

基于PPMA-YOLO的金属表面缺陷检测

王达¹, 李绍丽¹, 陆鹏², 葛兵²

¹沈阳工业大学信息学院, 辽宁 沈阳

²沈阳翰熙机械设备有限公司, 辽宁 沈阳

发布日期: 2026年01月30日

摘要

金属板材广泛应用于航空航天、汽车制造、桥梁建筑等支柱行业中, 其质量关系到终端产品的质量和寿命。然而, 金属板材在工业生产过程中, 受到板材表面杂质和加工设备磨损等影响, 可能出现各类表面缺陷, 这些缺陷不仅损害板材本身的外观, 而且可能进一步出现破裂和锈蚀, 影响产品最终的性能。随着现代工业产品加工要求的不断提高, 表面缺陷的检测已成为制造过程中的必要环节, 在保证产品质量的同时, 也有利于及时分析缺陷成因、发现轧机故障、改善生产工艺。对此, 本文提出了一种基于PPMA-YOLO的金属表面缺陷检测方法。在YOLOv8检测模型主干方面, 我们提出了采用PP-LCNET的轻量级网络替换YOLOv8模型中的主干网络, 旨在轻量化网络, 提高检测效率, 并采用H-Swish激活函数提升检测性能。同时, 在主干网络尾部添加SE模块和大卷积核做改进, 进一步兼顾准确度和检测速度。在颈部方面, 我们添加并改进了多重上下文注意力(MCALayer)机制。同时提出在MCA尾部添加多尺度特征融合以及小型残差块, 进行残差学习, 避免梯度爆炸和梯度消失问题。将Mca注意力机制、轻量级多尺度特征融合以及小型残差块整合成为MAFR模块, 嵌入YOLOv8检测头内, 输出精炼特征图。最后改进损失函数SIoU, 引入自适应权重因子, 基于目标框大小比例调整距离成本。添加纵横比成本, 评估边界框形状相似度, 使其在金属表面缺陷尺寸不一的情况下, 能更好地适应目标形状和位置的变化。实验证明, 在金属表面缺陷复杂的情况下, 使用本文改进的YOLO目标检测方法对东北大学钢带表面缺陷数据集(NEU-DET)的1800张200*200像素的缺陷图像进行训练并在200张验证集上进行验证, Map.50和模型大小分别为83.87以及4.8MB, 对比现存的金属表面缺陷检测方法, 展现出更准确的检测结果以及更高的检测效率, 满足工业实际场景对金属表面缺陷进行精准实时的检测要求。

关键词

金属表面缺陷; PPMA-YOLO; 目标检测; 轻量化

PPMA-YOLO: A Method for Metal Surface Defect Detection

Da Wang¹, Shaoli Li¹, Peng Lu² and Bing Ge²

¹Shenyang University of Technology, Shenyang, 110870, China

²Shenyang Hanxi Mechanical Equipment LLC, Shenyang Liaoning, China

文章引用: 王达, 李绍, 陆鹏, 葛兵. 基于 PPMA-YOLO 的金属表面缺陷检测[J]. 计算机科学辑要, 2026, 1(3): 1-13.

ISSN Print: 3079-5362; ISSN Online: 3079-5370

DOI: 10.63313/CS.8022

Abstract

Metal sheets are vital to core industries like aerospace, automotive, and bridge construction, where their quality directly determines the safety and lifespan of final products. During production, factors like material impurities and equipment wear often cause surface defects such as stains, scratches, and cracks, which can lead to structural failure and corrosion. As manufacturing standards rise, detecting these defects has become essential for maintaining quality and optimizing production processes. To address this, this paper introduces a detection method called PPMA-YOLO. We first replaced the standard YOLOv8 backbone with a lightweight PP-LCNet structure to make the model faster and more efficient, using the H-Swish activation function and adding an SE attention module with large-kernel convolutions to balance precision and speed. In the neck section, we improved the Multi-Contextual Attention (MCALayer) and combined it with multi-scale feature fusion and small residual blocks to create the MAFR module. This module, embedded in the detection head, helps the model capture refined features while avoiding common training issues like gradient explosion. Finally, we updated the Sigmoid loss function with an adaptive weight and a shape-matching cost, allowing the model to better locate defects of various sizes and shapes. Tests on the NEU-DET dataset show that our method is highly effective; with a model size of only 4.8 MB, it achieved an 83.87% mAP@0.5. Compared to existing methods, PPMA-YOLO is both more accurate and faster, making it an ideal solution for real-time defect detection in industrial settings.

Keywords

Metal Surface Defects; Ppma-Yolo; Object Detection; Lightweight

Copyright © 2026 by author(s) and Erytis Publishing Limited.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



1. 引言

为适应当前制造业高速发展的迫切需要, 加强质量控制、推进智能化转型等举措都成为当前重要任务。我国提出的“中国制造 2025[1]”将质量为先、智能制造作为基本方针。“中国机械工程技术路线图[2]”同样将提高产品质量作为重要目标。因此, 如何以自动化的方式保证产品出厂质量, 已成为工业界和学术界共同关注的热点问题。在工业生产中, 表面缺陷是一种十分常见的质量问题。尤其是在机械领域, 识别金属加工表面缺陷显得尤为重要。由于诸如机床振动、工件原料材质以及生产环境等因素的影响, 金属工件在加工过程中很容易产生各种表面缺陷。

传统的金属表面缺陷检测主要依赖人工检测的方式完成。其优点对检测设备的要求较少, 但该方法所需要劳动量大, 且伴随着生产速度的提高以及检测标准的愈发严格, 人工检测的准确率低、实时性差、受人工经验以及主观因素影响等弊端逐渐凸显。随着计算机技术的快速发展, 传统机器视觉技术替代人工检测。然而基于深度学习的金属缺陷检测更为高效。深度学习主要分为两个类别: 一类是单阶段检测算法, 例如 YOLO 系列算法; 另一类为双阶段检测算法, 例如 Faster-RCNN。然而无论是单阶段还是双阶段的检测算法都存在着模型效率低、小目标检测效果不好、特征提取能力不足等原因导致最终检测结果精度不高的问题, 这也是目前金属表面缺陷检测面临的问题。

因此, 为了满足工业生产中金属表面缺陷检测的要求, 并实现综合性模型, 本文提出了一种基于 PPMA-YOLO 结构的金属表面缺陷检测算法。该算法在单阶段检测算法 YOLO 系列中的 YOLOv8 模型基

础上进行优化, 并将其应用至 NEU-DET 金属缺陷数据集中。YOLOv8 相较于原先版本设计了新的骨干网络、无锚框方法和解耦检测头, 具有更好的性能, 因此我们选择在 YOLOv8 基础上进行优化改进。

本文工作的主要贡献主要如下:

1) 我们使用轻量级骨干网络 PPLC-NET 替换了 YOLOv8 模型中骨干网络, 避免了过多的 C2f 卷积网络, 减少了模型处理中的冗余计算, 大幅度轻量化模型。

2) 我们基于 Mca 注意力机制在其尾部添加了轻量级多尺度特征融合模块 MSFFM, 将 Mca 与 MSFFM 串联可以使模型具备单尺度的上下文和跨尺度的多分辨率的感知能力, 有更强的表征能力。在二者尾部串联融合小型残差块, 避免梯度爆炸问题。我们将其融合为一个 MAFR 模块, 使网络能更精准地关注到关键特征, 从而提升检测的准确性。

3) 我们改进了 YOLOv8 的 IoU 损失函数, 改进为自适应形状感知损失函数 Adaptive and Shape-Aware SIOU Loss(ASA-SIOU)。金属表面缺陷种类复杂, 尺度变化大, 传统 IoU 损失函数在处理时, 常出现收敛速度慢、定位不准确的问题。本文引入的 ASA-SIOU 引入了角度惩罚机制以应对方向随机的缺陷, 设计了自适应距离权重来统一优化不同大小的缺陷目标, 并采用解耦的形状代价来精准匹配不同形状的缺陷, 提高了模型的收敛速度, 同时增强了对复杂缺陷的定位能力。

本文的其余结构安排如下: 在第 2 节种, 介绍了相关工作, 重点讨论了其他研究者在应对金属表面复杂缺陷采用的优化方法。第 3 节则详细阐述了本文提出的方法, 首先概述了模型总体架构, 随后分别对 PPLC-NET、MAFP、daptive and Shape-Aware SIOU Loss 模块进行深入介绍。第 4 节对本文方法进行了详细的理论论证和实验对比, 并进行可视化分析。最后第 5 节总结了本文的研究成果, 并提出了未来的研究方向和展望。

2. 相关工作

2.1. 模型效率低问题

对于模型效率低问题, 通常采用模型结构优化、减小参数量以及剪枝等方式解决。针对不同模型选择合适的方法, 可以有效提高模型效率和实时性。在现有的研究成果中, 针对金属板材检测, 可使用已经成熟的轻量化网络解决模型效率低问题。突出的例子有: LONG 等人[3]提出了 GDetect 模块, 其根据参数共享的原理, 使特征图先经过两个参数共享的卷积操作, 再经过解耦, 进行回归分支操作, 采用分组卷积替换原始模型中 3×3 卷积, 减少模型参数数量。LI[4]等人[4]结合深度可分离卷积(DW-Conv)和普通卷积(CSB)的特点, 提出了一种轻量化的鬼影混洗卷积(GSConv), 该卷积不仅保留了 CBS 高效的特点, 且有效降低参数量和计算复杂度。赵等人[5]受 GSConv 启发, 将其三个检测头每个分支的第一个 3×3 卷积替换为 3×3 的 GSConv, 进一步降低参数量和复杂度。GUO 等人[6]依据 MobileNetV3 网络原理, 通过结合深度可分离卷积、线性瓶颈结构的逆残差模块以及基于压缩激励机制(SE)的轻量化注意力模型替换 YOLO 主干网络, 设计了具有更快检测速度的特征提取网络。然而这些方法虽能有效改善模型效率低的问题, 但在检测精度上仍有一定提升空间。因此我们受 MobileNetV3 模块的设计理念, 提出将 PP-LCNET 模块引入 YOLO 主干网络中, 同时在网络尾部 Strategic 添加 SE 模块和大卷积核, 平衡精度和速度。

2.2. 模型精度不高问题

针对模型检测精度不高问题, 模型特征提取能力不足是主要原因, 会直接影响模型精度。目前在金属表面缺陷检测领域, 一些研究者在模型特征提取能力提升上取得了一些成果。LONG 等人[3]使用 CG(context guided)[7]模块替换 Backbone 的下采样卷积, 使得主干网络能够同时处理局部特征、周围上下文和全局上下文, 提高模型对特征的定位精度以及周围环境的理解和识别。ZHAO 等人[8]设计了 TCM-YOLO, 采用三通道可变形卷积(TDCN)替换原模型骨干网络 C2f 中的标准卷积, 其以 CA[9]为基础, 对输入图像进行两个以为全局池化操作的基础上, 再加入全局自适应池化, 变成三通道, 增强了网络对不规则缺陷目标学习特征的表达能力。TAO 等人[10]在 Neck 网络中利用 CARAFE 算子替换最近邻插值上采样 Upsample 模块, 减少上采样过程中特征信息的损失, 提高网络的特征提取能力。REN 等人[11]在 YOLOv3 基础上采用动态修正线性单元替换 Darknet-53 中所有残差块的激活函数, 同时融合 CBAM 卷积注意力模块, 以此增强对复杂缺陷目标的特征提取能力。

然而仅提升模型特征提取能力不足以完全解决金属表面缺陷检测精度不高问题。由于金属表面缺陷复杂程度高, 且缺陷尺寸较小。例如以 NEU-DET 数据集为例, 图像整体为 200×200 个像素, 而微小缺陷可能仅为 10 个像素的面积, 此类小目标缺陷给特征提取带来了极大的挑战。因此需要采取进一步措施, 来解决小目标问题。目前针对金属表面缺陷检测小目标问题突出的例子有: REN 等人[11]在 FPN 中加入特征检测尺度, 通过上采样跨层连接所有特征层, 实现不同层级的特征融合, 增强对小目标的特征提取能力。叶等人[12]使用 K-means 聚类算法根据金属表面缺陷数据集的标注文件进行重新聚类, 使得到的新锚框更适合金属表面缺陷检测, 提高小目标检测的准确率。王等人[13]将损失函数替换为 NWD, 由于 YOLO 中一般使用 CIoU 计算损失值, 其对微小物体的偏差非常敏感, 而 NWD 对目标尺度不敏感, 对微小目标的定位更加准确, 以此提高对小目标缺陷的检测精度。同时加入了极化自注意力机制(PSA), 包含通道和空间两个自注意力部分, 更有效进行特征融合, 进一步提高精度。然而, 这些方法的模型虽然达到了较高的精度, 但同时也加大了模型参数量, 在一定程度上影响了模型的均衡性。因此我们设计了一款 MAFR 模块, 借鉴 PSA 极化自注意力机制, 我们采用 MCA 多重上下文注意力, 包含高度、宽度、通道三方向注意力部分, 同时融合轻量级多尺度特征融合 MSFFM, 以及小型残差块, 有效进行特征融合, 同时不增加模型参数负担。最后我们改进 YOLO 传统 IoU, 以 IoU 为基础, 引入自适应权重因子, 调整距离成本, 添加纵横比成本, 评估边界框形状相似度, 综合 IoU、距离成本、形状成本和纵横比成本, 得出最终损失, 提高对小目标缺陷检测精度。

3. 方法

3.1. 整体架构

本文所提出的 PPMA-YOLOv8 结构如图 1 所示, 主要包括三个部分, 负责从输入图像中提取目标特征的主干网络(Backbone), 负责融合来自主干网络不同尺度的特征的颈部网络(Neck), 以及将网络输出的特征图转换为包含类别、置信度等信息的预测结果的预测层 Head。具体而言, 在 Backbone 中引入我们设计的 PP-LCNET 模块替换 YOLOv8 基线模型中的主干网络, 采用深度可分离卷积作为基本模块, 构建轻量级网络结构, 并在网络尾部 strategic 添加 SE 模块和大卷积核, 平衡准确度和速度, 在全局平均池化后添加大维度 1×1 卷积, 增强特征组合能力。在 Neck 颈部添加 MAFR 模块, 该模块融合了 MCALayer 多重上下文注意力机制, 进行多维度特征交互, 引入 LightweightMSFFM 进行多尺度特征融合, 结合 MiniResidualBlock 进行残差学习, 实现自适应特征细化。

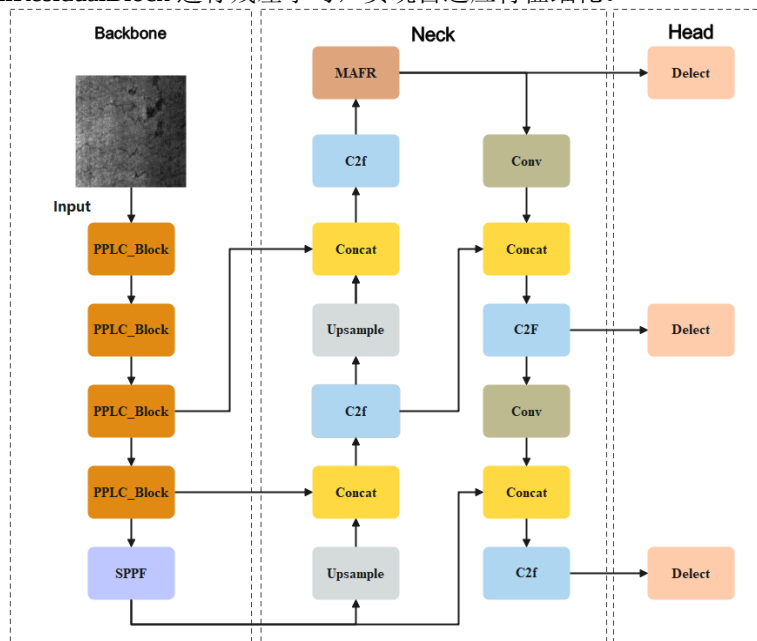


Figure 1. PPMA-YOLOv8 Architecture Diagram

图 1. PPMA-YOLOv8 结构图

3.2. 轻量级骨干网络 PPLCNet

在金属板材缺陷检测的场景中, 特征提取对于提高检测精度具有重要意义。金属板材缺陷通常包含复杂多样的缺陷种类且缺陷可能非常微小, 深度学习网络需要具备强大的特征提取能力, 以便应对多种类且微小的缺陷。在 YOLOv8 中通过 C2f 模块增加网络的深度和感受野, 提高网络特征提取能力。它将特征图进行一次卷积, 通过多个 Bottleneck 模块逐步提取特征, 输出的特征图与原始特征图拼接, 最后通过一次卷积将拼接后的特征图进行压缩, 输出目标通数的特征图。在特征提取通过一次 C2f 模块时, 产生过多的参数需要消耗过量内存, 限制模型运行效率, 延长训练时间以及处理速度。为解决上述问题, 我们受 MobileNet 启发, 将 YOLOv8 主干替换为轻量级骨干网络 PPLCNet。PPLCNet 是基于百度开源深度学习框架 PaddlePaddle 开发的检测网络, 相较于 MobileNetV2、MobileNetV3, 取得了更优的延迟与精度的均衡。PPLCNet 使用了类似 MobileNetV1 中的深度可分离卷积作为基础, 通过不同尺寸的深度可分离卷积构建 BaseNet。使用 H-Swish 激活函数提升性能, 在网络尾部添加 SE 模块和大卷积核, 平衡精度与速度。最后在全局平均池化后添加大维度 1×1 卷积, 增强特征组合能力, 具体结构图如图 2。

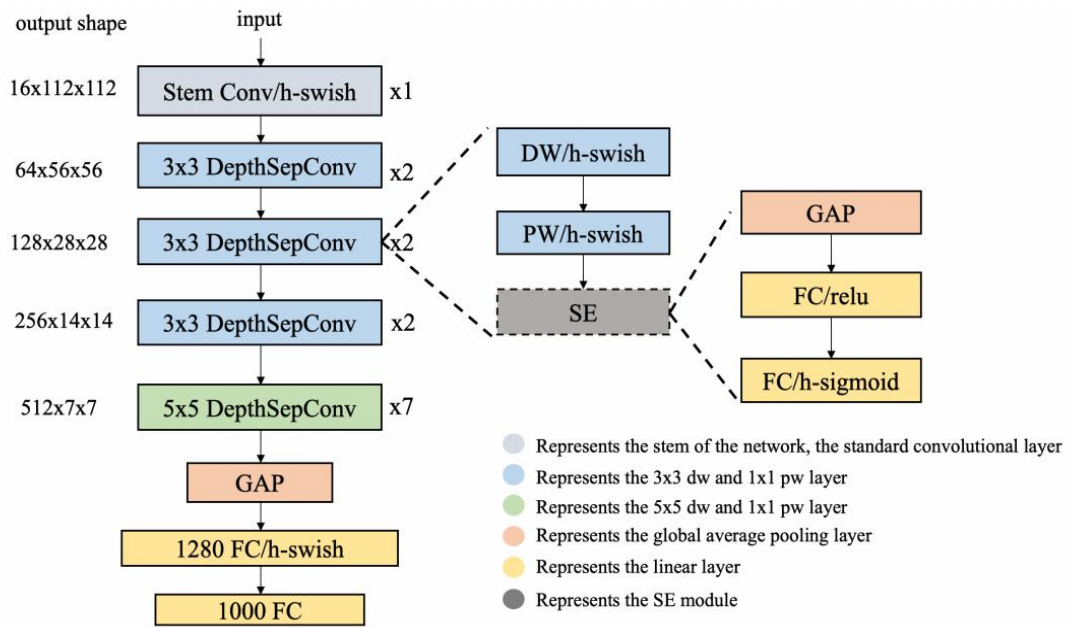


Figure 2. PPLCNet Architecture Diagram

图 2. PPLCNet 结构图

使用深度可分离卷积代替标准卷积可以更有效提取特征信息。深度可分离卷积由深度卷积(DW Conv)和逐点卷积(PW Conv)组成, 与标准卷积相比, 其通过分解卷积的方式减少了模型的数量和复杂程度, 有效提高了模型推理速度。分别采用 3×3 和 5×5 的卷积核进行调整提取, 不仅可以捕捉局部细节信息, 同时可以获得更大感受野以及更多全局信息。

H-Swish 激活函数是 Swish 的一个变体, Swish 激活函数在正向传播过程中能够产生较大的梯度, 有助于缓解梯度消失问题同时其具有非单调性, 意味着它能够提升模型的表达能力, 有助于提高模型性能, 对应公式为:

$$\text{Swish}(x) = x\sigma(\beta x) \quad (1)$$

其中 σ 是 Sigmoid 函数, β 是一个可学习参数。

由于 Swish 公式包含 Sigmoid 函数, 增加计算负担。H-Swish 用一个分段性函数替换 Sigmoid 函数, 以此减少计算量。同时 H-Swish 在输入接近 0 时, 表现得更为平滑, 能够在反向传播过程中提供更精确的梯度信息, 有助于网络优化, 对应公式为:

$$H - \text{Swish}(x) = x \frac{\text{ReLU}(x+3)}{6} \quad (2)$$

SE 注意力模块, 首先通过全局平均池化对输入特征进行压缩, 接着通过两个全连接层对压缩后的特征图进行激励, 分别进行非线性处理, 并生成权重。最后, 将生成的权重与原始特征图逐通道相乘, 实现对原始特征的重新加权。然而 SE 模块会增加模型推理时间, 所以不能在整个网络中使用。只在网络尾部添加 SE 模块, 以增强模型对不同特征通道的敏感性, 使其能聚焦对任务更重要的特征通道, 从而提升性能, 且不增加模型计算负担, SE 模块结构如图 3。

PPLCNet 在经过全局平均池化 (GAP) 后, 网络输出维度较小。直接添加最后的分类层可能会导致特征丢失。为了增强网络的拟合能力, 在最后的 GAP 层之后添加了一个 1280 维的 1×1 卷积。该方法可以在不增加太多计算时间的情况下存储更多特征。

3.3. 基于 Mca 注意力机制和多尺度特征融合模块 (MAFR)

在卷积神经网络中, 通过增加网络深度、宽度和基数来提升模型性能存在训练优化困难、存储和计算开销大等限制同时检测目标可能存在复杂背景干扰, 如何使模型突出关注我们需要检测的目标, 减少干扰, 是重中之重。神经网络在处理输入特征时, 会将输入序列压缩成一个固定长度的上下文向量。无论输入长短, 某型都只能根据这一个向量来代表所有信息。这会导致模型出现信息瓶颈以及长程依赖问题。而注意力机制的核心理念是动态聚焦它允许模型在输出的每一步有选择地、动态地关注输入序列中最相关的部分而不是被迫使用整个序列的粗糙摘要。添加注意力机制可以显著提醒模型性能, 使模型能捕捉到更精细的关联, 生成更准确、更流畅的结果。在针对金属板材缺陷检测时, 金属表面背景复杂、缺陷种类多, 且存在微小缺陷。单纯使用原始模型进行检测, 复杂的金属背景会干扰模型对缺陷信息的捕捉能力, 难以应对多种类且微小缺陷的繁琐特征信息。

为了应对这些挑战, 本文引入了 MCA 多维协同注意力机制。现有注意力机制分为通道注意力、空间注意力和混合域注意力三类虽然在一定程度上提升了网络性能, 但仍存在问题, 多数方法忽略在通道和空间维度同时建模注意力, 或引入较高模型复杂度和计算负担, 无法有效利用维度间关系。MCA 注意力机制是通过多维度的注意力机制来增强特征表示。其中包含三个并行卷积分支, 分别处理完整的输入特征图。 1×1 卷积主要进行跨通道的信息交互与整合, 感受野最小, 专注于捕获局部的、点与点之间的特征关系。 3×3 卷积是标准的卷积操作, 拥有中等大小感受野, 能捕获像素周围一个小区域的模式。 3×3 空洞卷积, 使用空洞卷积能在不增加参数量的情况下扩大了感受野。可以获得更广泛区域的、更全局的上下文信息。将三个分支的输出进行逐元素相加, 融合多尺度信息, 如等式 3、4 所示。

$$F = B1 \oplus B2 \oplus B3 \quad (3)$$

$$F_{c,h,w} = B1_{c,h,w} + B2_{c,h,w} + B3_{c,h,w} \quad (4)$$

将融合后的特征图 F 通过一个 Sigmoid 激活函数, 生成最终的空间注意力权重图 A 。权重图将每个元素值控制在 0, 1 之间, 明确指示了在通道中, 每个空间位置的重要性, 如等式 5 所示。将生成的注意力权重图 A 与原始输入特征图 X 进行逐通道的逐元素乘法, 得到最终的输出特征图。使用 MCA 注意力机制增强了重要特征区域的响应, 抑制了不重要的特征区域。

$$A_{c,h,w} = \frac{1}{1 + e^{-F_{c,h,w}}} \quad (5)$$

在针对金属板材缺陷检测中, 其缺陷背景、种类、尺寸复杂程度高, 且在同一张缺陷图像中, 同一种类缺陷的尺度变化较大。MCA 已经从多维度生成了权重图, 强调了重要特征, 将输出特征图与原始输入特征再次融合, 此时模型同时具备了单尺度内的多上下文和跨尺度的多分辨率的感知能力, 能够应对尺度变化大的金属缺陷。因此我们引入了在 MCA 注意力机制尾部添加轻量级多尺度特征融合 LightweightMSFFM。将 MCA 与 MSFFM 串联组合, 可以使模型获得更强的表征能力和更高的性能。虽然增加了计算量, 但 MCA 主要使用轻量的卷积, 且 MSFFM 的核心是 1×1 卷积和上下采样操作, 两者参数量相对较少, 而且由于 MCA 对特征图的净化作用, MSFFM 则不需要学习如何忽略金属背景噪声, 使网络能够更快收敛。MCA 和 MSFFM 主要由卷积、上采样、线性相加等操作构成, 非线性表达能力有

限。因此我们引入小型残差快, 使用 ReLU 激活函数, 获得了更强的非线性映射能力, 是网络能够学习更复杂、更抽象的特征组合, 从而更好地拟合 MSFFM 输出的复杂特征分布。

我们基于 MCA 注意力机制, 创新融合 MSFFM 多尺度特征融合以及小型残差块, 构成 MAFR 模块, 添加在 Neck 中, 首先用 MCA 对当前层特征进行提纯, 突出重要信息, 然后用 MSFFM 将提纯后的特征与其他层级特征融合, 构建丰富、多尺度的上下文信息, 最后用小型残差快对融合后可能粗糙的特征进行精细加工、深度提取, 并通过残差连接确保信息无损和梯度流畅, 为后续检测头提供高质量的特征基础, MAFR 结构图如图 3 所示。

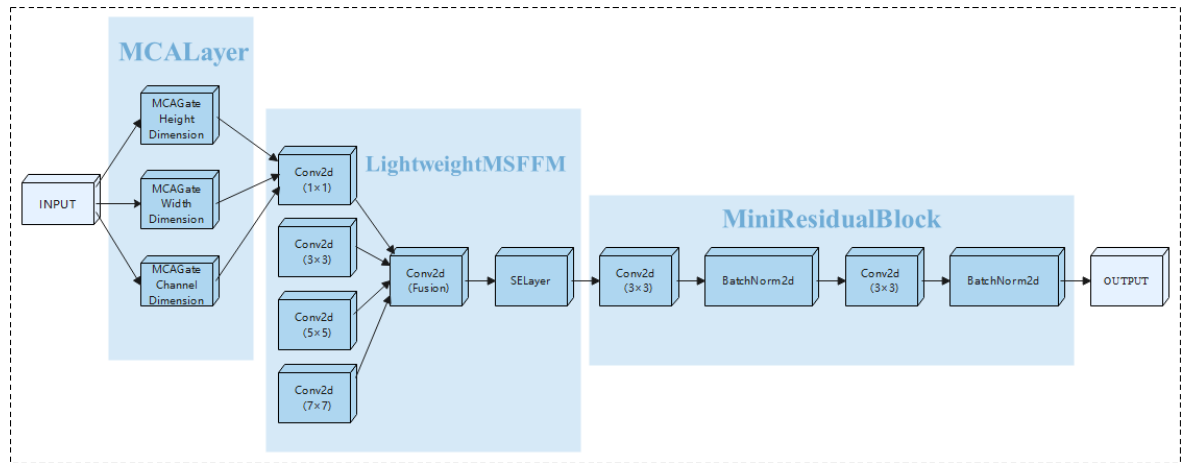


Figure 3. MAFR Architecture Diagram

图 3. PPMA-YOLOv8 结构图

3.4. 自适应形状感知损失函数 Adaptive and Shape-Aware SIoU Loss(ASA-SIoU)

金属板材表面缺陷具有形态多样性、尺度极端性和方向随机性的特点。这对目标检测模型中边界框的回归精度提出了极高要求。传统的 IoU 损失函数在处理此类复杂缺陷时, 常出现收敛速度慢、定位不准确的问题。本文提出一种针对性的尺度自适应 SIoU 损失函数, 其核心在于引入了角度惩罚项以应对方向随机的缺陷, 设计了基于缺陷尺度的自适应距离权重来统一优化不同大小的目标, 并采用解耦的形状代价来精准匹配不同形状的缺陷。

我们以 IoU 成本作为基础重叠度量, 确保预测区与真实缺陷区域基本吻合, 是判断是否检出的根本依据。以角度成本针对性解决划痕等方向性缺陷的回归问题。该成本会首先引导模型旋转预测框的方向, 使其与真实缺陷的主轴方向对其。这一步骤为后续距离和形状的优化提供了正确的初始姿态, 极大地加速了收敛过程, 避免了模型在错误方向上进行无效优化。使用自适应距离成本解决不同尺寸缺陷的距离敏感度一致性问题。该成本意味着对于大缺陷, 其中心点在 X 轴上的绝对距离容忍度更高。而对于微小缺陷, 其中心点偏移的惩罚则更为严厉。以解耦形状成本分别独立地优化缺陷的宽度和高度, 精准匹配各类形状。添加纵横比成本, 作为形状成本的补充, 从整体比例上防止模型预测出形状完全畸变的预测框, 确保合理性。

综合所有组件, 最终的 SIoU 损失函数为等式 6, 具体展开为等式 7, 其中 β 是一个超参数, 用于控制纵横比成本项的权重。SIoU 结构图如图 4 所示。

$$L_{SIoU} = L_{IoU} + \Delta + \Omega + L_{asp} \quad (6)$$

$$L_{SIoU} = 1 - IoU + \gamma \cdot (\rho x + \rho y) + [(1 - e^{-\omega w})^\beta + (1 - e^{-\omega h})^\beta] + \lambda \left| \frac{b_w}{b_h + \epsilon} - \frac{b_w^{gt}}{b_h^{gt} + \epsilon} \right| \quad (7)$$

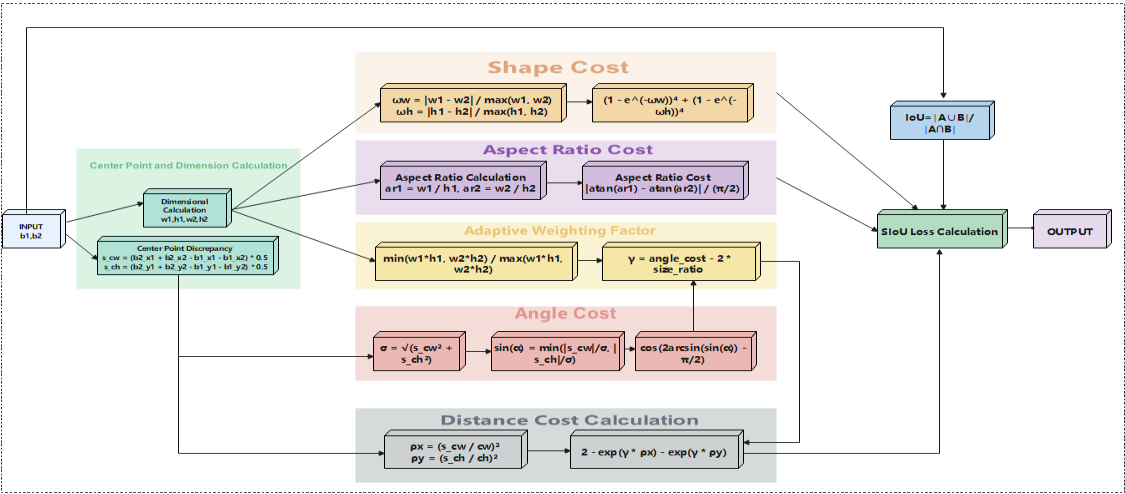


Figure 4. ASA-SIoU Architecture Diagram

图 4. ASA-SIOU 结构图

4. 结果与分析

4.1. 实验环境

本文选择的试验硬件及配置如表 1 所示：

Table 1. Training Environment

表 1. 训练环境

软件、硬件配置	版本参数
操作系统	Windows 10
GPU	nvidia GeForce GTX 4060 8G
CPU	英特尔 第十三代酷睿 i7-13560HX 十四核
内存	32GB DDR4 3600MHz
Python	3.8.19
Torch	2.3.0
CUDA	13.0
CuDNN	8.8.1

在训练之前需修改 yolov8n.yaml 数据集配置文件。本文选择的 YOLOv8 参数设置如表 2 所示：

Table 2. Training Parameters

表 2. 训练参数

实验参数	参数值
迭代次数	100
批量大小	8
初始学习率	0.01
权重衰减	0.0005
输入尺寸	640×640
优化器	SGD (Stochastic Gradient Descen)

4.2. 实验数据集

在金属板材生产过程中，会产生各类复杂缺陷。检测模型需要具备能够应对复杂缺陷的能力以及泛化能力。因此本文选择东北大学公开的 NEU-DET 金属缺陷数据集。其聚焦于冷轧带钢表面，包含了在

工业生产过程中的六种典型缺陷类型。数据集共包含 1800 张灰度图像，均匀涵盖了六种不同缺陷类别，分别是裂纹(Crazing)、夹杂(Inclusion)、斑块(Patches)、点蚀(Pitted Surface)、轧扎氧化皮(Rooled-in Scale)和划痕(Scratches)。缺陷样式如图 5 所示。

该数据集具有以下几个显著特点，同时也构成了算法研究的挑战。类别不平衡性：每类缺陷虽然数量均匀，但内部形态多变，且特征细微。缺陷尺寸多样性：数据集中缺陷的尺度变化很大，对模型的多尺度特征提取能力提出了较高要求。背景纹理干扰：金属表面本身具有复杂的纹理背景，难以区分，极易造成漏检。形态多样性：同类缺陷在形状、大小、方向和位置上存在差异，要求模型具备强大的泛化能力。为公平评估模型性能并避免过拟合，本文将数据集按 8：1：1 的比例随即划分为训练集、验证集和测试集。

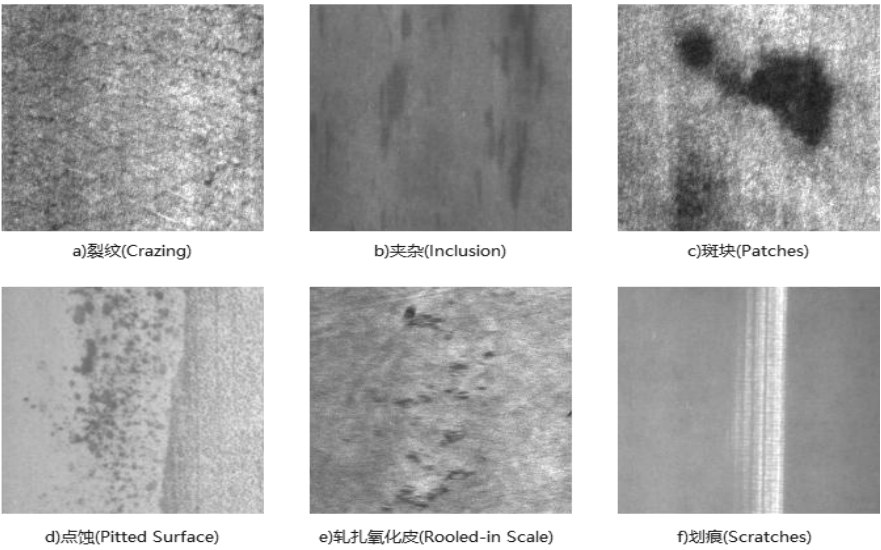


Figure 5. Dataset defect categories

图 5. 数据集缺陷种类

4.3. 实验评价指标

为了检验模型性能，本文在模型检测效果上采用精确率 P (Precision)、召回率 R (Recall)以及 mAP_{50} ($IOU=0.5$ 的均值平均精度)的值来作为评价指标；在模型轻量化检测速度上以参数量、模型体积 (Model Size)、浮点运算量(Flops, Floating Point Operations per Second)

以及每秒帧数(FPS)作为评价网络性能的指标。

评估缺陷检测质量的混淆矩阵，如表 3 所示：

Table 3. Detection metrics confusion matrix

表 3. 检测指标混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

4.4. 实验结果与分析

4.4.1. NEU-DET 对比试验

为了验证本改进网络的整体性能，我们将 PPMA-YOLO 与主流算法对 NEU-DET 数据集下进行对比实验。结果如表 4、表 5 所示。

Table 4. Comparison of lightweight experiment results for various algorithms

表 4. 各类算法轻量化实验结果对比

Model	Parameters/M	FLOPs/G	FPS/s
Faster-RCNN	137.09	190	13
SSD	26.285	62.7	22.6
YOLOv3	103.669	282.2	46.1
YOLOv5	2.504	7.1	48.8
YOLOv6	4.24	4.3	72.46
YOLOv7	37.620	106.5	27.9
YOLOv8	3.007	8.1	45.5
PPMA-YOLO(Ours)	2.367	6.44	111.11

Table 5. Comparison of detection experiment results for various algorithms

表 5. 各类算法检测实验结果对比

Model	mAP ₅₀ /%	Precision / %	Cr	In	Pa	Ps	Rs	Sc
Faster-RCNN	75.47	42.28	36.0	84.0	95.0	86.0	64.0	88.0
SSD	69.2	57.8	36.8	79.2	85.4	69.5	61.9	82.2
YOLOv3	64.9	63.7	41.4	78.7	86.5	51.0	50.1	81.7
YOLOv5	73.7	65.4	45.5	82.3	90.8	77.0	63.1	83.7
YOLOv6	72.7	73.0	40.6	84.3	89.4	80.5	63	90.2
YOLOv7	75.3	66.1	45.1	81.5	92.0	75.1	62.9	89.1
YOLOv8	75.57	71.27	46.8	79.5	88.7	74.5	64.2	90.1
PPMA-YOLO(Ours)	83.87	80.4	65.8	79.1	89.8	87.4	67.7	92.5

从表 4 实验结果可以看出, 本算法与上述经典和主流算法相比, 模型的参数量都小于其他经典算法, 浮点数较 YOLOv6 高 2.14, 但 FPS 均高于上述算法, 检测速度有大幅提升, 且模型参数量小更便于部署。从表 5 实验结果可以看出, 在精度 mAP50 方面, 本文提出的 PPMA-YOLO 表现最佳, 达到了 84.5%, 精度达到了 80.4。其次是 YOLOv7, 它的精度 (Precision) 和平均精度均值 (mAP50) 也相对较高。相较于本文模型, 双阶段 Faster-RCNN 检测算法不仅复杂度较高, 且检测效果较差, 难以满足对金属表面缺陷快速检测的需求。而 YOLO 系列虽然检测精度和速度都相较于双阶段有所提升, 但本文 PPMA-YOLO 算法基于 YOLOv8 模型基础上改进, 精度达到 80.4, FPS 达到 111.11, 兼顾了精度和速度, 效果高于 YOLO 基础模型。

4.4.2. 消融实验

为了验证每个改进模块的有效性, 本实验采用逐步增添改进模块的方法进行消融实验, 模型检测精度结果如表 6, 模型轻量化结果如表 7。

Table 6. Ablation results of detection accuracy

表 6. 检测精度消融结果

Num	YOLOv8n	PPLC-NET	MAFR	SIOU	mAP ₅₀ /%	mAP ₉₅ /%	P%	R%
1	√				75.57	42.27	71.27	66.05
2	√	√			73.79	49.56	73.54	72.78
3	√		√		79.36	44.75	80.73	67.91
4	√			√	77.80	44.28	68.07	73.80
5	√	√	√	√	83.87	50.33	80.60	76.98

Table 7. Ablation results of model lightweighting

表 7. 模型轻量化消融结果

Num	YOLOv8n	PPLC-NET	MAFR	SIOU	Model size/MB	Flops/G	Params/M	FPS
1	√				5.38	8.1	3.007	45.5
2	√	√			4.63	6.21	2.285	132.73
3	√		√		5.55	7.54	2.774	93.26
4	√			√	5.48	7.32	2.691	155.06
5	√	√	√	√	4.8	6.44	2.367	111.11

通过实验结果可以看出, 实验 1 为基线模型, 检测精度和速度都较低。实验 2 为将 PPLC-NET 模块替换主干, 主要用于轻量化模型, 检测精度虽然下降 1.78%, 但 Fps 有较高提升。实验 3 将整合后的 MAFR 模块添加到颈部, 旨在提高检测精度, mAP50% 高于基线模型 3.79, 同时 Fps 有所提升, 但相较于实验 2, Fps 较低。实验 4 为改进 IoU 为自适应形状感知损失函数, 加快模型应对复杂缺陷时的收敛速度, 如表 7 所示, 实验 4 的 Fps 最高。实验 5 将上述模块共同添加到基线模型中, 相较于原始 YOLOv8n, 在精度和速度上都有所提升。精度方面, P(Precision)、R(Recall)、mAP50% 分别提高了 9.33%、10.93%、8.3%。同时模型大小、浮点数运算量(Flops)分别减少了 0.58/MB、1.66/G。FPS 提高了 65.61。这些改进使得新的网络结构更加轻量化, 便于部署, 同时检测精度高, 更满足于金属表面缺陷检测要求。

4.4.3. 结果可视化分析

为了更直观地可视化分析本文模型与 YOLOv8n 原始模型对 NEU-DET 数据集下的金属表面缺陷特征的不同关注程度, 使用 Grad-CAM 方队生成热力图, 如图 6。Grad-CAM(Gradient-weighted class activation mapping) 通过计算目标类别相对于卷积层最终输出特征图的梯度, 来得到该特征图中每个通道(channel)的重要性权重, 然后对这些特征图进行加权求和, 生成一张能够突出显示图像中对网络决策最重要区域的热力图, 以此增加模型的可信度和透明度。

通过图 6, 可以直观看出, 原模型热力图存在漏检、误检现象, 如裂纹缺陷(Cr)、点蚀缺陷(Ps)。相比之下, 本文 PPMA-YOLO 模型更加注重缺陷部分, 相较于原模型几乎没有漏检和误检现象。针对其他四类缺陷, 原模型同样检测出缺陷位置, 但相较于本文模型, 热力图中针对缺陷位置颜色更深且缺陷显示更加细腻。此现象表明本文改进模型更加在缺陷处关注度更高, 对缺陷位置特征输出权重分配更为突出。

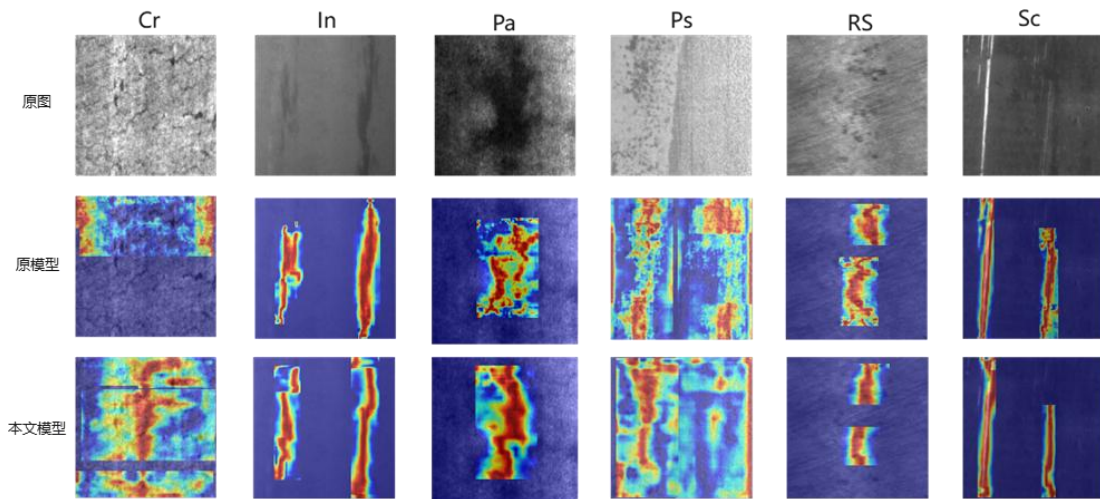


Figure 6. Heatmap comparison

图 6. 热力图对比

图 7 为原模型与本文模型检测结果的对比。图片中方框处为缺陷位置, 与图 6 热力图对应。通过对比可以发现, 原模型检测结果的置信度低于本文模型检测结果的置信度。即便如点蚀(Ps)二者的检测结果相同且置信度都为 0.86, 但通过图 6 热力图对比可以发现本文模型对缺陷处的关注度更高。因此本文模型更适合针对金属表面缺陷提取特征信息, 对缺陷位置检测更精准。

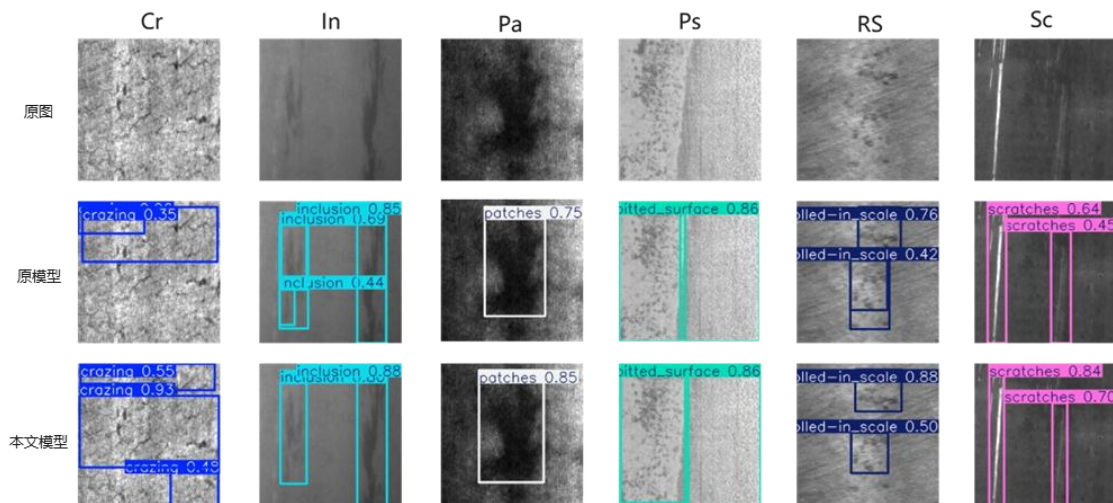


Figure 7. Comparison of results between the original model and the improved model

图 7. 原模型与改进模型检测结果对比

5. 结束语

本文提出了一种基于 YOLOv8n 模型的轻量化金属表面缺陷检测模型, 名为 PPMA-YOLO。通过借鉴 MobileNet 的思想, 我们采用更轻量化的网络结构 PPLC-NET 替换主干网络, 我们设计了基于 Mca 和多尺度特征融合的 MAFR 模块, 使网络对金属缺陷种类复杂, 尺度变化大等挑战有更强的应对能力。我们改进了模型原始 IoU, 更改为自适应形状感知损失函数, 即使面对复杂缺陷, 模型也能精准检测, 同时可以更快收敛, 提高检测速度。通过以上综合优化策略兼顾了模型在金属表面缺陷检测的精度和速度以及部署能力。

总体而言, PPMA-YOLO 模型在金属表面缺陷检测领域取得了显著进展, 不仅满足了工业生产中对高精度检测的需求, 还为移动端部署提供了可行的解决方案。未来的研究方向包括进一步优化模型的实时性能, 并在更广泛的工业场景中进行验证和推广。

参考文献

- [1] 周济. 智能制造 —— “中国制造 2025” 的主攻方向[J]. 中国机械工程, 2015, 26(17): 2273-2284.
- [2] 刘忠明, 王长路, 张元国, 等. 中国齿轮工业的现状、挑战与 2030 年愿景—《机械工业工程技术路线图》齿轮部分解读[J]. 机械传动, 2011, 35(12): 1-6.
- [3] 龙阳, 肖小玲. 改进 YOLOv8 的金属表面缺陷检测模型[J]. 制造技术与机床, 2024, (08): 187-194. DOI:10.19287/j.mtmt.1005-2402.2024.08.027.
- [4] Li H, Li J, Wei H, et al. Slim-neck by GSConv: a lightweight-design for real-time detector architectures[J]. Journal of Real-Time Image Processing, 2024, 21(3). DOI:10.1007/s11554-024-01436-6.
- [5] 赵佰亭, 张敏, 贾晓芬. YOLOv8n-CSG: 轻量化钢材表面缺陷检测算法[J]. 电子测量与仪器学报, 2025, 39(08): 115-125. DOI:10.13382/j.jemi.B2508110.
- [6] Guo S, Zhong P, Sun Y, et al. Fast detection algorithm for surface defects of metal parts based on YOLOv4-mobilenet network[C]//International Conference on Intelligent Equipment and Special Robots (ICIESR 2021). SPIE, 2021, 12127: 528-534.
- [7] Wu T Y, Tang S, Zhang R, et al. Cgnet: a light-weight context guided network for semantic segmentation[J]. IEEE Transactions on Image Processing, 2020, 30: 1169-1179.
- [8] 赵小虎, 谢礼逊, 慕灯聪, 等. 基于 TCM-YOLO 网络的金属表面缺陷检测方法[J]. 计算机工程, 2025, 51(06): 338-348. DOI:10.19678/j.issn.1000-3428.0069285.

-
- [9] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.
 - [10] 陶友凤,姜麟,大妹.基于 RCD-YOLO 的钢材表面缺陷检测算法[J/OL].控制工程,1-9
 - [11] 任伟建,陈明文,康朝海,等.基于 YOLOv3 的金属表面缺陷检测研究[J].控制工程,2024,31(07):1219-1228. DOI:10.14107/j.cnki.kzgc.20220773.
 - [12] 叶卓勋.基于深度学习的金属小目标外观缺陷检测.2023.浙江大学,MA thesis. doi:10.27461/d.cnki.gzjdx.2023.001743.
 - [13] 王少杰,朱皓然.基于轻量化 YOLOv8s 算法的金属表面缺陷检测研究[J].邵阳学院学报(自然科学版), 2025,22(02):11-18.