http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2024.0426

改进轻量高效 EHH-YOLOv8s 的带钢表面缺陷检测算法

梁礼明*，龙鹏威，卢宝贺，欧燕燕，曾璐

(江西理工大学 电气工程与自动化学院，赣州 341000)

摘 要：针对当前钢材表面缺陷检测存在效率低和精度差等问题，以及现有缺陷检测模型参数量庞大、检测精度和实时性较差的挑战，提出一种基于 YOLOv8s 的轻量级钢材表面缺陷检测算法 EHH-YOLOv8s。采用轻量级的 EfficientViT 网络作为骨干网络，以降低模型复杂度并更好地处理跨尺度特征信息，从而提升检测性能；重构特征交互模块 H-C2f，有效保留空间和通道特征，抑制冗余信息，以提高检测精度和速度；采用 Haar 小波下采样 (HWD) 模块，以扩大感受野、降低模型复杂度、提升训练和检测速度。在 NEU-DET 和 Severstal 钢材缺陷数据集上进行实验验证，结果表明，相较于 YOLOv8s 算法，EHH-YOLOv8 算法的平均精度均值 (mAP) 分别提升 6.9% 和 5.4%，参数量和浮点运算量分别降低 4.1×10^6 和 13.2×10^9 次，检测速度分别达到 136 帧/s 和 118 帧/s，表明所提算法在检测精度、速度和轻量化方面取得了更好的平衡，为边缘终端设备提供高精度和实时性的可靠参考。在 PCB 板缺陷数据集上进行验证，结果显示，所提算法具备良好的泛化性能。

关 键 词：缺陷检测；轻量化 YOLOv8s；EfficientViT 网络；交换模块 H-C2f；Haar 小波下采样
中图分类号：TP391.4
文献标志码：A **文章编号**：1001-5965(2026)11-3699-13

钢材作为建筑行业中常见的材料之一，确保高质量的钢材生产对居民生命和财产安全至关重要。钢材表面缺陷检测算法可分为传统算法和基于深度学习的算法。传统算法存在耗时、费力等问题，导致误检率和漏检率较高，难以满足高效和高质量的生产需求。随着深度学习[1]技术的快速发展，基于图像的物体检测[2]算法在工业领域得到广泛应用[3]。然而，深度学习模型的训练和测试需要占用大量的计算资源，而终端设备的计算能力有限，这给钢材表面缺陷检测的实际应用带来一定困难和挑战[4]。因此，研究和应用轻量级、高效和准确的钢材表面缺陷检测技术具有重要的现实需求。

目前，基于深度学习的表面缺陷检测主流算法分为两阶段和单阶段的检测算法。两阶段目标检测算法具有较高的检测精度，但检测速度较低且模型参数量较大，不利于边缘终端设备部署，典型算法包括 R-CNN[5]、Fast R-CNN[6]、Faster R-CNN[7] 和 Mask R-CNN[8] 等。单阶段目标检测算法虽具备较快的检测速度，但检测精度不如两阶段目标检测算法，典型算法有 SSD[9] 和 YOLO[10-11] 系列等。然而，在钢材表面缺陷检测领域，因终端设备计算能力有限，过量的参数不利于模型在边缘终端设备上部署。因此，在实际应用场景中，缺陷检测的实时性也成为评估网络性能的重要指标。

为应对深度学习算法在终端设备上部署和实时性等挑战，一些研究者提出一系列轻量级带钢表面缺陷检测模型。例如，文献 [12] 基于 YOLOv5 构建了一种轻量级带钢表面缺陷检测模型 YOLOv5s-GCE，

收稿日期：2024-06-12；**录用日期**：2024-07-30；**网络出版时间**：2024-08-08 08:50
网络出版地址：link.cnki.net/urlid/11.2625.V.20240807.1403.003
基金项目：国家自然科学基金 (51365017,61463018)；江西省自然科学基金面上项目 (20192BAB205084)；江西省教育厅科学技术研究重点项目 (GJJ170491,GJJ2200848)；江西省研究生创新专项资金 (YC2022-S676)
* **通信作者**. E-mail：9119890012@jxust.edu.cn

引用格式：*梁礼明，龙鹏威，卢宝贺，等. 改进轻量高效 EHH-YOLOv8s 的带钢表面缺陷检测算法 [J]. 北京航空航天大学学报，2026，52（11）：3699-3711. Liang L M，Long P W，Lu B H，et al. Improved lightweight and efficient EHH-YOLOv8s algorithm for strip steel surface defect detection [J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52（11）：3699-3711 （in Chinese）.*

该模型引入 GhostConv 卷积和坐标注意力机制，以减少参数量，实现轻量化目的，但在小目标检测方面表现较为不足。文献 [13] 采用 EfficientNet 网络作为骨干特征提取网络，设计 EDDNet 模型，有效减少计算开销。文献 [14] 使用 MobilenetV2 模块和卷积块注意力模块 (convolutional block attention module, CBAM) 构建 CBAM-MobilenetV2-YOLOv5 模型，实现 YOLOv5 模型轻量化。文献 [15] 基于 YOLOv5 算法构造 SPP_Res 特征金字塔结构，同时加入多头注意力机制，有效提升检测精度，但该算法对部分小尺度目标 (如裂缝类) 检测效果较差。文献 [16] 通过使用轻量化 ShuffleNetv2 作为 YOLOv5 模型的骨干网络，有效压缩模型体积并提升检测速度。文献 [17] 将 MobileNet 引入 Mask R-CNN 目标检测框架，作为骨干特征提取网络，成功降低模型参数量和计算量。文献 [18] 在 YOLOv7 算法基础上，通过 SPPCSPC_Group 对空间金字塔池化模块进行分组，设计出轻量级 YOLOv7 模型。上述研究对参数量庞大的目标检测算法具有启发意义。但轻量化操作虽有利于减少模型体积，却会降低网络的非线性建模能力，导致模型精度损失。因此，尽管上述文献在缺陷检测领域取得一定进展，但依然无法较好地平衡检测精度、速度和轻量级问题。

鉴于上述研究存在的局限性，本文在 YOLOv8s 算法的基础上，提出一种轻量级高效的钢材表面缺陷检测算法 EHH-YOLOv8s (EfficientViT+H-C2f+HWD，简称 EHH)，并实现对缺陷目标较高精度和速度的检测。主要研究工作如下：

1）采用轻量级的 EfficientViT 作为骨干特征提取网络，融入高效的多尺度注意力 (multi-scale attention, MSA) 机制，高效处理跨尺度信息，且显著压缩模型体积，加快模型检测速度。

2）设计轻量级特征交互模块 H-C2f，旨在优化模型参数量，充分融合空间和通道的语义信息，提高检测精度和速度。

3）使用 Haar 小波下采样 (Haar wavelet down-sampling, HWD) 模块作为整体网络的下采样操作，扩大模型感受野，同时压缩模型体积，提升模型性能。

4）在 NEU-DET 数据集上进行实验，验证本文算法的优越性。相较于主流算法，本文算法具备更低的计算资源、更快的检测速度和更高的检测精度。在 Severstal 数据集上的实验结果显示，EHH-YOLOv8 算法具有良好的鲁棒性。此外，在 PCB 数据集上的验证进一步展示了本文算法的泛化性能。

1 YOLOv8s 模型

YOLOv8 是 Ultralyticsgongs 团队开源的网络，集成检测、分割和实时追踪等任务。相较于现有 YOLO 系列算法，其最大的改进在于采用中心检测算法的无锚框检测方式，通过解耦头结构将分类与回归分离，减少了因锚框数量和选择尺寸而导致的干扰，并使用解耦头作为最终检测头，有效提升网络性能。此外，将 C3 模块换成梯度流丰富且轻量的 C2f 模块，该模块综合了 C3 模块和 YOLOv7 中 ELAN 的优点。YOLOv8 提供了 5 种不同尺寸的模型，为平衡检测精度和速度，本文选用 YOLOv8s 作为钢材表面缺陷检测的基线模型。该网络结构包括输入端、骨干网络、特征融合网络和预测网络 4 部分，分别用于特征输入、特征提取、特征融合和预测输出。YOLOv8s 网络结构如图 1 所示。

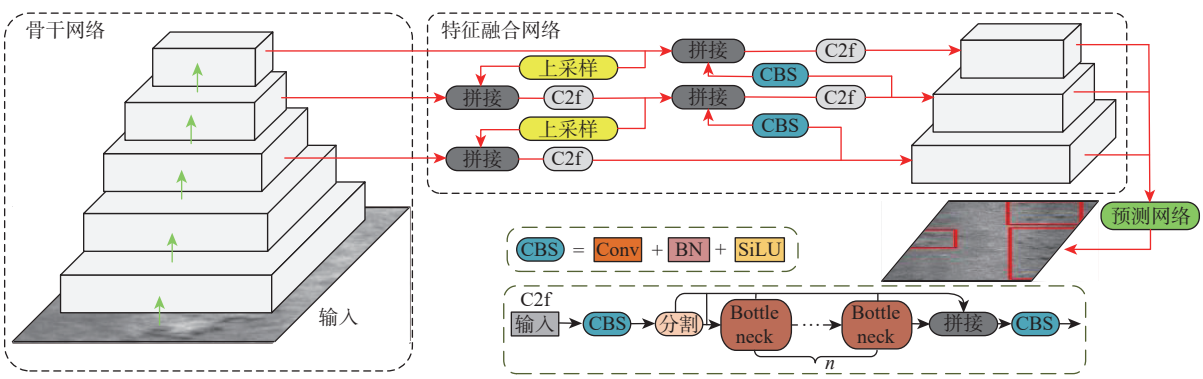


图 1 YOLOv8s 网络结构

Fig. 1 YOLOv8s network structure

2 本文模型

2.1 EHH-YOLOv8s 模型

针对现有钢材表面缺陷检测精度和效率低、通用算法参数量大等问题，本文提出一种轻量级的 EHH-YOLOv8s 检测模型，用于快速准确地检测钢材表面形态多样、结构复杂的缺陷。首先，将轻量级的 EfficientViT 网络替代原 YOLOv8s 的骨干网

络, 在维持较高精度的同时有效压缩模型体积。其次, 重新设计特征交互模块 H-C2f 取代原特征融合网络中的 C2f 模块, 抑制空间和通道维度上的冗余信息, 提升网络检测精度。最后, 将原网络颈部的下采样操作替换为 HWD 模块, 扩大感受野, 充分保留特征图细粒度信息, 提升模型的检测性能。EHH-YOLOv8s 网络结构如图 2 所示, 改进部分用红框标注。图中: SPPF 表示快速空间金字塔池化。

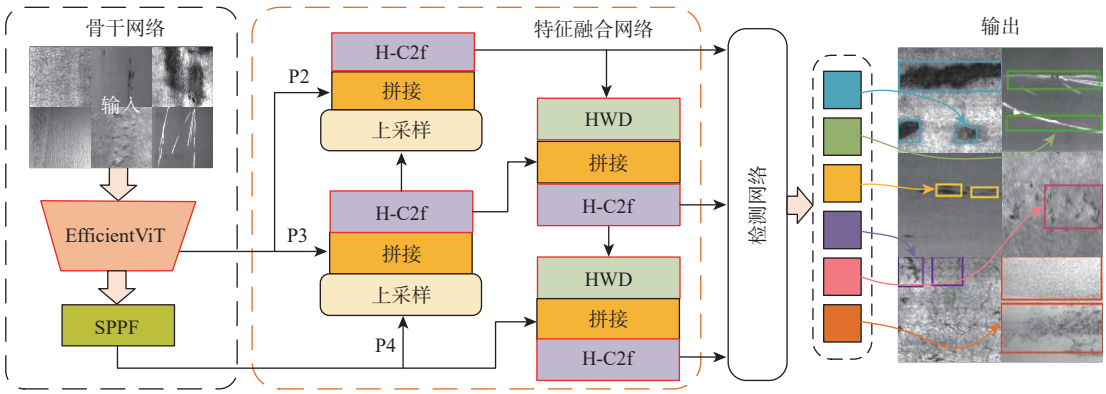


图 2　EHH-YOLOv8s 网络结构

Fig. 2　EHH-YOLOv8s network structure

2.2　EfficientViT 骨干网络

YOLOv8 的骨干网络由一系列卷积层和残差模块堆叠而成, 导致模型体积过大, 限制训练和推理速度, 不适合边缘设备部署; 且传统的骨干网络结构在处理跨尺度信息方面存在局限性, 难以准确捕捉钢材表面缺陷的全局特征。为解决上述问题, Cai 等[19]提出一种 EfficientViT 网络, 其结构如图 3 所示。图中: MBConv 表示倒置瓶颈卷积模块, GConv 表示分组卷积, DSConv 表示完整深度可分离卷积, FPN 表示特征金字塔网络。该网络采用轻量化设计, 减少网络的参数量和计算量, 使模型更易于部署在移动设备或资源受限的环境中。此外, EfficientViT 网络结合轻量级 MSA 机制, 有助于处理钢材表面缺陷的跨尺度信息, 提升模型的检测准确性。

MSA 摒弃复杂的 Softmax 注意力, 首先, 采用 ReLU 线性注意力对全局感受野进行操作。假设给定输入特征 $\boldsymbol{x}\in\mathbf{R}$, Softmax 注意力的广义形式在数学可表示为

$$\boldsymbol{Q}_i=\sum_{j=1}^{N}\frac{\mathrm{Sim}(\boldsymbol{Q}_i,\boldsymbol{K}_i)}{\sum_{j=1}^{N}\mathrm{Sim}(\boldsymbol{Q}_i,\boldsymbol{K}_i)}\boldsymbol{V}_j \tag{1}$$

式中: $\mathrm{Sim}(\boldsymbol{Q},\boldsymbol{K})=\mathrm{ReLU}(\boldsymbol{Q})\cdot\mathrm{ReLU}(\boldsymbol{K})^{\mathrm{T}}$ 为 ReLU 线性注意力的相似性函数; $\boldsymbol{Q}=\boldsymbol{x}\boldsymbol{W}_{\boldsymbol{Q}},\boldsymbol{K}=\boldsymbol{x}\boldsymbol{W}_{\boldsymbol{K}},\boldsymbol{V}=\boldsymbol{x}\boldsymbol{W}_{\boldsymbol{V}}$, 且 $\boldsymbol{W}_{\boldsymbol{Q}},\boldsymbol{W}_{\boldsymbol{K}},\boldsymbol{W}_{\boldsymbol{V}}\in\mathbf{R}$ 为线性投影矩阵; $\boldsymbol{Q}_i$ 表示矩阵 $\boldsymbol{Q}$ 的第 i 行。

然后, 利用矩阵乘法的结合性, 成功将计算复杂度和内存占用从二次降到线性, 同时保持其性能

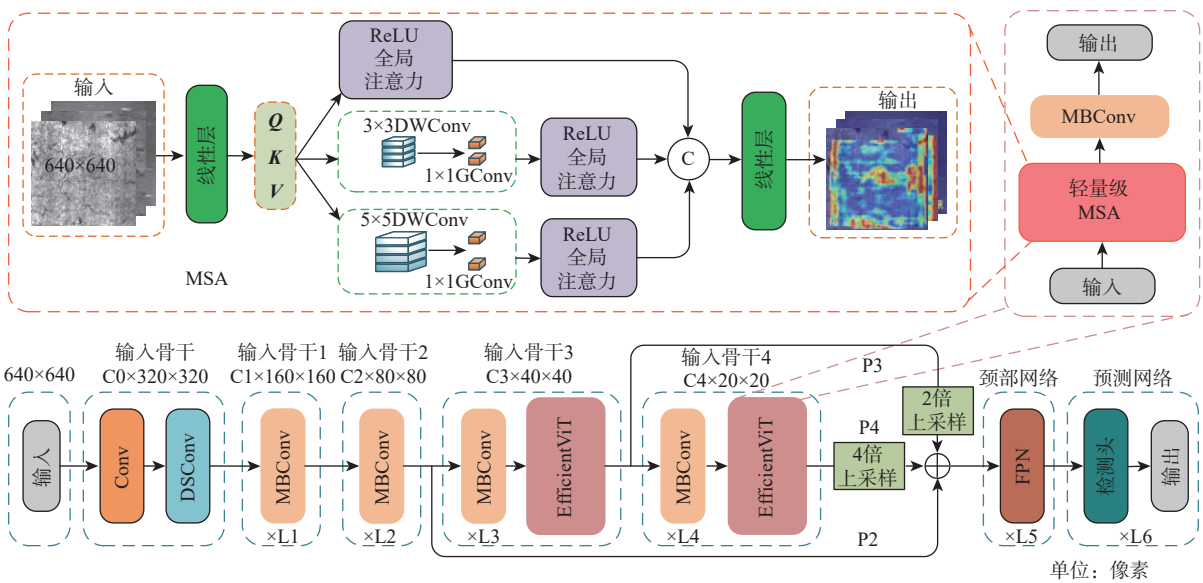


图 3　EfficientViT 宏观网络结构

Fig. 3　EfficientViT macro network structure

不受影响。最后，ReLU线性注意力的计算表达式在数学上可表示为

$$\boldsymbol{O}_i=\frac{\mathrm{ReLU}(\boldsymbol{Q}_i)\left[\sum_{j=1}^{N}\mathrm{ReLU}(\boldsymbol{K}_j)^{\mathrm{T}}\boldsymbol{V}_j\right]}{\mathrm{ReLU}(\boldsymbol{Q}_i)\sum_{j=1}^{N}\mathrm{ReLU}(\boldsymbol{K}_j)^{\mathrm{T}}} \tag{2}$$

可知，ReLU只需计算一次$\sum_{j=1}^{N}\mathrm{ReLU}(\boldsymbol{K}_j)^{\mathrm{T}}$和$\sum_{j=1}^{N}\left[\mathrm{ReLU}(\boldsymbol{K}_j)^{\mathrm{T}}\boldsymbol{V}_j\right]$，在每个查询中可对其重复使用，该过程只需消耗$\boldsymbol{O}(N)$的计算成本和内存，无需涉及Softmax，避免了硬件效率低和延时等问题。

为了在GPU上高效执行上述操作，MSA先将所有的深度可分离卷积(DWConv)融合为单个DWConv，并采用组卷积范式，再通过全局注意力操作提取多尺度的全局特征，最终将来自不同尺度的特征沿头部维度连接，输入线性投影层进行融合。

EfficientViT宏观网络主要由输入骨干和4个阶段组成，采用MSA和MBConv模块进行设计。信息传递过程中，特征图逐渐减小，通道数逐渐增加。其中，MSA模块用于捕获上下文信息，MBConv模块的步长为2，用于捕获本地信息和下采样操作。此外，该网络共设计了6种大小不同版本的模型(V0～V5)，以满足各种效率约束。

2.3 H-C2f模块

C2f是一种通道压缩和特征融合模块。语义信息经C2f模块后，因通道压缩易导致信息丢失，降低模型在复杂场景下的检测性能。此外，C2f模块采用深层网络结构，导致计算复杂度较高，影响模型的运行效率和速度。Li等[20]提出混合注意力模块(hybrid attention module, HAM)，将通道注意力模块(channel attention module, CAM)和空间注意力模块(spatial attention module, SAM)进行融合，有效提升模型对关键特征的提取能力，且该模块采用快速一维卷积实现跨通道的信息交互，保证信息传递过程不受通道降维影响，同时降低模型的复杂度。本文受此启发，将C2f模块与HAM模块进行融合，设计了H-C2f模块。HAM和H-C2f模块网络结构如图4所示。

假设输入特征为$\boldsymbol{F}\in\mathbf{R}^{H\times W\times C}$，HAM模块首先导出一维通道注意张量$\boldsymbol{A}_{\mathrm{C}}$，并与原始输入张量相乘得到通道细化特征$\boldsymbol{F}'$。其次，SAM模块沿通道轴线将$\boldsymbol{F}'$分为$\boldsymbol{F}'_1$和$\boldsymbol{F}'_2$ 2组，并分别产生二维空间注意力张量$\boldsymbol{A}_{\mathrm{S1}}$和$\boldsymbol{A}_{\mathrm{S2}}$。然后，与$\boldsymbol{F}'_1$和$\boldsymbol{F}'_2$相乘得到一对空间细化特征$\boldsymbol{F}''_1$和$\boldsymbol{F}''_2$。最后，将两者相加得到最终的细化特征$\boldsymbol{F}''$。上述过程在数学上可分别表示为

$$\begin{cases}\boldsymbol{F}'=\boldsymbol{A}_{\mathrm{C}}(\boldsymbol{F})\otimes\boldsymbol{F}\\ \boldsymbol{F}'=\boldsymbol{F}'_1\oplus\boldsymbol{F}'_2\\ \boldsymbol{F}''_1=\boldsymbol{A}_{\mathrm{S1}}(\boldsymbol{F}'_1)\otimes\boldsymbol{F}'_1\\ \boldsymbol{F}''_2=\boldsymbol{A}_{\mathrm{S2}}(\boldsymbol{F}'_2)\otimes\boldsymbol{F}'_2\\ \boldsymbol{F}''=\boldsymbol{F}''_1\oplus\boldsymbol{F}''_2\end{cases} \tag{3}$$

式中：⊗和⊕分别表示元素的乘法和加法。

CAM网络结构如图5所示。首先，通过平均池化和最大池化操作对空间维度信息进行聚合，生成平均池化特征$\boldsymbol{F}_{\mathrm{C}}^{\mathrm{avg}}$和最大池化特征$\boldsymbol{F}_{\mathrm{C}}^{\mathrm{max}}$。然后，通过自适应机制模块，将$\boldsymbol{F}_{\mathrm{C}}^{\mathrm{avg}}$和$\boldsymbol{F}_{\mathrm{C}}^{\mathrm{max}}$分别与可训练参数$\alpha$和$\beta$相乘，得到信息提取特征$\boldsymbol{F}_{\mathrm{C}}^{\mathrm{add}}$。该特征不仅保留了原始信息，且通过自适应调整特征倾向，使特征更加丰富。CAM模块采用快速一维卷积对跨通道的交互信息进行处理，降低模型复杂度，并避免通道降维对性能的影响。假设一维卷积的核大小为k，大小由通道数C决定，其计算式可表示为

$$k=\phi(C)=\left|\frac{\log_2 C}{\gamma}+\frac{b}{\gamma}\right|_{\mathrm{odd}} \tag{4}$$

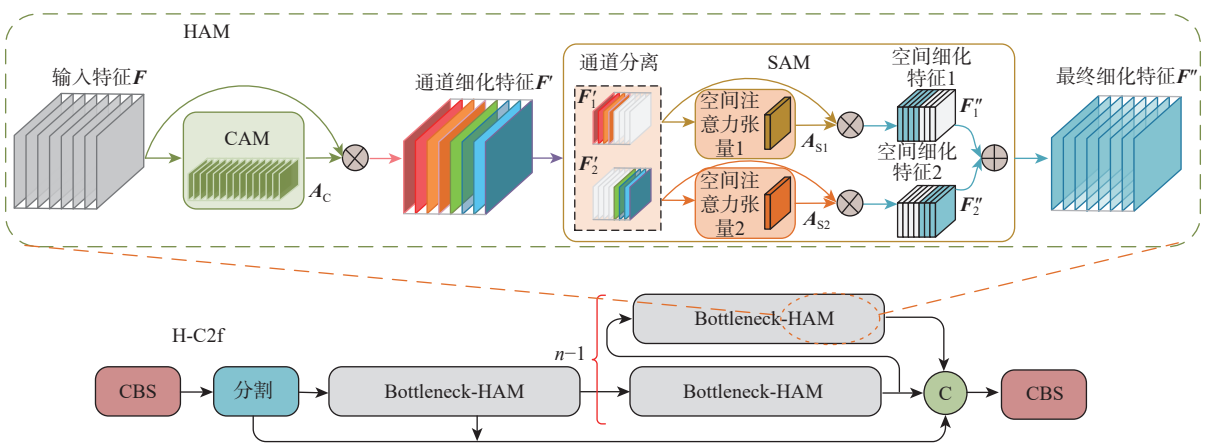


图4 HAM和H-C2f网络结构

Fig. 4 HAM and H-C2f network structure

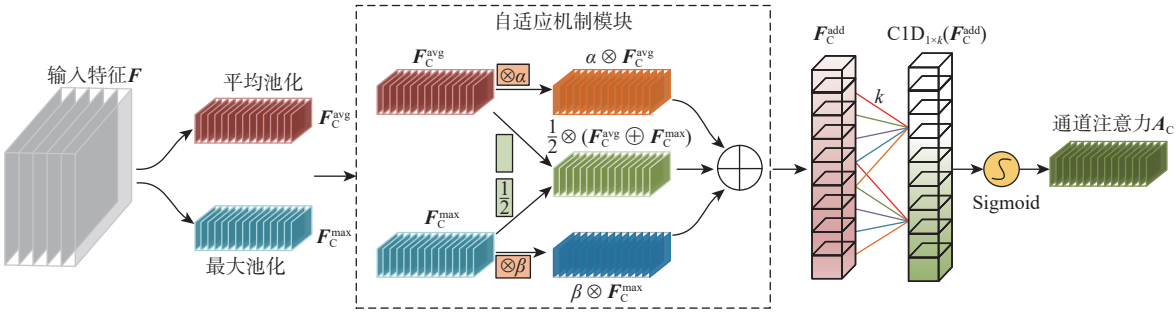


图 5　CAM 网络结构

Fig. 5　CAM network structure

式中: $|t|_{\text{odd}}$表示t最接近的奇数; γ和b为超参数。

最后, 特征信息$\boldsymbol{F}_{\text{C}}^{\text{add}}$经过一维卷积后, 采用 Sigmoid 函数激活输出特征张量$\boldsymbol{A}_{\text{C}}$, 得到通道细化特征$\boldsymbol{F}'$, 并将其作为 SAM 模块的输入。

上述过程在数学上可分别表示为

$$\begin{cases} \boldsymbol{F}_{\text{am}} = \boldsymbol{F}_{\text{C}}^{\text{avg}} \oplus \boldsymbol{F}_{\text{C}}^{\text{max}} \\ \boldsymbol{F}_{\alpha\beta} = \alpha \otimes \boldsymbol{F}_{\text{C}}^{\text{avg}} \oplus \beta \otimes \boldsymbol{F}_{\text{C}}^{\text{max}} \\ \boldsymbol{F}_{\text{C}}^{\text{add}} = \dfrac{1}{2} \otimes \boldsymbol{F}_{\text{am}} \oplus \boldsymbol{F}_{\alpha\beta} \end{cases} \tag{5}$$

$$\boldsymbol{A}_{\text{C}} = \sigma(\text{C1D}_{1\times k}(\boldsymbol{F}_{\text{C}}^{\text{add}})) \tag{6}$$

式中: α和β为大于 0 且小于 1 的浮动实数; σ表示 Sigmoid 函数; $\text{C1D}_{1\times k}$表示核大小为k的一维卷积。

SAM 网络结构如图 6 所示。首先, 对通道细化特征$\boldsymbol{F}'$进行通道分离, 得到$\boldsymbol{F}_1'$和$\boldsymbol{F}_2'$。其次, 采用平均池化和最大池化操作对其进行聚合, 得到 2 对二维特征$\boldsymbol{F}_{\text{S1}}^{\text{max}}$和$\boldsymbol{F}_{\text{S1}}^{\text{avg}}$及$\boldsymbol{F}_{\text{S2}}^{\text{max}}$和$\boldsymbol{F}_{\text{S2}}^{\text{avg}}$, 并将 2 对特征输出进行连接, 生成一对特征描述符。然后, 经过 7×7 的共享卷积层, 得到一对二维注意力图。最后, 通过归一化和激活函数操作, 生成空间注意力特征$\boldsymbol{A}_{\text{S1}}$和$\boldsymbol{A}_{\text{S2}}$。

上述过程在数学上可分别表示为

$$\begin{aligned} \boldsymbol{A}_{\text{S1}}(\boldsymbol{F}') &= \Phi(\text{C2D}_{7\times7}([\text{AvgPool}(\boldsymbol{F}_1');\text{MaxPool}(\boldsymbol{F}_1')])) = \\ &\Phi(\text{C2D}_{7\times7}([\boldsymbol{F}_{\text{S1}}^{\text{avg}};\boldsymbol{F}_{\text{S1}}^{\text{max}}])) \end{aligned} \tag{7}$$

$$\begin{aligned} \boldsymbol{A}_{\text{S2}}(\boldsymbol{F}') &= \Phi(\text{C2D}_{7\times7}([\text{AvgPool}(\boldsymbol{F}_2');\text{MaxPool}(\boldsymbol{F}_2')])) = \\ &\Phi(\text{C2D}_{7\times7}([\boldsymbol{F}_{\text{S2}}^{\text{avg}};\boldsymbol{F}_{\text{S2}}^{\text{max}}])) \end{aligned} \tag{8}$$

式中: Φ表示一系列非线性操作, 依次为批归一化、ReLU 函数和 Sigmoid 函数; $\text{C2D}_{7\times7}$表示 7×7 的共享卷积。

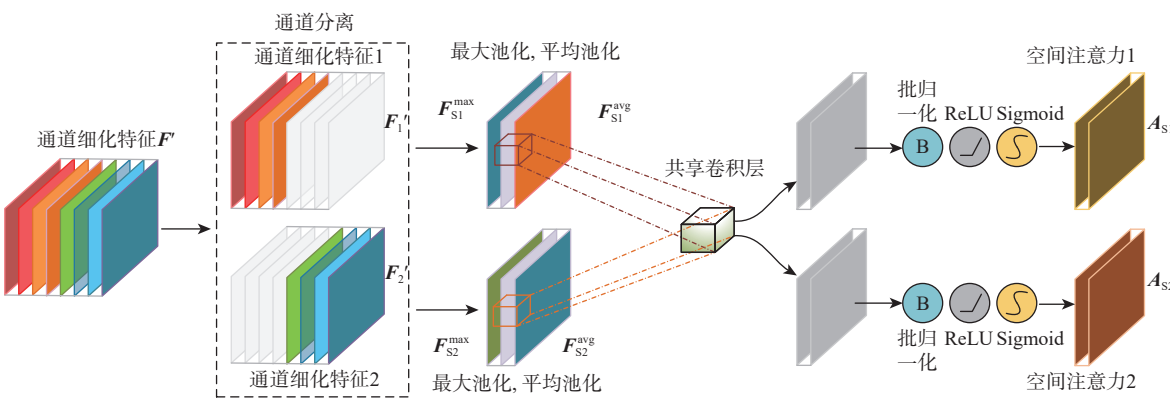


图 6　SAM 网络结构

Fig. 6　SAM network structure

2.4　HWD 模块

YOLOv8s 的下采样采用步长为 2 的卷积操作, 但该过程会使特征图尺寸迅速减小, 丢失细节信息, 导致感受野不足, 从而降低模型对细小缺陷或局部特征的识别准确性。Xu 等[21]提出 HWD 模型, 其网络结构如图 7 所示。HWD 利用 Haar 小波变换进行下采样操作, 充分保留图像细节的同时增加特征图的感受野, 使模型能更好地理解和捕捉图像整体特征, 提高检测准确性。此外, 相比于跨步卷积, HWD 设计简单高效, 能有效减少计算开销, 提升训练速度, 并提高检测精度。本文受此启发, 将 YOLOv8s 中步长为 2 的卷积替换为 HWD 模块, 以充分改善模型的检测性能。

HWD 模块主要包括无损特征编码块和特征学习块。其中, 无损特征编码块利用 Haar 小波变换对特征图进行转换和降低空间分辨率。首先, 将分

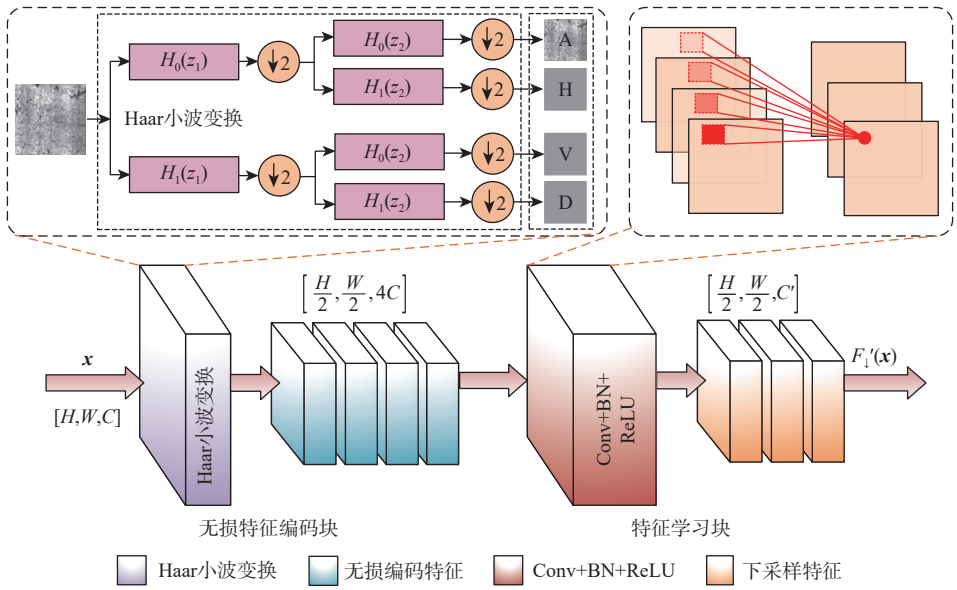


图 7 HWD 网络结构

Fig. 7 HWD network structure

辨率为 $H\times W$ 的图像进行分解，并利用高通(H_0)和低通(H_1)分解滤波器进行近似信息和高频信息的提取。然后，通过近似和细节分量的下采样操作(↓2)生成水平 (H)、垂直 (V) 和对角线 (D) 方向上的近似低频 (A) 和细节高频分量，从而有效保留细节特征信息。Haar 小波变换的基函数和尺度函数在数学上可表示为

$$\begin{cases}\phi_1(x')=\dfrac{1}{\sqrt{2}}\phi_{1,0}(x')+\dfrac{1}{\sqrt{2}}\phi_{1,1}(x')\\ \psi_1(x')=\dfrac{1}{\sqrt{2}}\phi_{1,0}(x')+\dfrac{1}{\sqrt{2}}\phi_{1,1}(x')\end{cases} \tag{9}$$

$$\phi_{1,k}(x')=\sqrt{2^j}\phi(2^jx'-k)\qquad k=0,1,\cdots,2^j-1 \tag{10}$$

式中：j和k为基函数的阶。

特征学习块由卷积层、批归一化层和 ReLU 激活层组成，用于调整特征映射的通道数，并尽可能过滤掉冗余信息，使后续层更高效地学习代表性特征。

假设特征图和下采样特征图的输入大小分别为$H\times W\times C$和$H/2\times W/2\times C$，则 HWD 模块与传统跨步卷积的参数量和计算复杂度在数学上可分别表示为

$$\begin{cases}P_{\mathrm{H}}=4C^2\\ F_{\mathrm{H}}=2HWC^2+3.75HWC\end{cases} \tag{11}$$

$$\begin{cases}P_{\mathrm{C}}=9C^2\\ F_{\mathrm{C}}=4.5HWC^2-0.25HWC\end{cases} \tag{12}$$

式中：P_{H}、P_{C}和F_{H}、F_{C}分别为参数量和计算复杂度。

由式(11)和式(12)可知，跨步卷积的参数量接近 HWD 模块的2 倍，而当$C>1$时，其计算开销也高于 HWD 模块。综上，HWD 模块不仅输出更丰富的细节特征映射信息，而且在参数量和计算量之间取得了较好的平衡。

3 实验结果及分析

3.1 数据集

本文实验数据集来源于东北大学 NEU-DET[16] 钢材缺陷数据集和 Kaggle 钢铁缺陷比赛数据集，后者由 Severstal 公司提供。2 个数据集分别包含 1 800 张和 1 120 张图像，涵盖市面上常见的裂缝、夹杂物、斑块、点蚀表面、氧化皮和划痕等 6 种缺陷类型。为增强模型的泛化性能和鲁棒性，减少过拟合的影响，对 Severstal 数据集进行数据增强，包括水平翻转、垂直翻转和水平垂直翻转，扩充比例为 1∶3。图 8 展示了原图和增强后的图像示例，对于训练集、验证集和测试集，2 个数据集均采用 6∶2∶2 的比例进行随机划分。

3.2 实验环境与参数设置

实验操作系统为 Windows11，采用 Pytorch1.11.0 作为深度学习框架，CPU 为 Intel Core i7-12700H，GPU 为 RTX4060，显卡为 8 GB。在训练阶段，采用 Adam 作为优化器，初始学习率为 0.01，动量为 0.973，衰减系数为 0.000 5。输入图像尺寸为 640 像素×640 像素，batch size 大小为 16，epoch 为 300 轮。

3.3 评价指标

实验采用平均精度均值 (mean average precision, mAP)、参数量、浮点运算量和检测速度 4 个指标对网络性能进行评估。mAP 表示所有类别的检测准确率的均值，由精确度 P、召回率 R 和平均精度 (average precision, AP) 计算所得。上述指标的定义式分别为

$$A_{\mathrm{P}}=\int_0^1 P\mathrm{d}R \tag{13}$$

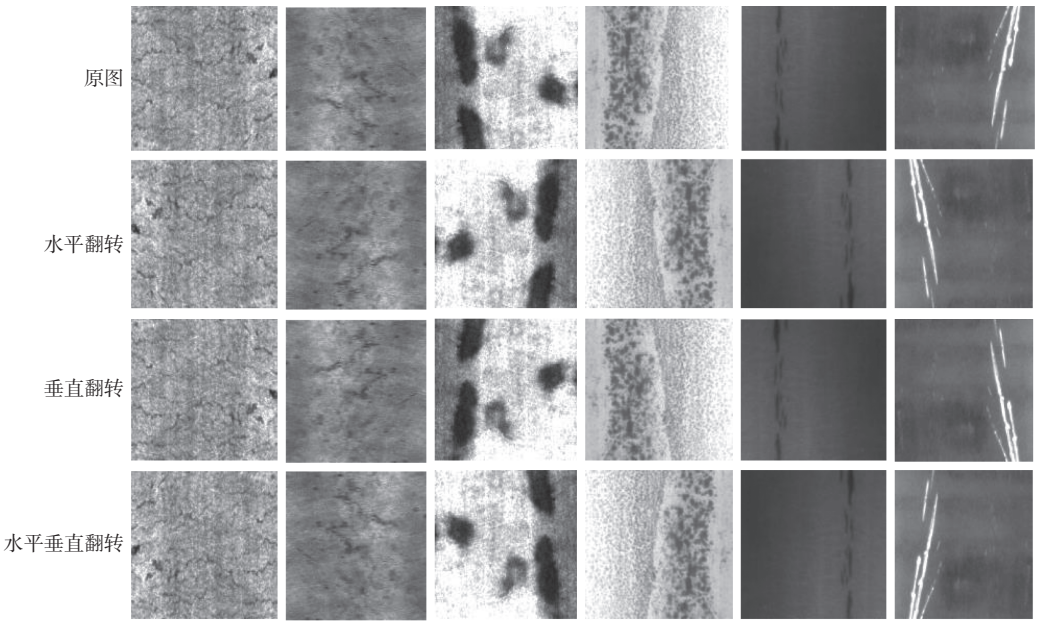


图 8　数据增强前后各图像示例

Fig. 8　Examples of each image before and after data enhancement

$$A_{\mathrm{mP}}=\frac{\sum_{i=1}^{N}A_{\mathrm{P}i}}{N} \tag{14}$$

$$P=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}} \tag{15}$$

$$R=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}} \tag{16}$$

$$F_{\mathrm{PS}}=\frac{N_{\mathrm{frame}}}{T_{\mathrm{elapsed}}} \tag{17}$$

式中: A_{P} 表示 *P-R* 曲线与坐标轴围成的面积; N 为待检测目标的类别数; N_{TP}、N_{FP} 和 N_{FN} 分别为正确检测的真阳性、误报阳性和漏报阳性的数量; N_{frame} 为检测图像总数; T_{elapsed} 为检测所需总时间。

3.4　改进模块实验

3.4.1　EfficientViT 网络实验

在 NEU-DET 数据集上, 对 EfficientViT 网络的 6 个版本 (V0～V5) 进行实验, 验证对指标提升效果最佳的版本, 实验结果如表 1 所示。

从表 1 结果可知, 随着 EfficientViT 网络版本从 V0 到 V5 逐步增加, 参数量和计算复杂度逐渐增加, 对 GPU 的性能要求也更高。选择 V0 版本时, 整体网络的参数量和计算开销相较于基线模型分别降低 2.8×10^6 和 4.9×10^9 次, 同时, mAP 和检测速度达到最优, 分别为 76.5% 和 126 帧/s, 优于其他版本。其中, EfficientViT-V0 是最基础的版本, 模型参数量和计算量较小, 适合在资源受限的环境提高模型的效率和速度。相较于其他版本, EfficientViT-V0 更适合捕捉和定位钢材缺陷的小尺度场景, 故而提升效果更为显著。

表 1　EfficientViT 网络不同版本实验结果

Table 1　Experimental results of different versions of EfficientViT network

模型	mAP%	参数量	浮点运算量/10^9次	检测速度/(帧·s^{-1})
基线模型	72.8	11.1×10^6	28.8	109
V0	76.5	8.3×10^6	23.9	126
V1	75.0	9.0×10^6	25.2	113
V2	75.2	10.2×10^6	26.9	101
V3	76.2	13.0×10^6	32.8	80
V4	75.3	14.9×10^6	37.7	75
V5	72.4	18.6×10^6	42.5	67

3.4.2　H-C2f 模块实验

为验证 HAM 与 C2f 模块融合的有效性, 在 NEU-DET 数据集上, 将其与 SE、CBAM、Shuffle-Attention、CA、SimAM 注意力机制与 C2f 模块融合后的效果进行对比, 实验结果如表 2 所示。

从表 2 结果可知, 将上述注意力与 C2f 模块融合后, mAP 值相比基线模型均得到一定程度的提升。选择 HAM 模块与 C2f 模块融合后, 对网络的整体贡献能力最佳, 检测精度和速度均达到最优水平。相较于基线模型, 参数量和计算复杂分别减少 0.9×10^6 和 2.2×10^9 次, 优于其余注意力机制。HAM 模块采用一维快速卷积实现跨通道信息交互, 有效提升模型对关键特征的提取能力, 同时大幅降低模型的复杂度。综上所述, 本文设计的 H-C2f 模块具备较高的优越性。

3.5　消融实验

以 YOLOv8s 为基线模型, 对其围绕 3 个方面

表2 H-C2f模块实验结果

Table 2 Experimental results of H-C2f module

模型	mAP%	参数量	浮点运算量/10^9次	检测速度/(帧·s^{-1})
基线模型	72.8	11.1×10^6	28.8	109
H-C2f	75.7	10.2×10^6	26.6	116
C2f+SE	75.3	19.8×10^6	46.2	90
C2f+CBAM	75.2	11.2×10^6	28.5	105
C2f+Shuffle-Attention	75.6	11.1×10^6	28.4	108
C2f+CA	74.3	11.1×10^6	28.5	113
C2f+SimAM	73.6	11.1×10^6	28.4	109

进行改进，设计了轻量高效的EHH-YOLOv8s模型，为验证每项改进的有效性，本文在NEU-DET和Severstal数据集上进行5组消融实验。其中，M1表示将原骨干网络替换为EfficientViT-V0网络，M2表示将颈部的C2f模块替换为H-C2f模块，M3表示将颈部下采样操作替换为HWD模块。实验结果如表3所示。

从表3结果可知，在NEU-DET和Severstal数据集上，首先，将骨干网络替换为轻量级的EfficientViT网络，mAP值分别提升3.7%和3.3%，参数量和计算复杂度相应减少2.8×10^6和4.9×10^9次，检测速度分别提升17帧/s和5帧/s。表明轻量级的EfficientViT-V0网络能够减少冗余信息，优化模型参数量和计算复杂度，并提升检测速度。EfficientViT网络采用MSA机制，更好地处理钢材表面缺陷的跨尺度信息，提高检测能力。其次，将颈部的C2f模块替换为更轻量的H-C2f模块，参数量和计算量分别降低0.9×10^6和2.2×10^9次，检测精度和速度也有所提升。HAM模块通过空间和通道注意力机制，提升对细节特征信息的提取能力。同时采用快速一维卷积来实现跨通道信息交互，降低模型复杂度。然后，将颈部的下采样操作替换为HWD模块，mAP值分别提升2.0%和2.9%，模型体积和检测速度进一步得到优化。HWD模块采用Haar小波变换保留图像细节信息，扩大特征图感受野，其简单高效的设计节省了计算开销，提升训练速度。最后，同时添加3个模块，即本文所提EHH-YOLOv8s算法，mAP值分别提升6.9%和5.4%，检测速度分别提升27帧/s和15帧/s，参数量和计算量仅为基线模型的63.1%和54.2%，精确度分别达到77.1%和73.8%。消融实验结果表明，本文算法具有合理性，较好地平衡了检测精度、速度和轻量化。

同时，为验证Severstal数据集中各项性能指标是因数据增强还是本文算法的效果所致，在相同实验条件下，将本文算法和YOLOv8s算法在扩充前后数据集上进行实验对比，其中，SeverstalA、SeverstalB分别表示扩充前、后的数据集。实验结果如表4所示。从表4实验结果可知，Severstal数据集扩充前mAP值达到74.3%，而扩充后mAP值反而降低4.4%。由于数据增强前的样本过于简单，导致模型在训练过程中出现过拟合。相反，本文算法的mAP值在数据增强前后均得到提升，表明Severstal数据集检测精度的提升源于本文算法的效果，同时其他指标也得到一定优化。实验结果表明，数据增强在一定程度上能避免过拟合，本文算法具有良好的鲁棒性。

3.6 对比实验

为验证本文算法的优越性，在相同实验条件下，在NEU-DET和Severstal数据集上，将改进后的算法与其他算法进行对比，包括Faster R-CNN、SSD、RetinaNet、YOLOv3-tiny、YOLOv4-tiny、YOLOv5s、YOLOv5m、YOLOX-s、YOLOv7-tiny、YOLOv8s及文献[15-16,22-24]中的算法。由于Severstal数据集扩充后样本数量庞大，本文实验设备的计算能力和内存容量有限，而Faster R-CNN算法参数量过大，因此无法对其进行实验。此外，文献[15-16,22-24]算法主要针对NEU-DET数据集，因此，无法在Sever-

表3 消融实验结果

Table 3 Ablation test results

数据集	模型方法	mAP%	参数量	浮点运算量/10^9次	检测速度/(帧·s^{-1})	*P*%	*R*%
NEU-DET	YOLOv8s	72.8	11.1×10^6	28.8	109	72.8	70.9
	YOLOv8s+M1	76.5	8.3×10^6	23.9	126	68.8	76.1
	YOLOv8s+M2	75.7	10.2×10^6	26.6	116	70.0	73.5
	YOLOv8s+M3	74.8	10.7×10^6	27.9	113	73.5	70.6
	YOLOv8s+M1+M2+M3	79.7	7.0×10^6	15.6	136	77.1	74.6
Severstal	YOLOv8s	69.9	11.1×10^6	28.8	103	71.1	68.4
	YOLOv8s+M1	73.2	8.3×10^6	23.9	108	71.8	64.7
	YOLOv8s+M2	72.3	10.2×10^6	26.6	105	70.8	68.5
	YOLOv8s+M3	72.8	10.7×10^6	27.9	104	67.5	68.2
	YOLOv8s+M1+M2+M3	75.3	7.0×10^6	15.6	118	73.8	68.6

表 4　Severstal 数据集扩充前后实验对比

Table 4　Comparison of experiments before and after enrichment of Severstal dataset

数据集	模型方法	mAP%	参数量	浮点运算量/10^9次	检测速度/(帧·s^{-1})
SeverstalA	YOLOv8s	74.3	11.1×10^6	28.8	120
	EHH-YOLOv8s	78.6	7.0×10^6	15.6	125
SeverstalB	YOLOv8s	69.9	11.1×10^6	28.8	103
	EHH-YOLOv8s	75.3	7.0×10^6	15.6	118

stal 数据集上进行对比分析。表 5 展示了实验结果。

从表 5 实验结果可知, EHH-YOLOv8s 算法在 NEU-DET 和 Severstal 数据集上表现出显著优势。相较于上述所提经典模型算法, EHH-YOLOv8s 算法具有较高的检测精度, 其 mAP 值分别达到 79.7% 和 75.3%。此外, EHH-YOLOv8s 算法的检测速度分别为 136 帧/s 和 118 帧/s, 优于其他对比算法。相比之下, 文献 [16] 在 NEU-DET 数据集上具有较高的检测精度, 模型参数量和计算量也具备优势, 但检测精度与实验设备、数据集划分及实验参数设置也具有相关性, 其 49 帧/s 的检测速度较低, 实时性较差。文献 [15,22-23] 的检测速度相比于文献 [16] 有所提升, 但参数量随之增加, 检测精度也有所降低。而文献 [24] 算法占用过多的计算资源, 同时检测精度和速度较低。上述文献均无法较好地平衡检测精度、速度和轻量级。相反, 本文算法参数量和计算量仅为 7.0×10^6 和 15.6×10^9 次, 检测速度达到最优, 极大地满足了实时性需求。因此, EHH-YOLOv8s 算法可更好地平衡检测精度、速度和轻量级, 具有较高的通用性和实用价值。

3.7　可视化对比实验

为直观比较本文算法和基线模型在钢材表面缺陷检测性能上的差异, 使用 EHH-YOLOv8s 算法与 YOLOv8s 模型对 NEU-DET 和 Severstal 数据集

表 5　对比实验结果

Table 5　Compare experimental results

数据集	模型方法	mAP0.5/%	参数量	浮点运算量/10^9次	检测速度/(帧·s^{-1})
NEU-DET	Faster R-CNN	65.7	72.00×10^6	167.3	17
	SSD	61.0	41.10×10^6	145.3	41
	RetinaNet	69.5	18.30×10^6	45.1	15
	YOLOv3-tiny	46.5	8.60×10^6	12.9	142
	YOLOv4-tiny	54.6	5.90×10^6	16.1	128
	YOLOv5s	70.1	7.07×10^6	16.4	102
	YOLOX-s	71.8	8.00×10^6	21.6	46
	YOLOv7-tiny	68.7	6.02×10^6	13.1	108
	YOLOv8s	72.8	11.10×10^6	28.8	120
	文献[15]	74.1	23.90×10^6		75
	文献[16]	78.5	5.80×10^6	10.9	49
	文献[22]	77.2	7.03×10^6		115
	文献[23]	75.5	8.12×10^6		96
	文献[24]	75.2	90.80×10^6		30
	EHH-YOLOv8s(本文)	79.7	7.00×10^6	15.6	136
Severstal	SSD	65.3	41.10×10^6	145.3	12
	RetinaNet	71.2	18.30×10^6	45.1	11
	YOLOv3-tiny	56.4	8.60×10^6	12.9	117
	YOLOv4-tiny	59.6	5.90×10^6	16.1	103
	YOLOv7-tiny	68.7	6.02×10^6	13.1	108
	YOLOv5s	72.4	7.07×10^6	16.4	59
	YOLOv5m	73.2	21.00×10^6	50.3	52.6
	YOLOX-s	73.8	8.00×10^6	21.6	42
	YOLOv8s	69.8	11.10×10^6	28.8	103
	EHH-YOLOv8s(本文)	75.3	7.00×10^6	15.6	118

的6类缺陷类别进行定性分析，并通过可视化展示检测效果，如图9所示。可以看出，YOLOv8s算法存在多处漏检现象，难以满足工厂生产高质量钢材的需求。相较于基线模型，本文算法能够更准确地检测出有效的缺陷区域，极大地避免了漏检和误检情况，尤其是对于难以检测的裂缝、氧化皮和划痕类别。还可以看出，在NEU-DET数据集上，本文算法在裂缝、夹杂物、点蚀表面和氧化皮等类别上表现出较好的提升效果；而在Severstal数据集上，对裂缝、氧化皮和划痕3类缺陷的提升效果较为显著。此外，本文算法检测到的缺陷相较于基线模型具有较高的缺陷置信度。尽管存在部分漏检情况(如裂缝、点蚀表面和划痕缺陷)，但总体检测效果较为突出。可视化结果表明，EHH-YOLOv8s算法具备更强的浅层特征提取能力，对小目标的检测能力较强，能更好地避免漏检和误检情况，同时具备显著的检测性能优势。

同时，采用伽马变换对可视化中的6类缺陷图像亮度进行增强和减弱操作，模拟本文算法和基线模型在不同光照条件下的表现，图10展示了本文算法和基线模型在NEU-DET数据集上对不同亮度缺陷类别的检测效果。可见，基线模型在不同亮度的缺陷类别中存在多处漏检现象，尤其对裂缝、夹杂物和氧化皮的漏检较为明显。相比之下，EHH-YOLOv8s算法对各类缺陷在不同亮度条件下均能准确检测出有效区域，且漏检和误检情况较少。实验结果表明，本文算法在面对环境因素时具有较强的抗干扰和鲁棒性能力。

此外，对NEU-DET和Severstal数据集的6类缺陷，采用本文算法和基线模型生成相应的可视化热力图。为更直观地展示其效果，在实验中将热力图区域指定为缺陷区域，最终效果如图11所示。可以看出，热力图的可视化结果与图9的检测效果大致相符。对于部分较难检测的缺陷类别(如图中用红框标注的裂缝、夹杂物、点蚀表面和细小划痕等)，EHH-YOLOv8s算法相较于YOLOv8s算法，在缺陷区域的热力效果更佳，对于大部分缺陷区域热力效果更为显著，漏检率和误检率较低。表明本文算法在小目标检测方面表现出较强的性能，具备较高的鲁棒性。

3.8 泛化性验证

采用开放PCB数据集[25]对本文算法(EHH-YOLOv8s)进行泛化性能验证。该数据集共包含693张图像，涵盖焊点漏焊、缺口、开路、短路、毛刺和余铜等6种缺陷类型，每张图像均包含3处以上缺陷，如图12所示，其中，部分缺陷已用红圈标出。数据集划分、实验参数设置和评价指标与3.2节和3.3节保持一致。

将基线模型和本文算法在PCB板上进行实验对比，其实验结果如表6所示。从表6实验结果可知，EHH-YOLOv8s算法在PCB数据集上mAP值达到94.2%，相较于基线模型提升2.4%，检测速度提升24帧/s，表现出较强的实时性。表明本文算法具备更强的小目标检测能力，同时参数量和计算量

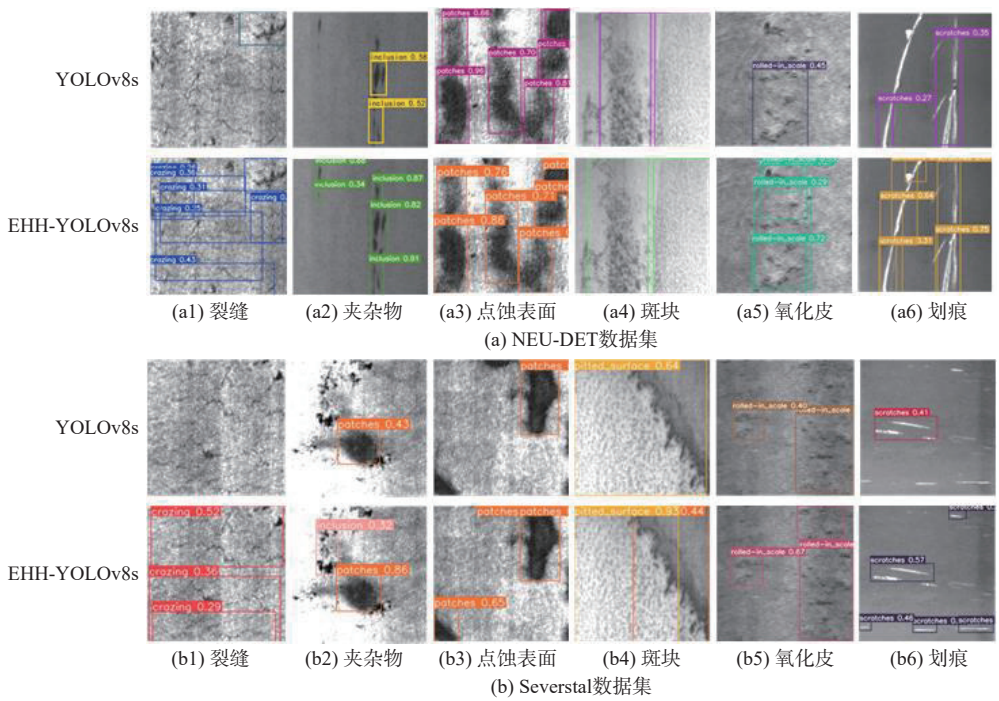


图9 本文算法与原模型可视化结果对比

Fig. 9 Visualization results comparison between the proposed algorithm and the original model

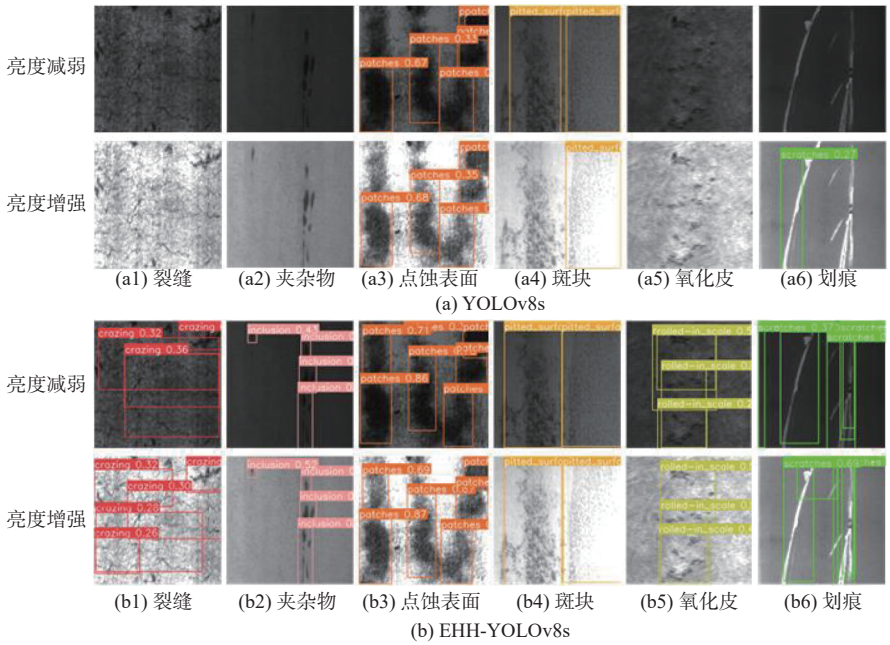


图 10　不同亮度检测效果对比

Fig. 10　Comparison of detection effect of different brightness

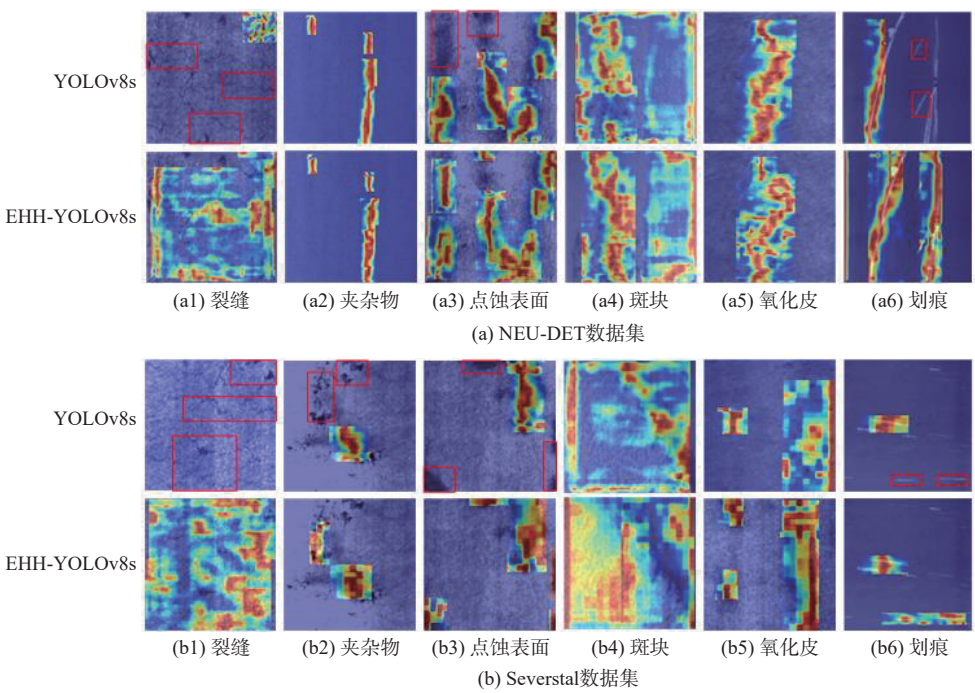


图 11　本文算法与原模型热力图检测效果对比

Fig. 11　Comparison of heat map detection effect between the proposed algorithm and original model

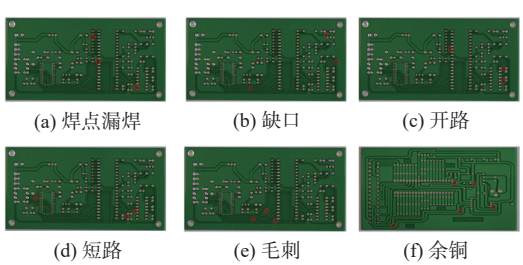


图 12　PCB 板表面各类缺陷图像

Fig. 12　Image of various defects on PCB surface

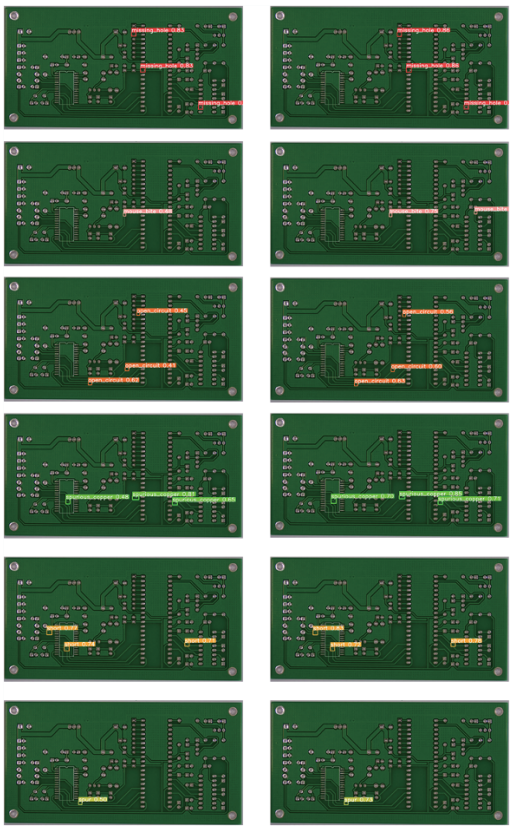
也具有显著优势。实验结果表明, 本文算法具备较强的泛化性能。

本文算法与基线模型在 PCB 板数据集上的可视化结果如图 13 所示。可以看出, 相比于基线模型, 本文算法能够更准确地检测出更多的缺陷目标。同时, 尽管 2 种算法都存在漏检情况, 但 EHH-YOLOv8 算法相对于 YOLOv8s 算法而言漏检情况较少, 且置信度更高, 进一步凸显了本文算法的优越性。

表6 PCB板数据集检测结果对比

Table 6 Comparison of detection results of PCB board dataset

模型方法	mAP%	参数量	浮点运算量/10^9次	检测速度/(帧·s^{-1})
YOLOv8s	91.8	11.1×10^6	28.8	101
EHH-YOLOv8s	94.2	7.0×10^6	15.6	125



(a) YOLOv8s (b) EHH-YOLOv8s

图13 本文算法和基线模型可视化结果对比

Fig. 13 Comparison of visualization results between the proposed algorithm and baseline model

4 结 论

针对现有缺陷检测模型参数量大、检测精度和实时性较差等问题，以及钢材表面缺陷检测效率低和精度差的挑战，本文提出EHH-YOLOv8s算法。

1）对Severstal数据集图像进行数据增强处理，以避免模型训练过拟合。

2）采用轻量级骨干网络，压缩模型体积，实现轻量化操作。

3）设计特征交互模块，引入HWD模块，保证减低模型体积的同时有效平衡检测精度和速度。

4）EHH-YOLOv8s算法在NEU-DET、Severstal和PCB板数据集上，在保持较高检测精度的同时，所需的计算资源更少，检测速度更快，较好地平衡了检测精度、速度和轻量级，且具备优越的泛化性能。

目前，EHH-YOLOv8s模型的部署尚在实验阶段，未来研究将考虑模型在实际部署时的寿命、维护和能源消耗等因素，并将其应用于工业钢材缺陷检测领域。

参考文献（References）

[1] 梁礼明, 阳渊, 朱晨锟, 等. 融合Mobile Vit和倒置门控编解码的视网膜血管分割算法[J]. 北京航空航天大学学报, 2025, 51(3): 712-723.
Liang L M, Yang Y, Zhu C K, et al. Fusion of Mobile Vit and inverted gated codec retinal vessel segmentation algorithm[J]. Journal of Beijing University of Aeronautics and Astronautics, 2025, 51(3): 712-723(in Chinese).

[2] 李彬, 汪诚, 丁相玉, 等. 改进YOLOv4的表面缺陷检测算法[J]. 北京航空航天大学学报, 2023, 49(3): 710-717.
Li B, Wang C, Ding X Y, et al. Surface defect detection algorithm based on improved YOLOv4[J]. Journal of Beijing University of Aeronautics and Astronautics, 2023, 49(3): 710-717(in Chinese).

[3] Singh S A, Desai K A. Automated surface defect detection framework using machine vision and convolutional neural networks[J]. Journal of Intelligent Manufacturing, 2023, 34(4): 1995-2011.

[4] Wang C M, Liu H. YOLOv8-VSC: lightweight algorithm for strip surface defect detection[J]. Journal of Frontiers of Computer Science and Technology, 2024, 18(1): 151-160.

[5] Liang F T, Zhou Y, Chen X, et al. Review of target detection technology based on deep learning[C]//Proceedings of the 5th International Conference on Control Engineering and Artificial Intelligence. New York: ACM, 2021: 132-135.

[6] Girshick R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: ACM, 2015: 1440-1448.

[7] Hu B, Wang J. Detection of PCB surface defects with improved Faster-RCNN and feature pyramid network[J]. IEEE Access, 2020, 8: 108335-108345.

[8] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2980-2988.

[9] Qian H M, Wang H L, Feng S, et al. FESSD: SSD target detection based on feature fusion and feature enhancement[J]. Journal of Real-Time Image Processing, 2023, 20(1): 2.

[10] Jiang P Y, Ergu D, Liu F Y, et al. A review of YOLO algorithm developments[J]. Procedia Computer Science, 2022, 199: 1066-1073.

[11] Hussain M. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection[J]. Machines, 2023, 11(7): 677.

[12] Zhou S Y, Zeng Y L, Li S C, et al. Surface defect detection of rolled steel based on lightweight model[J]. Applied Sciences, 2022, 12(17): 8905.

[13] Qin R B, Chen N J, Huang Y H. EDDNet: an efficient and accurate defect detection network for the industrial edge environment [C]//Proceedings of the IEEE 22nd International Conference on Software Quality, Reliability and Security. Piscataway: IEEE Press, 2022: 854-863.

[14] Yang L Y, Huang X B, Ren Y C, et al. Steel plate surface defect detection based on dataset enhancement and lightweight convolution

neural network[J]. Machines, 2022, 10(7): 523.

[15] 曹义亲, 伍铭林, 徐露. 基于改进 YOLOv5 算法的钢材表面缺陷检测[J]. 图学学报, 2023, 44(2): 335-345.
Cao Y Q, Wu M L, Xu L. Steel surface defect detection based on improved YOLOv5 algorithm[J]. Journal of Graphics, 2023, 44(2): 335-345(in Chinese).

[16] 张政超. 改进 YOLOv5 的轻量级带钢表面缺陷检测[J]. 计算机系统应用, 2023, 32(6): 278-285.
Zhang Z C. Lightweight strip steel defect detection based on improved YOLOv5[J]. Computer Systems and Applications, 2023, 32(6): 278-285(in Chinese).

[17] 蔡剑锋, 柏俊杰, 张雪, 等. 基于改进 Mask R-CNN 的金属板材表面缺陷检测[J]. 重庆科技学院学报 (自然科学版), 2023(2): 110-116.
Cai J F, Bai J J, Zhang X, et al. Research on surface defect recognition of metal sheet based on improved mask R-CNN[J]. Journal of Chongqing University of Science and Technology (Natural Sciences Edition), 2023(2): 110-116(in Chinese).

[18] Huang P L, Wang S H, Chen J Y, et al. Lightweight model for pavement defect detection based on improved YOLOv7[J]. Sensors, 2023, 23(16): 7112.

[19] Cai H, Li J Y, Hu M Y, et al. EfficientViT: multi-scale linear attention for high-resolution dense prediction[PP/OL]. arXiv (2024-02-06) [2024-06-01]. https://doi.org/10.48550/arXiv.2205.14756.

[20] Li G Q, Fang Q, Zha L L, et al. HAM: hybrid attention module in deep convolutional neural networks for image classification[J]. Pattern Recognition, 2022, 129: 108785.

[21] Xu G P, Liao W T, Zhang X, et al. Haar wavelet downsampling: a simple but effective downsampling module for semantic segmentation[J]. Pattern Recognition, 2023, 143: 109819.

[22] 周彦, 孟江南, 王冬丽, 等. 基于多尺度轻量化注意力的钢材缺陷检测[J]. 控制与决策, 2024, 39(3): 901-909.
Zhou Y, Meng J N, Wang D L, et al. Steel defect detection based on multi-scale lightweight attention[J]. Control and Decision, 2024, 39(3): 901-909(in Chinese).

[23] 赵林熔, 甄国涌, 储成群, 等. 基于 CBE-YOLOv5 的钢材表面缺陷检测方法[J]. 电子测量技术, 2023, 46(15): 73-80.
Zhao L R, Zhen G Y, Chu C Q, et al. Detection method of steel surface defects based on CBE-YOLOv5[J]. Electronic Measurement Technology, 2023, 46(15): 73-80(in Chinese).

[24] Guo Z X, Wang C S, Yang G, et al. MSFT-YOLO: improved YOLOv5 based on transformer for detecting defects of steel surface[J]. Sensors, 2022, 22(9): 3467.

[25] 周贤勇, 祝俊辉, 王圆, 等. 基于改进 YOLOv7 算法的 PCB 裸板缺陷检测[J]. 无线电工程, 2023, 53(12): 2791-2797.
Zhou X Y, Zhu J H, Wang Y, et al. PCB bare board defect detection based on improved YOLOv7 algorithm[J]. Radio Engineering, 2023, 53(12): 2791-2797(in Chinese).

Improved lightweight and efficient EHH-YOLOv8s algorithm for strip steel surface defect detection

LIANG Liming*, LONG Pengwei, LU Baohe, OU Yanyan, ZENG Lu

(School of Electrical Engineering and Automation, Jiangxi University of Science and Technology, Ganzhou 341000, China)

Abstract: To address the issues of low efficiency, and poor accuracy in current steel surface defect detection, and the challenges of large model parameters, low detection accuracy, and poor real-time performance of existing defect detection models, a lightweight steel defect detection algorithm based on YOLOv8s (EHH-YOLOv8s) is proposed. The reconstruction feature interaction module H-C2f, which successfully preserves spatial and channel features, suppresses redundant information, and improves detection accuracy and speed; the Haar wavelet downsampling (HWD) module, which expands the receptive field, reduces model complexity, and speeds up training and inference; and the lightweight EfficientViT network, which serves as the backbone network to reduce model complexity and better handle cross-scale feature information. Experimental validation on the NEU-DET and Severstal steel defect datasets shows that compared to the YOLOv8s algorithm, the EHH-YOLOv8 algorithm achieves an improvement in mean average precision (mAP) of 6.9% and 5.4%, respectively, with reductions in parameters and computations of 4.1×10^6 and 13.2×10^9, and achieves detection speed of 136 frames/s and 118 frames/s. These findings show that the algorithm strikes a superior balance between speed, lightweight design, and detection accuracy, offering a dependable reference for high precision and real-time performance on edge devices. Finally, validation on the PCB board defect dataset shows that the algorithm possesses good generalization performance.

Keywords: defect detection; lightweight YOLOv8s; EfficientViT; H-C2f; Haar wavelet downsampling

Received: 2024-06-12; **Accepted**: 2024-07-30; **Published Online**: 2024-08-08 08:50

URL: link.cnki.net/urlid/11.2625.V.20240807.1403.003

Foundation items: National Natural Science Foundation of China (51365017,61463018); General Project of Jiangxi Provincial Natural Science Foundation (20192BAB205084); Jiangxi Provincial Department of Education Science and Technology Research Key Project (GJJ170491,GJJ2200848); Jiangxi Province Graduate Innovation Special Fund (YC2022-S676)

* **Corresponding author**. E-mail: 9119890012@jxust.edu.cn