

DOI: 10.13382/j.jemi.B2508626

基于 CNN-Transformer 的多级特征融合晶圆缺陷分类网络^{*}

陈晓雷 李正成 沈星阳 马平川 杨富龙

(兰州理工大学微电子现代产业学院 兰州 730000)

摘要:晶圆缺陷分类是芯片制造过程中至关重要的一环,准确分类各种类型的缺陷模式对快速定位芯片制造问题并加以修复具有重要意义。针对晶圆缺陷分类中存在的多尺度特征表征不足和跨层次信息融合效率低等问题,提出了基于卷积神经网络(CNN)结合 Transformer 的多级特征融合晶圆缺陷分类网络(CNN-Transformer multi-level feature fusion network, CTM-Net)。该网络采用双分支并行结构,CNN 分支负责提取局部形态和细微缺陷特征,Transformer 分支利用自注意力机制建模全局上下文关系以获取全局结构特征,在此基础上,CTM-Net 通过构建多级特征交互模块,实现跨尺度、跨层次信息的自适应融合与增强,从而有效弥补单一分支在特征表达上的不足,提升模型的判别能力。CTM-Net 在公开 WM-811K 数据集上的分类准确度达到 99.31%,在整体分类性能上优于当前主流的 5 种代表性先进晶圆缺陷分类方法,尤其在 Edge-Loc、Edge-Ring、Loc 等关键缺陷类别上表现出了更高的识别精度与稳定性。这一结果证明了 CNN 和 Transformer 多级融合方法在处理晶圆缺陷分类任务中的有效性和优势。

关键词:晶圆缺陷分类;单类型缺陷;特征融合

中图分类号: TN305;TP18;TP391.41

文献标识码: A

国家标准学科分类代码: 510.4050

Multi-level feature fusion wafer defect classification network based on CNN-transformer

Chen Xiaolei Li Zhengcheng Shen Xingyang Ma Pingchuan Yang Fulong

(School of Microelectronics Industry-education Integration, Lanzhou University of Technology, Lanzhou 730000, China)

Abstract: Wafer defect classification is a critical task in the chip manufacturing process, as accurately identifying various types of defect patterns is essential for promptly locating and resolving manufacturing issues. To overcome the limitations of insufficient multi-scale feature representation and low efficiency in cross-level information fusion, this paper proposes a novel multi-level feature fusion network based on convolutional neural networks (CNN) and Transformers, named CTM-Net. The network adopts a dual-branch parallel architecture; the CNN branch focuses on extracting local morphological structures and fine-grained defect details, while the Transformer branch employs self-attention mechanisms to capture global contextual dependencies and structural information. Based on these complementary representations, a multi-level feature interaction module is further designed to adaptively integrate and enhance cross-scale features, effectively compensating for the limitations of single-branch representations and improving the model's discriminative capability. Experiments conducted on the publicly available WM-811K dataset show that CTM-Net achieves an accuracy of 99.31%, outperforming five representative state-of-the-art wafer defect classification approaches. Notably, the proposed model demonstrates superior recognition accuracy and stability in critical defect categories such as Edge-Loc, Edge-Ring, and Loc, highlighting the effectiveness and advantages of combining CNN and Transformer architectures for multi-level feature fusion in wafer defect classification tasks.

Keywords: wafer defect classification; single-type defects; feature fusion

0 引言

半导体芯片是微电子制造业的核心,是智能系统和自动化技术发展的基础,在未来科技中具有关键作用,广泛应用于信息、通信、医疗、交通等领域。随着 5G 网络、物联网(IoT)、人工智能等新兴技术的发展,半导体芯片需求将会持续增长^[1]。晶圆作为制造半导体芯片的基础材料,是半导体工艺中不可或缺的一部分。在芯片代工中,晶圆需经过光刻、刻蚀、沉积等多道复杂工序,每一步都可能造成缺陷,这些缺陷不仅影响芯片良率,还可能导致功能失效^[2]。因此,晶圆缺陷检测成为保障产品质量的重要环节。晶圆图(wafer map)是晶圆经过电气测试后的结果,表面上的缺陷图案对帮助工程师识别缺陷产生原因、促进问题解决和提高晶圆生产良率具有关键作用。晶圆缺陷分类技术通过自动识别和分类晶圆图上的多种缺陷类型(如边缘缺陷、环状缺陷、划痕缺陷等),实现对缺陷模式的准确区分,帮助追溯缺陷的工艺来源,实现精准定位和快速工艺优化。该技术不仅是提升生产效率和产品可靠性的关键手段,也是推动智能制造和良率提升的重要支撑。由于人工目视检测存在效率低、精度不足、成本高、劳动强度大以及标准化欠缺等弊端,晶圆缺陷分类技术早已成为替代人工目测的主流方案并得到了广泛应用^[3]。

最初,机器学习方法被用于晶圆缺陷分类。Xie 等^[4]提出了一种基于支持向量机(support vector machines, SVM)的晶圆缺陷分类方法。Yuan 等^[5]提出了一种多步骤缺陷分类方法,采用 K 最近邻(K-nearest neighbors, KNN)去噪技术区分局部与全局缺陷,并通过聚类算法根据空间分布对缺陷进行分类。Piao 等^[6]提出了一种基于决策树的晶圆缺陷分类方法,该方法首先采用 Radon 变换提取缺陷特征,然后度量特征之间的相关性,最后通过聚合决策树的预测结果做出最终分类。然而,随着集成电路层数和晶圆集成密度的快速提升,传统机器学习方法在晶圆缺陷分类中的性能显著下降^[7]。

随着卷积神经网络(convolutional neural networks, CNN)的迅速普及,其在图像处理和模式识别领域展现出强大的能力,为晶圆缺陷分类提供了新的解决方案^[8]。Saqlain 等^[9]提出了一种通过数据增强和正则化等技术来提升晶圆分类模型性能的方法。Nakazawa 等^[10]提出了一种基于 CNN 的晶圆缺陷模式分类方法。通过合成 28 600 个缺陷类别的晶圆图,用于 CNN 的训练、验证和测试,测试数据集的分类准确率为 98.2%。Wang 等^[11]提出了一种基于 CNN 的新型方法,用于自动识别晶圆图上的缺陷模式。通过在训练前使用极坐标映射将圆形晶圆转换为矩阵,解决了晶圆尺寸和芯片尺寸变化的问题,

同时减小了输入尺寸。Lee 等^[12]提出了一种名为 FDC-CNN 的模型,该方法能够将卷积层的输出与原始数据的结构意义相关联,进而定位故障变量和时间信息。Cheon 等^[2]采用 CNN 模型,直接从图像中提取有效特征进行分类,省去额外的特征提取步骤。Kyeong 等^[13]使用 CNN 在单独分类模型框架下对晶圆箱图(wafer bin map, WBM)中的混合型缺陷模式进行分类。Yu 等^[14]通过调整 PeleeNet 网络中密集块的结构,降低了晶圆缺陷分类模型复杂度,并在隐藏层中加入随机高斯噪声,采用卷积自动编码器进行数据增强,解决了数据不平衡问题。Han 等^[15]重点提取晶圆图旋转不变和芯片尺寸不变的稳健特征,相较于传统方法,该方法不仅提高了分类准确性,还增强了鲁棒性。Manivannan^[16]提出了一种基于 CNN 集成的半监督方法,用于自动识别 WBM 中的缺陷模式。该方法结合标记和未标记数据进行训练,克服了全监督方法对大量标注数据的依赖。同时采用 Label-Smoothing 和 Ensembling 技术缓解 CNN 高置信度错误分类问题。陈晓雷等^[17]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,显著提高了晶圆缺陷分类精度。他们还提出了一种基于全局和局部多尺度特征融合的混合型晶圆缺陷分类网络,证明了全局与局部特征融合在处理混合型晶圆缺陷分类任务中的重要性和有效性^[18]。

尽管 CNN 在提取局部特征方面表现优越,但其感受野有限,在捕捉晶圆图像全局结构信息方面仍存在一定不足,尤其在面对具有跨区域延展性或局部结构重叠的复杂缺陷时,容易出现误判,暴露出其全局建模能力的不足。因此,为弥补 CNN 的局限性,研究者开始将目光转向基于自注意力机制的 Transformer 模型。

Transformer 模型通过多头自注意力机制对输入序列进行全局建模,克服了 CNN 依赖局部感受野和逐层提取的局限性,尤其在捕捉长程依赖关系方面具有优势^[19]。目前,在晶圆缺陷分类领域,研究者也探索了多种基于 Transformer 的方法^[20]。Nafi 等^[20]研究了 4 种基于 Transformer 的模型,分别为 BEiT(BERT pre-training of image Transformers)、FNet(Fourier network)、ViT(vision Transformer)和 Swin Transformer(shifted window based Transformer)在晶圆缺陷分类任务中的表现,结果表明 Swin Transformer 在 MixedWM38 数据集上取得了最优性能。Fan 等^[21]基于 ViT 模型中多头注意力机制可生成多样特征热图的特点,提出了一种新型数据增强框架。Mohammad 等^[22]提出了一种针对晶圆缺陷分类优化的轻量级 ViT 框架——ViT-Tiny,与 ViT-Base 相比,该模型在保持较低计算复杂度的同时提升了分类精度。Zhang 等^[23]提出了一种多级中继 ViT(MLR-WM-ViT),该方法利用增强的卷积加法自注意力机制实现线性复杂度的全

局上下文建模,结合 MobileOne Stem 和空洞空间金字塔池化模块,提升多尺度特征提取能力。另外,该方法能够有效利用基本缺陷、缺陷数量和缺陷类型之间的关系,显著提高检测性能。Chang 等^[24]提出了一种基于多头注意力机制的特征提取网络,融合晶圆图的全局与局部信息,结合对比学习机制构建无监督异常检测算法,提出 ConViNet 框架提升特征提取与异常检测的鲁棒性,在真实数据集上表现出优异的性能和强稳定性。

然而,由于 Transformer 缺乏卷积操作中固有的归纳偏置(如局部性和平移不变性),其建模过程更依赖大规模训练数据和复杂的训练策略,导致 Transformer 中的自注意力机制往往无法充分捕捉局部特征细节。当局部结构信息较弱或缺陷边界模糊时,Transformer 可能难以有效区分目标缺陷与背景,从而影响分类性能。尤其在晶圆缺陷分类任务中,许多缺陷模式仅在细节上存在微小差异,若模型缺乏对局部区域的敏锐感知能力,便可能产生误判或漏检。

为克服单独使用 CNN 或 Transformer 进行晶圆缺陷分类的局限性,近年来有研究人员提出了联合使用 CNN 或 Transformer 的方法,旨在同时发挥 CNN 在局部特征提取方面的优势和 Transformer 在全局建模方面的能力,从而提升模型的整体性能和泛化能力。Wei 等^[25]提出了一种混合型缺陷分类框架 MSF-Trans,首先利用 CNN 提取高级语义区域特征及邻域关系,有效帮助网络聚焦于区域相关信息,从而减少背景噪声干扰,增强特征表达。在此基础上,将提取的特征图输入 6 个串联的 Transformer 编码器,并为特征图中的每个位置分配独立的注意力权重。然而,该方法仅在 Transformer 输入前利用 CNN 强化了缺陷的局部特征提取能力,并未从根本上优化 Transformer 结构本身在晶圆图像建模中存在的局限性。同样地,Xia 等^[26]提出了一种基于 CNN-Transformer 结构的晶圆缺陷分类方法,该方法首先采用 U-Net 结合可变形卷积来增强 CNN 主干网络对不同类型缺陷形态特征的捕捉能力。在完成特征提取与重建后,再将其输入 Transformer 进一步提取特征。尽管该方法在处理小目标缺陷方面展现出更好的细节保留能力,但其依然未对 Transformer 自身的结构缺陷进行改进,仅依赖 CNN 前置模块增强特征表达。

现有 CNN 与 Transformer 结合的晶圆缺陷检测方法在面对缺陷类型多样、形态复杂以及具有长距离依赖特征的图像时,相比单纯依赖 CNN 或者 Transformer 的方法具有更强的建模能力。然而,这些方法大多数采用串联结构,即先利用 CNN 进行特征提取,再将特征输入 Transformer 进行全局建模。这种串联结构虽然增强了 Transformer 对局部细节的感知能力,但仍存在如下局限: 1) 特征提取路径单一,缺乏局部与全局特征的同步建模

能力;2) 特征融合大多集中在网络后期,难以实现跨层次的深度语义交互;3) 在处理小目标缺陷、类别不平衡等实际问题时,泛化能力和鲁棒性仍有待提升。

为克服上述问题,本文提出了一种基于 CNN 与 Transformer 的多级特征融合晶圆缺陷分类网络(CNN-Transformer multi-level feature fusion network, CTM-Net)。该网络采用并联双分支结构,分别由 CNN 与 Transformer 并行提取局部细节与全局上下文信息,并在每一层下采样阶段实现两分支的逐层融合,从而更充分地挖掘不同尺度与语义层次的互补特征,提升模型对复杂缺陷的识别能力。

本文结合 CNN 和 Transformer 的优点,构建一个既能提取局部细节又能捕捉全局上下文信息的双分支并联晶圆缺陷分类网络;提出了一种多级特征交互模块(multi-level feature interaction block, MFIB),用于有效整合不同级别特征之间的语义信息,该模块结合了空间注意力机制、通道注意力机制以及倒置残差(inverted residual mobile block, iRMB^[27]);针对晶圆图像中的数据不平衡问题,结合 CNN 和 Transformer 更有效地进行数据增强和自适应学习。

1 本文方法

1.1 CTM-Net 总体结构和工作原理

CTM-Net 的总体结构如图 1 所示,主要由 CNN-Transformer 特征融合网络和分类网络组成。模型首先在初始阶段通过卷积操作对输入图像进行预处理,生成初始特征($56 \times 56, 96$ 通道)。然后 CNN 分支提取局部特征,Transformer 分支提取全局语义信息。最后,利用 MFIB 模块融合 CNN 和 Transformer 提取的特征。第 1 阶段保持输出尺寸为 56×56 ,通道数为 96,随后在第 2~4 阶段通过下采样(分别降至 28×28 、 14×14 和 7×7)逐步提取更深层次的特征,并增加通道数(192、384 和 768)。每个阶段均通过空间注意力和通道注意力机制在 MFIB 模块中融合多层次特征,并通过局部和全局特征的有效结合增强模型的特征表达能力。最终,经过全局平均池化和线性分类器进行分类。

1.2 MFIB 模块

为了提升晶圆缺陷分类的准确性,需要融合不同层次的局部空间特征和全局语义信息。为此,本文提出了多级特征交互模块 MFIB。MFIB 的整体结构如图 2 所示,其中 C_i 表示由 CNN 分支生成的特征矩阵,负责提取晶圆图像的局部空间特征。 T_i 表示由 Transformer 分支生成的特征矩阵,负责提取晶圆图像的全局语义信息。 M_i 表示 MFIB 单元上一级生成的特征矩阵, M_{i+1} 表示在该

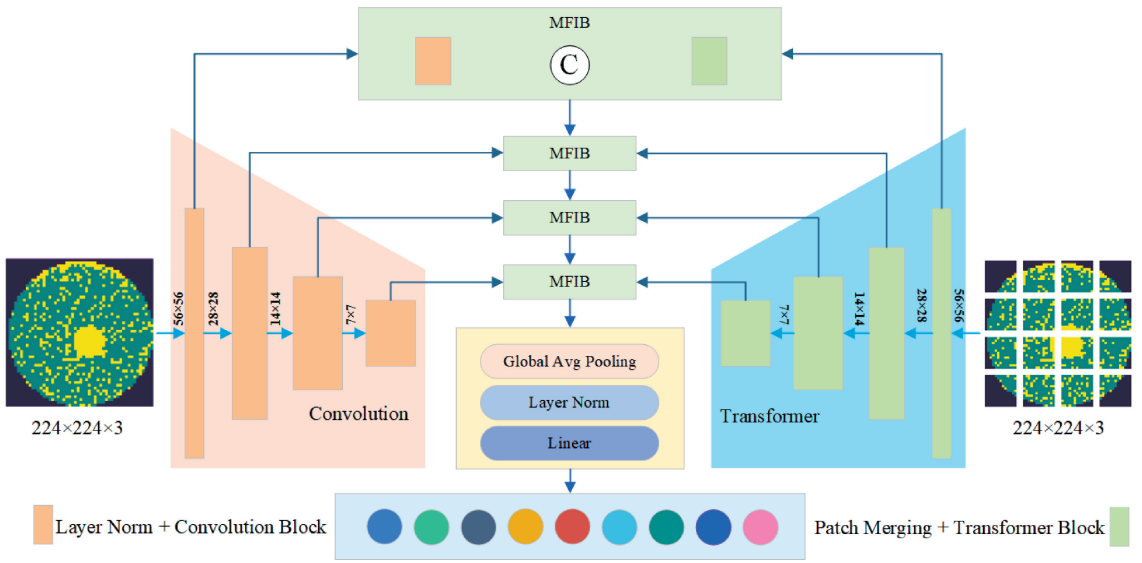


图 1 CTM-Net 总体架构

Fig. 1 Overall architecture of CTM-Net

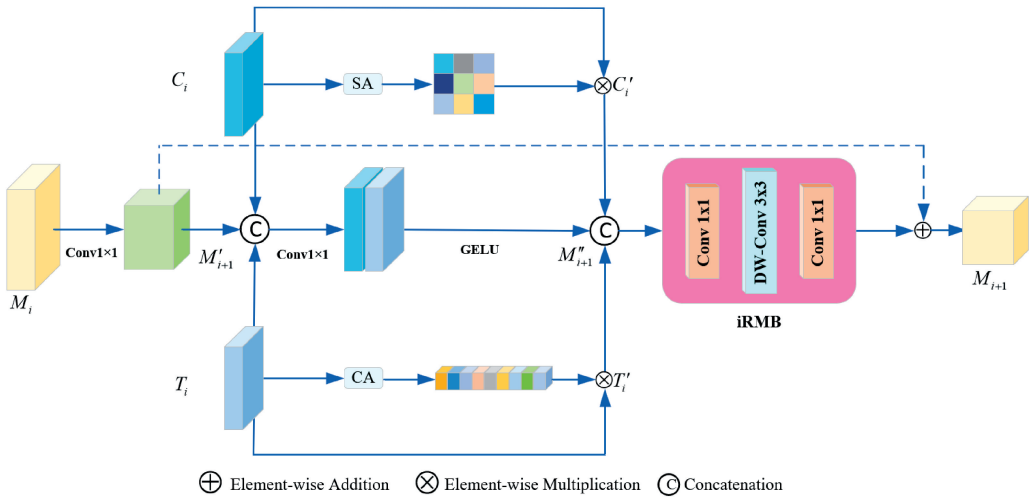


图 2 MFIB 单元结构

Fig. 2 Structure of the MFIB unit

级生成的特征矩阵。MFIB 通过多次融合上一级特征以及同级的局部空间特征和全局语义信息,并结合一系列卷积操作进一步强化特征表达能力。经过四级逐级融合后,最终生成更为丰富且具有代表性的特征矩阵,为后续的缺陷分类任务提供了强有力的特征支持。特征融合过程如下公式所示:

$$C'_i = SA(C_i) \otimes C_i \quad (1)$$

$$T'_i = CA(T_i) \otimes T_i \quad (2)$$

$$M'_{i+1} = Avgpool(Conv1 \times 1(M_i)) \quad (3)$$

$$M''_{i+1} = Conv1 \times 1(Concat[C_i, T_i, M'_{i+1}]) \quad (4)$$

$$M_{i+1} = iRMB(LN(Concat[T'_i, C'_i, M''_{i+1}])) + M'_{i+1} \quad (5)$$

式中:⊗表示元素相乘; C'_i 由空间注意力生成; T'_i 由通道注意力生成; M'_{i+1} 由 MFIB 单元上一阶段下采样生成; M''_{i+1} 是 CNN 和 Transformer 及上一阶段生成的特征; LN 是 LayerNorm; M_{i+1} 是 MFIB 模块的输出特征。

1.3 CNN 分支

晶圆图像中的局部空间特征至关重要, CNN 和 Transformer 分支如图 3 所示。图 3(a)中, CNN 分支使用 3×3 深度卷积进行局部空间特征提取。深度卷积能够将分组卷积中的分组维度与输入通道维度保持严格一致(即分组数 C = 通道数 C)。这种轻量化设计在保持空间特征捕获能力的同时,显著降低了模型的计算复杂度。随后,通过线性变换层建立跨通道特征关联。最终,提取

到的局部特征被输入到 MFIB 模块中进行多特征融合。该过程如式(6)所示。

$$C_{i+1} = DConv1 \times 1(LN(DConv3 \times 3(C_i))) + C_i \quad (6)$$

式中: C_i, C_{i+1} 分别表示 CNN 分支的输入、输出特征; $DConv3 \times 3, DConv1 \times 1$ 分别代表使用 $3 \times 3, 1 \times 1$ 卷积核的深度卷积操作。

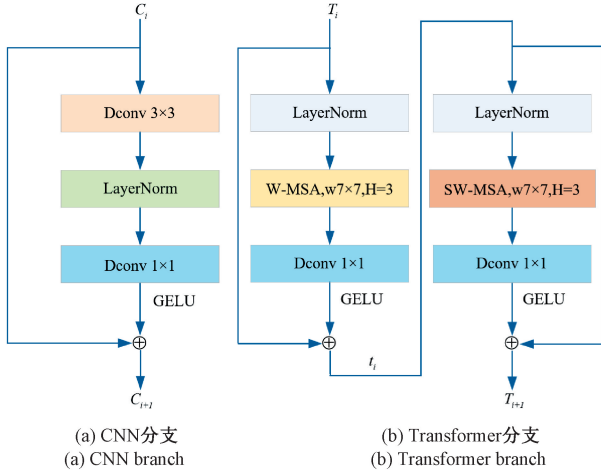


图 3 CNN 和 Transformer 分支

Fig. 3 CNN branch and Transformer branch

1.4 Transformer 分支

晶圆图像的多种特征存在类相似性,这使得全局语义信息的提取尤为重要。Transformer 模型因其独特的自注意力机制和并行计算能力,非常适合处理全局信息。其核心的自注意力机制允许模型在处理序列时直接考虑所有位置之间的关系,实现真正的全局交互,从而捕捉长距离依赖和全局特征。其多头注意力机制进一步增强了模型从多视角提取全局信息的能力。因此,本文引入了窗口多头自注意力(W-MSA)模块。W-MSA 模块源自 Swin-Transformer^[28],与传统的多头自注意力(MSA)相比,W-MSA 可有效降低计算复杂度,通过将特征图划分为固定大小的窗口并对每个窗口独立执行自注意力操作来实现。其计算复杂度如式(7)和(8)所示。

$$\Omega(MSA) = 4hwc^2 + 2(hw)^2C \quad (7)$$

$$\Omega(W-MSA) = 4hwc^2 + 2M^2hwC \quad (8)$$

式中: h 表示特征图的高度; w 表示宽度; c 表示深度; M 为每个窗口的大小。

如图 3(b) 所示,在特征处理阶段,输入张量首先经过补丁合并(patch merging)操作实现空间维度的压缩与通道维度的扩展,随后进入级联的特征增强模块。依次通过 LayerNorm 层、W-MSA 模块和具有 GeLU 激活函数的线性层。本文用 GELU 激活函数替代传统 Transformer 中的前馈神经网络,优化非线性表达方式以降低计算复

杂度,并引入 MFIB 模块进行跨模态参数化补偿,从而在保障模型表征能力的同时实现高效计算。每个模块均采用残差连接以促进信息传递,结合相对位置偏置机制,并在模块间交替应用 W-MSA 和 SW-MSA 以进一步增强特征提取能力,其过程如式(9)、(10)所示:

$$t_i = DConv1 \times 1(W-MSA(LN(T_i))) + T_i \quad (9)$$

$$T_{i+1} = DConv1 \times 1(SW-MSA(LN(t_i))) + t_i \quad (10)$$

式中: t_i 和 T_{i+1} 分别表示 W-MSA 和 SW-MSA 的输出特征; $DConv1 \times 1$ 代表使用 1×1 卷积核的深度卷积操作; LN 为 LayerNorm。最终,提取的特征被输入到 MFIB 模块中,实现多尺度融合,进一步提升模型的检测能力。

1.5 iRMB

如图 4 所示,与传统的残差结构不同,倒置残差结构在中间层扩展特征通道数,而在输入和输出层压缩通道数,从而在保持模型性能的同时显著降低计算复杂度。iRMB 采用三路协同的特征融合架构,通过差异化特征处理实现计算优化与性能提升。

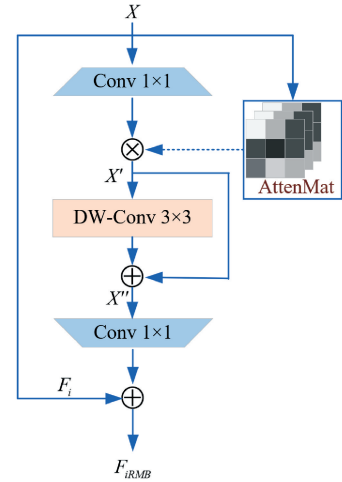


图 4 倒置残差模块

Fig. 4 The iRMB module

左侧支路直接通过残差连接传递原始特征,维持特征分布稳定性,有效缓解深层网络中的梯度消失问题,公式为:

$$F_i = X \quad (11)$$

右侧支路将输入特征映射为查询矩阵(query, Q)与键矩阵(key, K),经过 softmax 操作得到注意力矩阵 *AttenMat*,公式为:

$$AttenMat = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (12)$$

其中, d_k 表示 K 的维度,平方根是为了防止在计算过程中点积的结果过大,从而避免梯度消失问题。

中间支路首先通过一个 1×1 卷积升维调整通道数,并与注意力矩阵 *AttenMat* 加权融合,得到加权后的特征

表示 X' ;接着,该特征通过一个 3×3 深度可分离卷积提取局部空间特征,并与加权后的特征表示 X' 通过残差连接融合,得到增强后的特征表示 X'' ;最后,再通过一个 1×1 卷积降维回原始通道数,并与初始输入特征 F_i 进行连接,生成最终输出特征 F_{iRMB} 。

公式为:

$$X' = \text{Conv}_{1 \times 1}(X) \otimes \text{AttenMat} \quad (13)$$

$$X'' = X' + \text{DW} - \text{Conv}_{3 \times 3}(X') \quad (14)$$

$$F_{iRMB} = F_i + \text{Conv}_{1 \times 1}(X'') \quad (15)$$

式中: X 为原始特征; $\text{DW} - \text{Conv}_{3 \times 3}$ 表示 3×3 深度可分离卷积。

2 实验结果及分析

本文将所提出的 CTM-Net 在 WM-811K 数据集上进行验证与测试,并与现有代表性单类型晶圆缺陷分类先进方法进行了对比。

2.1 数据集预处理

WM-811K 数据集中的数据除了晶圆图和对应的缺陷标签外,还包含其他附加信息,如数据索引、批次名称、晶圆尺寸、缺陷类型、训练标签和测试标签等。总数据中,共有 172 950 张晶圆图带有缺陷标签,占比约 21.3%。各类缺陷的分布情况如图 5 所示,数据明显不平衡,例如“Edge-Ring”类型的数据量远多于其他类型,而“Donut”、“Random”、“Scratch”和“Near-Full”类型的数据量则相对较少。

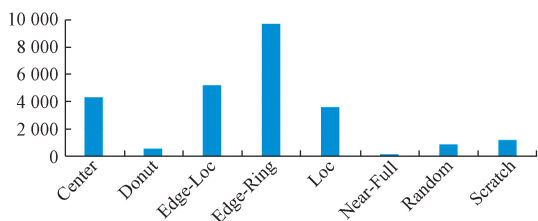


图5 晶圆缺陷种类分布

Fig. 5 Defect type distribution on wafer

针对 WM-811K 数据集存在的晶圆图像尺寸不一致和类别分布不平衡问题,本文实施了系统的数据预处理与增强方案。首先将原始单通道晶圆图像的分辨率(像素值 0 表示未知、1 表示正常、2 表示缺陷)统一调整为 224×224 ,并将单通道图像扩展为三通道 RGB 图像。考虑到缺陷类别标签的不平衡性对分类性能的影响,本文基于自编码器网络对 172 950 张带标签样本进行数据增强,通过特征空间重构生成新样本,使每种缺陷类型的样本量均稳定在 2 000 张左右,最终构建了一个包含 19 261 张晶圆图像的数据集(含正常类别),所有增强图像均保持与原始图像一致的规格化处理。该数据集的 RGB 图

像示例如图 6 所示。实验时,对该数据集按 80% 训练集(15 409 张)和 20% 测试集(3 852 张)进行划分,确保模型评估的可靠性。该预处理流程有效解决了原始数据中存在的维度差异和类别不平衡问题,为后续分类任务奠定了数据基础。

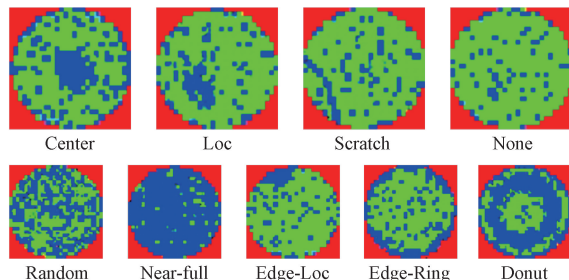


图6 三通道 RGB 图示例

Fig. 6 Example of three-channel RGB image

2.2 实验环境

本文使用 PyTorch 框架实现所提出的模型,所有实验均在配备 AMD EPYC 32-Core CPU 和 NVIDIA RTX 3090 GPU 上实现。在模型训练上,CTM-Net 采用 200 个 epoch 训练模型,学习率设置为 0.001,采用 Adam 优化器,损失函数为交叉熵损失函数。

2.3 客观指标对比实验

为了全面评估本文模型性能,将其与 5 种代表性先进方法进行了比较,这些方法包括 VAE-GAN^[29]、CNN-WDI^[9]、MRWA-Net^[30]、CWDR-Net^[31]、VAEDLM^[32],实验结果如表 1 所示,加粗字体为最优值,为 WM-811K 数据集的整体性能指标。从表 1 可以看出,本文方法在 4 个整体性能指标中都取得了最佳值,优于其他模型。相较于 2023 年提出的 CWDR-Net(准确率 98.21%, F1 分数 98.14%),CTM-Net 的准确率(99.31%)与 F1 分数(99.31%)分别提升 1.10% 和 1.17%,其召回率(99.34%)更显著优于传统模型如 CNN-WDI(95.62%)和 VAEDLM(96.16%)。尽管 VAE-GAN(2023)在精确率(96.96%)上表现尚可,但其 F1 分数(96.49%)比 CTM-Net 低 2.82%,凸显 CTM-Net 在分类均衡性上的改进。以上实验结果表明,CTM-Net 通过动态多尺度特征融合,有效克服了现有模型对复杂缺陷形态敏感度不足(如 MRWA-Net)及类别不平衡适应性弱(如 CWDR-Net)的缺陷,表明本文提出的模型在处理晶圆缺陷分类任务中具有优势。

表 2 为 CTM-Net、MRWA-Net 与 CWDR-Net 在不同缺陷类型上的检测表现。实验结果表明,CTM-Net 在所有缺陷类型上的 3 项指标均超过 99%,说明本文模型对各类典型模式均具备较强的分类能力;其中 Dount、Edge-Ring、Near-Full、Random 及 Scratch 等类别的指标几乎均

表 1 CTM-Net 与其他模型的整体指标比较

Table 1 Overall performance comparison of

CTM-Net and state-of-the-art models (%)

模型	准确率	精确率	召回率	F1 分数
VAE-GAN	95. 87	96. 96	96. 73	96. 49
CNN-WDI	96. 12	95. 90	95. 62	95. 35
MRWA-Net	99. 17	99. 20	99. 20	99. 20
CWDR-Net	98. 73	98. 78	98. 77	98. 77
VAEDLM	96. 18	96. 52	96. 16	95. 69
CTM-Net(本文)	99. 31	99. 29	99. 34	99. 31

逼近 100%,进一步验证了模型所提取的大尺度全局特征

表 2 CTM-Net、MRWA-Net 与 CWDR-Net 针对不同缺陷类型的检测结果

Table 2 Detection results of CTM-Net, MRWA-Net, and CWDR-Net on different defect types

(%)

类别	CTM-Net			MRWA-Net			CWDR-Net		
	精确率	召回率	F1 分数	精确率	召回率	F1 分数	精确率	召回率	F1 分数
Center	99. 34	99. 46	99. 49	99. 53	99. 30	99. 42	99. 76	98. 37	99. 06
Donut	99. 99	100. 00	100. 00	100. 00	100. 00	100. 00	100. 00	100. 00	100. 00
Edge-Loc	99. 08	99. 24	99. 13	98. 16	98. 16	98. 16	97. 91	97. 01	97. 46
Edge-Ring	99. 76	99. 87	99. 86	99. 53	98. 83	99. 18	99. 3	100	99. 65
Loc	99. 60	99. 49	99. 57	98. 68	97. 40	98. 04	97. 42	98. 05	97. 73
Near-Full	100. 00	99. 93	99. 96	100. 00	100. 00	100. 00	100. 00	100. 00	100. 00
Random	99. 89	99. 92	99. 92	99. 77	100. 00	99. 88	99. 31	99. 77	99. 54
Scratch	100. 00	99. 98	100. 00	100. 00	100. 00	100. 00	99. 32	98. 65	98. 99
None	99. 14	99. 10	99. 07	97. 12	99. 10	98. 10	95. 97	97. 06	96. 51

综合上述实验结果和分析,可以证明 CTM-Net 的分类性能优于现有代表性先进方法。

2.4 消融实验

CTM-Net 模型为双分支结构,本文进行了消融实验来证明各个分支及 MFIB 模块的有效性。本文将 CNN 和 Transformer 分支作为 Baseline,将多级特征交互 MFIB 作为模块。实验结果如表 3 所示,加粗字体为最优值。

表 3 消融实验结果

Table 3 Ablation study results

(%)

CNN	Transformer	MFIB	准确率	精确率	召回率	F1-Score
√	√		98. 41	98. 23	98. 271	98. 34
√		√	98. 39	98. 20	98. 49	98. 34
	√	√	98. 64	98. 23	98. 56	98. 44
√	√	√	99. 31	99. 29	99. 34	99. 31

从表 3 可以看出,当同时使用 CNN 和 Transformer 两个分支及 MFIB 模块时,模型各项指标都达到了最佳值,说明 CNN 和 Transformer 分支相互配合,能够为模型提供更强大的特征提取和分类能力,从而提升了 CTM-Net 的整体性能。当单独使用 CNN 或 Transformer 分支时,检测性能低于同时使用两个分支及 MFIB 模块的模型。这表明,CNN 和 Transformer 分支各自对模型性能都有提升作用,但是提升性能依然低于本文完备模型。

具有高度稳定的检测能力。尽管 Edge-Loc (F1 为 99. 13%) 和 None(F1 为 99. 07%) 类别的性能略低,但其召回率仍分别保持在 99. 24%和 99. 10%,表明 CTM-Net 在边缘定位缺陷及非缺陷样本上的漏检风险依旧可控。

与 MRWA-Net 和 CWDR-Net 相比,CTM-Net 在大多数类别中均取得更高的综合指标,并展现出更好的类别均衡性。尤其在 Edge-Loc、Loc、None 等类别上,CTM-Net 相较于其他模型提升幅度超过 1%。

上述结果表明,CTM-Net 在保留高分度缺陷检测优势的同时,通过自适应特征融合机制有效缓解了模型在复杂低对比度缺陷检测中的敏感性衰减问题。

当去除 MFIB 模块时,模型性能明显下降,证实了 MFIB 模块对本文模型 CTM-Net 的重要性。以上实验结果表明,对于晶圆缺陷分类,各个通道的信息都很重要,需要各个分支及模块的相互融合,获得的晶圆图像特征信息才会更全面,模型检测的效率才会更高。

为了进一步展示 CTM-Net 的分类能力,绘制了其混淆矩阵,如图 7 所示。混淆矩阵的行表示真实标签(true label),列表示预测标签(predicted label)。每个单元格中的数值表示属于某个真实类别的样本被模型预测为某个类别的数量。从混淆矩阵可以看出,本文模型在大多数类别上表现良好,尤其是在“Donut”和“Edge-Ring”类别上,分别有 391 和 404 个样本被正确分类,没有误分类情况。“Center”类别的分类效果也较好,有 469 个样本被正确分类,但有少量样本被误分为“Edge-Loc”和“none”类别。“Edge-Loc”类别的分类效果也较为理想,有 448 个样本被正确分类,但也有少量样本被误分为“Edge-Ring”和“none”类别。总体而言,模型在大多数类别上表现出较高的准确性,但在处理“Near-full”、“Random”和“Scratch”类别时存在一定的误分类问题,还需要进一步优化以提高分类性能。

2.5 主观可视化结果

如图 8 所示,通过分析图中所呈现的真实数据与预

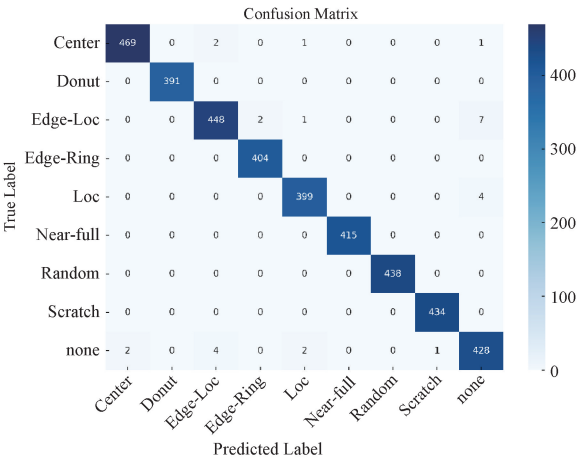


图 7 混淆矩阵
Fig. 7 Confusion matrix

测数据之间的差距,揭示了模型在不同模式下的预测性能。在“Scratch”、“Random”、“Edge-Ring”和“Near-Full”等模式中,预测结果与真实数据的分布模式展现出高度的一致性,这表明本文模型在这些特定场景下展现出了较高的预测准确性和可靠性,能够精准地捕捉并预测数据的真实分布特征。然而,在部分预测结果中,预测数据与真实数据之间仍存在一定的偏差。将“Random”预测为“Near-Full”、“Loc”预测为“Center”、“Edge-Loc”预测为“Edge-Ring”等。这种偏差可能源于模型在区分随机分布与近满分布时,对某些特征的敏感度不够,也可能未能准确识别局部特征与中心特征的区别,或者在训练过程中未能充分学习到两者之间的细微差异,从而导致了分类的错误。综合来看,本文所提模型整体分类性能已经优于现有代表性先进方法,但在处理某些特定数据模式时,其预测性能仍有待进一步提升。

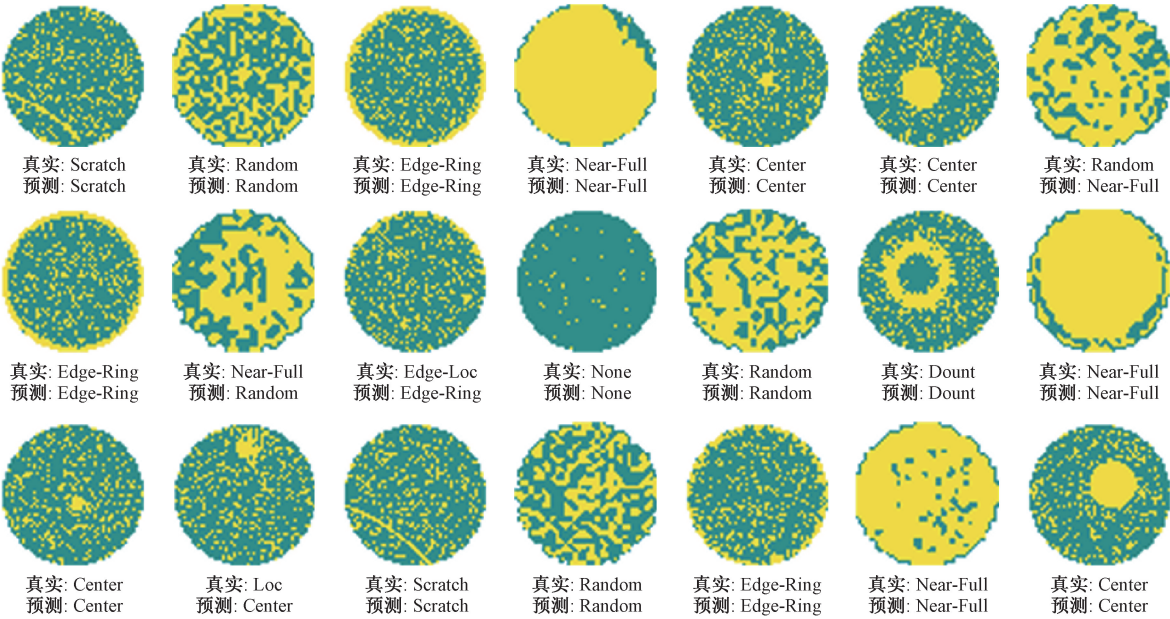


图 8 模型主观可视化结果预测
Fig. 8 Subjective visualization of model predictions

3 结 论

在半导体芯片制造中,晶圆缺陷的精准检测与分类是提升芯片良率的关键环节。针对单一类型缺陷分类中存在的尺度差异敏感性与微小缺陷易受背景干扰问题,本文提出了一种基于 CNN-Transformer 的 CTM-Net。该网络融合了 CNN 的局部细节捕捉能力与 Transformer 的全局上下文建模优势,通过多级特征交互模块实现跨尺度特征的有效整合。在 WM-811K 数据集上的实验结果证明,CTM-Net 通过深度融合局部细节感知与全局语义

理解,在 Edge-Loc、Edge-Ring、Loc 等类别上的表现均取得明显提升,显著提高了复杂晶圆缺陷模式的分类能力。综上,本研究在晶圆缺陷识别任务中构建了一种具有良好扩展性与稳健性的特征学习范式,为相关智能检测技术的进一步发展提供了可靠借鉴与理论支撑。

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] 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.
- [3] 胡志强, 吴一全. 基于机器视觉的芯片缺陷检测研究进展[J]. 仪器仪表学报, 2024, 45(7): 1-26.
- HU ZH Q, WU Y Q. Research progress in chip defect detection based on machine vision[J]. Chinese Journal of Scientific Instrument, 2024, 45(7): 1-26.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] ABD AL RAHMAN M, MOUSAVI A. A review and analysis of automatic optical inspection and quality monitoring methods in electronics industry [J]. IEEE Access, 2020, 8: 183192-183271.
- [8] 胡志强, 吴一全. 基于机器视觉的半导体晶圆缺陷检测方法综述[J]. 中国图象图形学报, 2025, 30(1): 25-50.
- HU ZH Q, WU Y Q. Survey of semiconductor wafer defect detection method based on machine vision [J]. Journal of Image and Graphics, 2025, 30(1): 25-50.
- [9] 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.
- [10] 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.
- [11] WANG R, CHEN N. Defect pattern recognition on wafers using convolutional neural networks [J]. Quality and Reliability Engineering International, 2020, 36(4): 1245-1257.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] HAN X, JIA X, JI D Y, et al. Designing robust topological features for wafer map pattern classification [C]. 34th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC). IEEE, 2023: 1-6.
- [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] 陈晓雷, 温润玉, 杨富龙, 等. 基于空频特征融合的双晶圆缺陷分类网络[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 and frequency domains feature fusion [J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(8): 56-67.
- [18] 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.
- [19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. ArXiv preprint arXiv: 1706.03762, 2017.
- [20] NAFI T I, HAQUE E, FARHAN F, et al. High accuracy Swin transformers for image-based wafer map defect detection[J]. International Journal of Engineering and Manufacturing, 2022, 12(5): 10-21.
- [21] FAN S K S, CHIU S H. A new ViT-Based augmentation framework for wafer map defect classification to enhance the resilience of semiconductor supply chains [J]. International Journal of Production Economics, 2024, 273(4): 109275.
- [22] MOHAMMAD F, RYU D. Semiconductor wafer map

defect classification with tiny vision transformers [J].
Arxiv preprint arXiv:2504.02494,2025.

- [23] ZHANG L Y, LIANG X X, ZHANG Y, et al. MLR-WM-ViT: Global high-performance classification of mixed-type wafer map defect using a multi-level relay vision transformer [J]. Expert Systems with Applications, 2025, 277: 127121.
- [24] CHANG Z H, CHEN K Y, HAN T, et al. Wafer map anomaly detection driven by contrast learning and vision transformer model [C]. 2024 Global Reliability and Prognostics and Health Management Conference (PHM-Beijing). IEEE, 2024.
- [25] WEI Y X, WANG H. Mixed-type wafer defect recognition with multi-scale information fusion transformer[J]. IEEE Transactions on Semiconductor Manufacturing, 2022, 35 (2): 341-352.
- [26] XIA M, MU X B, WU Z H. An end-to-end wafer map defect recognition model [C]. 5th International Conference on Computer Engineering and Application (ICCEA). IEEE, 2024.
- [27] ZHANG J, LI X, LI J, et al. Rethinking mobile block for efficient attention-based models[C]. 2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society, 2023: 1389-1400.
- [28] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted win dows[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022.
- [29] WANG Y, WEI Y, WANG H. A class imbalanced wafer defect classification framework based on variational autoencoder generative adversarial network [J]. Measurement Science and Technology, 2022, 34(2): 024008.
- [30] 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.
- [31] 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.
- [32] WANG S, ZHONG Z, ZHAO Y, et al. A variational autoencoder enhanced deep learning model for wafer defect imbalanced classification[J]. IEEE Transactions on Components, Packaging and Manufacturing Technology, 2021, 11(12): 2055-2060.

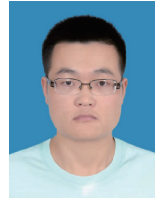
作者简介



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



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.



马平川,2024 年于中国民航大学获得学士学位,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: 2721500866@qq.com

Ma Pingchuan received his B. Sc. degree from Civil Aviation University of China in 2024. Now he is a M. Sc. candidate in Lanzhou University of Technology. His main research interest includes computer vision.



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.