

DOI: 10.13382/j.jemi.B2407345

基于空频特征融合的双流晶圆缺陷分类网络*

陈晓雷 温润玉 杨富龙 李正成 沈星阳

(兰州理工大学电气工程与信息工程学院 兰州 730000)

摘要:晶圆缺陷模式分类在晶圆制造过程中扮演着至关重要的角色,准确识别晶圆缺陷能够确定缺陷产生的根本原因,进而定位生产流程中的问题。然而,现有深度学习晶圆缺陷分类方法仅从空间域或者频率域出发进行网络设计,未能实现空频信息的相互补充与融合,限制了晶圆缺陷分类准确性的进一步提高。针对这一问题,提出了一种基于空间域和频率域特征融合的双流晶圆缺陷分类网络—SFWD-Net。该网络利用提出的多尺度特征提取卷积模块和多视角注意力模块构成空间流分支提取晶圆图的空间信息,利用离散小波变换构成频率流分支提取晶圆图的频率信息,空频信息融合后再进行缺陷分类。在大规模半导体晶圆图数据集 WM-811K 的实验证明, SFWD-Net 由于同时从空间域和频率域出发进行网络设计,其分类准确度达到 99.299 2%, 优于其他 5 种先进方法,能够显著提高晶圆缺陷分类的精度。

关键词: 晶圆缺陷分类; 双流网络; 小波变换; 注意力机制; 卷积神经网络

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

Dual-stream wafer defect classification network based on spatial and frequency domains feature fusion

Chen Xiaolei Wen Runyu Yang Fulong Li Zhengcheng Shen Xingyang

(College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730000, China)

Abstract: The classification of wafer defect patterns plays a crucial role in the wafer manufacturing process. Accurate identification of wafer defects enables the determination of the root causes of defects, thereby pinpointing issues in the production process. However, existing deep learning-based wafer defect classification methods are designed solely from the spatial or frequency domain, failing to achieve mutual supplementation and integration of spatial and frequency information. This limitation constrains the improvement of wafer defect classification accuracy. To address this issue, a dual-stream wafer defect classification network based on the fusion of spatial and frequency domain features, named SFWD-Net, is proposed. The network utilizes the proposed multi-scale feature extraction convolution module and multi-view attention module to form the spatial stream branch, which extracts spatial information from wafer images. The frequency stream branch, utilizing discrete wavelet transform, extracts frequency information from wafer images. After integrating spatial and frequency information, defect classification is performed. Experiments on the large-scale semiconductor wafer image dataset WM-811K demonstrate that SFWD-Net, by simultaneously designing the network from both spatial and frequency domains, achieves a classification accuracy of 99.299 2%, outperforming five other state-of-the-art methods and significantly improving the accuracy of wafer defect classification.

Keywords: wafer defect classification; dual-stream network; wavelet transform; attention mechanism; convolutional neural network

0 引言

当前互联网技术和数字经济的快速发展下,半导体行业取得了显著进步,并通过更先进的技术和更高效的产量将芯片制造技术不断升级,使芯片产业成为支撑数字经济的重要基础^[1]。晶圆(wafer)是制作硅半导体电路所用的硅晶片,是芯片制造的载体。晶圆图(wafer map)详细展示了芯片制造过程中,晶圆上的单个裸片(die)状态。为了保证晶圆的高良率,缺陷检测工序^[2]显得尤为重要且不可或缺。系统性晶圆缺陷往往以特定模式出现在晶圆图上,通过分析晶圆图缺陷图像,能够更好的查明晶圆缺陷产生的原因,提高产品良率^[3]。传统人工晶圆缺陷识别需要专业人员进行数小时的工作^[4],然而,由于疲劳、情绪等因素,检查人员很难做到长时间集中注意力,致使人工晶圆缺陷识别精度小于45%^[5-6],因此如何利用深度学习等人工智能技术实现高效的晶圆缺陷分类成为亟待解决的问题。

早期,研究人员尝试使用机器学习方法识别晶圆缺陷。例如,Yuan等^[7]提出了一种多步缺陷分析方法对晶圆上的缺陷模式进行聚类 and 识别。该方法首先利用K最近邻(k-nearest neighbor, KNN)算法去噪,将局部缺陷和全局缺陷分开,然后利用聚类技术根据缺陷空间位置将局部缺陷分成不同的簇,最后通过各种模型选择准则,确定每个局部缺陷簇的模式。Chang等^[8]采用人工神经网络(artificial neural network, ANN)研发了一个晶圆缺陷自动检测系统。Xie等^[9]提出了一种支持向量机(support vector machine, SVM)的晶圆缺陷模式检测的方案。然而,由于晶圆缺陷特征的维度高、噪声多,利用传统机器学习方法进行晶圆缺陷分类的准确率较低。

随着深度学习在图像分类任务中出色的表现,研究人员开始利用深度学习从空间域出发进行晶圆缺陷分类。卷积神经网络(convolutional neural network, CNN)作为最流行的深度学习方法之一,被广泛应用于晶圆缺陷识别。Lee等^[10]设计了一种名为FDC-CNN的模型,该模型与递归神经网络相结合,可以根据多元传感器信号提取缺陷特征的结构信息。Kyeong等^[11]提出了一种对晶圆图中的混合类型缺陷进行分类的方法,该方法为每个缺陷提出了单独的分类模型,最后将多个模型的结果组合起来确定最终的类别。Nakazawa等^[12]使用合成晶圆图训练了一个由3层CNN、2层全连接层和1个Softmax层组成的简单模型,该模型在测试集上的准确率达到了98.2%。Cheon等^[13]提出了一种结合CNN和KNN的晶圆表面缺陷检测模型,该模型可以在不需要额外训练提取特征的前提下,提高缺陷分类的性能。Wang等^[14]使用具有噪声的基于密度的聚类方法(density-based spatial

clustering of applications with noise, DBSCAN)过滤掉晶圆图中的随机缺陷,再使用CNN模型对过滤后的晶圆图提取特征。Saqlain等^[15]采用了一种CNN模型用于自动识别晶圆缺陷。Manivannan等^[16]提出一种基于CNN的深度半监督学习方法。该方法利用有缺陷标签数据和无缺陷标签数据同时进行训练,在使用相对较少的有缺陷标签数据进行训练的情况下,也能够超越当时其他方法的性能。Yu等^[17]通过基于PeeleNet模块^[18]的轻量级网络WM-PeeleNet对训练集中样本数量较少的晶圆缺陷进行分类。Chen等^[19]提出了一种基于多尺度特征融合的模型,该模型能够在不同尺度处理和聚合空间信息,从而提取更丰富的特征,提高晶圆缺陷识别的准确度和鲁棒性。Han等^[20]构造了一种基于图形(graph-based)的晶圆缺陷模式分类方法,利用图形骨架来表示晶圆图的拓扑信息,从而实现旋转不变和尺寸不变的特征提取。Chen等^[21]设计了一种基于改进的卷积块注意力机制(ICBAM)和深度卷积神经网络ResNeXt50的晶圆缺陷分类方法,提高了识别性能。Cha等^[22]探索了一种改进的U-Net模型,并使用注意力机制和残差块的结合来提取特征,减少无关信息。Xu等^[23]设计了一个基于ResNet的深度卷积神经网络方法,结合改进的CBAM注意力机制和余弦归一化。上述方法均为从空间域出发利用CNN进行晶圆缺陷分类,虽然能够达到较好的效果,但存在两方面的限制:首先,空间域方法的关注点仅限于图像的几何和像素信息,忽略了图像的高级语义信息和全局信息,这会导致空间域方法面对复杂场景或具有细微特征的晶圆图时的性能下降。其次,空间域方法对图像的旋转和变形相当敏感,导致模型在真实环境中表现不佳。

近年来,研究人员开始探索从频率域出发提取晶圆图特征再进行分类。Jizat等^[24]使用快速傅里叶变换(fast Fourier transform, FFT)和离散小波变换(discrete wavelet transform, DWT)将空间域转换到频率域从而实现图像数据的去噪,再使用InceptionV3和逻辑回归对处理后的图像进行特征提取。Fan等^[25]提出利用图像处理技术和傅里叶变换来识别半导体制造过程中的关键参数和晶圆缺陷的方法。该方法利用傅里叶变换消除图像中的水平纹理,再利用逆傅里叶变换将频率域图像还原为空间域图像,从而突出晶圆缺陷的异常纹理。Wei等^[26]为了增加CNN在频率域提取信息的能力,提出了一种基于小波变换的晶圆缺陷识别方法(MRWA-Net),该方法从小波域的不同频域分量提取缺陷的全局和细节信息。相较于空间域方法,频率域方法具有一些明显的优点:首先,频率域方法能够更好地捕捉晶圆图的高级语义信息,提取更为丰富的频率信息。其次,频率域方法对于尺度变化和旋转等几何变换具有较好的鲁棒性,有助于应对同类型不同形态和方向的缺陷。但是现有频率域方法仅

仅从频率域出发,没有深入研究频率域方法与空间域方法的结合。而且仅使用频率域方法容易忽略晶圆图像中的重要细节信息,使网络对缺陷的边缘和形状提取不足。

综上所述,现有基于深度学习的晶圆缺陷分类方法要么从空间域出发,要么从频率域出发,尚未同时从空间域和频率域出发进行网络设计,限制了分类性能。为了进一步提高晶圆缺陷分类性能,提出了基于空间域和频率域特征融合的双流晶圆缺陷分类网络,主要创新点如下:

1) 提出了一种基于空间域和频率域特征融合的双流晶圆缺陷分类网络 SFWD-Net。空间流用于捕捉晶圆图中缺陷的空间分布,频率流用于获取缺陷的频域维度信息。空间流和频率流两个维度的信息相互补充和融合,能够显著提升晶圆缺陷的分类性能。

2) 利用多尺度卷积网络组成了一种多尺度特征提取卷积模块(MSFE Conv Block)。该模块采用了一系列不同大小的卷积核,以涵盖不同尺度的感受野,从而使网络能够更全面地理解晶圆图缺陷信息的空间结构。

3) 提出了一种新的多视角注意力模块(multiview attention block, MVA Block),它能够增强模型的信息集合能力,能够从晶圆图的通道、高和宽 3 个角度提取缺陷特征。

4) 所提出的双流晶圆缺陷分类网络 SFWD-Net 优于

目前 5 种先进的晶圆缺陷分类方法。

1 晶圆缺陷分类网络 SFWD-Net

1.1 整体网络结构

提出的 SFWD-Net 整体网络结构如图 1 所示。SFWD-Net 主要由空频融合双流特征提取网络和分类网络组成。其中空频融合双流特征提取网络包含空间流和频率流两个分支。空间流分支主要利用多个多尺度特征提取卷积模块(MSFE Conv Block)和多视角注意力模块(MVA Block)提取晶圆图缺陷的空间特征与通道特征。该分支首先用 1×1 卷积将输入的晶圆图通道数转换为 48,再由多个 MSFE Conv 和 MVA 模块组合提取空间特征,最后输出为 128 通道的 7×7 大小的特征图像 F_s 。频率流分支则主要采用 DWT 提取晶圆图的频率特征。具体而言,频率分支首先通过卷积滤波器对输入的晶圆图进行离散小波变换,然后将变换后的图像分解为 LL、LH、HL 和 HH 4 幅子带图像并选取其中 3 个具有较为显著特征的子带图像进行融合,最后通过对融合后的图像进行残差操作得到频率域分支的输出 F_f 。空间流的输出 F_s 和频率流的输出 F_f 融合后得到空频融合双流特征提取网络的输出特征 F_c , F_c 输入分类网络进行分类,完成晶圆缺陷分类。

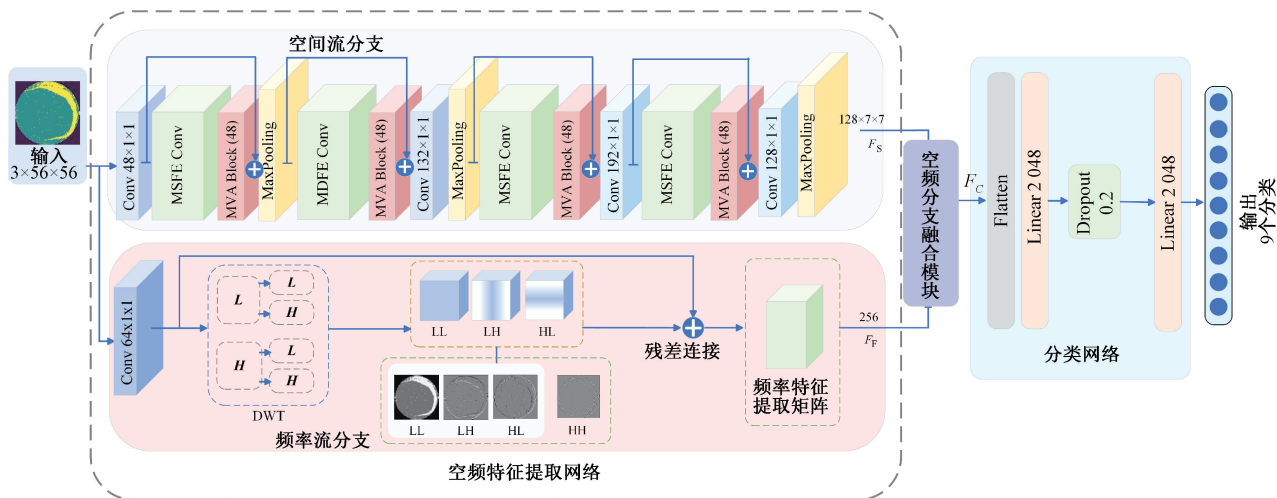


图 1 SFWD-Net 模型结构图

Fig. 1 Structure diagram of SFWD-Net model

1.2 空间流分支

空间流分支负责获取晶圆图中缺陷的空间信息,用于理解缺陷的空间分布特征。空间流分支首先利用 48 个 1×1 卷积核将原始 2D 晶圆图进行卷积将其转换为通道数为 48 的 3D 矩阵。然后主要通过多个 MSFE Conv

和 MVA Block 组合的形式进行特征提取,该组合有针对性地捕捉晶圆图中的全局关联信息,以提高对晶圆缺陷空间分布的感知能力。同时在每个组合使用了残差连接,以促进信息的有效传递和梯度的稳定传播,进而提高网络的训练效果和泛化性能。最终该分支提取的新特征矩阵传递到下一个阶段。

1) 多尺度卷积特征提取模块 (MSFE Conv Block)

传统的晶圆缺陷分类方法采用单个卷积核进行特征提取,但单个卷积核仅能捕捉固定窗口的缺陷特征。然而,晶圆缺陷类型通常存在不同大小、形状和位置等方面的差异。因此,为了捕捉复杂多样的特征,有必要引入多尺度卷积核。通过采用不同大小的卷积核,可以在感受野内提取不同尺度的特征,从而更好地满足缺陷分类任务的需求。这种多尺度卷积核的应用能够更全面地考虑晶圆缺陷的多样性,为分类模型提供更为丰富和准确的特征表示,使得模型能够更有效地适应晶圆缺陷的多样性和复杂性,提高分类性能。

提出的 MSFE Conv 模块的结构如图 2 所示。首先,将特征图 F 分为 6 组 ($F_1 \sim F_6$),每组对应一个不同大小的卷积核 ($1 \times 1, 3 \times 3, 5 \times 5, 7 \times 7, 9 \times 9$ 和 11×11)。输入特征图按通道分为 6 组,每组包含 $C/6$ 通道。6 组卷积核涵盖了不同大小的感受野能够获得比传统单核小视野 CNN 更为优越的特征,从而提高整体分类精度,提高模型对晶圆缺陷的信息提取。为了降低计算复杂度,该模块采用了将大尺寸卷积核替换为多个 3×3 卷积核的策略。具体而言,例如, 11×11 的原始卷积核被替换为 5 个 3×3 卷积核, 9×9 的原始卷积核被替换为 4 个 3×3 卷积核,以此类推。这种替代方案不仅保持了有效感受野,同时增加了网络的非线性表达能力,降低了参数计算量。

输入特征图可表示为 $F \in \mathbf{R}^{C \times H \times W}$,其中 C 和 H 和 W 分别表示输入特征图的通道数、高度和宽度。将 F 按通道分为 6 组,每组包含 $C/6$ 通道。用 $F_1 \sim F_6$ 表示每组输入, $F'_1 \sim F'_6$ 表示每组输出,多尺度卷积 F'_1, F'_2 分支可以表示为:

$$\begin{aligned} F'_1 &= \text{Conv}_{1 \times 1}(F_1) \\ F'_2 &= \text{Conv}_{3 \times 3}(F_2) \end{aligned} \quad (1)$$

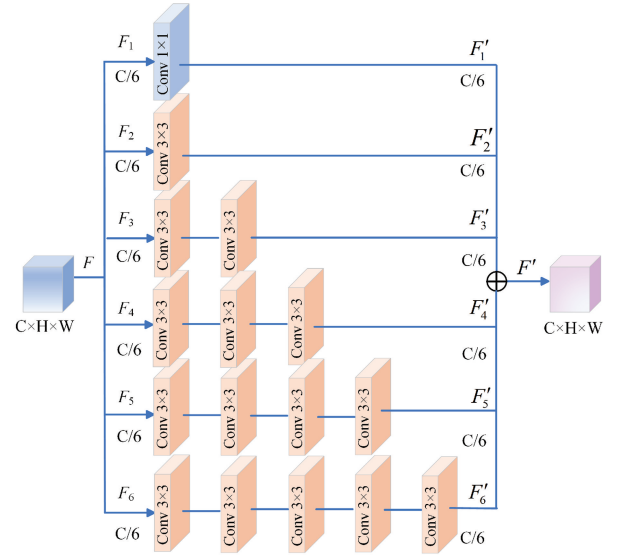


图 2 MSFE Conv 模块结构图

Fig. 2 Structure of the MSFE ConvBlock

其中, $\text{Conv}_{1 \times 1}$ 和 $\text{Conv}_{3 \times 3}$ 表示 1×1 和 3×3 的卷积运算。

F'_3, F'_4 分支过程则如式 (2) 和 (3) 所示。

$$\begin{aligned} F'_{31} &= \text{Conv}_{3 \times 3}(F_3) \\ F'_3 &= \text{Conv}_{3 \times 3}(F'_{31}) \end{aligned} \quad (2)$$

$$\begin{aligned} F'_{41} &= \text{Conv}_{3 \times 3}(F_4) \\ F'_{42} &= \text{Conv}_{3 \times 3}(F'_{41}) \\ F'_4 &= \text{Conv}_{3 \times 3}(F'_{42}) \end{aligned} \quad (3)$$

同理可得出 F'_5, F'_6 , 之后将获得的 $F'_1, F'_2, \dots, F'_6$ 拼接起来:

$$F' = \text{Concat}(F'_1, F'_2, F'_3, F'_4, F'_5, F'_6) \quad (4)$$

其中, Concat 表示拼接。

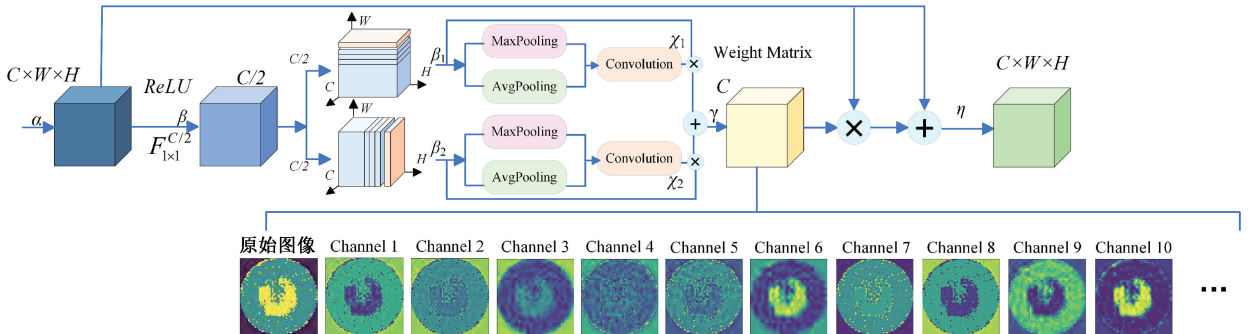


图 3 MVA Block 结构图,其中 $F_{1 \times 1}^{C/2}$ 表示 1×1 卷积核和输出为 $C/2$ 通道的卷积运算

Fig. 3 Structure of the MVA Block, where $F_{1 \times 1}^{C/2}$ denotes a 1×1 convolutional kernel and the output is a convolutional operation for $C/2$ channels

2) 多视角注意力模块 (MVA Block)

提出的 MVA Block 利用多视角注意力机制,能同时

关注空间特征和通道特征,可以增强网络的信息集成能力,其结构如图 3 所示。

MVA Block 模块的第一阶段负责捕捉通道方向特征,首先,将原始特征矩阵 $R^{C \times W \times H}$ 使用 1×1 的卷积核将通道 C 压缩为原来的一半,以便可以集成与通道相关的功能。其过程可以描述为:

$$\beta = \text{ReLU}(\text{BN}(\text{Conv}_{1 \times 1}(\alpha))) \tag{5}$$

式中: α 是输入, β 是输出, $\text{Conv}_{1 \times 1}$ 是内核为 1×1 的卷积运算, BN 和 ReLU 分别为批量归一化和 ReLU 激活函数。

MVA Block 模块的第 2 阶段将分为两个分支,分别负责获取通道维度 C 与空间维度 W 和 H 的关系。在处理完每个分支的信息之后,细化的张量会继续整合为原始数据的尺寸 $R^{C \times W \times H}$ 。图 3 中的权重矩阵 (weight matrix) 可展示晶圆图每个通道的热图,反映其注意力分布。具体而言,对于每个分支,首先在各自的维度方向利用最大池化和平均池化操作将张量压缩为特征图,以更全面地捕捉特征矩阵中的信息。然后将生成的特征图拼接在一起,形成一个更具丰富信息的联合特征图。接着该联合特征图经过允许动态调整其参数的卷积层处理以有选择地增强有价值的信息。最后使用 Sigmoid 函数将最终特征值投影到 $[0, 1]$ 范围内。以上计算过程可以用式 (6) 描述:

$$\chi_i = \text{Sigmoid}(\text{Conv}(\text{Concat}(\text{MaxPooling}(b_i), \text{AvgPooling}(b_i)))) \tag{6}$$

式中: $\beta_i (i=1, 2)$ 、下标 i 、 MaxPooling 、 AvgPooling 、 Concat 、 Conv 、 Sigmoid 分别表示输入张量、分支索引、最大池化操作、平均池化操作、拼接操作、卷积操作、Sigmoid 函数。

MVA Block 模块的第 3 阶段引入了残差连接进行重新校准,以减轻 SFWD-Net 训练时的梯度消失问题,进一步提升网络性能。该过程可以表示为:

$$\gamma = \text{Concat}(\beta_1 \otimes \chi_1 + \beta_2 \otimes \chi_2) \tag{7}$$

其中, $\otimes$ 代表逐元素相乘。

在两个分支拼接之后,引入两个残差连接,得到 MVA Block 模块最终的输出为:

$$\eta = \alpha + \alpha \otimes \gamma \tag{8}$$

3) 空间流分支网络模型及参数

表 1 为空间流分支的网络模型及参数,概述了该分支所使用的层和模块的顺序。

其中,Conv2d、MSFE Conv、MVA Block、MaxPool2d 分别表示二维卷积、多尺度特征提取卷积模块、多视角注意力模块和最大池化层。

1.3 频率流分支

频率流分支负责获取晶圆图中的频率信息,用于识别晶圆图中缺陷的细微纹理和特定频率模式。该分支首先使用 64 个 1×1 卷积核对原始晶圆图像进行处理,然后利用 DWT 进一步对处理后的图像进行频率特征的提取,接着采用残差连接保留原始信息并促进不同层次特征的有效传递,最后将生成的频率流新特征矩阵 F_f 送入下一

阶段进行后续处理。

表 1 空间流分支网络模型及参数

Table 1 Spatial stream branch network model and parameters

层	输入大小	输出大小
Conv2d	56×56×3	56×56×48
MSFE Conv	56×56×48	56×56×48
MVA Block	56×56×48	56×56×48
MaxPool2d	56×56×48	27×27×48
Conv2d	27×27×48	27×27×132
MSFE Conv	27×27×132	27×27×132
MVA Block	27×27×132	27×27×132
MaxPool2d	27×27×132	13×13×132
Conv2d	13×13×132	13×13×192
MSFE Conv	13×13×192	13×13×192
MVA Block	13×13×192	13×13×192
Conv2d	13×13×192	13×13×128
MaxPool2d	13×13×128	7×7×128

DWT 是一种成熟且应用广泛的信号处理工具,它可以将输入信号投射在小波域中,从而更好地揭示不同的特征。在实际应用中,DWT 主要采用双通道子带编码的方式进行信号处理。图 4 展示了对晶圆缺陷图像进行 DWT 分解的实验结果。

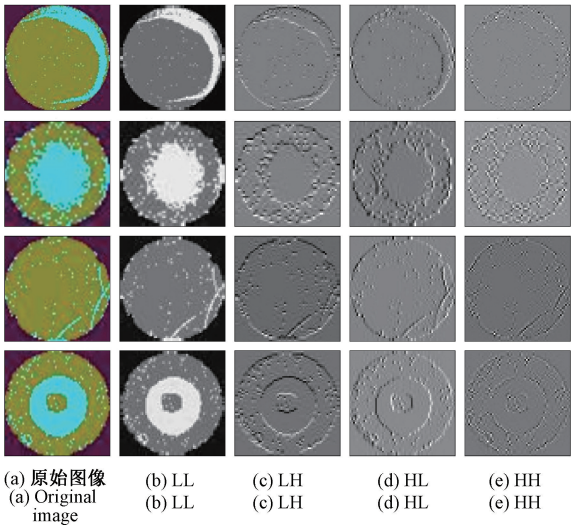


图 4 对晶圆缺陷图像进行 DWT 分解
Fig. 4 DWT decomposition of wafer defect images

如图 4 所示,通过 DWT 原始晶圆缺陷图像被分解为 4 个子频段:LL、LH、HL 和 HH。其中,LL 代表图像的低频部分,具有原始图像一半的分辨率,映射了图像的整体信息并消除了冗余。LH、HL 和 HH 分别代表图像的水平、垂直和对角线方向的高频部分,反映出图像的细节信息。

实验结果表明,由于原始晶圆缺陷图像的对角线信

息相对不够明显,DWT 分解后的 HH 子频段的信息几乎可以忽略。综合考虑频率流分支的特征提取性能和模型复杂性,该模型选择保留 LL、LH、HL 3 个子频段,并将它们连接起来形成频率域特征矩阵 F_F 。

1.4 空频分支融合模块

空频分支融合模块对空间流输出特征 F_S 和频率流输出特征 F_F 使用类得分融合(class score fusion)^[27]方法进行融合,如图 5 所示。具体来说,首先将 F_S 和 F_F 转换为分类分数 S_S 和 S_F ,然后整合两个分支的贡献,最后两者相加获得融合分类分数 S 为:

$$S = \frac{S_S + S_F}{2}$$

(9)

此时的 S 包含全部的信息,将其输入到分类网络进行对晶圆缺陷的分类。

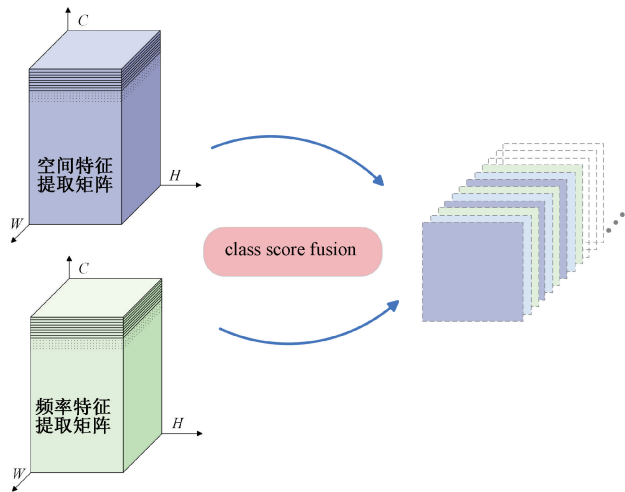


图 5 空频分支融合模块结构图

Fig. 5 Schematic diagram of the fusion module

1.5 空频融合双流特征提取网络优势

由于空间流和频率流两个分支分别捕捉了空间域和频率域的特征信息,通过融合两个分支的输出,网络能够更全面地理解输入数据,并综合考虑空间域和频率域的关键信息,这种互补性质使得 SFWD-Net 能够更好的对晶圆图中的缺陷进行提取,为后续缺陷分类提供更好的特征基础。

1.6 分类网络

分类网络是 SFWD-Net 的决策和输出阶段,其网络模型和参数如表 2 所示。

其中,Flatten 层负责将来自空频融合双流特征提取网络的特征矩阵转化为一维向量,旨在为后续 Linear 层的全连接操作做准备。Dropout(0.2)层用来增强 SFWD-Net 的泛化能力和防止过拟合,它以 20% 的概率随机丢弃网络中的部分神经元,从而在训练过程中防止网络对

特定的输入模式过度依赖,以提高网络对不同输入的适应型。Linear 层将来自 Flatten 层的一维特征向量映射到最终的输出空间,使得网络能够学习并捕捉输入特征之间的非线性关系。

表 2 分类网络模型及参数

Table 2 Classification network model and parameters

Layers	Inputsize	Output size
Flatten	7×7×(128+256)	18 816×1×1
Linear	18 816×1×1	2 048×1×1
Dropout(0.2)	2 048×1×1	2 048×1×1
Linear	2 048×1×1	2 048×1×1
Linear	2 048×1×1	9×1×1

2 实验结果与分析

2.1 数据集

本实验采用 WM-811K 数据集开展,WM-811K 数据集是一个大规模的半导体晶圆图数据集,由 Wu 等^[28]于 2015 年公开。这是目前已知最大规模的公开晶圆图数据集,包含 811 457 张真实世界的晶圆图,来自实际晶圆制造过程中的 46 293 个生产批次。该数据集不仅包含晶圆图和对应的缺陷标签,还提供了额外信息,如数据索引、批次名称、晶圆尺寸、缺陷类型、训练标签、测试标签等。其中,带有标签的晶圆图共有 172 950 张,约占总数据集的 21.3%。如图 6 所示,该数据集共有 9 种标签,其中标签“None”表示晶圆图无缺陷,占有标签数据集的 85.2%。其他标签表示晶圆图有缺陷,具体而言,“Donut”、“Center”、“Scratch”、“Loc”、“Edge-Ring”、“Edge-Loc”、“Near-Full”和“Random”分别表示“环形”、“中心”、“划痕”、“局部”、“边缘环”、“边缘局部”、“近满”和“随机”缺陷。图 7 为 WM-811K 数据集缺陷种类分布直方图。可以看出,该数据集缺陷类别极度不平衡,如“Edge-Ring”类型的数据量远远高于其他类型的数据量,而“Donut”、“Random”、“Scratch”、“Near-Full”的数据量相对较少。

2.2 数据预处理

在分类任务中,数据集类别不平衡会削弱网络的分类性能。从 WM-811K 数据集中选取 172 950 张带有标签的晶圆图并采用自编码器(auto-encoder)网络进行数据增强并使每类型的缺陷图像保持在 2 000 张左右,最终共得到 19 261 张图像。其中,80%用于训练,20%用于测试。统一调整为(56×56)的尺寸以供模型使用。为适应模型处理,将单通道图像扩展为三通道 RGB 图像。图 8 显示了对 7 种缺陷类型图像进行数据增强的示例结果。其中第 1 行表示原始缺陷图像,第 2 行表示使用数

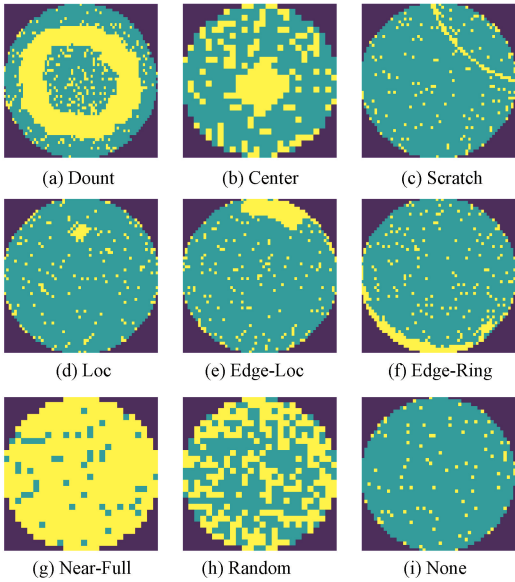


图 6 WM-811K 数据集缺陷种类

Fig. 6 Types of defects in WM-811K dataset

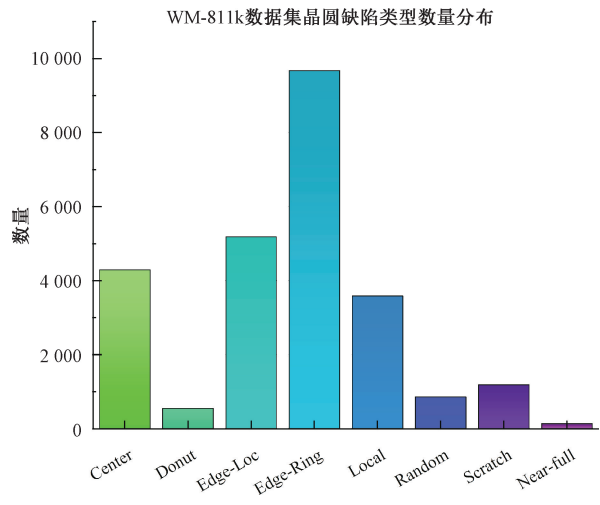


图 7 晶圆缺陷种类分布直方图

Fig. 7 Histogram of the distribution of wafer defect types

据增强后生成的新缺陷图像。

2.3 实验环境

实验的环境配置如表 3 所示,操作系统采用 Windows 11,CPU 为 Intel Core i7-12700H 处理器,内存容量为 16 GB,GPU 为 NVIDIA GeForce RTX3060 (Laptop) 6GB,深度学习框架为 Pytorch-1.10.1,解释器版本为 Python 3.8.16。在所提出的训练模型中,学习率设定为 0.0005,训练周期(epoch)设置为 50,采用 Adam 优化器。

2.4 评价指标

为了评价模型性能,采用图像分类任务中常见的四种评价指标:准确率(Accuracy),精确率(Precision),召

表 3 实验环境参数

Table 3 Parameters of experimental environment

标题	参数
操作系统	Windows 11
CPU	IntelCore i7-12700H
GPU	NVIDIA GeForce RTX3060 (Laptop) 6 GB
Python 版本	3.8.16
深度学习框架	Pytorch-1.10.1

回率(Recall)和 F1 分数(F1-Score)进行实验。各评价指标的公式如式(10)~(13)所示。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$
 (10)

$$Precision = \frac{TP}{TP + FP}$$
 (11)

$$Recall = \frac{TP}{TP + FN}$$
 (12)

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$
 (13)

表 4 真实与预测标签混淆矩阵示意图

Table 4 Schematic representation of the true vs. predicted label confusion matrix

		预测	
		正样本	负样本
真实	正样本	TP	TN
	负样本	FP	FN

其中 TP(true positive samples)为模型正确地将正样本预测为正样本的数量;TN(true negative samples)为模型正确地将负样本预测为负样本的数量;FP(false positive samples)为模型错误地将负样本预测为正样本的数量;FN(false negative samples)为模型错误地将正样本预测为负样本的数量。这 4 种分类结果示意如表 4 所示。准确度为模型对所有预测样本准确预测的比例;精确度表示模型预测正样本中,真正为正样本的比例;召回率衡量的是模型在所有实际正样本中,成功预测正样本的比例;F1 分数是精确度和召回率的调和平均,可以综合地衡量模型整体的测试结果。

2.5 实验结果和分析

图 9 展示了提出模型在训练集和测试集的准确度曲线。可以看出,经过 10 个周期后,模型的训练准确度和测试准确度即可收敛到较高且稳定的值。图 10 为提出模型在训练集和测试集的损失曲线,由图 10 可知,即使初始损失较大,提出模型的测试损失曲线也能够 5 个周期内收敛到一个较低的值并接近训练损失曲线。

图 11 显示了提出模型测试分类结果的混淆矩阵。可以看出,该模型对各类缺陷的分类都比较准确。

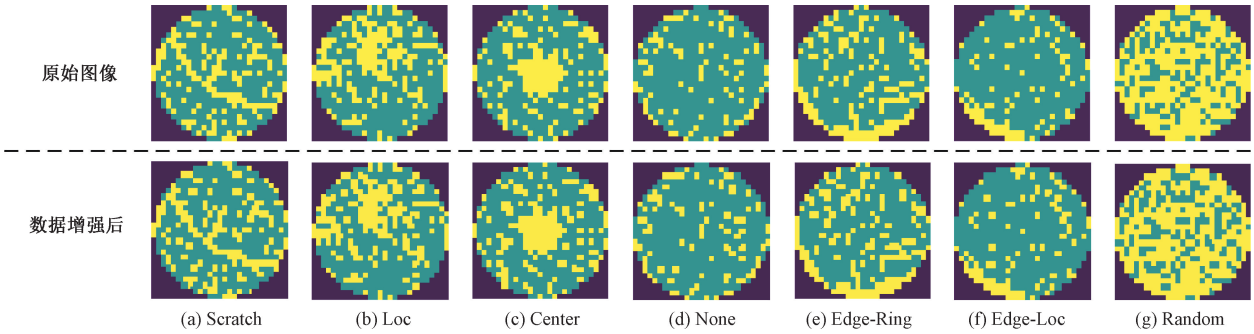


图 8 数据增强结果图,其中第一行为原始数据,第二行为对应的数据增强合成图

Fig. 8 Data enhancement result graph, wherein the first row is the original data and the second row is the corresponding data enhancement synthesis graph.

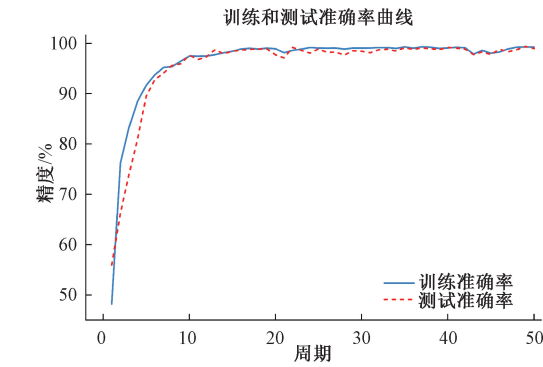


图 9 训练和测试的准确率曲线

Fig. 9 Accuracy curve for training and testing

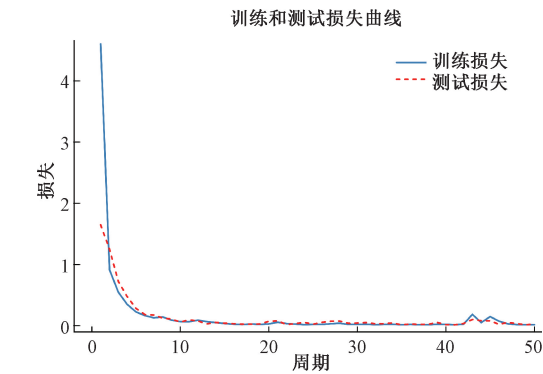


图 10 训练和测试的损失曲线

Fig. 10 Loss curves for training and testing

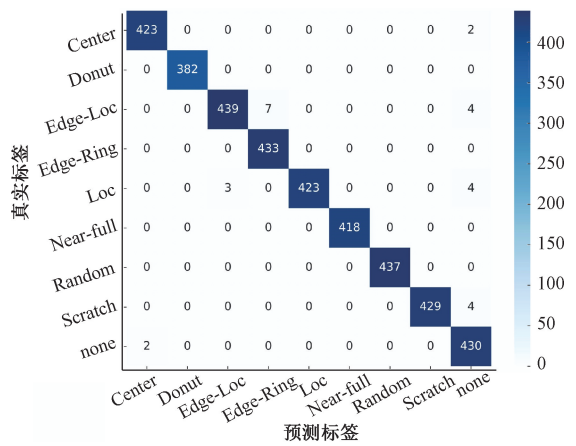


图 11 测试集混淆矩阵

Fig. 11 Confusion Matrix for testing

和 99.451 7%,超过第 2 名 0.181 6%、0.462 0%、0.332 5%和 0.313 4%。

表 5 不同方法检测结果对比

方法	Accuracy	Precision	Recall	F1-Score
K. P. et al. [29] (2023)	98.572 5	98.950 5	98.944 2	98.933 2
CNN-WDI [15] (2020)	97.806 1	97.216 4	97.194 8	97.193 1
CWDR-Net [30] (2023)	98.624 4	98.997 3	98.985 4	98.968 9
DCNN [31] (2021)	98.163 9	98.823 5	97.109 8	97.959 1
MRWA-Net [26] (2023)	99.117 6	99.174 0	99.129 4	99.138 3
SFWD-Net (Ours)	99.299 2	99.636 0	99.461 9	99.451 7

2.6 对比实验

1)不同方法检测结果对比

将所提出的 SFWD-Net 模型与晶圆缺陷识别方向的 5 种代表性先进方法进行了对比,实验结果如表 5 所示。

从表 5 可以看出,提出的 SFWD-Net 在所有 4 个指标上都超过了 5 种对比方法,分别在准确度,精确度、召回率和 F1 分数达到了 99.299 2%、99.636 0%、99.461 9%

2)不同缺陷类型检测结果对比

表 6 展示了 SFWD-Net 针对不同缺陷类型的检测结果。结果表明,SFWD-Net 对各类缺陷类型都具有优秀的检测性能,对于 Donut 和 Scratch 这两种缺陷类型,准确率、精确度、召回率和 F1 分数均达到了 100%,即使对于相对较为复杂的 Edge-Loc 缺陷类型,模型也展现出令人满意的性能。

表 6 不同缺陷类型检测结果对比

Table 6 Comparison of detection results of different defect types (%)

缺陷类型	Accuracy	Precision	Recall	F1-Score
Center	98.997 5	99.540 0	99.458 0	99.488 9
Donut	100.000 0	100.000 0	100.000 0	100.000 0
Edge-Loc	97.830 8	99.078 0	99.299 0	99.169 7
Edge-Ring	100.000 0	99.888 7	99.782 4	99.833 7
Loc	99.544 4	99.599 1	99.508 5	99.543 3
Near-Full	100.000 0	100.000 0	99.946 2	99.972 7
Random	100.000 0	99.861 1	99.953 7	99.904 1
Scratch	100.000 0	100.000 0	100.000 0	100.000 0
None	98.398 2	99.112 7	99.083 2	99.075 6

3)不同方法的模型复杂度对比

表 7 展示了不同方法的模型复杂度对比结果。选取了 2023 年以来的 3 种新方法进行对比。可以看出,提出方法在 FLOPs (G) 指标上具有显著的优势。SFWD-Net 的 FLOPs 为 0.393 G,远低于其他对比方法 K. P. et al. (4.023 G)、CWDR-Net (0.715 G) 和 MRWA-Net (1.880 G)。这表明 SFWD-Net 在计算效率和处理速度方面具有优势,适合应用于实时性要求高或计算资源有限的场景。此外,低 FLOPs 意味着 SFWD-Net 在实际运行过程中对硬件资源的需求较少,能够显著节约计算成本。

表 7 不同模型复杂度对比

Table 7 Comparison of complexity of different models

复杂度指标	K. P. et al. [29] (2023)	CWDR-Net [30] (2023)	MRWA-Net [26] (2023)	SFWD-Net (Ours)
FLOPs (G)	4.023	0.715	1.880	0.393
Params (M)	25.56	26.95	17.91	45.91

然而,SFWD-Net 的参数量为 45.91 M,相较于其他网络 K. P. et al. (25.56 M)、CWDR-Net(26.95 M)和和 MRWA-Net(17.91 M)较高。这意味着 SFWD-Net 在训练和推理阶段需要更多的存储空间,可能会增加模型的训练时间和部署难度。

2.7 可视化分析

为了更加深入地分析 MVA Block 模块的功能,对 5 种缺陷类型生成了 10 个通道的注意力权重热图。

如图 12 所示,MVA Block 能够根据晶圆图中的重要性为每个位置自动分配不同的注意力权重,从而有选择地强调相关特征并抑制无用的特征。例如图 12(a)所示,对于“Scratch”缺陷类型的权重矩阵的所有通道,MVA Block 对于晶圆图中的缺陷部分给予了明确的关注,并勾勒出了缺陷模式与背景之间的清晰边界。通过以上 5 种缺陷的例子,可以看到注意力权重矩阵中显示的所有模式几乎与原始模式相同。这表明 MVA Block 能够准确地突显缺陷模式并有助于网络准确区分缺陷类型。

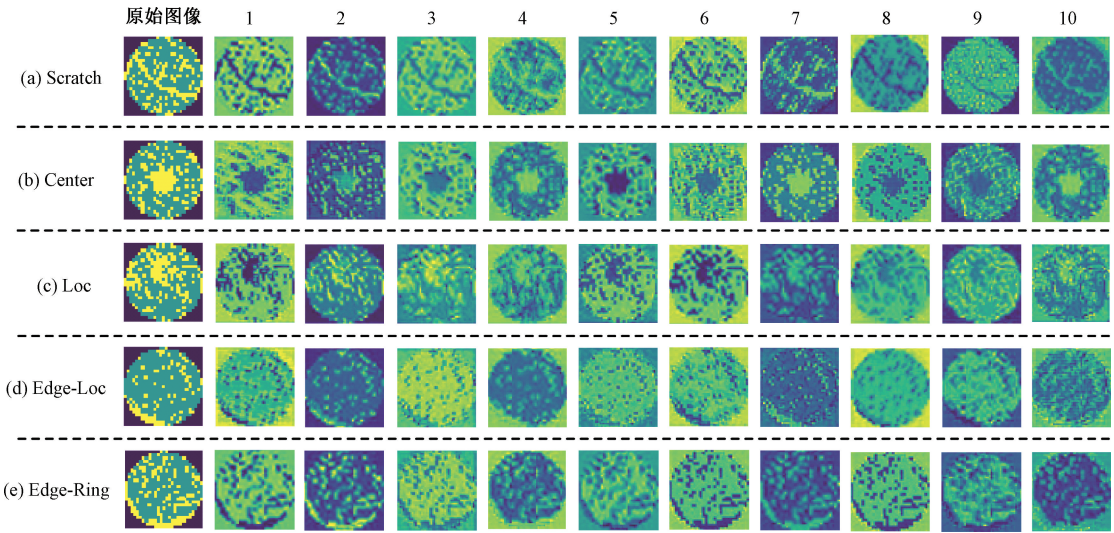


图 12 MVA Block 权重矩阵的通道可视化

Fig. 12 Channel visualization of the MVA Block weight matrix

2.8 消融实验

为了证明所提出的双流特征提取网络结构、MSFE Conv 模块和 MVA Block 模块的有效性,开展了消融实验。在实验中,将空间流分支的主干网络设定为基线

(Baseline),结果如表 8 所示。

实验结果显示,在将 MSFE Conv 模块引入基线模型后,准确度、精确度、召回率和 F1 分数均分别提升了 1.842 7%,1.916 5%,1.861 8%和 1.949 6%。加入 MVA

模块至基线模型中的效果与引入 MSFE Conv 模块相比,性能指标提升几乎相当。同时引入这两个模块后,评价指标显著上升,充分验证了 MSFE Conv 模块和 MVA 模块对空间流分支的性能具有良好的改进效果。在空间流基础上引入频率流后,获得最佳性能,证明了空频融合双流特征提取网络的有效性。

表 8 消融实验结果

Table 8 Results of ablation experiments (%)

Baseline	Spatial Stream		Frequency Stream	Accuracy	Precision	Recall	F1-Score
	MSFE Conv	MVA Block					
✓				96.340 5	96.738 6	96.742 2	96.658 7
✓	✓			98.183 2	98.655 1	98.604 0	98.608 3
✓		✓		98.078 4	98.477 8	98.426 4	98.423 6
✓	✓	✓		98.728 3	99.108 5	99.119 9	99.096 9
			✓	98.468 7	98.782 9	98.504 0	98.763 2
✓	✓	✓	✓	99.299 2	99.636 0	99.461 9	99.451 7

3 结 论

针对晶圆缺陷分类问题,通过创新性地设计双流特征提取网络结构、集成多尺度特征提取卷积模块与多视角注意力模块,构建了基于空间域与频率域特征融合的双晶圆缺陷分类网络 SFWD-Net,实现了对晶圆图像中缺陷特征的全面而精细的捕捉与融合,有效提升了晶圆缺陷识别的准确度。大量实验结果证明,SFWD-Net 的性能优于 5 种代表性先进晶圆缺陷分类方法。在计算复杂度方面,SFWD-Net 在计算效率和速度上具有显著优势,但在参数量方面仍有优化空间。未来的工作将在保持计算效率的同时,通过模型压缩和优化技术,进一步减少参数量,提高模型在各种应用场景中的适用性。另外将研究能够适用于不同数据集和制造环境的晶圆缺陷分类网络,以应对更为复杂的实际使用场景和新型缺陷模式。

参考文献

[1] HUANG A C, MENG S 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] 吴一全,赵朗月,苑玉彬,等. 基于机器视觉的 PCB 缺陷检测算法研究现状及展望 [J]. 仪器仪表学报, 2022,43(8): 1-17.

WU Y Q, ZHAO L Y, YUAN Y B, et al. Research status and the prospect of PCB defect detection algorithm based on machine vision [J]. Chinese Journal of Scientific Instrument, 2022, 43(8): 1-17.

[3] 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 based on deep learning [J]. Journal of Electronic Measurement and Instrumentation,

2022,36(11): 79-90.

[4] WANG J, XU C, YANG Z, et al. Deformable convolutional networks for efficient mixed-type wafer defect pattern recognition [J]. IEEE Transactions on Semiconductor Manufacturing, 2020, 33(4): 587-596.

[5] FAN M, WANG Q, VAN D W B. Wafer defect patterns recognition based on OPTICS and multi-label classification [C]. 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). IEEE, 2016: 912-915.

[6] 李阳,蒋三新. 基于改进生成对抗网络的无监督晶圆缺陷检测 [J]. 电子测量技术, 2023,46(6): 91-99.

LI Y, JIANG S X. Unsupervised wafer defect detection based on improved generative adversarial network [J]. Electric Measurement Technology, 2024, 46(6): 91-99.

[7] YUAN T, KUO W, BAE S J. Detection of spatial defect patterns generated in semiconductor fabrication processes [J]. IEEE Transactions on Semiconductor Manufacturing, 2011, 24(3): 392-403.

[8] CHANG C Y, LI C H, CHANG Y C, et al. Wafer defect inspection by neural analysis of region features [J]. Journal of Intelligent Manufacturing, 2011, 22: 953-964.

[9] XIE L, HUANG R, GU N, et al. A novel defect detection and identification method in optical inspection [J]. Neural Computing and Applications, 2014, 24: 1953-1962.

[10] LEE K B, CHEON S, KIM C O. A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes [J]. IEEE Transactions on Semiconductor Manufacturing, 2017, 30(2): 135-142.

[11] KYEONG K, KIM H. Classification of mixed-type defect patterns in wafer bin maps using convolutional neural networks [J]. IEEE Transactions on Semiconductor Manufacturing, 2018, 31(3): 395-402.

- [12] NAKAZAWA T, KULKARNI 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.
- [13] CHEON S, LEE H, KIM C O, et al. Convolutional neural network for wafer surface defect classification and the detection of unknown defect class [J]. IEEE Transactions on Semiconductor Manufacturing, 2019, 32(2): 163-170.
- [14] WANG R, CHEN N. Defect pattern recognition on wafers using convolutional neural networks [J]. Quality and Reliability Engineering International, 2020, 36(4): 1245-1257.
- [15] SAQLAIN M, ABBAS Q, LEE J Y. A deep convolutional neural network for wafer defect identification on an imbalanced dataset in semiconductor manufacturing processes [J]. IEEE Transactions on Semiconductor Manufacturing, 2020, 33(3): 436-444.
- [16] MANIVANNAN S. An ensemble-based deep semi-supervised learning for the classification of wafer bin maps defect patterns [J]. Computers & Industrial Engineering, 2022, 172: 108614.
- [17] YU N, CHEN H, XU Q, et al. Wafer map defect patterns classification based on a lightweight network and data augmentation [J]. CAAI Transactions on Intelligence Technology, 2023, 8(3): 1029-1042.
- [18] WANG R J, LI X, LING C X. Pelee: A real-time object detection system on mobile devices [J]. Advances in Neural Information Processing Systems, 2018, 31, DOI: 10.48550/arXiv.1804.06882.
- [19] CHEN Y, XU Z, JI J. Wafer defect recognition method based on multi-scale feature fusion [J]. Frontiers in Neuroscience, 2023, 17: 1202985.
- [20] HAN X, JIA X, JI D Y, et al. Designing robust topological features for wafer map pattern classification[C]. 2023 34th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC). IEEE, 2023: 1-6.
- [21] CHEN S, LIU M, HOU X, et al. Wafer map defect pattern detection method based on improved attention mechanism [J]. Expert Systems with Applications, 2023, 230: 120544.
- [22] CHA J, JEONG J. Improved U-Net with residual attention block for mixed-defect wafer maps[J]. Applied Sciences, 2022, 12(4): 2209.
- [23] XU Q, YU N, ESSAF F. Improved wafer map inspection using attention mechanism and cosine normalization[J]. Machines, 2022, 10(2): 146.
- [24] JIZAT J A M, NASIR A F A, MAJEED A P P A, et al. Effect of image compression using fast fourier transformation and discrete wavelet transformation on transfer learning wafer defect image classification [J]. Mekatronika, 2020, 2(1): 16-22.
- [25] FAN S K S, TSAI D M, HE F, et al. Key parameter identification and defective wafer detection of semiconductor manufacturing processes using image processing techniques [J]. IEEE Transactions on Semiconductor Manufacturing, 2019, 32(4): 544-552.
- [26] WEI Y, WANG H. Wavelet integrated attention network with multi-resolution frequency learning for mixed-type wafer defect recognition[J]. Engineering Applications of Artificial Intelligence, 2023, 121: 105975.
- [27] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]. Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1, 2014: 568-576.
- [28] 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, 2014, 28(1): 1-12.
- [29] REMYA K P, SAJITH V. Machine learning approach for mixed type wafer defect pattern recognition by ResNet architecture [C]. 2023 International Conference on Control, Communication and Computing (ICCC). IEEE, 2023: 1-6.
- [30] LUO W, WANG H. Composite wafer defect recognition framework based on multiview dynamic feature enhancement with class-specific classifier [J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-12.
- [31] CHEN S, ZHANG Y, YI M, et al. AI classification of wafer map defect patterns by using dual-channel convolutional neural network [J]. Engineering Failure Analysis, 2021, 130: 105756.

作者简介



陈晓雷(通信作者),2003年于兰州大学获得学士学位,2006年于兰州大学获得硕士学位,2014年于兰州理工大学获得博士学位,现为兰州理工大学副教授,主要研究方向为人工智能、计算机视觉和虚拟现实。

E-mail: chenxl703@lut.edu.cn

Chen Xiaolei (Corresponding author) received his B. Sc. degree from Lanzhou University in 2003, M. Sc. Degree from Lanzhou University in 2006, and Ph. D. degree from Lanzhou University of Technology in 2014, respectively. Now he is an

associate professor at Lanzhou University of Technology. His main research interests include artificial intelligence, computer vision, and virtual reality.



温润玉, 2016 年于延安大学获得学士学位, 现为兰州理工大学硕士研究生, 主要研究方向为计算机视觉和图像处理。

E-mail: wenrunyu@hotmail.com

Wen Runyu received his B. Sc. degree from Yanan University in 2016. Now he is a M. Sc. candidate in Lanzhou University of Technology. His main research interests include computer vision and image processing.



杨富龙, 2006 年于兰州大学获得学士学位, 2009 年于兰州大学获得硕士学位, 2019 年于兰州大学获得博士学位, 现为兰州理工大学副教授, 主要研究方向为超材料、智能信息处理和电磁传感。

E-mail: yangfulong1982@126.com

Yang Fulong received his B. Sc. degree from Lanzhou University in 2006, M. Sc. Degree from Lanzhou University in 2009, and Ph. D. degree from Lanzhou University in 2019,

respectively. Now he is an associate professor at Lanzhou University of Technology. His main research interests include metamaterials, intelligent information processing, and electromagnetic sensing.



李正成, 2019 年于陇东学院获得学士学位, 现为兰州理工大学硕士研究生, 主要研究方向为计算机视觉。

E-mail: 1749492931@qq.com

Li Zhengcheng received his B. Sc. degree from Longdong University in 2019. Now he is a M. Sc. candidate in Lanzhou University of Technology. His main research interest includes computer vision.



沈星阳, 2023 年于兰州理工大学获得学士学位, 现为兰州理工大学硕士研究生, 主要研究方向为计算机视觉。

E-mail: 1071579761@qq.com

Shen Xingyang received his B. Sc. degree from Lanzhou University of Technology in 2023. Now he is a M. Sc. candidate in Lanzhou University of Technology. His main research interest includes computer vision.