



仪器仪表学报
Chinese Journal of Scientific Instrument
ISSN 0254-3087, CN 11-2179/TH

《仪器仪表学报》网络首发论文

题目：基于因子图优化的晶圆缺陷跨尺度检测方法
作者：陈瑞，龙兵，戴志坚，詹文法，李荣杰
DOI：10.19650/j.cnki.cjsi.J2615301
收稿日期：2026-05-15
网络首发日期：2026-09-09
引用格式：陈瑞，龙兵，戴志坚，詹文法，李荣杰. 基于因子图优化的晶圆缺陷跨尺度检测方法[J/OL]. 仪器仪表学报. <https://doi.org/10.19650/j.cnki.cjsi.J2615301>



网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

DOI: 10.19650/j.cnki.cjsi.J2615301

基于因子图优化的晶圆缺陷跨尺度检测方法^{*}

陈 瑞^{1,2}, 龙 兵¹, 戴志坚^{1,3}, 詹文法², 李荣杰¹

(1. 电子科技大学自动化工程学院 成都 611731; 2. 安庆师范大学电子信息与集成电路学院 安庆 246133;
3. 电子科技大学(深圳)高等研究院 深圳 518110)

摘 要:晶圆图像缺陷检测是保障先进半导体制造良率与工艺稳定性的关键环节。针对先进制程中晶圆图像分辨率与缺陷尺寸跨度极大、微裂纹等弱边缘缺陷易漏检的问题,提出了一种基于因子图优化的晶圆缺陷跨尺度检测模型 FLC-DETR。首先,引入拉普拉斯边缘增强模块,利用方向可分离拉普拉斯核的频域选择性机制强化微缺陷边缘梯度响应,抑制背景噪声干扰。其次,设计因子图优化模块,在增强后的多尺度特征上构建跨尺度特征对齐的概率图模型,通过先验因子、观测因子与运动因子的联合推理,实现目标候选框全局一致性优化。此外,在解码器回归分支中引入通道梯度增强模块,将 Scharr 梯度特征嵌入边界框回归过程,通过可学习门控机制实现梯度信息的自适应加权融合,进一步提升了微小缺陷的定位精度。实验结果表明,在自建晶圆缺陷数据集上,FLC-DETR 模型的 mAP 和 AP@0.5 分别达到 81.7% 和 98.6%,相较于 RT-DETRv4-S 模型分别提升了 12.5% 和 10.7%,FNR 由 12.2% 降至 10.1%,有效缓解了微裂纹漏检问题。本方法在 3 种尺度等级缺陷上的 AP@0.5 分别达到 69.56%、79.18% 与 90.27%,验证了其跨尺度缺陷的检测能力。FLC-DETR 模型实现了晶圆图像跨尺度缺陷的鲁棒检测,有效缓解了微裂纹漏检风险,为先进半导体制造与封装工艺中的在线质量监测提供了可行的技术方案。

关键词: 晶圆缺陷;跨尺度检测;边缘增强;因子图优化;梯度增强

中图分类号: TH165 TP391 **文献标识码:** A **国家标准学科分类代码:** 520.20

Factor graph optimization for cross-scale wafer defect detection

Chen Rui^{1,2}, Long Bing¹, Dai Zhijian^{1,3}, Zhan Wenfa², Li Rongjie¹

(1. School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China;

2. School of Electronic Information and Integrated Circuits, Anqing Normal University, Anqing 246133, China;

3. Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518110, China)

Abstract: Wafer surface defect detection is a critical step for ensuring yield and process stability in advanced semiconductor manufacturing. To address the challenges posed by the extreme span between wafer image resolution and defect sizes in advanced processes, as well as the high miss-detection rate of weak-edge defects such as micro-cracks, we propose FLC-DETR, a cross-scale wafer defect detection model based on factor graph optimization. First, a Laplacian edge enhancement module is introduced to strengthen gradient responses at micro-defect edges by exploiting the frequency-domain selectivity of directionally separable Laplacian kernels, thereby suppressing background noise. Secondly, a factor graph optimization module is designed to formulate a probabilistic graphical model for cross-scale feature alignment upon the enhanced multi-scale features; through joint inference of prior, observation, and motion factors, global consistency of object proposals is optimized. Furthermore, a channel gradient enhancement module is incorporated into the decoder regression branch to embed Scharr gradient features into the bounding-box regression process; a learnable gating mechanism enables adaptive weighted fusion of gradient information, further improving localization accuracy for tiny defects. Experimental results on a self-built wafer defect dataset show that FLC-DETR achieves an mAP of 81.7% and an AP@0.5 of 98.6%, representing improvements of 12.5% and 10.7% over RT-DETRv4-S, while reducing the FNR from 12.2% to 10.1%, effectively alleviating the micro-crack miss-detection problem. The AP@0.5 values for defects across three scale levels reach 69.56%, 79.18%, and 90.27%,

收稿日期:2026-05-15 Received Date: 2026-05-15

^{*} 基金项目:国家自然科学基金面上项目(62474002)、国家自然科学基金重大研究计划项目(92573109)资助

respectively, validating the model's capability for cross-scale defect detection. FLC-DETR achieves robust cross-scale defect detection on wafer images, effectively mitigating the risk of micro-crack miss-detection and providing a feasible technical solution for in-line quality monitoring in advanced semiconductor manufacturing and packaging processes.

Keywords: wafer defect detection; cross-scale detection; edge enhancement; factor graph optimization; gradient enhancement

0 引言

在半导体制造与封装测试环节,晶圆缺陷检测是保障产品质量与工艺稳定性的核心任务^[1]。晶圆制造涉及硅锭生长、切割、研磨、抛光、光刻、化学蚀刻等多道精密工艺,任何环节的工艺波动均可能在晶圆表面形成微裂纹、划痕或颗粒污染等不同程度的损伤^[2]。表面缺陷即便初始几何尺度微小,也可能在后续热循环或电应力作用下逐渐扩展,进而引发器件短路、开路或性能退化,最终导致产品可靠性下降,批次良率降低甚至整批报废^[3]。因此,在晶圆测试环节实施高灵敏度、高精度的缺陷检测,是识别工艺异常、保障制造与封装工艺稳定性的必要前提。

当前晶圆检测主要采用自动光学检测(automated optical inspection, AOI)设备作为核心硬件^[4]。其工作流程分为两个阶段:首先对晶圆进行全视场快速扫描,自动提取疑似缺陷区域并生成候选图像块;随后对上述局部图像块实施后续智能分类与人工复检。在上述流程中,形态学运算、边缘检测及缺陷模式分析等传统基于手工特征规则的方法仍被广泛采用^[5]。然而,此类方法依赖于固定阈值与人工设计的特征模板,其泛化能力受限于预设规则,难以自适应地识别微小、不规则及低对比度缺陷^[6]。尤其在晶圆表面纹理复杂、光照不均的条件下,传统算法易产生大量误检与漏检,导致复检工作量显著增加、检测效率严重受限。

近年来,深度学习方法在目标检测领域取得了突破性进展,基于卷积神经网络(convolutional neural network, CNN)^[7]与Transformer^[8]架构的端到端模型已被成功应用于晶圆缺陷检测。两阶段检测器通过区域提议网络实现了较高定位精度^[9],单阶段检测器以全卷积结构满足了实时性需求^[10],而基于注意力机制的DETR(Detection Transformer)系列模型则通过全局交互缓解了复杂背景下的漏检问题^[11]。相较于传统手工特征,深度网络通过多层非线性变换自动学习缺陷的层次化表征,在复杂缺陷特征提取与语义建模方面表现出显著优势^[12]。然而,现有主流检测模型大多针对固定输入分辨率设计,其感受野与特征图尺度相对单一。在图像分辨率与缺陷尺寸跨度极大的晶圆检测场景下,常规的多尺度特征融合策略中深层语义特征经多次下采样后空间分辨率锐减,微裂纹等极小目标特征极易消失;浅层细

节特征虽保留高分辨率空间信息,却缺乏语义上下文支撑。这使得现有模型难以避免微裂纹漏检、宏观缺陷边界回归失准等问题,无法满足先进制程对检测精度与效率的严苛要求。

针对以上挑战,本文提出一种基于因子图优化的晶圆缺陷跨尺度检测方法 FLC-DETR,旨在实现晶圆图像极端尺度缺陷的高效、鲁棒检测。主要贡献为:

1) 针对目标检测模型后处理阶段候选框独立打分导致的结构化信息缺失问题,构建面向检测场景的因子图优化(factor graph optimization, FGO)跨尺度关联机制。通过先验因子、观测因子与运动因子的联合计算推理,实现非邻层特征的直接交互,增强浅层细节信息与深层语义信息的传播与融合。

2) 设计拉普拉斯边缘增强模块(Laplacian edge-enhancement module, LEM),利用边缘增强机制强化微裂纹类低对比度缺陷的边缘特征表达,有效提升主干网络对微弱缺陷的边界感知能力,丰富浅层细节语义特征。

3) 设计通道梯度增强模块(channel gradient-enhancement module, CGM),利用Scharr梯度增强机制细化边界梯度信息,提升浅层细节信息与深层语义信息的传播与融合效率,优化缺陷定位回归精度。

4) 在自建晶圆缺陷数据集上对所提模型进行多尺度、不同分辨率输入的实验验证,并与主流检测模型进行对比分析,实验结果表明 FLC-DETR 模型的检测精度显著提升,证明了其在晶圆图像极端尺度缺陷检测中的有效性。

1 相关工作

根据缺陷产生因素,晶圆缺陷可分为随机缺陷与系统缺陷^[13]。随机缺陷由表面颗粒污染物引发,分布随机;系统缺陷源于光刻掩模与曝光工艺的系统误差,多出现于亚分辨率结构区,并在同晶圆各芯片对应位置重复出现。为提升晶圆成品率,准确识别上述缺陷至关重要。Park 等^[14]提出了一种基于模糊逻辑的晶圆图缺陷模式识别框架,利用中心区域密度等定量指标有效降低了不完美标签环境下的过拟合风险。Wang 等^[15]提出了一种基于子区域一对一映射的检测算法,通过标准晶圆基准自动校准分割、三通道 RGB(red-green-blue)灰度相似度决策函数实现缺陷判别,显著提高检测精度。传统 AOI 设备依赖规则驱动图像算法与模板匹配进行缺陷检

测,但易受光照变化、微小缺陷及工艺变化影响,导致检测性能下降。此外,该方法需大量人工定义规则,对新型复杂缺陷的适应能力差,因此,主流方案逐步转向数据驱动的检测策略^[16]。

监督式深度学习模型经端到端训练,逐层抽取从低级边缘到高级语义的分层纹理特征,自适应学习多尺度缺陷形态与背景杂波的统计分布,实现高效特征建模。Chi 等^[17]提出了一种轻量级模型 CISC-YOLO,通过 GhostConv 模块减少冗余计算,构建新型跨尺度特征融合颈部网络,实现低延迟的实时缺陷检测方案。Thuan 等^[18]提出利用单缺陷热图提取感兴趣区域合成自适应混合缺陷样本,并采用自注意力 CNN 捕捉缺陷间复杂交互以提升分类精度与泛化性,提高混合缺陷检测效率。基于 CNN 的检测器虽推理速度较快,然而有限的感受野易遗漏不规则的细微缺陷。Xu 等^[19]提出一种改进的 RT-DETR 方法,通过引入动态蛇形卷积层增强细长划痕特征提取,并设计基于注意力的编码器模块与残差网络缓解小目标信息丢失。Zhang 等^[20]提出轻量级检测器 FALCO-WAFER,通过深度可分离模块与对角注意力头实现高效纹理编码与鲁棒特征细化,实现高吞吐低延迟的部署方案。

现有研究多采用固定输入分辨率,难以同时兼顾微小缺陷的精细纹理与大尺度缺陷的全局上下文。在晶圆制造场景中,基于固定采样率的特征金字塔难以建立有效的跨尺度语义关联,导致微小缺陷漏检与大尺度缺陷定位漂移。针对该问题,本文提出 FLC-DETR 模型,通过因子图优化实现跨尺度特征对齐,并结合拉普拉斯边缘增强与通道梯度增强模块强化缺陷特征表达与边界锐化,实现晶圆极端尺度缺陷的精确检测。

2 晶圆缺陷检测模型

2.1 晶圆缺陷检测框架

随着芯片制程迭代与封装形式多元化,AOI 设备在不同放大倍率下采集的晶圆图像呈现显著的分辨率差异与尺度变化。晶圆裂纹由多物理场耦合应力诱发,材料热膨胀系数不匹配引发热应力失配,易导致界面开裂;减薄与划片等机械工序在边缘引入微裂纹;前道工艺缺陷则在封装应力释放过程中扩展为宏观断裂。上述成像条件与裂纹成因共同导致缺陷尺度跨越亚微米至毫米量级,大幅增加了晶圆图像特征的复杂性。针对晶圆图像跨尺度检测难题,本文设计了 FLC-DETR 检测模型,结构如图 1 所示,核心组件包括:

1) 搭载拉普拉斯边缘增强模块的骨干网络:在网络浅层嵌入可学习的拉普拉斯门控卷积单元,通过二阶微分算子显式增强缺陷边缘梯度响应,在深层特征提取过程中动态抑制背景噪声,有效防止亚微米级微裂纹因多次下采样导致的特征消失与梯度弥散;

2) 高效混合编码器:采用跨尺度特征金字塔融合策略,通过高效层聚合路径对百倍分辨率跨度下的多层级特征进行自适应融合。编码器在保持较低计算开销的同时,构建从细节到语义的多尺度特征表示,为候选查询生成与目标解码提供精准的多尺度输入;

3) 因子图优化模块:作为迭代式后处理优化器,通过构建因子图模型对初始检测结果进行一致性重评分。模块融合先验因子、运动因子与观测因子,并通过消息传递机制实现检测结果的全局优化,提升边界框筛选的稳定性和亚分辨率结构区的定位准确性;

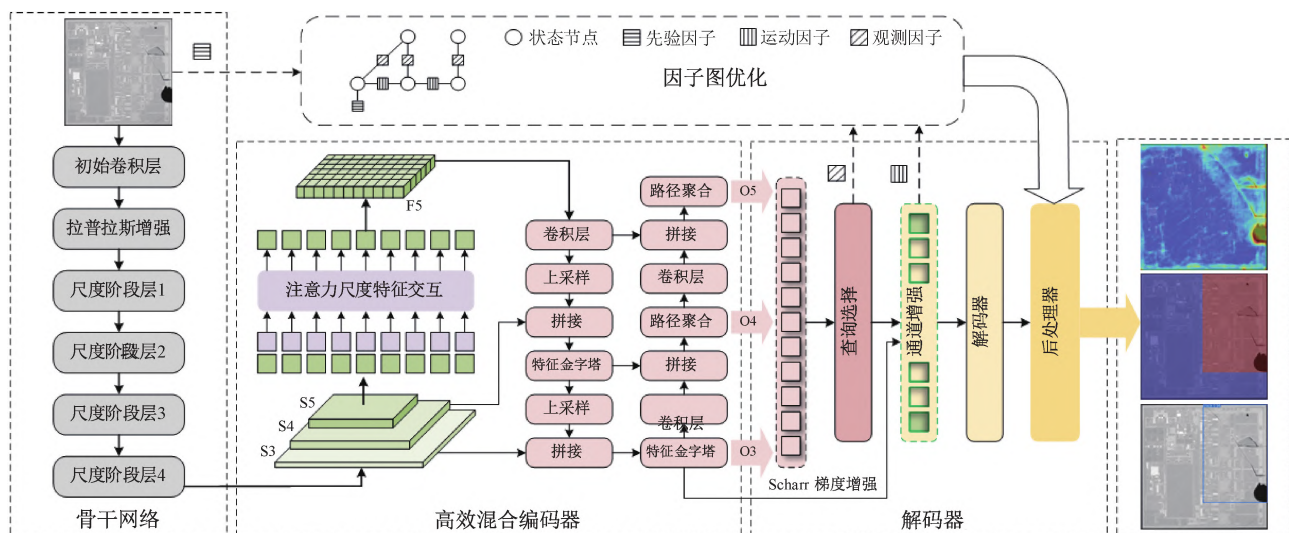


图 1 FLC-DETR 检测模型结构

Fig. 1 The architecture of the FLC-DETR detector

4) 候选与解码模块: 基于置信度阈值筛选高分查询目标以抑制冗余候选框, 编码器输出的多尺度特征与通道梯度增强模块进行深度融合, 进一步锐化缺陷与背景的判别边界。解码器在此基础上同步细化分类概率与回归参数, 实现定位精度与分类的协同提升。

2.2 拉普拉斯边缘增强模块

为了强化晶圆图像缺陷弱边界表征, 本文在骨干网络 HGNetv2 的浅层阶段引入轻量拉普拉斯边缘增强模块 LEM。如图 2 所示, LEM 在主干路径上采用高效通道注意力 (efficient channel attention, ECA) 对通道重标定^[21], 获得语义校准后的基线特征。该设计仅依赖全局平均池化与小核一维卷积, 参数与计算开销极低, 可在不破坏实时性的前提下提升通道响应的判别性。

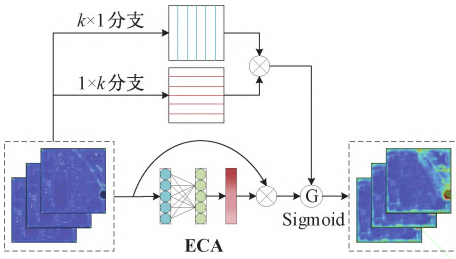


图 2 拉普拉斯边缘增强模块结构

Fig. 2 The architecture of the LEM

在 LEM 增强路径中, 构建双分支定向深度卷积。垂直分支使用 $k \times 1$ 拉普拉斯条形核, 强调纵向结构变化, 水平分支使用 $1 \times k$ 拉普拉斯条形核, 关注横向连续边界。当 $k=3$ 时, 两分支的核矩阵构造如式 (1) 所示。

$$\mathbf{K}_v = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \mathbf{K}_h = \begin{bmatrix} 0 & 0 & 0 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad (1)$$

二者均为零均值核, 满足可分离性, 且 LEM 的分离条形核在频域具有更强的轴向频率选择性。对 $\mathbf{K}_v$ 和 $\mathbf{K}_h$ 分别作二维离散傅里叶变换, 利用可分离核的频域乘积性质, 其频域传递函数如式 (2) ~ (3) 所示。

$$H_v(\omega_x, \omega_y) = -4\gamma \sin^2\left(\frac{\omega_y}{2}\right) \quad (2)$$

$$H_h(\omega_x, \omega_y) = -4\gamma \sin^2\left(\frac{\omega_x}{2}\right) \quad (3)$$

其中, 垂直和水平两分支的幅度响应仅在单一频率轴上具有选择性, $\mathbf{K}_v$ 的响应在 $\omega_y = \pm\pi$ 处取得峰值, 对应垂直边缘检测; $\mathbf{K}_h$ 的响应在 $\omega_x = \pm\pi$ 处取得峰值, 对应水平边缘检测; γ 是条形拉普拉斯核的增益系数。从频域幅频特性分析, 如图 3 所示, LEM 双分支的曲面呈现出严格的轴向分离特征, $\mathbf{K}_v$ 沿着 ω_x 方向全通, 而在 ω_y 方向呈现山谷形高通特性, $\mathbf{K}_h$ 则正好互换。

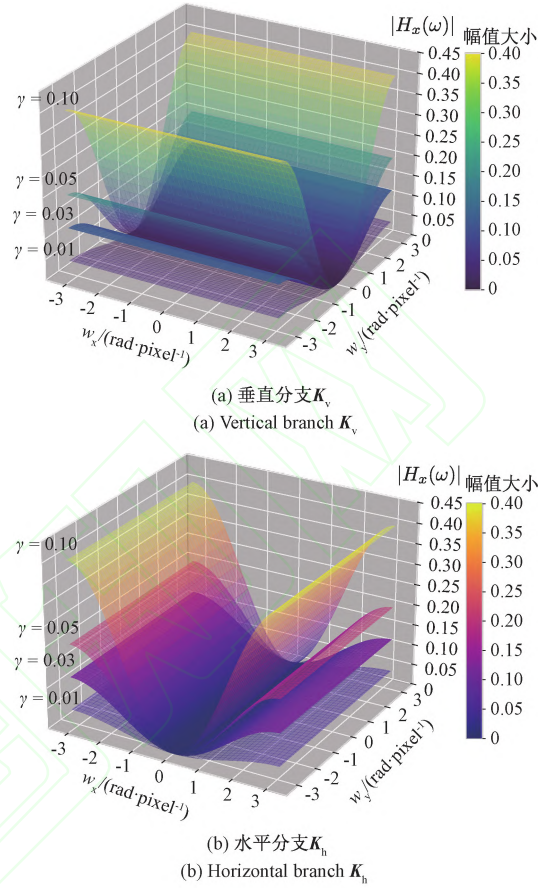


图 3 可变 γ 参数下 LEM 双分支频率响应

Fig. 3 LEM dual-branch frequency response under varying γ

引入条形拉普拉斯核的增益系数 γ 后, 随着从 0.01 增至 0.10, 频域峰值线性增长。但时域核能量呈二次增长, 形成天然的频率选择性正则化。在 $\gamma=0.03$ 时, 峰值响应 0.12 足以增强微裂纹边缘, 而能量开销可忽略; $\gamma>0.05$ 时, 边际能量效率急剧下降, 过增强导致的噪声放大。这种轴向频率选择性使得 LEM 在频域将缺陷的轴向频率成分与背景噪声解耦, 实现对方向性弱边缘的定向放大, 提升对目标缺陷的感知能力。

LEM 的两分支先进行方向融合, 再经双曲正切压缩得到高频增强项, 随后通过可学习总门控系数自适应调节, 避免对预训练主干特征的剧烈扰动, 并与主干特征残差融合输出, 如式 (4) ~ (5) 所示。

$$H = \tanh(\text{DWConv}_{1 \times k}^{\text{Lap}}(x) + \text{DWConv}_{k \times 1}^{\text{Lap}}(x)) \quad (4)$$

$$y = \text{ECA}(x) + \sigma(\theta)H \quad (5)$$

其中, $x, y \in \mathbf{R}^{C \times H \times W}$ 分别代表输入、输出特征, $\text{ECA}(\cdot)$ 为高效通道注意力, $\text{DWConv}_{1 \times k}^{\text{Lap}}(x)$ 和 $\text{DWConv}_{k \times 1}^{\text{Lap}}(x)$ 分别为水平和垂直拉普拉斯深度卷积分支, k 为条形卷积核长度, $\sigma(\cdot)$ 为 Sigmoid 操作, α 为可学习参数, $\tanh(\cdot)$ 为边缘响应幅值限制, θ 为可学习门控参数。

FLC-DETR 通过 LEM 增强局部边缘保持能力,适用于晶圆缺陷检测场景中低对比缺陷的稳健表征。LEM 作为骨干网络中高效的特征编码单元,通过方向卷积与边缘先验的协同引导,增强水平与垂直方向通道特征传输效率,使裂纹等具有显著方向性纹理的弱边缘缺陷在特征空间中获得更高的信噪比。此外,LEM 对方向性边缘的增强并非以牺牲各向同性缺陷特征为代价;对于圆形颗粒污染、划痕等非方向性或弱方向性缺陷,LEM 模块残差形式的结构设计与通道特征增强机制保障了网络的通用特征提取能力。

2.3 因子图优化模块

因子图优化算法以概率图模型为理论基础,在最大后验估计框架下联合建模先验知识、观测证据与邻域一致性,实现对结构化约束的高效推理^[22]。在变量节点与因子节点构成的二分图中,通过迭代式的置信度传播聚合多源信息,将全局优化目标解耦为可并行计算的局部更新,从而提升推断效率与鲁棒性。传统基于 Transformer 检测框架通常对候选目标独立打分与回归,候选框之间缺乏显式的结构化关联,导致推理过程割裂。这种独立处理的机制在面对晶圆图像中微小缺陷与大尺度异常并存、低对比度背景干扰强的复杂场景时,容易引发误检率上升与跨尺度目标框定位不稳定的问题^[23]。

因子图优化的结构化建模能力为缓解上述问题提供了理论途径,但在晶圆缺陷检测场景中,需在变量定义、因子设计与推理机制 3 个层面进行针对性适配。在变量定义中,节点对应候选框参数,包含中心坐标、宽高与置信度,节点间无显式时序关系,需构建类别感知的 KNN (k -nearest-neighbors) 图,从而建立语义相关的邻域关联。在因子设计中,需涵盖先验因子、运动因子与观测因子这 3 类联合约束,其中观测因子显式惩罚分类置信度与 IoU 定位质量之间的不一致,确保分类与回归任务在优化层面保持协同。图结构的推理机制随每帧图像独立构建,节点数量随候选框规模动态变化,采用固定迭代轮次的闭式消息传递机制,在利用批处理并行性的同时,需应对极端尺度下节点稀疏与分布不均衡的问题。

因此,本文在此基础上设计因子图优化模块 FGO,如图 4 所示。该模块的核心思想源于最大后验估计与互信息增强,通过在训练与推理阶段联合建模先验知识、观测证据及邻域一致性,有效抑制冗余背景干扰,仅保留与缺陷语义高度相关的高效信息。

FGO 将目标检测的后处理建模为因子图上的最大后验概率估计问题,定义因子图 $\mathcal{G} = (\nu, \mathcal{F})$, 其中变量节点集合 $\nu = \{v_1, \dots, v_n\}$ 表示待估计的边界框位置、尺寸以及置信度,因子节点集合 $\mathcal{F} = \{f_1, \dots, f_m\}$ 编码观测约束、先验几何约束与邻域运动约束。联合概率分布分解为各因子的乘积,如式(6)所示。

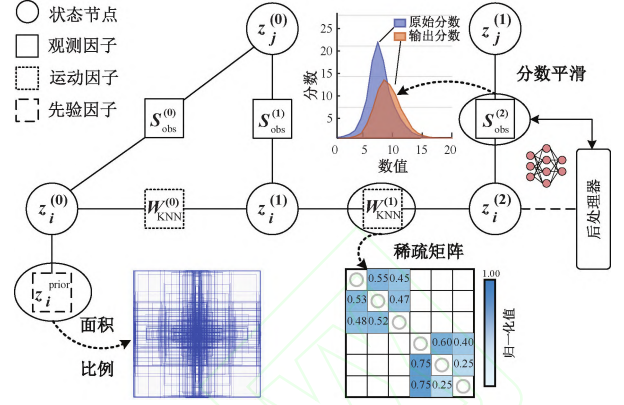


图 4 因子图优化模块结构

Fig. 4 The architecture of the FGO model

$$P(y|x) = \frac{1}{Z} \prod_{j=1}^m f_j(y_{N(j)}) \quad (6)$$

其中, $y_{N(j)}$ 为与 f_j 因子相连的变量子集, Z 为归一化常数,最大后验概率估计等价于最小化能量函数如式(7)所示。

$$y^* = \operatorname{argmin}_y \sum_{j=1}^m -\log f_j(y_{N(j)}) \quad (7)$$

FGO 以候选目标为图节点,在目标中心坐标空间构建 k 近邻关系,并结合类别一致性得到稀疏权重,再通过 3 类因子联合约束。先验因子约束缺陷边界框的几何参数分布,抑制不合理的尺度漂移。设变量节点 v_i 的状态为 $y_i = [x_i, y_i, w_i, h_i, s_i]^T$, 采用对数变换后的面积与长宽比统计先验信息,先验能量定义如式(8)所示。

$$E_{\text{prior},i} = \frac{1}{2} \left[\left(\frac{\ln A_i - \mu_a}{\sigma_a} \right)^2 + \left(\frac{\ln r_i - \mu_r}{\sigma_r} \right)^2 \right] \quad (8)$$

其中, $A_i = w_i h_i$ 为第 i 个预测框面积; $r_i = w_i / h_i$ 为长宽比; μ_a, σ_a 与 μ_r, σ_r 分别为缺陷面积和长宽比的先验均值与标准差。

运动因子基于 KNN 约束邻域候选框的空间关联,增强局部结构连续性。以目标中心坐标 $c_i = [x_i, y_i]$ 构建 KNN 图,并结合类别一致性得到稀疏权重。运动能量编码邻域几何与置信度的局部平滑性如式(9)所示。

$$E_{\text{motion},i} = \left\| b_i - \sum_{j \in \kappa(i)} w_{ij} b_j \right\|_1 + \eta \left(s_i - \sum_{j \in \kappa(i)} w_{ij} s_j \right)^2 \quad (9)$$

其中, b_i 为边界框向量, s_i 为置信度分数, $\kappa(i)$ 为节点 i 的 k 近邻集合; w_{ij} 为 KNN 自适应权重; η 为置信度平滑系数。

观测因子将检测头的分类置信与定位质量对齐,编码网络输出的观测信息如式(10)所示。

$$E_{\text{obs},i} = \ell_{\text{bce}}(z_i, \text{IoU}_i) \quad (10)$$

其中, z_i 为检测头输出的分类置信度, IoU_i 为预测框与真值框的交并比, $\ell_{\text{bce}}(\cdot)$ 为二元交叉熵损失。

三类因子联合构成总能量函数,引入可学习权重以

平衡各约束项贡献,如式(11)所示。

$$E_{\text{FGO}} = \lambda_p \sum_{i=1}^N E_{\text{prior},i} + \lambda_m \sum_{i=1}^N E_{\text{motion},i} + \lambda_o \sum_{i=1}^N E_{\text{obs},i} \quad (11)$$

其中, $\lambda_p, \lambda_m, \lambda_o$ 分别为先验因子、运动因子和观测因子权重。

直接对 E_{FGO} 执行完整的非线性二阶优化涉及大规模矩阵求逆,计算开销较高,并且单帧图像内候选目标的图结构为静态且规模有限,这一特性使得 FGO 能够直接在该稀疏拓扑结构上对能量进行线性化,进而执行基于迭代的消息传递,以极低的计算开销逼近能量极小点。每个节点聚合 3 类消息更新,再进行置信度更新,如式(12)~(14)所示。

$$\psi_i^{\text{prior}} = -E_{\text{prior},i} \quad (12)$$

$$w_{ij} = \text{Softmax}_{j \in \kappa(i)} \left(-\frac{\|\mathbf{c}_i - \mathbf{c}_j\|_2^2}{2\sigma_c^2} \right) \quad (13)$$

$$z_i^{(t+1)} = \frac{\lambda_o z_i^{(0)} + \lambda_p \psi_i^{\text{prior}} + \lambda_m \sum_{j \in \kappa(i)} w_{ij} z_j^{(t)}}{\lambda_o + \lambda_p + \lambda_m + \varepsilon} \quad (14)$$

其中, ψ_i^{prior} 为先验状态; $\|\mathbf{c}_i - \mathbf{c}_j\|_2^2$ 为中心坐标向量的欧氏距离平方; σ_c 为距离衰减带宽; $\text{Softmax}(\cdot)$ 为归一化处理; $z_i^{(0)}$ 为初始观测状态; $z_j^{(t)}$ 为邻居节点 j 在 t 轮的状态; ε 为极小稳定正数。

为清晰阐述因子图优化的推理流程,将跨尺度特征对齐形式化为能量最小化问题,采用迭代消息传递算法显式求解,详细流程如算法 1 所示。

算法 1: 基于消息传递的因子图优化 (FGO)

输入: 初始节点置信度与几何状态集合, 先验、观测、运动因子权重以及邻域节点数量等系数

输出: 优化后的节点置信度与几何状态

```

1: 数据初始化, 基于各节点中心坐标构建  $k$  近邻关联图
2: while 迭代次数未达到阈值 do
3:   for 每个候选节点 do
4:     先验能量项取负值更新, 运动消息初始化为零
5:     for 每个邻域节点 do
6:       根据两节点中心坐标距离计算空间权重
7:       运动消息聚合邻域节点权重更新
8:     end for
9:     观测消息更新检测头分类与定位置信度, 节点更新
10:   end for
11:   计算所有节点新旧置信度之间差值的最大值
12:   if 该最大值小于收敛阈值 then
13:     break
14:   end if
15:   迭代计数器加 1, 继续下一轮迭代
16: end while
17: return 所有节点的最终置信度与几何状态

```

每个节点每轮迭代聚合先验、运动与观测 3 类消息。先验消息编码检测器初始语义约束, 运动消息经空间权重由邻域传递实现局部平滑, 观测消息融合检测头分类置信度与定位质量。3 类消息按因子权重加权聚合后更新节点置信度。迭代终止条件为相邻两轮置信度最大变化量小于阈值或达到预设轮次。消息传递仅涉及节点级局部运算与邻居聚合, 计算复杂度取决于每帧候选节点数 n , KNN 近邻数 k 与迭代轮数 T 。基于候选框中心坐标构建 k - d 树, 其时间复杂度为 $\mathcal{O}(n \log n)$, 每轮迭代中, 先验消息与观测消息均为节点独立计算, 复杂度为 $\mathcal{O}(n)$ 。运动消息需要遍历每个节点的 k 邻居进行加权聚合, 复杂度为 $\mathcal{O}(nk)$, 置信度更新为 $\mathcal{O}(n)$ 。因此, 单轮迭代复杂度为 $\mathcal{O}(nk)$, FGO 总时间复杂度为 $\mathcal{O}_{\text{FGO}} = \mathcal{O}(n \log n + T \cdot nk)$ 。

在计算过程中, 运动消息聚合通过邻居索引直接访问并加权求和, 先验与观测消息为逐元素运算, 不涉及矩阵分解。相较于高斯-牛顿等方法需构建并分解 Hessian 矩阵, 稠密情形下时间复杂度为 $\mathcal{O}(n^3)$, 稀疏情况下为 $\mathcal{O}(n^2)$, 且内存开销随节点数平方增长。FGO 模块以迭代式消息传递替代二阶优化中全局矩阵求解, 将计算开销控制在 $\mathcal{O}(nk)$ 量级。该模块以结构先验、运动约束与观测似然三者联合建模为基础, 通过因子图显式编码缺陷间的空间关联, 从而在微小缺陷与剧烈尺度变化并存的复杂工况下维持定位的稳定性, 是平衡检测精度与实时性的核心组件。

2.4 通道梯度增强模块

在复杂噪声环境中, 微小、细长及各向异性缺陷因其边缘对比度低、几何形态复杂, 导致传统基于绝对强度或语义特征的检测器在边界回归阶段难以建立稳定的梯度响应, 定位精度显著下降。为此, 本文提出通道梯度增强模块 CGM, 如图 5 所示, 该模块以通道级梯度信息作为辅助监督信号, 通过轻量化的边缘提取与特征对齐, 以可学习门控机制将梯度特征嵌入边界回归分支, 实现定位精度与推理效率的平衡。

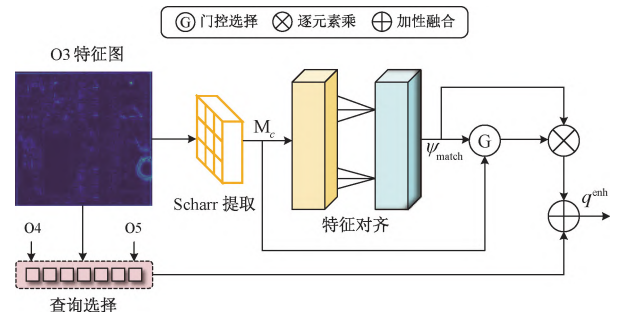


图 5 通道梯度增强模块结构

Fig. 5 The architecture of the CGM

通道梯度增强模块编码器输出的高分辨率特征图 $O_3 \in \mathbf{R}^{C \times H/8 \times W/8}$ 逐通道执行 Scharr 梯度幅值计算,定义水平与垂直卷积核如式 (15) 所示。

$$\mathbf{K}_x = \begin{bmatrix} -3 & 0 & 3 \\ -10 & 0 & 10 \\ -3 & 0 & 3 \end{bmatrix}, \quad \mathbf{K}_y = \begin{bmatrix} -3 & -10 & -3 \\ 0 & 0 & 0 \\ 3 & 10 & 3 \end{bmatrix} \quad (15)$$

第 c 通道的梯度幅值图 M_c 计算如式 (16) 所示。

$$M_c = \sqrt{(O_{3c} * \mathbf{K}_x)^2 + (O_{3c} * \mathbf{K}_y)^2} \quad (16)$$

其中, $*$ 表示卷积计算, $O_{3,c}$ 为第 c 通道特征。

为使梯度特征与解码器查询空间兼容,CGM 引入分辨率对齐层,通过双线性插值将边缘图下采样,再经深度可分离卷积、批归一化及逐点卷积映射至与查询兼容的表示空间,得到对齐特征,确保后续跨模态融合时的通道一致性与语义适配性。如式 (17) 所示。

$$\mathbf{G} = W_{1 \times 1} \cdot \sigma(\text{BN}(\text{DWConv}_{3 \times 3}(\mathbf{M}))) + \mathbf{M}_{\text{prior}} \quad (17)$$

其中, $\mathbf{G}$ 为特征图张量, $W_{1 \times 1}$ 为卷积权重, $\sigma(\cdot)$ 为 ReLU 激活函数, $\mathbf{M}_{\text{prior}}$ 为残差分支张量。

空间采样单元基于查询参考点从对齐特征中提取局部梯度令牌,经通道匹配后与回归分支输入维度对齐。门控融合单元通过可训练参数生成标量门控权重,利用加权融合获得增强查询。该动态加权机制使模型在噪声区域自动抑制梯度响应以避免伪边缘放大,而在真实缺陷边界处保持高增益以强化几何监督,从而解决固定融合比例导致的噪声放大或有效特征弱化问题,实现梯度特征的自适应加权融合。

3 实验与结果分析

3.1 晶圆缺陷数据集

本文采用自定义数据集评估 FLC-DETR 模型,从真实工业产线采集晶圆图像,并引入了数据增强策略,以扩展样本规模与多样性。采用随机裁剪、旋转反射及透视变换等操作,模拟实际场景中的位置偏移与目标变形。在光度增强方面,采用高斯噪声、图像压缩、亮度对比度扰动与灰度变换,以模拟不同照明与成像条件下的图像退化。构建 4 660 张人工标注的晶圆缺陷图像数据集,类别为 crack。数据集按照 8 : 1 : 1 比例划分训练集 (3 728)、验证集 (466) 和测试集 (466)。划分后的训练集缺陷尺寸分布统计图如图 6 所示。

根据晶圆缺陷面积分布图 6 所示,数据集在 $\leq 100 \text{ pixel}^2$ 、 $100 \sim 9\,000 \text{ pixel}^2$ 及 $\geq 9\,000 \text{ pixel}^2$ 这 3 个面积区间的缺陷数量分别为 294、3 126 与 1 624,整体呈现以中等尺度缺陷为核心的层级化分布趋势。中等面积缺陷占比约 62%,大尺度缺陷占比约 32%,共同构成样本主

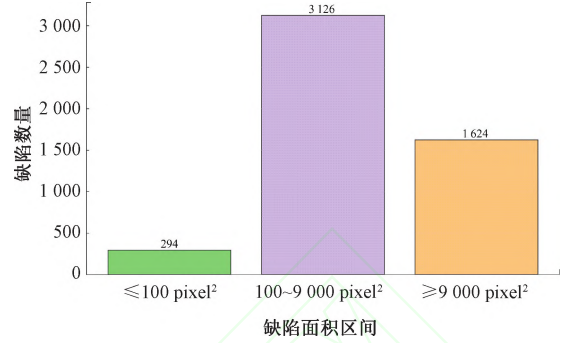


图 6 晶圆缺陷面积分布统计

Fig. 6 Wafer defect area distribution statistics

体,极小缺陷占比约 6%。该分布形态表明数据集在面积维度上覆盖了从微小到显著的完整尺度范围,能够满足多尺度缺陷检测模型对不同大小目标的训练需求。

3.2 实验环境与评估指标

为综合评价模型的性能,本文采用 6 项评估指标,涵盖精度与效率两个维度。包括参数量 (params)、千兆浮点运算 (giga floating-point operations, GFLOPs)、延迟 (latency)、平均精度均值 (mean average precision, mAP)、平均精度 (average precision, AP) 和漏检率 (false negative rate, FNR)。精度维度指标用于量化模型的检测质量,平均精度 AP 通过对 Precision-Recall 曲线进行插值积分,以曲线面积度量单类别的检测性能,平均精度均值 mAP 为所有类别 AP 的算术平均,漏检率 FNR 用于衡量实际缺陷被错误判定为背景的比例,直接反映缺陷漏检风险。具体定义如式 (18) ~ (21) 所示。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (18)$$

$$\text{AP} = \int_0^1 \text{P}(R) \text{d}R \quad (19)$$

$$\text{mAP} = \frac{1}{n} \sum_{i=1}^n \text{AP}_i \quad (20)$$

$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}} \quad (21)$$

其中,TP 为正类判定为正类的样本,FP 为负类误判为正类的样本,FN 为正类漏判为负类的样本。 $\text{P}(R)$ 为精度随召回率变化关系, AP_i 为第 i 类的平均精度。

效率维度指标用于评估模型的计算与部署开销。参数量表示模型中可训练参数的总数,决定存储占用大小。千兆浮点运算衡量模型完成一次前向传播所需的浮点运算总量。延迟定义为模型处理单张图像所需的推理时间,是衡量实时性的关键指标。因实际产线需求,本文实验训练环境硬件配置:CPU 为 Intel Xeon (R) Platinum 8370C, GPU 为 NVIDIA RTX PRO 6000 (96 GB)。推理阶段硬件配置:CPU 为 Intel Core (TM) Ultra 7 265KF,

GPU 为 NVIDIA GeForce RTX 5060 Ti (16 GB)。为确保对比实验的公平性,所有模型采用官方推荐配置统一训练,均在相同硬件环境下训练至收敛。环境基于 Python 3.11,Pytorch 2.11,Batch size 设置为 16,数据增强策略为 Mosaic,采用自动混合精度 AMP,详细配置如表 1 示。

表 1 实验环境信息

Table 1 Experimental environment information		
实验环境	训练配置	推理配置
编程环境	Python 3.11	
	Pytorch 2.11	
CPU	Intel Xeon (R)	Intel Core(TM)
	Platinum 8370C	Ultra 7 265KF
GPU	NVIDIA RTX	NVIDIA GeForce RTX
	PRO 6000 (96 GB)	5060 Ti (16 GB)
训练周期	132 epochs	-
Batch size	16	1
优化器	AdamW	-
初始学习率	0.000 4	-

3.3 模型对比

为验证所提模型在晶圆缺陷检测任务中的性能,在自建缺陷数据集上进行对比实验,结果如表 2 所示。YOLOv26-S 与 YOLOv26-M 的 mAP 分别仅为 43.2% 和 44.3%,FNR 高达 35.2% 和 33.9%,表明单纯增加骨干网络深度对于精度提升十分有限,无法准确获取尺度差异较大的缺陷纹理信息。

RT-DETR 系列模型具备全局上下文的感知能力,但未针对晶圆缺陷场景进行尺度优化,导致漏检率较高,无法直接部署在产线上。其中 RT-DETRv2-S 的 mAP 为 68.6%,FNR 为 15.0%;RT-DETRv4-S 的 mAP 提升至 69.2%,FNR 降至 12.2%,但仍存在明显的漏检问题。D-FINE-S 和 DEIMv2-S 等改进型 DETR 模型虽然通过优化解码器结构取得了一定进展,但 FNR 仍维持在 31.7% 以上,难以满足半导体制造的高可靠性需求。

本文提出的 FLC-DETR 模型参数量为 10.49 M,计算量为 24.89 GFLOPs,在保持轻量化的同时实现了最高 mAP 81.7% 和最低 FNR 10.1%,较 RT-DETRv4-S 分别提升 12.5% 和降低 2.1%。如图 7 所示,归一化混淆矩阵表明 FLC-DETR 模型对缺陷具有高召回率 90%,背景误

表 2 不同模型检测性能实验结果

Table 2 Experimental results of detection performance for different models						
模型	参数量/M	GFLOPs/G	延迟/ms	mAP/%	AP@0.5/%	FNR/%
Cascade-RCNN-R50	69.29	93.65	6.47	44.5	73.2	34.8
ATSS-R50	38.91	43.26	3.50	48.5	75.7	34.2
YOLOv11-S	9.41	21.33	2.10	37.7	67.2	34.0
YOLOv13-S	9.02	20.04	4.40	39.9	65.9	35.2
YOLOv26-S	9.47	20.54	2.21	43.2	73.1	35.2
YOLOv26-M	20.35	67.80	4.74	44.3	73.8	33.9
YOLOv26-L	24.75	86.13	6.17	46.7	76.1	32.1
D-FINE-S	10.18	24.81	10.46	53.9	77.6	35.2
DEIMv2-S	9.67	25.33	10.99	52.2	75.2	31.7
RT-DETRv2-S	20.08	59.86	12.17	68.6	90.8	15.0
RT-DETRv4-S	10.42	24.82	10.52	69.2	87.9	12.2
RT-DETRv4-M	19.67	56.33	12.57	75.5	96.6	13.6
FLC-DETR	10.49	24.89	10.99	81.7	98.6	10.1

报率为 1%,漏报率为 10%,表明 FLC-DETR 在晶圆缺陷场景下实现高效、鲁棒的检测。

如图 8 所示,所提方法在检测精度与漏检控制上均显著优于现有主流方法。相较于 RT-DETRv4-S 模型,FLC-DETR 模型的 mAP 由 69.2% 提升至 81.7%,增加 12.5%,FNR 由 12.2% 降至 10.1%,减少 2.1%;相较于 YOLOv26-S 模型,mAP 由 43.2% 提升至 81.7%,增加

38.5%,FNR 减少 25.1%。上述结果验证了所提方法在晶圆缺陷跨尺度检测任务的准确检测。

为直观显示在复杂晶圆缺陷场景下的检测行为差异,选取包含裂纹、区域损伤及划痕等类型缺陷的测试样本进行可视化分析,结果如图 9 所示。其中,左列为含预测框与置信度的检测图像,右列为对应热力图,用以表征模型对缺陷区域的关注程度。图 9(a)的检测结果显示,

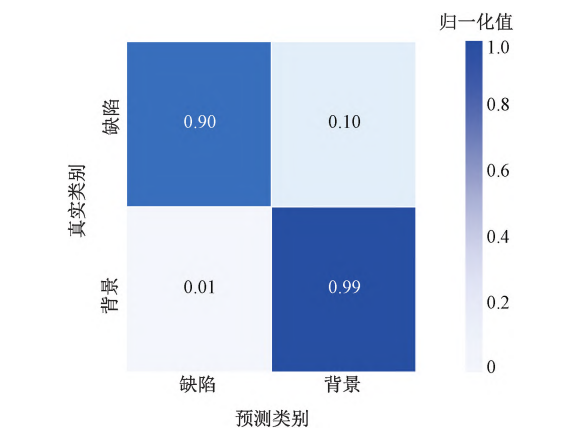


图 7 FLC-DETR 模型的归一化混淆矩阵

Fig. 7 Normalized confusion matrix of the FLC-DETR model

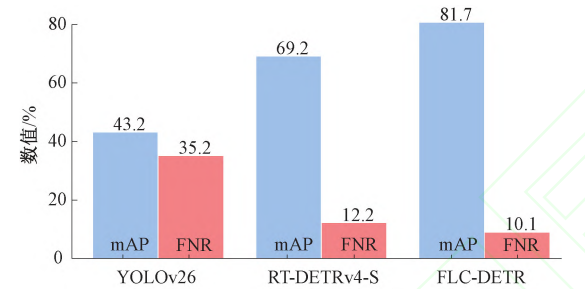


图 8 不同模型的性能对比

Fig. 8 Performance comparison of different models

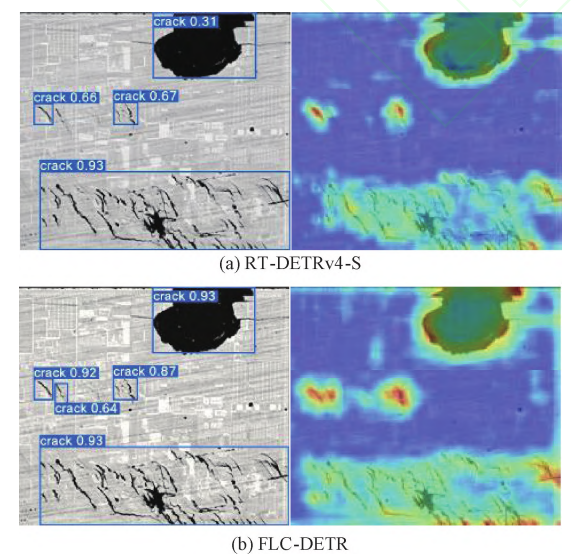


图 9 不同模型晶圆检测结果可视化

Fig. 9 Visualization of wafer detection across different models

RT-DETRv4-S 对显著缺陷定位基本准确,然而低对比度裂纹与弱边缘划痕存在漏检,且预测置信度偏低。其热力图呈现弥散响应,背景中可见大量非目标激活,表明模

型易受晶圆表面复杂纹理干扰。对于各向同性缺陷,热力图未能形成紧凑的高响应区,反映出全局注意力机制在局部边缘保持方面存在局限。

相较而言,图 9(b)中 FLC-DETR 展现出更优的定位精度与置信度。LEM 和 CGM 模块对水平与垂直方向通道特征传输的增强,使模型对细长裂纹和弱边缘划痕实现高置信度稳定检测。通过观察微小裂纹区域,模型可捕捉细弱而连续的边界,形成尖锐连续的激活轨迹。与此同时,LEM 的双分支增强结构未对各向同性缺陷特征产生抑制,以圆形区域损伤为例,FLC-DETR 的热力图呈现与缺陷轮廓高度契合的紧凑激活区,表明方向性边缘增强与通用特征提取实现了协同计算。对于大尺度差异缺陷,模型基于因子图优化的结构化约束,聚合多尺度语义信息为紧凑响应中心,避免了传统 Transformer 独立解码引发的定位漂移。

3.4 消融实验

本文通过消融实验评估各模块对模型性能的贡献程度,不同模型的训练精度曲线如图 10 所示。方法 1 为基线模型 RT-DETRv4-S,前期精度快速上升,后期曲线逐渐平稳时 mAP 为 69.2%,表明模型已充分训练。但在晶圆检测场景下,微裂纹表现为低对比度、边界模糊且尺度极端变异,与通用目标检测存在显著差异,导致预训练模型在特征迁移时面临严重的分布偏移,难以直接适配微裂纹的弱判别性特征。方法 6 建立针对微裂纹的判别性特征表示,收敛速度优于基线配置。在训练中期,由于各模块间特征空间的重组与参数适配,精度提升速率出现阶段性放缓。在训练后期,经过充分协同优化的模型迅速超越所有对比方案,并稳定收敛至最高 mAP 为 81.7%。这表明多模块优化弥补了基线模型在晶圆缺陷场景中的特征提取短板,加速早期有效学习,保障模型的最终收敛质量与检测上限。

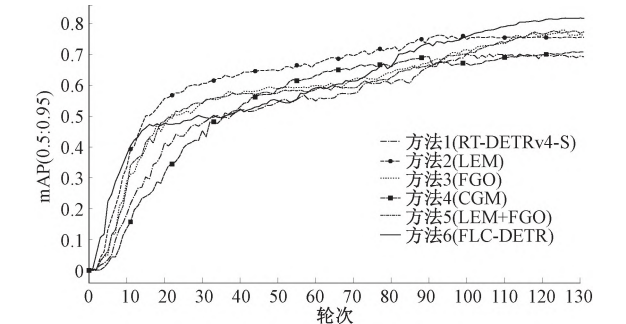


图 10 对比模型训练 mAP 曲线

Fig. 10 Compare the mAP curves of model training

消融实验结果如表 3 所示,方法 2 引入 LEM 后, mAP 提升至 75.5%,较基线提高 6.3%,FNR 降至 10.9%,降低 1.3%,表明边缘增强机制能够有效补偿微

表 3 消融实验结果

Table 3 Results of ablation study

方法	LEM	FGO	CGM	mAP/%	AP@ 0. 5/%	FNR/%
1				69. 2	87. 9	12. 2
2	✓			75. 5	93. 3	10. 9
3		✓		76. 2	95. 2	12. 6
4			✓	70. 8	90. 1	11. 4
5	✓	✓		77. 3	95. 9	12. 0
6	✓	✓	✓	81. 7	98. 6	10. 1

为验证跨尺度非对称模型检测机制的有效性,在由边缘响应、频率响应与特征幅值构成的三维特征空间中,对比

了 RT-DETRv4-S 与 FLC-DETR 在测试集上的特征分布差异,如图 11 所示。为系统评估模型对不同尺度缺陷的表征能力,依据先验统计分析将缺陷面积区间划分为三类:小尺度 ($\leq 100 \text{ pixel}^2$)、中尺度 ($100 \text{ pixel}^2 \sim 9\,000 \text{ pixel}^2$) 以及大尺度 ($\geq 9\,000 \text{ pixel}^2$)。

在小尺度缺陷场景下,稀疏的 KNN 连接抑制了 FGO 运动因子的激活,使得观测项与先验项主导能量最小化过程。如图 11(a) 所示,FLC-DETR 在特征幅值维度上集中分布于 190~200 区间,显著高于 RT-DETRv4-S 的 175~190 区间,且点云方差更小。同时,FLC-DETR 与灰色背景样本点的分布边界清晰,分离度明显提升。这表明观测先验约束能够有效放大微弱缺陷信号并抑制背景噪声,从而增强幅值维度的可区分性。

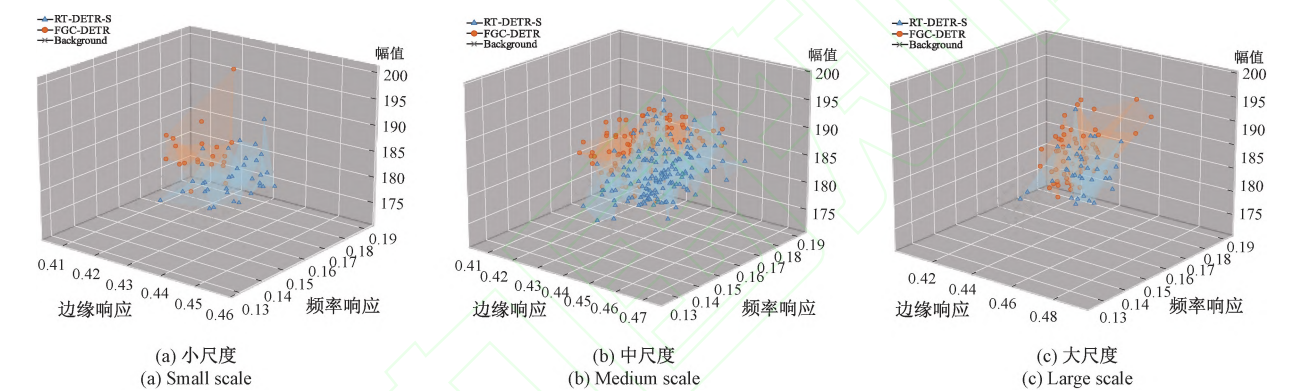


图 11 多尺度三维频域特征分布对比

Fig. 11 Multi-scale 3D frequency domain feature distribution comparison

随着缺陷尺度增大,KNN 邻域内有效节点数量增加,运动因子被激活并参与推理。如图 11(b) 所示,尽管两类模型的点云存在部分重叠,但 FLC-DETR 在频域维度上呈现更紧密的聚集趋势,相较于 RT-DETRv4-S 的分散分布,其方差显著衰减。说明邻域消息传递机制压缩了频域维度的散布范围,体现了局部一致性约束对中等尺度目标的稳定化作用。此时 LEM 提供的边缘先验与 FGO 的结构化约束形成互补,避免了单一运动因子因缺乏边缘支撑而导致的过度抑制。

在大尺度缺陷场景下,密集候选框带来了丰富的 KNN 连接,运动因子成为主导项。如图 11(c) 所示,FLC-DETR 在时域边缘响应 0.42~0.47、频域响应 0.14~0.18,双维度上均形成紧凑流形,点云分散度显著降低。而 RT-DETRv4-S 时域边缘响应 0.42~0.48、频域响应 0.13~0.19,仍保持分散状态并存在孤立离群点。这证实了 FGO 模块通过三因子联合约束,聚合独立候选框的分散响应,有效缓解了独立解码固有的随机漂移问题。

不同缺陷尺度下的检测性能对比如图 12 所示。所提方法在小、中、大 3 种尺度缺陷上的 AP@0.5 分别达到

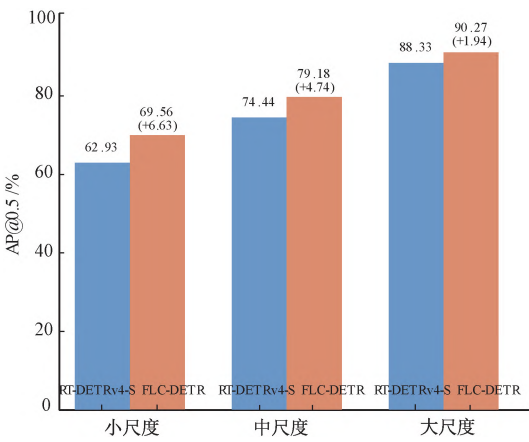


图 12 不同尺度下模型性能对比

Fig. 12 Comparison of model performance at different scales

69.56%、79.18% 与 90.27%,相较于 RT-DETRv4-S 分别提升 6.63%、4.74% 与 1.94%。其中小尺度缺陷的增益最为显著,表明 LEM 的拉普拉斯边缘增强与 CGM 的梯度嵌入机制有效缓解了微裂纹特征衰减与定位漂移问

题。中尺度增益体现了 FGO 因子图优化通过 KNN 邻域消息传递压缩分散响应、建立局部一致性的作用。FLC-DETR 通过运动因子的邻域共识机制进一步降低了边界框随机分散,压缩了剩余误差空间。总体而言,FLC-DETR 在跨尺度全区间内均实现了稳定增益,尤其显著改善了小尺度缺陷的检测瓶颈。

3.5 不同分辨率适应性分析

先进制程 AOI 设备采集的晶圆全视场图像分辨率跨度极大。在推理计算中,多次下采样操作会显著损失亚微米级微裂纹的空间细节,而高分辨率整图推理则对显存容量提出严苛要求。为评估 FLC-DETR 在真实 AOI 硬件上的部署可行性,需系统分析输入分辨率与显存占用、检测精度之间的权衡关系。在 NVIDIA GeForce RTX 5060 Ti (16GB) 平台上,采用 FP32 和自动混合精度 (automatic mixed precision, AMP) 实验对比,测试不同输入分辨率的 mAP 数值变化,结果如图 13 所示。

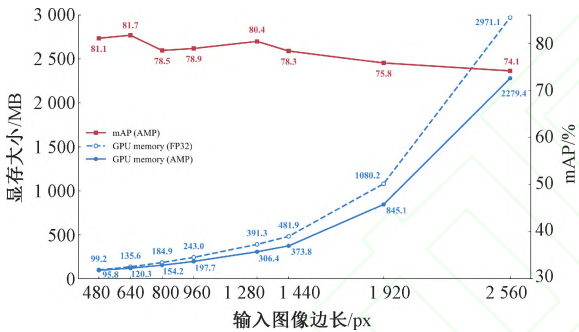


图 13 不同分辨率下显存与 mAP 对比

Fig. 13 GPU Memory vs. mAP under Different Resolutions

随着输入分辨率提升,所提模型 mAP 在 640×640 pixels 下达到峰值 81.7%, 1280×1280 pixels 出现次高峰 80.4%,此后随着输入分辨率增大而逐渐递减,在 2560×2560 pixels 降至 74.1%。该趋势表明,当输入分辨率接近训练多尺度分布上限时模型泛化性能最优;继续增大虽保留了更细粒度的像素信息,但超出训练分布导致的域偏移与自注意力梯度优化难度加剧,反而抑制了检测精度。

在显存占用中,实验仅统计已分配张量实际占用的显存,涵盖模型参数、输入数据、各层激活值及中间计算产生的临时张量,不含 CUDA 驱动与缓存池容量。FP32 与 AMP 模式下的显存均随分辨率近似线性增长,但增长斜率存在显著差异。在低分辨率区间,两种模式的显存占用接近,AMP 需要额外维护 FP32 主参数副本,显存节省有限。从 1280×1280 pixels 开始,激活内存成为显存计算主体,AMP 的 FP16 激活存储优势充分释放,平均可节省约 20% 显存占用。此外,FLC-DETR 模型支持

ONNX 导出与 CUDA 推理,无需专用加速硬件。综合显存与精度约束, 640×640 pixels 可用于在线高速检测,通过切片策略支持任意分辨率,验证了模型面向真实 AOI 设备部署的高效率与分辨率适应性。

4 结 论

针对晶圆图像分辨率与缺陷尺寸跨度极大的检测需求,基于实际 AOI 产线图像构建了 4 660 张图像的晶圆缺陷数据集,为模型训练与评估提供了可靠的数据支持。在此基础上,提出一种基于因子图优化的晶圆缺陷跨尺度检测模型 FLC-DETR,通过融合拉普拉斯边缘增强模块 LEM,增强微缺陷边缘特征提取能力;利用因子图优化模块 FGO,提升跨尺度特征对齐效率;并引入通道梯度增强模块 CGM,优化边界回归阶段的定位精度。最后,在自建数据集上开展的多尺度与不同输入分辨率实验结果表明,FLC-DETR 在检测精度和漏检控制方面均优于主流实时检测模型,证明了所提方法的有效性和模型的实用价值。未来,可进一步探索模型的自适应邻域动态调整机制,并引入多模态传感数据融合以提升复杂产线环境下的模型泛化能力。

参考文献

[1] Fang T Y, An J S, Qi C, et al. Progress and comparison in nondestructive detection, imaging and recognition technology for defects of wafers, chips and solder joints[J]. Nondestructive Testing and Evaluation, 2023, 39(6): 1599-1654.

[2] 杜先君, 贾龙. 融合改进多头注意力与残差结构的 VGGNet 晶圆缺陷检测[J]. 电子测量与仪器学报, 2025, 39(8): 1-12.

Du Xianjun, Jia Long. VGGNet wafer defect detection with improved multi-head attention and residual structure[J]. Journal of Electronic Measurement and Instrumentation, 2025, 39(8): 1-12.

[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 Zhiqiang, Wu Yiquan. Research progress in chip defect detection based on machine vision[J]. Chinese Journal of Scientific Instrument, 2024, 45(7): 1-26.

[4] Mei S, Diao Z L, Liu X Y, et al. BDS-Net: An efficient and high-precision anomaly detector for real-time semiconductor wafer vision inspection [J]. IEEE Transactions on Semiconductor Manufacturing, 2025, 38(3): 675-686.

- [5] Jin C H, Na H J, Piao M H, et al. A novel DBSCAN-based defect pattern detection and classification framework for wafer bin map[J]. IEEE Transactions on Semiconductor Manufacturing, 2019, 32(3): 286-292.
- [6] Yu R Y, Li H Y, Guo B Y, et al. Background-weaken generalization network for few-shot industrial metal defect segmentation[J]. IEEE Transactions on Instrumentation and Measurement, 2025, 74: 5010811.
- [7] Chou J H, Wang F K, Chen H K. Adaptive transfer learning with deformable convolution for wafer defect pattern recognition[J]. IEEE Transactions on Semiconductor Manufacturing, 2025, 38(4): 840-849.
- [8] Dekhovich A, Soloviev O. SupConWI-RL: Wafer inspection with reinforcement learning enhanced by supervised contrastive learning [C]//2025 IEEE International Conference on Computer Vision Workshops, 2025: 1396-1405.
- [9] Zheng J B, Dang J T, Zhang T. Wafer surface defect detection based on feature enhancement and predicted box aggregation[J]. Electronics, 2022, 12(1): 76.
- [10] Wang Z Y, Yang Y C, Chu J N, et al. YOLO-LA: Prototype-based vision-language alignment for silicon wafer defect pattern detection [J]. Micromachines, 2026, 17(1): 67.
- [11] Liao Z J, Zhao Y A, Shan X, et al. RT-DETRv4: Painlessly furthering real-time object detection with vision foundation models[PP/OL]. V1. (2025-10-29)[2026-07-09]. <https://doi.org/10.48550/arXiv.2510.25257>.
- [12] Zhong X, Sun Y, Zhou G C, et al. Statistical-analysis-based imaging-free method for efficient and fast wafer defect detection [J]. IEEE Transactions on Instrumentation and Measurement, 2025, 74: 2537306.
- [13] Liu J M, Zhao H, Wu Q Z, et al. Patterned wafer defect inspection at advanced technology nodes [J]. Laser & Optoelectronics Progress, 2023, 60(3): 0312003.
- [14] Park S Y, Kim T S. Fuzzy inference system for interpretable classification of wafer map defect patterns[J]. Electronics, 2026, 15(1): 130.
- [15] Wang J, Yu Z Y, Duan Z Z, et al. A sub-region one-to-one mapping (SOM) detection algorithm for glass passivation parts wafer surface low-contrast texture defects[J]. Multimedia Tools and Applications, 2021, 80: 28879-28896.
- [16] Zhao Z Y, Wang J, Tao Q, et al. An unknown wafer surface defect detection approach based on incremental learning for reliability analysis [J]. Reliability Engineering & System Safety, 2024, 244: 109966.
- [17] Chi Y L, Gong X Y, Zhao B, et al. CISC-YOLO: A lightweight network for micron-level defect detection on wafers via efficient cross-scale feature fusion [J]. Electronics, 2025, 14(19): 3960.
- [18] Thuan N D. Inspection of mixed-type wafer defects from single-defect data via adaptive synthesis and self-attention CNN[J]. Measurement, 2025, 256: 118306.
- [19] Xu A, Li Y W, Xie H B, et al. Optimization and validation of wafer surface defect detection algorithm based on RT-DETR[J]. IEEE Access, 2025, 13: 39727-39737.
- [20] Zhang H T, Cao S R, Zou N M, et al. FALCO-WAFER: Feature-aware lightweight contextual detector for wafer defect detection[C]//2025 IEEE International Test Conference in Asia, 2025: 43-47.
- [21] Wang Q L, Wu B G, Zhu P F, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11534-11542.
- [22] Shabani H, Balasubramanian M. A novel factor graph-based optimization technique for stereo correspondence estimation[J]. Scientific Reports, 2022, 12: 15613.
- [23] Zong Z F, Song G L, Liu Y. DETRs with collaborative hybrid assignments training [C]//2023 IEEE/CVF International Conference on Computer Vision, 2023: 6748-6758.

作者简介



陈瑞, 2018 年于安徽建筑大学获得学士学位, 2022 年于西安科技大学获得硕士学位, 现为电子科技大学博士研究生, 主要研究方向为集成电路可靠性与测试技术。

E-mail: ruichen26@std.uestc.edu.cn

Chen Rui received his B.Sc. degree from Anhui Jianzhu University in 2018, and his M.Sc. degree from Xi'an University of Science and Technology in 2022. He is currently a Ph.D. candidate at the University of Electronic Science and Technology of China. His main research interests include integrated circuit

reliability and test technology.



龙兵 (通信作者), 分别于 1997 年和 2002 年在哈尔滨工程大学获得学士和硕士学位, 2005 年于哈尔滨工业大学获得博士学位, 现为电子科技大学教授, 主要研究方向为电路与系统测试、诊断技术。

E-mail: longbing@uestc.edu.cn

Long Bing (Corresponding author) received his B. Sc. and M. Sc. degrees both from Harbin Engineering University in 1997 and 2002, respectively, and received his Ph. D. degree from Harbin Institute of Technology in 2005. He is currently a professor at the University of Electronic Science and Technology of China. His main research interests include testing and diagnosis technology for circuits and systems.



戴志坚, 2002 年于电子科技大学获得硕士学位, 现为电子科技大学教授, 主要研究方向为集成电路测试。

E-mail: daizhijian@uestc.edu.cn

Dai Zhijian received his M. Sc. degree from the University of Electronic Science and Technology of China in 2002. He is currently a professor at the University of Electronic Science and Technology of China. His main research interest is integrated circuit testing.



詹文法, 2000 年于安庆师范学院获得学士学位, 2004 年于合肥工业大学获得硕士学位, 2009 年于合肥工业大学获得博士学位, 现为安庆师范大学教授, 主要研究方向为集成电路测试。

E-mail: zhanwf@aqnu.edu.cn

Zhan Wenfa received his B. Sc. degree from Anqing Normal University in 2000, his M. Sc. degree from Hefei University of Technology in 2004, and his Ph. D. degree from Hefei University of Technology in 2009. He is currently a professor at Anqing Normal University. His main research interest is integrated circuit testing.



李荣杰, 2019 年于安徽工业大学获得学士学位, 2022 年于电子科技大学获得硕士学位, 现为电子科技大学博士研究生, 主要研究方向为集成电路测试。

E-mail: 1720313118@qq.com

Li Rongjie received his B. Sc. degree from Anhui University of Technology in 2019, and his M. Sc. degree from the University of Electronic Science and Technology of China in 2022. He is currently a Ph. D. candidate at the University of Electronic Science and Technology of China. His main research interest is integrated circuit testing.