比例;精确率(P)表示预测正确的样本占预测为正样本的比例;IoU 阈值为 0.5 时的平均精度均值(mAP:0.5);P、R 和 mAP:0.5 的值越高,表示模型的性能越好。

(2)模型参数量(Params)也是评价模型性能的重要指标。一般来说,模型参数量越小,模型越轻量化。

(3)用 FLOPs 每秒可作浮点操作的数量衡量模型的计算量,FLOPs 越小说明算法计算量小,对设备硬件要求也越小。

(4)FPS 用来评估模型的处理数据能力,在同一时间内 FPS 值越高,说明模型可以处理缺陷图像的数量越多,其处理性能越好。

上述评价指标的计算公式为

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times 100\% \tag{6}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100\% \tag{7}$$

$$\mathrm{AP} = \int_0^1 P(R)\,\mathrm{d}R \tag{8}$$

$$\mathrm{mAP} = \frac{\sum_{i=1}^{m} \mathrm{AP}_i}{m} \times 100\% \tag{9}$$

式中:FP 为正确预测的负样本数;TP 为正确预测的正样本数;FN 为错误预测的正样本数;AP 为 P-R 曲线与各坐标轴围成的区域面积;mAP 为在各 IoU 阈值之间对 m 类标签的均值。

3.4　实验结果与分析

3.4.1　网络训练

实验模型训练时的超参设置为 YOLOv8n 模型默认值,初始学习率动量为 0.937,为了防止模型过拟合,权重衰减系数配置为 0.000 5;学习率为 0.01,随着训练 epoch 的迭代而逐渐减少。使用 SGD 优化器,epoch 设置为 500,在 50 个迭代次数内如果模型性能没有提升则停止训练,达到节省训练时间的目的。改进后的模型训练过程图如图 9 所示。图片上面部分依次为训练的 IOU 损失、置信度损失和分类损失、准确率和召回率;图片下半部分为验证的 IOU 损失、置信度损失和分类损失、mAP 值以及 mAP0.5:0.95 值。其中候选框训练和验证的 loss 值有升高的趋势,说明 YOLOv8 自带的 3 种先验框与本文使用的钢材表面缺陷数据标签框尺寸的匹配度较低,接下来的工作拟使用聚类等方法对先验框进行改进,使之匹配本文的数据集标签,从而进一步提高模型的检测精度。其他指标 loss 训练过程平稳,说明模型拟合能力强。

图 10 展示了模型检测结果的混淆矩阵,混淆矩阵反映了对所有检测标签分类准确率的具体结果,对角线元素表示检测预测分类正确的结果,标签分类准确率最高可达 93%,表明改进后的模型能够准确识别出 6 种钢材表面缺陷,说明了该模型的有效性。

3.4.2　改进后模型与原始模型的检测性能对比

经过反复实验,改进 YOLOv8n 算法前后的缺陷检测总体性能对比结果如表 1 所示,并绘制改进前后模型的 P-R 曲线对比,如图 11 所示。由表 1 可知,改进后的 YOLOv8n 算法 FLOPs 计算量减少到 2.1 G,减少了 74.1%;模型参数由 3.01 M 减少到 1.02 M,减少了 66.1%。其次,图 11 展示了对所有钢材表面缺陷类别检测的 mAP:0.5 值,改进后模型对裂缝、点蚀表面、氧化皮的检测精度分别提高了 20.1%、6.9% 和 23.9%,而且模型整体的检测平均精度 mAP:0.5 值增加了 5%。最后,YOLOv8n-MDC 的模型大小仅为基线的 36.5%,在同一显卡下 FPS 达208.3,几乎在各个方面本文算法均达到较好的

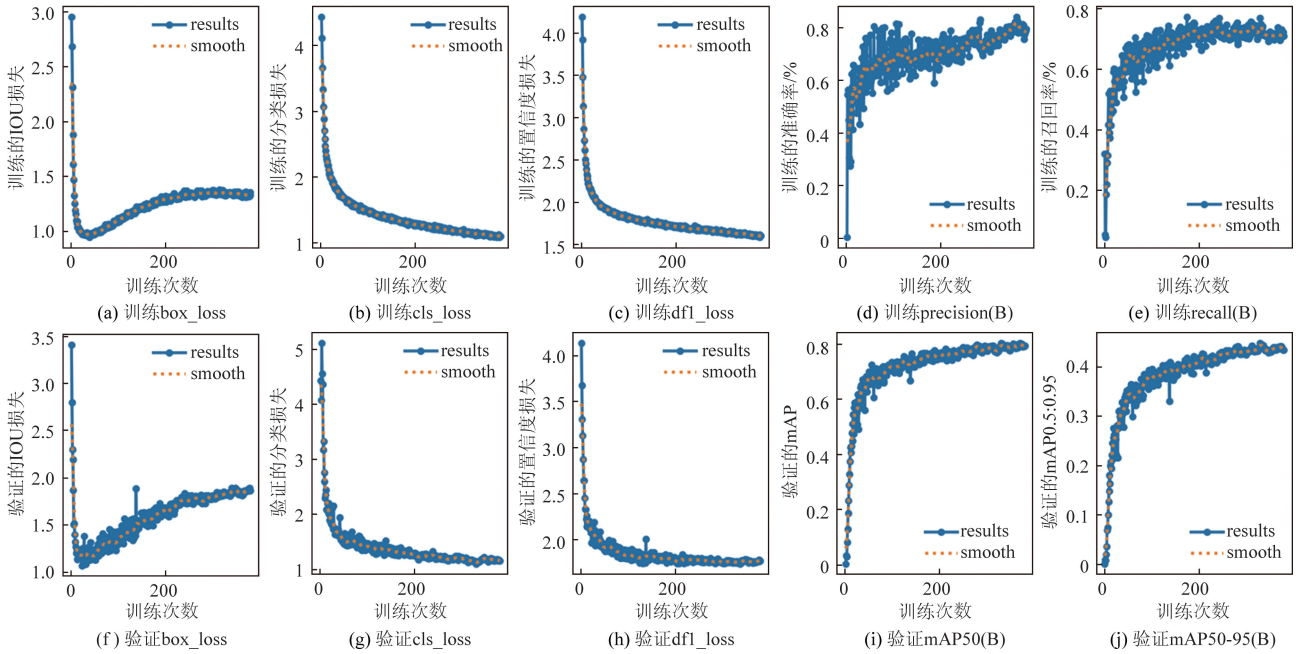


图 9 改进后模型训练过程

Fig. 9 The improved model training process

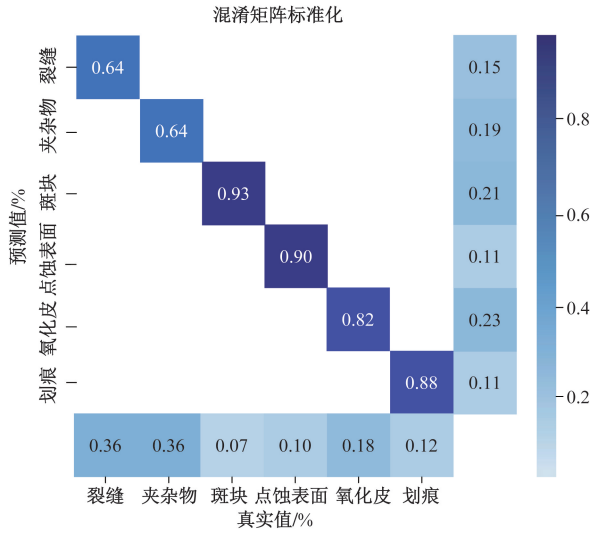


图 10 改进后模型检测结果混淆矩阵

Fig. 10 The improved model detection result confusion matrix

表 1 原始模型与改进模型的性能对比

Table 1 Performance comparison between the original model and the improved model

模型	*P*/%	*R*/%	mAP:0.5/%	参数量/M	FLOPs/G	FPS	模型大小/MB
YOLOv8n	0.805	0.669	0.763	3.01	8.1	136.9	6.3
YOLOv8n-MDC	0.714	0.790	0.813	1.02	2.1	208.3	2.3

性能。综合来看,本文算法有效达到了轻量化模型的目的,既提高了检测的准确率,又满足了现实中对钢材表面缺陷实时检测的要求。

3.4.3 消融实验

通过设置消融实验,可以探究所引入的各网络模块对钢材表面缺陷检测效果的影响,同时能够对比出各模块对 YOLOv8n 模型改进前后的作用。以 YOLOv8n 模型为基础,逐步引入改进的模块,来验证模型的检测性能,消融实验结果如表 2 所示。

由表 2 可知,使用 WIoU 候选框函数得到的 mAP 值比使用 YOLOv8n 自带的 IoU 函数得到的 mAP 值提升 0.3%,表明使用恰当的方法获取适合数据的候选框函数能提高模型的检测性能;在此基础上,在保持模型的检测精度不变的前提下,单独引入 DW + C3Ghost 模块能使模型的参数量减少到 1.42 M,FLOPs 减少到 4.4 G,使用此方法极大程度上减轻了模型的负载量,使网络性能明显提升,说明 DW + C3Ghost 模块轻量化模块的有效性;同理,单独引入 MobileNetv3 模块能使模型参数量减少到 1.19 M,FLOPs 减少到 2.8 G,FPS 明显提升至 196,模型检测精度也有略微提升,表明该模块在提高检测精度的同时又能有效轻量化算法。最后,在 WIou 函数的加持下,DW + C3Ghost 和 MobileNetv3 模块同时使用,能够使模型轻量到极致,即 1.02M,FLOPs 减少到 2.1 G,同时检测精度提升了 5%,表明这 3 个模块结合使用能够显著轻量化模型。

综上,引入所有改进模块,YOLOv8n-MDC 模型的实验结果最好,mAP 值最高达 0.813,而且模型达到轻量最大化,表明所提出的轻量化思路对钢材表面缺陷检测性能的提升是有效的。

3.4.4　改进后模型与其他模型的检测效果对比

为了进一步验证模型的检测性能和有效性,对所改进后的算法与目标检测领域的主流算法做检测效果对比研究。使用 SSD、Faster RCNN 和 YOLOv5 等在所采集的 NEU-DET 实验数据集上进行实验,同时,对比近期在钢材表面缺陷检测方面的研究者发布的YOLOv8-VSC[13]算法、DIN[15]检测方法、MCB-FAH-YOLOv8[14]算法和 YOLOv5s-FCS 算法[25],保证了实验的可靠性。由目标检测领域的主流算法对比结果(表 3)得出,所提出的改进算法的参数量和 FLOPs 均为最低,具有最好的轻量化效果;mAP 仅低于 MCB-FAH-YOLOv8 算法,但参数量只占其 16.8%,仅牺牲 0.005 的训练精度获得算法轻量化的大幅提升。整体来说,所提出的改进模型具有更好的检测性能,表明该算法具有一定的先进性。

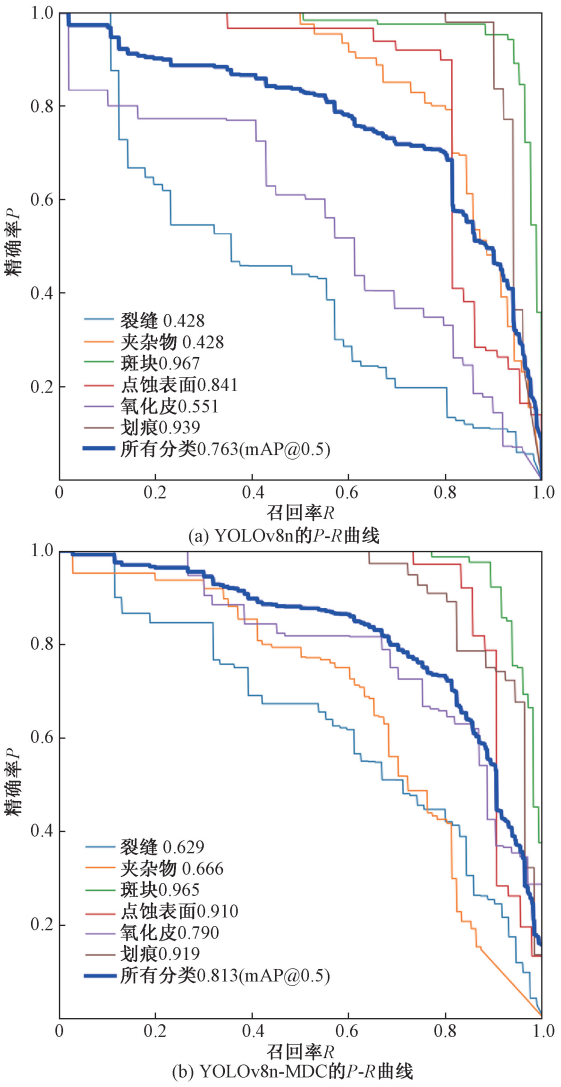


图 11　YOLOv8n 与 YOLOv8n-MDC 的 *P-R* 曲线对比

Fig. 11　Comparison of *P-R* curves between YOLOv8n and YOLOv8n-MDC

4　结论

在钢铁制造和加工业,钢材表面缺陷检测是质量把控的重要一环。人工检测方法成本高且精度不足,而传统目标检测方法模型复杂、计算资源消耗大,难以满足工业现场对实时性和低功耗的需求。随着机器学习的发展,轻量化模型成为实现工业设备智能化和便携化的重要手段,能够降低计算负担、提高检测效率,并支持实时响应和低功耗运行。因此,开发适用于钢材表面缺陷检测的轻量化模型具有重要意义。对此,提出一种轻量级的钢材表面缺陷检测算法 YOLOv8n-MDC。在实现轻量化目标的同时还提高了模型在公开数据集上的检测精度,达 81.3%。相较于常见的目标检测算法,YOLOv8n-MDC 在保持高准确性的同时,需要更少的计算资源,说明本文方法在钢材表面缺陷检测领域有所创新。不仅解决了当前钢材表面缺陷检测

表 3　目标检测领域主流算法对比

Table 3　Comparison of mainstream algorithms in the field of object detection

模型	参数量/M	FLOPs/G	mAP:0.5/%
SSD	26.2	62.7	0.732
Faster RCNN	138	371.2	0.759
YOLOv5	7.3	15.9	0.767
YOLOv8n	3.01	8.1	0.763
YOLOv8-VSC	1.96	6.0	0.808
DIN	—	—	0.805
MCB-FAH-YOLOv8	6.06	—	**0.818**
YOLOv5s-FCS	8.32	18.5	0.747
本文模型	**1.02**	**2.1**	0.813

注:加粗数值表示在对比实验中该指标的最佳结果。

表 2　消融实验结果

Table 2　Ablation experimental results

WIou	DW + C3Ghost	MobileNetv3	*P*/%	*R*/%	mAP:0.5/%	参数量/M	FLOPs/G	FPS
			0.805	0.669	0.763	3.01	8.1	136.9
√			0.764	0.763	0.786	3.01	8.1	135.0
√	√		0.736	0.788	0.789	1.42	4.4	156.0
√		√	0.76	0.744	0.791	1.19	2.8	196.0
√	√	√	0.714	**0.790**	**0.813**	**1.02**	**2.1**	**208.3**

注:加粗数值表示在对比实验中该指标的最佳结果;√表示使用该方法。

成本昂贵、精度低的难题,还提出了一种具有实用价值的轻量化算法,为工业智能化检测提供了新的解决方案,也为工业实际应用提供了可靠的技术支持。

但在 IoU 损失方面还需使用聚类等方法对先验框进行改进,从而进一步提高模型的检测精度。本文算法对于裂缝和夹杂物的检测精度还可以进一步提高,具体可以添加一些注意力机制来提高特征融合效率,但会造成模型参数量变大,计算过程变复杂,因此找到适合的改进方法是本文的改进目标。

参 考 文 献

[1] He Z, Liu Q. Deep regression neural network for industrial surface defect detection[J]. IEEE Access, 2020, 8: 35583-35591.

[2] 张涛, 刘玉婷, 杨亚宁, 等. 基于机器视觉的表面缺陷检测研究综述[J]. 科学技术与工程, 2020, 20(35): 14366-14376.
Zhang Tao, Liu Yuting, Yang Yaning, et al. Review of surface defect detection based on machine vision[J]. Science Technology and Engineering, 2020, 20(35): 14366-14376.

[3] Liu Y, Xiao H, Xu J, et al. A rail surface defect detectionmethod based on pyramid feature and lightweight convolu-tional neural network[J]. IEEE Transactions on Instrumen-tation and Measurement, 2022, 71: 1-10.

[4] 邵延华, 张铎, 楚红雨, 等. 基于深度学习的 YOLO 目标检测综述[J]. 电子与信息学报, 2022, 44(10): 3697-3708.
Shao Yanhua, Zhang Duo, Chu Hongyu, et al. A review of YOLO target detection based ondeep learning[J]. Journal of Electronics and Information, 2022, 44(10): 3697-3708.

[5] Zhang K, Shen H. Solder joint defect detection in the connectors using improved faster-RCNN algorithm[J]. Applied Sciences, 2021, 11(2): 576.

[6] Jiang P Y, Ergu D, Liu F Y, et al. A review of YOLO algorithm developments[J]. Procedia Computer Science, 2022, 199: 1066-1073.

[7] Ren J S, Wang Y. Overview of object detection algorithms using convolutional neural networks[J]. Journal of Computer and Communications, 2022, 10(1): 115-123.

[8] Xiao Y, Tian Z, Yu J, et al. A review of object detection based on deep learning[J]. Multimedia Tools and Applications, 2020, 79(33/34): DOI: 10.1007/s11042-020-08976-6.

[9] 陈科圻, 朱志亮, 邓小明, 等. 多尺度目标检测的深度学习研究综述[J]. 软件学报, 2021, 32(4): 1201-1227.
Chen Keqi, Zhu Zhiliang, Deng Xiaoming, et al. A review of deep learning research on multi-scale target detection[J]. Journal of Software, 2021, 32(4): 1201-1227.

[10] Tong K, Wu Y, Zhou F. Recent advances in small object detection based on deep learning: a review[J]. Image and Vision Computing, 2020, 97: DOI: 10.1016/j.imavis.2020.103910.

[11] Han K, Wang Y, Chen H, et al. A survey on vision Transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45: DOI: 10.1109/TPAMI.2022.3152247.

[12] 许德刚, 王露, 李凡. 深度学习的典型目标检测算法研究综述[J]. 计算机工程与应用, 2021, 57(8): 10-25.
Xu Degang, Wang Lu, Li Fan. A review of research on typical targetdetection algorithms for deep learning[J]. Computer Engineering and Applications, 2021, 57(8): 10-25.

[13] 王春梅, 刘欢. YOLOv8-VSC: 一种轻量级的带钢表面缺陷检测算法[J]. 计算机科学与探索, 2024, 18(1): 151-160.
Wang Chunmei, Liu Huan. YOLOv8-VSC: a lightweight surface defect detection algorithm for strip steel[J]. Computer Science and Exploration, 2024, 18(1): 151-160.

[14] 崔克彬, 焦静颐. 基于 MCB-FAH-YOLOv8 的钢材表面缺陷检测算法[J]. 图学学报, 2024, 45(1): 112-125.
Cui Kebin, Jiao Jingyi. Algorithm for steel surface defect detection based on MCB-FAHYOLOv8[J]. Journal of Graphics, 2024, 45(1): 112-125.

[15] Hao R, Lu B, Cheng Y, et al. A steel surface defect inspection approach towards smart industrial monitoring[J]. Journal of Intelligent Manufacturing, 2021, 32(7): 1833-1843.

[16] Zhou S, Zeng Y, Li S, et al. Surface defect detection of rolled steel based on lightweight model[J]. Applied Sciences, 2022, 12(17): 8905.

[17] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 779-788.

[18] Suryarasmi A, Chang C, Akhmalia R, et al. FN-Net: a lightweight CNN-based architecture for fabric defect detection with adaptive threshold-based class determination[J]. Displays, 2022, 73: 102241.

[19] Ding K, Niu Z, Hui J, et al. A weld surface defect recognition method based on improved MobileNetV2 algorithm[J]. Mathematics, 2022, 10: 553-575.

[20] Shi C, Lin L, Sun J, et al. A lightweight YOLOv5 transmission line defect detection method based on coordinate attention[C]// IEEE 6th Information Technology and Mechatronics Engineering Conference(ITOEC). New York: IEEE, 2022: 851-866.

[21] Ma Z, Huang Y, Huang M, et al. Automated real-time detection of surface defects in manufacturing processes of aluminum alloy strip using a lightweight network architecture[J]. Journal of Intelligent Manufacturing, 2022, 34(5): 2431-2447.

[22] Tong Z J, Chen Y H, Xu Z, et al. Wise-IoU: bounding box regression loss with dynamic focusing mechanism[J]. Journal of Real-Time Image Processing, 2024, 21(2): 1-12.

[23] Howard A, Sandler M, Chu G, et al. Searching for MobileNetv3[C]//IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2020: DOI: 10.1109/ICCV.2019.00140.2019.

[24] Han K, Wang Y, Tian Q, et al. GhostNet: more features from cheap operations[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). New York: IEEE, 2020: DOI: 10.1109/CVPR42600.2020.00165.

[25] 周孟然, 王昊男, 高立鹏, 等. 基于 YOLOv5s-FCS 的钢材表面缺陷检测[J]. 科学技术与工程, 2024, 24(14): 5901-5910.
Zhou Mengran, Wang Haonan, Gao Lipeng, et al. Steel surface defection based on YOLOv5s-FCS[J]. Science Technology and Engineering, 2024, 24(14): 5901-5910.