

基于改进 YOLOv8s-seg 的轻量化晶圆缺陷检测算法

余欣 缪佳欣 龙浩辰 郝彦哲 刘晓宇

(四川大学机械工程学院 成都 610000)

摘要: 晶圆缺陷检测是半导体制造流程中保障产品质量的关键环节,分析缺陷的分布区域与呈现模式,能够精准追溯生产工艺中的薄弱环节,为制造技术的优化提供重要依据。针对复杂场景下混合模式晶圆缺陷的识别检测高精度、实时性及轻量化的需求,提出一种基于改进 YOLOv8s-seg 的轻量化晶圆图缺陷检测算法。该算法首先在主干特征提取网络中引入上下文引导机制优化 C2f 模块,其次提出移位增强多分支多尺度融合特征金字塔网络(SEMF-FPN),提高模型的特征融合能力,最后引入轻量化非对称检测头思想优化头部网络。实验结果显示,改进模型边界框和掩码的平均精度均值 mAP@0.5:0.95 分别达 90.9% 和 84.5%,检测速度达到了 333.3 fps,在精度损失不超过 0.5% 的前提下实现了参数量和计算量 54.84% 和 37.18% 的大幅减少,模型轻量化、精度高且检测快速,满足工业生产的需要,具有一定的工程应用价值。

关键词: 半导体制造;晶圆图;缺陷检测;轻量化

中图分类号: TP391.41;TN405 **文献标识码:** A **国家标准学科分类代码:** 520.6040

Lightweight wafer defect detection algorithm based on improved YOLOv8s-seg

Yu Xin Miu Jiaxin Long Haochen Hao Yanzhe Liu Xiaoyu

(School of Mechanical Engineer, Sichuan University, Chengdu 610000, China)

Abstract: Wafer defect detection is a critical step in ensuring product quality within semiconductor manufacturing processes. Analyzing the distribution areas and presentation patterns of defects enables precise tracing of weak links in production processes, providing essential evidence for optimizing manufacturing technologies. To address the demands for high precision, real-time processing, and lightweight solutions in identifying hybrid wafer defects under complex scenarios, this paper proposes a lightweight wafer map defect detection algorithm based on an improved YOLOv8s-seg. The algorithm first introduces a context-guided mechanism within the backbone feature extraction network to optimize the C2f module. It then proposes a shift-enhanced multi-branch & scale fusion feature pyramid network (SEMF-FPN) to enhance the model's feature fusion capabilities. Finally, it incorporates lightweight asymmetric detector head concepts to optimize the head network. Experimental results demonstrate that the improved model achieves mean average precision (mAP@0.5:0.95) of 90.9% for bounding boxes and 84.5% for masks, with a detection speed of 333.3 fps. It significantly reduces parameters and computational load by 54.84% and 37.18%, respectively, while maintaining accuracy loss below 0.5%. This lightweight, high-precision and fast-detection model meets industrial production requirements and holds considerable engineering application value.

Keywords: semiconductor manufacturing;wafer map;defect detection;weight reduction

0 引言

在当前互联网技术与数字经济深度融合、蓬勃发展的背景下,芯片半导体产业已成为支撑现代工业体系的核心基石,其技术革新速度与生产效率水平直接决定着一个国家在全球科技竞争中的地位^[1]。作为半导体集成电路芯片

制造的基础材料,晶圆是由硅等半导体材料制成的圆形薄片,广泛应用于电子通讯、新能源、航空航天、人工智能等关键领域,其产能规模与良率水平对整个半导体产业链的发展质量具有决定性影响^[2]。随着芯片设计复杂度的持续提升,越来越多的精密电路被集成到晶圆表面,这不仅让芯片功能更强大、存储容量更大,也使得晶圆制造工艺变得异常

复杂——从光刻、蚀刻到沉积,每一个环节的微小偏差都可能导致缺陷产生,缺陷的出现概率与种类也随之大幅增加^[3]。因此,晶圆缺陷检测成为半导体制造流程中保障产品质量的关键环节;通过分析缺陷的分布区域与呈现模式,能够精准追溯生产工艺中的薄弱环节,为制造技术的优化提供重要依据。传统的人工目视检测方法因精度有限、效率低下、成本高昂,且检测标准受人为因素影响较大,已被自动化检测技术取代;早期的机器视觉技术通过人工设计特征提取与模式识别算法实现检测,虽然性能有一定的提升,但难以应对复杂缺陷模式^[4]。

近年来,深度学习凭借自动提取丰富特征的优势,逐渐成为晶圆缺陷检测的主流技术,为解决复杂缺陷识别问题提供了全新路径^[5]。2018 年 Nakazawa 等^[6]提出使用 3 个卷积层和 2 个全连接层构成的卷积神经网络对晶圆缺陷模式进行分类和图像检索,通过 28 600 张合成的晶圆图训练模型,分类准确率达 98.2%;2020 年 Wang 等^[7]提出了一种基于可变形卷积网络(deformable convolutional network, DC-Net)的混合类型晶圆缺陷模式分类识别方法,在 MixedWM38 数据集上平均识别准确率达 93.2%,显著优于传统卷积神经网络(convolutional neural network, CNN)模型;2021 年 Kim 等^[8]提出基于单阶段检测器(single shot multiBox detector, SSD)的混合类型晶圆缺陷检测框架,通过自动生成缺陷边界框并结合多尺度特征图实现检测,较传统 CNN 模型在混合缺陷检测 F1 分数提升 23.5%,并能同时输出缺陷位置信息;同年,Chiu 等^[9]提出一个集成了基于 Mask R-CNN 的实例分割模型以及复制-粘贴和旋转数据增强的晶圆缺陷检测框架,单类准确率为 97.7%,混合类型准确率为 82%;2022 年 Cha 等^[10]提出一种基于残差块和注意力机制的改进 U-Net 模型来识别分割整体混合缺陷,准确率达 98%,交并比(intersection over union, IoU)为 64.4%;李洋^[11]提出了一种基于改进 Mask R-CNN 实例分割模型的晶圆缺陷检测方法,边界框和掩码的平均精度均值(mean average precision, mAP)达 60.4%和 40.5%;2023 年 Yan 等^[12]提出一种基于语义分割的晶圆混合缺陷模式识别方法,在合成数据集上的平均准确率达 97%以上,在 MixedWM38 数据集的平均准确率上达 95.8%;2024 年邓广远等^[13]提出了一种基于通道混洗和深度可分离卷积的轻量化深度神经网络混合型晶圆缺陷识别方法,在 MixedWM38 数据集上的正确率达 97.32%;2025 年 Choi 等^[14]提出了基于活动轮廓的轻量级深度网络(active contour-based lightweight depthwise network, AC-LDN)模型,用于多缺陷晶圆图图案的分类,平均分类准确率超 98%。

当前国内外相关研究已从分类任务逐步向目标检测、语义分割及实例分割任务过渡,其研究对象也从单一缺陷模式拓展至混合缺陷模式^[15-16]。然而,当前实例分割领域的研究较为匮乏,且均将混合缺陷视作整体独立类别,未对

混合缺陷中所包含的基本缺陷的位置与形态信息进行独立分割识别。因此,针对混合模式晶圆缺陷的高效检测仍需深入研究。本文在 YOLOv8s-seg 的基础上提出一种改进 YOLOv8s-seg 的轻量化混合缺陷模式晶圆缺陷检测算法,通过优化特征提取网络中的 C2f 模块、重设计及优化颈部网络、改善检测头设计对原始网络模型进行了优化改进,在保证较高检测精度的同时实现了轻量、快速、准确的目的,满足工业检测中对高效、精准缺陷检测的需求,同时半导体晶圆表面缺陷检测与国家的“四个面向”紧密相关,对于推动科技进步、经济发展、满足国家战略需求以及保障人民健康都具有重要的意义。

1 YOLOv8 算法概述

YOLOv8^[17]是一种单阶段端到端目标检测算法,在继承 YOLO 系列高效推理核心优势的基础上,通过架构优化与功能扩展,实现了检测、分类、分割等多任务覆盖,其丰富的模型变体(v8-cls 分类模型、v8-seg 分割模型、v8-pose 姿态估计模型、v8-obb 旋转目标检测模型等)可灵活适配不同场景需求,尤其在工业视觉任务中展现出极强的适用性^[18-21],能够实现高精度与高速度的平衡,对于晶圆缺陷检测任务具有较强的应用价值。

YOLOv8-seg 的整体架构由主干网络(Backbone)、颈部网络(Neck)和检测头(Head)3 部分组成,提供 YOLOv8n/s/m/l/x 等多尺度模型,通过调整网络深度(层数)、宽度(通道数)和输入分辨率可实现速度与精度的灵活权衡。本文选择 YOLOv8s-seg 网络模型作为基础模型,其网络架构如图 1 所示。

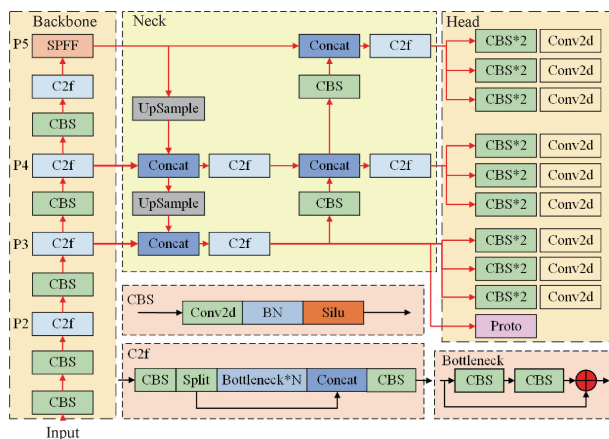


图 1 YOLOv8s-seg 网络架构

Fig. 1 YOLOv8s-seg network architecture

主干网络负责多尺度特征提取,包含复合卷积模块(convolution batchnormalization silu, CBS)、两卷积跨阶段部分瓶颈快速实现模块(faster implementation of cross stage partial bottleneck with 2 convolutions, C2f)与快速空间金字塔池化模块(spatial pyramid pooling fast, SPPF)。

CBS 由标准卷积、归一化及激活函数组成, C2f 模块通过卷积模块引入残差连接, 将特征分为两部分: 一部分直接参与残差连接, 另一部分经 n 个瓶颈模块 (Bottleneck) 处理, 每个中间 Bottleneck 的输出均与直接参与残差连接部分进行按通道拼接操作 (channel-wise concatenation, Concat), 借助分支结构实现了梯度与特征的高效流动, 与 YOLOv5 中的三卷积跨阶段部分瓶颈模块 (cross stage partial bottleneck with 3 convolutions, C3) 相比, 在相同计算量下提升了特征提取的丰富度; SPPF 模块则通过 3 个 5×5 的 MaxPool 串行计算结合残差连接, 再经 Concat 操作整合特征, 实现多尺度信息的高效捕捉。

颈部网络负责特征融合, 采用路径聚合特征金字塔网络 (path aggregation feature pyramid network, PAFPN) 结构, 在特征金字塔网络 (feature pyramid network, FPN) 的基础上增加一层自底向上的融合架构, 进一步强化全局语义信息的提取与聚合, 助力网络更深入理解图像的整体语义结构。

头部网络采用双分支架构, 协同完成实例分割任务。检测分支由基础 YOLOv8 检测头扩展而来, 保留多尺度特征处理逻辑与解耦合头 (decoupled head) 设计, 将分类、目标定位及掩码系数 (mask coefficient) 预测的特征路径分

离, 使各任务在优化过程中专注于自身目标、提升精度; 分割分支通过 80×80 尺度特征图生成原型掩码 (prototype mask), 检测分支输出的掩码系数与原型掩码经加权融合后, 生成像素级的最终分割掩码, 实现检测与分割任务的高效联合。

2 改进算法介绍

晶圆缺陷检测任务具有缺陷尺度变化大、形态多样、边缘精度要求高等独特挑战, 且由于晶圆由一个个晶粒组成, 图像具有一定的周期性结构, 因此改进模型需要具有多尺度感知能力、能够捕捉不同形状和方向的缺陷以及能够捕捉一定的空间规律的能力。针对晶圆缺陷检测中对模型轻量化与检测精度的双重需求, 本文基于 YOLOv8s-seg 算法进行多维度改进, 构建轻量化晶圆混合模式缺陷检测模型。改进聚焦主干网络、颈部网络与检测头 3 大核心模块: 通过引入轻量级上下文引导机制优化主干特征提取模块, 设计移位增强多分支多尺度融合特征金字塔网络 (shift-enhanced multi-branch & scale fusion feature pyramid network, SEMF-FPN) 增强颈部特征融合能力, 采用轻量化非对称检测头提升检测端的效率与精度, 最终实现晶圆缺陷的高效、精准检测。改进后的网络结构图如图 2 所示。

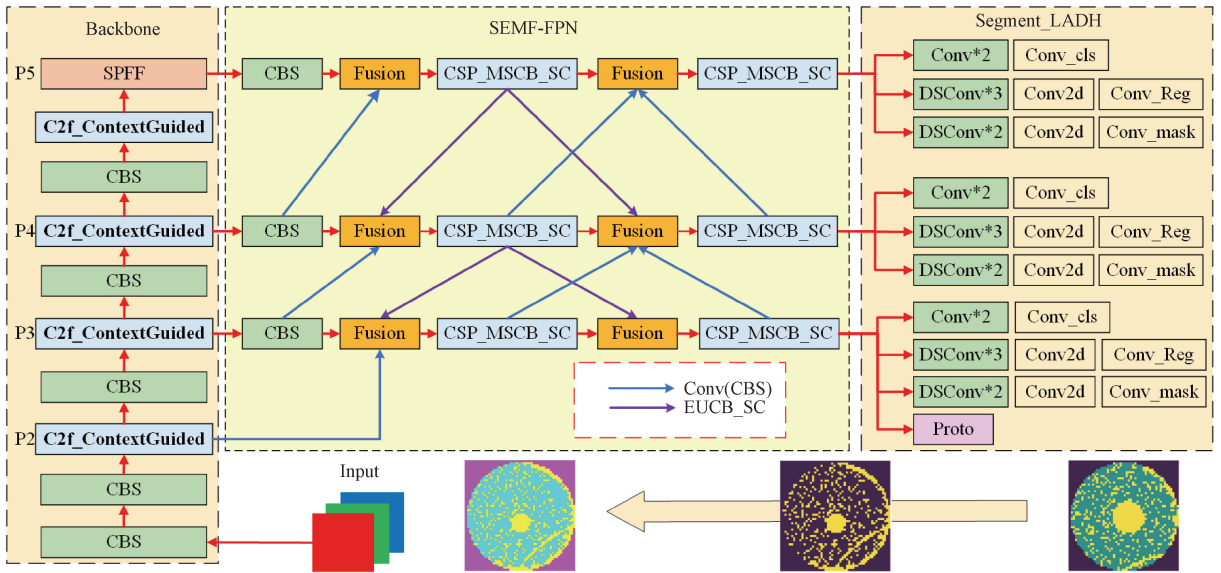


图 2 改进模型网络结构

Fig. 2 Improved model network structure

2.1 主干网络 C2f_ContextGuided 模块

YOLOv8 主干网络以 C2f 模块作为核心特征提取单元, 如图 3 所示。其中默认的 Bottleneck 单元主要依赖常规卷积进行局部特征建模, 其感受野局限于局部区域, 对复杂场景下的长距离依赖关系、多尺度目标关联及全局语义信息的捕捉能力有限。为弥补 C2f 模块在上下文建模能力上的不足, 强化主干网络对多尺度、多维度信息的融合能力, 本文提出将内容引导网络 (context grided network, CG-Net) 中的上下文引导机制引入 C2f 模块设计, 构建 C2f_ContextGuided 模块。

CG-Net) 中的上下文引导机制引入 C2f 模块设计, 构建 C2f_ContextGuided 模块。

C2f_ContextGuided 模块的结构组成如图 4 所示, 其中 Context Guided Block^[22] 由 4 部分组成: 局部特征提取器 (f_{loc}) 使用常规卷积 (standard convolution) 进行局部特征提取, 通过卷积操作捕捉输入特征图中的局部信息, 如边缘、纹理等细节特征, 为后续的上下文理解提供基础; 周围上下文特征提取器 (f_{sur}) 使用膨胀卷积 (dilated

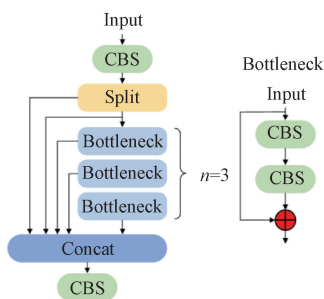


图 3 C2f 模块结构

Fig. 3 Structure of C2f module

convolution)进行周围上下文特征提取,膨胀卷积能够在不增加参数数量的情况下扩大感受野,从而捕捉更大范围的上下文信息;联合特征提取器(f_{join})通过简单的拼接层将局部和周围特征结合在一起,形成联合特征,并应用批归一化(batch normalization)和 SiLu 激活函数帮助加速训练和增强模型的非线性表达能力;全局特征提取器(f_{glo})通过全局平均池化层提取特征,然后连接两个全连接层生成一个权重向量用于指导联合特征的融合,帮助模型关注重要的特征,并在融合过程中增强有用信息。

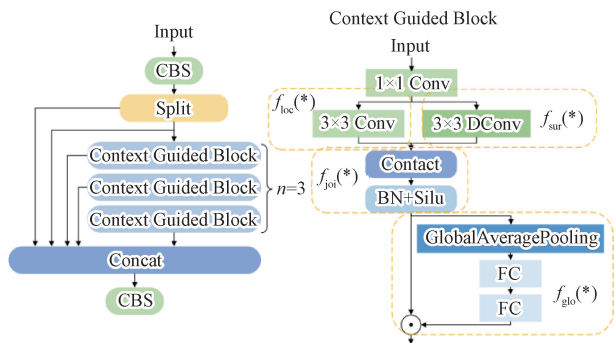


图 4 C2f_ContextGuided 模块结构

Fig. 4 Structure of C2f_ContextGuided module

2.2 颈部网络 SEMF-FPN 模块

原始 YOLOv8-seg 模型的颈部采用 PAFPN 结构进行多尺度特征融合,但在晶圆缺陷检测场景中存在一定局限。首先,其分支结构采用固定尺寸的卷积核,感受野难以适配较大的尺度跨度,且重复的卷积操作造成参数与计算量冗余,限制了模型的轻量化部署;其次,其特征融合机制仍有提升空间,且依赖简单的 Concat 操作,该策略在合并不同来源特征时未加筛选,易引入信息冗余,从而抑制了网络对关键缺陷特征的代表能力。

针对上述问题,本文提出 SEMF-FPN。该颈部网络以多分支辅助特征金字塔网络(multi-branch auxiliary feature pyramid network, MAFPN)全局异构核机制^[23]和双向特征金字塔网络(bi-directional feature pyramid network, BiFPN)多尺度加权融合思想^[24]为理论基础进行了全面的重设计,并在基于移位运算的二进制多层感知器

模块^[25]启发下提出一个无参的移位通道混合(Shift_channel_mix)模块用于优化多尺度卷积模块(multi-scale convolution block, MSCB)和高效上采样模块(efficient up-convolution block, EUCB)^[26],据此提出基于移位通道混合的跨阶段多尺度卷积模块(CSP_MSCB_SC)和高效上采样模块(EUCB_SC),分别用于替代原 YOLOv8-seg 颈部网络中的 C2f 特征提取模块和 Upsample 插值上采样模块,从而进一步提升颈部网络的特征表达与融合能力。

其中,Shift_channel_mix 模块旨在无额外参数的前提下强化特征的空间关联性与通道间信息交互。其具体流程为首先将特征图沿通道维度分为 4 个部分(x_1, x_2, x_3, x_4),然后分别在水平方向(宽度维度)和垂直方向(高度维度)上施加正负方向的循环移位, x_1, x_2 在高度维度进行正向和负向偏移, x_3, x_4 在宽度维度进行正向和负向偏移,最后拼接组合偏移后的特征块。对于晶圆缺陷,移位增强操作可以模拟缺陷在空间中的方向性,例如划痕缺陷可能是水平或垂直的,并通过混合不同方向的空间信息,增强模型对空间方向性和晶圆缺陷形态多样性的捕捉能力,提升特征判别性,并适配晶圆缺陷分割对边缘细节的精准捕捉需求。

SEMF-FPN 颈部网络结构如图 5 所示,由 1×1 CBS 卷积复合模块、Fusion 操作、基于移位通道混合的跨阶段多尺度卷积模块(CSP_MSCB_SC)和高效上采样模块(EUCB_SC)、Conv 下采样卷积模块构成。

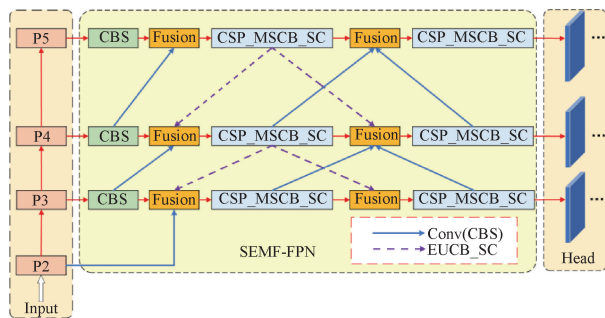


图 5 SEMF-FPN 模块结构

Fig. 5 Structure of SEMF-FPN module

输入图像首先通过主干网络的 4 个阶段(P2、P3、P4、P5)逐步提取特征,接着 P3、P4、P5 分别通过一个 1×1 卷积复合模块对齐通道信息,以适配后续元素级加权融合 Fusion 操作。SEMF-FPN 借鉴 BiFPN 的加权思想,引入快速归一化加权融合策略,用加权求和的 Fusion 操作替代了原始 YOLOv8-seg 颈部网络中的通道拼接 Concat 操作,Concat 操作会线性增加通道数,而 Fusion 操作通过加权求和保持通道数不变,显著降低了模型复杂度和计算量,有利于晶圆缺陷检测任务的轻量化,同时晶圆缺陷检测中,不同尺度的特征对最终检测的贡献不同,Fusion 加权融合通过先归一化权重、再加权融合的方式动态学习并分配不同来源特征的权重,从而达到更佳的多尺度目标特征融合

效果。归一化加权融合计算公式为^[24]:

$$O = \sum_i \frac{\omega_i}{\epsilon + \sum_j \omega_j} I_i \quad (1)$$

其中, I_i 为输入特征, O 为输出特征, i 和 j 都是输入特征图的索引, ω_i 和 ω_j 表示对应索引特征图的可学习权重, ϵ 为防止数值不稳定的小量。在本文实验中,可学习权重 ω_i 和 ω_j 的初始化为全 1 张量,并通过 ReLU 激活函数确保非负性。 ϵ 设置为 1×10^{-4} , 以在保证数值稳定性的同时避免过小值对权重分布的干扰。训练过程中,权重通过反向传播自动优化。

加权融合后的特征被送入基于移位通道混合的跨阶段多尺度卷积模块(CSP_MSCB_SC)进行特征提取,CSP_MSCB_SC 模块结构如图 6 所示。

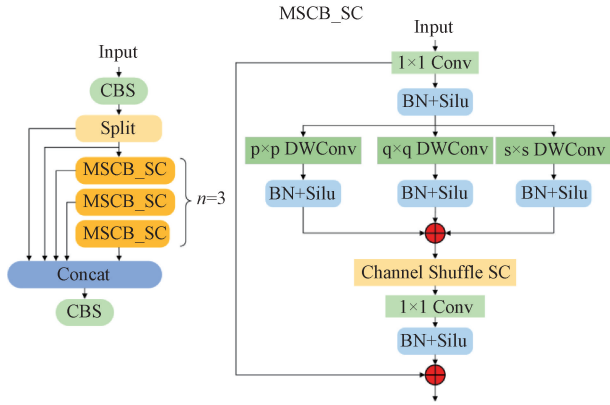


图 6 CSP_MSCB_SC 模块结构

Fig. 6 Structure of CSP_MSCB_SC module

该模块集成了多尺度深度卷积(MSCB)以捕捉不同规模的缺陷特征,适应了晶圆缺陷尺寸变化大的特点。通过并行不同尺寸深度卷积增强多尺度特征捕捉能力,并使用融合 Shift_channel_mix 模块改进的通道打乱模块 Channel Shuffle SC 强化特征图的空间关联性与通道间信息交互,弥补传统卷积局部感受野的局限,在轻量化前提下提升对晶圆缺陷特征的判别性,且通过跨阶段部分连接(cross stage partial connection, CSP)的多分支结构设计,在减少计算量的同时,促进梯度流动,使训练更稳定。

与此同时,SEMF-FPN 基于全局异构核选择机制^[27]设计。全局异构核选择机制是指在网络结构中的不同分辨率特征层上(如 P3、P4 和 P5)也使用不同大小卷积核的策略,自适应地扩大整个网络的感受范围。在本文中,通过在 P3、P4 和 P5 不同尺度上的 CSP_MSCB_SC 模块分别适配 $[1, 3, 5]$ 、 $[3, 5, 7]$ 、 $[5, 7, 9]$ 的卷积核实现全局异构核选择机制,使网络能自适应地调整感受野,应对晶圆多尺度缺陷检测,从而能够进一步高效、精准地处理各种尺度目标。

在上采样方面,本文通过改进 EUCB 模块提出 EUCB_SC 模块,其结构如图 7 所示。传统上采样方法存在明显瓶

颈:转置卷积易产生“棋盘效应”,破坏缺陷边缘的空间一致性;插值上采样因缺乏特征学习导致输出模糊,难以保留微小缺陷细节。EUCB 模块虽通过“上采样-深度卷积修复-通道洗牌- 1×1 聚合”流程实现轻量化改进,但特征交互依赖通道洗牌,对空间维度不同位置(如缺陷边缘与内部区域)的关联强化不足。为此,本文在 EUCB 的基础上使用融合 Shift_channel_mix 模块改进的通道打乱模块 Channel Shuffle SC,在轻量化的同时增强了空间和通道维度的特征交互,也有助于恢复晶圆缺陷的边缘细节。

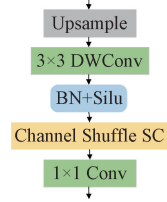


图 7 EUCB_SC 模块结构

Fig. 7 Structure of EUCB_SC module

2.3 检测头 Segment_LADH 模块

YOLOv8 原检测头虽采用解耦设计,但在晶圆缺陷检测任务中,其轻量化程度与对细粒度特征的捕捉能力仍有提升空间。本文基于轻量化非对称检测头(light weight asymmetric detection head, LADH)思想^[28]提出一种专为分割任务设计的 Segment_LADH 模块以替代原检测头。

该模块的核心设计如图 8 所示,其核心在于一种非对称路径设计。P3、P4、P5 分别对应 3 个分支:分类分支(cls)、回归分支(reg)、掩码系数分支(mask),且仅在 P3 层有一个原型掩码生成分支(proto)。分类分支通过 2 个 1×1 卷积输出维度为 $H \times W \times nc$ 的类别预测图,回归分支通过 3 个 3×3 深度可分离卷积和 1 个 1×1 卷积输出维度为 $H \times W \times 64$ 的回归特征图,掩码系数分支采用 2 个 3×3 深度可分离卷积和 1 个 1×1 卷积输出维度为 $H \times W \times 32$ 的掩码系数预测图,原型掩码生成分支通过 80×80 尺度的特征图生成一个大小为 $(1, 32, 80, 80)$ 特征图。掩码系数与原型掩码经加权融合后,生成像素级的最终分割掩码。

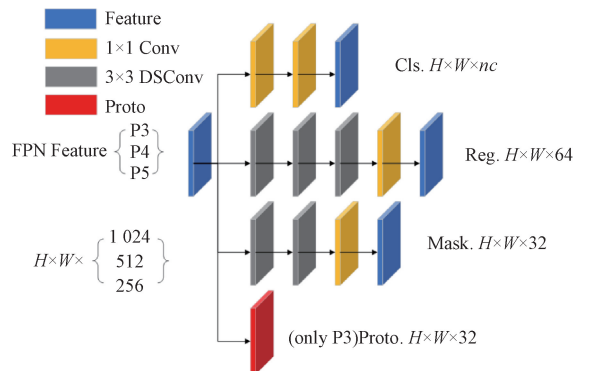


图 8 Segment_LADH 模块结构

Fig. 8 Structure of Segment_LADH module

Segment_LADH 模块通过非对称路径设计,将分类、回归与掩码预测任务在特征提取阶段进行分离,增强了各任务的专一性并减少了参数冗余,同时采用轻量级激活函数与深度可分离卷积,在保证对晶圆缺陷细粒度特征捕捉能力的前提下,显著降低检测头的计算负载,提升了模型整体运行效率。

3 实验结果与分析

3.1 数据集预处理

本文所使用的数据源是 MixedWM38 数据集,由东华大学人工智能研究院构建,是一个专注于晶圆图混合模式缺陷分类识别的专业数据集,其构建旨在为晶圆缺陷成因分析及制造工艺优化提供关键数据支持。该数据集包含 38 个缺陷模式:具体分为 1 个无故障模式、8 个单一缺陷模式、13 个 2 混合型模式、12 个 3 混合型缺陷模式和 4 个 4 混合型缺陷。该数据集的原始数据来源于某晶圆制造厂,通过对晶圆片上每个晶粒的电学性能测试采集获得大量真实晶圆图,由于实际采集的不同类型晶圆图存在显著的数量分布不平衡问题,该研究团队采用对抗生成网络(generative adversarial network, GAN)生成部分补充数据以平衡各类别样本量,最终形成包含约 38 000 张混合模式晶圆图的数据集。

MixedWM38 数据集中,晶圆图为单通道 0-1-2 三值图,以 npy 格式进行保存,其中 0 代表背景部分,1 代表通过电学测试的正常晶粒,2 代表未通过电学测试的失效晶粒。晶圆缺陷标签为 8 个基本缺陷类别:中心(Center)、圆环(Donut)、边缘局部(edge location, EL)、边缘环(edge ring, ER)、局部(location, LOC)、近满(near full, NF)、划痕(scratch, S)、随机(Random),如图 9 所示。而混合缺陷是由除 NF 和 Random 以外的 6 种基本缺陷组成的 2、3、4 三种混合型缺陷模式,如图 10 所示。

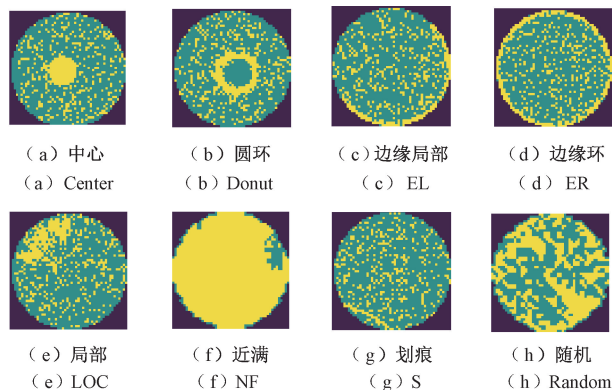


图 9 8 个晶圆图常见单一基本缺陷

Fig. 9 Eight common wafer map single-type defects

本文将晶圆混合缺陷模式识别定义为实例分割任务,而 MixedWM38 数据集原始仅包含缺陷模式的分类标签,未提供实例分割任务所需的缺陷区域像素级掩码和边界

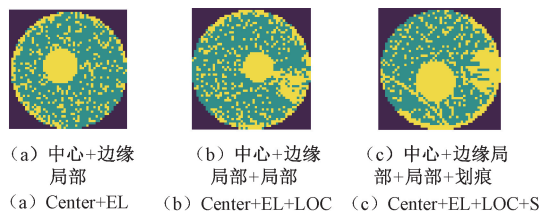


图 10 2、3、4 三种混合缺陷模式

Fig. 10 Three mixed-type defect patterns of 2, 3 and 4

框标注。为此,本文参考 Yan 等^[12]提出的方法在 MixedWM38 数据源基础上构建了本实验所需的数据格式,依据该数据集晶圆图特性设计算法生成了对应的缺陷掩码和边界框标注。

针对单一缺陷类别,首先单通道 0-1-2 三值晶圆图数据中的背景和正常区域像素设置为 0,缺陷区域像素设置为 255,得到只包含缺陷区域的掩膜图,然后使用连通区域分析算法提取最大连通区域获得不含随机噪声的掩膜图,如图 11 所示。最后根据不同的缺陷类型对掩膜图中的缺陷区域赋值,8 种基本缺陷的像素标签值分别为 1~8。



图 11 最大连通区域提取(以圆环缺陷为例)

Fig. 11 Extraction of the largest connected region (taking donut defects as an example)

而针对混合缺陷掩膜生成,由于直接的叠加单缺陷晶圆图会导致过多的随机缺陷,因此使用将单缺陷晶圆图与其它缺陷标签叠加的方法,叠加时各缺陷标签采用随机旋转的方式以增加混合缺陷晶圆图的多样性,同时设置缺陷优先级以减小对形状更容易破坏的缺陷的标注影响,混合缺陷掩膜生成效果如图 12 所示。

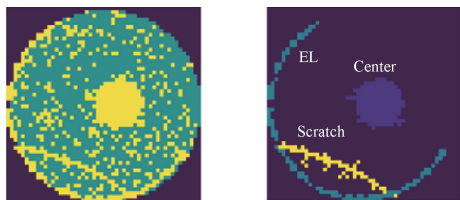


图 12 混合缺陷晶圆图及其多标签掩膜

Fig. 12 Mixed defect wafer map and its multi-label mask

由于 YOLOv8 要求输入为 RGB 三通道形式且需要的数据标签格式为 txt,本文将单通道 0-1-2 三值 npy 图像数据映射到 $[255, 0, 255]$, $[0, 255, 255]$, $[255, 255, 0]$ 的 RGB 格式的图像文件,并使用官方给出的 mask 分割掩码转 YOLO 分割 convert_segment_masks_to_yolo_seg 函数得到 txt 形式的标签文件,最终呈现的数据标注结果如图 13 所示。作为实例分割任务模型时, YOLOv8s-seg 同

时生成了掩码标注和矩形边界框标注。

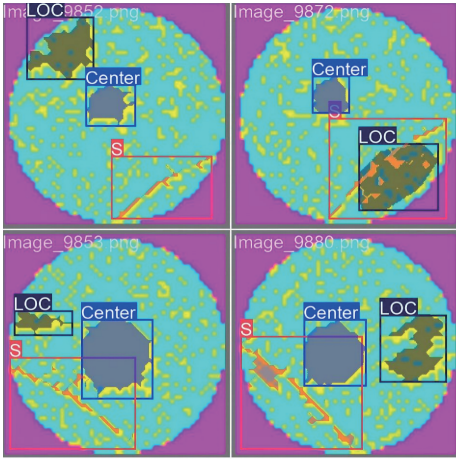


图 13 数据标注结果
Fig. 13 Dataset annotation results

本文所构建晶圆图混合缺陷模式实例分割数据集共计 38 000 张,按照 7:1:2 的比例对所构建数据集进行数据划分,分别得到训练集图像 26 600 张、验证集 3 800 张、测试集 7 600 张,满足模型训练和模型评估测试的需求。

3.2 实验环境及参数设置

为了保证实验数据的准确性,本文所有实验在相同的环境下进行,实验环境配置如表 1 所示。

表 1 实验环境配置

Table 1 Configuration for the experimental environment	
名称	环境配置
操作系统	Ubuntu 22.04
CPU	64 vCPU Intel(R) Xeon(R) Platinum 8481C
GPU	NVIDIA GeForce RTX 4090D(24 GB) * 4
内存	320 G
Python 版本	3.10
PyTorch 版本	2.1.0
CUDA 版本	12.1

为了合理调用配置硬件资源,本文实验超参数如表 2 所示。

3.3 评价指标

为评估所提晶圆缺陷混合模式检测算法的有效性,以平均精度均值(mAP)、每秒检测帧数(frames per second, FPS)、参数量(Parameters)、计算量(GFLOPs)作为模型的核心评价指标。

其中,mAP 是衡量多类别目标检测模型综合性能的核心指标。通过计算每个类别的精度后取平均值来衡量模型的性能。mAP 值越高,模型的性能就越好。其定义为:

表 2 实验参数配置

Table 2 Configuration of experimental parameters

超参数	数值
训练图片尺寸(Image size)	640×640
训练次数(Epochs)	200
批次训练数据量(Batch size)	256
进程数(workers)	4×4
优化器(optimizer)	SGD
动量因子(comentum)	0.937
初始学习率(lr0)	0.01
周期学习率(lrf)	0.01
权重衰减系数(weight decay)	0.000 5

$$mAP = \frac{\sum_{i=1}^k AP}{k} \tag{2}$$

其中, k 表示数据集中类别的数量。mAP@0.50 表示当交并比 IoU 设置为 0.50 时的平均精度,mAP@0.50:0.95 表示当 IoU 设置为从 0.50~0.95(以 0.50 的步长取值)时的平均精度。

$$FPS = \frac{1\,000\text{ ms}}{\text{speed}} = \frac{1\,000\text{ ms}}{\text{Pre} + I + \text{Post}} \tag{3}$$

其中,FPS 用来评估检测速度,FPS 值越高,单位时间内检测的帧数越高,模型检测速度越快。speed 单位为 ms,等于图像预处理、推理与后处理时间之和。

3.4 消融实验

为了验证改进模型设计的合理性与有效性,本文将输入图像尺寸设置为 640×640,在原始 YOLOv8s-seg 网络模型的基础上,分别引入 C2f_ContextGuided 模块、MF-FPN 特征融合网络结构、SEMF-FPN 特征融合网络结构、Segment_LADH 检测头进行实验,通过消融实验验证不同改进方法对模型性能的影响。消融实验选择的观察指标为边界框任务和分割掩码任务下的 mAP@0.5 和 mAP@0.5:0.95、检测速度(speed)、模型参数量(Parameters)和计算量(GFLOPs)。所有实验均在相同软硬件环境配置下进行,为了保证实验的真实性,在训练网络过程中,未使用任何预训练权重。

消融实验设计如表 3 所示。序号 1 表示原 YOLOv8s-seg 模型;序号 2 表示在主干网络引入 C2f_ContextGuided 模块;序号 3 表示使用不含移位增强的多分支多尺度融合特征金字塔网络 MF-FPN;序号 4 表示使用移位增强多分支多尺度融合特征金字塔网络 SEMF-FPN;序号 5 表示采用轻量化非对称检测头网络 Segment_LADH 模块;序号 6 表示同时采用 C2f_ContextGuided 模块和 SEMF-FPN 颈部网络;序号 7 表示同时使用 SEMF-FPN 颈部网络和非对称检测头网络 Segment_LADH 模块;序号 8 表示同时使用对主干网络、颈部网络、头部网络 3 处的改进优化模块。

表 3 改进算法的消融实验设计

Table 3 Design of ablation experiments with improved algorithm					
模型	YOLOv8s-seg	C2f_ ContextGuided	MF- FPN	SEMF- FPN	Segment_ LADH
1	✓				
2	✓	✓			
3	✓		✓		
4	✓			✓	
5	✓				✓
6	✓	✓		✓	
7	✓			✓	✓
8	✓	✓		✓	✓

改进算法的消融实验结果如表 4 所示。相比于基准模型 YOLOv8s-seg, 当在主干网络引入 C2f_
ContextGuided 模块时, 掩码平均精度损失 0.8%, 但模型参数和计算量分别下降了 1.86 M 和 5.5 GFOLPs。当引入 MF-FPN 时, 掩码平均精度提升 0.3% 且模型参数量显著下降了 3.44 M, 计算量下降了 3.9 GFOLPs。当引入 SEMF-FPN 时, 掩码平均精度提升 1.2%, 而模型参数和计算量与不含移位增强操作时持平, 证明移位增强操作和多分支多尺度设计对颈部网络中 CSP_MSCB 模块和 EUCB 模块改进的性能提升具有显著效果。当引入 Segment_LADH 模块时, 掩码平均精度提升 0.2% 且模型参数量和计算复杂度分别下降了 2.01 M 和 7.8 GFOLPs, 证明了该模块的轻量化优势和精度性能提升。

表 4 改进算法的消融实验结果

Table 4 Results of ablation experiments with improved algorithm							
模型	Box		Mask		Speed/	Param/	GFLOPs
	mAP@0.5	mAP@0.5:0.95	mAP@0.5	mAP@0.5:0.95	ms	M	
1	0.984	0.913	0.960 0	0.846	2.7	11.78	42.5
2	0.983	0.907	0.956 0	0.838	2.6	9.92	37.0
3	0.984	0.913	0.962 0	0.849	3.0	8.34	38.6
4	0.985	0.915	0.964 0	0.858	3.2	8.34	38.6
5	0.983	0.906	0.958 0	0.848	2.7	9.71	34.7
6	0.985	0.912	0.956 0	0.844	3.2	6.48	33.2
7	0.984	0.913	0.959 0	0.852	3.3	7.18	32.1
8	0.983	0.909	0.958 0	0.845	3.0	5.32	26.7

将本文提出的改进方法同时应用于 YOLOv8s-seg 模型时, 实验结果显示在参数量和计算量分别大幅减少 54.84% 和 37.18% 的同时, 边界框和掩码的平均精度 mAP@0.5 和 mAP@0.5:0.95 分别仅产生了 0.1%、0.4%、0.2% 和 0.1% 的波动, 而 FPS 由检测速度计算可得为 333.33 fps, 充分满足工业检测的实时性速度需求。

由此可知, 本文提出的改进的检测模型能有效地满足晶圆缺陷检测的轻量化、实时性及检测精度目标, 证实了改进方法的有效性。图 14 为部分晶圆缺陷图检测结果示意图。

3.5 对比实验

为进一步验证所提改进算法的性能, 本文分别进行与基准模型、同量级轻量化模型、其它实例分割模型三方面的对比。当前主流的实例分割模型主要有两类: 单阶段 YOLO 系列和两阶段 Mask R-CNN 系列, 其中 YOLO 系列实例分割功能的引入仅存于 YOLOv5 的 7.0 以上的版本、YOLOv8、YOLOv9 和 YOLO11 及以上, 由于本文基准模型为 YOLOv8s-seg, 且 YOLOv9c-seg 模型参数量为 27.67 M 过大, 故本文选取 YOLOv5s-seg 和 YOLOv8n-seg 作为同量级轻量化对比模型。另外, 当前国内外有关

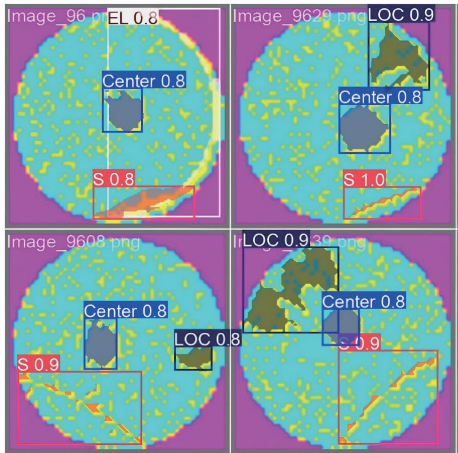


图 14 检测结果
Fig. 14 Detection result

晶圆图混合缺陷模式识别的实例分割研究较少, 暂无针对 MixedeWM38 数据集上进行电学测试得到的三值单通道晶圆数据格式的 YOLO 系列的相关应用研究, 仅李洋^[11]和 Chiu 等^[9]利用两阶段实例分割模型 Mask R-CNN 开展了对晶圆缺陷识别的实例分割研究, 因此本文也对使用

Mask R-CNN 的这两篇研究进行了对比。

对比实验结果如表 5 所示。轻量化的目的在于在尽量低的精度损失范围内显著降低模型的参数量和计算量,从对比实验结果可以看出:除 Mask R-CNN 外,YOLO 系列的模型均具备较强的相应效率,场景实时性佳,且本文改进模型相对基准模型在精度损失不超过 0.5% 的情况下实现了模型参数量和计算量的大幅降低,实现了较好的轻量化效果,且相比于其它轻量化模型基本具备更高的检测

精度,mAP@0.5:0.95 在目标掩码上的指标比 YOLOv5s-seg、YOLOv8n-seg 分别高 11.5%、2.6%,精度显著高于 Mask R-CNN 模型研究结果,参数量和计算量显著低于现有研究中的 Mask R-CNN 模型及其改进,较低于 YOLOv5s-seg,而相比于 YOLOv8n-seg,本文改进算法在精度和轻量化的平衡上具有更高的综合性能。由此,对比实验验证了本文改进算法具有一定的轻量化和检测精度的双重优势。

表 5 不同研究实验结果对比

Table 5 Comparison of results from different studies

研究	方法	mAP (Box)	mAP (Mask)	Speed/ ms	Param/ M	GFLOPs
文献[9]	改进 Mask R-CNN	—	0.637	—	—	—
文献[11]	改进 Mask R-CNN	0.604	0.405	43.0	—	—
文献[11]	Mask R-CNN	0.583	0.382	58.0	44.00	228.0
本文	YOLOv8n-seg	0.898	0.819	1.8	3.26	12.0
本文	YOLOv5s-seg	0.866	0.730	2.2	7.42	25.7
本文	改进 YOLOv8s-seg	0.909	0.845	3.0	5.32	26.7
本文	YOLOv8s-seg	0.913	0.846	2.7	11.78	42.5

4 结 论

为实现对晶圆混合模式缺陷识别的高精度检测,且同时降低算法复杂度,增强模型的可部署性,本文提出一种改进 YOLOv8s-seg 的轻量级实例分割算法。首先在特征提取层通过引入轻量级上下文引导机制,提出 C2f_ContextGuided 模块,通过优化主干特征提取 C2f 模块达到轻量化和增强语义的目的;其次,提出 SEMF-FPN 模块,在轻量化的同时提升网络对不同尺寸和复杂对象的处理能力;最后,改进检测头部网络,引入轻量化非对称检测头思想,提出 Segment_LADH 模块,达到头部网络轻量化的目的且不造成检测精度的损失。消融实验和对比实验结果表明,相较于基准模型 YOLOv8s-seg、其它同量级轻量化模型以及两阶段检测模型 Mask R-CNN,一方面实现了模型边界框和掩码的 mAP@0.5:0.95 分别为 90.9% 和 84.5% 的高精度检测,另一方面在改进模型检测精度损失仅 0.5% 范围内的情况下实现了模型复杂度的有效降低,且检测速度为 333.33 fps,整体而言模型检测精度高、轻量且快速,满足工业生产检测的需要,具有一定的工程应用价值,且对其它晶圆缺陷检测算法改进的研究工作也具有一定的参考借鉴价值。

参考文献

[1] HUANG AN CH, MENG SH H, HUANG T J. A survey on machine and deep learning in semiconductor industry: Methods, opportunities, and challenges[J]. Cluster Computing, 2023, 26(6): 3437-3472.

[2] 苏昊,李文豪,李俊龙,等. 晶圆级 Micro-LED 芯片检测技术研究进展[J]. 液晶与显示, 2023, 38(5): 582-594.
SU H, LI W H, LI J L, et al. Recent progress of wafer level Micro-LED chip inspection technology[J]. Chinese Journal of Liquid Crystals and Displays, 2023, 38(5): 582-594.
[3] 胡志强,吴一全. 基于机器视觉的芯片缺陷检测研究进展[J]. 仪器仪表学报, 2024, 45(7): 1-26.
HU ZH Q, WU Y Q. Research progress in chip defect detection based on machine vision[J]. Chinese Journal of Scientific Instrument, 2024, 45(7): 1-26.
[4] 王雨芹,栗雅娟,苏晓菁,等. 面向晶圆图缺陷模式识别的机器学习方法综述[J]. 微纳电子与智能制造, 2023, 5(2): 22-29.
WANG Y Q, SU Y J, SU X J, et al. Research of machine learning approaches for wafer defect pattern recognition[J]. Micro/nano Electronics and Intelligent Manufacturing, 2023, 5(2): 22-29.
[5] GHABRAMANI M, QIAO Y, ZHOU M CH, et al. AI-based modeling and data-driven evaluation for smart manufacturing processes [J]. IEEE-CAA Journal of Automatica Sinica, 2020, 7(4): 1026-1037.
[6] NAKAZAWA T, KULKARMI D V. Wafer map defect pattern classification and image retrieval using convolutional neural network[J]. IEEE Transactions

- on Semiconductor Manufacturing, 2018, 31(2): 309-314.
- [7] WANG J L, XU CH Q, YANG ZH L, et al. Deformable convolutional networks for efficient mixed-type wafer defect pattern recognition [J]. IEEE Transactions on Semiconductor Manufacturing, 2020, 33(4): 587-596.
- [8] KIM T S, LEE J W, LEE W K, et al. Novel method for detection of mixed-type defect patterns in wafer maps based on a single shot detector algorithm [J]. Journal of Intelligent Manufacturing, 2021, 33(6): 1715-1724.
- [9] CHIU M CH, CHEN T M. Applying data augmentation and mask R-CNN-based instance segmentation method for mixed-type wafer maps defect patterns classification [J]. IEEE Transactions on Semiconductor Manufacturing, 2021, 34(4): 455-463.
- [10] CHA J, JEONG J. Improved U-Net with residual attention block for mixed-defect wafer maps [J]. Applied Sciences, 2022, 12(4): 2209.
- [11] 李洋. 基于深度学习的晶圆缺陷检测算法研究[D]. 西安: 西安工业大学, 2022.
- LI Y. Research on wafer map detection based on deep learning[D]. Xi'an: Xi'an Technological University, 2022.
- [12] YAN J D, SHENG Y, PIAO M H. Semantic segmentation-based wafer map mixed-type defect pattern recognition [J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2023, 42(11): 4065-4074.
- [13] 邓广远, 王红成. 基于通道混洗和深度可分离卷积的混合型晶圆缺陷识别[J]. 东莞理工学院学报, 2024, 31(3): 17-23.
- DENG G Y, WANG H CH. Mixed-type wafer map recognition based on channel shuffle and depthwise separable convolution [J]. Journal of Dongguan University of Technology, 2024, 31(3): 17-23.
- [14] CHOI J, SUH D. A depthwise convolutional neural network model based on active contour for multi-defect wafer map pattern classification [J]. Engineering Applications of Artificial Intelligence, 2025, 139: 109707.
- [15] KIM T, BEHDINAN K. Advances in machine learning and deep learning applications towards wafer map defect recognition and classification: A review [J]. Journal of Intelligent Manufacturing, 2022, 34(8): 3215-3247.
- [16] MA J H, ZHANG T, YANG C, et al. Review of wafer surface defect detection methods [J]. Electronics, 2023, 12(8): 1787.
- [17] VARGHESE R, SAMBATH M. YOLOv8: A novel object detection algorithm with enhanced performance and robustness [C]. 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), 2024: 1-6.
- [18] 王彩霞, 郭鑫鹏, 刘鹏. 改进 YOLOv8n 的电路板缺陷检测算法 [J]. 电子测量与仪器学报, 2025, 39(8): 65-78.
- WANG C X, GUO X P, LIU P. Improved YOLOv8n algorithm for PCB flaws detection [J]. Journal of Electronic Measurement and Instrumentation, 2025, 39(8): 65-78.
- [19] 窦维嘉, 陈凯, 王娟平, 等. 基于 YOLOv8 的钛棒表面缺陷检测 [J]. 电子测量技术, 2025, 48(14): 56-64.
- DOU W J, CHEN K, WANG J P, et al. Improved surface defect detection of titanium rod based on YOLOv8 [J]. Electronic Measurement Technology, 2025, 48(14): 56-64.
- [20] 陈俊生, 陈沂蒙, 刘明杰, 等. 基于 MFES-YOLOv8n 的光伏电池缺陷检测方法 [J]. 仪器仪表学报, 2025, 46(6): 251-262.
- CHEN J SH, CHEN Y M, LIU M J, et al. A defect detection method for photovoltaic cells based on MFES-YOLOv8n [J]. Chinese Journal of Scientific Instrument, 2025, 46(6): 251-262.
- [21] 陈方彬, 赵仲勇, 王建, 等. 基于 YOLO-MCSL 的轻量化智能电能表热缺陷目标检测方法 [J]. 仪器仪表学报, 2025, 46(8): 108-119.
- CHEN F B, ZHAO ZH Y, WANG J, et al. A lightweight thermal defect detection method for smart electricity meters based on YOLO-MCSL [J]. Chinese Journal of Scientific Instrument, 2025, 46(8): 108-119.
- [22] WU T Y, TANG SH, ZHANG R, et al. CGNet: A light-weight context guided network for semantic segmentation [J]. IEEE Transactions on Image Processing, 2021, 30: 1169-1179.
- [23] YANG ZH Q, GUAN Q, ZHAO K ER, et al. Multi-branch auxiliary fusion YOLO with re-parameterization heterogeneous convolutional for accurate object detection [J]. Lecture Notes in Computer Science, 2024, 15042: 492-505.
- [24] TAN M X, PANG R M, LE Q. EfficientDet: Scalable and efficient object detection [C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 10778-10787.
- [25] GAO T, ZHANG Y, ZHANG ZH Y, et al. BHViT:

Binarized hybrid vision transformer[C]. 2025 IEEE/ CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2025: 3563-3572.

[26] RAHMAN M M, MUNIR M, MARCULESCU R. EMCAD: Efficient multi-scale convolutional attention decoding for medical image segmentation[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2024: 11769-11779.

[27] CHEN Y M, YUAN X B, WU R Q, et al. YOLO-MS: Rethinking multi-scale representation learning for real-time object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023(47): 4240-4252.

[28] ZHANG J R, CHEN ZH H, YANG X, et al. Faster and lightweight: An improved YOLOv5 object detector for remote sensing images [J]. Remote

Sensing, 2023, 15(20): 4974.

作者简介

余欣, 硕士研究生, 主要研究方向为图像处理、深度学习。
E-mail: 2236438320@qq.com

缪佳欣, 硕士研究生, 主要研究方向为图像处理、深度学习。
E-mail: 752203007@qq.com

龙浩辰, 硕士研究生, 主要研究方向为图像处理、深度学习。
E-mail: 15183537596@163.com

郝彦哲, 硕士研究生, 主要研究方向为图像处理、深度学习。
E-mail: 2695532178@qq.com

刘晓宇(通信作者), 博士, 副教授, 硕士生导师, 主要研究方向为测控技术与仪器、图像处理、深度学习。
E-mail: liuxiaoyu@scu.edu.cn