DOI:10.19651/j.cnki.emt.2519121

面向边缘计算设备的轻量化带钢缺陷检测算法*

任 帅[1] 杨思念[1] 曹立佳[2,3] 郭川东[2,3] 刘艳菊[2,3]
(1.四川轻化工大学计算机科学与工程学院 宜宾 644000;2.四川轻化工大学自动化与信息工程学院 宜宾 644000;
3.企业信息化与物联网测控技术四川省高校重点实验室 宜宾 644000)

摘 要:针对带钢缺陷检测过程中小目标难以识别以及目标检测算法参数量过大影响模型部署的问题,本文提出了一种基于 YOLOv11 框架的轻量化带钢缺陷检测算法。算法通过集成双向特征金字塔网络(BiFPN)和全局注意力机制(GAM),优化模型对小目标的识别能力。同时,采用轻量化模型 SlimNeck 加入颈部网络,在保持精度的同时降低模型参数和复杂度。最后,在 NEU-DET 数据集上对模型进行实验验证,结果表明,改进后的 BSG-LiteYOLO 检测模型性能显著优于基线算法 YOLOv11n,BSG-LiteYOLO 在参数量上降低 25.6%,模型权重大小降低 22.6%,运算量降低 7.94%,并且 mAP@0.5 提高 4.19%。实验验证 BSG-LiteYOLO 在解决带钢表面缺陷检测问题上的可行性,并将 BSG-LiteYOLO 部署于 Jetson Orin Nx 边缘计算设备,FPS 达到 34.3,符合实际生产要求。

关键词:目标检测;BiFPN;GAM;缺陷检测;边缘计算设备
中图分类号:TN99 **文献标识码**:A **国家标准学科分类代码**:510.1050

Lightweight strip steel defect detection algorithm for edge computing devices

Ren Shuai[1] Yang Sinian[1] Cao Lijia[2,3] Guo Chuandong[2,3] Liu Yanju[2,3]
(1. School of Computing Science and Engineering, Sichuan University of Science and Engineering, Yibin 644000, China;
2. School of Automation and Information Engineering, Sichuan University of Science and Engineering, Yibin 644000, China;
3. Key Laboratory of Higher Education of Sichuan Province for Enterprise Informationalization and Internet of Things, Yibin 644000, China)

Abstract: To address the problem of small targets being difficult to identify during strip defect detection and the large number of target detection algorithm parameters affecting model deployment, this paper proposes a lightweight strip defect detection algorithm based on the YOLOv11 framework. The algorithm enhances small target recognition capability through the integration of a bidirectional feature pyramid network (BiFPN) and GAM attention mechanism. Simultaneously, a lightweight SlimNeck model is adopted to integrate the neck network to reduce model parameters and complexity while maintaining detection accuracy. Experimental validation on the NEU-DET dataset demonstrates that the improved BSG-LiteYOLO detection model based on YOLOv11n significantly outperforms the original YOLOv11n. The optimized model achieves a 25.6% reduction in parameters, 22.6% decrease in model weight size, 7.94% reduction in floating point operations (FLOPs), while improving mAP@0.5 by 4.19%. Experimental results demonstrate the feasibility of the improved model for steel strip surface defect detection, with the optimized algorithm successfully deployed on Jetson Orin Nx edge computing devices achieving 34.3 FPS, which meets practical production requirements.

Keywords: object detection;BiFPN;GAM;defect detection;edge computing device

0 引 言

带钢作为现代工业中的重要材料,广泛应用于汽车制造、家电生产、建筑结构等领域。其表面质量直接影响到最终产品的性能和安全性。然而,在带钢生产过程中,由于轧制工艺、设备磨损、环境因素等影响,带钢表面常会出现裂纹、划痕、夹杂物、孔洞等缺陷,对产品性能产生直接影响[1]。这些缺陷不仅降低产品的机械性能,还可能引发严

收稿日期:2025-06-16
*基金项目:四川轻化工大学自然科学基金(2024RC03)、企业信息化与物联网测控技术四川省高校重点实验室项目(2024WYJ05)、四川省科技计划(2024NSFSC2048)、四川轻化工大学科研创新团队计划(SUSE652A011)、四川轻化工大学研究生创新基金(Y2025099)项目资助

重的安全事故。因此,开发高效、准确的带钢表面缺陷检测技术具有重要现实意义。

传统带钢缺陷检测方法主要依赖于人工目检,这种方法虽然简单直观,但存在主观性强、效率低、易疲劳等问题,难以满足大规模生产的需求。基于传统计算机视觉的检测方法,依赖根据预先定义好的模板特征进行匹配识别,需要人工精确选择提取区域和特征点位,过程繁杂适应性差,易受外部环境干扰,而且方法的泛化性普遍较低,检测速度与性能成反比。近年来,随着计算机视觉和深度学习技术的快速发展,基于深度学习的带钢缺陷检测方法逐渐成为研究热点。与传统方法相比,基于深度学习的检测方法能够自动提取缺陷特征,具有更高的检测精度和更强的适应性[2]。目前,基于深度学习的缺陷检测算法主要分为两类:两阶段检测算法和一阶段检测算法。两阶段检测算法以 Faster R-CNN[3] 和 Mask R-CNN[4] 为代表,通过首先生成候选区域,再对候选区域进行分类和回归,具有较高的检测精度。然而,这类算法计算复杂度高,难以满足实时检测的需求。一阶段检测算法则以 YOLO 系列[5]、SSD[6] 和 RetinaNet[7] 为代表,通过直接预测目标的位置和类别,实现检测速度显著提升,但在小目标缺陷检测方面仍存在不足。

针对上述问题,研究者们提出多种改进方法。例如,Lin 等[8]提出特征金字塔网络(feature pyramid networks, FPN),通过自上而下的结构增强对小目标特征的提取能力,但单向信息流限制特征融合的效果。刚帅等[9]根据加权双向特征金字塔网络(weighted bi-directional feature pyramid network, BiFPN)的特征融合方式,在路径聚合网络上(path aggregation network, PANet)增加自顶向上的多尺度特征融合,更好地平衡不同尺度的特征息,但同样模型精度丢失明显。胡玮等[10]提出一种轻量化的 YOLOv7 算法,引入多分支重参数化卷积块重构聚合模块,提升模型检测精度,但是模型参数量较大。崔丽莎等[11]提出一种基于人类视觉认知机制的表面缺陷检测网络,模拟人类视网膜的工作原理,结合缺陷图像的高空间频率局部细节信息和低空间频率全局语义信息,增强浅层与深层特征的融合能力,在检测精度和速度之间取得较好的平衡,但模型参数量有所增加。徐莲蓉等[12]提出一种改进的 YOLOv8 算法,在颈部网络中引入空间和通道重构卷积 SCConv 模块,提高小目标检测精度。Song 等[13]提出一种基于 YOLOv8 算法的多向优化改进模型 PIC2f-YOLO,虽然精度有所提高,但是可变形卷积 DCN 带来的计算复杂度上升,也显著增加模型的计算复杂度和参数量。胡依伦等[14]提出一种改进 YOLOv8n 的金属表面缺陷检测轻量化方法,设计局部卷积倒置交叉融合模块,减少算法的参数量并提升模型整体的特征提取能力,但是模型的检测性能较差。王林琳等[15]提出一种基于 YOLOv5s 的改进算法,通过更丰富的位置信息提升小目标缺陷的检测效果,但是尽管增加大尺度预测层,但对于极端小目标(如像素极少的缺陷),检测效果可能仍然有限。Lu 等[16]提出一种基于 YOLOv8 的模型 WSS-YOLO,用于工业钢材表面缺陷的精确检测,采用动态蛇形卷积,使模型能够自适应调整感受野,增强对复杂缺陷的捕捉能力。Shen 等[17]提出一种轻量化的 YOLO 模型 Mobile-YOLO-SDD,用于实时、高精度的钢材表面缺陷检测。将主干网络替换为 MobileNetV2 减少模型大小和计算复杂度。使用 K-Means++算法重新生成锚框并确定最佳尺寸,提升锚框对实际目标的适应性。但该模型的实时性较差。

尽管当前基于深度学习的带钢缺陷检测方法已取得显著进展,但在小目标检测精度、模型部署和计算效率方面仍存在提升空间。因此本文提出一种基于 YOLOv11n 的轻量化模型 BSG-LiteYOLO,用于实时、高效的处理缺陷数据。在满足工业部署要求的前提下,通过以下技术路线实现性能优化:首先,采用多尺度特征融合网络增强模型对钢材表面复杂背景适应能力;其次,引入全局注意力机制提升对小尺度缺陷检测精度;最后,使用 SlimNeck 结构在降低模型参数量的同时保持实时处理能力。本文其余部分安排如下:第 1、2 节为算法介绍,分别是 YOLOv11 算法与改进后的算法,第 3 节进行对比和消融实验,最后总结本文的研究成果以及对未来发展方向提出展望。

1　YOLOv11 目标检测算法

YOLOv11[18] 是由 Ultralytics 于 2024 年发布的 YOLO 系列最新算法,YOLOv11n 的网络结构由骨干网络(Backbone)、颈部网络(Neck)和检测头(Head)3 部分组成,通过模块化设计平衡性能与效率。YOLOv11 的 Backbone 层通过 C2PSA 模块实现多层次语义特征的精准提取,该模块融合跨阶段部分连接(cross stage partial, CSP)与多尺度注意力机制,显著提升对钢材表面缺陷(如裂纹、斑块)的敏感度。Backbone 层通过从钢材缺陷的图像中获取高级语义信息,为后续的缺陷识别提供有力支撑。Neck 层通过上采样层和 C3K2 模块进一步提取和融合不同尺度的特征信息,增强模型的泛化能力。与 YOLOv8 相比,YOLOv11 在分类检测头中加入两个深度可分离卷积(depthwise convolution, DWConv),从而大幅度减少模型的参数量和计算量。同时 Head 采用基于无锚框的方法,预测目标边界框,使用解耦合的方法将分类和检测任务分开,提高目标预测准确性。

2　改进 YOLOv11 算法

为进一步提升模型的检测性能和移动设备的部署效率,本文在 YOLOv11 模型的基础上提出改进后的钢材缺陷检测算法 BSG-LiteYOLO。其整体结构如图 1 所示:在 Backbone 部分,首先通过常规卷积与轻量化的 GSConv 提取浅层特征,并结合多次 C3k2 模块 逐步加深网络深度;在高层语义阶段引入 SPPF 获取大感受野信息,并通过

C2PSA 模块进行空间与通道注意力增强,从而获得更具判别力的多尺度特征。在 Neck 部分,将 BiFPN[19] 与 SlimNeck[20] 相结合,通过多层 BiFPN Add 节点实现自顶向下与自底向上的双向信息流动,并在关键融合位置引入 VoV-GSCSP 结构,以保证梯度传递与特征表达的高效性,同时在横向连接处采用 GSConv 以减轻计算开销。最终在 Head 部分,输出三路多尺度特征分别进入检测头,其中高分辨率分支在检测前增加了全局注意力机制(global attention mechanism, GAM)[21],以突出小目标特征并抑制复杂背景干扰。通过上述设计,BSG-LiteYOLO 在保证检测精度的同时显著降低了计算量,更适合在移动端设备上部署。

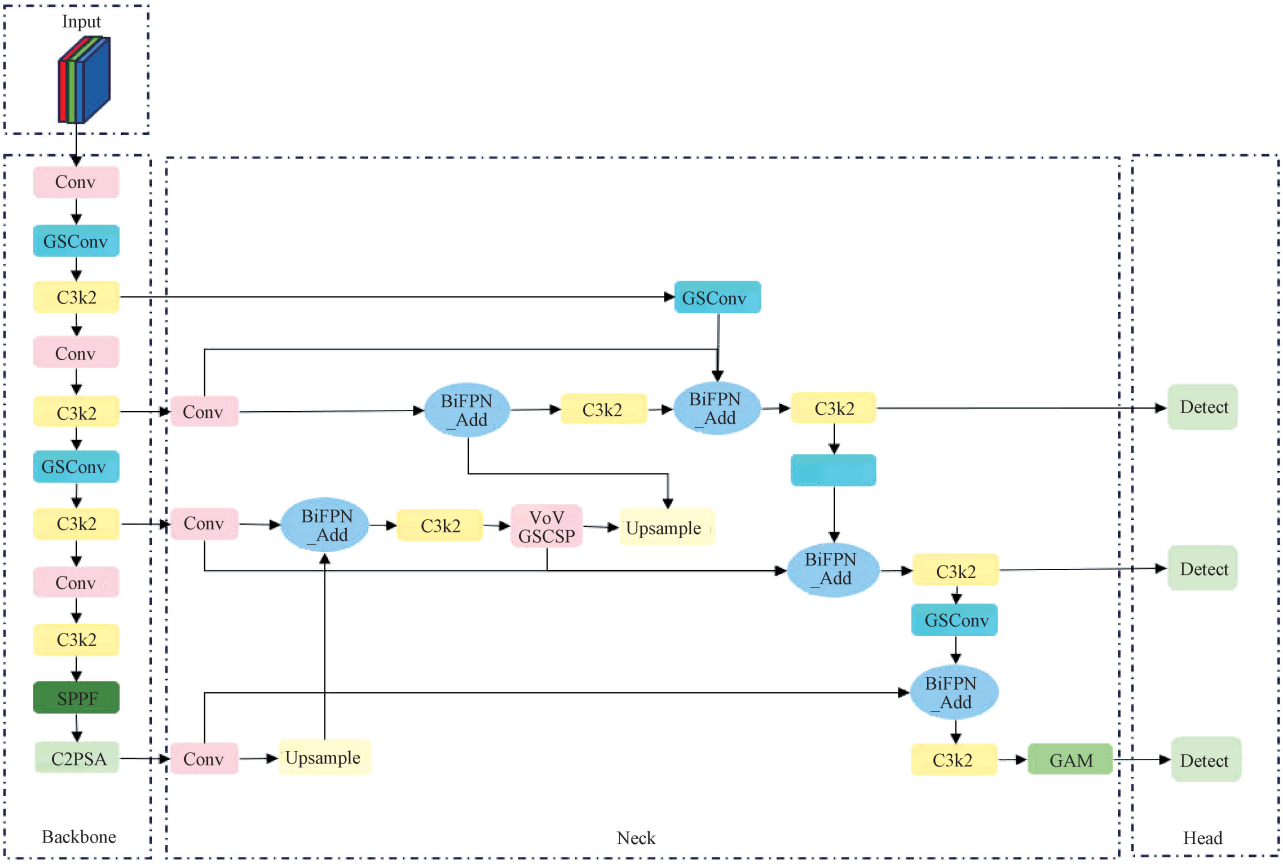

图 1　BSG-LiteYOLO 网络结构

Fig. 1　BSG-LiteYOLO network architecture

2.1　双向特征金字塔网络

在钢材缺陷检测任务中,缺陷的大小、形状各异,需要依赖于模型的多尺度特征提取以及融合能力。YOLOv11 网络模型 Neck 层在特征提取过程中,对于小目标检测的注意力比较匮乏,因此在本文提出在 YOLOv11 模型中加入 BiFPN,采用加权融合的方式融合不同分辨率特征图中的特征信息。对于每个输入都加入一个权值,网络在训练过程中对融合的权值进行优化。BiFPN 是由谷歌大脑团队在 EfficientDet 目标检测算法中引入 BiFPN,该网络是在 PaNet 网络的基础的进一步改进,其网络结构如图 2 所示。通过跳跃连接丰富特征传输过程中的语义信息,具备更高效的特征融合和表示能力,从而增强多尺度目标检测性能。如图 3 所示。描述 P_3 层如何融合相邻层的特征信息。特征融合公式如式(1)、(2)所示,仍然以 P_3 层的特征融合为例:

$$P_3^m = Conv\left(\frac{w_1 * P_3^{in} + w_2 * Deconv(P_4^{in})}{w_1 + w_2 + \mu}\right) \tag{1}$$

$$P_3^{out} = Conv\left(\frac{w'_1 * P_3^{in} + w'_2 * P_3^m + w'_3 * P_2^{out}}{w'_1 + w'_2 + w'_3 + \mu}\right) \tag{2}$$

式中:P_3^{in} 表示第 3 层的输入特征信息;P_3^{out} 表示第 3 层的输出特征信息;w 与 w' 分别表示网络融合的权值以及优化后的权值;Deconv 表示反卷积操作,将 P_4 特征信息增加分辨率与 P_3 特征信息分辨率达到一致,增强特征融合效率;Conv 代表卷积操作;μ 是一个值为 0.000 1 的常数,防止分母为 0。在式(1)中通过加权融合的方式,结合 P_3 和 P_4 的特征信息,经过卷积操作得到中间特征层。式(2)使用网络优化后的权值进一步融合当前输入特征层、上一输出特征层以及当前中间特征层的特征信息,从而得到最终输出的特征信息。

2.2　SlimNeck 结构

SlimNeck 是由分组空间瓶颈(grouped spatial bottleneck,

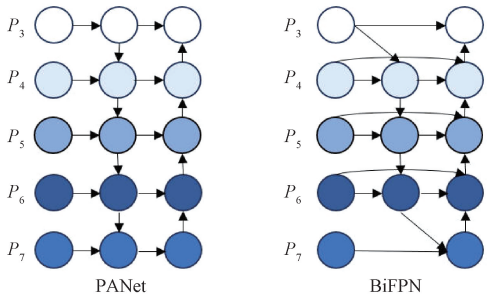

图 2　PANet 与 BiFPN 网络结构表示

Fig. 2　PANet and BiFPN network structures

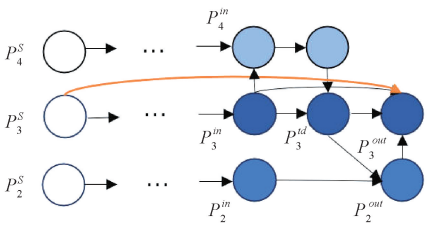

图 3　融合特征层过程

Fig. 3　Fused feature layer process

GSBottleneck)组成的特征网络轻量化策略。GSConv 模块的输入输出与普通卷积模块相同,经过普通卷积和 DWConv 得到的特征图的通道数均为 $C1/2$,其中,DWConv 使得该模块实现接近标准卷积的性能,同时降低计算复杂度。GSConv 模块如图 4 所示。在普通卷积中,卷积核的维度等于输入维度,输出维度等于卷积核的个数,但在 DWConv 中,一个通道仅被一个卷积核卷积,因此卷积核个数等于输入维度,也等于输出维度。GSConv 模块的计算复杂度如式(3)所示。

$$O_{FLOPs2}=W\times H\times\frac{C_1}{2}\times s\times s\times C_0+W\times H\times\frac{C_1}{2}\times s\times s \tag{3}$$

将普通卷积和 GSConv 的计算复杂度进行比较,如式(4)所示。

$$\frac{O_{FLOPs1}}{O_{FLOPs2}}=\frac{W\times H\times C_1\times s\times s\times C_0}{W\times H\times\frac{C_1}{2}\times s\times s\times C_0+W\times H\times\frac{C_1}{2}\times s\times s}=\frac{2C_0}{C_0+1}\approx 2 \tag{4}$$

式中:O_{FLOPs1} 和 O_{FLOPs2} 分别代表标准卷积与 GSConv 模块的计算复杂度,W 和 H 分别为输入特征图的宽度和高度,C_0 和 C_1 分别为输入特征图的通道数与输出特征图的通道数,s 为卷积核的尺寸。

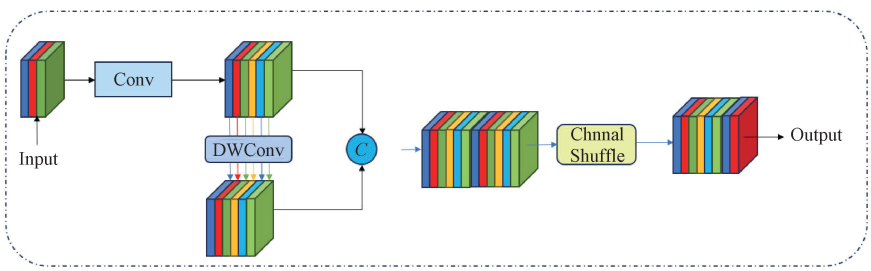

图 4　GSConv 结构

Fig. 4　GSConv structure

当 $C_0\geqslant 1$ 时,普通卷积的计算复杂度是 GSConv 模块的 2 倍。VOVGSCSP 结构采用一次性聚合方法,由 GSConv 构建而成,结构如图 5 所示,首先将输入分为两部分,一部分经过卷积操作实现通道数缩减,另一部分在改变通道数后经过 GSBottleneck,然后将这两部分输出拼接在一起,再进行卷积操作以实现输出通道的改变。该结构使得不同位置的梯度能够交叉混合,从而增强网络的梯度表现和学习能力。

2.3　全局注意力机制

由于带钢样例图像中缺陷目标通常较小,增加检测的难度。为提高模型的检测性能,在颈部网络中引入全局注意力机制(global attention mechanism,GAM),实现对关键信息的捕获,显著降低小目标缺陷的检测难度。GAM 注意力机制强调目标全局维度的相互作用,更加关注关键信息。GAM 提出全新的空间通道注意力机制并取代 CBAM[22] 模型的子模块。特征提取过程如图 6 所示。

下面的公式建立从输入特征到中间特征表示,再到最终输出特征的非线性映射关系。

$$F_2=M_c(F_1)\times F_1 \tag{5}$$

$$F_3=M_s(F_2)\times F_2 \tag{6}$$

式中:M_c 和 M_s 分别为通道注意模块和空间注意模块;F_1、F_2 和 F_3 分别表示输入特征、中间特征和输出特征。

1)通道注意力子模块

在 GAM 模型结构中,通道注意力子模块采用三维排列结构以保留 3 个维度的信息,随后通过双层多层感知机(MLP)结构模块来增强不同维度上通道和空间元素的关联性。该模块结构如图 7 所示。

2)空间注意力子模块

在空间注意力子模块中,为更好地聚焦空间信息,该

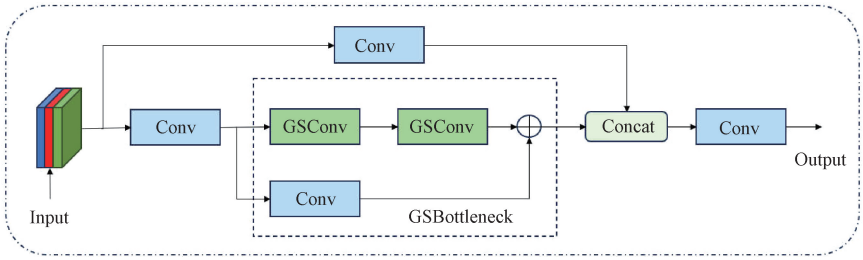

图 5　VOVGSCSP 结构

Fig. 5　VOVGSCSP structure

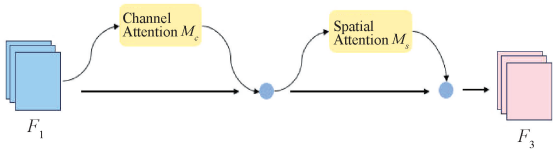

图 6　GAM 结构

Fig. 6　GAM structure

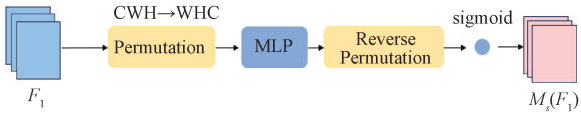

图 7　通道注意力子模块

Fig. 7　Channel attention sub-module

模型采用两个卷积层进行空间维度上的信息融合。该结构放弃池化操作以进一步保护局部特征信息,因为最大池化操作会导致部分空间信息丢失。虽然这可能增加参数量,但能更完整地收集空间信息,不易遗漏部分特征映射。模块结构如图 8 所示。在带钢场景的目标检测任务中,目标物体尺寸小、背景噪声过大,这会造成大量无效区域的训练,进而影响网络训练效率。在 YOLOv11 目标检测模型的特征提取层加入 GAM 后,通过收集不同维度的特征信息并反馈,减少采样过程中成像特征信息的损失,充分运用各维度感受野的视觉表征能力,从而能在实际采集过程中实现性能提升。

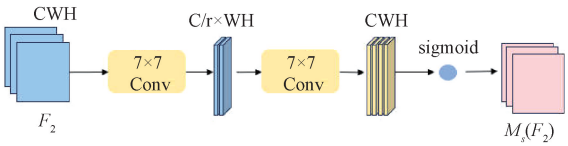

图 8　空间注意力子模块

Fig. 8　Spatial attention sub-module

3　实验与分析

3.1　实验平台与评价指标

本文算法基于 PyTorch 1.11.0 深度学习框架与 Python 3.8.19 编程环境构建,算法运行环境基于 Ubuntu 18.04 操作系统。部署测试基于 Jetson Orin Nx 嵌入式平台。所使用 Jetpack 版本为 5.1.2。详细参数如表 1 所示。

表 1　训练及测试环境

Table 1　Training and testing environment

类型	训练服务器	边缘计算设备
设备	DELL Precision R7920	Jetson Orin Nx
CPU	Intel Xeon(R) Gold 6254 CPU	8-core Arm Cortex ® 64-bit
GPU	Quadro RTX 8000 48 G×2	1024-core NVIDAI Ampere GPU
RAM	128 GB	16 GB

定义平均准确率(mAP)与回归率(R)评估指标为

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \times 100\% \tag{7}$$

$$R = \frac{TP}{TP + FP} \times 100\% \tag{8}$$

式中:TP 和 FP 代表真阳性和假阳性,分别表示模型正确预测样本与错误预测样本的阳性结果。FN 代表假阴性,表示模型错误预测样本的硬性结果。N 为缺陷类别的总数。初始训练参数如表 2 所示。

表 2　初始训练参数

Table 2　Initial training parameters

参数	值
输入尺寸	640×640
优化器	SGD
训练轮数	200
学习率	1×10^{-4}
预热批次	10

3.2　数据集

本文使用 NEU-DET 热轧钢带表面缺陷数据集[23],该数据集共 1 800 张图像,包含 6 种典型的热轧钢带生产过程中产生的缺陷类型:裂纹(crazing)、内含物(inclusion)、斑块(patches)、点蚀表面(pitted surface)、轧制氧化皮(rolled-in scale)和划痕(scratches),每种类型各 300 张图像。训练集、验证集和测试集按照 7∶2∶1的比例划分,其中

训练集有 1 260 张图像，验证集有 360 张图像，测试集有 180 张图像。图 9[23]展示了 6 种典型缺陷类型的图像。

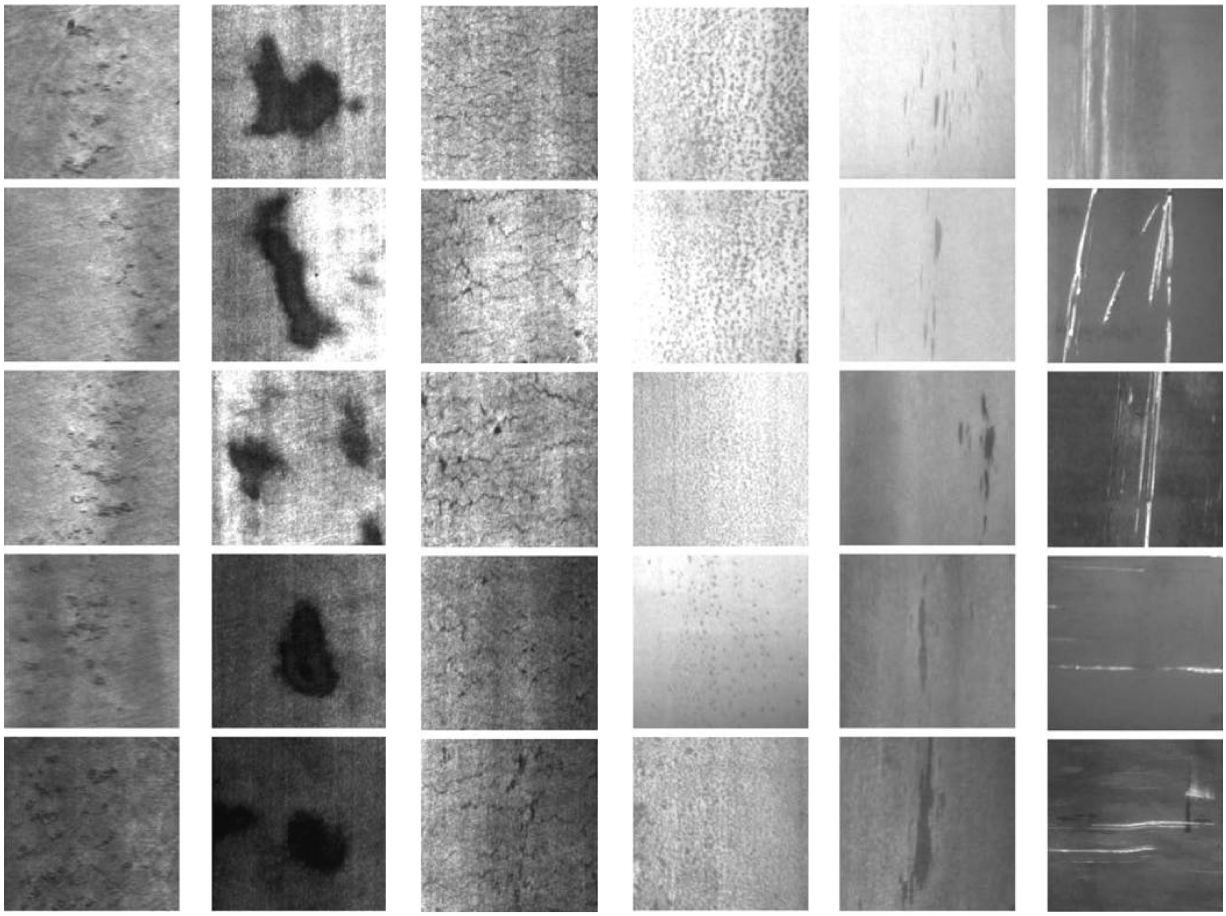

图 9　6 种钢材缺陷

Fig. 9　Six types of steel defects

3.3　对比试验

1）模型对比试验结果与分析

为了进一步验证改进算法的有效性，本实验对比本文方法与其他模型的性能，实验结果如表 3 所示。其中模型大小是指模型部署所需占用内存大小，参数量是指模型可训练参数的总数，计算量表示模型计算复杂度的指标。

表 3　各模型实验对比

Table 3　Comparison of various model experiments

模型	mAP@0.5	mAP@0.5:0.95	模型大小/MB	参数量/10^6	计算量/10^9
YOLOv5n	0.738	0.392	4.5	2.18	5.8
YOLOv6n	0.739	0.404	16.2	4.15	11.5
YOLOv8n	0.736	0.386	5.4	2.68	6.8
RTDETR-r18	0.703	0.379	38.6	19.87	57.0
Mobile-YOLO-SDD	**0.773**	—	5.0	2.50	6.3
YOLOv11n	0.740	0.407	5.3	2.58	6.3
本文算法	0.771	**0.422**	**4.1**	**1.92**	**5.8**

由表 3 可知，基于带钢缺陷检测任务的实验数据对比分析，本文算法在轻量化设计与检测精度方面展现出显著优势。从模型效率来看，本文算法的参数量仅为 1.92 M，远低于 RTDETR-r18，同时较 YOLOv8 减少 28.4%。模

型体积压缩至 4.1 MB,较 YOLOv5 和 YOLOv8n 分别降低 8.9%和 24.1%,更仅为 YOLOv6n 的 25.3%。在计算复杂度方面,本文算法的 5.8 GFLOPs 与 YOLOv5n 持平,但显著低于 YOLOv8n 和 YOLOv6,尤其 RTDETR-r18 的计算量高达 57GFLOPs,约为其 9.8 倍,凸显了本文算法在工业场景中低功耗部署的潜力。在检测性能方面,本文算法的 mAP@0.5 达到 0.771,较 YOLOv8 提升 4.8%,较 RTDETR-r18 提升 9.7%。对于更严格的 mAP@0.5:0.95 指标,其 0.422 的精度同样领先于所有对比模型,较 YOLOv11 提升 3.7%。本文还对比了近期研究人员提出的改进方法,本文算法在精度上与 Shen 等[17]提出的 Mobile-YOLO-SDD 方法精度相近,但本文算法的模型尺寸较其减少 18%,参数量降低 23.2%,计算量减少 7.9%。综上所述,本文方法在较低的参数量和计算复杂度下显著提升了带钢缺陷检测的性能,因此总体优于其他检测模型。

2) 特征融合方式对比试验结果与分析

为评估不同特征金字塔架构对检测性能的影响,本研究在统一实验设置下对比 PAFPN[24]、BIMAFPN[25]、GFPN[26]和 BiFPN 4 种模型的性能表现。如表 4 所示,采用双向递归融合机制的 BiFPN 展现出显著的综合性优势。在检测精度方面,BiFPN 以 0.762 的 mAP@0.5 指标位居首位,较基准 PAFPN 提升 2.97%。值得注意的是,BiFPN 在保持 6.3 GFLOPs 轻量级计算消耗的同时,将参数量压缩至 1.9×10^6,较 PAFPN 减少 26.4%,这得益于其创新的参数共享机制和精简的跨尺度连接拓扑。尽管 BIMAFPN 实现 4.8 MB 的较小模型体积,但其 0.728 的 mAP@0.5 暴露出精度与效率的失衡问题。相比之下,BiFPN 在模型复杂度控制方面表现更为优异,将 mAP@0.5:0.95 指标提升至 0.42。该实验结果验证双向递归融合机制在平衡检测精度与计算效率方面的有效性。

表 4　特征融合方式对比实验结果

Table 4　Comparative experimental results of feature fusion methods

模型	mAP@0.5	mAP@0.5:0.95	模型大小/MB	参数量/10^6	计算量/10^9
YOLO11n-PAFPN	0.740	0.407	5.3	2.58	6.3
YOLO11n-BIMAFPN	0.728	0.391	4.8	2.30	7.1
YOLO11n-GFPN	0.728	0.394	14.7	3.60	8.2
YOLO11n-BiFPN	**0.762**	**0.420**	**4.0**	**1.90**	**6.3**

3.4　消融实验

通过系统性消融实验验证 SlimNeck(A)、GAM 注意力(B)和 BiFPN(C)3 个模块的协同优化效果。如表 5 所示,单独应用时,模块 C 在精度-效率平衡性上表现最优,其 mAP@0.5 达到 0.762 的同时维持 4.0 MB 模型体积与 6.3 计算量,较 A 的 0.764 mAP@0.5 虽略有降低,但参数量减少 26%。当模块组合使用时,模块 A 与模块 C 的融合展现出显著的协同效应,在维持 mAP@0.5 为 0.764 精度的前提下,将计算量压缩至 5.7 GFLOPs,较单模块 C 降低 9.5%,验证了窄颈结构与双向特征金字塔结构的架构兼容性。值得注意的是,3 模块联合架构 A+B+C 以 1.92×10^6 参数量实现 mAP@0.5 为 0.771 的峰值精度,较单模块最佳结果提升 0.9%,其 4.1 MB 的紧凑模型尺寸较原始模块 B 单独应用的 8.4 MB 缩减 51.2%,表明全局注意力机制在多层特征精炼中的增效作用。因此,相较于基线模型 YOLOv11n,本文模型参数量减少了 25.6%,计算量减少了 7.94%,模型大小减少了 22.6%,mAP@0.5 提高了 4.19%,各模块能够轻量化网络结构,同时提升带钢缺陷检测的精度。

表 5　消融实验对比结果

Table 5　Comparison results of ablation experiments

模块	mAP@0.5	mAP@0.5:0.95	模型大小/MB	参数量/10^6	计算量/10^9
A	0.764	0.409	5.3	2.57	5.9
B	0.751	0.395	8.4	4.20	7.6
C	0.762	0.420	4.0	1.90	6.3
A+B	0.767	0.417	8.4	4.20	7.3
A+C	0.764	0.421	**4.0**	**1.82**	5.7
B+C	0.760	0.407	4.2	2.02	6.4
A+B+C	**0.771**	**0.422**	4.1	1.92	**5.8**

3.5　部署验证

将本文所提出的 BSG-LiteYOLO 带钢缺陷检测算法,部署于 Nvidia 公司所生产的 Jetson Orin Nx 边缘设备,验证改进算法的性能。

随机抽取 300 张图片进行边缘计算设备与 PC 端对比试验。如表 6 所示，BSG-LiteYOLO-jetson 在 Jetson 边缘计算平台上展现出显著的综合优势。在检测精度方面，该模型以 0.771 的 mAP@0.5 超越 YOLOv11-jetson 的 0.739，实现 4.2%的性能提升，这意味着在实际应用中能够提供更准确的检测结果。在模型效率方面，BSG-LiteYOLO-jetson 仅占用 4.1 MB 存储空间，较 YOLOv11-jetson 的 5.3 MB 减小 22.6%，这种轻量化特性使其特别适合资源受限的边缘部署场景。尽管在推理速度上 34.3 FPS 略低于对比模型，但其通过更优的精度-速度平衡设计，在保证检测质量的同时仍能满足实时性要求，符合实际生产要求。

表 6　边缘设备对比试验

Table 6　Edge device comparison test

模型	mAP@0.5	模型大小/MB	FPS
YOLOv11-jetson	0.739	5.3	53.8
BSG-LiteYOLO-jetson	0.771	4.1	34.3

3.6　实验结果及展示

为更具体的展示本文算法与其他算法的比较结果，分别对 YOLOv11n 和改进后的 BSG-LiteYOLO 在测试集上进行测试，图 10 展示在 NEU-DET 数据集上的检测结果。从中可以看出每个框代表检测到的缺陷区域，每个缺陷区域上方标注缺陷类别、缺陷的置信度，并且改进之后的模型能够更精确地定位缺陷目标的位置，实现对缺陷的定位和分类，更接近实际目标的位置和大小，对图像的分析能展示出改进之后模型的高效性和精确性。

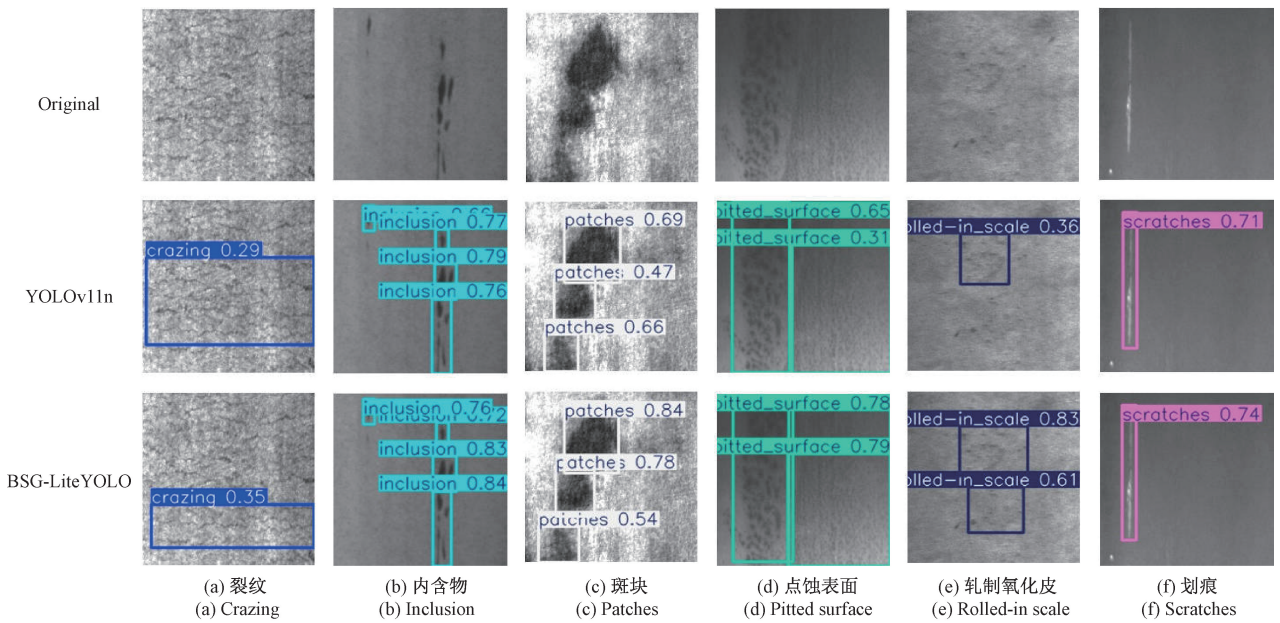

图 10　NEU-DET 数据集检测结果

Fig. 10　Detection results on the NEU-DET dataset

4　结　　论

本文针对带钢缺陷检测问题，提出一种基于改进 YOLOv11 的轻量化带钢缺陷检测算法 BSG-LiteYOLO。该方法首先加入双向特征金字塔网络 BiFPN，然后在颈部网络中加入 SlimNeck 结构，最后在颈部网络中最后一层加入全局注意力机制。通过以上 3 个方面的轻量化改进，显著减少模型的参数量以及计算复杂度，同时提高模型的检测精度。本文方法不仅在准确度上不逊色传统方法，而且还降低模型在计算资源受限设备上的部署难度。在未来的工作中，计划进一步探索蒸馏、剪枝等轻量化技术，增加模型的性能和适用范围，推动模型在实际生产中的应用。

参考文献

[1]　周李洪，龚金科，李兵. 基于稀疏表示的车用带钢表面图像信息修复[J]. 湖南大学学报(自然科学版)，2021，48(8)：141-148.
ZHOU L H，GONG J K，LI B. Image information restoration of automotive strip steel surface based on sparse representation[J]. Journal of Hunan University (Natural Sciences)，2021，48(8)：141-148.

[2]　张睿，高美蓉，傅留虎，等. 基于多域多尺度深度特征自适应融合的焊缝缺陷检测研究[J]. 振动与冲击，2023，42(17)：294-305.
ZHANG R，GAO M R，FU L H，et al. Weld defect detection based on adaptive fusion of multi-domain and multi-scale deep features[J]. Journal of Vibration and Shock，2023，42(17)：294-305.

[3]　REN SH，HE K，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, 2016, 39(6): 1137-1149.

[4] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]. IEEE International Conference on Computer Vision, 2017: 2961-2969.

[5] 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, 2016: 779-788.

[6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot MultiBox detector [C]. European Conference on Computer Vision, 2016: 21-37.

[7] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]. IEEE International Conference on Computer Vision, 2017: 2980-2988.

[8] LIN T, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117-2125.

[9] 刚帅, 刘培胜, 郭希旺. 改进基于 YOLOv8n 的轻量化钢材表面缺陷检测算法[J]. 电子测量技术, 2025, 48(3):74-82.
GANG SH, LIU P SH, GUO X W. Improved lightweight steel surface defect detection algorithm based on YOLOv8n [J]. Electronic Measurement Technology, 2025, 48(3): 74-82.

[10] 胡玮, 赵菊敏, 李灯熬. 基于改进 YOLOv7-Tiny 的轻量化激光器芯片缺陷检测算法[J]. 太原理工大学学报, 2025, 56(1): 137-147.
HU W, ZHAO J M, LI D AO. A lightweight laser chip defect detection algorithm based on improved YOLOv7-Tiny[J]. Journal of Taiyuan University of Technology, 2025, 56(1): 137-147.

[11] 崔丽莎, 代润鹏, 姜晓恒, 等. 基于人类视觉认知机制的表面缺陷检测[J]. 浙江大学学报(理学版), 2025, 52(1): 38-49.
CUI L SH, DAI R P, JIANG X H, et al. A human visual cognitive mechanism based network for surface defect detection[J]. Journal of Zhejiang University (Science Edition), 2025, 52(1): 38-49.

[12] 徐莲蓉, 梁少华. 改进 YOLOv8 的钢材表面缺陷检测算法[J]. 现代电子技术, 2025, 48(4): 173-180.
XU L R, LIANG SH H. Improved YOLOv8 steel surface defect detection algorithm [J]. Modern Electronics Technique, 2025, 48(4): 173-180.

[13] SONG X, CAO SH ZH, ZHANG J, et al. Steel surface defect detection algorithm based on YOLOv8 [J]. Electronics, 2024, 13(5): 988.

[14] 胡依伦, 杨俊, 许聪源, 等. PIC2f-YOLO:金属表面缺陷检测轻量化方法[J]. 光电工程, 2025, 52(1): 89-103.
HU Y L, YANG J, XU C Y, et al. PIC2f-YOLO: A lightweight method for the detection of metal surface defects[J]. Opto-Electron Engineering, 2025, 52(1): 89-103.

[15] 王林琳, 龚昭昭, 梁泽启. 改进 YOLOv5s 算法的带钢表面缺陷检测[J]. 组合机床与自动化加工技术, 2024(12): 181-186.
WANG L L, GONG ZH ZH, LIANG Z Q. Improved YOLOv5s strip surface defect detection method[J]. Modular Machine Tool & Automatic Manufacturing Technique, 2024(12): 181-186.

[16] LU M, SHENG W, ZOU Y, et al. WSS-YOLO: An improved industrial defect detection network for steel surface defects[J]. Measurement, 2024, 236: 115060.

[17] SHEN L, XU Y, ZHU M, et al. Mobile-YOLO-SDD: A lightweight YOLO for real-time steel defect detection[J]. Procedia CIRP, 2024, 129: 228-233.

[18] KHANAM R, HUSSAIN M. YOLOv11: An overview of the key architectural enhancements[J]. ArXiv preprint arXiv: 2410.17725, 2024.

[19] TAN M, PANG R, LE Q V. EfficientDet: Scalable and efficient object detection [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10778-10787.

[20] LI H, LI J, WEI H, et al. Slim-Neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles[J]. ArXiv preprint arXiv: 2206.02424, 2022.

[21] LIU Y, SHAO Z, HOFFMANN N. Global attention mechanism: Retain information to enhance channel-spatial interactions [J]. ArXiv preprint arXiv: 2112.05561, 2021.

[22] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]. European Conference on Computer Vision, 2018: 3-19.

[23] HE Y, SONG K, MENG Q, 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.

[24] LIU SH, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 2157-2167.

[25] YANG ZH Q, GUAN Q, ZHAO K, et al. Multi-

branch auxiliary fusion YOLO with re-parameterization heterogeneous convolutional for accurate object detection[C]. Chinese Conference on Pattern Recognition and Computer Vision, Singapore: Springer Nature Singapore, 2024: 492-505.

[26] XU X, JIANG Y, CHEN W, et al. DAMO-YOLO: A report on real-time object detection design[J]. ArXiv preprint arXiv: 2211.1544, 2022.

作者简介

任帅,硕士研究生,主要研究方向为目标识别。

E-mail:324085406125@stu.suse.edu.cn

杨思念,硕士研究生,主要研究方向为目标识别。

E-mail:323085406129@stu.suse.edu.cn

曹立佳,博士,教授,主要研究方向为目标识别、无人系统导航与控制。

E-mail:caolj@suse.edu.cn

郭川东,博士,讲师,主要研究方向为复杂系统建模与控制和机器视觉。

E-mial: guochuandong@suse.edu.cn

刘艳菊(通信作者),硕士,讲师,主要研究方向为无人机智能飞行控制。

E-mail:liuyanju_980930@163.com