DOI: 10.13382/j.jemi.B2408021

基于多尺度特征提取和注意力机制的轻量化晶圆缺陷检测方法*

任 杰[1] 迟荣华[2] 李红旭[2]

(1.南京信息工程大学电子与信息工程学院 南京 210044;2.无锡学院江苏省通感融合光子器件及系统集成工程研究中心 无锡 214105)

摘 要:在半导体制造中,晶圆图缺陷检测至关重要,能够对缺陷进行快速定位,实现对缺陷的识别,对于提升晶圆产品质量和生产效率具有意义。然而,现有方法存在局限性,如模型过于庞大,网络模型深度过深,难以充分利用多层次特征进行精确分类。为了解决这些问题,结合了 Stem-Dense 特征提取模块和多尺度注意力特征融合结构,提出了一种新型网络结构——MSD-DFE。MSD-DFE 通过 Stem-Dense 的密集连接结构和多尺度注意力特征融合技术,有效提取丰富的浅层特征信息,同时显著降低模型的参数量和计算复杂度。多尺度特征提取模块融合了不同尺度下的晶圆图信息,增强了模型对不同层次缺陷特征的提取能力。此外,引入的注意力机制使得模型能够更关注晶圆图存在缺陷区域,从而提升分类精度。实验结果表明,在减少参数量和计算量的前提下,MSD-DFE 在 WM-811K 数据集上达到了 97.4%的平均准确率,优于现有主流方法,表明其在实际生产环境中具有较高的应用潜力。

关键词:晶圆缺陷;多尺度特征提取;注意力机制;深度学习

中图分类号:TP391.41;TN405 **文献标识码**:A **国家标准学科分类代码**:510.4050

Lightweight wafer defect detection method based on multi-scale feature extraction and attention mechanism

Ren Jie[1] Chi Ronghua[2] Li Hongxu[2]

(1. School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; 2. Jiangsu Engineering Research Center for Sensor Fusion Photonic Devices and System Integration, Wuxi University, Wuxi 214105, China)

Abstract: In semiconductor manufacturing, wafer map defect detection is crucial for the rapid localization and identification of defects, which is significant for enhancing wafer product quality and production efficiency. However, existing methods have limitations, such as overly complex models and excessively deep network structures that struggle to leverage multi-level features for accurate classification. To address these issues, this paper combines a Stem-Dense feature extraction module with a multi-scale attention feature fusion module to propose a novel network architecture—multi-scale defect detection network with enhanced feature extraction (MSD-DFE). MSD-DFE effectively captures rich shallow feature information through the dense connection structure of Stem-Dense and multi-scale attention-based feature fusion technology, while significantly reducing the number of parameters and computational complexity of the model. The multi-scale feature extraction module integrates wafer map information from various scales, enhancing the model's ability to extract defect features. Additionally, the introduced attention mechanism allows the model to focus more on defect areas, thereby improving classification accuracy. Experimental results show that MSD-DFE achieves an average accuracy of 97.4% on the WM-811K dataset, outperforming current mainstream methods, indicating its high potential for practical application in industrial settings.

Keywords: wafer defect; multi-scale feature extraction; attention mechanism; deep learning

收稿日期:2024-12-06 Received Date: 2024-12-06

*基金项目:江苏省基础研究计划重点项目(BK20243021)、江苏省产学研合作项目(BY20230745)、江苏省高等学校基础科学研究面上项目(22KJB510043)、无锡市科技创新创业资金"太湖之光"科技攻关计划(K20241049)、无锡学院引进人才科研启动专项经费(550222001,550223012)项目资助

0 引 言

半导体制造是现代电子工业中的核心环节,而晶圆缺陷检测是确保芯片质量和生产效率的关键任务[1-3]。随着半导体技术的快速发展,晶圆缺陷检测面临着越来越复杂的挑战,尤其是缺陷类型的多样性和图像数据量的迅速增加[4]。传统的缺陷检测方法多依赖人工检查或基于简单图像处理技术,如阈值分割和边缘检测,这些方法不仅效率低下,而且难以应对形态复杂、分布不均的缺陷[5-6]。因此,开发高效、精确的自动化检测方法成为当前的研究热点[7-8]。

近年来,学术界和工业界提出了多种自动化、智能化的方法,主要分为无监督学习和有监督学习两大类[9]。在无监督学习方面,Jin 等[10]采用了密度聚类算法(density-based spatial clustering of applications with noise,DBSCAN),通过分析数据的密度对缺陷进行聚类。这种方法能够有效地处理随机分布的缺陷,但由于缺乏对特征深层语义信息的充分挖掘,DBSCAN 在面对复杂缺陷模式时,其分类精度较低。Kim 等[11]提出了一种基于连通路径的过滤算法,结合无限混合模型(infinite mixture model)进行混合缺陷的聚类。尽管该方法在处理复杂缺陷时有所提升,但其较强的先验知识依赖性使得在缺乏明显特征或特征分布高度重叠的情况下,性能往往不理想。此外,其他方法如自适应共振理论(adaptive resonance theory, ART)[12]、K 均值聚类算法(K-means clustering algorithm)[13]以及自组织映射(self-organizing map, SOM)[14]等,也在动态聚类领域进行了探索,但这些方法普遍依赖于样本特征的相似性,难以有效捕捉特征间的深层次关联。在处理复杂、多变的缺陷模式时,这些方法存在显著的局限性,无法充分应对缺陷检测任务中的多样性和复杂性。

相较于无监督学习,有监督学习方法通常在分类精度上更为出色。深度学习尤其是卷积神经网络(convolutional neural network, CNN)已被广泛应用于图像分类任务中,并在晶圆缺陷检测中取得了显著成果。CNN 通过自动学习图像中的特征,能够有效提高检测精度。近些年,许多学者提出了基于 CNN 的缺陷检测模型。Nakazawa 等[15]提出的 5 层卷积神经网络能够识别 22 种晶圆图缺陷模式;Tsai 等[16]开发了一种轻量级缺陷识别卷积网络,实现了在部分缺陷模型下的识别;Kang 等[17]提出了结合手工与卷积特征识别的多类型缺陷的特征融合模型;Manivannan[18]提出了基于轻量级 ResNet-10 的双头 CNN 的晶圆缺陷检测模型。然而,上述提到的模型虽然使用了 CNN 对于晶圆图缺陷检测识别的优秀特征提取能力,但是存在着无法兼具提高检测精度和解决模型过于庞大、计算复杂度高和参数量巨大的问题。对此,Chen 等[19]提出了使用堆叠网络层数的深度卷积神经网络(deep convolutional neural network,DCNN)识别晶圆图缺陷检测模型,使用 19 层端到端的网络解决精度过低的问题,但是在一些缺陷的监测还是准确率偏低。付强等[20]则是针对模型计算复杂度和参数量,提出了一种基于可分离和注意力机制的晶圆缺陷检测方法(wafer defect detection-separable convolution and attention,WDD-SCA),此方法使用深度可分离卷积降低模型的参数量,提高模型的推理速度,但是由于其模型主体还是建立在传统的 CNN 模型上,模型的精度还是有待提升。

随着晶圆检测场景对实时性、泛化能力和检测精度的要求不断提升,研究者们开始尝试在轻量化设计与多尺度特征融合方向进行改进。轻量化模型的研究如 MobileNet[21]、PeleeNet[22]等,通常通过深度可分离卷积等技术来降低模型参数和计算量,具备嵌入式设备的实际部署潜力。然而,在应用中,这些模型在特征提取和细节捕捉方面的能力仍不够完善。为解决上述问题,在提高模型精度的同时尽可能的降低模型的计算复杂度与参数量,本文提出了一种基于多尺度特征提取和 SE(squeeze-and-excitation)注意力机制[23]的轻量化网络模型——MSD-DFE(multi-scale defect detection network with enhanced feature extraction)。该模型通过引入 Stem-Dense 特征提取模块和多尺度融合结构,有效提升了对晶圆缺陷的检测精度,同时保持较低的参数量和计算量。本文提出的网络结合了多尺度特征提取模块和 SE 注意力机制,有效提升了晶圆缺陷的检测精度,同时在减少参数量和计算复杂度方面表现出色,适用于资源受限的应用场景。提出了一种网络宽度扩展策略,有效提升了 CNN 的分类性能,同时减轻了过拟合风险,并加快了训练速度,为提高晶圆缺陷检测的效率和准确性提供了新途径。

1 轻量化多尺度网络模型设计

1.1 整体网络架构设计

从以往研究中得知,增加 CNN 网络的深度可以提升其特征学习能力。然而,网络深度的增加同时带来学习速度减慢、过拟合风险提升及特征提取能力受限等问题。因此,单纯依靠增加网络深度并非提升性能的最佳方法。本文从提升网络宽度的设计入手,提出了一种多尺度特征提取与融合结构。通过扩大网络的宽度,提高了网络的分类性能,降低了网络深化带来的过拟合风险,提高了网络训练速度。图 1 为本文提出的轻量化多尺度网络模型结构。该结构通过 3 种不同大小的卷积核提取晶圆图的多尺度特征,提高了晶圆缺陷检测中不同尺度特征的

利用率,从而提升了模型的准确性和效率。利用 3 个不同尺寸卷积(3×3、5×5 和 7×7)的 Conv1-SE 层进行特征并行提取,卷积次数分别设置为 3 次、2 次和 1 次,不仅扩展了网络宽度,还有效降低了因网络加深导致的过拟合风险。通过多尺度卷积核设计,特征提取模块可获取不同感受野的特征信息,有助于捕捉多样化的缺陷特征,并增强模型对复杂缺陷模式的适应性。最终,3 个不同尺寸卷积核的输出与 Stem-Dense 模块经过上采样调整特征图大小后的结果进行残差密集连接与融合,再经卷积、Spatial-DropOut 进行 DropOut、全连接及 SoftMax 层完成晶圆缺陷的分类。这种结合了残差连接的多尺度特征提取设计确保了信息传递与特征重用的最大化。此外,为了解决晶圆缺陷检测中的类别不平衡问题,本文采用 Focal Loss 作为损失函数[24],旨在提高模型对稀有缺陷类别的检测精度。相较于传统的交叉熵损失,Focal Loss 能够动态调整样本权重,通过重点关注难以分类的样本,显著缓解数据不平衡对模型性能的负面影响。通过这些改进,MSD-DFE 在复杂缺陷的检测任务中表现出色,有效提升了分类精度和模型鲁棒性。

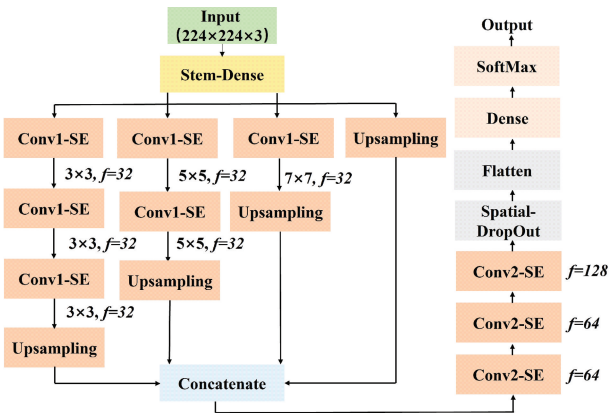


图 1　MSD-DFE 结构示意图

Fig. 1　Diagram of the MSD-DFE architecture

1.2　Stem-Dense 特征提取模块

Stem-Dense 特征提取模块来自于一种轻量级卷积神经网络——PeleeNet,借鉴了 DenseNet[25] 的设计思路,并在结构上进一步优化,以提升计算效率并实现模型的压缩。PeleeNet 在显著减少参数量和计算量的同时,依然保持了出色的分类精度和性能。Stem-Dense 特征提取模块结构如图 2 所示,Stem-Dense 模块在初步特征提取中,显著减少了参数量和计算复杂度,从而实现模型的轻量化设计。

Stem 模块作为初始卷积层,将输入图像转换为特征表示。不同于传统卷积网络,Stem 模块采用高效的双路径密集结构,结合不同步长和卷积核尺寸的并行特征提取,以及最大池化操作,从而保留输入图像中的关键信息。双路径密集连接结构使得多尺度特征图在融合后被有效整合,不仅显著降低了模块参数量,还能够高效提取多个尺度的浅层特征。此设计在确保计算效率的同时,有效捕捉了多尺度特征信息,为后续卷积层提供了丰富的上下文特征,从而显著提升分类性能。Stem 模块的结构如图 3(a)所示。

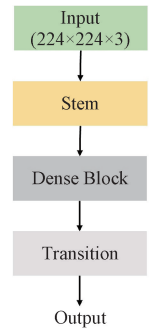
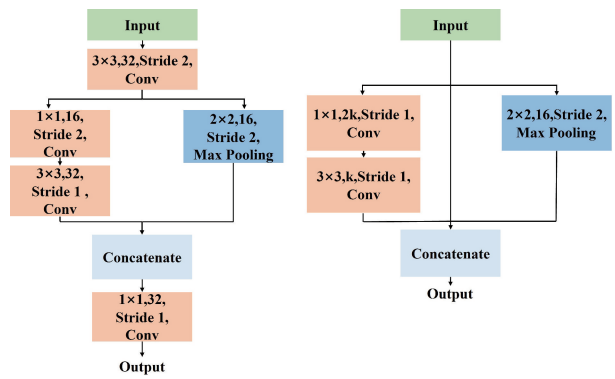


图 2　Stem-Dense 特征提取模块示意图

Fig. 2　Diagram of the Stem-Dense feature extraction module

Dense 模块借鉴 DenseNet 的核心思想,通过密集连接实现信息传递与特征重用的最大化。图 3(b)所示为 Dense 模块的结构。每一层的输入不仅来自直接前一层的输出,还包括所有先前层的输出。这样的密集连接使得每一层都能直接访问最初的输入和所有中间层的特征图,从而确保前层提取的特征能被后续层充分利用。此外,由于每层都与之前的所有层直接相连,Dense 模块在训练过程中还具有显著的正则化效果,有效减少了过拟合现象,从而提升了模型的泛化性能。通过这种特征重用和梯度流动设计,能够在保持较少参数量的情况下实现较高的模型性能,特别适用于计算效率与高精度需求并存的应用场景。



(a) Stem模块　(b) Dense模块

(a) Stem module　(b) Dense module

图 3　Stem-Dense 模块结构

Fig. 3　Structure diagram of the Stem-Dense module

此外,在 Stem-Dense 模块最后一层引入了 Transition

层来控制网络的深度和宽度。Transition 层通过降低特征图的空间尺寸,在保留关键信息的同时有效地减少了计算开销。通过减少通道数量和特征图尺寸,Transition 层有效缓解了参数量和计算量的爆炸性增长,实现了性能与效率之间的良好平衡。

1.3 Conv-SE 模块

在 Stem-Dense 初步特征提取后,本文采用了一种结合 SE 注意力机制的卷积模块——Conv-SE 模块,以实现深层特征图的高效信息提取。SE 注意力机制在特征提取中通过动态赋予不同通道权重,使模型聚焦于重要特征通道,从而提升学习效率和泛化能力。值得一提的是,该机制在显著提升模型性能的同时,计算量几乎未增加。

图 4 所示为 Conv-SE 模块的结构,其中图 4(a)为 Conv1-SE,图 4(b)为 Conv2-SE,W 表示卷积核大小,f 表示卷积核个数,*Stride* 表示步长。Conv1-SE 和 Conv2-SE 模块的工作原理如下:首先,输入特征图经过特定大小卷积核的卷积、激活、正则化并在空间维度上通过最大池化,调整输出的通道数;其次,这些描述经由 SE 注意力层处理,生成通道权重,这些权重被作用于原始特征图的各通道,对关键特征进行加权强化,同时抑制无关信息。

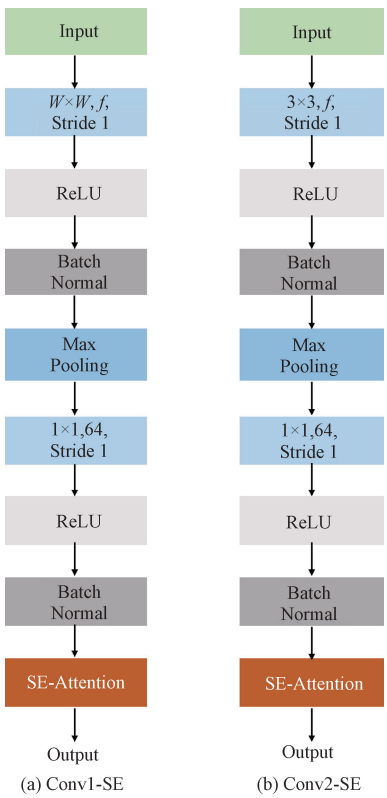


图 4 Conv1-SE 和 Conv2-SE 结构示意图

Fig. 4 Diagram of the Conv1-SE and Conv2-SE structures

图 5 所示为 SE 注意力机制的结构,其中 Scale 操作表示逐通道相乘。该机制通过自适应调整通道权重,增强了卷积神经网络对有用特征的表达能力,同时有效抑制无关或噪声特征。SE 注意力机制操作如下:首先,对输入特征图在空间维度上进行全局平均池化,将每通道的全局信息压缩为单个数值,生成通道级描述;随后,这些描述经由两个全连接层的 Excitation 处理,生成通道权重;最后,权重作用于原特征图的各通道,对关键特征进行加权强化,抑制无关信息。这一设计显著提升了模型对重要特征的关注,增强了特征表达能力。

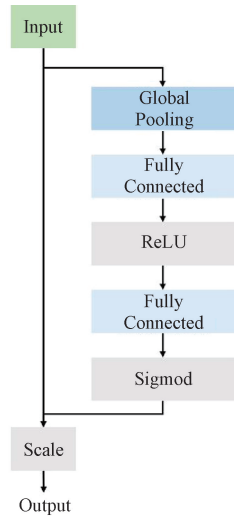


图 5 SE 注意力机制结构示意图

Fig. 5 Diagram of the SE attention mechanism structure

2 WM-811K 数据集介绍及数据预处理

2.1 WM-811K 数据集

WM-811K 数据集[26]是当前全球规模最大且公开可访问的真实半导体制造环境的晶圆图像数据集,包含 811 457 张晶圆图像,这些数据来自 46 293 个批次的采集,每个批次通常包含约 25 张晶圆图像。理论上,数据集应包括 1 157 325 张图像,但由于传感器故障或其他未知原因,部分批次图像缺失,最终形成了当前的规模。WM-811K 数据集不仅提供了丰富的晶圆图像数据和缺陷类别信息,还附带了批次名称、晶圆尺寸、训练标签和测试标签等附加信息,为研究者提供了全面的数据支持。数据集中的晶圆图像尺寸多达 632 种,从(6×21)到(300×202)不等,体现了极大的多样性。这种特性使 WM-811K 成为研究晶圆缺陷检测和分类的重要基准,为不同尺寸的图像处理提供了广泛的应用场景。

该数据集存在明显的数据不平衡现象。在 811 457 张图像中,仅有 172 950 张被标注了缺陷标签,而具有明确缺陷图案的图像数量为 255 519 张。数据集中包含的缺陷标签共分为 9 种类别,各类别的样本数量分布如表 1 所示。这种数据不平衡特性为模型的训练和评价提出了

挑战,同时也为研究高效处理不平衡数据的算法提供了宝贵的实践机会。

表 1　WM-811K 数据集不同缺陷类别的数量分布

Table 1　The distribution of the number of defect categories in the WM-811K dataset

缺陷类别	数量
Center	4 294
Donut	555
Edge-Loc	5 189
Edge-Ring	9 680
Local	3 593
Random	866
Scratch	1 193
Near-full	149
None	147 431

2.2　数据集预处理

WM-811K 数据集的显著类别不平衡性对模型训练提出了重大挑战,尤其是缺陷样本数量远少于无缺陷(none)类别。这种不平衡性导致模型在训练过程中更倾向于预测无缺陷类别,从而显著影响分类的准确性。为解决这一问题,本文在数据预处理中引入了中值滤波去噪和多种数据增强策略,以提升模型在少数类缺陷图像上的识别能力。

WM811-K 中的不同晶圆缺陷如图 6 所示。为了统一数据集的输入格式,将 WM-811K 中尺寸不一的图像调整为固定大小(224×224×3)。原始晶圆图像为单通道形式,其中每个像素的值表示 3 种状态:0 表示未知状态;1 表示正常状态;2 表示缺陷状态。为适应模型需求,采用独热编码将单通道图像扩展为三通道图像。考虑到数据中的随机噪声可能导致误检问题,对独热编码后的图像应用中值滤波操作,以有效去除随机噪声并提高模型对实际缺陷模式的关注。中值滤波处理后的晶圆图像如图 7 所示。

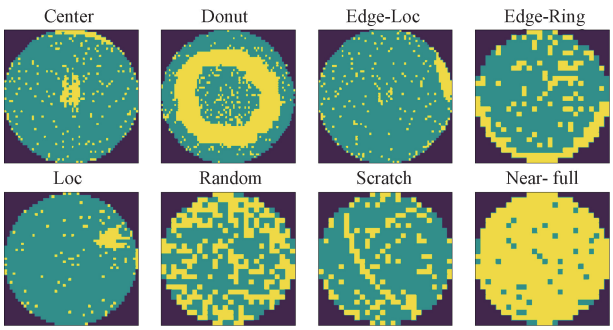


图 6　原始晶圆图

Fig. 6　Original wafer map

此外,为了缓解类别样本不平衡问题,尤其是除无缺陷类别外,其余 8 类样本数量较少的情况,本文通过数据增强技术扩充少数类别的样本量至每类 10 000 张。数据增强方法包括图像的垂直和水平翻转、平移、轻微旋转、缩放以及裁剪等。这些增强操作不仅扩大了样本规模,还增加了样本的多样性,从而提升了模型的泛化能力。

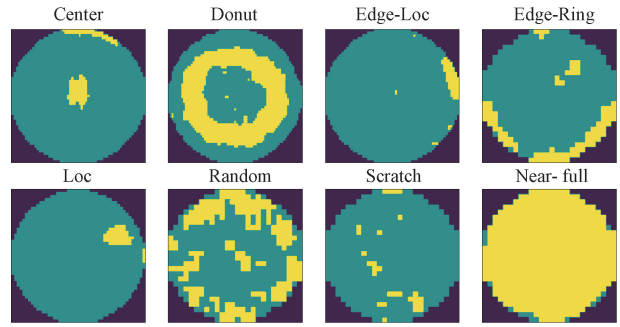


图 7　中值滤波后的晶圆图

Fig. 7　Wafer map after median filtering

3　实验与分析

3.1　实验环境与参数设置

本文实验使用了 TensorFlow 框架构建网络,实验环境包括 Windows11 操作系统、Cuda11.8、Python3.9 以及 TensorFlow 2.9.1,硬件平台为 NVIDIA RTX 4090 GPU。实验所用的数据集包含 9 个类别,每类 10 000 张晶圆图像,总计 90 000 张。数据集按照 70∶15∶15 的比例划分为训练集、验证集和测试集,确保实验结果具有可靠性和普适性。实验图像输入的大小为 224×224×3,学习率设定为 1×10^{-3},批量大小为 100。网络训练过程中进行了 40 个周期,优化器选择了 Adam 梯度优化器,以其良好的收敛性能确保训练稳定性和效率。

损失函数采用 Focal Loss,以有效应对类别不平衡问题,其表达式如式(1)所示。

$$F_L(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

式中:α 和 γ 调节因子;p_t 是模型对类别的正确预测概率。参数 α_t 用于平衡正负样本的权重,通常设置 α_t 为 $\alpha_t = \alpha$ 或 $1-\alpha$,以降低易分类样本对总损失的贡献;参数 γ 控制模型对难分类样本的关注程度,γ 值越大,模型越倾向于聚焦于难分类样本。本文实验中,γ 通常设为 2,以实现较优的样本权重平衡和分类性能。

3.2　模型指标

在图像分类任务中,评价模型性能的常用指标包括准确率(accuracy, Acc),精确率(precision, P),召回率(recall, R),$F1$ 分数(F1-Score),其表达式如式(2)~(5)所示。

$$Acc = \frac{TN + TP}{FP + TN + TP + FN} \tag{2}$$

$$P = \frac{TP}{FP + TP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - Score = \frac{2 \times P \times R}{P + R} \tag{5}$$

式中：TP(true positive)表示预测为正样本且实际为正样本的数量；FP(false positive)为预测为正样本但实际为负样本的数量；TN(true negative)为预测为负样本且实际为负样本的数量；FN(false negative)为预测为负样本但实际为正样本的数量。$F1$ 分数是精确率与召回率的加权平均，其值介于 0 和 1 之间，值越大表示模型在分类任务中表现越好。$F1$ 分数尤其适合在数据类别不平衡的情况下，综合衡量模型对正负样本的分类性能。

3.3 实验结果分析

将本文方法与支持向量机(SVM)、人工神经网络(ANN)、VGG16 等传统模型以及近年来的晶圆缺陷检测算法进行了对比，结果如表 2 所示。实验结果表明，本文方法在参数量、精确率、召回率、准确率和 $F1$ 分数等多个指标上都展现了显著优势。

从表 2 可以看出，除 SVM 模型无法进行参数量化外，本文所提出的方法在其他 5 个对比模型中，从参数量、计算量和模型准确度等角度，本文方法具有更好的性价比。本文模型的参数量为 0.646×10^6，远低于传统的大型神经网络模型，如 VGG16(134.297×10^6)和 ANN(102.762×10^6)。即使与一些轻量化网络，如 MobileNetV2(3.4×10^6)和 WM-PeleeNet(0.169×10^6)相比，本文模型的参数量依然保持在极低水平，展示了较高的参数效率，这使得该模型特别适合在资源受限的环境中部署。在计算量方面，尽管由于结构中采用了不同尺寸的卷积核计算，导致本文模型的浮点计算量为 1.634 GFLOPS，相较于一些轻量化网络模型(如 WM-PeleeNet 为 0.316 GFLOPS 和 MobileNetV2 为 0.313 GFLOPS)，计算量略大，但远低于 VGG16(18.48 G)等大型模型。通过这种设计，本文模型在保证高性能的同时，有效降低了计算复杂度，实现了计算效率与准确率之间的良好平衡。在分类精度方面，本文模型达到了 97.4%的准确率，在所有对比模型中表现最佳。相比之下，VGG16 和 MobileNetV2 的准确率分别为 92.8%和 95.8%，而 WDD-SCA 和 WM-PeleeNet 的准确率分别为 96.5%和 93.6%。这些结果表明，本文模型不仅在减少参数量和计算量方面表现出色，还在精度上超越了现有的轻量化模型。本文提出的模型在性能和计算效率之间达到了良好的平衡，说明其在实际应用中巨大的潜力和优势。

表 2 各个模型参数量及准确率比较

Table 2 Comparison of the number of parameters and accuracy for each model

模型	总参数量/10^6	总计算量/GFLOPs	准确率/%
SVM	-	-	70.4
ANN	102.762	-	88.7
VGG16	134.297	18.48	92.8
MobileNetV2	3.400	0.313	95.8
WDD-SCA	75.068	0.640	96.5
WM-PeleeNet	**0.169**	**0.316**	93.6
PeleeNet	23.090	0.768	93.7
本文	0.646	1.634	**97.4**

表 3 为这些模型在精确率、召回率、准确率和 $F1$ 分数等指标上的对比结果。可以看出，本文提出的模型在各项指标上均有显著提升，表现出更强的分类性能。

表 3 不同模型的准确率、查准率、查全率、$F1$ 分数对比

Table 3 Comparison of accuracy, precision, recall, and $F1$-score for each model (%)

模型	训练集准确率	验证集准确率	测试集准确率	精确率	召回率	F1 分数
SVM	71.5	70.4	70.4	72.8	70.4	67.6
ANN	93.9	88.8	88.7	88.7	88.7	88.7
VGG16	**99.9**	92.8	92.8	92.7	92.7	92.6
MobileNetV2	**99.9**	95.8	95.8	95.8	95.8	95.8
WDD-SCA	99.8	96.5	96.5	96.4	96.5	96.5
WM-PeleeNet	99.9	96.3	96.3	96.2	96.5	96.3
本文	**99.9**	**97.4**	**97.4**	**97.4**	**96.9**	**97.3**

表 4 为不同模型在晶圆缺陷分类任务中的精度对比。图 8 所示为本文模型的混淆矩阵，其中横轴代表预测标签，纵轴代表真实标签。每一行表示某类真实标签的晶圆图像在被正确预测的类别数量、错误预测的类别数量，以及被错误预测为其他类别的数量。从表 4 及图 8 可以看出，本文提出的 MSD-DFE 模型在 WM-811K 数据集的 8 个缺陷类型及无缺陷类型(包括 Center、Donut、E-L(Edge-Loc)、E-R(Edge-Ring)、Loc(Local)、Random、Scratch、N-F(Near-Full)、None)的分类准确率分别为 96.2%、95.2%、94.8%、100%、94.9%、100%、95.8%、100%、100%。本文提出的模型尽管在 Center、Donut 和 Scratch 类别的准确率相较于其他模型略有不足，但在其余缺陷类别及无缺陷类别的分类准确率上均优于其他方法，且大部分类别的准确率均达到 95%以上。本文提出的 MSD-DFE 方法在整体性能上展现了最佳表现，证明其在晶圆缺陷检测中的优越性。

3.4 模型的收敛性分析

模型的损失曲线以及在训练集和验证集上的准确率变化曲线如图 9 和 10 所示。从图 9 可以看出，模型的损失值在训练初期迅速下降，并在第 30 个周期后趋于

表 4　不同模型在各个缺陷类别的正确率

Table 4　Accuracy of different models for each defect category　　(%)

模型	Center	Donut	E-L	E-R	Loc	Random	Scratch	N-F	None
SVM	92.7	95.2	54.3	97.1	11.3	**100**	64.7	73.0	45.1
ANN	93.7	99.9	77.2	96.7	66.4	**100**	98.6	86.8	79.2
VGG16	96.0	99.2	85.6	97.9	74.9	**100**	95.3	93.7	92.0
MobileNetV2	97.9	**100**	90.5	97.0	89.7	**100**	**99.2**	94.6	45.1
WDD-SCA	97.9	**100**	91.5	96.7	92.6	**100**	99.1	96.7	94.0
WM-PeleeNet	**98.2**	90.8	90.1	97.8	87.2	89.2	92.3	87.2	**100**
本文	96.2	95.2	**94.8**	**100**	**94.9**	**100**	95.8	**100**	**100**

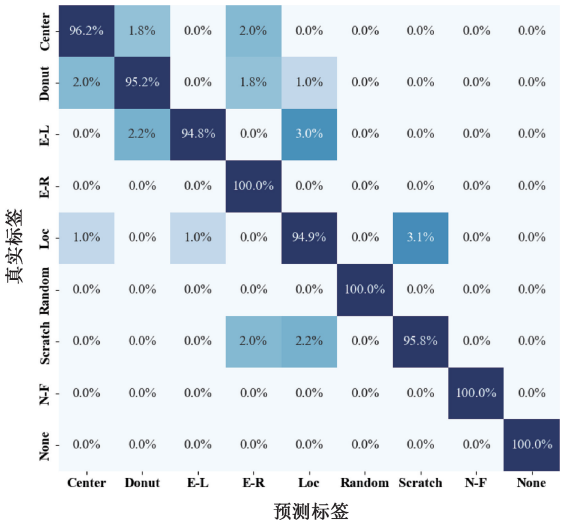


图 8　模型的混淆矩阵

Fig. 8　Confusion matrix of the proposed model

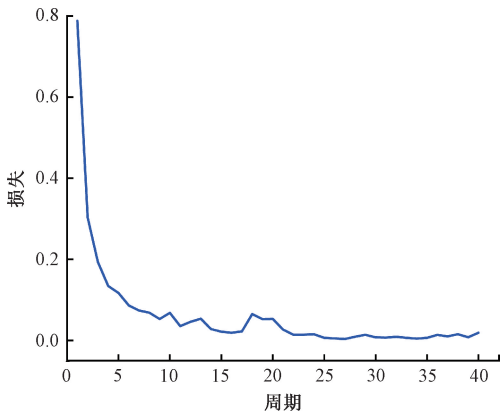


图 9　训练模型的损失曲线

Fig. 9　Loss curve of the trained model

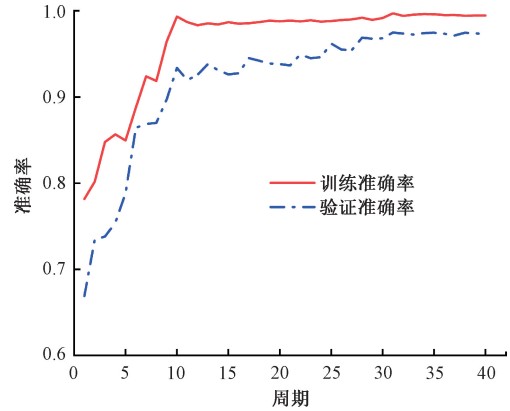


图 10　训练集和验证集上准确率变化曲线

Fig. 10　The accuracy variation curve on the training and validation sets

平稳,表明模型在训练过程中快速收敛。从图 10 可以看出,模型在训练集和验证集上的准确率均呈现出先上升后平稳的趋势。在训练集上,模型的准确率在第 31 个周期达到了最高值 99.9%,并在随后的周期中保持稳定;在验证集上,模型的准确率在第 31 个周期达到了最大值 97.4%。这些结果表明,模型具有较强的收敛性,能够在较少的训练周期内迅速达到高准确率,并在后续训练中保持相对稳定的表现。这表明模型在训练效率和稳定性方面都表现良好。

3.5　消融实验

消融实验对比如表 5 所示,包含所有模块的情况下,模型的总参数量为 0.646×10^6,计算量为 1.634 GFLOPs,并达到了最高平均分类准确率 97.4%。这表明各个模块的协同作用有效提升了模型的分类性能。与不使用 SE 注意力机制的模型相比,采用 SE 注意力机制的模型在准确率上有了显著提升,同时可以观察到,SE-Attention 几乎不增加额外的参数量和计算量。

表 5　MSD-DFE 消融实验对比

Table 5　Ablation study comparison of MSD-DFE

SE	Stem-Dense	Multi-scale Fusion Structure	总参数量/$\times10^6$	计算量/GFLOPs	准确率/%
–	–	–	1.04	2.344	89.4
√	–	–	1.05	2.355	91.2
–	√	–	0.274	0.594	92.4
–	–	√	0.417	1.166	92.6
√	√	–	0.276	0.595	94.3
√	–	√	0.448	1.166	94.4
–	√	√	0.635	1.633	95.8
√	√	√	0.646	1.634	**97.4**

在不包含任何模块的情况下,模型的参数量和计算量分别为 1.04×10^6 和 2.344 GFLOPs,且准确率最低,仅为 89.4%,而使用 Stem-Dense 模块进行特征提取后,模型的参数量和计算量分别为 0.274×10^6 和 0.594 GFLOPs,同时准确率也达到了 92.4%,可以看出 Stem-Dense 模块可以大幅降低模型的参数量和计算量。而与仅使用 Stem-Dense 模块作为初步特征提取模块的模型相比,使用多尺度融合结构后,虽然会使得在计算量和参数量上有所增加,但多尺度融合结构的引入使得模型能够学习到不同尺度的特征信息,特别是小卷积核有助于提取细节信息,而大卷积核则擅长捕捉全局结构。这种多尺度特征融合结构显著提高了模型的鲁棒性和表现力。消融实验结果表明,去除该多尺度融合结构后,进行卷积特征提取时,准确率平均下降了 2.6%。这表明,多尺度融合结构对于捕获全面的特征信息以及提升分类性能至关重要。尽管该结构增加了模型的参数量和计算量,但其对准确率的显著提升证明了其在提升模型性能中的关键作用。

4 结 论

本文提出了一种基于多尺度特征提取与 SE 注意力机制的轻量化晶圆缺陷检测网络——MSD-DFE。通过设计高效的 Stem-Dense 模块、多尺度特征融合结构和 SE 注意力机制,本文模型在保持较低参数量和计算量的同时,实现了高精度的缺陷检测与分类。实验结果表明,MSD-DFE 模型在准确率、精确率、召回率和 $F1$ 分数等指标上,对比传统模型和现有的轻量化网络具有一定优势。通过引入多尺度特征提取和融合,模型能够从不同尺度的特征中捕捉到更多的缺陷信息,增强了对复杂缺陷模式的适应性。消融实验进一步验证了各模块的有效性,表明多尺度融合结构在提升分类性能和鲁棒性方面发挥了关键作用。此外,SE 注意力机制几乎不增加额外的计算负担,却显著提高了模型的性能。尽管本文模型与最轻量化的模型相比增加了参数量和计算量,但整体性能的提升使得该模型在资源受限的环境中具有良好的应用前景。总体而言,本文方法通过精心设计和模块优化,实现了计算效率与分类精度之间的平衡,为晶圆缺陷检测领域提供了一种高效、精准的解决方案,具备较高的实际应用价值。但受限于 Conv1-SE 卷积核大小的不同,模型在设计时并不能实现并行处理,这也导致了模型在降低计算量和参数量方面仍有可优化的空间。

参考文献

[1] CHEN S H, LIU M Q, HOU X N, et al. Wafer map defect pattern detection method based on improved attention mechanism [J]. Expert Systems with Applications, 2023, 230: 120544.

[2] 史浩琛, 金致远, 唐文婧, 等. 基于深度学习的高精度晶圆缺陷检测方法研究[J]. 电子测量与仪器学报, 2022, 36(11): 79-90.
SHI H CH, JIN ZH Y, TANG W J, et al. Research on high-precision wafer defect detection method based on deep learning[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(11): 79-90.

[3] FANG T Y, AN J SH, CHEN Q, 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.

[4] 陈晓雷, 温润玉, 杨富龙, 等. 基于空频特征融合的双流晶圆缺陷分类网络[J]. 电子测量与仪器学报, 2024, 38(8): 56-67.
CHEN X L, WEN R Y, YANG F L, et al. Dual-stream wafer defect classification network based on spatial-frequency feature fusion [J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(8): 56-67.

[5] HANSEN C K, THYREGOD P. Use of wafer maps in integrated circuit manufacturing [J]. Microelectronics Reliability, 1998, 38(6-8): 1155-1164.

[6] JHA S B, BABICEANU R F. Deep CNN-based visual defect detection: Survey of current literature [J]. Computers in Industry, 2023, 148: 103911.

[7] CHIEN C F, CHANG K H, WANG W C. An empirical study of design-of-experiment data mining for yield-loss diagnosis for semiconductor manufacturing[J]. Journal of Intelligent Manufacturing, 2014, 25: 961-972.

[8] 王宸, 杨帅, 周林, 等. 基于自适应多尺度特征融合网络的金属齿轮端面缺陷检测方法研究[J]. 电子测量与仪器学报, 2023, 37(10): 153-163.
WANG CH, YANG SH, ZHOU L, et al. Research on metal gear end face defect detection method based on adaptive multi-scale feature fusion network[J]. Journal of Electronic Measurement and Instrument, 2023, 37(10): 153-163.

[9] TULBURE A A, DULF E H. A review on modern defect detection models using DCNNs-Deep convolutional neural networks[J]. Journal of Advanced Research, 2022, 35: 33-48.

[10] 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.

[11] KIM J, LEE Y M, KIM H Y. Detection and clustering of mixed-type defect patterns in wafer bin maps[J]. IISE

Transactions, 2018, 50(2): 99-111.
[12] CHOI G, KIM S H, HA C H et al. Multi-step ART1 algorithm for recognition of defect patterns on semiconductor wafers [J]. International Journal of Production Research, 2012, 50(12): 3274-3287.
[13] LIUKKONEN M, HILTUNEN Y. Recognition of systematic spatial patterns in silicon wafers based on SOM and K-means[J]. IFAC-PapersOnLine, 2018, 51(2): 439-444.
[14] PALMA F D, NICOLAO G D, MIRAGLIA G, et al. Unsupervised spatial pattern classification of electrical-wafer-sorting maps in semiconductor manufacturing[J]. Pattern Recognition Letters, 2005, 26(12): 1857-1865.
[15] NAKAZAWA T, KULKARIN 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.
[16] TSAI T H, LEE Y C. A light-weight neural network for wafer map classification based on data augmentation[J]. IEEE Transactions on Semiconductor Manufacturing, 2020, 33(4): 663-672.
[17] KANG H, KANG S. A stacking ensemble classifier with handcrafted and convolutional features for wafer map pattern classification[J]. Computers in Industry, 2021, 129: 103450.
[18] MANIVANNAN S. Semi-supervised imbalanced classification of wafer bin map defects using a Dual-Head CNN[J]. Expert Systems with Applications, 2024, 238: 122301.
[19] CHEN S H, ZHANG Y X, HOU X N, et al. Wafer map failure pattern recognition based on deep convolutional neural network[J]. Expert Systems with Applications, 2022, 209: 118254.
[20] 付强, 王红成. 基于可分离卷积和注意力机制的晶圆缺陷检测[J]. 计算机系统应用, 2023, 32(5): 20-27.
FU Q, WANG H CH. Wafer defect detection based on separable convolution and attention mechanism [J]. Computer Systems and Applications, 2023, 32(5): 20-27.
[21] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications [DB/OL]. (2017-04-17) [2024-10-15]. https://arxiv.org/abs/1704.04861.
[22] WANG R J, LI X, LING C X. Pelee: A real-time object detection system on mobile devices[C]. Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018: 1957-1976.
[23] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.
[24] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection [C]. 2017 IEEE International Conference on Computer Vision (ICCV), 2017: 2999-3007.
[25] HUANG G, LIU Z, MAATEN L V D, et al. Densely connected convolutional networks [C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 2261-2269.
[26] WU M J, JANG J S R, CHEN J L. Wafer map failure pattern recognition and similarity ranking for large-scale data sets [J]. IEEE Transactions on Semiconductor Manufacturing, 2015, 28(1): 1-12.

作者简介

任杰,2023 年于烟台大学获得学士学位,现为南京信息工程大学硕士研究生,主要研究方向为晶圆缺陷检测、信号与信息处理。
E-mail: 202312490394@nuist.edu.cn

Ren Jie received his B. Sc. degree from Yantai university in 2023. He is currently a M. Sc. candidate at Nanjing University of Information Science and Technology. His main research interests include wafer defect detection, signal and information processing.

迟荣华,2003 年于南开大学获得博士学位,现为无锡学院教授,主要研究方向为光电检测与信号处理。
E-mail: ronghchi@cwxu.edu.cn

Chi Ronghua received her Ph. D. degree from Nankai University in 2003. She is c a professor at Wuxi University. Her main research interests include photoelectric detection and signal processing.

李红旭(通信作者),2021 年于南京信息工程大学获得博士学位,现为无锡学院讲师,主要研究方向为智能信号处理。
E-mail: hongxuli@cwxu.edu.cn

Li Hongxu (Corresponding author) received his Ph. D. degree from Nanjing University of Information Science and Technology in 2021. He is currently a lecturer at Wuxi University. His main research interests include intelligent information processing.