DOI: 10.13382/j.jemi.B2508281

融合改进多头注意力与残差结构的 VGGNet 晶圆缺陷检测*

杜先君[1,2] 贾 龙[1]

(1. 兰州理工大学微电子现代产业学院 兰州 730050;2. 兰州理工大学自动化与电气工程学院 兰州 730050)

摘 要:精准检测晶圆图像中的缺陷对于及时识别晶圆生产过程中的异常故障具有重要意义。在晶圆测试阶段,由于深度学习方法具备卓越的特征提取能力,其在晶圆缺陷检测中得到广泛应用。然而,传统深度学习模型通常依赖于大量标注充分且高质量的数据进行训练,而在实际应用中,均衡、充足的标注数据往往难以获得。针对这一问题,提出了一种融合改进多头注意力机制与残差结构的 VGGNet 深度学习模型,旨在从不平衡的数据集中提取更全面的特征,从而实现对晶圆表面缺陷的精准检测。具体而言,利用改进的多头注意力机制将输入的晶圆图像特征映射到多维子空间,显著提升了模型的表达能力和泛化性能;同时,在传统 VGGNet 的全连接层中引入残差连接(residual structure, RS),有效缓解了深层网络训练中的梯度消失问题。为验证融合改进多头注意力机制与残差结构的 VGGNet 的有效性,在数据集 WM811K 上进行大量实验,其分类准确率达到 94.3%,相较传统 VGGNet 准确率提高了 3%,相较现有类似模型准确率平均提高了 1%。实验结果表明,在真实数据集 WM811K 上,所提方法不仅提高了晶圆缺陷检测的鲁棒性,而且在非平衡数据集上的检测性能明显优于现有算法。

关键词:晶圆图像分类;卷积神经网络;不平衡数据集;VGGNet;注意力机制

中图分类号:TN305;TP18 **文献标识码**:A **国家标准学科分类代码**:510.4050

VGGNet wafer defect detection with improved multi-head attention and residual structure

Du Xianjun[1,2] Jia Long[1]

(1. School of Microelectronics Industry-education Integration, Lanzhou University of Technology, Lanzhou 730050, China;
2. School of Automation and Electrical Engineering, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract: Accurate detection of defects in wafer images is of great significance for timely identification of abnormal faults in wafer production. In the wafer testing phase, the deep learning method has been widely used in wafer defect detection due to its excellent feature extraction capability. However, traditional deep learning models often rely on a large number of adequately labeled and high-quality data for training, and in practical applications, balanced and sufficient labeled data is often difficult to obtain. To address this issue, we propose a VGGNet deep learning model that integrates an improved multi-head attention mechanism with a residual structure, aiming to extract more comprehensive features from imbalanced data sets to achieve accurate detection of wafer surface defects. Specifically, we use an improved multi-head attention mechanism to map the input wafer image features to multi-dimensional subspaces, which significantly improves the expressiveness and generalization performance of the model. At the same time, the residual connection is introduced into the full connection layer of traditional VGGNet, which effectively alleviates the problem of gradient disappearance in deep network training. To validate the effectiveness of the VGGNet with the improved multi-head attention mechanism and residual structure (RS), extensive experiments were conducted on the WM811K dataset, achieving a classification accuracy of 94.3%, which is 3% higher than the traditional VGGNet and 1% higher than existing similar models on average. The experimental results show that on the real data set WM811K, the proposed method not only improves the robustness of wafer defect detection, but also significantly outperforms the existing algorithms on the non-equilibrium data set.

Keywords: wafer image classification; convolutional neural network; imbalanced dataset; VGGNet; attention mechanism

收稿日期:2025-04-03 Received Date:2025-04-03

*基金项目:国家自然科学基金(62241307)、兰州市科技支撑计划(2024-3-47)项目资助

0 引 言

晶圆缺陷识别是半导体制造领域的重要研究方向之一[1]。现代半导体生产过程中,晶圆的制造需经过硅锭生长、切割、研磨、抛光、光刻、化学蚀刻等一系列复杂工艺。在此过程中,不可避免地会在晶圆表面产生不同程度的损伤,因此,晶圆表面缺陷检测在生产工艺中至关重要。

根据缺陷成因,晶圆缺陷可分为随机缺陷和系统缺陷[2]。其中,随机缺陷主要由晶圆表面附着的颗粒造成,其分布具有随机性;系统缺陷则主要源于光刻掩模及曝光工艺中的系统误差,通常出现在亚分辨率结构特征区域,并在同一晶圆上不同芯片区域(Die)的对应位置重复出现。为确保晶圆具备较高的成品率,缺陷检测环节格外关键[3]。光学检测技术在识别随机缺陷方面较为高效,通过对比相邻芯片区域的光学图像,即可检测随机缺陷。然而,该方法对系统缺陷的检测能力有限,因为相邻芯片区域的图像差分可能导致系统缺陷信号被抵消,无法通过简单的图像差分检测。相比之下,电子束成像技术因其高分辨率优势,可以直接检测随机缺陷和系统缺陷,其基本原理是将电子束扫描图像与无缺陷的参考图像或数据库进行对比。此外,不同类型的晶圆缺陷对检测方法的要求也存在差异。例如,裸晶圆上的缺陷通常表现为颗粒或划痕,并在高频散射分量上具有较高的灵敏度,因此适合采用暗场显微镜等光学检测系统。而对于已经图案化的晶圆,由于其结构复杂、材料多样,传统光学检测手段的有效性受到极大挑战,因而高精度设备、先进建模技术和图像后处理算法在晶圆缺陷检测中变得尤为关键。

目前,晶圆缺陷检测在很大程度上仍依赖于人工视觉检查。工程师通常根据缺陷类型推断晶圆制造过程中的故障原因,以优化生产工艺、提高产品良率[4]。例如,抛光垫的硬化可能导致晶圆表面产生划痕缺陷。然而,这种基于人工经验的检测方式存在成本高、效率低、易受主观因素影响等问题,亟需智能化的缺陷检测系统替代。早期的晶圆图缺陷模式分类方法主要基于缺陷簇的特征表征,通常采用特征提取与分类器学习相结合的两阶段策略。Hsu 等[5]在数据准备中通过增强信号方法建模并增强缺陷簇,之后采用修正豪斯多夫距离进行相似度计算。常见的传统方法包括基于 Radon 变换的投影特征分析以及缺陷簇的几何形态特征提取。Wu 等[6]采用上述特征,并利用支持向量机(support vector machines, SVM)分类器,在 WM-811K[7]数据集上实现了 83.1%的平均准确率。Piao 等[8]提出了一种基于决策树的集成学习方法,在 WM-811K 数据集的 Center、Donut、Random 和 None 模式识别方面表现良好,但在其他模式上的识别效果仍有不足。Saqlain 等[9]认为单一分类器难以应对复杂多变的缺陷模式,因此提出了一种结合几何特征、Radon 变换特征及密度信息的集成学习方案。该方法采用逻辑回归、随机森林、梯度增强和人工神经网络等多种分类器进行集成,显著提高了分类精度。此外,基于缺陷特征的聚类方法也被广泛应用,包括基于密度的聚类[10]、K-means 聚类[11]和层次聚类[12]。尽管早期方法在有限场景下有效,但是仍存在以下问题。

1)在晶圆缺陷中,投影特征、缺陷簇的几何特征提取和缺陷特征的聚类方法都是在单一维度中进行特征提取和特征分析,无法做到特征信息全面融合。

2)传统方法在复杂模式下准确率仍有限。

近年来深度学习方法在图像识别任务中取得了显著进展,如 Nakazawa 等[13]率先将卷积神经网络(convolutional neural network,CNN)应用于晶圆缺陷检测,利用浅层 CNN 训练模拟数据集,并在真实数据上取得了良好的分类效果。陈晓雷等[14]提出的全局与局部多尺度特征融合晶圆缺陷分类网络对多种缺陷分类表现优良。虽然基于深度学习的方法在晶圆缺陷分类中的有效性被大大证实,但也存在以下问题。

1)在晶圆缺陷的提取中欠考虑数据集不平衡的影响,虽然对个别种类缺陷识别准确率很高,但是无法排除这几种类别是否因为数据量少造成的。

2)深度学习的方法易发生梯度消失、爆炸等问题。在晶圆缺陷检测领域,工业生产实际获取的数据往往数量有限且类别分布不均衡,制约了数据驱动的深度学习技术的性能。以往研究尝试通过数据增强技术扩展少数类别样本,以缓解数据不平衡问题。然而,本文发现在基于原始数据集训练的深度学习模型中,少数类别的检测精度可能较高,而某些样本数量较多的类别反而识别率较低。例如,WM811K 数据集的标注样本统计结果表明,尽管 Near-Full 模式的样本量较小,但识别率较高,而 Edge-Local 和 Local 模式的样本量较大,但识别率反而较低。

根据缺陷的分布特征,本文推测这主要源于不同缺陷模式的特征复杂度差异。Near-Full 模式的特征较为明显,易于被模型识别,因此即使样本数量较少,也能获得较好的分类效果;而 Edge-Local 和 Local 模式的特征较为复杂,类别间的差异较小,导致模型难以准确区分。因此,这种特征复杂度引起的识别难度差异启发本文将其定义为特征分布不平衡,并将其作为研究数据集不平衡问题的新视角。

在工业实际应用中,获取的可用数据往往数量有限且分布不均,这极大地制约了数据驱动深度学习技术的发挥。在图像识别领域,VGGNet 相较于其他深度学习网

络具有显著优势。首先,与早期的网络架构(如 LeNet 和 AlexNet)相比,VGGNet 在网络深度和参数量上表现更为出色。LeNet 和 AlexNet 结构较浅、参数较少,因而在处理复杂图像时存在一定局限性。其次,VGGNet 采用了简单而高效的模块化设计,将网络划分为多个连续的卷积层和池化层,这不仅便于理解和实现,还使得网络能够更充分地捕捉图像特征,从而提升图像识别的准确性与性能。最后,VGGNet 展现出了较好的通用性和可迁移性,其简单的架构使得该模型能够较为轻松地适应不同的图像识别任务,并在多个领域和数据集上均取得了良好的表现。

基于对数据不平衡问题的深入分析,本文针对数据丢失、模型鲁棒性不足以及特征提取不充分导致的分类准确率低下问题,提出了一种融合改进多头注意力机制(improvement multi-head attention, IMHA)与改进 VGGNet(IVGGNet)的方法。

本文采用了一种基于 Transformer 的数据处理算法,以增强 VGGNet 的特征学习能力;提出改进的多头注意力机制,有效融合全局特征信息并有效缓解模型训练过程中因数据不平衡引发的信息丢失问题;对 VGGNet 进行改进,通过引入残差结构增强模型的鲁棒性,进而提升分类准确率。

1　相关工作

1.1　晶圆图缺陷模式分类

尽管传统方法在晶圆图缺陷模式分类中取得了一定进展,但仍然面临诸多挑战。首先,特征提取依赖于人工设计,而手工特征的质量对模型性能影响显著。其次,分类器的选择和参数调优较为复杂,集成学习方法虽然提高了分类精度,但也增加了计算成本和模型复杂性。近年来,深度学习的快速发展推动了智能制造技术的进步,为晶圆缺陷检测提供了新的解决方案。Nakazawa 等[15]在基于 CNN 的晶圆缺陷检测中还提出了一种基于深度学习的缺陷聚类分割方法,进一步验证了 VGGNet 在晶圆图缺陷检测中的可行性。Park 等[16]采用 Siamese 网络学习晶圆图的特征空间,并结合高斯均值聚类和离群点检测方法,以降低数据标注误差带来的不确定性。此外,集成卷积神经网络(Ensemble CNN)[17]结合 LeNet、AlexNet 和 GoogleNet 的主要权重,提高了晶圆图缺陷模式的识别率,尽管这也增加了计算开销。

然而,深度学习模型高度依赖于数据标注质量和样本分布均衡性。数据集的不平衡可能导致特征表达能力下降,进而影响分类性能。针对深度卷积神经网络(deep convolutional neural network, DCNN)的扩展研究已提出多种改进方案。Maksim 等[18]、Saqlain 等[19]和 Wang 等[20]通过合成样本和数据增强技术扩展少数类数据集,有效缓解了模型的泛化能力不足和过拟合问题。此外,基于 DenseNet 的迁移学习方法(T-DenseNet)[21]通过预训练权重实现快速泛化,无需大量数据即可达到良好的分类性能。尽管这些方法从数据数量分布的角度缓解了不平衡问题,但它们忽略了类间特征相似性较高时对分类模型的影响。

本文在此基础上,从特征提取和数据分布的双重视角,综合考虑数据不平衡问题及模型鲁棒性,提出融合改进多头注意力机制和改进 VGGNet 的晶圆缺陷分类方法,以提升分类精度并增强模型的泛化能力。

1.2　多头注意力机制

当人们观察一个物体时,通常会先快速浏览整体图像,再将注意力集中到关键区域,而忽略其他不重要的信息,这种现象即为视觉注意机制[22]。在晶圆图像处理中,视觉注意机制有助于 VGGNet 等网络捕捉不同类别之间的重要特征,从而提高对类间细微差异的识别能力。Jaderberg 等[23]提出的空间变换器层(spatial transformer layer)展现出出色的移位、旋转和尺度不变性,能够将原始空间信息映射到新的空间域,同时保留关键特征;而坐标注意机制则致力于挖掘特征之间的相对位置信息[24]。此外,挤压激励网络(squeeze-and-excitation networks, SENet)通过建立通道间依赖关系,有效增强了关键特征在模型决策中的影响[25]。上述方法均属于空间域的注意机制。

多头注意力机制(multi-head attention, MHA)[26]是在自注意力(self-attention)机制基础上发展而来的技术,最初在自然语言处理领域取得显著成效,随后逐渐在深度学习的各个方向得到广泛应用。其核心思想是利用多个注意力头(attention head)并行地对输入数据进行注意力计算,再将各头的计算结果进行融合。通过这种方式,模型可以从不同角度学习到多种特征表示,从而在捕捉长距离依赖关系、丰富特征组合和提升模型泛化能力等方面均表现出优势。

多头注意力机制的主要特点和优势体现在以下 5 个方面。

1)并行计算,多个注意力头可同时对输入序列进行处理,加快计算速度并提高效率。

2)特征融合,不同注意力头能够捕捉到不同的关注点,通过融合多头计算结果,可获得更丰富的特征表示,进一步提升模型性能。

3)抗噪性,各个注意力头独立工作,在一定程度上降低了噪声对整体特征表达的干扰。

4)解释性增强,每个注意力头对应不同的关注区域,有助于揭示模型决策过程中的关键因素。

5)泛化能力提升,通过学习不同粒度的特征表示,模

型在处理多样化任务和数据集时能够表现得更为稳健。

综上,多头注意力机制通过引入并行计算和特征组合的方法,有效提升了模型对 WM811K 图像数据集的特征提取能力,增强了对序列数据的建模和表达能力,特别在晶圆图像处理任务中显示出显著效果。

2 本文方法

基于不平衡晶圆数据集训练深度学习模型的方法主要步骤如下:1) 骨干网选择 VggNet-16。根据 Simonyan 等[27]的研究,本文采用了加入改进多头注意力机制和残差结构方案,结构如图 1 所示。原始 VGG 仅使用连续的卷积层和池化层。在引入混合注意力机制的基础上,特征提取层被拆分成低级特征提取层和高级特征提取层两部分。2)对多头注意力机制进行残差连接,有助于缓解深度模型中的梯度消失问题,并促进信息的流动。对于数据集全局信息和局部信息的处理方面对多头注意力机制采用混合注意力的改进,使其不仅对晶圆图像进行全局大范围模式识别,同时也能有效提取局部纹理信息和边缘特征。并且在注意力机制中加入了层归一化(layer norm),这对稳定训练过程和提高模型的收敛性能有很大的帮助;在注意力权重上引入 Dropout,进一步减少过拟合风险。3)在数据集处理方面增加了多种数据增强技术如随机裁剪、水平翻转、旋转等增强方法,提升泛化能力。

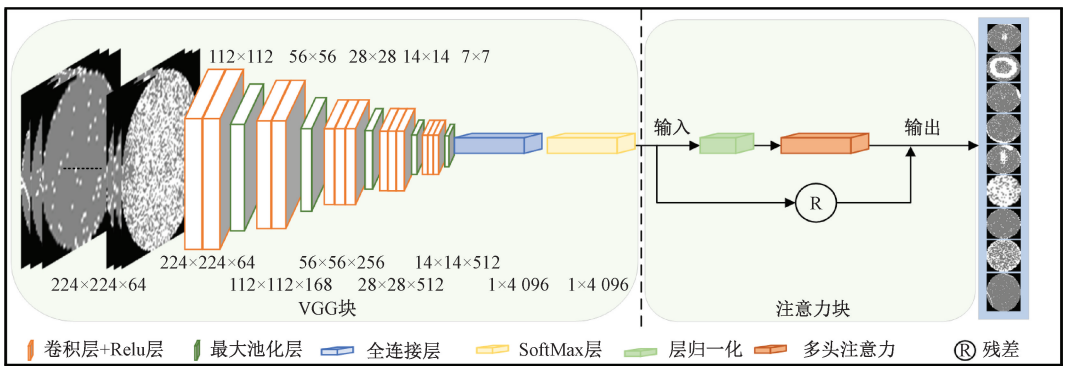


图 1 VGGNet 结构

Fig. 1 VGGNet structure diagram

2.1 骨干网络

在计算机视觉领域,许多优秀的 DCNN[28]架构被提出并应用于图像分类任务。本文选择 VGGNet 作为晶圆图缺陷模式分类的主干。VGGNet 作为图像处理的经典方法也是晶圆分类的有效方法之一。VGGNet 的核心是使用多个 3 × 3 的卷积核来代替较大的卷积核,使得网络层数更深。通过增加网络的深度来增强模型的特征提取能力。因此通过小卷积核和池化层使得模型的参数数量显著减少降低模型运行时间。

VGGNet-16 原理如图 1 所示。设输入特征图为 $\boldsymbol{X}$,卷积核为 $\boldsymbol{W}$,偏置为 b。卷积操作的输出 $\boldsymbol{Y}$ 表示为:

$$\boldsymbol{Y}=f(\boldsymbol{X}*\boldsymbol{W}+b) \tag{1}$$

式中:$*$ 表示卷积操作;f 表示激活函数。设池化窗口为 $k\times k$,步长为 s。池化操作的输出 $\boldsymbol{P}$ 可表示为:

$$P(i,j)=\max_{m,n\in k\times k}X(is+m,js+n) \tag{2}$$

式中:max 表示最大池化操作。设输入维度为 d,输出维度为 o,输入特征为 $\boldsymbol{X}$,权重矩阵为 $\boldsymbol{W}$,偏置为 b。全连接层的输出 $\boldsymbol{F}$ 可表示为:

$$\boldsymbol{F}=f(\boldsymbol{W}*\boldsymbol{X}+b) \tag{3}$$

在检测过程中,由于 WM811K 数据集中晶圆仓图已经被处理过,因此,图像特征的提取相对简单不需要复杂 DCNN 网络进行深度、重复提取,VGGNet 的小卷积核可以精确提取到晶圆仓图中的细小特征,更符合检测需求。鉴于晶圆图像相对简单的特点,本文采用了较为轻量的 VGGNet-16 作为网络的主干架构。该网络的 5 层卷积结构如图 2 所示。以网络中负责低级特征提取的初始卷积层为例,本文可以看到该层采用了 3×3 大小的卷积核,步长为 1,并进行了相同大小的填充,从而保证了输出特征图的尺寸不变,同时该层输出了 64 个通道的特征图。

为了进一步提升特征的表现力,本文在低级特征上应用了混合注意力机制,该机制采用了 256 维的嵌入表示和 8 个注意力头;对于更抽象的高级特征,本文也采用了混合注意力机制,使用了更高维度的 512 维嵌入和同样数量的 8 个注意力头,以捕捉更复杂的特征信息。

在特征提取的最后阶段,本文通过 7×7 的最大池化操作来降低特征图的维度,同时保留最显著的特征。为了促进模型的快速收敛并避免梯度消失问题,本文在每个卷积层之后引入了批量归一化(batch normalization)处理,并采用 ReLU 作为激活函数,增强了网络的训练效率和稳定性。本文设计的 VGGNet 不仅优化了网络性能,也保证了训练过程的顺利进行。

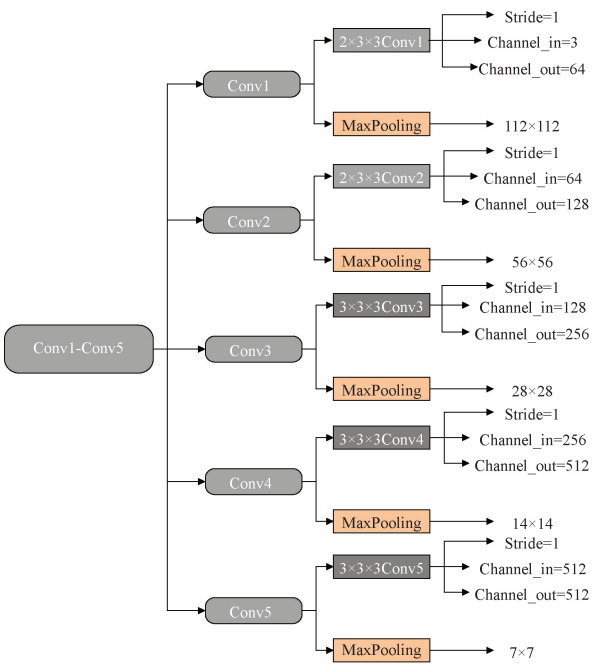


图 2　5 层卷积结构

Fig. 2　Five-layer convolutional structure diagram

本文采用 VGGNet-16 作为主干网络,针对晶圆图像的简单性进行了优化。在这个框架下,深度学习模型训练针对不平衡数据集的问题,可以被分解为特征表示学习和分类器学习两个独立阶段。本文中,位于 VGGNet-16 网络后端的全连接层承担了分类器模块的角色,而前面的卷积层以及与 SoftMax 相连的混合注意力机制构成了特征表示模块。此结构设计使得网络能够有效地提取图像特征,并通过分类器模块对九种缺陷准确分类。

2.2　改进多头注意力机制

计算机视觉中的注意机制[29]可以放大关键特征的影响,从而使模型能够抑制无关信息,增强特征表征。MHA 机制是许多现代深度学习架构,尤其是 Transformer 中的关键组成部分。多头注意力机制能够从不同的子空间中并行地提取信息,提高了模型的表达能力和学习特征的效果。具体流程如图 3 所示。

传统的注意力机制的核心思想是为输入的每个元素分配不同的权重,从而让模型能够关注到输入中不同部分的重要性,尤其是与当前任务相关的重要部分,其主要通过计算查询向量(query,Q)、键向量(key,K)和值向量(value,V)之间的相关性来实现。假设输入为 X,查询、键和值向量的线性投影为 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$,则多头注意力机制的计算如下。

线性投影:

$$\boldsymbol{Q}_i = X\boldsymbol{W}_i^Q \tag{4}$$

$$\boldsymbol{K}_i = X\boldsymbol{W}_i^K \tag{5}$$

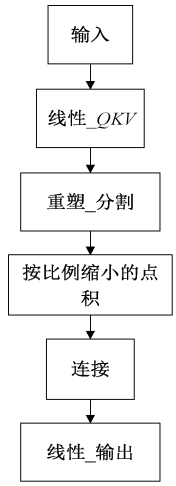


图 3　多头注意力机制流程

Fig. 3　Multi-head attention mechanism process

$$\boldsymbol{V}_i = X\boldsymbol{W}_i^V \tag{6}$$

式中:$\boldsymbol{W}_i^Q$、$\boldsymbol{W}_i^K$、$\boldsymbol{W}_i^V$ 是第 i 个头的投影矩阵。

缩放点积注意力:

$$Attention(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i) = softmax\left(\frac{\boldsymbol{Q}_i \times K_i^T}{\sqrt{dK}}\right)\boldsymbol{V}_i \tag{7}$$

式中:dK 是键向量的维度。

本文引入改进多头注意力机制,结构如图 4 所示,实现多维特征提取,并且有效缓解梯度消失。首先引入层归一化,这一步将 VGGNet 训练后输出的数据进行归一化处理,不仅提高了模型的稳定性和训练效率,还有效减轻梯度消失和爆炸的问题。其次通过分割多头并行计算,将从晶圆图中提取到的不同特征提高计算效率,丰富特征表示。最后将各个头分别学习到的特征信息进行连接(Concat),既增强了特征的组合性,也提升了模型的学习能力和表达能力。最终,通过线性层将晶圆图中的特征进行映射得到对应晶圆缺陷的注意力输出。

2.3　残差结构

深度网络[30]可以提高模型的表达能力,但容易造成梯度消失或梯度爆炸。本文的研究对象晶圆图虽然几乎没有语义信息,但纹理信息非常重要,并且浅层特征容易在深层结构中丢失。ResNet 是解决这些问题的有效方式,其中的残差学习 ResNet 的核心,残差单元[31]的示意图如图 5 所示。输入向量定义为 x,输出定义为 y,$F(x)$ 是残差函数。因此,残差单元的输出可表示为:

$$y = F(x) + x \tag{8}$$

本文采用的残差单元的核心是通过跳跃连接实现的。在传统的神经网络中,每一层的输出仅作为下一层的输入。而在本文残差网络中,将 VGGNet 输出与多头注意力输出相连接形成跳跃残差结构。这种设计允许梯度直接通过这些跳跃连接流动,从而有效地缓解了在深

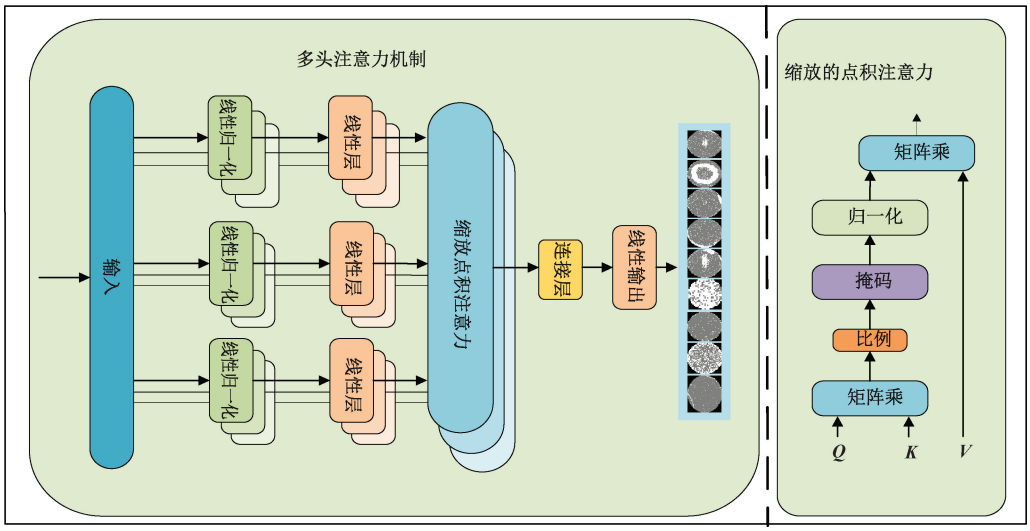


图 4 改进多头注意力机制结构

Fig. 4 Improved multi-head attention mechanism structure diagram

层网络中出现的梯度消失问题。本文残差单元的结构可以包含两个卷积层,以及一个跨层的直接连接。对于两个卷积层的残差块,输入先经过一个卷积层,然后通过 ReLU 激活函数,再经过另一个卷积层,最后与原始输入相加(通过恒等映射或 1×1 卷积进行维度匹配)。对于 3 个卷积层的残差块(bottleneck block),中间还包含一个 1×1 卷积层用于减少通道数,从而减少计算量。

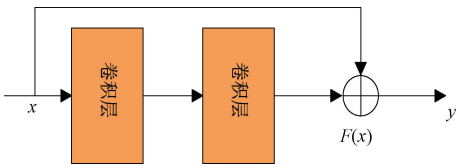


图 5 残差结构

Fig. 5 Residual structure diagram

本文加入残差单元不仅使得训练过程加速,而且允许 VGGNet 学习更多晶圆图像信息、捕捉更多特征使得分类和回归任务的表现取得更优质的提升。通过残差连接将 VGGNet 训练后的高纬度特征输出与多头注意力机制集成进行更细化维度的提取特征,使得网络有效减缓梯度消失梯度爆炸的问题,而且让网络对于晶圆缺陷分类更加精确化。

2.4 数据处理

在大多数图像分类任务中,通常只需关注目标类别的区分(如区分猫、狗或汽车),这主要依赖于 DCNNs 在通道域中提取的关键特征;然而,在晶圆图缺陷检测任务中,除了需要识别缺陷簇的几何属性(如面积、长度、形状等)外,更需要关注缺陷簇在晶圆图像中的绝对位置。

基于上述考虑,本文采用了一种基于 Transformer 的数据处理算法,以增强 VGGNet 的特征学习能力。对数据集施加了一系列变换(如图像反转、裁剪和尺寸调整等)[32],从而实现了 9 种缺陷模式图像数量的平衡——既减少了那些易于识别的图像数量,又增加了难以识别的图像数量[33]。这种数据处理策略不仅缓解了数据分布不均的问题,还提高了模型对各缺陷模式的区分能力。

3 实验与结果

3.1 WM-811K 数据集

本文采用的 WM-811K 晶圆地图数据集是迄今为止公开规模最大的晶圆地图数据集,其数据来源于实际生产过程。该数据集共包含 811 457 个样本,涵盖 9 种缺陷模式,其中仅约 21%的样本获得了标注。典型缺陷模式如图 6 所示,每种模式均反映了特定的制造过程故障信息,数据集中 9 种晶圆缺陷的分布情况如图 7 所示。

本文的实验均基于这些有标签样本进行。需要说明的是,由于某些类别样本数量较多,为平衡各类别样本数,实验仅选取了 10 000 个 None 模式样本。由于此数据集提供的晶圆图本质是按 Die 排列成的二维矩阵,不包含真实物理直径(如 200 mm/300 mm)。原始数据中的晶圆图尺寸从约 15×15、200×200 pixels 不等(代表不同的 Die 阵列规模)。为保证本模型对不同尺寸晶圆图都有准确的缺陷检测能力,在将晶圆图输入到模型前进行统一的预处理。由于常用的工业生产晶圆尺寸为 300 mm(即 12 in),本模型为确保采用的晶圆缺陷图像数据集更接近于真实晶圆图像,因此将每幅晶圆图预先处理为 256×256 的灰度图像。也就是说,图像中的每个像素代表一个晶粒,大约对应于毫米级别的物理尺寸(在

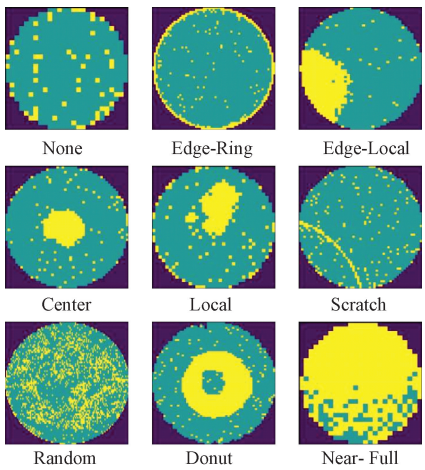


图 6　WM-811K 数据集中典型晶圆图缺陷模式

Fig. 6　Typical wafer defect patterns in the WM-811K dataset

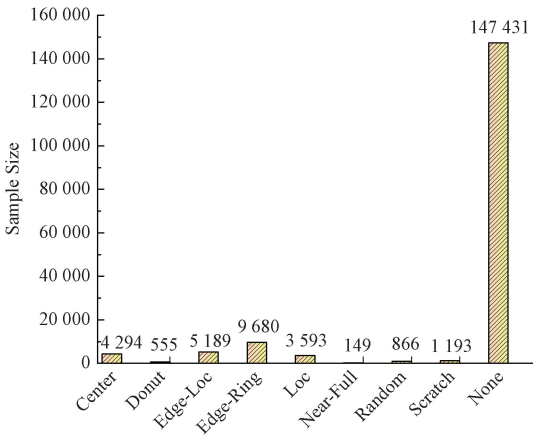


图 7　9 种晶圆缺陷图分布

Fig. 7　Distribution of nine types of wafer defects

300 mm 晶圆上约为 1.1 ~ 1.2 mm，每像素约 1 100 ~ 1 200 μm）。因此，本文模型能检测到的最小缺陷对应于单个不良晶粒（一个像素），其物理大小约为该量级。

模型的训练和测试均在 DELLT7920 工作站上进行，主要硬件配置包括一块 RTX 3060 显卡和 1 TB 内存。软件环境为 Python 3.8，基于 PyTorch 深度学习框架实现。模型训练采用交叉熵损失函数，初始学习率设为 0.000 1；当迭代次数达到总训练次数的 1/2 时，学习率降低至原来的 1/10。在表征学习阶段，结合改进的多头注意力机制，对 VGGNet-16 进行了 100 个 epoch 的训练；随后在分类器微调阶段，基于前一阶段的模型以相同学习率进行了 25 个 epoch 的微调。

3.2　评价参数和训练指标

在实验中，采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F1 评价指标来评价模型分类结果。准确率的定义如下：

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{9}$$

精确率和召回率的计算公式如下：

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

式中：真阳性（TP）是阳性样本和阳性预测的样本数量；真阴性（TN）是阴性样本和阴性预测的样本数量；假阳性（FP）是阴性样本但预测为阳性的样本数量；假阴性（FN）是阳性样本但预测为阴性的样本数量。

有时精确率和召回率指标之间存在矛盾，需要综合考虑这两个指标。结合精确率和召回率，可以得到 F1 分数的定义，如式（12）所示。

$$F_{1-score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

在模型训练阶段，由于 Loc 模式和 Edge-Loc 模式仅在缺陷的位置不同，因此不再使用随机裁剪等数据扩展方法。仅通过从训练集中随机旋转图像来扩展数据集，输入大小为 256×256。

3.3　与经典晶圆图检测算法比较

在 WM-811K 数据集中，本文模型在 100 个 epoch 的训练过程中，每个 epoch 的损失值如图 8 所示。从图 8 可以看出，仅经过 25 个批次的训练，损失曲线便已收敛，表明模型训练稳定。训练集与验证集的准确率曲线如图 9 所示，可以看出随着迭代次数的增加，训练集准确率逐渐趋近于 1，而验证集准确率最终稳定在约 94%。

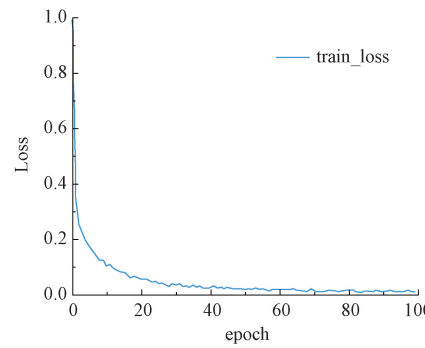


图 8　损失值曲线

Fig. 8　Loss value curve

本文所提出的改进多头注意力机制的 VGGNet 模型在 9 种晶圆图案上的分类结果如图 10 所示。结果表明，该模型在 9 种缺陷模式下均取得了较高的精度、召回率和 F1 分数。此外，为进一步直观反映各类别的分类效果，本文绘制了分类混淆矩阵（图 11），用于展示各缺陷模式中正确与错误分类的样本数量。

从混淆矩阵图 11 可以观察到，由于 Edge-Loc 与 Loc

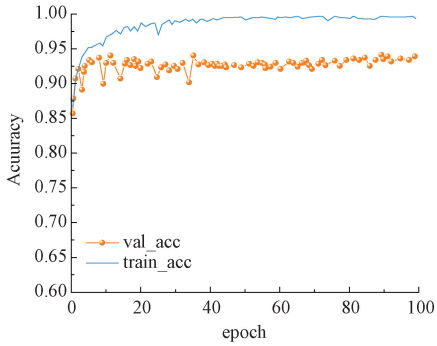


图 9 准确率曲线

Fig. 9 Accuracy curve

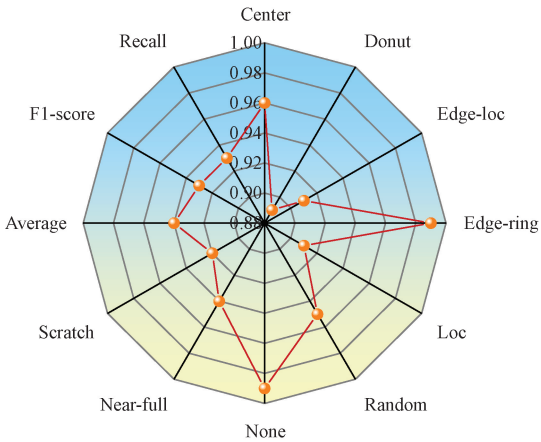


图 10 VGGNet 分类结果雷达图

Fig. 10 VGGNet classification results radar chart

两种缺陷图案形状较为相似,Edge-Loc 图案中分别有 0.77%和 3.62%的样本被误分类为 Edge-Ring 和 Loc 图案;而在 Loc 缺陷模式中,约有 1.45%的样本被错误归类为 Edge-Loc 模式。此外,Scratch 缺陷模式由于在晶圆图表面分布范围较广且分布不确定,经常被误判为其他类型;在 Loc 和 None 缺陷模式中,分别有约 4.18%和 12.13%的样本被误分类为 Scratch 模式。总体来看,除 Donut 模式外,本文方法在其他晶圆图案的分类准确率均超过 90%。

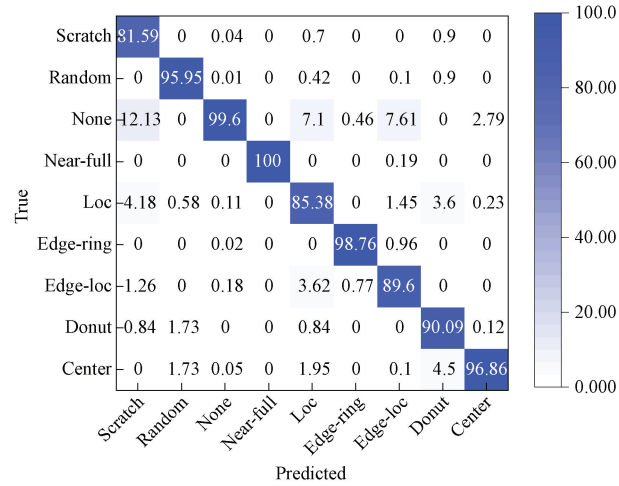


图 11 分类混淆矩阵

Fig. 11 Classification confusion matrix

将本文所提出的方法 IVGGNet 与近年来在 WM-811K 数据集上研究的几种模型进行了对比实验,比较对象包括 AlexNet、ResNet、GoogLeNet 和 MobileNet。实验中,分别对各缺陷模式的识别率和平均准确率进行了评估,结果如表 1 所示。结果表明,本文方法在 Edge-Local、Center、Local 和 Scratch 模式上均表现出较好的平均性能,而其他缺陷模式的检测结果与经典算法相当。

表 1 经典晶圆图检测算法比较

Table 1 Comparison of classic wafer pattern detection algorithms

模型	None	Edge-Ring	Edge-Local	Center	Local	Scratch	Random	Donut	Near-Full	Average
IVGGNet	**0.99**	**0.99**	**0.91**	**0.96**	**0.91**	**0.92**	0.95	0.89	0.94	**0.94**
AlexNet	0.99	0.98	0.84	0.94	0.79	0.91	0.99	**0.91**	0.88	0.91
ResNet	0.99	0.97	0.82	0.93	0.64	0.75	0.80	0.66	0.83	0.82
GooleNet	0.99	0.97	0.87	0.93	0.83	0.78	0.94	0.81	**1.00**	0.90
MobileNet	0.98	0.99	0.82	0.93	0.82	0.81	**1.00**	0.90	0.88	0.90

3.4 与最新晶圆图检测算法比较

将本文提出的模型与近年来的晶圆缺陷模式分类方法进行比较,如图 12 所示,包括 WM-PeleeNe[9]、WMDP[34]、DTE-WMFPR[35]。实验结果如表 2 所示。对比结果表明,本文提出的模型在 9 种晶圆图案中平均分类准确率为 94.3%,总体上优于其他所有模型。

表 2 对比方法

Table 2 Comparison methods

模型	*Accuracy*	*Recall*	*F*1-Score
IVggNet	0.94	0.93	0.93
WM-PeleeNet	0.93	0.94	0.93
WMDP	0.90	0.93	0.94
DTE-WMFPR	0.90	0.92	0.93

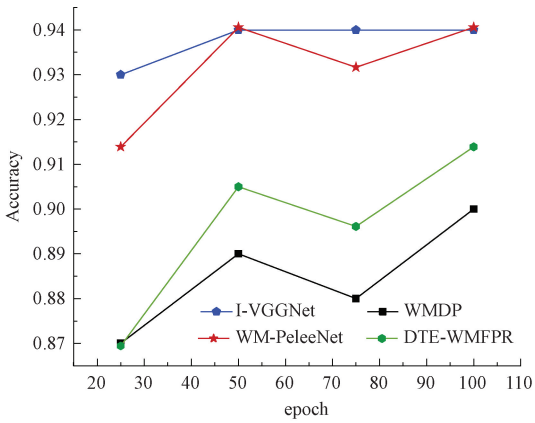


图 12　方法对比

Fig. 12　Comparison of methods

3.5　消融实验

为了验证所提出方法的有效性，本文在测试集上分别测试了 VGGNet 模型在不同改进方案下的晶圆图分类性能。分类结果对应的混淆矩阵如图 13 所示。实验结果表明，相较于原始 VGGNet 结构（图 13(b)），引入改进的多头注意力机制和残差结构后，缺陷模式的分类准确率显著提升（图 13(a)）。

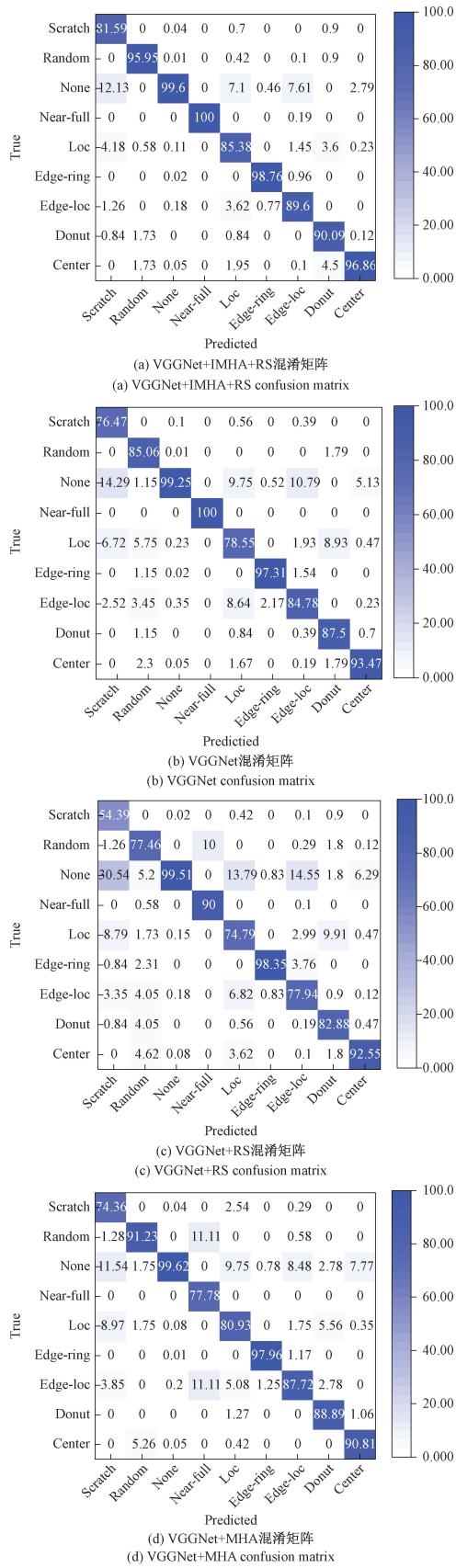
进一步分析混淆矩阵可得，相较于残差模块（图 13(c)），改进的多头注意力机制（图 13(d)）对 VGGNet 在晶圆图分类任务中的影响更为显著。这表明，引入多头注意力机制能够有效增强模型对关键特征的捕捉能力，提高分类精度。而残差模块的引入虽然对分类精度的提升相对较小，但在增强网络的稳定性和鲁棒性方面发挥了重要作用，使模型在面对复杂缺陷模式时具有更强的泛化能力。

表 3　消融对比

Table 3　Ablation comparison

模型	*Accuracy*	*Loss*	*F*1-Score
VGGNet	0.91	0.07	0.90
VGGNet+RS	0.91	0.04	0.91
VGGNet+MHA	0.92	0.06	0.90
VGGNet+IMHA+RS	0.94	0.06	0.93

通过对图 13 的混淆矩阵进行对比分析，可以观察到以下趋势。在引入本文所提出的 IMHA 机制和 RS 后（图 13(a)），大多数类别的预测准确性较高，特别是在 Center 类别上表现尤为出色。然而，在 Edge-Ring 和 Scratch 类别上仍然存在一定程度的误分类。相比之下，基础模型（图 13(b)）在各类别上的分类性能较差，尤其在 Edge-Ring、Near-Full 和 Random 类别中出现了较多的错误分类。仅引入 RS 后（图 13(c)），模型在 Edge-Ring 和 Scratch 类别上的分类精度有所提升，但误分类情况仍然



(a) VGGNet+IMHA+RS混淆矩阵

(a) VGGNet+IMHA+RS confusion matrix

(b) VGGNet混淆矩阵

(b) VGGNet confusion matrix

(c) VGGNet+RS混淆矩阵

(c) VGGNet+RS confusion matrix

(d) VGGNet+MHA混淆矩阵

(d) VGGNet+MHA confusion matrix

图 13　混淆矩阵对比

Fig. 13　Confusion matrix comparison

存在。这表明 RS 结构能够在一定程度上改善模型的特征提取能力,但对整体分类性能的提升相对有限。

仅引入 IMHA 后(图 13(d)),模型在 Center 和 Donut 类别上的分类效果显著提升,几乎实现了完全正确分类。然而,在 Edge-Ring 和 Scratch 类别上仍存在较多误分类,这表明 IMHA 对于中心区域和完整结构的特征识别较为有效,但在处理边缘区域和随机噪声方面仍存在一定局限性。

综上所述,图 13 的混淆矩阵对比表明,引入 IMHA 和 RS 后(图 13(a)),大多数类别的分类准确率均有所提升,尤其在 Center 类别上表现最优。然而,Edge-Ring 和 Scratch 仍然是较难分类的类别。基础模型(图 13(b))整体性能较差,特别是在 Edge-Ring、Near-Full 和 Random 类别上误分类较多。仅引入 RS(图 13(c)),在 Edge-Ring 和 Scratch 类别上的分类效果有所改善,但整体提升有限。仅引入 IMHA(图 13(d)),在 Center 和 Donut 类别上的分类表现显著提升,但对 Edge-Ring 和 Scratch 仍存在较大误分类率。综合来看,IMHA 和 RS 的联合应用能有效提升模型的整体分类性能,其中 RS 对于减少特定类别的误分类具有一定作用,而 IMHA 更擅长捕捉中心区域的特征,但对边缘和随机噪声的处理能力相对较弱。

4 结 论

晶圆图检测是半导体制造中缺陷检测的重要环节。尽管基于 DCNN 的方法在晶圆图缺陷模式识别方面取得了显著进展,但由于训练过程中数据集的不平衡性,DCNN 的性能仍受到一定限制。为此,本文提出了一种基于 VGGNet 结构的改进神经网络模型 IVGGNet,以提升不平衡数据下的缺陷检测能力。

首先,在 VGGNet 结构中引入了 IMHA,增强了网络的特征表示能力,使其能够更深入地挖掘困难样本的特征,并通过不同子空间的线性变换并行执行注意力计算。这不仅有效缓解了特征分布不均衡的问题,还优化了梯度传播,避免了梯度消失和梯度爆炸,提高了训练的稳定性。此外,为了加速收敛并进一步提升模型性能,在 VGGNet 的全连接层中加入了残差结构模块,以改善信息流动并增强特征学习能力。其次,在数据预处理阶段,引入数据增强策略,以缓解数据不平衡问题。通过增加数据集的多样性,使其更接近于实际应用场景,从而提高模型的泛化能力和鲁棒性。该方法在不平衡的 WM-811K 数据集上进行了实验验证,结果表明 IVGGNet 在晶圆缺陷模式分类任务中表现优越,最终准确率达到 94.3%,显著优于近期提出的其他方法。此外,由于该方法无需修改原始数据集,因此具有较高的可移植性和应用价值。

实验结果表明,模型在 Scratch、Loc 和 Edge-Loc 模式上仍然存在一定程度的误分类问题。此外,当前方法仍然依赖于大量标注数据,在少样本缺陷检测场景下识别能力有限。因此,未来的研究将重点关注提升 Scratch、Loc 和 Edge-Loc 模式的分类精度,并探索小样本晶圆缺陷检测方法,以进一步增强模型在实际工业环境中的适应性。半导体发展表明,其产业对晶圆缺陷检测容忍度降低,验证研究显示缺陷漏检率需控制在千分之几甚至更低级别。换言之,要实现行业要求,识别率应接近 99%。本文提出的识别率为基于便捷模型、易于部署达到的优良识别率和识别效率。虽然本文提出的 94.3%较高,但与理想水平仍有差距。后续将通过增强稀有类别样本、数据增强和继续进行模型优化等手段继续提升识别率及其在工业环境实际应用能力。

参考文献

[1] XIA M, MU X, WU Z. An end-to-end wafer map defect recognition model[C]. 2024 International Conference on Computer Engineering and Application (ICCEA). IEEE, 2024: 1205-1210.

[2] LIU J, ZHAO H, WU Q, et al. Patterned wafer defect inspection at advanced technology nodes[J]. Laser & Optoelectronics Progress, 2023, 60(3): 0312003.

[3] 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.

[4] ADLY F, ALHUSSEIN O, YOO P D, et al. Simplified subspaced regression network for identification of defect patterns in semiconductor wafer maps [J]. IEEE Transactions on Industrial Informatics, 2015, 11(6): 1267-1276.

[5] HSU C Y, CHEN W J, CHIEN J C. Similarity matching of wafer bin maps for manufacturing intelligence to empower industry 3.5 for semiconductor manufacturing [J]. Computers & Industrial Engineering, 2020, 142: 106358.

[6] 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.

[7] 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, 2023, 34(8): 3215-3247.

[8] PIAO M, JIN C H, LEE J Y, et al. Decision tree ensemble-based wafer map failure pattern recognition based on radon transform-based features [J]. IEEE Transactions on Semiconductor Manufacturing, 2018, 31(2): 250-257.

[9] SAQLAIN M, JARGALSAIKHAN B, LEE J Y. A voting ensemble classifier for wafer map defect patterns identification in semiconductor manufacturing[J]. IEEE Transactions on Semiconductor Manufacturing, 2019, 32(2): 171-182.

[10] JIN C H, NA H J, PIAO M, 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] CHEN X, ZHAO C, CHEN J, et al. K-means clustering with morphological filtering for silicon wafer grain defect detection[C]. 2020IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 2020, 1: 1251-1255.

[12] LEE S, KIM D. Distributed-based hierarchical clustering system for large-scale semiconductor wafers[C]. 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). IEEE, 2018: 1528-1532.

[13] 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.

[14] 陈晓雷,李正成,杨富龙,等. 全局与局部多尺度特征融合晶圆缺陷分类网络[J]. 电子测量与仪器学报, 2024,38(10):159-169.
CHEN X L, LI ZH CH, YANG F L, et al. Wafer defect classification network with global and local multi-scale feature fusion[J]. Journal of Electronic Measurement and Instrumentation, 2024,38(10):159-169.

[15] NAKAZAWA T, KULKARNI D V. Anomaly detection and segmentation for wafer defect patterns using deep convolutional encoder-decoder neural network architectures in semiconductor manufacturing[J]. IEEE Transactions on Semiconductor Manufacturing, 2019, 32(2): 250-256.

[16] PARK S, JANG J, KIM C O. Discriminative feature learning and cluster-based defect label reconstruction for reducing uncertainty in wafer bin map labels[J]. Journal of Intelligent Manufacturing, 2021, 32(1): 251-263.

[17] HSU C Y, CHIEN J C. Ensemble convolutional neural networks with weighted majority for wafer bin map pattern classification[J]. Journal of Intelligent Manufacturing, 2022, 33(3): 831-844.

[18] MAKSIM K, KIRILL B, EDUARD Z, et al. Classification of wafer maps defect based on deep learning methods with small amount of data [C]. 2019 International Conference on Engineering and Telecommunication (EnT). IEEE, 2019: 1-5.

[19] 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.

[20] WANG R, CHEN N. Defect pattern recognition on wafers using convolutional neural networks [J]. Quality and Reliability Engineering International, 2020, 36(4): 1245-1257.

[21] SHEN Z, YU J. Wafer map defect recognition based on deep transfer learning [C]. 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). IEEE, 2019: 1568-1572.

[22] MNIH V, NICOLAS G, ALEX, et al. Recurrent models of visual attention [C]. 28th Conference on Neural Information Processing Systems (NIPS), 2014: 2204-2212.

[23] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks [C]. 29th Annual Conference on Neural Information Processing Systems (NIPS),2015:2017-2025.

[24] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021: 13713-13722.

[25] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[26] GAO G. Survey on attention mechanisms in deep learning recommendation models[J]. Computer Engineering and Applications, 2022, 58(9): 9-18.

[27] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [C]. 3rd International Conference on Learning Representations, 2015.

[28] HOU X, YI M, CHEN S, et al. Recognition and classification of mixed defect pattern wafer map based on multi-path DCNN [J]. IEEE Transactions on Semiconductor Manufacturing, 2024, 37(3): 316-328.

[29] DENG G, WANG H. Efficient mixed-type wafer defect pattern recognition based on light-weight neural network[J]. Micromachines, 2024, 15(7): 836.

[30] SUN X, ZHANG B, WANG Y, et al. A multiscale attention mechanism super-resolution confocal microscopy for wafer defect detection [J]. IEEE Transactions on Automation Science and Engineering, 2024, 22: 1016-1027.

[31] LI C, LI X, LI H, et al. A remote sensing image denoising method fused with multi-scale features [J]. Electronics Optics & Control, 2024, 31(6): 74.

[32] SUN S, FAN J, SUN Z, et al. A survey of image data augmentation based on deep learning [J]. Journal of Computational Science, 2024, 51: 150-167.

[33] MOLES L, ANDRES A, ECHEGARAY G, et al. Exploring data augmentation and active learning benefits in imbalanced datasets [J]. Mathematics, 2024, 12(12): 1898.

[34] 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.

[35] PIAO M, JIN C H, LEE J Y, et al. Decision tree ensemble-based wafer map failure pattern recognition based on radon transform-based features [J]. IEEE Transactions on Semiconductor Manufacturing, 2018, 31(2): 250-257.

作者简介

杜先君(通信作者),2002 年于兰州理工大学获得学士学位,2008 年于兰州理工大学获得硕士学位,2013 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及其在建模、优化、控制与诊断中的应用,包括污水处理过程的软测量建模;污水处理过程的多目标优化控制;设备/元件故障诊断与寿命预测;元启发式搜索算法;遥感图像云检测、厄尔尼诺/拉尼娜预测、短临预测等。

E-mail: xdu@ lut. edu. cn

Du Xianjun (Corresponding author) received his B. Sc. degree from Lanzhou University of Technology in 2003, M. Sc. Degree from Lanzhou University of Technology in 2008, and Ph. D. degree from Lanzhou University of Technology in 2013, respectively. Now he is an associate professor at Lanzhou University of Technology. His main research interests include artificial intelligence and its applications in modeling, optimization, control, and diagnosis, including soft sensor modeling of wastewater treatment processes; multi-objective optimization control of wastewater treatment processes; equipment/component fault diagnosis and life prediction; meta-heuristic search algorithms; remote sensing image cloud detection, El Niño/La Niña prediction, short-term forecasting, etc.

贾龙,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: 1429016512@ qq. com

Jia Long 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.