DOI:10.19651/j.cnki.emt.2417674

用于缺陷检测的 YOLOv8 轻量化设计方法*

艾 峰 邓耀华
(广东工业大学机电工程学院 广州 510006)

摘 要:在大规模制造的端侧产线工业质检应用中,由于算力、成本和功耗等因素的限制,将深度学习模型裁剪并部署到小型算力的边缘设备上变得尤为重要。针对铝型材复杂缺陷检测这一应用场景,基于 YOLOv8 设计了缺陷检测模型。首先,通过轻量化结构设计,结合局部自注意力机制提升细微缺陷提取能力;采用空间通道下采样替代传统下采样卷积;并提出结合混合局部通道注意力机制的 C2f-M 模块。然后,基于双向特征金字塔网络设计了 SC-BiFPN 颈部网络,增强了多尺度特征融合能力。接着,设计任务动态对齐的特征检测头 TDADH,充分利用多层次特征,实现更精准的目标定位与分类;采用 MPDIoU 损失函数增强边界框回归能力。最后,通过 Taylor 方法对 YOLOv8 进行裁剪,显著减少模型参数量和计算成本。实验结果表明,轻量化 YOLOv8 模型在铝材表面缺陷数据集上的参数量降低至原模型的 36.7%,计算量减少 40%,模型体积缩小 62%;同时,检测精确度、召回率及 mAP@50-95 分别提升 0.3%、1.1%、4.8%。该方法有效解决了端侧部署中的计算复杂度与检测性能平衡问题,为小型算力硬件上的高效缺陷检测提供了可行方案。

关键词:缺陷检测;双向特征金字塔;损失函数;任务动态对齐;剪枝
中图分类号:TP391.4;TN40 **文献标识码**:A **国家标准学科分类代码**:520.6040

YOLOv8 lightweight design approach for defect detection

Ai Feng Deng Yaohua
(College of Mechanical and Electrical Engineering, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: In end-side production line industrial quality inspection applications for large-scale manufacturing, tailoring and deploying deep learning models to edge devices with small arithmetic power becomes particularly important due to the limitations of arithmetic power, cost and power consumption. Based on the application scenario of complex defect detection of aluminum profile, the defect detection model is designed based on YOLOv8. First of all, through the lightweight structure design, combined with the partial self-attention mechanism to improve the ability of subtle defect extraction; the use of spatial channel downsample instead of the traditional downsampling convolution; and proposed a combination of mixed local channel attention mechanism of the C2f-M module. Then, SC-BiFPN neck network is designed based on bidirectional feature pyramid network, which enhances the multi-scale feature fusion capability. Then, the task dynamic align detection head is designed to make full use of multilevel features for more accurate target localisation and classification; the MPDIoU loss function is used to enhance the bounding box regression capability. Finally, YOLOv8 is trimmed by Taylor's method to significantly reduce the number of model parameters and computational cost. The experimental results show that the lightweight YOLOv8 model reduces the number of parameters to 36.7% of the original model on the aluminium surface defects dataset, reduces the computational effort by 40%, and reduces the model volume by 62%; at the same time, the detection accuracy, the recall rate and the mAP@50-95 are improved by 0.3%, 1.1%, and 4.8%, respectively. The method effectively solves the problem of balancing computational complexity and detection performance in end-side deployment, and provides a feasible solution for efficient defect detection on small arithmetic hardware.

Keywords: defect detection; bifpn; loss function; task dynamic alignment; pruning

0 引 言

近年来,随着中国制造业的快速发展和智能化升级,工业生产领域对质量控制的要求也日益提高。工业制造过程复杂,细微的表面划痕、裂纹或尺寸偏差等缺陷不仅会影响产品外观,还可能降低其机械性能,从而对产品质量和使用

收稿日期:2024-12-19
*基金项目:广东省省级科技计划项目(2023A0505050151)资助

寿命造成威胁。因此工业缺陷检测至关重要[1]。

传统的人工目视检测方法效率低下且精度有限，基于机器视觉的检测方法尽管能够实现一定程度的自动化，但仍需依赖手动提取特征，操作复杂且检测速度较慢。而基于深度学习的缺陷检测方法通过神经网络自动学习图像特征，不仅能够提取更复杂的特征，还能显著提升检测的精度与鲁棒性，因此逐渐应用于工业质检领域。李澄非等[2]在特征提取网络中引入挤压与激励注意力机制(squeeze and excitation，SE)，有效增强了 YOLOv4 对铝材表面缺陷特征的提取能力，然而该方法检测帧率仅为 27 fps，实时检测性能不足；邓刚等[3]将 K-Means＋＋算法应用到自适应锚框算法中改进 YOLOv5，提高了铝型材表面缺陷检测的平均精度，然而该方法在 GPU 上单张图片推理速度为 17.9 ms，部署至资源受限的端侧设备则会更慢。尽管许多改进方法在检测精度上表现出色，但其庞大的模型规模、众多的参数数量以及高计算需求严重限制了推理速度，从而降低了检测效率，因此深度学习神经网络模型的轻量化设计显得尤为重要。

深度学习神经网络模型的轻量化设计主要包括模块高效轻量化和模型剪枝两种技术。吕秀丽等[4]研究轻量级的多尺度混合卷积模块(multi-scale mixed convolution，MSMC)并将其嵌入 YOLOv8 模型，成功应用于 PCB 表面缺陷检测；胡惠娟等[5]通过在 YOLOv8 主干网络中引入轻量级的上下文引导模块 Context Guided，使其能在资源受限的环境中运行。赵佰亭等[6]设计了一种轻量化特征增强模块 RLKM，有效提高 YOLOv5s 模型的检测速度。张上等[7]则提出轻量化卷积模块 GSConvns，使用 LAMP 剪枝策略去除不重要的权重参数，改进后的 YOLOv8 参数量和计算量均减少 80％以上；赵亚凤等[8]将 YOLOv7 的主干网络替换成轻量化 FasterNet，同时使用通道剪枝技术对改进模型进行剪枝，实时检测速度 FPS 提高了近 60；徐尽达等[9]设计了一种轻量化的颈部网络并使用 LAMP 剪枝技术，模型大小降至基线的 22.22％，实现了轻量化小目标检测。

尽管上述方法能够有效降低模型的计算复杂度和模型大小，但模型在轻量化后检测精度会受较大影响，在平衡低计算量与高准确性方面的研究仍有欠缺。本文以铝型材复杂缺陷检测为应用场景，提出一种用于缺陷检测的 YOLOv8 轻量化设计方法。研究同时降低深度学习模型计算量、参数量以及提高模型检测性能的轻量化设计方法。本文的主要研究包括：

1)为了在增强模型对多尺度缺陷的检测能力的同时降低模型参数量，提出了结合混合局部通道注意力机制(mixed local channel attetion，MLCA)的 C2f-M(c2f module based on MLCA attention)模块，并将主干网络中的下采样卷积替换为轻量化的空间通道下采样(spatial channel downsample，SCDown)模块。然后，在主干网络最后一层引入局部自注意力机制(partial self-attention，PSA)，兼顾多尺度缺陷检测。

2)为进一步充分融合不同尺度缺陷特征信息，设计了 SC-BIFPN(bifpn with SCDown)颈部网络。首先，借鉴双向特征金字塔网络(bidirectional feature pyramid network，BiFPN)的思想改进颈部网络；然后，新增一条浅层特征融合路径；最后，将颈部网络中的下采样卷积替换为轻量化的 SCDown 模块。

3)为了提高缺陷信息预测的准确性，设计一种任务动态对齐的特征检测头(task dynamic align detection head，TDADH)，实现分类与定位任务的动态对齐与交互。

4)为提升网络模型训练时的边界框回归性能，加速损失函数收敛，采用最小点距离损失函数(minimum point distance-iou，MPDIoU)作为模型训练时的损失函数。

5)为进一步压缩模型大小，降低计算复杂度，采用 Taylor 展开的思想研究 YOLOv8 模型的剪枝，以便部署至边缘计算平台。

1　缺陷检测 YOLOv8 模型的轻量化结构设计思路及实现

本文提出的改进 YOLOv8 模型结构如图 1 所示。首先，引入 C2f-M 模块与 SCDown 下采样模块，用于降低模型的参数量和计算复杂度。然后，在主干网络最后一层引入 PSA 注意力机制，强化模型对多尺度缺陷的捕获能力。其次，借鉴 BIFPN 的思想，设计了 SC-BIFPN 颈部网络，确保多尺度特征信息充分融合，同时在颈部网络也增加一条融合 P2 和 P3 特征层信息的特征融合路径，增强模型对细微缺陷的检测能力。最后，设计了一种任务动态对齐的特征检测头，实现分类与定位任务的动态对齐与交互。

1.1　C2f-M 模块

C2f 模块因其良好的缺陷特征提取性能被频繁使用，为提升其局部与全局特征提取能力的同时不增加模型的参数量和计算量，本文对 C2f 模块进行结构改进，提出 C2f-M 模块，如图 2 所示。首先，采用轻量化的 SCDown 模块对输入特征进行下采样，然后，将 C2f 中的 Bottleneck 块替换为 MLCABlock。能在降低参数量和计算量的前提下有效提升模型对于局部和全部特征的提取能力。其中 SCDown 是文献[10]提出的新型轻量化下采样模块，将空间与通道的下采样操作进行解耦，其结构如图 3(b)所示，对于输入的特征，首先使用 1×1 点卷积压缩通道数，减少后续计算量，然后采用卷积核为 3、步长为 2 的深度可分离卷积(depthwise separable convolution，DSConv)[11]分别对每个通道进行进行空间下采样，在轻量化的同时最大限度维持了特征信息。

MLCABlock 则受文献[12]提出的 MLCA 注意力启发，将混合局部通道注意力融入 Bottleneck，如图 4 所示，MLCA 注意力首先采用局部特征池化(local average

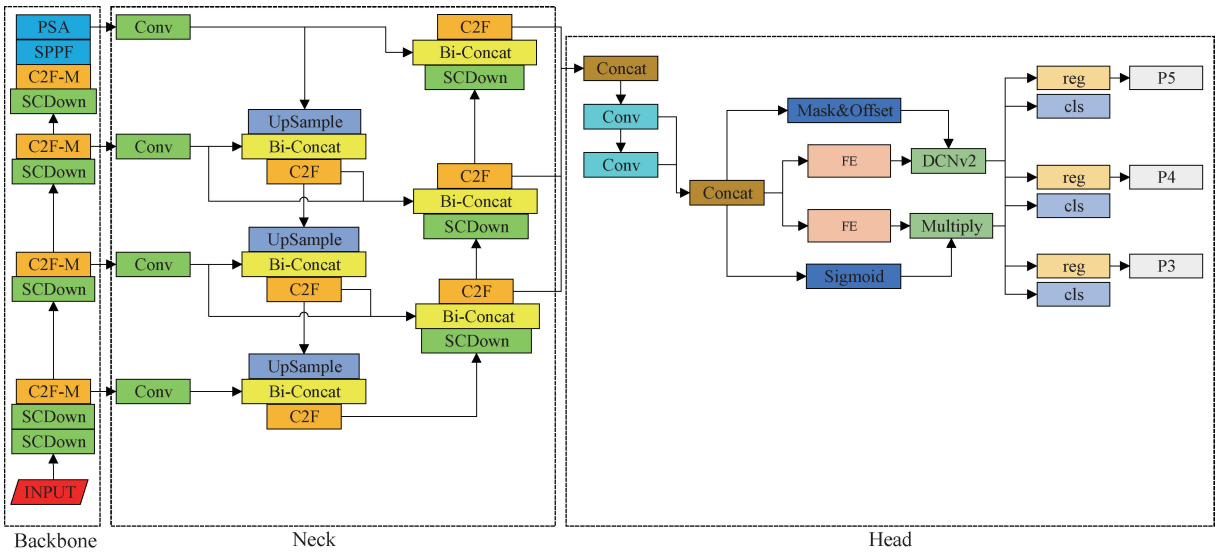


图 1　改进 YOLOv8 模型结构

Fig. 1　Improvement of YOLOv8 model structure

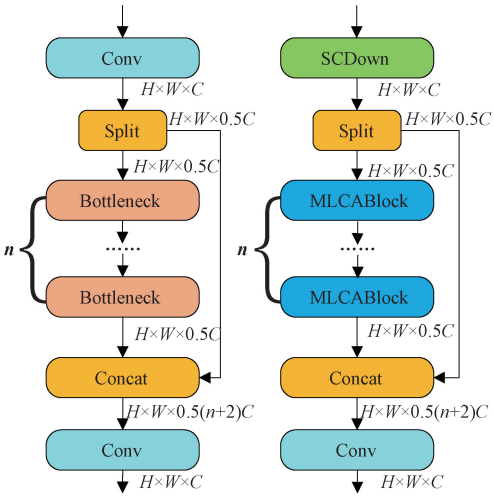


图 2　改进的 C2f 模块结构

Fig. 2　Improved C2f module structure

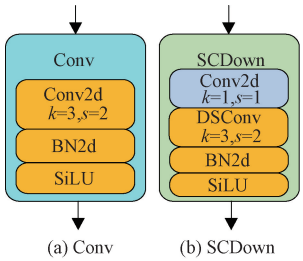


图 3　改进的下采样模块结构

Fig. 3　Improved downsampling module structure

pooling,LAP)与全局特征池化(global average poolong,GAP)提取输入特征图的局部特征与全局特征,然后,分别对局部和全局权重进行反平均池化(unaverage pooling),得到归一化的注意力权重并通过加权融合得到最终的注意力权重,最后,将最终注意力权重与原始特征图逐通道相乘,完成特征增强。

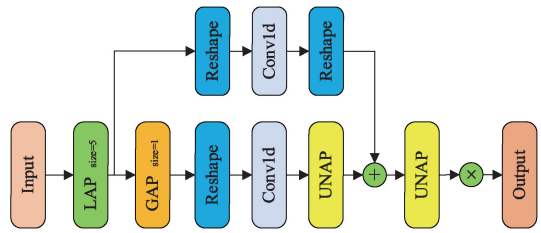


图 4　MLCA 网络结构

Fig. 4　Structure of MLCA network

1.2　PSA 注意力网络

铝型材的缺陷通常以局部异常或纹理差异的形式出现,而这些特征在复杂背景下容易被掩盖,从而导致漏检或误检的发生。为了更有效地检测这些缺陷,本文在 YOLOv8 的主干网络中引入了一种融合自注意力机制的局部自注意模块(PSA)[10],该模块能够增强模型对局部特征的捕捉能力,并在特征表达上实现显著提升。如图 5 所示,首先,PSA 模块对输入特征通道进行均分,一部分通过多头自注意机制(multi-head self-attention,MHSA)进行处理,有效捕捉局部特征中的细微差异,然后,通过前馈神经网络(feed-forward network,FFN)对局部特征进行非线性映射,使得模型能够更清晰地分离缺陷与背景之间的特征差异,同时也能增强模型地泛化能力,最后,将另一部分未经自注意力处理地全局特征与局部特征进行拼接融合,实现在强化局部特征的同时保留全局上下文信息,从而进一步提升缺陷检测的准确性。此外,为减少计算开销,将 PSA 注意力放置在特征分辨率最低的第 4 阶段,实现计算效率与检测性能的平衡。

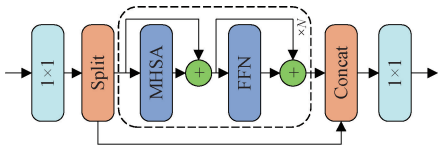


图 5　PSA 注意力网络结构

Fig. 5　Structure of PSA attention network

1.3　SC-BiFPN 颈部网络

为了优化层次信息的融合，增强细微缺陷的特征检测能力，设计了 SC-BiFPN 颈部网络。首先，受文献[13]的启发，采用 BiFPN 带权特征融合网络替换原颈部网络，高效融合深浅层网络信息；然后，新增一条 P2 与 P3 融合路径，能够获取更浅层的细微缺陷信息；最后，将下采样的卷积替换成 SCDown 实现轻量化下采样。

如图 6(a)所示，BiFPN 包含自顶向下和自底向上两条路径，确保不同层次的特征信息得到充分利用。该结构剔除了对整体特征网络贡献较小的节点，并在同一层级的原始输入和输出节点间添加额外的特征信息传送路径，能在不增加额外成本的情况下融合更多有用信息。此外，在 BiFPN 中，采用带权的特征融合算法学习不同特征的重要性。以 BiFPN 第 i 层为例，第 i 层的融合特征公式如下：

$$\begin{cases} P_i^{td} = Conv\left(\dfrac{w_1 \cdot P_i^{in} + w_2 \cdot Resize(P_{i+1}^{in})}{w_1 + w_2 + \varepsilon}\right) \\ P_i^{out} = Conv\left(\dfrac{w'_1 \cdot P_i^{in} + w'_2 \cdot P_i^{td} + w'_3 \cdot Resize(P_{i-1}^{in})}{w'_1 + w'_2 + w'_3 + \varepsilon}\right) \end{cases} \tag{1}$$

式中：P_i^{in}、P_i^{td}、P_i^{out} 分别表示第 i 层的输入特征、中间特征和输出特征，$Resize$ 为与分辨率匹配的上采样或下采样操作 w_i 为输入特征对应的学习权重，其中每个 w_i 后使用 RELU 激活函数来确保 $w_i \geqslant 0$。同时设置 $\varepsilon = 0.000\ 1$，以避免数值不稳定。

为了进一步提升细微缺陷的检测性能，如图 6(b)所示，在颈部网络的 P2 层新增了一条特征融合路径，与 P3 层进行信息交互，使模型能够更充分地利用浅层特征中的小目标信息，增强对小目标的检测能力。

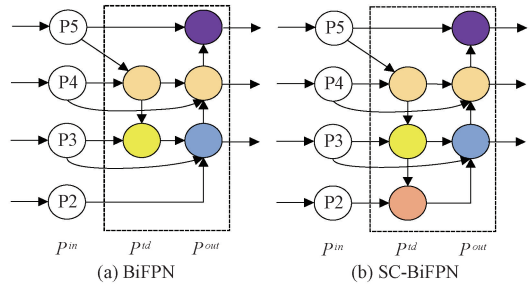
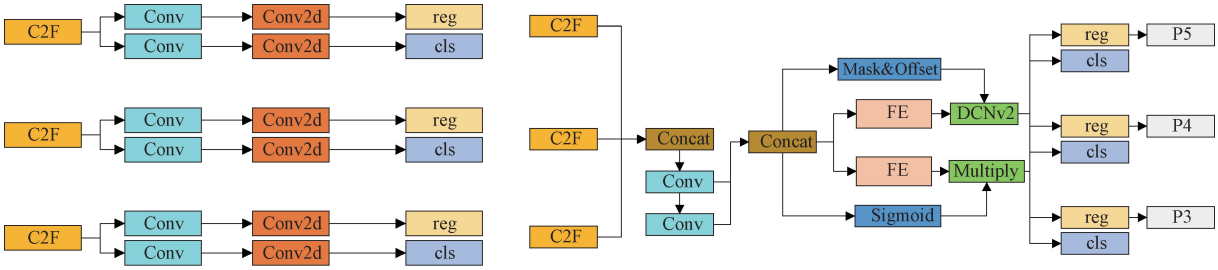


图 6　特征融合模块图

Fig. 6　Feature fusion module diagram

1.4　任务动态对齐的特征检测头 TDADH

YOLOv8 的检测头网络由 3 个独立尺度的检测路径组成，如图 7(a)所示。然而，在缺陷检测任务中，相互独立的检测路径容易忽视不同尺度特征之间的关联信息，从而导致定位不准确和分类效果不佳。此外，由于每个检测头都需要执行多次卷积操作，其计算量复杂度较高，显著增加了参数量，进而拖慢实时检测速度。为解决上述问题，本文设计了一个任务动态对齐的特征检测头 TDADH。



(a) YOLOv8检测头
(a) YOLOv8 detection head structure

(b) TDADH结构
(b) TDADH structure

图 7　检测头部网络对比

Fig. 7　Comparison of detection header networks

TDADH 结构如图 7(b)所示，首先，输入的 3 个不同尺度的特征经过拼接(concatenate，Concat)操作进行特征融合，并通过共享卷积层进行特征传递，大幅减少模型的参数数量；随后，融合后的特征进入特征提取器(feature extraction，FE)模块，如图 8 所示，该模块通过 N 个带激活函数的连续转换层提取不同尺度特征之间的交互特征，并通过学习权重 w 来获得联合特征，从而有效增强分类与定位任务的交互。

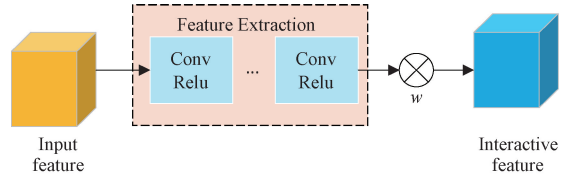


图 8　FE 结构

Fig. 8　FE structure

TDADH 在分类和定位部分,通过两条路径进行处理,首先,联合特征通过 Mask&Offset 模块生成用于定位的权重掩码和偏移量,输入到可变形卷积(deformable ConvNets v2,DCNv2)中,DCNv2 引入了调制机制,使其能够根据特征的空间偏移量和权重进行自适应调整,从而精确定位缺陷[14]。在另一条路径上,交互特征通过 Sigmoid 函数进行动态特征选择,用于缺陷分类。最后,模型通过 Scale 层将特征缩放至 P3、P4、P5 检测头,以适应不同尺度的检测需求。TDADH 有效减少了检测头的参数计算量,同时增强了不同尺度下缺陷定位和分类任务之间的信息交互,通过引入可变形卷积 DCNv2 和特征动态选择机制,既实现了网络轻量化,又提升了检测精度和鲁棒性。

1.5　优化损失函数

YOLOv8 中,采用完全交并比(complete-iou,CIoU)作为损失函数[15],CIoU 相较于 IoU(intersection over union),它考虑了检测框的中心点距离和宽高的差异,使得 CIoU 在训练时能更准确地衡量两个框之间的重叠程度,提高缺陷检测的准确性和鲁棒性。CIoU 损失函数定义如下:

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

$$\alpha=\frac{V}{1-IoU+V} \tag{3}$$

$$L_{CIoU}=1-IoU+\frac{\rho^2(B_{gt},B_{prd})}{C^2}+\alpha V \tag{4}$$

其中,w^{gt}、h^{gt} 为实际框宽高,w^{prd}、h^{prd} 为预测框宽高,IoU 表示预测框和实际框的重复程度,ρ 表示预测框中心点和实际框中心点之间的欧氏距离。

然而,当预测框与边界框具有相同的高宽比,即 $V=0$ 时,CIoU 损失函数无法进行优化,因此引入最小点距离 MPDIoU 损失函数[16],其计算示意图如图 9 所示。

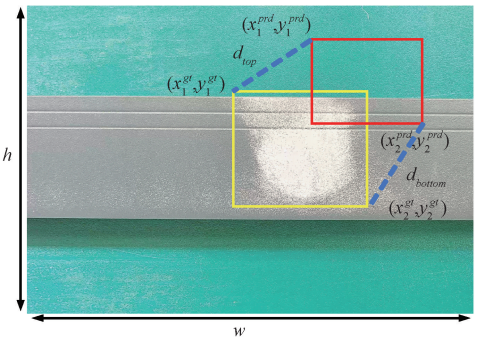


图 9　MPDIoU 计算示意图

Fig. 9　Schematic diagram of MPDIoU calculation

通过最小化预测边界框和真实边界框之间的左上和右下点距离,充分挖掘了水平矩形的几何特征,有效提高了边界框回归性能。

$$d_{top}^2=(x_1^{gt}-x_1^{prd})^2+(y_1^{gt}-y_1^{prd})^2 \tag{5}$$

$$d_{bottom}^2=(x_2^{gt}-x_2^{prd})^2+(y_2^{gt}-y_2^{prd})^2 \tag{6}$$

$$L_{MPDIoU}=1-IoU+\frac{d_{top}^2}{w^2+h^2}+\frac{d_{bottom}^2}{w^2+h^2} \tag{7}$$

其中,d_{top}、d_{bottom} 表示实际框和预测框左上点和右下点的欧氏距离,w 和 h 表示输入图片的宽高。

2　缺陷检测 YOLOv8 模型的 Taylor 剪枝方法

2.1　卷积核贡献度评估准则

在剪枝前有一个重要的环节,对卷积核进行贡献度评估,即计算在去除特定卷积核后目标损失函数的变化,由于目标损失函数的原始形式通常较为复杂,直接计算其变化量难度较,因此使用一阶 Taylor 展开式去近似目标损失函数[17],从而有效地简化了卷积核贡献度计算过程。根据这一方法,剪枝通过计算卷积核对目标函数的近似影响,优先去除贡献度较低的卷积核,其剪枝原理如图 10 所示。

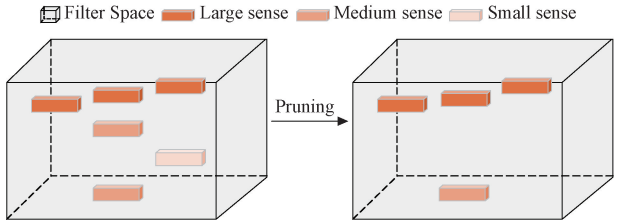


图 10　Taylor 剪枝原理

Fig. 10　Taylor pruning principle

贡献度评估可以被看作一个优化问题,即如何使目标损失函数变化最小:

$$|\Delta L(w_i)|=|L(D,w_i=0)-L(D,w_i)| \tag{8}$$

式中:w_i 表示神经网络参数 i 的权重,D 代表网络的输入输出,$L(D,w_i=0)$ 和 $L(D,w_i)$ 分别表示卷积核 h_i 被剪枝前后的损失值。为了近似目标损失函数 $|\Delta L(w_i)|$,使用一阶 Taylor 多项式逼近,其原型为:

$$f(x)=\sum_{p=0}^{P}\frac{f^{(p)}}{p!}(x-a)^p+R_p(x) \tag{9}$$

其中,$f^{(p)}$ 是 f 在点 a 出的 p 阶导数,$R_p(x)$ 是 p 阶余项。因此在 $h_i=0$ 附近逼近 $L(D,w_i=0)$ 的一阶 Taylor 多项式为:

$$L(D,w_i=0)=L(D,w_i)-\frac{\delta L}{\delta h_i}w_i+R_1(w_i=0) \tag{10}$$

余项 $R_1(h_i=0)$ 可以通过拉格朗日形式计算:

$$R_1(w_i=0)=\frac{\delta^2 L}{\delta(w_i^2=\xi)}\cdot\frac{w_i^2}{2} \tag{11}$$

其中,ξ 是 0～w_i 之间的实数,将式(15)代入目标函数式(13)并忽略余项以简便计算,则可以用一阶梯度近似每个卷积核对损失函数的影响:

$$|\Delta L(w_i)|=\left|\frac{\delta L}{\delta w_i}w_i\right| \tag{12}$$

如果一个卷积核对特征图的梯度是平坦的,那么该卷

积核对损失函数的影响很小，可以去除。通过计算对特征图的一阶梯度与特征图本身的乘积并累积，可得贡献度评估准则：

$$\Theta_T(f_j^{(k)}) = \frac{1}{M} \cdot \left| \sum_m \frac{\delta L}{\delta f_{j,m}^{(k)}} \cdot f_{j,m}^{(k)} \right| \tag{13}$$

其中，$f_l^{(k)}$ 表示第 j 层的第 k 个特征图，$\frac{\delta L}{\delta f_{j,m}^{(k)}}$ 表示损失函数 L 对于特征图上位置 m 处的梯度，$\delta f_{j,m}^{(k)}$ 表示特征图位置 m 处的值，M 表示特征图的长度。通过计算累积值，可以评估每个特征图对损失函数的影响，从而确定是否要裁剪该特征图。

2.2　正则化 FLOPs 剪枝

考虑到剪枝过程中不同层的特征图和卷积核的大小不一，计算需求也不同，因此引入计算量（floating point operations，FLOPs）正则项对计算量正则化，优化模型的计算效率。

$$\Theta(f_j^{(k)}) = \Theta(f_j^{(k)}) - \lambda \Theta_j^{FLOPs} \tag{14}$$

其中，λ 用于控制正则化程度。对卷积核与全连接层的 $FLOPs$ 计算公式如下：

$$FLOPs = 2HW(C_{in}K^2 + 1)C_{out} \tag{15}$$

$$FLOPs = (2I - 1)O \tag{16}$$

其中，H、W 和 C_{in} 分别表示输入特征图的宽高和通道数，K 为卷积核宽度，C_{in} 表示输出通道数；I 和 O 表示全连接层的输入和输出维度。

通过使用基于 Taylor 展开的贡献度评估准则，进行全局迭代裁剪贡献最小的特征图，达到剪枝比例后对模型进行微调（fine-tune），使检测精度回升。剪枝比例公式如下：

$$Speed\_up = \frac{FLOPs_{ori}}{FLOPs_{prune}} \tag{17}$$

式中：采用剪枝前后计算量的比值作为剪枝比例的值。

3　实验设计

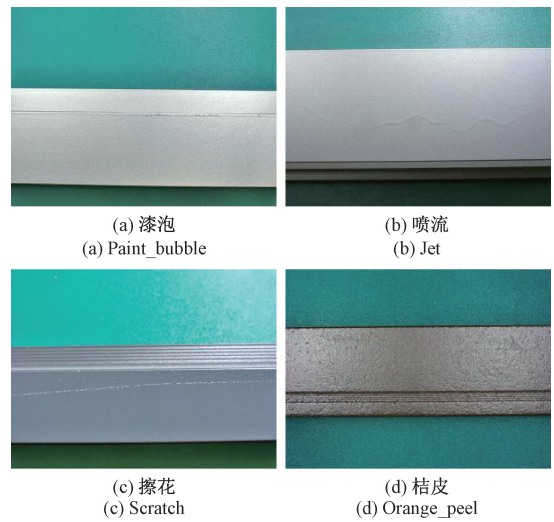
3.1　数据集构建

实验使用的数据集为阿里云天池平台从南海铝型材标杆企业采集的实际生产中有瑕疵的铝型材监测影像，分辨率为 2 560×1 920，包含漆泡（Paint_bubble）、擦花（Scratch）、桔皮（Orange_peel）、喷流（Jet）等 10 种缺陷，共 3005 张图片，为满足实验要求，以 6∶2∶2比例划分为训练集、验证集和测试集。各类缺陷图像数量分布如表 1 所示，部分铝材缺陷图例如图 11 所示。

表 1　铝材缺陷数量分布表

Table 1　Distribution of the number of aluminum defects

类别	数量/张
擦花	128
杂色	365
漏底	538
不导电	390
桔皮	173
喷流	86
漆泡	82
起坑	407
脏点	261
角位漏底	346
多瑕疵	229



(a) 漆泡　(a) Paint_bubble　(b) 喷流　(b) Jet

(c) 擦花　(c) Scratch　(d) 桔皮　(d) Orange_peel

图 11　铝材缺陷样本

Fig. 11　Aluminum Defect Sample

3.2　实验环境配置

本实验本实验基于 PyTorch 深度学习框架，操作系统是 Ubuntu20.04，CPU 为 Intel(R) Xeon(R) Silver 4214R，GPU 为 RTX 3080Ti（12 GB），开发语言为 Python3.8，CUDA 版本为 11.6。参数设置如表 2 所示。

表 2　参数设置

Table 2　Parameter settings

参数	配置
Batchsize	16
Imgsz	640×640
Epochs	300
Workers	8
Optimizer	SGD
lr0/lrf	0.01
Momentum	0.937
Weight_decay	0.000 5
Speed_up	1.3
Fine-tune	300

3.3　模型评价标准

为了验证模型对铝材缺陷的检测性能，本研究采用精确度（precision，P）、召回率（recall，R）、平均精度（average

precision,AP)、均值平均精度(mean average precision,mAP)、参数量(parameters,Param)、计算量 FLOPs 和模型体积(Weight)对模型进行评价。精确度的含义是在全部预测为正样本的结果中,预测正确的正样本所占的比例,即误检率,召回率的含义是预测正确的正样本占所有正样本的比例,即漏检率。平均精度是模型在所有置信度阈值下的值,基于精确度和召回率进行计算。均值平均精度 mAP@50-95 表示 IoU 在 0.5～0.95 上的 mAP。参数量、计算量和模型体积用于衡量模型大小,值越低模型越轻量。本实验使用其中 P、R、mAP、FPS 的计算公式如下:

$$P = \frac{T_P}{T_P + F_P} \tag{18}$$

$$R = \frac{T_P}{T_P + F_N} \tag{19}$$

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{20}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP \tag{21}$$

式中:T 代表实际为正样本,F 代表实际为负样本,下标 P 表示预测为正样本,下标 N 代表预测为负样本;故 T_P 代表正确判断为正样本的数量,F_P 代表错误判断为正样本的数量;F_N 代表错误判断为负样本的数量;n 代表缺陷总类别数。

4　实验结果与分析

4.1　消融实验

为了验证本研究中改进模块与剪枝算法对模型性能提升以及轻量化的贡献,以原始 YOLOv8n 为基线模型,并以精确度、召回率、mAP@50-95、参数量、计算量和模型体积作为评价指标,在保持相同实验条件基础上开展消融实验,实验结果如表 3 所示。

表 3　模型消融实验

Table 3　Model ablation experiments

PSA	C2f-M	SC-BiFPN	TDADH	MPDIoU	Prune	P/%	R/%	mAP@50-95/%	Param/10^6	FLOPs/G	Weights/MB
×	×	×	×	×	×	78.5	65.9	48.9	3.0	8.1	6.0
√	×	×	×	×	×	82.7	66.2	49.3	3.3	8.3	6.5
×	√	×	×	×	×	85.4	64.0	49.3	3.0	8.1	6.0
×	×	√	×	×	×	75.2	67.0	52.0	1.5	6.8	3.2
×	×	×	√	×	×	71.6	70.7	54.6	2.2	8.6	4.5
×	×	×	×	√	×	79.5	68.6	50.0	3.0	8.1	6.0
√	√	√	√	√	×	80.8	66.7	54.0	1.5	6.4	3.1
√	√	√	√	√	√	78.8	67.0	53.8	1.1	4.9	2.3

从表 3 可以看出,单独引入 PSA 注意力和 MLCA 注意力对于检测的准确率均有提升,分别提升了 4.2% 和 6.9%,其中 PSA 注意力使得模型体积增大 0.5 MB;SC-BiFPN 模块由于去除了网络中冗余的节点,参数量和计算量分别降低了 1.5 M 和 1.3 G;TDADH 模块采用共享卷积有效降低了 0.8 M 的参数量,mAP@50-95 也显著提升了 5.7%,但由于其特征提取器 FE 模块包含 N 次转换操作,单独加入时计算量有些许提高;MPDIoU 得益于其出色的边界框回归能力,召回率提升了 2.7%,同时在不影响模型大小的情况下提高了模型的检测效果。在此基础上,引入 Taylor 剪枝技术,对改进模型进行剪枝,进一步降低模型的计算量以及模型体积,此外,剪枝后的模型经过微调后在精确率、召回率和 mAP 均有提升。实验证明,各个模块在网络中均有正向作用。

4.2　剪枝方法对比实验

为验证本文所采用剪枝方法对模型检测性能的影响,本文横向对比了 5 种经典剪枝方法,包括 Slim[18]、Lamp[19]、Group Norm[20]、Group Sl[20]、Growing Reg[21]。所有实验均基于改进后的 YOLOv8 模型进行剪枝,并将未改进的 YOLOv8 模型作为基线进行性能评估。通过对比不同方法在相同实验配置下的检测精度 P、召回率 R、平均精度 mAP、参数量 Param、计算量 FLOPs 以及模型体积,评估各剪枝方法的优越性。实验结果如表 4 所示。

由表 4 可知,Taylor 剪枝方法在检测性能指标上表现最为优异,在精确率 P、召回率 R 以及平均精度 mAP 上均达到最高水平。Lamp 和 Slim 剪枝方法由于忽略了权重间的关系,导致检测精确度大幅下降。得益于其基于数学分析来评估权重对损失函数贡献的原理,Taylor 方法显著区别于 Lamp 的权重绝对值排序和 Group Norm 的组归一化评估,通过引入梯度信息,克服了基于经验或简单规则的剪枝方法可能带来的不确定性。同时,Taylor 方法充分利用了一阶梯度信息,相较于 Group Sl 的梯度敏感性计算,能够更加精确地量化权重的重要性,有效减少了剪枝对模型性能的影响。此外,Taylor 方法在稀疏化过程中具备较高的稳定性,不像 Growing Reg 依赖正则化项的动态调整,也无需像 Group Norm 受限于组划分的超参数影响,其适用性更加广泛。综上所述,Taylor 方法在高精度剪枝优化领域提供了一种稳定且高效的方案。

表 4　不同剪枝方法对比实验结果

Table 4　Comparative experimental results of different pruning methods

剪枝方法	P/%	R/%	mAP@50-95/%	Param/10^6	FLOPs/G	Weights/MB
基线模型	78.5	65.9	48.9	3.0	8.1	6.0
Lamp	65.4	66.8	52.5	1.1	4.9	2.3
Slim	68.5	64.8	46.9	1.1	4.9	2.3
Group Norm	78	65.0	52.3	1.1	4.9	2.3
Group Sl	79	66.9	52.5	1.1	4.9	2.3
Growing Reg	78.5	66.4	53	1.1	4.9	2.3
本文	78.8	67.0	53.8	1.1	4.9	2.3

4.3　不同检测模型对比实验

为验证本文改进算法的综合检测能力，以 YOLOv8n 为基线，将 YOLOv3-Tiny、YOLOv5n、YOLOv7-Tiny、YOLOv9-Tiny、YOLOv10 n 以及其他学者的轻量化改进算法在相同的实验条件下进行对比实验，并增加推理速度评价指标进行实时性对比，结果如表 5 所示。

表 5　不同算法对比实验结果

Table 5　Comparative experimental results of different algorithms

模型	P/%	R/%	mAP@50-95/%	推理速度/ms	Param/M	FLOPs/G	Weights/MB
YOLOv3-Tiny	47.0	51.9	20.4	3.6	8.7	12.9	16.6
YOLOv5n	76.5	66.8	52.7	2.5	1.8	4.2	3.6
YOLOv7-Tiny	76.2	53.7	42.6	1.8	6.0	13.1	12.0
邓刚改进 YOLOv5[3]	×	×	53.2	17.9	×	×	×
谢昆改进 YOLOv5[22]	×	×	×	10.1	×	×	1.9
席凌飞改进 YOLOv7-Tiny[23]	×	×	×	14.9	5.8	12.2	11.6
YOLOv8n(基线)	78.5	65.9	48.9	1.8	3.0	8.1	6.0
YOLOv9-Tiny	81.8	66.9	54.5	3.6	2.6	10.7	22.0
YOLOv10n	70	66.9	42.8	1.6	2.3	6.5	5.5
本文	78.8	67.0	53.8	1.8	1.1	4.9	2.3

注：“×”表示原文中未给出数据

实验结果中，引用了邓刚与谢昆分别改进的 YOLOv5 模型，和席凌飞等改进的 YOLOv7-Tiny 作为参考，尽管其都基于相同的数据集开展实验，但由于实验环境差异以及数据增强等因素，本文仅将其视为辅助对比。在全面对比中，YOLOv3-Tiny 模型的参数量和计算量都远高于基线模型，且检测效果一般；YOLOv5n 模型在 mAP@50-95 上高于基线模型 3.8%，但推理速度低于基线模型；YOLOv7-Tiny 推理速度与基线模型一致，但其检测精度均低于基线模型；YOLOv9-Tiny 相对基线模型，精确率 P、召回率 R 和 mAP@50-95 均有提升，但代价是计算量成倍增加，导致实时性大打折扣。而最新的 SOTA 模型 YOLOv10n 相比基线模型，虽然在推理速度上有优势，且参数和计算量更少，但其对于铝材的缺陷检测精确度远不如基线模型。综合来看，首先，本文的改进模型参数量、计算量和模型体积均最低，实现了模型轻量化，其次兼顾了检测速度和检测准确度，在铝材缺陷检测上具有优越性。

4.4　可视化实验

为直观展示改进后的模型对铝材缺陷的检测效果。从图 12 中可以看到，改进模型通过更为精确的多尺度特征融合方法（如 SC-BiFPN）和任务动态对齐的检测头（TDADH），有效地缓解了 YOLOv8n 对于喷流、不导电、划痕、漆泡等缺陷的漏检问题，对于细微缺陷能够更准确地识别出来。

4.5　边缘部署测试

为验证为验证本文改进算法在端/边设备上的部署情况，将算法部署在 RK3568 计算平台上进行测试。

首先将 PC 上训练好的权重文件（. pt 格式）通过 ONNX 框架转换为. onnx 格式的中间推理文件。ONNX 框架具有良好的兼容性，支持多种文件格式的转换，便于模型在不同推理框架之间迁移。接着在 PC 环境下使用瑞芯微神经网络（rockchip neural network，RKNN）推理框架将. onnx 文件转换为量化后的. rknn 文件，采用 int8 精度以减少计算资源消耗并提升推理效率。最后通过数据传输线将. rknn 文件及相关测试数据传输至 RK3568 设备中，运行推理程序并输出检测结果。

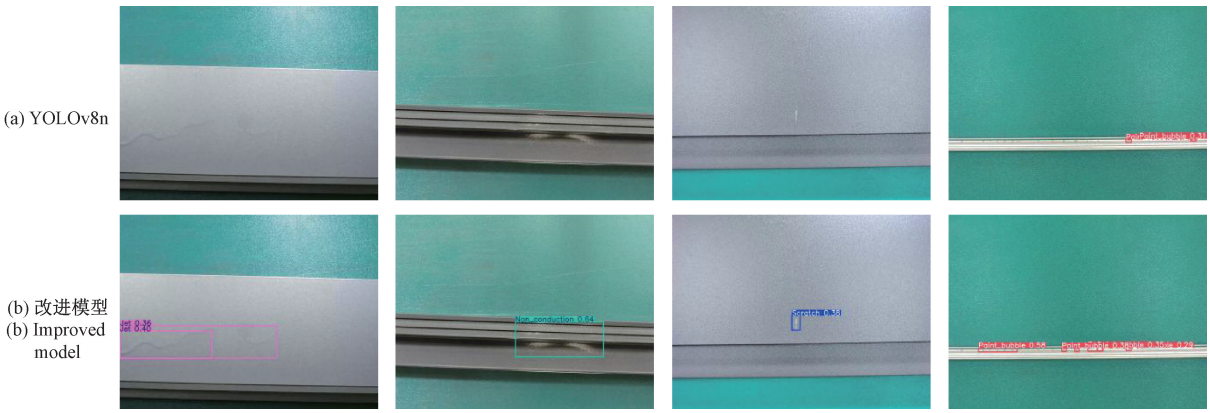


图 12　铝材缺陷检测效果对比

Fig. 12　Comparison of aluminum defect detection effect

由图 13 可以直观地看出,部署于 RK3568 计算平台的模型依旧能够准确定位并识别铝材不导电缺陷。在 RK3568 端侧设备上的推理速度如表 6 所示。

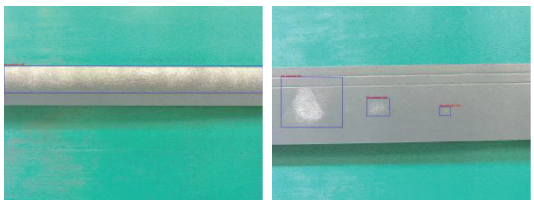


图 13　端侧部署检测推理结果

Fig. 13　End-side deployment detects inference results

表 6　端侧推理速度对比

Table 6　Comparison of inference speed on the edge device

模型	推理速度/ms
YOLOv8	45
改进模型	60

从部署结果可以看出,本文改进算法不仅能够保持优异的检测性能,还具备良好的端侧部署效果,满足实时性与资源受限环境下的应用需求。

5　结　　论

本文提出了一种适用于端侧部署的轻量化 YOLOv8 模型。通过在主干网络引入 PSA 注意力机制和 SCDown 下采样,有效提升了模型对细微缺陷的提取能力,并通过 MLCA 注意力机制增强了特征提取的表现;同时,基于 BiFPN 改进的 SC-BiFPN 颈部网络与任务动态对齐的 TDADH 特征检测头,显著优化了多尺度特征融合与目标定位精度;引入的 MPDIoU 损失函数,提升了边界框回归性能;最后,采用 Taylor 方法对 YOLOv8 模型进行裁剪,显著降低了参数量和计算成本,以便在资源受限的边缘计算设备上实现高效部署。实验结果表明,该轻量化 YOLOv8 模型在大幅降低模型规模的同时保持了较高的检测精度和召回率,适用于小型算力硬件的缺陷检测需求。未来的研究将聚焦于进一步优化模型的量化操作和后处理流程,以提升轻量化 YOLOv8 模型在端/边设备上的检测效果和运行效率。

参考文献

[1]　伍麟,郝鸿宇,宋友. 基于计算机视觉的工业金属表面缺陷检测综述[J]. 自动化学报, 2024, 50(7): 1261-1283.
WU L, HAO H Y, SONG Y. A review of computer vision based defect detection on industrial metal surfaces[J]. Acta Automatica Sinica, 2024, 50(7): 1261-1283.

[2]　李澄非,蔡嘉伦,邱世汉,等. 基于改进 YOLOv4 的铝材表面缺陷检测方法[J]. 中国测试, 2024, 50(5): 160-166.
LI CH F, CAI J L, QIU SH H, et al. Defect detection method in aluminum material surface based on improved YOLOv4[J]. China Measurement Test, 2024, 50(5): 160-166.

[3]　邓钢,赵庆华,祁光威,等. 基于 YOLOv5 的铝合金型材表面缺陷检测方法研究[J]. 现代制造工程, 2023(11): 120-128.
DENG G, ZHAO Q H, QI G W, et al. Research on surface defect detection method of aluminium alloy profiles based on YOLOv5[J]. Modern Manufacturing Engineering, 2023(11): 120-128.

[4]　吕秀丽,杨昕升,曹志民. 改进 YOLOv8 的 PCB 表面缺陷检测算法[J]. 电子测量技术, 2024, 47(12): 100-108.
LYU X L, YANG X SH, CAO ZH M. Improved PCB surface defect detection algorithm for YOLOv8[J]. Electronic Measurement Technology, 2024, 47(12): 100-108.

[5]　胡惠娟,秦一锋,徐鹤,等. 面向无人机航拍图像的 YOLOv8 目标检测改进算法[J/OL]. 计算机科学, 1-

18[2024-10-22]. http://kns. cnki. net/kcms/detail/50. 1075. TP. 20240925. 1400. 013. html.
HU H J, QIN Y F, XU H, et al. An improved YOLOv8 target detection algorithm for UAV aerial images[J/OL]. Computer Science, 1-18[2024-10-22]. http://kns. cnki. net/kcms/detail/50. 1075. TP. 20240925. 1400. 013. html.

[6] 赵佰亭,吴俊东,贾晓芬. 融合特征增强的轻量化罐道缺陷检测算法[J]. 电子测量与仪器学报, 2023, 37(6): 159-168.
ZHAO B T, WU J D, JIA X F. Lightweight tank channel defect detection algorithm incorporating feature enhancement[J]. Journal of Electronic Measurement and Instrument, 2023, 37(6): 159-168.

[7] 张上,许欢,张岳. 轻量级锻件表面裂纹检测算法[J]. 电子测量技术, 2024, 47(11): 123-130.
ZHANG SH, XU H, ZHANG Y. Algorithm for surface crack detection on lightweight forgings[J]. Electronic Measurement Technology, 2024, 47(11): 123-130.

[8] 赵亚凤,宋文华,刘晓璐,等. YOLO-FCA:面向钢轨缺陷的实时轻量化检测模型[J/OL]. 电子测量技术, 1-10[2024-10-22]. http://kns. cnki. net/kcms/detail/11. 2175. tn. 20241019. 1115. 022. html.
ZHAO Y F, SONG W H, LIU X L, et al. YOLO-FCA: A real-time lightweight inspection model for rail defects[J/OL]. Electronic Measurement Technology, 1-10[2024-10-22]. http://kns. cnki. net/kcms/detail/11. 2175. tn. 20241019. 1115. 022. html.

[9] 徐尽达,陈慈发,张上. 基于轻量级算法的水上垃圾小目标检测研究[J]. 电子测量技术, 2024, 47(18): 145-154.
XU J D, CHEN C F, ZHANG SH. Research on small target detection of waterborne garbage based on lightweight algorithm[J]. Electronic Measurement Technology, 2024, 47(18): 145-154.

[10] WANG A, CHEN H, LIU L H, et al. YOLOv10: Real-time end-to-end object detection[J]. ArXiv preprint arXiv:2405. 14458, 2024.

[11] GUO Y H, LI Y D, WANG L Q, et al. Depthwise convolution is all you need for learning multiple visual domains[C]. AAAI Conference on Artificial Intelligence, 2019, 33(1): 8368-8375.

[12] WAN D H, LU R SH, SHEN S Y, et al. Mixed local channel attention for object detection[J]. Engineering Applications of Artificial Intelligence, 2023, 123: 106442.

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

[14] ZHU X ZH, HU H, LIN S, et al. Deformable convnets v2: More deformable, better results[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 9308-9316.

[15] ZHENG ZH H, WANG P, REN D W, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation[J]. IEEE Transactions on Cybernetics, 2021, 52(8): 8574-8586.

[16] MA S L, XU Y. MPDIoU: A loss for efficient and accurate bounding box regression[J]. ArXiv preprint arXiv:2307. 07662, 2023.

[17] MOLCHANOV P, MALLYA A, TYREE S, et al. Importance estimation for neural network pruning[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 11264-11272.

[18] LIU ZH, LI J G, SHEN ZH Q, et al. Learning efficient convolutional networks through network slimming[C]. IEEE International Conference on Computer Vision, 2017: 2736-2744.

[19] LEE J, PARK S, MO S, et al. Layer-adaptive sparsity for the magnitude-based pruning[J]. ArXiv preprint arXiv:2010. 07611, 2020.

[20] FANG G F, MA X Y, SONG M L, et al. Depgraph: Towards any structural pruning[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 16091-16101.

[21] WANG H, QIN C, ZHANG Y L, et al. Neural pruning via growing regularization[J]. ArXiv preprint arXiv:2012. 09243, 2020.

[22] 谢昆,方凯,陈娟,等. 基于改进 YOLOv5s 的铝材表面缺陷检测方法[J]. 制造技术与机床, 2024(1): 179-184.
XIE K, FANG K, CHEN J, et al. Aluminium surface defects detection method based on improved YOLOv5s[J]. Manufacturing Technology and Machine Tools, 2024(1): 179-184.

[23] 席凌飞,伊力哈木·亚尔买买提. 基于轻量化 YOLOv7-tiny 的铝材表面缺陷检测方法[J]. 科学技术与工程, 2024, 24(27): 11786-11794.
XI L F, YILIHAMU Y. Detection of surface defects in aluminium based on lightweight YOLOv7-tiny[J]. Science Technology and Engineering, 2024, 24(27): 11786-11794.

作者简介

艾峰,硕士研究生,主要研究方向为目标检测、模型轻量化设计方法。
E-mail:2112201024@mail2. gdut. edu. cn

邓耀华(通信作者),博士,教授,博士生导师,主要研究方向为智能视觉传感、图像处理、智能检测技术等。
E-mail:dengyaohua@gdut. edu. cn