DOI:10.19651/j.cnki.emt.2416098

轻量级锻件表面裂纹检测算法*

张 上[1,2,3] 许 欢[1,2,3] 张 岳[1,2,3]
(1.水电工程智能视觉监测湖北省重点实验室 宜昌 443002; 2.三峡大学湖北省建筑质量检测装备工程技术研究中心 宜昌 443002; 3.三峡大学计算机与信息学院 宜昌 443002)

摘 要: 针对复杂场景下缺陷检测算法占用内存大、计算复杂度高和检测速度难以满足实时需求等问题,本文提出一种基于 YOLOv8 的轻量级锻件缺陷检测算法。首先,采集重卡转向节生产流水线探伤车间的磁粉检测图像,构建锻件表面裂纹数据集;然后,提出轻量化卷积模块 GSConvns,以增强特征交互并降低计算量;同时,引入 Shape-IOU 损失函数,优化训练效果;最后,利用 LAMP 剪枝策略去除不重要的权重参数,减少模型体积并提高检测速度。实验结果表明,模型的 mAP 值为 83.8%,参数量和计算量分别减少 85.05%和 80.25%,检测速度从 38.7 FPS 提升至 65.6 FPS,显著优于其他主流算法,更适用于实时检测。在公开数据集上的测试进一步验证了其泛化能力,与基准算法相比,未剪枝的改进算法 mAP 值提升了 2.0%。综上,本文算法能在不显著降低检测精度的前提下,大幅度提升了检测速度和资源利用效率。

关键词: 表面缺陷检测;YOLOv8 算法;轻量化模型;损失函数;模型剪枝
中图分类号: TP391.4; TN98 **文献标识码**: A **国家标准学科分类代码**: 510.4050

Lightweight forged part surface crack detection algorithm

Zhang Shang[1,2,3] Xu Huan[1,2,3] Zhang Yue[1,2,3]
(1. Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang 443002, China; 2. Hubei Province Engineering Technology Research Center for Construction Quality Testing Equipment, China Three Gorges University, Yichang 443002, China; 3. College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China)

Abstract: To address issues of high memory usage, computational complexity, and inadequate detection speed in defect detection algorithms for complex scenarios, this paper proposes a lightweight forged defect detection algorithm based on YOLOv8. First, magnetic particle inspection images from the production line of heavy truck steering knuckles were collected to construct a forged surface crack dataset. Then, a lightweight convolution module, GSConvns, was introduced to enhance feature interaction and reduce computational load. The Shape-IOU loss function was employed to optimize training performance. Finally, the LAMP pruning strategy was used to remove unnecessary weight parameters, reducing model size and increasing detection speed. Experimental results show that the model achieves a mAP of 83.8%, with parameter and computational reductions of 85.05% and 80.25%, respectively. Detection speed improved from 38.7 FPS to 65.6 FPS, significantly outperforming other mainstream algorithms, making it more suitable for real-time detection. The algorithm's generalization capability was further verified on a public dataset, with the unpruned improved algorithm's mAP value increasing by 2.0% compared to the baseline. In summary, this algorithm significantly enhances detection speed and resource efficiency without substantially compromising detection accuracy.

Keywords: surface defect detection; YOLOv8 algorithm; lightweight model; loss function; model pruning

0 引 言

在钢铁生产中,品质保障至关重要,直接影响成品质量和生成过程中的安全性。钢铁被普遍用于航空航天、汽车零部件、建筑施工和传统制造业等领域,其质量要求日益提高。劣质原材料可能导致经济损失和安全威胁,因此提升

收稿日期:2024-05-22
* 基金项目:省级大学生创新创业计划(S202311075047)、国家级大学生创新创业训练计划(202011075013、202111075012)项目资助

钢铁缺陷检测准确性至关重要。

缺陷检测方法主要包括传统方法、机器视觉和深度学习技术。传统方法如人工抽检和红外检测存在一定的局限性，而机器视觉虽有所改进，但在多缺陷分类和特征提取上仍面临挑战。因此，基于深度学习的缺陷检测[1-2]成为研究焦点，以解决这些问题。

随着深度学习技术的快速发展，目标检测方法也取得了显著的进步。目标检测算法的发展经历了从两阶段的 R-CNN[3]和 Fast R-CNN[4]到单阶段的 YOLO[5]和 SSD[6]，后者以速度和精度优势成为实时检测的主流选择。马燕婷等[7]基于 YOLOv5 提出了一种新的多尺度特征融合算法，在 NEU-DET 数据集上平均检测精度（mean average precision，mAP）达到了 82.4%，然而模型大小达到了 29.7 M，不易部署在实际生产环境中。徐国伟等[8]在利用生成对抗网络进行数据增强的基础上，采用轻量级网络 MobileNetV3-large[9]替换 YOLOv5 主干结构，同时引入坐标注意力机制 CA，提高了检测精度，但在 NEU-DET 数集上检测精度表现不佳，mAP 仅为 76.4%。熊聪等[10]在 YOLOX 网络的基础上，通过引入 Swin Transformer 模块、BiFPN 网络[11]和改进的目标定位损失函数，实现了高精度的钢材表面缺陷检测，但模型检测速度却下降了 56.3%，实时检测能力较差。

以上方法在模型轻量化和实时性方面仍有不足，推理过程的计算资源消耗过大，这限制了其在资源受限的生产环境中的应用。针对这一问题，本文采集了重卡转向节生产线的磁粉检测数据，制作了表面裂纹数据集。为了满足工业环境中对金属零件表面缺陷检测的高性能要求，提出了一种基于 YOLOv8 的改进轻量级目标检测算法，旨在提升流水生产线上的实时缺陷检测能力。本文主要贡献如下：

本文提出一种轻量化卷积模块 GSConvns，旨在增强特征交互的同时降低计算量，并基于 GSConvns 改进 YOLOv8 颈部结构，实现模型轻量化的同时有效提升精度；其次，针对数据集中存在一定数量小目标，引入 Shape-IOU[12]损失函数，优化数据集训练效果；最后，通过基于层自适应幅度的修剪（layer-adaptive magnitude-based pruning，LAMP）[13]剪枝方案，去除模型冗余参数与通道，提高模型的推理速度。

1　YOLOv8 网络模型

YOLOv8 算法模型主要由主干特征提取模块（Backbone）、特征加强模块（Neck）、检测模块（Detect）三个部分构成，如图 1 所示。

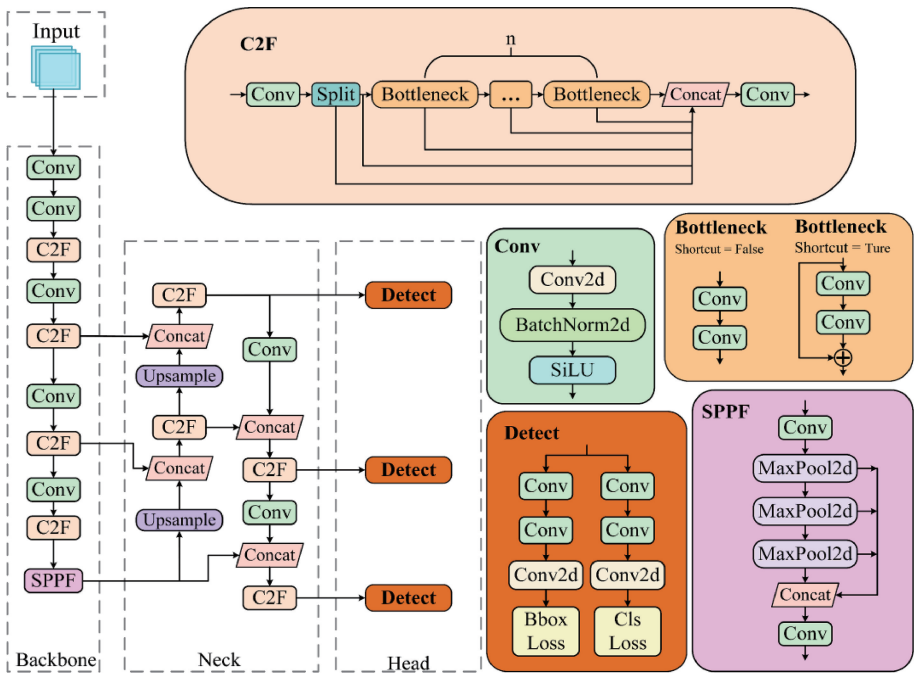


图 1　YOLOv8 算法模型结构

Fig. 1　Network structure of the YOLOv8 algorithm model

YOLOv8 在开源发布后，经过广泛测试，其准确性和速度都达到了最先进（state-of-the-art）算法水平。在 YOLOv8 中，Backbone 类似于 YOLOv5 的结构，将 C3 模块更换为 C2F 模块。在 Neck 模块中，使用了 PANet 结构，通过上采样和通道融合来强化特征。检测模块采用了解耦头结构，将回归和分类任务分离，以提高模型的收敛速度和检测效率。此外，YOLOv8 还引入了 Anchor Free 方法，使得模型更适合处理密集目标检测任务。

2　YOLOv8 网络模型的改进

本文提出算法基于 YOLOv8 算法进行改进,使用轻量级卷积模块 GSConvns 和 VoVGSCSPns 改进颈部网络;使用 Shape-IOU 替换原有的 CIOU。具体模型网络结构如图 2 所示。

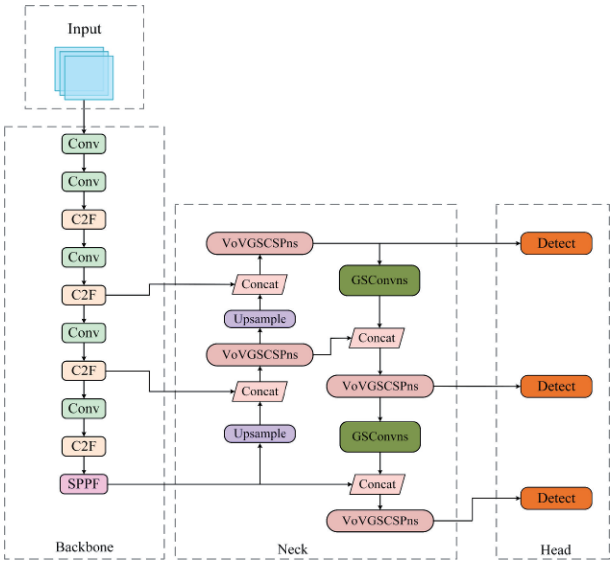


图 2　改进后的 YOLOv8 算法模型结构

Fig. 2　Network structure of the enhanced YOLOv8 algorithm model

2.1　基于轻量化卷积 GSConvns 改进颈部网络

为了适应工业生产线上对锻件表面裂纹实时检测的需求,本文引入了 GSConvns 卷积模块,该模块在增强特征交互的同时降低计算量,设计上充分考虑了计算资源受限的实际环境。在典型的深度神经网络中,所提取的特征图包含大量信息,支撑着网络对输入数据的深入分析。然而,这些网络往往会产生一些多余的特征映射。Ghost 模块[14]为此提供了一种高效且资源消耗低的解决方案。该模块先创建一系列核心特征映射,随后通过线性变换扩充这些特征图,从而增加通道维度。这不仅缩减了模型的大小,还有效挖掘了隐藏在核心特征背后的细节。2022 年,Li 等[15]基于 Ghost 模块提出了一种轻量级卷积模块 GSConv。该模块使用通道混洗[16](channel shuffle)技术将标准卷积(standard convolution)生成信息渗透到深度可分离卷积(depthwise separable convolution)生成信息的每个部分。GSConv 结构图如图 3 所示。

GSConv 模块时间复杂度为:

$$Time_{GSConv} \sim O\left[W \cdot H \cdot K_1 \cdot K_2 \cdot \frac{C_2}{2}(C_1+1)\right] \tag{1}$$

然而,标准卷积模块时间复杂度为:

$$Time_{SC} \sim O[W \cdot H \cdot K_1 \cdot K_2 \cdot C_1 \cdot C_2] \tag{2}$$

其中,W,H 分别代表输入特征图的宽度和高度;K_1,K_2 分别代表第一个卷积核和第二个卷积核的大小;C_1 代表输入特征图的通道数;C_2 代表输出特征图的通道数。由此可见,GSConv 模块显著减少了计算开销。

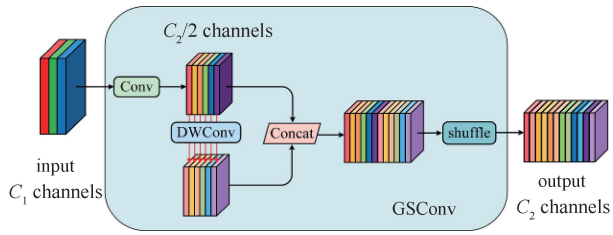


图 3　GSConv 结构图

Fig. 3　Schematic of the GSConv structure

GSConv 提供了一定的性能提升,但其通道混洗技术可能会导致信息损失。特别是在通道重组后,相邻特征通道被分开,破坏了原有的空间信息和通道之间的关系,从而影响了模型的性能和泛化能力。为解决这一问题,本文提出了一种名为 GSConvns 的卷积模块。GSConvns 模块与 GSConv 的不同之处在于:特征通道拼接之后,通过一个 1×1 卷积实现特征融合,最后应用 ReLU 激活函数,将处理后的特征图作为最终输出。这样的设计旨在保持网络性能的同时,减少计算成本,使得模型更适合资源受限的环境。GSConvns 结构如图 4 所示。

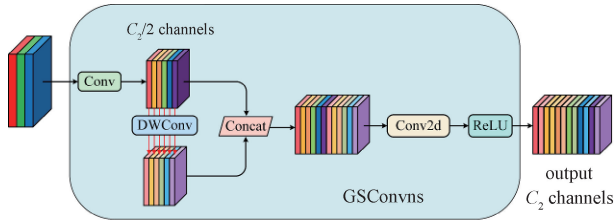


图 4　GSConvns 结构图

Fig. 4　Schematic of the GSConvns structure

采用 VoVGSCSPns 模块替代 YOLOv8 颈部网络中的 C2f 模块。通过应用 GSConvns 模块,Neck 模块能够更有效地保留输入数据的理解,从而在不增加网络参数的情况下,提高了模型识别冗余信息的速度。此结构调整不仅缩小了模型体积和提升了处理速度,同时保持了高水平的检测精度,满足了实时缺陷检测的需求。图 5 展示了使用 GSConvns 优化后的 VoVGSCSPns 结构。

2.2　Shape-IOU 损失函数

在 YOLOv8 的框架中,采用了 Complete-IoU(CIoU)损失函数,该损失函数相较于传统的 IoU 损失函数,具有显著的优势。CIoU 损失函数引入了完整性信息的考虑,包括中心点距离、宽高比和重叠面积等因素,这些因素能够更全面地反映出锚框和目标框之间的相对位置和形状关系。这种设计使得 CIoU 在训练过程中能够更有效地引导锚框回归到目标框的准确位置,从而提高了目标检测的准确性和鲁棒性。CIoU 公式如下:

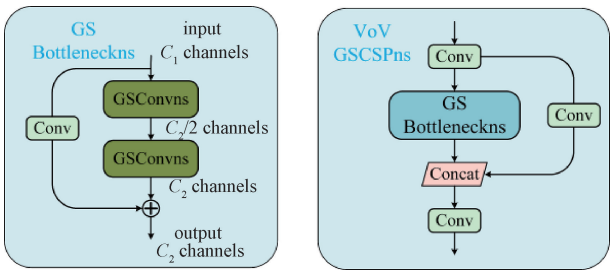


图 5　VoVGSCSPns 结构图

Fig. 5　Schematic of the VoVGSCSPns structure

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{3}$$

$$\alpha = \frac{v}{(1-IOU)+v} \tag{4}$$

$$L_{CIoU} = 1 - IoU + \frac{(x_c - x_{gt})^2 + (y_c - y_{gt})^2}{W_g^2 + H_g^2} + \alpha v \tag{5}$$

其中，w^{gt}，h^{gt} 为实际框尺寸；w，h 为预测框尺寸；IoU 是衡量预测框和实际框重叠程度的指标；x_c，y_c 代表预测框中心坐标；x_{gt}，y_{gt} 代表实际框中心坐标；W_g，H_g 是指包围预测和实际框的最小矩形框的尺寸。

然而，CIoU 忽略了边界框本身的形状和尺度也会对边界框回归产生影响，为了进一步提高回归的准确性，本文引入了 Shape-IoU 损失函数。Shape-IoU 的公式如下所示。

$$ww = \frac{2\times(w^{gt})^{scale}}{(w^{gt})^{scale}+(h^{gt})^{scale}} \tag{6}$$

$$hh = \frac{2\times(h^{gt})^{scale}}{(w^{gt})^{scale}+(h^{gt})^{scale}} \tag{7}$$

$$distance^{shape} = hh\times\frac{(x_c - x_{gt})^2}{W_g^2+H_g^2} + ww\times\frac{(y_c - y_{gt})^2}{W_g^2+H_g^2} \tag{8}$$

$$\begin{cases}\omega_w = hh\times\dfrac{|w - w^{gt}|}{\max(w, w^{gt})}\\ \omega_h = ww\times\dfrac{|h - h^{gt}|}{\max(h, h^{gt})}\end{cases} \tag{9}$$

$$\Omega^{shape} = \sum_{t=w,h}(1-e^{\omega_t})^{\theta}, \theta = 4 \tag{10}$$

$$L_{Shape-IoU} = 1 - IoU + distance^{shape} + 0.5\cdot\Omega^{shape} \tag{11}$$

其中，部分参数解释同上；$scale$ 为尺度因子，与数据集中目标的尺度有关；ww，hh 为水平方向和垂直方向的权重系数，其值与真实框形状有关。

2.3　基于 LAMP 分数的剪枝算法

深度神经网络在各个领域表现出色，但计算力和内存的高需求限制了其应用。为在有限硬件资源中发挥大模型的优势，剪枝算法越来越受到关注。神经网络剪枝通过移除网络中占用大量资源的冗余部分来优化模型。本文采用 LAMP 评分策略进行剪枝。LAMP 是一种新型剪枝方法，核心在于通过计算移除不必要的连接，实现模型压缩和性能提升。在神经网络中，每个连接的权重反映了输入信号的重要性。传统方法根据权重绝对值判断其重要性，绝对值较小的权重被认为对网络贡献有限，因此可以剪枝。然而，全局剪枝方法可能导致某些层的功能丧失。

为解决这一问题，LAMP 采用独特评分机制。首先，将每层的权重张量压平为一维向量，并根据权重大小排序。然后，通过权重幅值平方与同层其他权重幅值平方之和的比值来计算目标连接的重要性。这种方法有效评估并选择需要剪枝的连接，优化网络结构。具体公式如下：

$$S(u;W_t) = \frac{(W_t[u])^2}{\sum_{v\geqslant u}(W_t[v])^2} \tag{12}$$

其中，$W_t[u]$，$W_t[v]$ 分别表示索引 u，v 对应的权重项。

LAMP 评分反映了连接在网络中的重要性，高评分表示重要连接，低评分的部分将被剪枝。为避免网络结构不稳定，每层至少保留一个评分为 1 的通道。LAMP 根据目标修剪比率选择最小评分的连接进行剪枝，实现全局稀疏性。这种设计融合了全局和局部剪枝的优点，能准确衡量相对重要性，实现自适应逐层稀疏。LAMP 的剪枝方式如图 6 所示。

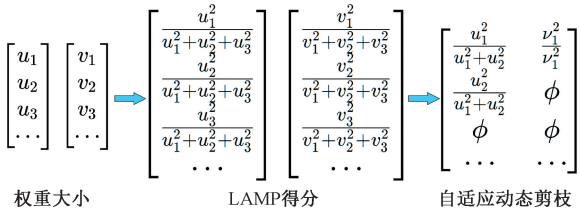


图 6　LAMP 剪枝流程图

Fig. 6　Pruning procedure diagram of LAMP

3　实验数据与分析

3.1　实验环境

模型训练环境为 Ubuntu 22.04 操作系统、AMD Ryzen 5 5600、32 G 内存，和 NVDIA GeForce RTX 4060 Ti GPU，显存为 8 GB。

实验参数设置主要沿用了 YOLOv8 官方推荐参数。输入图像尺寸为 640×640，批量大小(batch size)为 16，总训练周期(epoch)为 500。训练期间使用随机梯度下降(SGD)优化网络参数，学习率设为 0.01，动量设为 0.937，权重衰减为 0.000 5。

3.2　实验数据集

本文的数据集[17]来源于湖北三环锻造有限公司东风重型卡车转向节生产流水线的探伤车间，通过相机拍摄缺陷图片。初始数据集包含 362 张图片，数量不足以满足神经网络训练需求。因此，进行了标注和数据增强，包括旋转、翻转和亮度调整。旋转角度在−20°～20°之间，翻转包

括上下和左右翻转,比例各为 0.5,以提升模型的检测性能和鲁棒性。亮度调整范围为原始亮度的 0.8～1.2 倍,以模拟不同光照条件,增强模型适应性。数据增强后,训练集扩展至 3 206 张图片,包含 11 700 个缺陷目标;验证集包含 1 416 张图片,共 3 792 个缺陷目标,所有目标均为"defect"类别。数据集中部分图片如图 7 所示。其中,矩形框标注为缺陷目标。

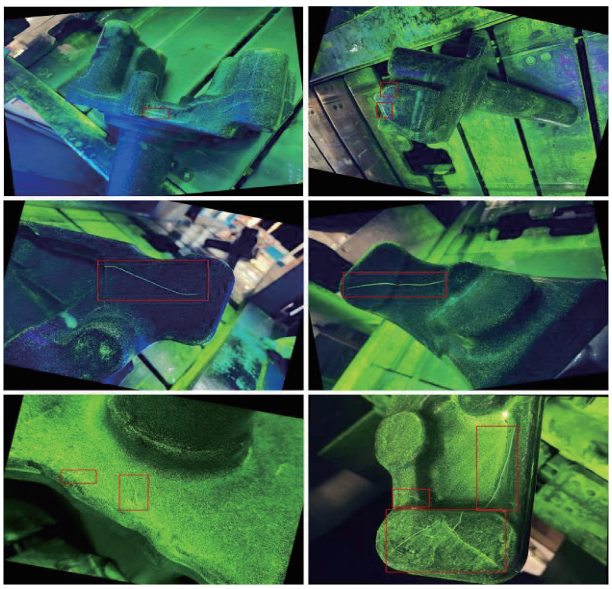


图 7　转向节表面裂纹数据集内部分图片

Fig. 7　Example images from the steering knuckle surface crack dataset

3.3　评价指标

本文采用多种评价指标,包括准确率(precision,P)、召回率(recall,R)、平均精度均值 mAP、参数量(Params)、浮点运算每秒(GFLOPs)和每秒检测帧数(FPS)等。部分指标计算公式如下:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{13}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{14}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP(R)_i \times 100\% \tag{15}$$

其中,TP 表示被正确预测为正例的样本,FP 表示被错误预测为正例的样本,FN 表示被错误预测为负例的样本,而 $P(R)$则是精确率 P 关于召回率 R 的函数。

GFLOPs 用于度量模型或算法的复杂性,它表示模型在进行一次前向推断时所需的浮点运算次数。而 Params 则表示模型的大小,即模型中的可学习参数数量。FPS 用来衡量模型检测速度。一般来说,Params 和 GFLOPs 越小,表示模型所需的计算能力就越小,对硬件的性能要求就越低,更适合在低端设备上进行部署和应用。而较高的 FPS 值则表明模型在实际应用中的处理速度更快,能更好地满足实时性要求。

3.4　剪枝结果及分析

为增强目标检测的识别速度,采用 LAMP 分数的剪枝策略对优化后的网络进行剪枝。通过调整 Speed-up 参数实现对网络不同程度的剪枝。Speed-up(加速比)是指剪枝前模型和剪枝后模型计算量的比值。当 Speed-up 为 1 时,模型处于未剪枝的状态;当 Speed-up 为 2 时,意味着剪枝后的模型运行速度是剪枝前模型的两倍。表 1 展示了剪枝后的实验结果。

表 1　不同加速比对应的剪枝实验结果

Table 1　Pruning experiment results corresponding to different acceleration ratios

Speed-up	P/%	R/%	mAP/%	Parameters/M	GFLOPs	FPS (bs=16)
1.0	86.0	80.7	85.2	2.87	7.4	37.3
1.5	85.6	75.8	82.3	1.37	4.9	40.2
2.0	84.8	77.5	83.0	0.97	3.6	49.3
2.5	84.4	77.8	83.2	0.76	2.9	51.4
3.0	83.3	77.6	83.2	0.62	2.4	54.9
3.5	82.8	76.1	82.4	0.55	2.1	58.5
4.0	81.9	79.0	82.7	0.49	1.8	61.0
4.5	84.6	77.9	83.8	0.45	1.6	65.6
5.0	82.3	76.0	81.8	0.43	1.5	66.2

由表 1 可知,随着 Speed-up 值的增大,检测速度随之提升,同时模型的参数量和计算量减少,但检测精度有所下降。通过调整 Speed-up 值,可以在模型性能和计算效率之间找到一个平衡点,以满足锻件缺陷检测的需求。当 Speed-up 比提高至 4.5 时,检测速度增至 65.6 FPS,满足实时处理需求。此时,模型参数量减至 0.45 M,计算量降至 1.6 GFLOPs,显著优化了资源消耗。尽管 mAP 从 85.2%降至 83.8%,整体性能仍维持在较高水平。因此,当 Speed-up 值为 4.5 时,实现了检测精度、速度和资源消耗之间的平衡。

3.5　消融实验

为验证所提出算法改进的有效性,本文在转向节表面裂纹数据集上进行实验。基准模型(BASE)为 YOLOv8n。改进颈部网络表示使用 VoVGSCSPns 替换颈部网络中的 C2f,使用 GSConvns 替换原颈部网络中的 Conv;Shape-IOU 表示将边界框损失函数替换为 Shape-IOU;LAMP 表示使用 LAMP 剪枝方法对模型进行剪枝。实验结果如表 2 所示。

算法 A 相较于基准算法在精确率、召回率和 mAP 方面分别取得显著提升,分别提升 0.8%,2.9%,和 0.9%。与此同时,参数量从基准算法的 3.01 M 减少至 2.87 M,计算量从基准算法的 8.1 GFLOPs 减少至 7.4 GFLOPs。这一结果充分证明 GSConvns 和 VoVGSCSPns 的有效性,尤

表 2　消融实验结果

Table 2　Ablation experiment results

Algorithms	改进颈部网络	Shape-IOU	LAMP	P/%	R/%	mAP/%	Parameters/M	GFLOPs	FPS(bs=16)
BASE	—	—	—	84.7	78.0	83.8	3.01	8.1	38.7
A	√	—	—	85.5	**80.9**	84.7	2.87	7.4	37.2
B	—	√	—	85.0	80.8	85.1	3.01	8.1	38.8
C	√	√	—	**86.0**	80.7	**85.2**	2.87	7.4	37.3
D(加速比 4.5)	√	√	√	84.6	77.9	83.8	**0.45**	**1.6**	**65.6**

其在提高颈部网络的特征交互能力方面。GSConvns 卷积模块的引入增强了特征之间的交互，使得网络能够更好地捕获锻件表面裂纹的特征，从而提高了检测的准确性。同时，VoVGSCSPns 的应用进一步优化网络结构，减少参数量和计算量，保证性能同时提升效率。

算法 B 相较于基准算法实现了精确率提升 0.3%，召回率提升 2.8%，mAP 提升 1.3%。这表明算法 B 在目标检测性能上取得了一定的改进，尤其是在召回率方面的提升更为显著。证明 Shape-IOU 损失函数更有效地捕获所有目标对象，从而降低了漏检率。

算法 C 验证了改进颈部网络和 Shape-IOU 同时改进的有效性，相较于基准模型 YOLOv8n 准确率提升 1.3%，召回率提升 2.7%，mAP 提升 1.4%。

使用加速比为 4.5 的 LAMP 剪枝方法对模型进行剪枝后，算法 D 在保持与基准模型相当的检测精度的情况下，参数量下降了 85.05%，计算量降低了 80.25%，检测速度提高了 69.5%，达到了 65.6 FPS。

综上所述，算法 D 通过 LAMP 剪枝方法成功实现了轻量化，并在保持检测精度的同时显著提升了检测速度，同时在参数量和计算量上取得了显著降低。这证明了算法 D 在轻量化和实时检测方面的有效性，为其在各种应用场景中的广泛应用提供坚实基础。

3.6　对比实验

为了验证剪枝后网络的有效性，在相同的数据集和实验环境条件下，将本文模型与其他单阶段主流模型进行对比实验，对比实验结果如表 3 所示。

表 3　对比实验结果

Table 3　Comparative experimental results

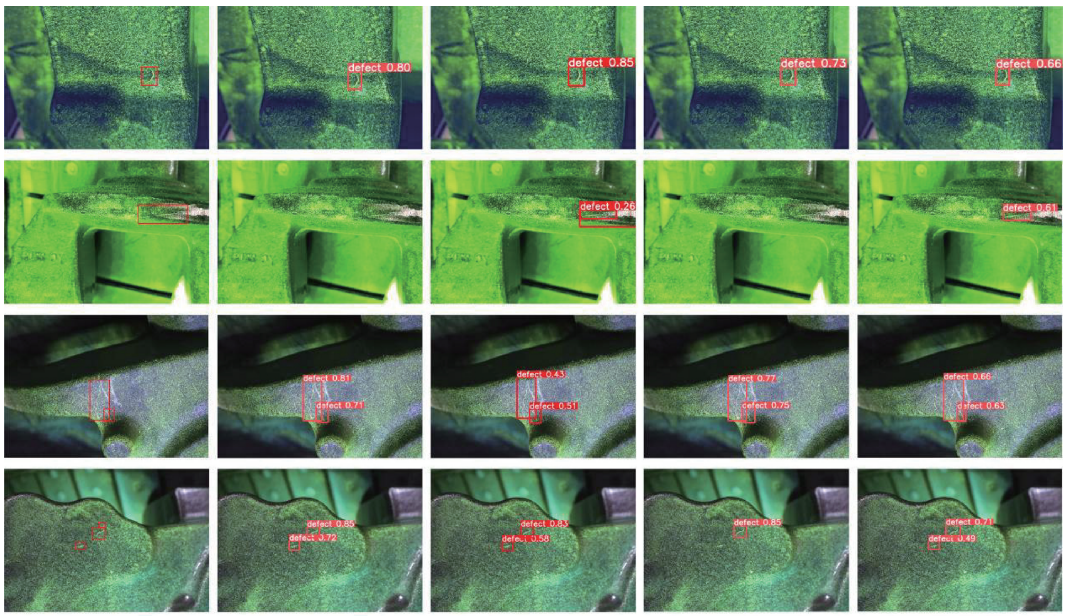
Algorithms	P/%	R/%	mAP/%	Parameters/M	GFLOPs	FPS(bs=16)
YOLOv3-tiny	79.0	72.3	76.6	12.1	18.9	25.2
YOLOv5s	**86.6**	79.4	**85.9**	9.11	23.8	17.1
YOLOv7-tiny	82.8	76.8	83.5	6.01	13.0	24.3
YOLOv8n	84.7	78.0	83.8	3.01	8.1	38.7
Our(未剪枝)	86.0	**80.7**	85.2	2.87	7.4	37.3
Our(加速比 4.5)	84.6	77.9	83.8	**0.45**	**1.6**	**65.6**

本文算法在模型复杂度和检测速度上，取得了明显优势，与原始网络相比参数量下降了 85.05%，计算量降低了 80.25%，检测速度提高了 69.5%。同时检测精度仅次于 YOLOv8n。而 mAP 最高的 YOLOv5s 模型，相较于本文模型，计算量为 14.9 倍，参数量为 20.4 倍，FPS 下降 73.9%。综上所述，本文算法更为适合部署在资源受约束的嵌入式检测设备中。

图 8 展示了不同算法在本文数据集上的检测效果。第一行中，所有模型均能准确地识别小缺陷，证明本文模型并没有因为轻量化而导致漏检现象；在第二行的场景中，仅有本文模型和 YOLOv7-tiny 成功检测到缺陷，且本文模型的锚框置信度最高；第三行中，所有模型均能准确地检测出重叠缺陷，且 YOLOv5s 的锚框置信度最高；第四行中，YOLOv8n 出现了漏检。综上，本文模型仍保持较高精度，满足流水线作业中的实时检测需求。

3.7　NEU-DET 数据集实验

为了进一步验证算法的泛化性，本文使用 NEU-DET 数据集[18]进行实验。该数据集是由东北大学制作的带钢表面缺陷数据集，共 1 800 张灰度图像，按 4∶1 随机划分训练集和验证集，即训练集 1 440 张，验证集 360 张。NEU-DET 数据集的缺陷类别包括六类：裂缝(crazing，Cr)、夹杂物(inclusion，In)、斑块(patches，Pa)、点蚀表面(pitted-surface，Ps)、氧化皮(rolled-in-scale，Rs)和划痕(scratches，Sc)。在不使用 LAMP 剪枝的情况下，分别采用 YOLOv8n 和本文改进的 YOLOv8n 模型进行实验。泛化实验结果如表 4 所示。结果显示，改进后的模型在除 Sc 缺陷之外的各类缺陷上的检测精度均有所提升，*mAP* 从 76.7%提高至 78.7%。



(a) 真实框
(a) Ground truth

(b) YOLOv5s算法
(b) YOLOv5s algorithm

(c) YOLOv7-tiny算法
(c) YOLOv7-tiny algorithm

(d) YOLOv8n算法
(d) YOLOv8n algorithm

(e) 本文算法
(e) Our algorithm

图 8　缺陷检测结果对比图

Fig. 8　Comparative results of defect detection

表 4　泛化实验结果

Table 4　Generalization experiment results　%

Algorithms	mAP	Cr	In	Pa	Ps	Rs	Sc
YOLOv8n	76.7	44.8	79.2	87.5	88.8	64.6	**95.4**
本文(未剪枝)	**78.7**	**53.0**	**81.3**	**88.1**	**89.1**	**66.1**	94.4

4　结　　论

针对传统人工方法在重卡转向节表面裂纹磁粉检测场景中存在的效率问题与安全隐患,本文提出了一种改进 YOLOv8 的轻量级缺陷检测算法。通过使用 VoVGSCSPns 和 GSConvns 模块改进 YOLOv8 的颈部网络,实现模型轻量化的同时,保证了检测准确性;引入 Shape-IoU,提高了模型对小目标的关注度,从而提高了回归的准确性;同时引入了基于 LAMP 分数的剪枝算法,在实验中选用了 Speed-up 为 4.5 的剪枝效果,结果显示与原始网络相比,参数量和计算量大幅度下降,有效提高了网络的识别速度并保证了模型的检测精度,满足工业实时检测需求。

研究实验证明,本文模型在参数量和计算量方面更为轻量,同时在检测准确度上并没有大幅度下降,能够达到实时处理的标准。与其他现存单阶段模型相比,该模型不仅在检测精度上占有优势,还显著减轻了对计算和存储资源的依赖,使其更易于在资源有限的设备上实施。同时,采用 YOLOv8n 和未剪枝的 YOLOv8n 改进模型在 NEU-DET 数据集上进行实验,验证了其泛化能力。目前,本文研究仅停留在算法研究阶段,未来的研究方向将集中在模型部署到实际应用场景中,同时进一步完善数据集,以增强模型检测精度。

参考文献

[1]　陶显,侯伟,徐德. 基于深度学习的表面缺陷检测方法综述[J]. 自动化学报,2021,47(5):1017-1034.
TAO X, HOU W, XU D. A survey of surface defect detection methods based on deep learning[J]. Acta Automatica Sinica, 2021,47(5):1017-1034.

[2]　赵朗月,吴一全. 基于机器视觉的表面缺陷检测方法研究进展[J]. 仪器仪表学报,2022,43(1):198-219.
ZHAO L Y, WU Y Q. Research progress of surface defect detection methods based on machine vision[J]. Chinese Journal of Scientific Instrument, 2022,43(1): 198-219.

[3]　GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

[4]　GIRSHICK R. Fast R-CNN[C]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[5]　REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the 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]. Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[7] 马燕婷,赵红东,阎超,等.改进 YOLOv5 网络的带钢表面缺陷检测方法[J].电子测量与仪器学报,2022,36(8):150-157.
MA Y T, ZHAO H D, YAN CH, et al. Strip steel surface defect detection method by improved YOLOv5 network[J]. Journal of Electronic Measurement and Instrumentation, 2022,36(8):150-157.

[8] 徐国伟,林辉,修春波,等.基于改进 YOLOv5 的动车组关键部件缺陷检测[J].光电子·激光,2023,34(7):752-761.
XU G W, LIN H, XIU CH B, et al. Defect detection of key components of electric multiple units based on improved YOLOv5[J]. Journal of Optoelectronics·Laser, 2023,34(7):752-761.

[9] HOWARD A, SANDLER M, CHU G, et al. Searching for mobilenetv3[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1314-1324.

[10] 熊聪,于安宁,高兴华,等.基于改进 YOLOX 的钢材表面缺陷检测算法[J].电子测量技术,2023,46(9):151-157.
XIONG C, YU A N, GAO X H, et al. Steel surface defect detection algorithm based on improved YOLOX[J]. Electronic Measurement Technology, 2023, 46(9): 151-157.

[11] TAN M, PANG R, LE Q V. Efficientdet: Scalable and efficient object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10781-10790.

[12] ZHANG H, ZHANG S. Shape-IoU: More accurate metric considering bounding box shape and scale[J]. ArXiv preprint arXiv:2312.17663, 2023.

[13] LEE J, PARK S, MO S, et al. Layer-adaptive sparsity for the magnitude-based pruning[J]. ArXiv preprint arXiv:2010.07611, 2020.

[14] HAN K, WANG Y, TIAN Q, et al. Ghostnet: More features from cheap operations[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1580-1589.

[15] 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.

[16] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 6848-6856.

[17] 张岳,张上,王恒涛,等.基于跨尺度特征提取的锻件表面裂纹检测算法[J/OL].计算机集成制造系统:1-24. https://doi.org/10.13196/j.cims.2023.069.
ZHANG Y, ZHANG SH, WANG H T, et al. Crack detection algorithm for forging surface based on cross-scale feature extraction[J/OL]. Computer Integrated Manufacturing Systems: 1-24. https://doi.org/10.13196/j.cims.2023.069.

[18] 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, 2019, 69(4): 1493-1504.

作者简介

张上,博士,副教授,主要研究方向为物联网、计算机应用、图像处理。

E-mail:zhangshang@ctgu.edu.cn

许欢,硕士研究生,主要研究方向为目标检测、嵌入式技术。

E-mail:xu_huan@ctgu.edu.cn

张岳(通信作者),硕士研究生,主要研究方向为图像识别、智能制造。

E-mail:zhangyue980202@ctgu.edu.cn