doi:10.12068/j.issn.1005－3026.2026.20259062

QHN-YOLO:面向端侧部署的钢材表面缺陷检测算法

宋 鹏[1,2]，陈 霄[1]，何 玄[1]，孔德昊[1]
（1.昆明理工大学 民航与航空学院，云南 昆明 650500；2.昆明理工大学 材料科学与工程学院，云南 昆明 650093）

摘 要：针对钢材表面缺陷在复杂纹理、弱对比及端侧部署条件下易发生漏检和误检的问题，提出一种面向端侧部署的检测算法QHN-YOLO. 首先，构建量化友好卷积QFConv以替换原有卷积模块，从而稳定激活分布并提高端侧部署效率. 其次，将混合膨胀残差注意力块(HDRAB)嵌入C3k2模块以构成C3k2_HDRAB，用于增强多尺度特征提取与细小缺陷表征能力. 最后，设计噪声方差先验检测头(NVPD)，以抑制背景干扰并提高缺陷响应的判别性. 实验结果表明，改进后模型的mAP达到80.2%，同时参数量仅为2.8×10^6，浮点计算量为7.1 GFLOPs. 在端侧部署条件下，模型推理时延为28.9 ms，检测速度达到34.6 帧/s，具备工业场景下的实时检测能力. 改进后的模型兼顾了检测精度与轻量化，提升了钢材表面缺陷检测的端侧应用性能.

关 键 词：YOLOv11；钢材缺陷；深度学习；有界激活；端侧部署
中图分类号：TP 391.4 **文献标志码**：A **文章编号**：1005-3026(2026)06-0130-13

QHN-YOLO: Detection Algorithm for Steel Surface Defects Oriented to Edge Deployment

SONG Peng[1,2], *CHEN Xiao*[1], *HE Xuan*[1], *KONG De-hao*[1]
（1. Faculty of Civil Aviation and Aeronautics, Kunming University of Science and Technology, Kunming 650500, China; 2. Faculty of Materials Science and Engineering, Kunming University of Science and Technology, Kunming 650093, China. Corresponding author: CHEN Xiao, E-mail: mitori1027@163.com）

Abstract: To address the issue of missed and false detections of steel surface defects under conditions of complex textures, weak contrast, and edge deployment, a detection algorithm oriented to edge deployment named QHN-YOLO was proposed. Firstly, a quantization-friendly convolution (QFConv) was constructed to replace the original convolution module, so as to stabilize activation distribution and improve edge deployment efficiency. Secondly, a hybrid dilated residual attention block (HDRAB) was embedded into the C3k2 module to form C3k2_HDRAB, which was used to enhance multi-scale feature extraction and small defect representation ability. Finally, a noise-variance prior detection head (NVPD) was designed to suppress background interference and improve defect response discriminability. Experimental results indicate that the mAP of the improved model reaches 80.2%, with 2.8×10^6 parameters and 7.1 GFLOPs of floating-point computations. Under edge deployment conditions, the model inference latency is 28.9 ms, and the detection speed reaches 34.6 fps, indicating real-time detection capability in industrial scenarios. The improved model balances detection accuracy and lightweighting and improves the edge application performance of steel surface defect detection.

Key words: YOLOv11; steel defect; deep learning; bounded activation; edge deployment

钢材作为现代工业的基础材料，凭借其高强 度、耐腐蚀及易加工等优良性能，在建筑、汽车、

收稿日期：2025-12-31
基金项目：国家级大学生创新创业训练计划项目(202410674017S).
作者简介：宋 鹏(1979—)，男，山东青州人，昆明理工大学教授，博士生导师.
通信作者：陈 霄，E-mail:mitori1027@163.com.

航空航天及机械制造等领域得到了广泛应用[1].然而,钢产品在生产过程中不可避免地会产生缺陷,这不仅严重影响了产品的外观质量与可靠性,若未得到有效处理,甚至可能引发安全隐患并造成重大经济损失[2].因此,面向生产线的在线检测需要在轧制纹理密集、对比度不均与尺度变化显著等条件下实现稳定、实时的缺陷识别.

传统的缺陷检测手段主要依赖人工目视检查及漏磁、涡流等无损检测技术[3].近年来,随着计算机视觉技术的发展,机器学习方法逐渐被应用于缺陷检测领域[4].然而,此类方法在处理钢材表面复杂背景及微小缺陷时,仍难以保证检测的可靠性,检测精度偏低且误报率较高[5].

随着深度学习技术的引入,缺陷检测的精度与效率得到显著提升.当前目标检测算法主要分为两阶段和单阶段两大类[6].两阶段算法以 R-CNN[7]系列为代表,如 Faster R-CNN[8],Mask R-CNN[9]等,其流程为先生成候选区域,再进行分类与回归的级联策略,检测精度较高,但计算复杂度大,难以满足实时性要求.相比两阶段检测算法,单阶段检测算法在计算复杂度与实时性方面更适于生产线部署,其中 YOLO 系列在速度与精度之间实现了更好的平衡[10].

为了提升检测精度和效率,现有研究主要围绕特征提取能力、目标检测头以及注意力机制等方面进行优化.An 等[11]提出的 FSU-YOLO 在 YOLOv8 的基础上首次引入频域细节增强和时序注意力机制,提升微小缺陷检测能力并满足实时检测要求.Li 等[12]提出的 HDFA-YOLO 则采用轻量骨干网络、多分支注意力和浅层深层特征融合模块,加强钢材模糊缺陷的特征提取与表达,提高缺陷检测精度.郭业才等[13]在 YOLOv11 中提出 MEB 模块以提升模型的边缘信息提取能力,并设计 DESC 检测头以增强细节信息和减少模型计算量,从而提升了钢材缺陷的检测性能.梁礼明等[14]在 YOLOv8 的基础上设计了轻量化的多分支特征聚合网络,有效精简了模型体积并提升了检测速度.张泽毅等[15]在 YOLOv8 中设计了动态特征提取模块和高效轻量级检测头,通过动态调整各卷积权重,显著提高了不同尺度下缺陷的检测精度并大幅降低了计算成本.胡名琪等[16]针对钢材缺陷检测提出了一种轻量级缺陷检测算法 YOLOv8n-MDC,通过引入 WIoU 损失函数、采用 MobileNetv3 网络等,在检测精度提升的同时大大降低了算法的复杂度和计算开销.郝博等[17]通过引入可变形卷积、SimAM 注意力机制等,实现了目标特征的增强,在提高铆接孔表面缺陷检测精度的同时降低了漏检率.

尽管上述方法在精度或轻量化方面取得进展,但在复杂背景下的误检抑制与端侧部署效率方面仍有提升空间,且对细长缺陷的全局结构建模仍需进一步增强.为此,本文设计 3 个模型改进策略,提出检测模型 QHN-YOLO.

1)针对端侧设备推理的稳定性,提出量化友好卷积模块 QFConv(quantization-friendly convolution),将卷积、归一化和有界激活设计为可融合结构,并结合量化感知训练在训练阶段显式建模量化误差,使训练形态与端侧推理形态对齐,提高端侧设备部署的数值鲁棒性与推理效率.

2)为提高钢材表面微小缺陷的检测能力,引入混合膨胀残差注意力块(hybrid dilated residual attention block,HDRAB),通过多尺度空洞上下文与残差聚合增强微小与细长缺陷的全局结构表征,并以轻量注意力抑制轧制纹理干扰,实现对低对比缺陷更稳健的特征提取.

3)针对背景纹理干扰的问题,提出噪声方差先验检测头(noise-variance prior detection head,NVPD),在分类分支引入轻量门控先验并进行自适应校准,从而降低背景纹理的误检并提高分类鲁棒性.

4)在 NEU-DET 数据集上开展对比实验与消融实验,并在端侧设备上评估推理延迟与部署效率.

1　YOLOv11 算法介绍

YOLO 系列采用单阶段检测框架并支持端到端训练,在保证推理效率的同时可获得较高检测精度.YOLOv11 在网络结构与训练流程上进一步优化,在相近模型规模下提升了检测精度并改善了推理效率[18].此外,得益于其高效的计算机制,YOLOv11 能够便捷地部署于算力受限的设备.然而,针对钢材表面缺陷中常见的复杂纹理背景、弱对比边缘和细长缺陷目标,其特征提取能力、误检抑制能力以及端侧部署稳定性仍存在不足,尚难充分满足工业边缘场景对实时性与鲁棒性的要求.

2 QHN-YOLO模型介绍

YOLOv11在目标检测任务中具备良好的精度与效率,但在钢表面缺陷检测场景中仍面临背景纹理复杂、缺陷形态细长且尺度变化显著等挑战,导致缺陷与背景的特征可分性下降,细微裂纹、划痕等高频细节容易被背景纹理淹没,从而引发漏检与误检.为此,本研究以YOLOv11n为基准构建改进模型,在保持轻量化部署目标的前提下,从卷积算子、特征建模与检测头先验约束三方面进行协同增强.首先,引入量化友好卷积QFConv替换网络中的原有卷积单元,通过更适配量化与端侧设备推理的激活与算子组合,提升细粒度缺陷特征的表达稳定性与部署一致性.其次,将混合膨胀残差注意力块HDRAB引入模型中的C3k2模块,在多尺度残差与空洞上下文的联合建模下扩大有效感受野并强化纹理细节的聚合能力,以提升对不同尺度缺陷的鲁棒表征.最后,将原有检测头替换为噪声方差先验检测头(NVPD),通过噪声与方差先验门控对分类分支进行自适应重标定,抑制复杂背景引起的误检,提升缺陷类别判别的鲁棒性.改进后的QHN-YOLO模型结构如图1所示.

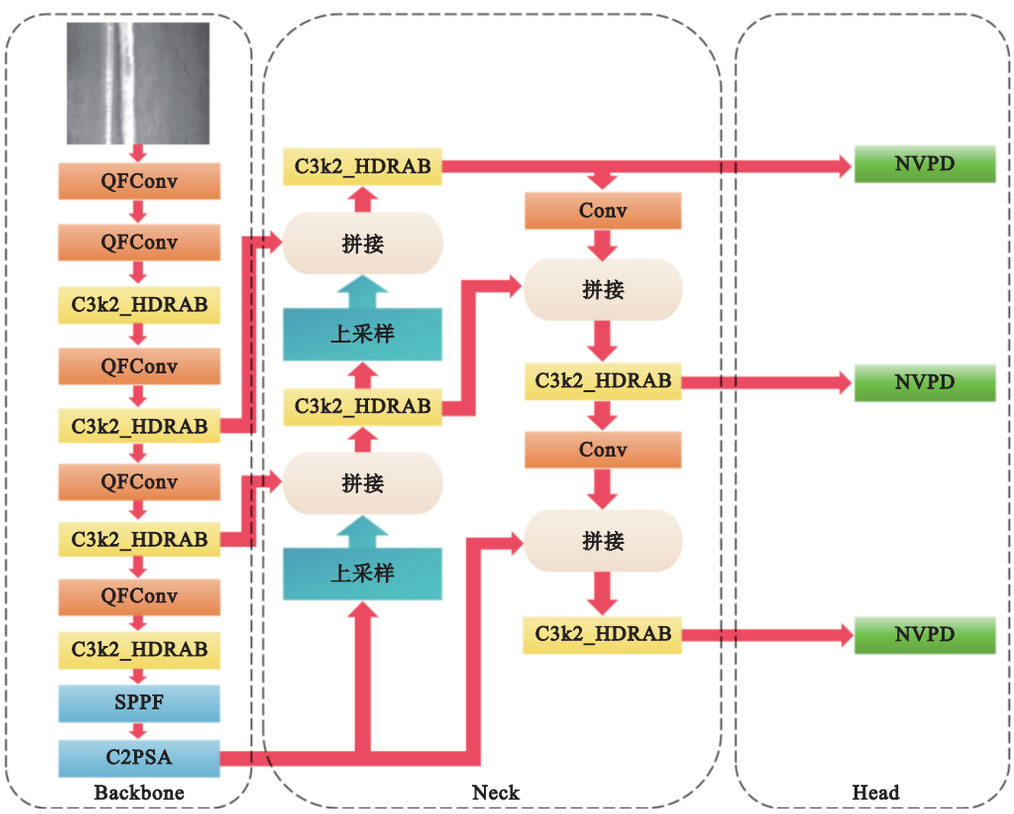


图1 QHN-YOLO模型结构

Fig. 1 Structure of QHN-YOLO model

2.1 量化友好卷积模块

在钢材表面缺陷中,缺陷常呈现弱对比、细小细长与形态不规则等特点,背景纹理与低对比度下易引入强干扰.当模型部署到端侧设备时,通常需要量化以降低延迟与功耗.为满足端侧设备推理的要求,原有卷积模块存在激活值范围过大容易导致截断饱和与量化尺度不稳定的问题,从而引发分类得分波动与检测性能下降.

为提升在端侧设备推理的稳定性,本文以Yu等[19]面向检测模型的全量化部署工作为启发,提出量化友好卷积模块QFConv.Yu等[19]的工作主要面向Transformer结构的特征变换与位置嵌入量化,而本文聚焦钢材缺陷检测中卷积检测器在端侧推理时更突出的算子问题,即复杂纹理背景下激活范围波动与离散化误差叠加导致端侧推理不稳定的问题.具体地,该模块在标准卷积结构中引入有界激活以抑制异常值,并在部署阶段将批归一化参数融合到卷积参数中,以减少算子数量与内存访问开销,同时在训练期采用伪量化模拟量化与反量化过程,使网络适应整数推理带来的离散化误差,从而提高边缘设备部署的鲁棒性与推理效率.QFConv模块的结构如图2所示.

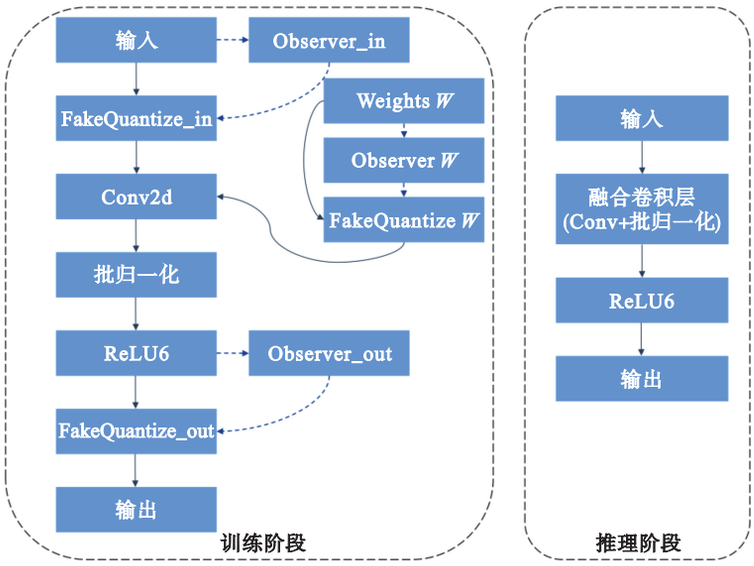


图2　QFConv模块结构

Fig. 2　Structure of QFConv module

QFConv通过卷积运算对输入特征图进行特征提取.卷积层包含一组可学习的卷积核参数,这些卷积核在训练阶段通过反向传播逐步优化,并在推理阶段保持固定不变,从而保证特征提取过程的稳定性与一致性.随后,卷积输出经过批归一化处理,通过对通道统计量进行标准化与可学习的仿射变换,减弱内部特征分布偏移并提升模型的泛化能力.

对于钢材表面缺陷检测,复杂纹理背景与局部高对比噪声易诱发特征峰值偏移,为使模型在细长裂纹与弱对比缺陷检测中保持更稳定的特征响应与端侧部署效率,本文采用有界激活函数ReLU6抑制由背景噪声导致的异常激活峰值.其核心优势在于限制特征的激活动态范围,从而降低截断饱和风险并提升数值稳定性.相比于无上界激活函数,引入的有界激活函数ReLU6将输出限制在有限区间内,使得激活尺度估计更稳定,提高推理的一致性与数值稳定性,满足端侧部署的鲁棒性要求.QFConv可表示为

$$\boldsymbol{Y}_{c_o,h,w}=\sum_{c_i=1}^{C_{in}}\sum_{u=1}^{k}\sum_{v=1}^{k}W_{c_o,c_i,u,v}\,\boldsymbol{X}_{c_i,h+u,w+v}+b_{c_o},\tag{1}$$

$$\boldsymbol{Z}=\gamma\odot\frac{\boldsymbol{Y}-\mu}{\sqrt{\sigma^2+\varepsilon}}+\beta,\tag{2}$$

$$\boldsymbol{O}=\min(\max(\boldsymbol{Z},0),6).\tag{3}$$

式中:$\boldsymbol{X}$表示输入特征图;$\boldsymbol{Y}$表示卷积输出;W为卷积核权重;$\boldsymbol{Z}$为归一化后输出的特征图;b为偏置;C_{in}为输入通道数;h,w为特征图空间尺寸;k为卷积核大小;c_i与c_o分别为输入与输出通道索引;u与v分别表示卷积核在高度方向和宽度方向上的空间索引;μ与σ^2分别为均值与方差;γ与β为可学习缩放与平移参数;ε为数值稳定项;$\odot$表示逐通道乘法;$\boldsymbol{O}$为最终激活输出的特征图.

QFConv的前向映射可用算子复合形式表示,如式(4)所示.

$$\mathrm{QFConv}(\boldsymbol{X})=\mathrm{ReLU6}(\mathrm{BN}(\mathrm{Conv}(\boldsymbol{X}))).\tag{4}$$

式中:Conv,BN和ReLU6分别表示卷积、批归一化与有界激活函数.

在量化感知训练阶段,QFConv通过伪量化在浮点前向中模拟整数推理带来的离散化误差,使网络在训练过程中提前适应量化引入的舍入与截断影响.该过程首先由观察器对张量的取值范围进行统计,并据此估计量化尺度与零点,然后利用伪量化算子对特征进行量化与反量化的数值仿真.在具体实现中,本文采用PyTorch提供的FakeQuantize伪量化算子对权重与激活进行仿真量化,以实现仿射伪量化,从而提高模型在端侧推理条件下的稳定性.其输出可表示为

$$x_{\mathrm{out}}=\big(c(r(x/s+z),q_{\min},q_{\max})-z\big)\cdot s.\tag{5}$$

式中:x与x_{out}分别为输入与输出;s为量化尺度;z为零点;$q_{\min}$与$q_{\max}$为整数表示范围;r表示四舍五入;c表示截断.

在推理融合阶段,QFConv将批归一化的线性变换折叠到卷积参数中,以减少推理时的算子数量并提升端侧计算效率.该过程中,批归一化的缩放与平移参数对卷积权重与偏置进行等效重参数化,使卷积输出在保持数值等价的前提下完成分布校准.融合后的卷积权重与偏置可写为

$$W'_{c_o,c_i,u,v}=\frac{\gamma_{c_o}}{\sqrt{\sigma^2_{c_o}+\varepsilon}}W_{c_o,c_i,u,v}, \tag{6}$$

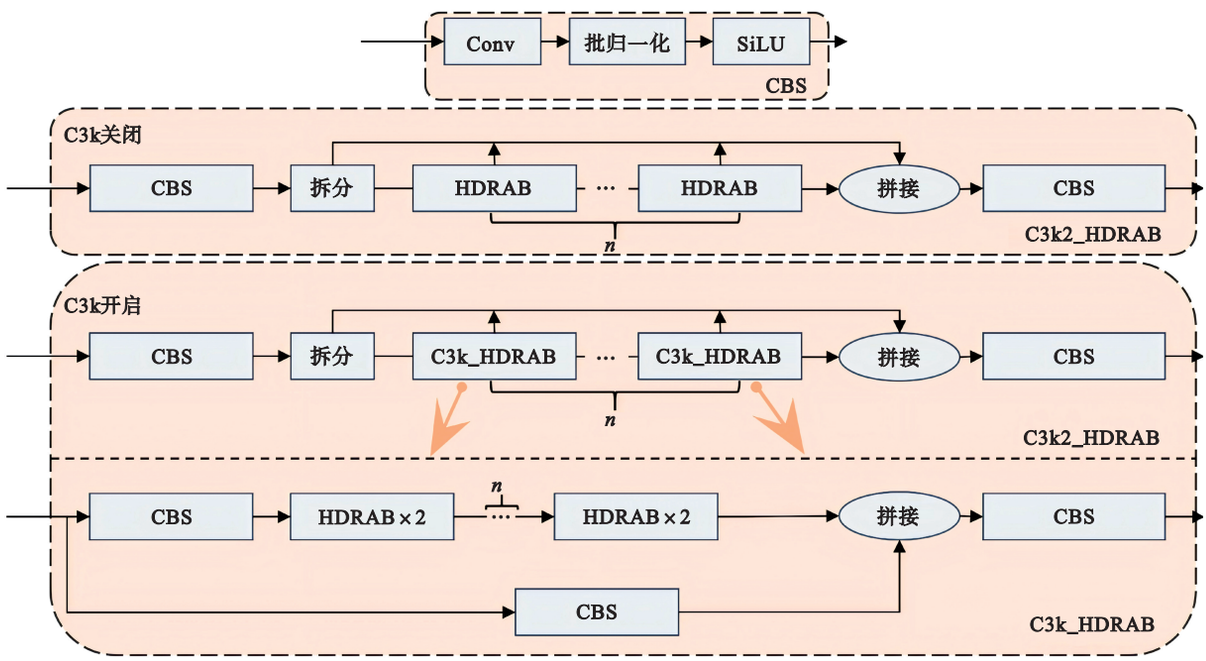
$$b'_{c_o}=\frac{\gamma_{c_o}}{\sqrt{\sigma^2_{c_o}+\varepsilon}}(b_{c_o}-\mu_{c_o})+\beta_{c_o}. \tag{7}$$

式中：W'与b'为融合后的卷积参数；$\mu_{c_o},\sigma^2_{c_o},\gamma_{c_o},\beta_{c_o}$为第$c_o$个输出通道对应的批归一化统计量与可学习参数.融合后，推理仅需执行1次卷积与1次激活，有助于减少算子数量并提升端侧吞吐.

QFConv将卷积、归一化与有界激活组织为可融合结构，并结合量化感知训练在训练阶段显式建模量化误差，使模型参数适应端侧推理.导出阶段进一步完成卷积与批归一化的融合折叠，使推理结构与端侧部署形式一致，从而形成结构、训练与导出协同的闭环改进.通过该融合方式，模型在推理阶段仅需执行1次卷积与1次激活即可完成等效计算，从而在不牺牲特征表达能力的前提下，进一步降低部署开销并提升推理效率和检测精度.

2.2 混合膨胀残差注意力块

在钢材表面缺陷中，裂纹与划痕等缺陷呈现细长连续形态，夹杂、麻点与氧化皮等缺陷又常伴随复杂纹理背景与尺度跨度.在工业场景中，特征需要同时具备更大的有效感受野以覆盖长距离结构，以及对局部细节的敏感性以捕捉边缘与微小异常.为此，本文在原有C3k2模块的基础上引入混合膨胀残差注意力块HDRAB[20]，形成C3k2_HDRAB模块.与通用注意力模块不同，HDRAB以多尺度空洞上下文为主、注意力为辅，通过多膨胀率卷积的级联与残差聚合扩大上下文感知范围，重点增强缺陷的跨尺度连贯表征，并在轻量约束下兼顾端侧计算复杂度.其结构如图3所示.



注：CBS是由卷积、批归一化和SiLU激活函数构成的基础卷积模块.

图3 C3k2_HDRAB模块结构

Fig. 3 Structure of C3k2_HDRAB module

C3k2_HDRAB模块沿用了C3k2的双分支特征流架构.输入特征首先经过CBS完成通道对齐与基础特征提取，随后在通道维度上进行解耦分流.其中，旁路分支保留浅层的纹理与边缘细节，避免特征在深层传递中发生退化；主干分支则由若干HDRAB增强单元堆叠而成，用于实现跨尺度的上下文信息建模.该设计在保持外部接口统一的前提下，有效增加了模型的非线性深度与多尺度聚合次数.最终，双分支的输出在通道维度进行拼接，并完成特征融合与降维输出.其具体的计算流程如下：

$$\boldsymbol{X}_0=\mathrm{CBS}(X), \tag{8}$$

$$\left[\boldsymbol{X}_\mathrm{a},\boldsymbol{X}_\mathrm{b}\right]=\mathrm{Split}(\boldsymbol{X}_0), \tag{9}$$

$$\boldsymbol{X}'_\mathrm{b}=M(\boldsymbol{X}_\mathrm{b}), \tag{10}$$

$$\boldsymbol{Y}=\mathrm{CBS}\left(\mathrm{Concat}(\boldsymbol{X}_\mathrm{a},\boldsymbol{X}'_\mathrm{b})\right). \tag{11}$$

式中：$\boldsymbol{X}_0$为通道对齐后的中间特征；$\boldsymbol{X}_\mathrm{a}$为旁路分支特征；$\boldsymbol{X}_\mathrm{b}$为主干分支特征，根据C3k的布尔值而定；Concat表示拼接.

HDRAB的核心在于混合膨胀卷积与残差聚合.膨胀卷积通过在卷积采样点之间引入间隔,在不增加参数量的情况下扩大感受野,适于捕获钢材表面长裂纹的连续结构与大尺度缺陷的上下文信息.与此同时,单一膨胀率可能产生栅格效应并削弱细节表征,HDRAB采用多膨胀率的对称级联,并在多个阶段引入残差相加以保持局部纹理信息与梯度传播稳定,从而在大尺度建模与细节保真之间取得平衡.在混合膨胀残差路径中,HDRAB按1,2,3,4,3,2,1的膨胀率顺序依次进行卷积变换,并在若干关键节点执行逐元素相加以聚合不同感受野的表征.使用该集合的目的是在保持计算开销可控的前提下实现稳定的多尺度上下文聚合.膨胀卷积能够在不降低特征分辨率的情况下扩大有效感受野,从而增强对缺陷形态与背景纹理差异的建模能力.该序列在结构上呈现先扩张后收敛的对称形态,前半段通过逐级增大的膨胀率增强长距离依赖与全局轮廓感知,后半段通过逐级减小的膨胀率回归局部细节以强化边界对齐与细粒度定位,使特征表达同时具备全局判别性与局部可分辨性,从而适应钢材缺陷中细长划痕、裂纹缺陷的长程连通性和边界敏感性.此外,为避免膨胀率过大造成采样稀疏与栅格效应,本文将最大膨胀率限制为4,并通过多层级对称融合抑制密集预测中的伪影传播,从而在提升感受野的同时维持稳定的空间一致性.

为强化缺陷相关通道响应并抑制背景干扰,HDRAB在卷积聚合后引入通道注意力调制.激活函数采用与QFConv一致的有界激活形式,以保持推理时的数值稳定性.HDRAB模块结构如图4所示,其计算过程如下:

$$\boldsymbol{F}_1=\phi\left(\mathrm{DConv}_1(\boldsymbol{F}_0)\right), \tag{12}$$

$$\boldsymbol{F}_2=\mathrm{DConv}_2(\boldsymbol{F}_1)\oplus\boldsymbol{F}_0, \tag{13}$$

$$\boldsymbol{F}_3=\phi\left(\mathrm{DConv}_3(\boldsymbol{F}_2)\right)\oplus\boldsymbol{F}_2, \tag{14}$$

$$\boldsymbol{F}_4=\mathrm{DConv}_4(\boldsymbol{F}_3)\oplus\boldsymbol{F}_3, \tag{15}$$

$$\boldsymbol{F}_5=\phi\left(\mathrm{DConv}_3(\boldsymbol{F}_4)\right)\oplus\boldsymbol{F}_4, \tag{16}$$

$$\boldsymbol{F}_6=\mathrm{DConv}_2(\boldsymbol{F}_5)\oplus\boldsymbol{F}_5, \tag{17}$$

$$\boldsymbol{F}_7=\phi\left(\mathrm{DConv}_1(\boldsymbol{F}_6)\right), \tag{18}$$

$$a=\sigma_{\mathrm{S}}\left(\mathrm{Conv}_2\left(\phi\left(\mathrm{Conv}_1\left(\mathrm{GAP}(\boldsymbol{F}_7)\right)\right)\right)\right), \tag{19}$$

$$\boldsymbol{Y}'=(\boldsymbol{F}_7\odot a)\oplus\boldsymbol{F}_0. \tag{20}$$

式中:$\boldsymbol{F}_0$表示模块输入特征图,并作为残差分支与主分支特征进行逐元素相加,以保留原始细节信息并提升训练稳定性;$\boldsymbol{F}_t$(t=1,2,…,7)为第t阶段的中间特征;DConv_d为膨胀率d的卷积变换;ϕ为ReLU6激活函数;⊕为逐元素相加,用于实现残差聚合;a为通道注意力权重;σ_{S}为Sigmoid函数;Conv_1与Conv_2为注意力分支中的1×1卷积变换;GAP为全局平均池化;$\boldsymbol{Y}'$为输出特征图.

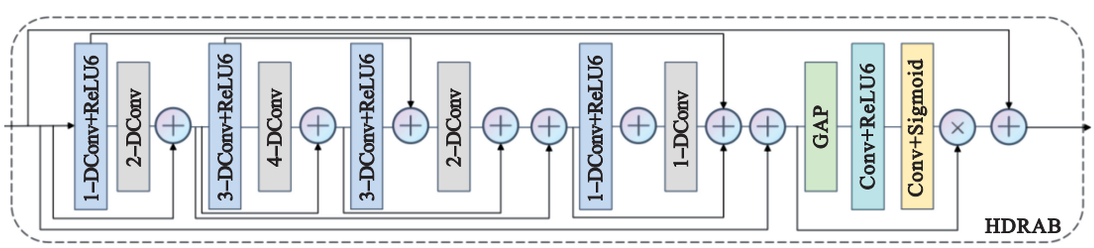


图4　HDRAB模块结构

Fig. 4　Structure of HDRAB module

综上,HDRAB通过多膨胀率卷积与残差聚合扩大了有效感受野,增强了对长裂纹与大尺度缺陷的跨区域关联建模能力,并减弱了复杂纹理背景下的断裂式响应.同时,通道注意力进一步强化了与缺陷边缘相关的有效响应,抑制了背景噪声引起的伪激活.上述设计使模型在多尺度缺陷定位、边缘保持与复杂背景抑制方面获得更稳定的特征表达,从而为后续检测头提供更具判别性的输入特征.

2.3　基于噪声与方差先验门控机制检测头

为提升钢材表面缺陷检测任务中对复杂背景纹理与噪声干扰的鲁棒性,本研究在YOLOv11原有的解耦检测头基础上引入噪声与方差先验门控机制,构建噪声方差先验检测头NVPD.其核心思想是在不改变边界框回归分支解码方式的前提下,通过一条轻量级的空间门控分支为分类分支提供可学习的先验权重图,从而对疑似背景纹理区域的分类响应进行抑制,并在细小裂纹、

划痕等高频细节区域保留较高的判别响应.该机制在缺陷表面对比度不均、轧制纹理密集及缺陷尺度跨度较大的场景中可有效抑制误检,并保持对细小缺陷的分类响应稳定性.其结构如图5所示.

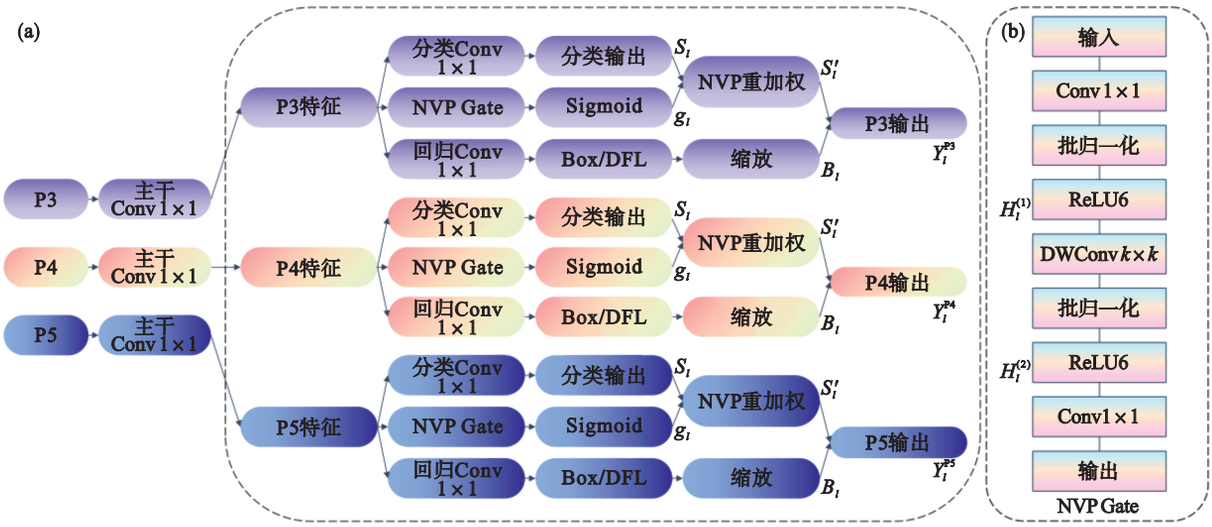


图5 NVPD结构

Fig. 5 Structure of NVPD

(a)—NVPD网络结构;(b)—NVP Gate模块结构.

NVPD以多尺度特征层P3,P4,P5为输入,每一尺度首先经过一层1×1卷积完成通道对齐与特征压缩,得到用于该尺度检测的特征图.随后,特征图被送入回归分支与分类分支形成解耦结构.回归分支通过1×1卷积产生边界框分布或偏移量表征,并在推理阶段沿用YOLOv11的解码流程以及尺度因子完成对不同尺度回归幅值的自适应校准.分类分支则通过1×1卷积生成各类别的分类对数概率,作为候选框的原始类别响应.

为抑制背景纹理引发的误检,本文在NVPD的每一尺度额外引入一条噪声与方差先验门控分支,名为NVP Gate分支.该分支以同尺度特征图为输入,采用串联的卷积结构提取局部噪声与纹理变化,并输出单通道的空间先验图.门控分支末端使用Sigmoid函数将先验图归一化,使其可作为稳定的乘性调制系数,通过Sigmoid门控权重对特征或响应进行增强与抑制,从而提升特征判别与抗噪能力.随后,分类分支输出在空间位置上按照该先验图进行重标定,形成校准后的分类响应.由于门控分支仅生成单通道空间权重且主要由深度可分离卷积构成,其额外参数与计算开销较小.在不显著增加参数量的条件下,该分支能有效刻画纹理起伏与噪声扰动特征,使门控权重能够敏感响应背景噪声与缺陷结构差异,从而提升模型在复杂工业场景下的稳定检出能力并抑制误检.

在前向传播中,NVPD对每一尺度分别计算回归输出、分类输出与门控权重,并将校准后的分类输出与回归输出在通道维拼接作为该尺度检测输出.训练阶段,拼接后的张量直接参与分类损失与回归损失的计算,门控分支的参数通过检测损失的反向传播自动学习到与缺陷先验相关的空间权重分布.而在推理阶段,门控重标定仅作用于分类分支的对数概率层面,不改变回归解码和后处理流程,从而在不增加计算与部署复杂度的情况下提升检测精度并抑制背景噪声.对应的计算关系表示为

$$\boldsymbol{S}_l=\mathrm{Conv}_{\mathrm{cls}}(\boldsymbol{F}_l), \tag{21}$$

$$\boldsymbol{B}_l=\mathrm{Conv}_{\mathrm{reg}}(\boldsymbol{F}_l), \tag{22}$$

$$\boldsymbol{g}_l=\sigma_{\mathrm{S}}\big(\mathrm{Gate}(\boldsymbol{F}_l)\big), \tag{23}$$

$$\boldsymbol{S}'_l=S_l\odot\big(1+\alpha\boldsymbol{g}_l\big), \tag{24}$$

$$\boldsymbol{Y}_l=\mathrm{Concat}\big(B_l,S'_l\big). \tag{25}$$

式中:$\boldsymbol{F}_l$为尺度l的输入特征图;$\mathrm{Conv}_{\mathrm{cls}}$与$\mathrm{Conv}_{\mathrm{reg}}$分别表示该尺度的分类与回归1×1卷积映射;$\boldsymbol{S}_l$为原始分类对数概率;$\boldsymbol{B}_l$为回归分支输出,包含Box与DFL所需的回归表征;Gate为门控映射;$\boldsymbol{g}_l$为单通道空间先验图;α为重标定强度系数,综合数据集与实验情况后取α=1.0,并在全部实验中保持不变;$\boldsymbol{S}'_l$为校准后的分类对数概率;$\boldsymbol{Y}_l$为该尺度的检测输出张量.

NVP Gate模块从同尺度特征图中提取与噪声强度、纹理起伏和局部方差相关的线索，并输出空间位置相关的先验权重．为兼顾特征表达能力与轻量化约束，门控分支采用1×1卷积进行通道混合与压缩，再以深度可分离卷积建模局部邻域变化，最后通过1×1卷积映射到单通道先验图并用Sigmoid函数归一化．对应的计算可表示为

$$\boldsymbol{H}_l^{(1)}=\phi\Big(\mathrm{BN}\Big(\mathrm{Conv}_{1\times1}\big(\boldsymbol{F}_l,W_1\big)\Big)\Big),\tag{26}$$

$$\boldsymbol{H}_l^{(2)}=\phi\Big(\mathrm{BN}\Big(\mathrm{DWConv}_{5\times5}\big(\boldsymbol{H}_l^{(1)},W_{\mathrm{d}}\big)\Big)\Big),\tag{27}$$

$$\boldsymbol{g}_l=\sigma_{\mathrm{S}}\big(\mathrm{Conv}_{1\times1}\big(\boldsymbol{H}_l^{(2)},W_2\big)+b_2\big).\tag{28}$$

式中：$\boldsymbol{H}_l^{(1)}$与$\boldsymbol{H}_l^{(2)}$为门控分支的中间特征；$\mathrm{DWConv}_{5\times5}$为核大小为5×5的深度卷积；$W_1$，$W_{\mathrm{d}}$，$W_2$分别为门控分支对应的卷积权重；$b_2$为1×1卷积的偏置；$k$为深度卷积核大小．

综上，NVPD通过引入轻量门控分支对分类分支进行空间自适应重标定，使背景纹理区域的虚高响应得到抑制，而与缺陷结构相关的位置获得更稳定的类别激活．该机制本质上是一种面向背景噪声抑制的分类校准策略，能够在不改变回归解码流程的前提下提高检测在复杂背景下的鲁棒性，减少由噪声与纹理扰动引起的误检．

3　实验结果及分析

3.1　数据集

为了验证所提出结构的可靠性，本文采用NEU-DET数据集[21]进行实验．该数据集收集了热轧带钢常见的6种缺陷，包括裂纹（crazing）、夹杂物（inclusion）、斑点（patches）、麻点（pitted surface）、压入氧化皮（rolled-in scale）和划痕（scratch），如图6所示，每种缺陷包含300张图片，共计1 800张图片．按照8∶1∶1的比例将数据集划分为训练集、验证集和测试集．

3.2　实验设置

实验平台基于Windows 10操作系统，搭载Intel Core i5-13400F CPU（32 GB RAM）及NVIDIA GeForce RTX 4060 Ti（16 GB）显卡；深度学习框架采用PyTorch 2.3.0（Python 3.10，CUDA 12.1）．模型训练参数设置如下：输入图像分辨率为640像素×640像素，批量大小为32，迭代轮数为300；采用SGD优化器，初始学习率设为0.01，未采用学习率衰减与早停策略．

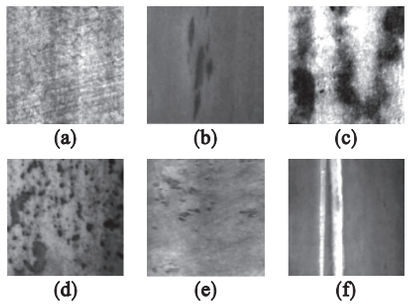


图6　NEU-DET 钢材表面缺陷

Fig. 6　NEU-DET steel surface defects

（a）—裂纹；（b）—夹杂物；（c）—斑点；（d）—麻点；（e）—压入氧化皮；（f）—划痕．

本文在训练过程中采用量化感知训练，以模拟端侧整型推理的离散化误差．训练阶段对权重与激活均按8 b进行伪量化，权重采用逐通道对称量化，激活采用逐张量仿射量化，并在前向传播中插入伪量化算子，以显式建模量化与反量化引入的舍入与截断误差．本文自训练开始即启用量化感知训练，并在模型导出阶段执行卷积与批归一化层的融合折叠，使推理结构与端侧部署形式一致，从而提升部署一致性与结果可复现性．

3.3　评价指标

为了验证本文方法的有效性，本次实验采用的评估指标包括精确率（P）、召回率（R）、平均精确率（AP）、平均精确率均值（mAP）、参数量、计算量和检测速度（v）．其中计算量用浮点运算次数（floating point operations，FLOPs）表示，GFLOPs指模型需要进行10^9次浮点运算；检测速度用每秒帧率（fps）表示，即模型每秒钟处理的图像帧数．各指标计算公式为

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\tag{29}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\tag{30}$$

$$\mathrm{AP}=\sum_{i=1}^{K}\int_0^1 P(R)\,\mathrm{d}R,\tag{31}$$

$$\mathrm{mAP}=\frac{\sum_{i=1}^{N}\int_0^1 P(R)\,\mathrm{d}R}{N},\tag{32}$$

$$O=HW\left(C_{\mathrm{in}}K^2\right)C_{\mathrm{out}},\tag{33}$$

$$v=\frac{1\,000}{t_{\mathrm{pre}}+t_{\mathrm{inf}}+t_{\mathrm{post}}}.\tag{34}$$

式中：TP表示实际为正样本且被模型正确预测为正样本的数量；FP表示实际为负样本但被模型错误预测为正样本的数量；FN表示实际为正样本但被模型错误预测为负样本的数量；O为计算量，单位为GFLOPs；N为缺陷类别数；H，W和C_{in}分别表

示输入特征图的高度、宽度及通道数；K为核宽度；C_{out}为输出通道数；t_{pre}，t_{inf}和t_{post}分别表示图像预处理时间、模型推理时间和结果后处理时间.

3.4 对比实验及分析

表1给出了QHN-YOLO与YOLO系列模型在NEU-DET数据集上的对比结果.总体来看，QHN-YOLOv11在保持轻量化规模的前提下取得了最高的mAP，达到了80.2%，说明所提出的3项改进在钢材表面缺陷检测任务中具有较好的协同提升效果.与基线YOLOv11n相比，QHN-YOLOv11的精确率、召回率和mAP分别提升至77.7%，74.0%和80.2%，同时参数量仅由2.6×10^6增加至2.8×10^6，计算量由6.3 GFLOPs增加至7.1 GFLOPs，表明模型在较小复杂度增量下获得了明显的检测性能提升.

与其他轻量模型相比，QHN-YOLOv11在参数量和计算量与YOLOv8n，YOLOv12n等接近的条件下取得了更高的检测精度，体现了更优的精度与复杂度权衡.与FSU-YOLO和HDFA-YOLO等面向缺陷检测优化的方法相比，QHN-YOLOv11在获得更高mAP的同时保持了更低的模型复杂度，说明其更适合端侧部署场景.各类别指标表明，QHN-YOLO在裂纹、麻点和压入氧化皮等复杂缺陷上检测精度的提升更为明显，说明所提方法对细长结构、弱对比边缘及复杂纹理背景具有更稳定的判别能力.综合来看，QHN-YOLO在精度、复杂度与端侧部署适应性之间实现了较为均衡的优化.

表1 YOLO系列模型对比结果

Table 1 Comparative results of YOLO series models

模型	P/%	R/%	mAP/%	检测精度/%						参数量$\times10^{-6}$	计算量
				裂纹	夹杂物	斑点	麻点	压入氧化皮	划痕		GFLOPs
YOLOv3[22]	69.3	72.3	76.0	44.5	86.9	88.7	84.4	58.7	93.2	103.7	282.5
YOLOv3-tiny[23]	60.5	66.2	66.1	30.6	84.2	87.3	64.7	55.6	73.8	12.1	19.0
YOLOv5n[24]	73.4	70.4	75.7	39.7	86.0	86.8	90.8	59.9	91.2	2.2	5.8
YOLOv6n[25]	66.8	70.2	74.2	45.3	81.4	91.4	83.9	53.1	90.3	4.2	11.9
YOLOv8n[26]	71.3	73.1	76.9	43.4	86.9	84.6	92.1	62.4	92.1	2.7	6.8
YOLOv9tiny[27]	68.9	70.1	75.8	38.4	88.5	91.8	87.3	58.0	90.7	1.7	6.4
YOLOv10n[28]	71.3	67.0	73.8	34.8	84.4	90.3	85.2	58.2	90.0	1.9	5.3
YOLOv11n	67.5	70.1	75.5	40.0	84.3	93.2	86.1	59.6	89.7	2.6	6.3
YOLOv12n[29]	67.6	71.8	74.9	39.1	87.5	87.8	81.6	59.7	93.5	2.5	5.8
FSU-YOLO[11]	—	—	79.5	47.2	86.9	94.2	92.9	63.3	92.3	8.6	22.0
HDFA-YOLO[12]	—	—	78.3	50.2	87.0	88.6	87.1	63.7	93.3	—	23.7
QHN-YOLOv5	69.1	77.7	77.8	48.1	83.6	91.8	90.0	61.7	91.5	2.9	7.5
QHN-YOLOv12	67.6	74.6	76.4	34.1	84.7	93.7	88.6	62.6	94.8	2.7	6.8
QHN-YOLOv11	77.7	74.0	80.2	54.0	86.5	93.3	92.1	63.4	92.0	2.8	7.1

3.5 消融实验及分析

为评估QHN-YOLO中QFConv，C3k2_HDRAB和NVPD改进模块对钢材表面缺陷检测性能的贡献，本文共设置8组消融实验，结构配置如表2所示，消融实验结果如表3所示，其中“√”表示该实验启用对应模块.

在单模块实验中，No.2引入QFConv后，精确率由67.5%提升至75.3%，召回率降至68.5%，mAP提升至77.1%，原因在于QFConv引入了有界激活约束，使特征的动态范围更稳定，减弱了噪声与强背景纹理导致的异常激活，从而提升了对细粒度纹理缺陷的响应一致性.No.3引入C3k2_HDRAB后，mAP提升至77.8%，且裂纹和划痕类别检测精度提升更明显，说明引入的C3k2_HDRAB通过多分支残差叠加与全局调制项抑制无效背景并强化缺陷边缘，对裂纹类与边界类缺陷具有更强的结构增强作用.No.4单独引入NVPD后，mAP提升至76.9%，在复杂纹理干扰较强的类别上改善效果更明显，说明其对背景诱发误检具有针对性抑制作用.

在双模块组合中，各模块的提升并非简单线性叠加.No.5组合QFConv与C3k2_HDRAB后，mAP为77.9%，整体提升有限，主要在于引入的C3k2_HDRAB倾向于放大局部对比度与边缘响应，使特征呈现更尖锐的峰值分布，而QFConv的有界激活函数会对峰值进行截断并重新塑形，表

明激活约束与强边缘增强在响应分布上可能存在一定的重整关系,因此双模块组合的提升未明显扩大.No.6 组合 QFConv 与 NVPD 后,mAP 为 77.3%.其中,压入氧化皮的检测精度达到 66.8%,为所有设置中的最高值,这主要是因为 QFConv 提供更稳定的激活分布,使 NVPD 的门控图更易收敛到合理范围,总体上对纹理干扰较强类别的检测精度有提升,但由于缺少结构增强,整体提升仍受限制.No.7 组合 C3k2_HDRAB 与 NVPD 后,mAP 达到 78.4%,召回率提升至 76.0%,表明该组合更有利于提升缺陷覆盖与召回.但压入氧化皮的检测精度却下降到 58.1%,说明 NVPD 的门控效果与前端特征分布密切相关,仅依赖前端增强与分类校准时,门控效果仍会受到输入特征稳定性的影响.

当 3 个模块同时引入后,QHN-YOLO 的精确率提升至 77.7%,召回率提升至 74.0%,mAP 提升到 80.2%,相较 No.1 分别提升 10.2%,3.9% 和 4.7%,并在 6 类缺陷上呈现更均衡的提升趋势.这说明 QFConv,C3k2_HDRAB 与 NVPD 在特征稳定、结构增强和分类校准 3 个层面形成了互补作用,使模型在误检抑制、缺陷覆盖与整体检测精度之间取得了更均衡的性能表现.与此同时,QHN-YOLO 的参数量与计算量仅分别增加到 2.8×10^6 与 7.1 GFLOPs,参数的增量主要来自 HDRAB 引入的结构分支,整体保持轻量水平,仍满足端侧部署场景对轻量化的要求.

表 2　消融实验模型结构配置

Table 2　Structural configuration of models for ablation experiments

消融实验	QFConv	C3k2_HDRAB	NVPD
No. 1	—	—	—
No. 2	√	—	—
No. 3	—	√	—
No. 4	—	—	√
No. 5	√	√	—
No. 6	√	—	√
No. 7	—	√	√
QHN-YOLO	√	√	√

表 3　消融实验结果

Table 3　Results of ablation experiments

消融实验	*P*/%	*R*/%	mAP/%	检测精度/%						参数量×10^{-6}	计算量
				裂纹	夹杂物	斑点	麻点	压入氧化皮	划痕		GFLOPs
No. 1	67.5	70.1	75.5	40.0	84.3	93.2	86.1	59.6	89.7	2.6	6.3
No. 2	75.3	68.5	77.1	41.6	82.4	91.4	90.9	62.3	93.8	2.6	6.3
No. 3	74.2	71.0	77.8	46.9	81.3	94.4	88.9	60.9	94.2	2.7	7.0
No. 4	72.9	69.2	76.9	44.6	85.6	88.7	85.4	64.4	92.6	2.6	6.4
No. 5	72.1	71.3	77.9	50.9	83.8	91.5	87.3	63.4	90.5	2.7	7.0
No. 6	72.1	73.2	77.3	41.0	85.0	91.6	87.2	66.8	92.3	2.6	6.4
No. 7	68.5	76.0	78.4	51.1	86.8	91.5	91.7	58.1	91.0	2.8	7.1
QHN-YOLO	77.7	74.0	80.2	54.0	86.5	93.3	92.1	63.4	92.0	2.8	7.1

3.6　模型可视化分析

为比较 QHN-YOLO 与 YOLOv11n 的定性检测差异,从 NEU-DET 数据集中选取各类别样本进行可视化对比.推理设置中,输入图像分辨率尺寸为 640 像素×640 像素、置信度阈值为 0.25,交并比阈值为 0.70.图 7 给出了两种模型的检测结果.相比基线模型,QHN-YOLO 在微小缺陷、复杂纹理背景与低对比度缺陷上的漏检与误检均有所减少,尤其在裂纹类缺陷上能够提供更完整的目标覆盖与更稳定的分类结果,表明所提的改进有助于提高复杂背景下的检测可靠性.

为了进一步验证 QHN-YOLO 的检测性能,本文采用梯度加权类激活映射[30](Grad-CAM)来比较两个模型的性能.在特征图的基础上,热力图可以直观地展现模型的关注区域.本文选取了所有 6 类缺陷进行热力图分析,对比结果如图 8 所示.图中绿色框标注真实缺陷,红色框标注当前模型在该图上的检测结果.

具体来看,对于细长条带状或边缘较弱的缺陷样本,YOLOv11n 的高激活区域呈离散分布,并在缺陷周围纹理区域出现外溢,难以沿缺陷走向形成连续响应;QHN-YOLO 则在缺陷延展方向上形成连续的激活响应,且背景激活相对降低,有助于保持缺陷形态的一致性.对于面积较大或对比度较高的缺陷样本,QHN-YOLO 的激活主要集中于缺陷内部及关键边界处,背景区域响应

较少;而YOLOv11n在部分样本中出现较大范围背景激活,降低了缺陷与背景的分离度.上述可视化结果表明,QHN-YOLO在缺陷区域聚焦、背景抑制与空间一致性方面优于基线模型,从而为复杂背景与微小缺陷场景提供更稳定的判别依据.

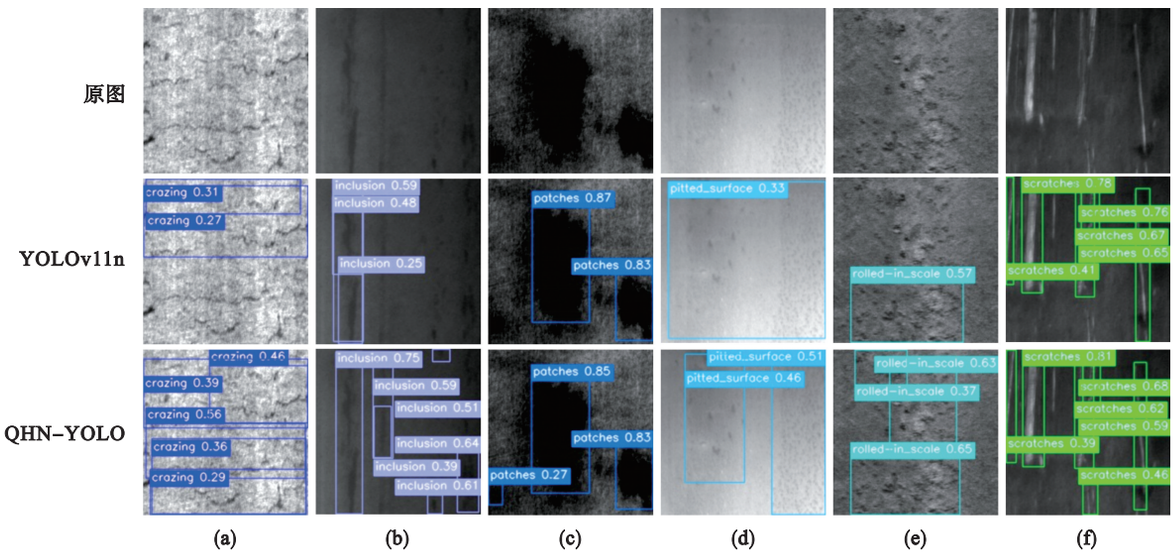


图7 检测结果可视化

Fig. 7 Visualization of detection results

(a)—裂纹;(b)—夹杂物;(c)—斑点;(d)—麻点;(e)—压入氧化皮;(f)—划痕.

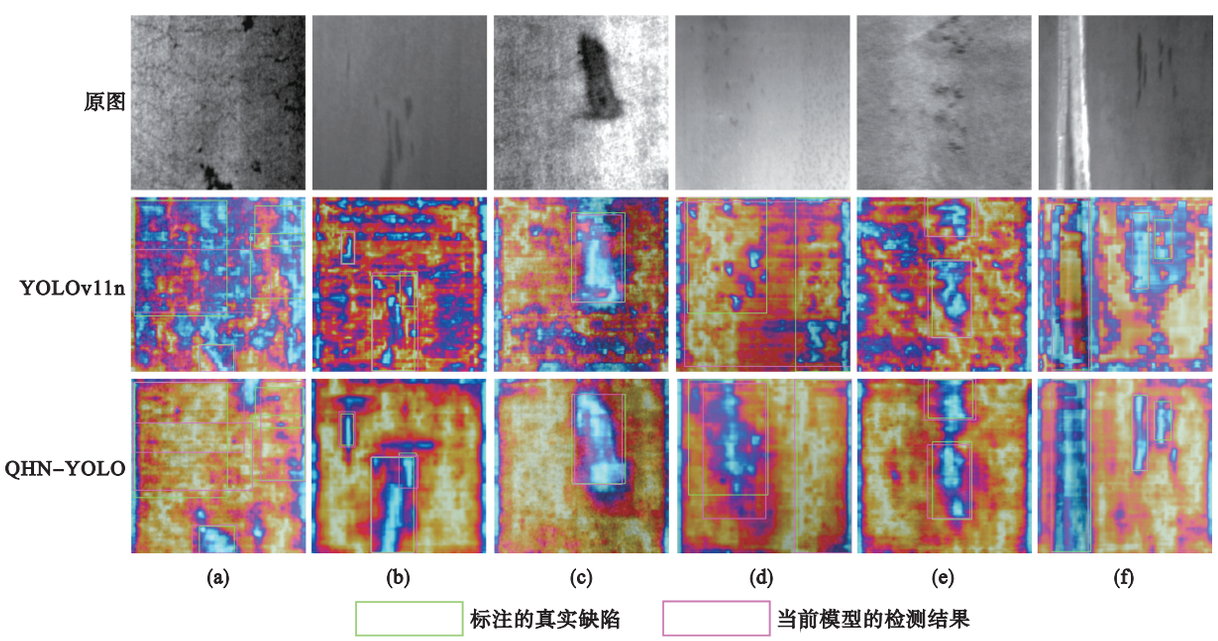


图8 Grad-CAM检测结果对比

Fig. 8 Comparison of Grad-CAM detection results

(a)—裂纹;(b)—夹杂物;(c)—斑点;(d)—麻点;(e)—压入氧化皮;(f)—划痕.

3.7 端侧部署分析

钢材表面缺陷检测通常在工业边缘端进行部署,如现场的便携式工业设备,以满足实时处理的要求.为评估QHN-YOLO在资源受限平台上的实际性能,本文选用搭载AMD Ryzen7 7840U的小型端侧设备作为测试平台.该平台为8核16线程的Zen 4架构低功耗移动处理器,集成Radeon 780M核显并与系统内存共享显存.在Windows 11和ONNX Runtime的推理栈下,可利用核显进行加速.在保持模型权重与推理配置一致的前提下,硬件平台差异主要影响推理时延与吞吐性能,因此重点分析端侧部署时的检测速度与时延表现.结果如表4所示.

实验结果表明,QHN-YOLO在该平台上的检测效率优于原模型,充分展示了其设计的可行性.尽管YOLO系列模型在提升推理速度方面已

有所优化,但仍受限于其内部较高的算子复杂度和资源消耗,难以满足工业场景的严格要求.相比于仅替换QFConv模块的模型,QHN−YOLO的检测效率略有降低,但mAP仍维持较高水平.综合来看,QHN−YOLO通过针对性优化,提高了推理阶段的检测效率,检测速度提升至34.6 帧/s,时延降低至28.9 ms.尽管仅替换QFConv的模型在检测速度上略优,但QHN−YOLO模型在精度与实时性之间实现了折中,验证了其端侧部署的可行性.

表4　端到端实验性能对比

Table 4　Performance comparison in end-to-end experiments

模型	输入尺寸/像素	输入批次大小	mAP/%	参数量×10^{-6}	时延/ms	检测速度/(帧·s^{-1})
YOLOv11n	640×640	1	75.5	2.6	36.0	27.8
YOLOv11n−QFConv	640×640	1	77.1	2.6	26.4	37.9
QHN−YOLO	640×640	1	80.2	2.8	28.9	34.6

4　结　　论

1) 本文在YOLOv11的基础上进行改进,通过设计量化友好卷积模块QFConv,在不增加模型复杂度的前提下稳定特征动态范围,减弱强背景纹理与噪声导致的异常激活,使端侧推理过程中的鲁棒性与检测速度得到提升.

2) 采用混合膨胀残差注意力块HDRAB优化原有网络中的C3k2模块,通过多膨胀率上下文与残差注意力协同扩大有效感受野并强化缺陷边缘与结构线索,使得模型对裂纹、划痕等多尺度缺陷的表征能力更强,检测精度更高.

3) 设计噪声方差先验检测头NVPD,利用先验门控对分类分支进行自适应重标定,抑制复杂背景纹理引起的虚高置信度与误检,提高跨尺度缺陷检测场景下的检测可靠性与误检抑制能力.

4) 在NEU−DET数据集上的实验结果表明,融合3项改进后的QHN−YOLO的mAP达到80.2%,较基线YOLOv11n提升4.7%,仅需参数量为2.8×10^6与计算量为7.1 GFLOPs.在端侧设备的验证上,其时延降低至28.9 ms且检测速度达到34.6 帧/s,说明该方法在保持较高检测精度的同时提升了端侧实时检测能力.

参考文献:

[1] Schlegel J. The world of steel [M]. Wiesbaden: Springer, 2023: 45−49.

[2] Palit P, Sri K P, Kayal N, et al. Failure investigation and crack characterization of spalled work roll in hot strip mill [J]. *Journal of Failure Analysis and Prevention*, 2024, 24(4): 1522−1532.

[3] Dwivedi S K, Vishwakarma M, Soni A. Advances and researches on non destructive testing: a review [J]. *Materials Today: Proceedings*, 2018, 5(2): 3690−3698.

[4] Saberironaghi A, Ren J, El-Gindy M. Defect detection methods for industrial products using deep learning techniques: a review [J]. *Algorithms*, 2023, 16(2): 95.

[5] Bai J, Wu D, Shelley T, et al. A comprehensive survey on machine learning driven material defect detection [J]. *ACM Computing Surveys*, 2025, 57(11): 1−36.

[6] Alhashmi S A, Al−azawi A. A review of the single-stage vs. two-stage detectors algorithm: comprehensive insights into object detection [J]. *International Journal of Environmental Sciences*, 2025, 11(3): 775−787.

[7] Bharati P, Pramanik A. Deep learning techniques—R−CNN to mask R−CNN: a survey [C]// Computational Intelligence in Pattern Recognition (CIPR). Singapore: Springer, 2020: 657−668.

[8] Ren S Q, He K M, Girshick R, et al. Faster R−CNN: towards real-time object detection with region proposal networks [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137−1149.

[9] He K M, Gkioxari G, Dollár P, et al. Mask R−CNN [C]// 2017 IEEE International Conference on Computer Vision (ICCV). Venice, 2017: 2980−2988.

[10] Kang S Z, Hu Z Y, Liu L J, et al. Object detection YOLO algorithms and their industrial applications: overview and comparative analysis [J]. *Electronics*, 2025, 14(6): 1104.

[11] An Y S, Chen X, He X, et al. FSU−YOLO: aero-engine blade defect detection based on improved YOLOv8 [J]. *Measurement Science and Technology*, 2025, 36(10): 105417.

[12] Li J, He X, Chen X, et al. HDFA−YOLO: a real-time steel surface defect detection model based on backbone lightweight design and multi-scale feature fusion [J]. *Measurement*, 2026, 258: 119390.

[13] 郭业才, 唐宇捷. 基于MHD−YOLO的钢表面缺陷检测算法[J]. 激光与光电子学进展, 2025, 62(20): 132−141.
(Guo Ye-cai, Tang Yu-jie. Steel surface defect detection algorithm based on MHD−YOLO [J]. *Laser & Optoelectronics Progress*, 2025, 62(20): 132−141.)

[14] 梁礼明, 陈康泉, 钟奕, 等. DCD−YOLOv8n:一种高效的钢材表面缺陷检测算法[J]. 计算机工程与应用, 2025, 61(7): 117−127.
(Liang Li-ming, Chen Kang-quan, Zhong Yi, et al. DCD−YOLOv8n: efficient algorithm for steel surface defect detection [J]. *Computer Engineering and Applications*, 2025, 61(7): 117−127.)

[15] 张泽毅, 陶志勇, 林森, 等. DFE−YOLO:高效且轻量化的钢材缺陷检测网络[J]. 激光与光电子学进展, 2025, 62(22): 395−407.
(Zhang Ze-yi, Tao Zhi-yong, Lin Sen, et al. DFE−YOLO:

efficient and lightweight steel defect detection network [J]. *Laser & Optoelectronics Progress*, 2025, 62(22): 395-407.)

[16] 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 Ming-qi, Chen Hui-ming, Xu Wei, et al. Surface defect detection on lightweight YOLOv8 steel incorporating MobileNetv3 [J]. *Science Technology and Engineering*, 2025, 25(16): 6831-6840.)

[17] 郝博, 徐新岩, 赵玉欣, 等. 基于改进YOLOv8的铆接孔表面缺陷检测[J]. 东北大学学报(自然科学版), 2024, 45(11): 1595-1603.
(Hao Bo, Xu Xin-yan, Zhao Yu-xin, et al. Surface defect detection of riveting holes based on improved YOLOv8 [J]. *Journal of Northeastern University* (*Natural Science*), 2024, 45(11): 1595-1603.)

[18] Khanam R, Hussain M. YOLOv11: an overview of the key architectural enhancements [EB/OL]. (2024-10-23) [2025-12-25]. https://arxiv.org/abs/2410.17725.

[19] Yu J Y, Shu C Y, Zhou S F, et al. FQ-PETR: fully quantized position embedding transformation for multi-view 3D object detection [EB/OL]. (2025-11-12) [2025-12-09]. https://arxiv.org/abs/2511.09347.

[20] Wu W C, Liu S J, Xia Y L, et al. Dual residual attention network for image denoising [J]. *Pattern Recognition*, 2024, 149: 110291.

[21] He Y, Song K C, Meng Q G, et al. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features [J]. *IEEE Transactions on Instrumentation and Measurement*, 2020, 69(4): 1493-1504.

[22] Redmon J, Farhadi A. YOLOv3: an incremental improvement [EB/OL]. (2018-04-08) [2025-12-25]. https://arxiv.org/abs/1804.02767.

[23] Adarsh P, Rathi P, Kumar M. YOLOv3-tiny: object detection and recognition using one stage improved model [C]// 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). Coimbatore, 2020: 687-694.

[24] Glenn J. YOLOv5 release v7.0 [EB/OL]. (2022-11-22) [2025-12-25]. https://github. com/ultralytics/YOLOv5/releases/tag/v7.0.

[25] Li C Y, Li L L, Jiang H L, et al. YOLOv6: a single-stage object detection framework for industrial applications [EB/OL]. (2022-09-07) [2025-12-25]. https://arxiv. org/abs/2209.02976.

[26] Glenn J. YOLOv8 release v8.1.0. [EB/OL]. (2024-01-10) [2025-12-25]. https://github. com/ultralytics/ultralytics/releases/tag/v8.1.0.

[27] Wang C Y, Yeh I H, Mark L H. YOLOv9: learning what you want to learn using programmable gradient information [C]// European Conference on Computer Vision (ECCV). Cham: Springer, 2025: 1-21.

[28] Chen H, Chen K, Ding G G, et al. YOLOv10: real-time end-to-end object detection [C]// Advances in Neural Information Processing Systems. Vancouver, 2024: 107984-108011.

[29] Tian Y J, Ye Q X, Doermann D. YOLOv12: attention-centric real-time object detectors [EB/OL]. (2025-02-18) [2025-12-25]. https://arxiv.org/abs/2502.12524.

[30] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization [J]. *International Journal of Computer Vision*, 2020, 128(2): 336-359.