OEE | 光电工程 Opto-Electronic Engineering

Article

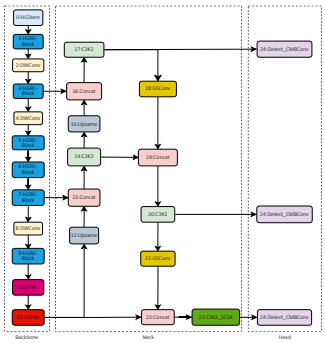
2026 年 3 月, 第 53 卷, 第 3 期

RGCE-YOLO：轻量化钢材表面缺陷检测算法

曾　志[1], 史晓亮[1], 严　赫[1], 吴超华[1*], 孙存甫[2]

[1]武汉理工大学机电工程学院，湖北 武汉 430070；

[2]武汉小出钢管有限公司，湖北 武汉 430090



摘要：针对工业钢材表面因缺陷种类多、缺陷尺寸跨度大，导致传统检测算法检测精度低、模型复杂度高、部署困难等问题，提出一种改进 YOLOv11n 的轻量化钢材表面缺陷检测算法 RGCE-YOLO。首先，重构主干网络，结合轻量化网络 HGNet (high performance GPU network) 与改进重参数化 RepLightConv 将多分支卷积过程转换为单路结构，保证检测精度的同时压缩模型体积。其次，颈部嵌入轻量化 GSConv (group shuffle convolution) 替代原有下采样减少冗余信息，融合空间与通道协同注意力机制 (spatial and channel synergistic attention, SCSA) 重新设计 C3K2 结构，提高多尺度目标检测精度。然后，结合倒置瓶颈结构与坐标注意力 (coordinate attention，CA) 机制设计 CMBC_Detect 模块引入回归分支中，压缩结构同时增强对微小目标检测能力。最后，通过逐通道蒸馏 (channel-wise knowledge distillation, CWD) 使模型聚焦于显著信息，减少特征损失，无损提高精度。实验结果表明，RGCE-YOLO 算法相比 YOLOv11n 基准模型精度提升了 3.9%，达到了 83.4%，mAP50 和 mAP50-95 分别提升了 3.9% 和 4.6%，参数量下降了 34.5%，FLOPs 下降了 33.3%，实现了精度与轻量化的双重优化，在 NEU-DET 与 GC10-DET 缺陷数据集上也获得了优异的泛化性能，为工业实时检测提供了可靠的技术解决方案。

关键词：钢材表面；缺陷检测；知识蒸馏；轻量化；注意力机制；YOLOv11n

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

DOI: 10.12086/oee.2026.250240 | CSTR: 32245.14.oee.2026.250240 | 中图分类号：TP391.4 | 文献标识码：A

曾志，史晓亮，严赫等. RGCE-YOLO：轻量化钢材表面缺陷检测算法 [J]. 光电工程，2026，**53**(3): 250240

Zeng Z, Shi X L, Yan H et al. RGCE-YOLO: lightweight steel surface defect detection algorithm[J]. *Opto-Electron Eng*, 2026, **53**(3): 250240

1　引　言

随着智能制造和工业 4.0 的快速发展，工业产品生产质量要求愈发严格。钢材作为工业制造基础性材料，在工业生产中被广泛使用。作为质检过程中的重要环节，钢材表面缺陷因具有尺度跨度大、纹理特征复杂等特点，人工检测耗时久、成本高且易漏检。传统检测方法如人工目检、基于机器视觉的阈值分割或纹理分析等不仅效率低、易误检且泛化性差，无法满足高精度、高效率的工业场景需求。

近年来，深度学习技术在缺陷检测方面飞速发展，能够自动提取复杂特征并实现高精度分类与定位。袁宏强等[1]通过知识蒸馏分别选用 ResNet50 和 MobileNetV3 作为教师模型和学生模型，通过教师模型的经验辅助学生模型提高学生模型精度。李鑫等[2]引入 GhostBottleneck 使网络体积缩小，嵌入通道注意力 (squeeze-and-excitation,

收稿日期：2025-08-16；修回日期：2025-11-11；录用日期：2025-11-12；网络出版日期：2026-03-19

基金项目：国家自然科学基金资助项目 (52375201)；武汉理工大学产学研科技合作项目 (202501hx0411)

*通信作者：吴超华，wuchaohua@whut.edu.cn

版权所有©2026 中国科学院光电技术研究所

SE) 机制加强特征提取能力。刘琪等[3] 使用融合残差网络和密集连接网络提高了钢材表面缺陷的提取能力。曹义亲等[4] 以带残差边的 SPP_Res 特征金字塔结构，提高模型收敛速度，通过采取多头注意力机制，提取目标更多信息。齐向明等[5] 将特征融合网络的张量拼接操作与加权双向特征金字塔结合，在 NEU-DET 数据集上取得了较高的精度。崔克彬等[6] 针对微小物体检测，提出了四头自适应特征融合 (adaptively spatial feature fusion, ASFF) 预测头，加入改进后的混合注意力 (convolutional block attention module，CBAM) 机制加强对钢材表面小目标缺陷的提取。梁礼明等[7] 通过重构轻量化特征交互模块结合多谱注意力机制，在提高检测精度的同时缩减了模型体积。高春艳等[8] 融合级联注意力机制和解耦头重新设计 YOLOv7 检测头网络，平衡各尺度缺陷样本对 Loss 的贡献，降低了各尺度缺陷的漏检率。王祺等[9] 采用 K-means++算法聚类生成自适应锚框，替换 Hard Swish 激活函数，减少了计算成本和参数量。Yan 等[10] 在主干网络和颈部网络之间集成通道先验卷积注意力机制，通过多尺度简单高效的锚点匹配头，缓解了不同类型缺陷之间的重叠而引起的精度问题。Zeng 等[11] 在颈部网络中加入轻量级 VovGscsp 跨级融合模块减少语义损失，嵌入自注意力机制提高了对小目标的检测能力。马玉洁等[12] 压缩 P2 层特征并将其与 P3 层特征深度融合，有效增强模型对小目标特征的敏感性，通过引入 SIoU 损失函数优化边界框回归，提升定位精度。

综上，基于深度学习的检测算法在钢材表面缺陷检测方面已经获得了许多成果，但在模型复杂度与检测精度的平衡性方面依旧存在进步空间。现有深度学习模型普遍体积大、计算复杂度高，难以部署在资源受限的工业边缘设备中，限制了实时检测场景中的应用。因此，本文基于 YOLOv11n 设计了一种轻量化钢材表面缺陷检测算法 RGCE-YOLO，旨在保证检测精度的同时，压缩模型体积，减少计算复杂度，解决现有深度学习模型在工业场景中资源率占用过高的问题。本文主要工作如下：

1) 主干网络构建一种新型结构 HGRCBlock，将轻量化网络 HGNet 与改进重参数化 RepLightConv 相结合替换原有网络结构，在推理时通过将不同尺寸的卷积与归一化步骤融合保持精度不降的同时降低运算量。

2) 颈部网络嵌入轻量化 GSConv 模块，引入混合卷积缓解深度可分离卷积精度瓶颈问题，针对钢材表面缺陷尺度跨度大的问题在深度特征图层添加 SCSA 注意力机制重构 C3K2 模块，通过共享多语义空间注意力 (shareable multi-semantic spatial attention, SMSA) 和顺序连接的渐进式通道自注意力 (progressive channel-wise self-attention, PCSA) 缓解不同子特征之间的语义差异，促进信息融合。

3) 回归分支中基于 EfficientNet 移动倒置瓶颈结构构建 CMBC_Detect 模块，通过扩展-变换-压缩的方式升维再降维增强特征表达能力。替换 SE 注意力为 CA 注意力以将信息嵌入通道注意力，强化空间特征提取能力，增强对微小目标的特征响应并减少计算成本。

4) 将构建的模型进行逐通道蒸馏，每个通道的特征图归一化得到软概率图，最小化两个网络的通道之间的 KL (Kullback-Leibler) 散度，使学生模型模仿教师模型关注每个通道最显著的区域，减少特征损失，提高模型在密集目标中的捕捉能力。

2 RGCE-YOLO 改进算法

2.1 RGCE-YOLO 整体架构简介

本文提出的 RGCE-YOLO 模型在 YOLOv11n 的基础上进行了多方面的优化，结构如图 1 所示。首先，在主干网络构建了一种新型结构 HGRCBlock，将轻量化网络 HGNet 与改进重参数化 RepLightConv 相结合替换原有网络结构。其次，在颈部网络引入 GSConv 模块减缓数据并针对钢材表面缺陷尺度跨度大的问题在深度特征图层嵌入 SCSA 注意力机制。头部采用 EfficientNet 中的移动倒置瓶颈结构并结合 CA 注意力机制进行重构，将信息嵌入通道注意力加强对空间特征的提取。

2.2 HGRCBlock 模块设计

YOLOv11n 主干采用 C2PSADarkNet 结构，它通过多层卷积与反卷积进行特征提取，利用残差连接与瓶颈结构减少模型复杂度[13]。但深层网络下采样次数过多，高语义特征图分辨率较低，易丢失小目标的细节特征，随模型深度增加，重复的残差结构将引入冗余计算。为提高特征提取能力减少计算成本，通过改进重参数化卷积 RepLightConv 结合轻量化网络 HGNet 设计了 HGRCBlock 模块替换原有主干，如图 2(a) 所示。HGNet 作为主干的基础骨架以层次化聚合的方式在输入前端创建了多个并行结构，以缓解原有链式结构中常见的梯度消失问题，不同阶段中生成的多尺度特征图可以更好地保留不同抽象层次的特征信息。

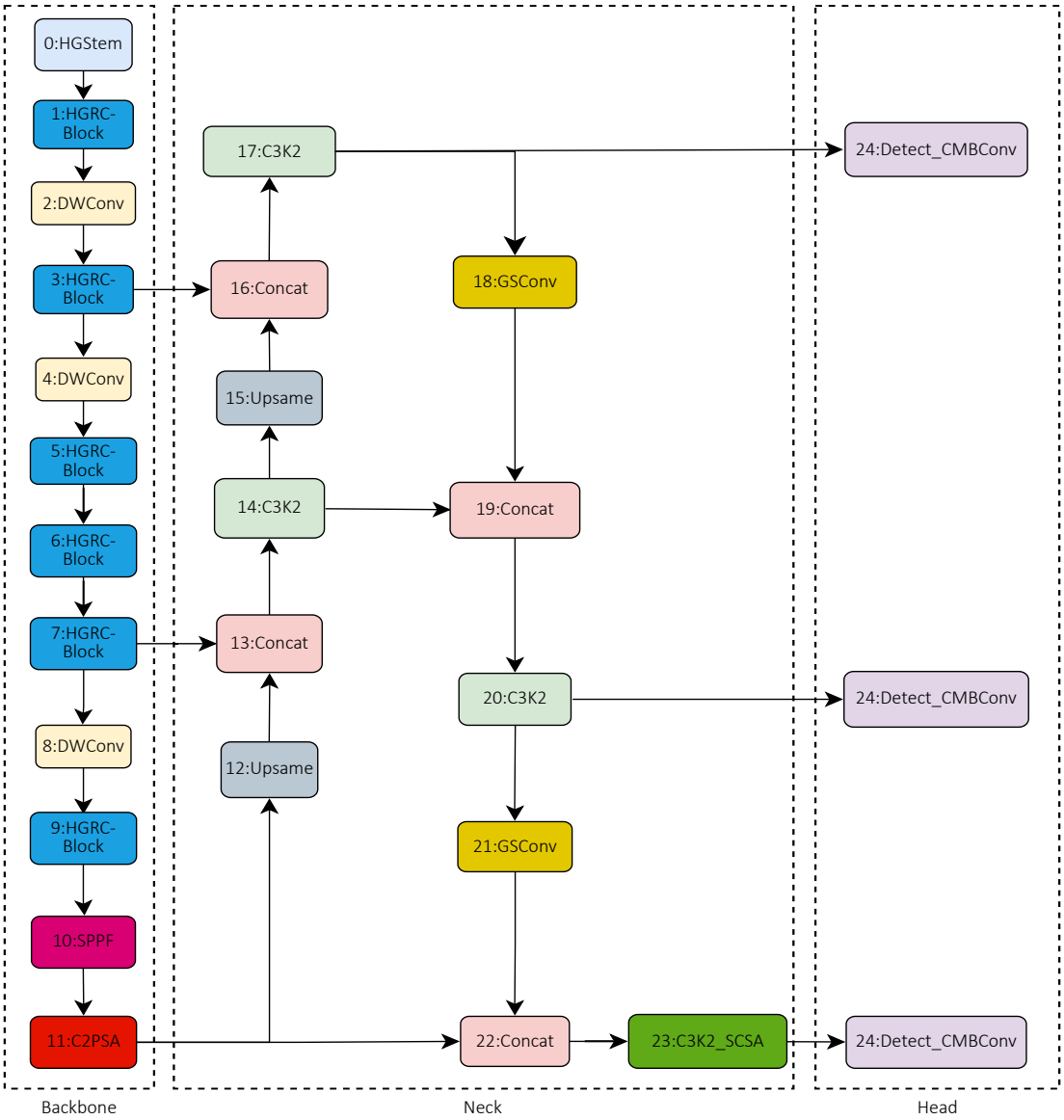


图 1 RGCE-YOLO 网络结构图

Fig. 1 Structure of the RGCE-YOLO network model

HGRCBlock 模块由 RepLightConv 组成的多分支卷积堆叠相互拼接内嵌逐点卷积构成，如图 2(b) 所示，原始的 RepConv 结构由一个并行的 3×3 卷积、1×1 卷积和批归一化 (batch normalization，BN) 分支组成，本文引入 LightConv 思想进行通道压缩，在每个分支的卷积操作之前，以一个 1×1 卷积的逐点卷积作为压缩层，通过减少输入 RepLightConv 的通道数，大幅降低后续所有并行分支的计算量和参数量。然后以多分支保持丰富的特征表征能力，并融合 BN 层转换成 3×3 的卷积层。卷积层计算公式为

$$\mathrm{Conv}(x)=\theta(x)+b\,, \tag{1}$$

式中：x 表示输入；θ表示卷积核权重；b 表示偏置。

BN 层计算公式为

$$BN(x)=\gamma\frac{x-\mu B}{\sqrt{\delta^2+\varepsilon}}+\beta\,, \tag{2}$$

式中：γ表示缩放因子；μB表示均值；ε表示一个很小的数值；δ^2表示方差；β表示偏差。

卷积层和 BN 层融合公式为

$$BN_{\gamma,\beta}(x)=\frac{\gamma W(x)}{\sqrt{\delta^2+\varepsilon}}+\left(\frac{\gamma(b-\mu B)}{\sqrt{\delta^2+\varepsilon}}+\beta\right). \tag{3}$$

重构后的 RepLightConv 在训练时通过多分支结构共同学习获得更强的表征能力，推理时三分支相加转换

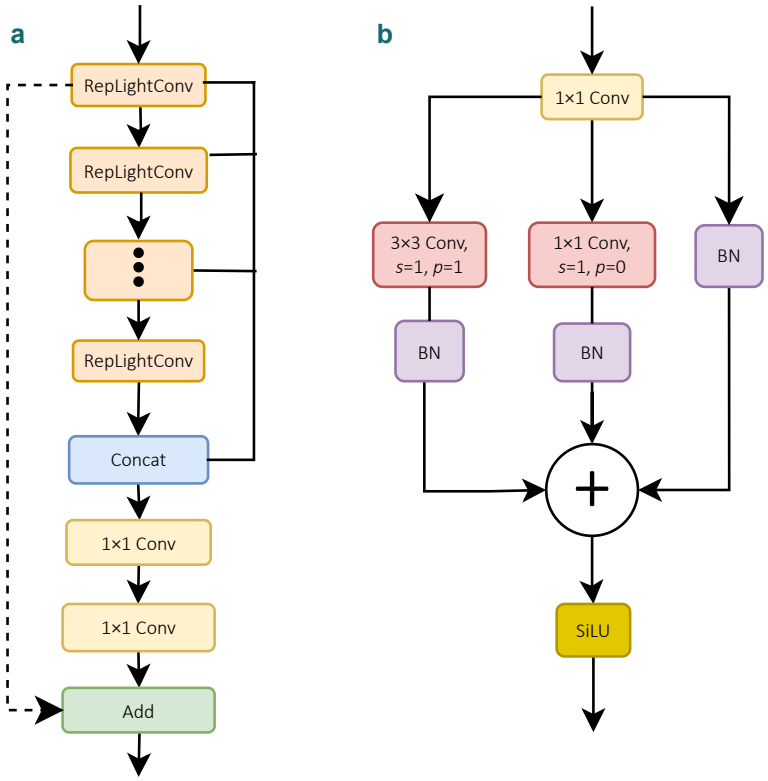


图 2 HGRC-Block 结构图。 (a) HGRC-Block 整体结构； (b) RepLightConv 结构

Fig. 2 Structure diagram of the HGRC-Block. (a) Overall structure of the HGRC-Block; (b) RepLightConv structure

为轻量化的单路结构，省去多分支内存访问与写入，计算浮点数大幅下降，使模型更快、更省内存且更加灵活。

2.3 颈部结构改进

YOLOv11 的颈部结构结合了特征金字塔和路径聚合网络的双向特征，通过融合主干不同层级特征，加强不同尺寸目标特征融合能力，但卷积的密集连接和跨层级融合会产生特征冗余，不同层级特征间的冲突也将降低模型鲁棒性。针对颈部结构特点与问题，对其改进以提高轻量化效果。

2.3.1 GSConv 轻量化混合卷积

GSConv 是一种同时使用 SC (standard convolution)、DSC (depthwise separable convolution) 和通道混洗结构的混合卷积[14]，如图 3 所示。它通过 DSC 层和随机运算增强非线性表达能力，特征图经多层级压缩后，通道维度最大化、空间分辨率最小化，信息密度达到峰值而冗余度显著降低，有效避免传统卷积在通道激增时的计算冗余，同时采用随机运算突破线性表达局限，在信息高度凝练的语义特征层激发出更精细的非线性表征能力。特征流的强语义属性与 GSConv 的跨通道交互特性形成协同，既能强化注意力机制对关键特征的聚焦精度，又避免了浅层部署导致的数据流阻力累积。

输入特征首先经过标准卷积压缩通道并进行初步提取，随后以深度可分离卷积 (DSC) 捕获局部细节。之后，将 DSC 输出的局部特征与标准卷积得到的全局特征相互融合，兼顾不同尺度的信息。最后利用通道混洗促进跨通道信息交互，实现低计算成本下高效的特征提取。SC、DSC 与 GSConv 的浮点运算量 FLOPS 分别如式 (4)、式 (5) 与式 (6) 所示。

$$F_{\mathrm{SC}} = W \times H \times K_1 \times K_2 \times C_1 \times C_2 , \tag{4}$$

$$F_{\mathrm{DSC}} = W \times H \times K_1 \times K_2 \times 1 \times C_2 , \tag{5}$$

$$F_{\mathrm{SC}} = W \times H \times K_1 \times K_2 \times \frac{C_1}{2} \times (C_2 + 1) , \tag{6}$$

式中：W 表示输出特征宽度；H 表示输出特征图高度；K_1 表示卷积核高度；K_2 表示卷积核宽度；C_1 表示卷积核的通道数；C_2 表示输出通道数。GSConv 与 SC 的运算量比率为

$$ration_C = \frac{W \times H \times K_1 \times K_2 \times C_2/2 \times (C_1 + 1)}{W \times H \times K_1 \times K_2 \times C_1 \times C_2} . \tag{7}$$

可得 GSConv 的计算量小于 SC，将颈部网络的

Conv 换成 GSConv 模块有效去除了多余的计算量，减少冗余信息的同时无需压缩，有效增强注意力。

2.3.2 改进 C3K2_SCSA 模块

C3K2 模块是颈部网络中的核心模块，通过选择 C3K 来针对不同特征的提取要求。它采用堆叠多个 Bottleneck 实现不同尺度的特征结合残差结构来保留更多的空间信息，但针对多目标的检测能力不足。为提高模型在多尺度目标的检测精度，对 C3K2 改进添加空间与通道协同注意机制 (SCSA) 聚焦不同尺寸目标，结构如图 4(a) 所示。

其中，SCSA 结构如图 4(b) 所示，它结合了多语义空间注意力 (SMSA) 与渐进式通道注意力机制 (PCSA)，通过多尺度卷积和自注意力机制动态调整特征权重[15]。其核心在于空间-通道协同感知，分阶段优化特征图的关键区域和重要通道。

SCSA 模块包含空间与通道信息处理分支。空间分支对池化后的水平与垂直特征进行分组，并行施加多尺度深度可分离卷积 (multi-Scale depthwise convolution, MS-DWConv) 以提取多尺度空间上下文，经拼接与组归一化生成注意力权重，MS-DWConv 结构如图 5 所示。通道分支则通过查询 ($\boldsymbol{Q}$)、键 ($\boldsymbol{K}$)、值 ($\boldsymbol{V}$) 交互聚合多尺度信息，动态调整通道权重。两部分协同工作，有效提升对尺度变化目标的特征表征能力。计算过程如下所示。

$$A_h = \delta\left(GN_h^k\left(Cat\left(D_h^1, D_h^2, \cdots D_h^i\right)\right)\right), \tag{8}$$

$$A_w = \delta\left(GN_w^k\left(Cat\left(D_w^1, D_w^2, \cdots D_w^i\right)\right)\right), \tag{9}$$

$$SMSA(X) = X_s = X \times A_h \times A_w, \tag{10}$$

式中：A_h 与 A_w 表示沿高度和宽度的权重矩阵；δ 表示 Sigmoid 激活函数；GN_h^k、GN_w^k 表示沿高度和宽度组的

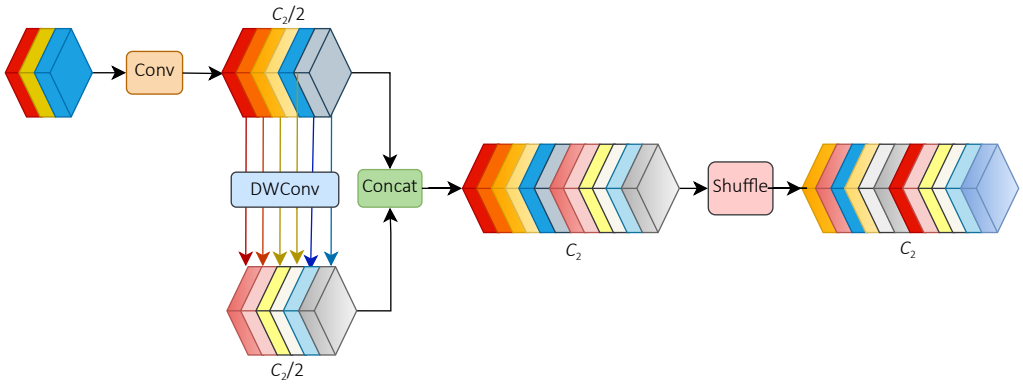


图 3 GSConv 结构图

Fig. 3 The structure of GSConv

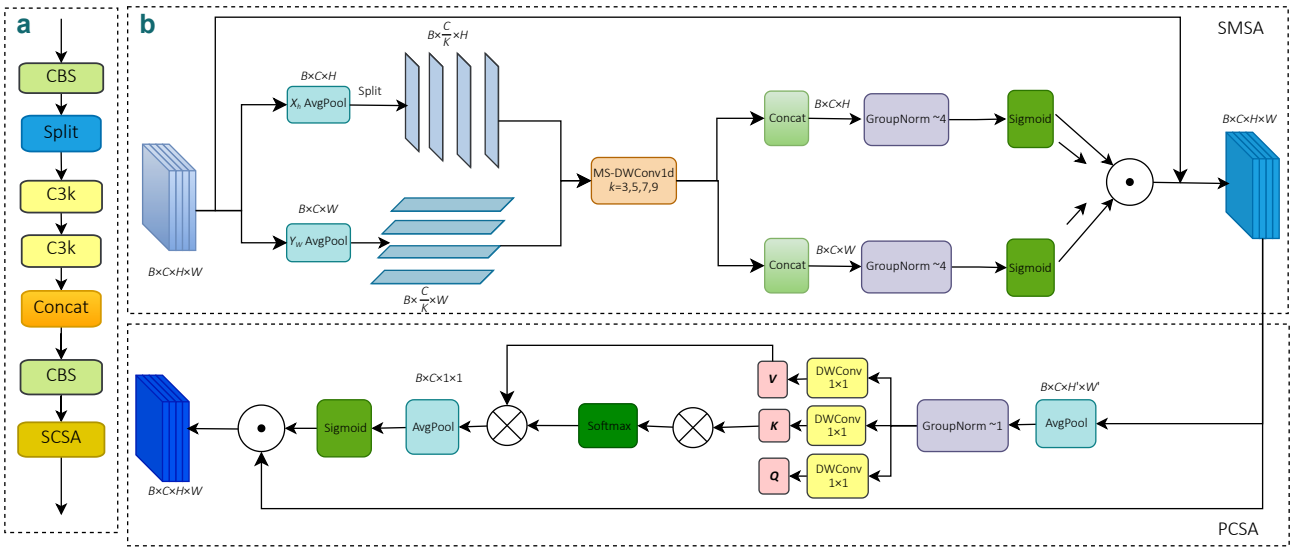


图 4 C3K2_SCSA 结构。(a) C3K2_SCSA 整体结构；(b) SCSA 结构

Fig. 4 Structure of C3K2_SCSA. (a) Overall structure of C3K2_SCSA; (b) SCSA structure

归一化；Cat表示拼接操作；D_h^i与D_w^i表示沿高度和宽度的子特征；i表示分组索引；k表示组归一化的组索引；X表示输入；h表示高度；w表示宽度；$SMSA(X)$为输入特征经过 SMSA 增强后的特征图。

最后，空间权重作用于原始特征通过双向权重 (水平+垂直) 的协同作用，实现精细化空间校准。

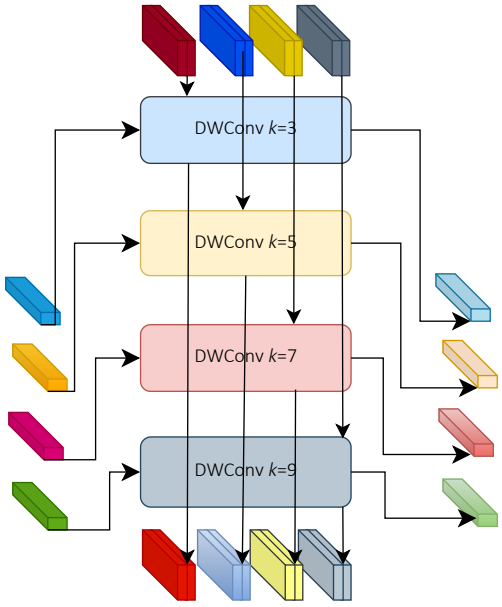


图 5 MS-DWConv 结构

Fig. 5 The structure of MS-DWConv

在 PCSA 模块中，特征在提取空间信息后进入通道注意力路径。首先经过平均池化与组归一化进行压缩，随后使用三组深度可分离卷积分别生成查询 ($\boldsymbol{Q}$)、键 ($\boldsymbol{K}$)、值 ($\boldsymbol{V}$) 张量，减少参数量。通过 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 的点积及 Softmax 归一化得到注意力权重，再与 $\boldsymbol{V}$ 加权求和，实现多尺度信息聚合。最终，经多头合并、全局平均池化与 Sigmoid 激活后，生成通道权重向量，该向量用于校准并增强空间特征。实现过程如下所示。

$$\begin{cases} \boldsymbol{Q} = F_{\text{proj}}^{\boldsymbol{Q}}(X_{\text{p}}) \\ \boldsymbol{K} = F_{\text{proj}}^{\boldsymbol{K}}(X_{\text{p}}) \\ \boldsymbol{V} = F_{\text{proj}}^{\boldsymbol{V}}(X_{\text{p}}) \end{cases}, \tag{11}$$

$$X_{\text{a}} = A(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{C}}\right)\boldsymbol{V}, \tag{12}$$

$$PCSA(X_{\text{c}}) = X_{\text{c}} = X_{\text{s}} \times \delta\left(Pool_{(h^1,w^1)}^{(h^1,w^1)\to(1,1)}(X_{\text{a}})\right), \tag{13}$$

式中：X_{p}表示池化后的特征图；X_{s}表示经过 SMSA 后的特征图；F_{proj}表示生成查询 ($\boldsymbol{Q}$)、键 ($\boldsymbol{K}$)、值 ($\boldsymbol{V}$) 的线性投影；$A(\cdot)$表示加权操作；X_{a}是注意力增强后的特征图；C 表示通道数；T 表示转置；X_{c}表示输出特征图。

特征经过空间注意力模块 SMSA 得到多语义的空间信息，提供给通道注意力模块 PCSA 准确的空间特性减少噪声的干扰。而 PCSA 模块增强特征表达能力进一步减轻冗余通道，二者互相联合共同增强空间注意力与通道注意力特征，表达式如下所示。

$$SCSA(X) = PCSA(SMSA(X)). \tag{14}$$

在 C3K2 结构中嵌入 SCSA 模块能有效增强跨尺度目标的显著性特征并抑制背景干扰。通过空间与通道注意力机制结合，增强全局特征提取能力，对多尺度目标关键特征增加响应，缓解缺陷差异导致的特征丢失，实现跨维度特征优化，提升多尺度目标识别鲁棒性。

2.4 CMBC_Detect 模块

在 YOLOv11 框架中，检测头作为网络的核心输出组件，负责将颈部的多尺度特征图转换为边界框与类别预测。本文设计了 CMBC_Detect 检测头模块，通过结构优化和注意力机制融合，在压缩模型规模的同时抑制漏检现象，增强小目标识别能力。

CMBC_Detect 模块的结构如图 6 所示，包含边界框回归分支和类别预测分支。在检测过程中，来自颈部的多尺度特征图分别输入到两个分支，分类分支采用两个深度可分离卷积与 1×1 卷积相结合的方式，减少参数量的同时增强特征非线性，提升对低对比度目标的分类鲁棒性，降低因特征表达能力不足导致的漏检；回归分支则专注于边界框的分布参数预测，采用两个串联的 CMBConv 模块替换传统卷积，用于提取高细节空间特征和精准编码边界框参数，从而有效改善定位精度与小目标检测效果。

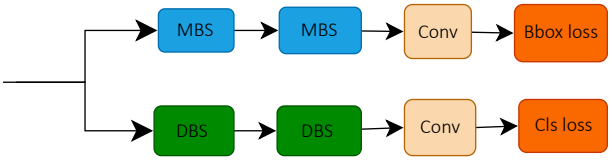


图 6 CMBC_Detect 结构

Fig. 6 The structure of CMBC_Detect

CMBConv 是在 MBConv 基础上面向检测任务的特化改进而来，如图 7 所示。原 MBConv 采用的扩展-深度卷积-压缩结构用于主干特征提取网络，功能定位是分类，其中的 SE 机制在压缩特征时会丢失空间细节信息，不利于小目标定位。为此，CMBConv 将 SE 机制替换

为 CA 注意力机制[16]，使模块能够同时感知通道关系与空间位置信息，从而更准确地捕获目标的位置。这一改进有效缓解了因空间信息丢失所造成的目标漏检，提升了对小尺寸目标和边缘目标的捕捉能力。CA 注意力结构如图 8 所示，处理过程可分为三个阶段，首先，对输入特征图分别沿高度和宽度方向进行特征编码，获取具有方向感知的特征表示，表达式为

$$Z_c^h(h)=\frac{1}{W}\sum_{0\leqslant i<W}x_c(h,i),$$
$$Z_c^w(w)=\frac{1}{H}\sum_{0\leqslant j<H}x_c(j,w), \tag{15}$$

式中：h 表示高度；w 表示宽度；x_c 表示第 c 个通道的特征；Z^h 与 Z^w 分别表示沿高度 h 和宽度 w 的输出。

输出拼接操作后通过一个 1×1 共享卷积进行变换生成特征图 f，f 沿着高度方向和宽度方向进行 Split 操作得到垂直权重 f^h 与水平权重 f^w，再通过逐点卷积和激活函数生成加权的输出特征。该机制使模型能够显式地建模位置敏感的特征响应，增强对细微目标和复杂背景中目标的辨别力，如下所示。

$$g^h=\delta(Fh(f^h)),$$
$$g^w=\delta(Fw(f^w)), \tag{16}$$

$$y_c(i,j)=x_c(i,j)\times g_c^h(i)\times g_c^w(j), \tag{17}$$

式中：g^w 表示沿宽度的注意力权重；g^h 表示沿高度的注意力权重；F 表示卷积操作；x_c 表示第 c 个通道的输入特征；y_c 表示第 c 个通道的输出特征。

改进后的 CMBC_Detect 检测头通过融合坐标注意力机制和分支结构优化，在减少参数量的同时有效增强了模型对目标的空间感知与判别能力，降低了对小目标的漏检率。

2.5 知识蒸馏

知识蒸馏作为一种知识迁移的方法，它可以将学习能力更强的教师模型训练的经验和成果转移到学生模型，在提升小模型精度的同时，还能使小模型的泛化学习能力增强且不增加计算成本。本文采用逐通道蒸馏的方法对模型进一步优化。通过将每个通道的特征图归一化来得到 soft 概率图，以不同模型之间的 KL 散度作为目标并最小化来对通道中关键特征进行关注。分为两个步骤，首先采用 Softmax 标准化，提取特征通道的概率值，用其代表这一通道的重要性或显著性。然后计算这两个模型对应通道概率分布之间的 asymmetric KL 散度作为损失，使学生模型在前景显著区域模仿教师模型。蒸馏过程数学表达式为

$$\phi(\varphi(y^{\mathrm{T}}),\varphi(y^{\mathrm{S}}))=\phi(\varphi(y_c^{\mathrm{T}}),\varphi(y_c^{\mathrm{S}})), \tag{18}$$

式中：y^{T} 与 y^{S} 分别代表教师模型和学生模型的特征图。φ 用于将特征值转换为概率分布，如下所示。

$$\phi(y_c)=\frac{\exp(y_{c,i}/T)}{\sum_{i=1}^{W\cdot H}\exp(y_{c,i}/T)}, \tag{19}$$

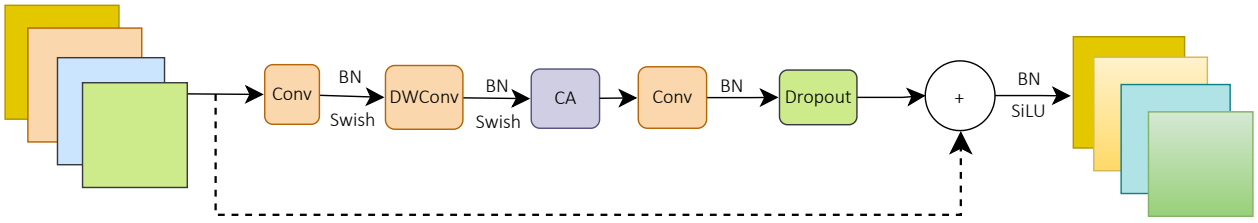


图 7 CMBConv 结构

Fig. 7 The structure of CMBConv

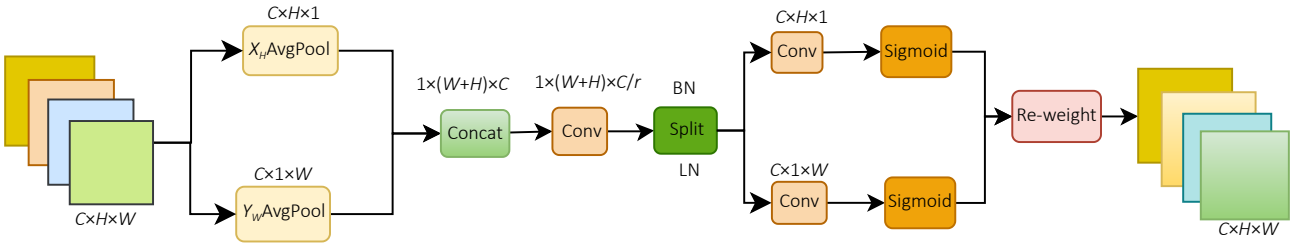


图 8 CA 注意力结构

Fig. 8 The structure of Coordinate Attention

式中：c 和 i 分别表示通道与通道索引；T 表示超参数温度。

特征蒸馏损失用 KL 散度表示，如下式所示：

$$\phi\left(y^{\mathrm{T}}, y^{S}\right)=\frac{T^{2}}{C} \sum_{c=1}^{C} \sum_{i=1}^{W \cdot H} \varphi\left(y_{c, i}^{\mathrm{T}}\right) \cdot \log \left[\frac{\varphi\left(y_{c, i}^{\mathrm{T}}\right)}{\varphi\left(y_{c, i}^{\mathrm{S}}\right)}\right] . \tag{20}$$

蒸馏时学生模型会更趋向于在前景突出特征的位置学习教师模型的分布，而教师模型分布的区域对学习过程造成的影响很小。

3 实验结果与分析

3.1 数据集

本文采用的数据集是在钢材企业在工业环境下通过海康 MV-CH050-10UM 工业相机 (图 9) 收集的常见钢材表面缺陷图片，经筛选后得到 1203 张缺陷图片包含生锈 (Ru)、划痕 (Sca)、凹坑 (Pit) 和裂痕 (Cr) 四种缺陷，如图 10 所示。为提升模型的鲁棒性与泛化能力，研究采用数据增强策略对原始图像进行扩充，随机采用随机裁剪 (保留原图 70%~90% 区域以增强尺度不变性)、动态亮度调整 (0.5~1.5 倍范围以模拟光照变化)、椒盐噪声 (5% 密度模拟脉冲干扰) 与高斯噪声 (标准差 25 模拟传感器噪声)。经扩充后共获得 2406 张缺陷图像，按类别分布为：生锈 604 张、划痕 606 张、凹坑 564 张及裂痕 672 张。所有图像按 8:1:1 比例随机划分为训练集、验证集与测试集，确保各类别分布均衡。

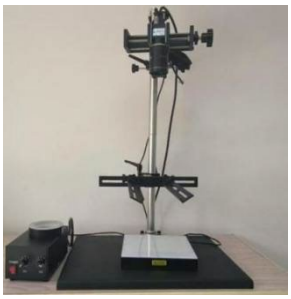


图 9 工业相机

Fig. 9 Industrial camera

3.2 实验环境配置

本文实验是在 Linux 平台上进行的，训练环境为 pytorch 2.0.0，cuda 11.8.0，python 版本为 3.10；采用的硬件设施是 NVIDIA RTX 4090，显存 24 GB，优化器采用随机梯度下降 SGD。训练参数如表 1 所示，为保证实验公平性与可重复性，本文实验所有参数均保持一致。

表 1 实验参数

Table 1 Experimental parameters

Name	Value
Epochs	300
Batch sizes	16
Input resolution	640 pixel×640 pixel
Learning rate	0.01
Momentum	0.937
Weignt decay coefficient	0.0005

3.3 评价指标

本文使用精度 (P)、平均精度均值 (mAP)、参数量 (parameters)、FPS (frames per second) 以及浮点运算次数 (FLOPs) 作为评价指标。mAP 表示在所有目标上的平均精度；mAP50 表示交并比阈值为 0.5 时的值；mAP50-95 表示步长为 5 时多个交并比阈值对应的 mAP 的平均值；Params 表示模型所有可训练参数数量，反映算法复杂度；FPS 是模型每秒处理帧数，值越高推理速度越快；Flops 表示前向传播中的浮点运算量，反映计算复杂程度。其公式为

$$P=\frac{TP}{TP+FP}, \tag{21}$$

$$mAP=\frac{1}{N} \sum_{i=1}^{N} AP_{i}, \tag{22}$$

式中：TP表示正确预测为正类的数量；N表示类别总数；

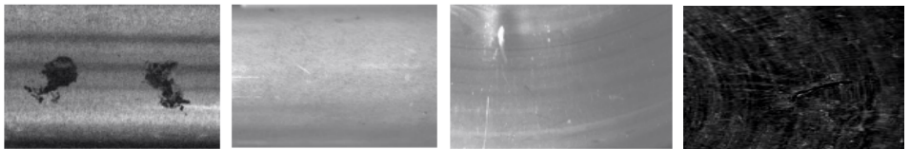


图 10 各类缺陷图片

Fig. 10 Images of various defects

*FP*表示错误预测为正类的数量；*AP*表示单个类别平均精度。

3.4 消融实验与结果分析

为验证改进算法有效性，实验条件相同的情况下，以 YOLOv11n 为基准模型，依次加入各个模块与知识蒸馏进行消融实验，训练过程中精度、mAP50、mAP50-95 变化曲线如图 11 所示，实验结果如表 2 所示，√表示采用该模块，×表示未采用该模块。

从实验结果可以看出，模型 2 在主干网络中引入 HGRCBlock 模块后，尽管精度下降 0.2%，推理速度也略有降低，但 mAP50 和 mAP50-95 分别提升了 0.8% 和 1%，同时参数量和计算量 (FLOPs) 均有所下降。这表明 HGRCBlock 通过压缩通道数有效降低了模型复杂度，其多分支转单分支的结构增强了特征提取能力。模型 7 在颈部网络中使用 GSConv 替代原下采样结构后，精度、mAP50 和 mAP50-95 均得到提高，参数量和 FLOPs 分别减少 0.54 M 和 1 G，冗余信息显著减少，验证了 GSConv 在轻量化方面的有效性。模型 10 采用 SCSA 注意力机制重构 C3K2 模块后，mAP50 达到 77.7%，mAP50-95 提升至 48.1%，参数量降至 2.01 M。SCSA 注意力机制增强了对多尺度目标的感知能力，提升了模型检测鲁棒性，但其动态调整通道权重的机制导致推理速度有所下降。模型 14 检测头采用 CMBC_Detect 模块进行优化后，CA 注意力机制与渐进式通道注意力机制之间的协同作用导致检测速度降低，但精度、mAP50 和 mAP50-95 分别提升至 81.9%、78.8% 和 47.9%，增强了对细小划痕等易漏检目标的捕捉能力，

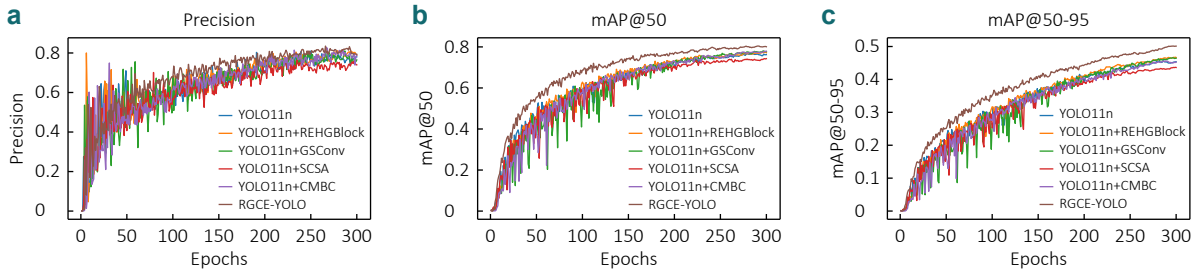


图 11 评价指标曲线图。(a) Precision；(b) mAP50；(c) mAP50-95

Fig. 11 Evaluation metric curves. (a) Precision; (b) mAP50; (c) mAP50-95

表 2 消融实验结果

Table 2 Results of ablation experiments

Model	HGRCBlock	GSConv	SCSA	CMBC	CWD	*P*/%	mAP50/%	mAP50-95/%	Params/M	FPS	FLOPs/G
1	×	×	×	×	×	79.5	76.6	45.7	2.58	208.3	6.6
2	√	×	×	×	×	79.3	77.4	46.7	2.13	185.2	5.7
3	×	√	×	×	×	75.3	78.0	46.6	2.49	204.1	6.3
4	×	×	√	×	×	74.4	74.1	43.1	2.47	188.7	6.2
5	×	×	×	√	×	79.1	77.6	45.6	2.26	212.8	5.1
6	×	×	×	×	√	80.5	78.7	48.7	2.58	201.6	6.6
7	√	√	×	×	×	81.1	77.1	47.4	2.04	181.8	5.6
8	×	√	√	×	×	80.6	76.7	46.5	2.17	180.5	5.0
9	√	×	×	√	×	82.1	78.4	48.1	1.82	175.4	4.5
10	√	√	√	×	×	79.3	77.7	48.1	2.01	187.6	5.6
11	×	×	√	√	×	79.6	78.1	48.6	2.24	181.8	5.1
12	√	√	×	√	×	83.1	78.2	48.5	1.72	175.4	4.4
13	×	√	√	√	×	76.5	74.0	42.8	2.14	191.6	5.0
14	√	√	√	√	×	81.9	78.8	47.9	1.69	175.4	4.4
15	√	√	√	√	√	83.4	80.5	50.3	1.69	185.8	4.4

参数量降至 1.69 M，引入模块前后漏检改善可视化结果如图 12 所示。该模块的倒置瓶颈结构增强了目标的空间感知与判别能力，保留精确的空间位置信息同时进一步压缩了模型规模。最终，通过对改进后的算法施加逐通道蒸馏，精度达到 83.4%，mAP50 和 mAP50-95 分别提升至 80.5% 和 50.3%，FPS 达到 185.8，且模型大小与计算复杂度保持不变，显著提高了模型的泛化能力。与 YOLOv11n 基准算法相比，本文提出的 RGCE-YOLO 算法在精度上提升 3.9%，mAP50 提升 3.9%，mAP50-95 提升 4.6%，参数量下降 34.5%，FLOPs 下降 33.3%。综上所述，RGCE-YOLO 算法在检测精度与轻量化方面实现了协同优化。

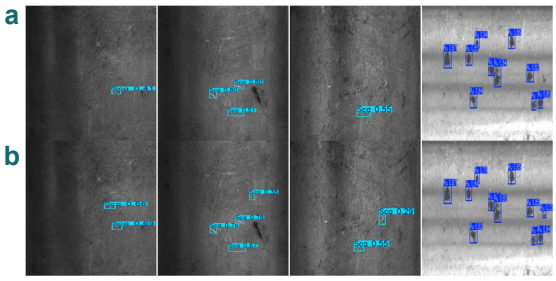


图 12 漏检与改善结果展示。(a) 漏检图片；(b) 改善后

Fig. 12 Missed detections and improvement results. (a) Missed-detection image; (b) Results after improvement

3.5 对比实验

3.5.1 蒸馏对比实验

教师模型作为知识蒸馏中的关键要素，影响着蒸馏的最终效果，为选择合适的教师模型，以 YOLOv11s、YOLOv11m、YOLOv11L 作为教师模型在相同环境下分别进行蒸馏实验，结果如表 3 所示，考虑模型性能与复杂度的平衡最终以 YOLOv11m 为教师模型。

为选择蒸馏方法，将改进后的模型分别采用逐通道蒸馏 CWD、相似性蒸馏 (similarity-preserving, SP)、跨层级蒸馏 (cross-hierarchy similarity distillation, CHSIM) 以及生成式蒸馏 (matching guided distillation, MGD) 在相同环境下进行对比实验，如表 4 所示。

根据实验结果可知模型经过逐通道蒸馏后精度达到了 83.4%，mAP50 达到了 80.5%，mAP50-95 达到了 50.3%，精度提高效果最好，它通过归一化使得模型在学习过程中不丢失空间信息的同时进一步提高了学生模型迁移学习的能力。

3.5.2 主流算法对比实验

为进一步验证改进算法在钢材表面缺陷检测中的有效性，本文选取了几种主流的检测算法在相同实验环境下进行对比实验，结果如表 5 所示。根据实验结果可得，传统 Faster R-CNN[17] 与 RetinaNet[18] 网络模型精度较低且参数量与运算量较大，RGCE-YOLO 在精度与计算方面远优于它们。SSD[19] 与 RT-DETR[20] 模型检测效果也不理想，YOLO 系列算法中 YOLOv8n 精度较高且检测速度较快但其参数量与 FLOPs 都较大，计算程度复杂，需要有较高算力的要求，不满足移动端部署的要求，mAP50 与 mAP50-95 也低于 RGCE-YOLO，模型鲁棒性较差。Mamba-YOLO-T[21] 精度高，检测效果好，但模型体积较大且运算量达到 14.3 G，不易部署。轻量网络

表 3 教师模型蒸馏实验结果

Table 3 Experimental results of teacher model distillation

Model	*P*/%	mAP50/%	mAP50-95/%
YOLOv11s	82.7	81.3	50.6
YOLOv11m	83.4	80.5	50.3
YOLOv11L	83.6	80.4	49.2

表 4 蒸馏方法对比实验结果

Table 4 Comparative experimental results of distillation methods

Distillation method	*P*/%	mAP50/%	mAP50-95/%
CWD	83.4	80.5	50.3
SP	83.0	80.3	49.2
CHSIM	82.1	79.9	49.3
MGD	79.3	80.1	50.1

StarNet[22] 体积较小，精度 mAP50-95 只有 42.0%，检测效果不稳定。轻型网络 MobileNetV4[23] 检测速度最快但检测精度较低只有 66.8%，而模型 NanoDet-Plus-m 参数量与运算量最小但精度只有 76.4%。本文提出的 RGCE-YOLO 速度上相较其他 YOLO 系列算法稍有降低但足以满足工业检测需求，在检测精度上优于其他算法，验证了改进算法检测有效性，并且参数量仅有 1.69 M，运算量 FLOPs 只有 4.4 G，轻量化效果好、压缩程度高，更适用于资源受限环境下的高精度检测需求。

3.6 泛化实验

为验证改进算法检测效果通用性，选取东北大学 NEU-DET 带钢表面缺陷公开数据集与 GC10-DET 工业表面缺陷公开数据集通过数据增强扩充后在相同条件下进行模型泛化性实验，结果如表 6 所示。相比 YOLOv11n 基准模型，在 NEU-DET 数据集上精度提高了 8.4%，达到了 81.2%，mAP50 与 mAP50-95 分别提高了 0.7% 与 0.8%；在 GC10-DET 数据集上精度、mAP50 与 mAP50-95 分别提高了 4.2%、2.2% 与 1.3%，证明 RGCE-YOLO 算法具有良好的适用性与泛化能力。

3.7 ONNX 模型部署验证

由于导出模型的规模相较于模型本身的规模有直接影响。为验证模型在边缘设备部署后实际效果，将模型转换为 ONNX 格式部署在 ARM 工控机上进行检测实验，工控机 CPU 型号为 R7-7745HX，系统为 Linux，采用 ONNX Runtime 进行加速。实验结果如表 7 所示，部署后检测精度为 81.7%、mAP50 为 79.3%、mAP50-95 为

表 5 对比实验结果

Table 5 Comparative experimental results

Model	*P*/%	mAP50/%	mAP50-95/%	Params/M	FPS	FLOPs/G
Faster R-CNN	55.2	56.1	26.1	41.40	43.1	91.10
RetinaNet	60.2	60.2	28.8	32.30	36.4	81.80
SSD	56.0	54.7	26.6	26.28	111.1	62.70
RT-DETR	71.5	64.5	31.0	19.02	178.6	50.60
YOLOv5n	75.4	77.3	43.0	2.18	212.8	5.80
YOLOv8n	82.4	77.4	47.3	3.10	217.4	8.10
YOLOv9t	78.1	73.6	43.8	1.73	205.0	6.40
Mamba-YOLO-T	80.0	76.3	44.3	6.10	131.2	14.30
StarNet	78.1	74.8	42.0	1.90	192.3	5.00
MobileNetV4	66.8	65.9	34.3	1.60	285.7	4.00
YOLOv10n	74.3	74.7	45.1	2.69	210.6	8.20
NanoDet-Plus-m	76.4	75.4	43.6	1.17	220.5	1.52
YOLOv11n	79.5	76.6	45.7	2.58	208.3	6.60
RGCE-YOLO	83.4	80.5	50.3	1.69	185.8	4.40

表 6 泛化性实验结果

Table 6 Experimental results on generalization

Dataset	Model	*P*/%	mAP50/%	mAP50-95/%
NEU-DET	YOLOv11n	72.8	76.3	49.3
	RGCE-YOLO	81.2	77.0	50.1
GC10-DET	YOLOv11n	68.9	66.5	34.0
	RGCE-YOLO	73.1	68.7	35.3

表 7 模型部署实验结果

Table 7 Experimental results of model deployment

Model	*P*/%	mAP50/%	mAP50-95/%	Params/M	FPS
YOLOv11n	77.7	73.2	42.7	2.62	37.3
RGCE-YOLO	81.2	79.3	48.6	1.74	35.1

48.6%，略有下降，参数量为 1.74 M，FPS 为 35.1 满足工业检测要求，证明改进模型能正常部署在边缘设备且检测效果无较大差别。

3.8 可视化分析

为便于具体展示改进后的 RGCE-YOLO 算法的检测效果，将 YOLOv11n 与 RGCE-YOLO 算法对各类缺陷图片检测进行可视化检测验证，如图 13 所示。改进后的算法在各类缺陷检测中都取得了更好的效果，漏检率更低，对小目标缺陷如划痕、生锈等特征提取能力更强，鲁棒性更好。从热力图上也体现出改进后模型在目标识别和定位上效果更好，且边界框位置更为合理，覆盖面积更精确。

4 结 论

针对工业钢材表面缺陷检测困难，传统检测算法无法满足工业检测部署要求的难题，本文提出一种基于 RGCE-YOLO 的轻量化钢材表面缺陷检测算法，对网络结构优化实现精度与轻量化双重提升。主干采用轻量化网络 HGNet 与改进重参数化 RepLightConv 结合降低运算量并提高特征提取能力；颈部引入 GSConv 混合卷积模块突破深度可分离卷积精度瓶颈，去除冗余信息，嵌入 SCSA 注意力机制缓解多尺度缺陷的语义差异强化特征融合，检测头 CMBConv 结构结合 CA 注意力引入回归分支优化定位精度，增强小目标检测能力；最后通过逐通道蒸馏在计算开销不变的同时提升检测精度。实验结果表明，RGCE-YOLO 算法检测精度优于多种主流算法且轻量化效果好，泛化性能强，实现了模型精度与轻量化的平衡。在未来，计划构建更多缺陷类别的数据集，进一步增强对钢材表面各种缺陷类型的识别能力，同时对细小目标进行针对性改进，提高综合检测能力。

作者贡献

曾志：研究资料的收集与分析，实验的设计与完成，论文写作；史晓亮、吴超华、孙存甫：指导论文选题，论文修改，实验器材提供；严赫：数据集的采集，实验结果分析。

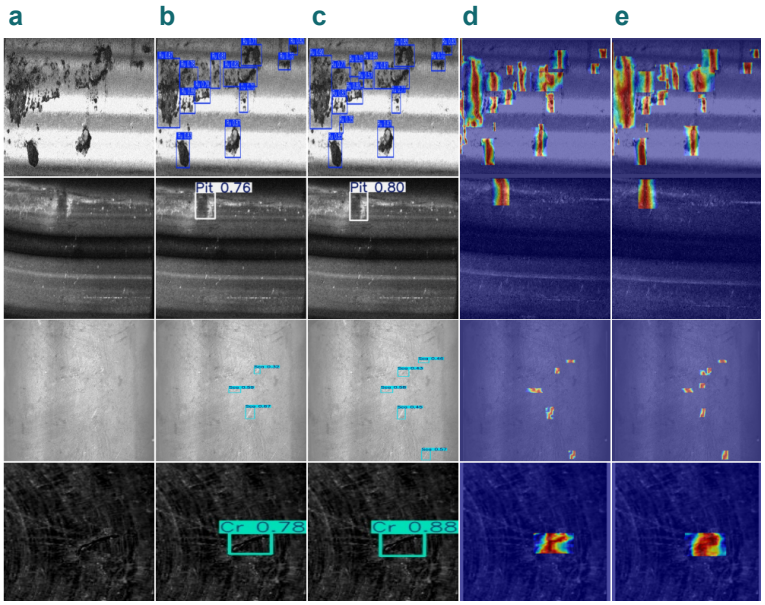


图 13 可视化结果。(a) 原始图像；(b) YOLOv11n；(c) RGCE-YOLO；(d) YOLOv11n；(e) RGCE-YOLO

Fig. 13 Visualization results. (a) Original image; (b) YOLOv11n; (c) RGCE-YOLO; (d) YOLOv11n; (e) RGCE-YOLO

利益冲突

所有作者声明无利益冲突

参考文献

1. Yuan H Q, Du G F, Yu Z Y et al. Fast identification of steel surface defects based on lightweight neural network[J]. *Sci Technol Eng*, 2021, **21** (34): 14651−14656.
袁洪强, 杜国锋, 余泽禹, 等. 基于轻量神经网络的钢材表面缺陷快速识别[J]. 科学技术与工程, 2021, **21** (34): 14651−14656.
2. Li X, Wang C, Li B et al. Steel surface defect detection algorithm based on improved YOLOv5[J]. *J Air Force Eng Univ (Nat Sci Ed)*, 2022, **23** (2): 26−33.
李鑫, 汪诚, 李彬, 等. 改进 YOLOv5 的钢材表面缺陷检测算法[J]. 空军工程大学学报 (自然科学版), 2022, **23** (2): 26−33.
3. Liu Q, Lei J S. Surface defect detection of steel based on improved deep network[J]. *Comput Eng Des*, 2022, **43** (9): 2654−2661.
刘琪, 雷景生. 基于改进深度网络的钢材表面缺陷检测[J]. 计算机工程与设计, 2022, **43** (9): 2654−2661.
4. Cao Y Q, Wu M L, Xu L. Steel surface defect detection based on Improved YOLOv5 algorithm[J]. *J Graphics*, 2023, **44** (2): 335−345.
曹义亲, 伍铭林, 徐露. 基于改进 YOLOv5 算法的钢材表面缺陷检测[J]. 图学学报, 2023, **44** (2): 335−345.
5. Qi X M, Dong X. Improved YOLOv7-tiny algorithm for steel surface defect detection[J]. *Comput Eng Appl*, 2023, **59** (12): 176−183.
齐向明, 董旭. 改进 YOLOv7-tiny 的钢材表面缺陷检测算法[J]. 计算机工程与应用, 2023, **59** (12): 176−183.
6. Cui K B, Jiao J Y. Steel surface defect detection algorithm based on MCB-FAH-YOLOv8[J]. *J Graphics*, 2024, **45** (1): 112−125.
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.
7. Liang L M, Chen K Q, Chen L J et al. Improving the lightweight FCM-YOLOv8n for steel surface defect detection[J]. *Opto-Electron Eng*, 2025, **52** (2): 240280.
梁礼明, 陈康泉, 陈林俊, 等. 改进轻量化的 FCM-YOLOv8n 钢材表面缺陷检测[J]. 光电工程, 2025, **52** (2): 240280.
8. Gao C Y, Qin S, Li M H et al. Research on steel surface defect detection with improved YOLOv7 algorithm[J]. *Comput Eng Appl*, 2024, **60** (7): 282−291.
高春艳, 秦燊, 李满宏, 等. 改进 YOLOv7 算法的钢材表面缺陷检测研究[J]. 计算机工程与应用, 2024, **60** (7): 282−291.
9. Wang Q, Ye R C, Ma G J et al. Lightweight surface defect detection method of steel[J]. *J Mech Eng*, 2025, **61** (8): 18−31.
王祺, 叶仁传, 马国杰, 等. 钢材表面缺陷轻量化识别算法研究[J]. 机械工程学报, 2025, **61** (8): 18−31.
10. Yan C M, Xu E E. ECM-YOLO: a real-time detection method of steel surface defects based on multiscale convolution[J]. *J Opt Soc Am A*, 2024, **41** (10): 1905−1914.
11. Zeng K, Xia Z B, Qian J L et al. Steel surface defect detection technology based on YOLOv8-MGVS[J]. *Metals*, 2025, **15** (2): 109.
12. 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.
13. Saqib S M, Iqbal M, Mazhar T et al. Effectiveness of teachable machine, mobile net, and YOLO for object detection: a comparative study on practical applications[J]. *Egypt Inf J*, 2025, **30**: 100680.
14. Li H L, Li J, Wei H B et al. Slim-neck by GSConv: a lightweight-design for real-time detector architectures[J]. *J Real-Time Image Process*, 2024, **21** (3): 62.
15. Si Y Z, Xu H Y, Zhu X Z et al. SCSA: exploring the synergistic effects between spatial and channel attention[J]. *Neurocomputing*, 2025, **634**: 129866.
16. Hou Q B, Zhou B Q, Feng J S. Coordinate attention for efficient mobile network design[C]//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 2021: 13708–13717.
https://doi.org/10.1109/CVPR46437.2021.01350.
17. 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.
18. Lin T Y, Goyal P, Girshick R et al. Focal loss for dense object detection[J]. *IEEE Trans Pattern Anal Mach Intell*, 2020, **42** (2): 318−327.
19. Liu W, Anguelov D, Erhan D et al. SSD: single shot multibox detector[C]//*Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, 2016: 21–27.
https://doi.org/10.1007/978-3-319-46448-0_2.
20. Zhu W M, Zhu C M. Improved safety helmet detection model based on RTDETR: RTDETR-IHD[C]//*Proceedings of 2024 7th International Conference on Computer Information Science and Application Technology (CISAT)*, Hangzhou, 2024: 895–899.
https://doi.org/10.1109/CISAT62382.2024.10695288.
21. Zhao K, Peng S F, Li Y J et al. A lightweight Xray-YOLO-Mamba model for prohibited item detection in X-ray images using selective state space models[J]. *Sci Rep*, 2025, **15** (1): 13171.
22. Ma X, Dai X Y, Bai Y et al. Rewrite the stars[C]//*Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Hangzhou, 2024: 5694–5703.
https://doi.org/10.1109/CVPR52733.2024.00544.
23. Qin D, Leichner C, Delakis M et al. MobileNetV4: universal models for the mobile ecosystem[C]//*Proceedings of the 18th European Conference on Computer Vision*, Milan, 2024: 78–96.
https://doi.org/10.1007/978-3-031-73661-2_5.

作者简介



曾志 (2000-)，男，硕士研究生，研究方向为智能装备、深度学习。

E-mail：zz346655@whut.edu.cn



【通信作者】吴超华 (1980-)，男，博士，副教授，研究方向为智能制造、机器视觉、材料界面优化设计。

E-mail：wuchaohua@whut.edu.cn

RGCE-YOLO: lightweight steel surface defect detection algorithm

Zeng Zhi[1], Shi Xiaoliang[1], Yan He[1], Wu Chaohua[1*], Sun Cunfu[2]

[1] School of Mechanical and Electronic Engineering, Wuhan University of Technology, Wuhan, Hubei 430070, China;
[2] Wuhan Koide kokan Co., Ltd., Wuhan, Hubei 430090, China

Abstract

Objective Industrial steel surface defect detection is a core step of industrial quality inspection. However, traditional detection algorithms suffer from low accuracy, high model complexity, and difficult deployment due to the various types and large scale variations of steel surface defects. Existing deep learning-based detection algorithms have improved detection accuracy to a certain extent, but most of them feature large model size and high computational complexity, which makes them hard to deploy on resource-constrained industrial edge devices and thus restricts their practical application in real-time industrial detection scenarios. This study aims to propose a lightweight steel surface defect detection algorithm that balances high detection accuracy and low computational complexity, realizes efficient feature extraction for multi-scale steel surface defects, and meets the requirements of real-time detection and edge device deployment in industrial scenarios.

Methods A lightweight steel surface defect detection algorithm named RGCE-YOLO was proposed based on an improved YOLOv11n. The backbone network was reconstructed with a novel HGRCBlock structure combining lightweight HGNet and improved reparameterized RepLightConv, and multi-branch convolution and normalization steps were fused during inference to reduce computational load while maintaining feature extraction capability. For the neck network, the lightweight GSConv replaced the original downsampling structure to reduce redundant information via mixed convolution and address the accuracy bottleneck of depthwise separable convolution; the C3K2 structure was redesigned by integrating the Spatial and channel synergistic attention (SCSA) mechanism to alleviate sub-feature semantic differences and enhance multi-scale object feature fusion. The CMBC_Detect module, constructed based on EfficientNet's inverted bottleneck structure and combined with the coordinate attention (CA) mechanism, was introduced into the detection head's regression branch; the SE attention was replaced with the CA mechanism to retain spatial detail information, enhance small object detection ability, and compress the structure simultaneously. Channel-wise knowledge distillation (CWD) was applied to the improved model: each channel's feature map was normalized to a soft probability map, and the asymmetric KL divergence between the teacher and student models was minimized to make the student model focus on salient regions of defect features, reducing feature loss without increasing computational cost. A self-built steel surface defect dataset with four defect types (rust, scratch, pit, crack) was established and expanded via data augmentation strategies including random cropping, dynamic brightness adjustment, salt-and-pepper noise, and Gaussian noise, and then divided into training, validation, and test sets at an 8:1:1 ratio. Ablation experiments verified the effectiveness of each improved module, and comparative experiments on distillation methods and mainstream detection algorithms were conducted under the same conditions. Generalization experiments were carried out on the public NEU-DET and GC10-DET datasets, and the model was converted to ONNX format for deployment verification on an ARM industrial computer. Precision, mAP50, mAP50-95, parameters, FPS, and FLOPs were used as key evaluation indicators to comprehensively assess the algorithm's performance.

Results and Discussions Ablation experiments confirmed the effective optimization role of each improved module: HGRCBlock reduced model parameters and FLOPs while enhancing multi-scale feature extraction; GSConv reduced the neck network's redundant information and computational complexity; SCSA improved the model's perception of multi-scale objects and detection robustness; CMBC_Detect strengthened the capture ability for small, easily missed defects and further compressed the model size; CWD improved model accuracy without changing model size and computational complexity. On the self-built dataset, RGCE-YOLO achieved a precision of 83.4% (3.9% higher than YOLOv11n), with mAP50 and mAP50-95 increased by 3.9% and 4.6% to 80.5% and 50.3%, respectively; parameters decreased by 34.5% to 1.69 M and FLOPs by 33.3% to 4.4 G, realizing dual optimization of accuracy and lightweight design. Distillation comparison experiments showed that CWD had a better optimization effect than other distillation methods, and YOLOv11m was selected as the optimal teacher model by balancing performance and complexity. Mainstream algorithm comparison experiments indicated that RGCE-YOLO outperformed traditional algorithms (Faster R-CNN, RetinaNet, SSD) and other lightweight algorithms (MobileNetV4, StarNet, NanoDet-Plus-m) in detection accuracy, and had obvious lightweight advantages over other YOLO series

Foundation item: National Natural Science Foundation of China (52375201), and the Industry-University-Research Collaboration Project of Wuhan University of Technology (202501hx0411)

*Correspondence: Wu C H, E-mail: wuchaohua@whut.edu.cn

algorithms, with an FPS of 185.8 meeting industrial real-time detection needs. Generalization experiments showed that RGCE-YOLO achieved a precision of 81.2% (8.4% higher than YOLOv11n) on the NEU-DET dataset and 73.1% (4.2% higher than YOLOv11n) on the GC10-DET dataset, with stable mAP improvements, demonstrating excellent generalization ability. ONNX deployment verification on the ARM industrial computer showed that the model maintained a precision of 81.7% and an FPS of 35.1 with no significant performance decline, proving its adaptability to edge devices. Visual results showed that RGCE-YOLO had a lower missed detection rate, stronger small object feature extraction ability, and more accurate bounding box positioning than YOLOv11n. The excellent performance of RGCE-YOLO stemmed from the synergy of all improved modules: RepLightConv's reparameterization balanced feature representation and computational efficiency; GSConv and SCSA solved feature redundancy and semantic differences in multi-scale defect detection; the CA mechanism in CMBC_Detect retained spatial details and solved small object detection information loss; CWD enhanced dense object feature capture and generalization without additional computational cost.

Conclusions RGCE-YOLO effectively solves the problems of low accuracy and poor lightweight performance of existing steel surface defect detection algorithms, and the synergy of the improved modules achieves dual optimization of detection accuracy and model lightweight design. On the self-built dataset, the algorithm not only has significantly higher precision, mAP50, and mAP50-95 than the YOLOv11n benchmark model, but also greatly reduces parameters and computational complexity, with its inference speed meeting industrial real-time detection requirements. It exhibits excellent generalization performance on public steel surface defect datasets and can be stably deployed on edge devices with no obvious detection performance decline, providing a reliable technical solution for real-time industrial steel surface defect detection. The design idea of combining a lightweight network structure, an attention mechanism, and knowledge distillation in this study also provides a reference for the lightweight optimization of other industrial object detection algorithms. In the future, constructing a dataset with more steel surface defect types and conducting targeted improvements for tiny object defects will further enhance the model's comprehensive detection ability, promoting its wider application in industrial quality inspection scenarios with more complex defect types and higher detection requirements.

Keywords: steel surface; defect detection; knowledge distillation; lightweight; attention mechanism; YOLOv11n

Zeng Z, Shi X L, Yan H et al. RGCE-YOLO: lightweight steel surface defect detection algorithm[J]. *Opto-Electron Eng*, 2026, **53**(3): 250240; DOI: 10.12086/oee.2026.250240

Opto-Electronic Journals Group
www.oejournal.org

扫描二维码，获取 PDF 全文