DOI: 10.13382/j.jemi.B2508347

拓扑数据分析在晶圆图缺陷模式分类中的高效应用*

杜先君[1,2] 丁家俊[1] 董明月[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:晶圆图的缺陷模式分类是半导体生产制造过程中的重要环节,对提高产品良品率与生产效率有着重要意义。针对现有深度学习晶圆图缺陷模式分类方法解释性差和资源消耗高等问题,改进了一种基于拓扑数据分析(topological data analysis, TDA)的特征提取方法,其依托持久同调理论,通过构建 Alpha 复形(alpha complex)以挖掘晶圆图的拓扑结构,并将其转化为可量化的拓扑特征。实验结果表明,在基于 WM-811K 数据集构建的模拟晶圆图数据集上,采用 Alpha 复形代替原 VR 复形(vietoris-rips complex),平均复形构建时间降低了约 82%,平均内存占用降低了约 10.09%。此外,将基于 TDA 的方法与 DenseNet121、Swin Transformer 以及新兴的 ConvNeXt 模型进行了对比,在特征提取方面,t-SNE 可视化结果显示基于 TDA 方法提取的特征向量取得了最佳的聚类效果,相比于次优的 ConvNeXt,轮廓系数提升了 17.24%,提取时间降低了约 75%,而内存峰值降低了约95%;在分类性能方面,结合支持向量机(SVM)分类器的实验表明,基于 TDA 的模型整体分类准确率高达 0.992,优于 DenseNet(0.989 3)和 Swin Transformer(0.982 0)。

关键词: 晶圆图缺陷模式分类;拓扑数据分析;持久同调;Alpha 复形;支持向量机;半导体制造
中图分类号: TP391.4;TN407 **文献标识码**: A **国家标准学科分类代码**: 510.4050

Efficient application of topological data analysis in wafer map defect pattern classification

Du Xianjun[1,2] Ding Jiajun[1] Dong Mingyue[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: Wafer map defect pattern classification is a critical step in semiconductor manufacturing, significantly impacting product yield and production efficiency. To address the limitations of existing deep learning-based wafer map defect pattern classification methods, such as poor interpretability and high computational resource consumption, this study proposes an improved feature extraction method based on topological data analysis (TDA). By leveraging persistent homology theory, the method constructs Alpha complexes to characterize topological structures in wafer maps and quantifies them into discriminative features. Experimental results on a synthetic wafer map dataset, generated by emulating the geometric distribution characteristics of the WM-811K dataset, demonstrate that replacing the conventional vietoris-rips (VR) complex with the Alpha complex reduces the average complex construction time by approximately 82% and decreases memory usage by 10.09%. Compared to state-of-the-art models including DenseNet121, Swin Transformer, and ConvNeXt, the TDA-based method achieves superior clustering performance, as evidenced by t-SNE visualizations, with a 17.24% improvement in Silhouette Coefficient over the suboptimal ConvNeXt model, along with a 75% reduction in feature extraction time and a 95% reduction in peak memory consumption. When integrated with a support vector machine (SVM) classifier, the TDA-based framework attains an overall classification accuracy of 0.992, outperforming DenseNet (0.989 3) and Swin Transformer (0.982 0).

Keywords: wafer map defect pattern classification; topological data analysis (TDA); persistent homology; Alpha complex; support vector machine (SVM); semiconductor manufacturing

收稿日期: 2025-04-29 Received Date: 2025-04-29
基金项目:国家自然科学基金(62241307)、兰州市科技支计划项目(2024-3-47)资助

0 引　言

近年来,智能手机、智能机器人、新能源汽车等下游产业市场的规模持续增长,驱动半导体行业飞速发展,晶圆的生产制造呈现出大规模化和工艺尺寸微缩化的同时,多工序叠加效应同样导致缺陷产生概率呈指数级上升。遵循电子系统检测的“十倍成本法则”,芯片级未检出的缺陷在封装后环节将造成10倍级经济损失,这使得晶圆级缺陷的早期检测成为质量管控的关键环节。在此背景下,基于测试结果生成的晶圆图(Wafer Map)通过可视化呈现缺陷空间分布特征,不仅能够直观地反映出晶圆的整体质量状况,其揭示的缺陷聚集模式(例如特定区域的频繁失效标记)能帮助工程师快速定位问题所在,从而指导针对性的改进措施,最终提升产品良率与经济效益。可以说,高效可靠的晶圆图缺陷模式分类技术已成为现代半导体制造流程中保障系统可靠性、提升良率和降低成本不可或缺的一环。

在神经网络兴起之前,晶圆图缺陷模式分类主要依赖传统机器学习方法,这类方法通过手工设计特征(如聚类特征、密度/几何/Radon特征)结合分类器(如支持向量机(support vector machine,SVM))、软投票集成)实现分类,在数据集WM-811K[1]上准确率可达94%以上[2-3]。但其性能高度依赖特征工程经验,且难以捕捉复杂缺陷的深层空间关联。部分学者尝试引入图论方法进行空间噪声过滤,如Ezzat等[4]提出的邻接聚类算法,虽在高复杂度缺陷处理中表现优异,但仍受限于小样本验证与参数敏感性等问题,未能实现工业级应用。

随着2012年深度学习方法在ImageNet竞赛中取得突破,以卷积神经网络(convolutional neural network,CNN)为代表的端到端模型开始广泛应用于晶圆图缺陷模式分类[5-7]。Chen等[8]提出多源双通道CNN特征融合框架,结合ECOC-SVM分类器,在WM-811K上实现了96.4%缺陷识别准确率;Nafi等[9]则验证了多种Transformer模型在晶圆图缺陷模式分类上的有效性,为该领域提供了新选择。此外,一些混合方法也展现出潜力,Kang等[10]的动态加权融合框架,通过动态加权融合手工特征分类器与卷积神经网络的优势,提升晶圆图缺陷模式分类的准确率。

然而,深度学习模型虽然表现出强大的特征学习能力,但在晶圆图缺陷模式分类中仍存在几个明显的不足。首先,CNN作为黑箱模型,其多层卷积与池化操作缺乏可解释性,导致缺陷判定依据难以追溯,从而阻碍缺陷成因分析与工艺优化;其次,深层网络提取的特征维度高且混杂大量冗余信息,不仅增加计算复杂度,还阻碍模型对关键缺陷特征的针对性学习,导致模型泛化性下降;最后,深层CNN模型复杂度较高,训练和推理过程需要大量计算资源,不仅导致检测速度慢,难以满足实时性要求高的检测场景,而且模型体积庞大,占用大量内存,不利于在资源受限的嵌入式设备或本地环境中部署。

在此背景下,计算拓扑学(computational topology)的发展为特征表示提供了新的理论工具。以持久同调(persistent homology)[11]为核心的计算框架,通过多尺度拓扑特征量化方法,成功建立了从复杂数据中提取拓扑不变量的数学基础。随着如GUDHI、Ripser等开源软件包的诞生,简化了将拓扑特征集成到深度学习模型中的过程,使得拓扑特征与深度学习的融合成为可能。与CNN相比,拓扑数据分析(topological data analysis, TDA)能够提取数据内在的拓扑结构信息,提供更具可解释性的缺陷特征,同时提取的特征维度较低,模型更加轻量化,无需额外的高性能硬件支持,易于在本地进行部署。2023年,Ko等[12]率先将TDA应用于晶圆图缺陷模式分类,并取得了显著的性能提升。尽管如此,其研究在复形构建、数据集的真实性以及模型评估的全面性方面仍有进一步优化的空间。本文在此基础上,提出一系列创新性的改进,以充分挖掘TDA在晶圆图缺陷分析中的潜力。

1)复形构建优化,将原方法的VR复形(vietoris-rips complex)替换为Alpha复形(Alpha complex),得到了更高的计算效率与更精简的拓扑结构,显著降低了VR复形固有的高维噪声干扰,实现5.7倍的构建速度提升与10.09%的内存压缩。

2)数据集重构与拓扑特征分析,以当前规模最大的晶圆图数据集WM-811K为参照,生成了基于真实缺陷分布特征的像素化模拟数据集,相比原方法采用的点云数据集更加真实。在此基础上,运用TDA的方法,对5种典型的晶圆图缺陷类型展开了深入的拓扑性质分析。

3)跨架构验证,最后,在实验对比中,除了原方法采用的传统CNN模型外,还引入了Swin Transformer模型以及新兴的ConvNeXt模型,以更全面地评估所提出方法的性能。

1 拓扑数据分析(TDA)

1.1 基础理论

单纯形(simplex)和单纯复形(simplicial complex)(图1)是拓扑学与几何学领域的核心基础概念[13],它们是实现原始数据向拓扑分析适用形式转化的关键工具,在数据的结构化表征与拓扑特征挖掘中发挥着不可或缺的作用。

单纯形的定义是在n-维欧几里得空间$\boldsymbol{R}^n$中,$k+1$个几何独立的点$v_0,v_1,\cdots,v_k(k\leqslant n)$的凸包称为一个$k$-

单纯形(k-simplex)。

单纯复形是欧几里得空间 R^n 中的有限单纯形集合,它需满足以下两个条件。

1)如果 $\delta \in K$ 是一个 k- 单纯形,那么 δ 的所有面(包括它本身)也属于 K。

2)对于任意两个单纯形 $\delta_1, \delta_2 \in K$,它们的交集 $\delta_1 \cap \delta_2$ 要么是空集,要么是 δ_1 和 δ_2 的一个公共面。

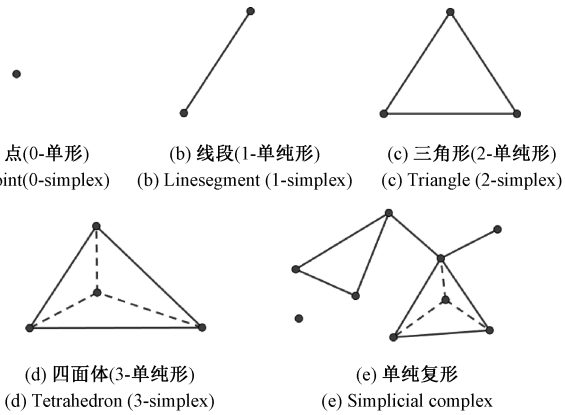


图 1　单纯形与单纯复形

Fig. 1　Simplex and simplicial complex

在拓扑数据分析中,提取数据的拓扑特征需要一种抽象方法,以构建能够反映数据内在结构的单纯复形。这种抽象的方法在持久同调理论[11]中被称为滤(filtration),其本质上是一个随参数变化的复形序列,常见的复形类型包括 VR 复形和 Alpha 复形。当参数 α(Alpha 复形)或者参数 r(VR 复形)逐渐增大时,单纯形会经历动态生成或消亡,这导致了拓扑结构中连通分支的合并以及环或孔的生成与消亡。通过追踪这些拓扑特征在不同尺度下的变化来捕捉数据拓扑性质,这便是持久同调的核心思想。为了可视化这些信息 Edelsbrunner 等[11]和 Cohen-steiner 等[14]提出将每个拓扑特征的诞生时间 b(birth)和消亡时间 d(death)作为坐标点绘制二维散点图,就得到了持久图(persistence diagram,PD)。

PD 以点集 (b,d) 形式呈现拓扑特征及其在特定滤下的生命周期,b 为生成时刻,d 为消亡时刻。点到对角线 $b=d$ 的距离(即持久性 $d-b$)是衡量特征稳定性的关键指标。通常认为,低持久性点代表噪声,而高持久性点则反映了数据集固有的拓扑结构。PD 中零维同调群(H_0)的点表示连通分支(connected components),一维同调群(H_1)的点代表环(loops)或孔(holes)。

为了将 PD 上的拓扑特征输入到深度学习模型中,Adams 等[15]提出将 PD 中的每个点映射到二维网格上,并使用高斯核函数对每个点的持久性进行加权,生成一种标准化的图像表示——持久图像(persistence image,PI)。PI 上每个像素的强度量化了 PD 中映射到该像素对应邻域内的点的集中程度,高强度区域标识了数据中更为显著或稳定的拓扑结构,该表示方法不仅实现了拓扑特征的结构化转换,还通过归一化处理使不同尺度的 PD 具有可比性,为后续拓扑特征分析提供了标准化输入。

1.2　VR 复形与 Alpha 复形的定义与构建过程

VR 复形是基于点集和距离参数构建的抽象单纯复形,其具体定义如下:给定点集 $X \in \boldsymbol{R}^n$ 和参数 $r \geqslant 0$,VR 复形 $VR(X,r)$ 是一个抽象单纯复形,包含元素如下。

1)顶点集,所有点 $x \in X$。

2)单纯形,对于 X 的一个非空子集 $\varphi = \{x_0, x_1, \cdots, x_k\}$ $(k \geqslant 0)$,若其任意两点满足 $d(x_i, x_j) \leqslant r$,则 φ 对应的单纯形属于 VR(X,r)。

Alpha 复形在 GUDHI[16]的定义是给定点集 $Y \in \boldsymbol{R}^n$ 和参数 $\alpha \geqslant 0$,Alpha 复形 $Alpha(Y,\alpha)$ 是德劳内三角剖分(delaunay triangulation)的一个子复形(subcomplex),它只包含那些外接圆(或球,在三维情况下)半径的平方不超过 α 的单纯形。Alpha 复形包含以下元素。

1)顶点集,所有点 $y \in Y$。

2)边,对于任何两个点 $y_i, y_j \in Y$,如果存在一个空心球 B,其半径平方 $r^2 \leqslant \alpha$,并且 B 恰好包含 y_i 和 y_j,同时 B 的内部不包含 Y 中的任何其他点,则 y_i 和 y_j 之间有一条边。

3) k - 单纯形($k \geqslant 2$),对于任意一组 $k+1$ 个点 $y_0, y_1, \cdots, y_k \in Y$,如果存在一个空心球 B,其半径平方 $r^2 \leqslant \alpha$,它可以包围这 $k+1$ 个点,并且 B 的内部不包含 Y 中的任何其他点,则这些点形成一个 k - 维单纯形。

以点集 $P = \{[4,4],[3,5],[5,5],[4,7],[6,2],[8,2],[8,7],[10,4],[10,8],[12,6],[6,8]\}$ 为例,图 2 所示为当 $\alpha = 0.5, 1, 1.5, 2, 2.5, 3$ 时的 Alpha 复形与当 $r = 1, 2, 3, 3.5, 4, 4.5$ 时的 VR 复形。

图 2 中,虚线圆表示半径为 $r/2$,用来判定点间距离;粉色虚线边框的三角形表示德劳内三角剖分的结果;蓝色表示 0-单纯形;橙红色表示 1-单纯形;绿色填充表示 2-单纯形;紫色填充表示 3-单纯形;红色填充表示 4-单纯形。受限于二维空间,高维单纯形(如 3-单纯形的四面体、4-单纯形的五胞体)通过顶点连接与面填充示意,其几何形态需通过组合拓扑性质理解。

从图 2 可以看出,VR 复形的构建存在维度膨胀问题:随着 r 值的增大,高维单纯形的过度生成易引入噪声干扰,导致拓扑特征失真。相较之下,Alpha 复形基于德劳内三角剖分,能更精确拟合数据的底层几何结构。其核心优势在于受数据维度约束,仅需处理顶点、边及三角形等低维结构,这使得单纯形数量显著少于 VR 复形,计算效率更高。

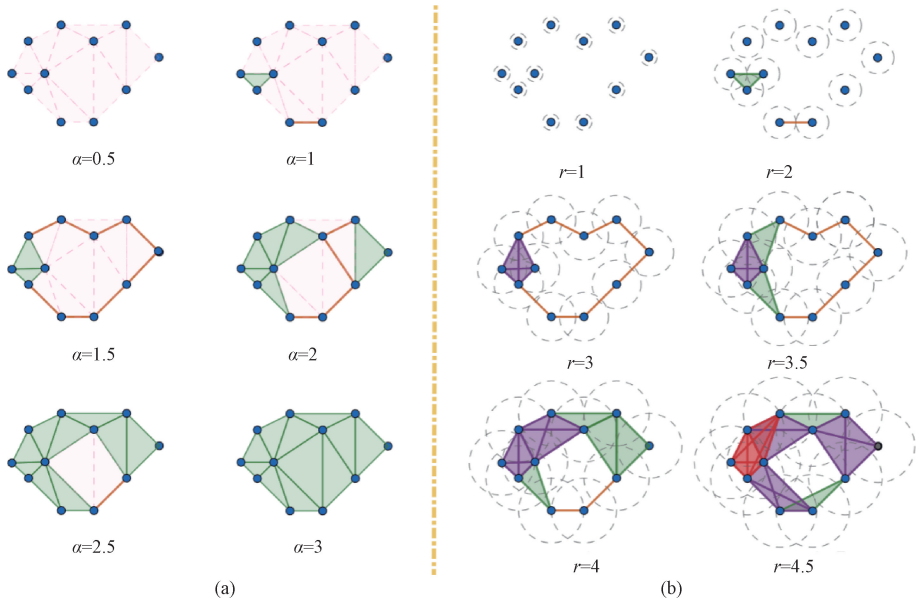


图2　Alpha复形的构建过程(a)与VR复形的构建过程(b)

Fig. 2　Construction process of the Alpha complex (a) and the VR complex (b)

1.3　VR复形与Alpha复形的比较

为了深入探讨VR复形与Alpha复形在处理不同特征数据时的性能差异,进行了对比实验。实验使用了自主制作的晶圆图模拟数据集,涵盖聚集、密集、稀疏、环形和划痕等五种代表性晶圆图缺陷类型,每种类型各300张,尺寸均为510×510。

在统一计算环境下,本文对每张晶圆图像分别构建VR复形和Alpha复形,记录其构建时间、内存占用、单纯形数量以及PD上的特征点数量,并按缺陷类别取均值。本实验在提取特征点后对坐标进行了归一化,VR复形的最大r值设为30,最大单纯形维度设为2,实验结果如表1所示,其中单纯形数量四舍五入取整,其他数据保留两位小数。

表1　不同缺陷类型下VR复形与Alpha复形构建性能对比

Table 1　Performance comparison of VR and Alpha complex construction across defect types

类别	构建时间/ms		占用内存/MB		单纯形数量		特征点数量	
	VR	Alpha	VR	Alpha	VR	Alpha	VR	Alpha
聚集	9.66	2.8	78.46	69.87	655 015	884	225.50	259.69
密集	12.31	2.03	87.29	78.1	894 612	973	207.55	299.12
稀疏	0.17	0.35	78.91	78.91	8 526	169	37.23	53.39
环	24.58	2.49	109.15	82.61	1 685 204	1 136	229.50	269.41
划痕	0.61	0.38	80.64	80.64	24 428	268	56.19	78
平均值	9.42	1.66	86.89	78.04	653 557	686	151.19	191.92

实验结果表明,相比Ko等[12]采用的VR复形,本文改进的Alpha复形在计算效率和特征提取方面均有显著优势。

1)计算效率显著提升。Alpha复形在聚集、密集和环类型中展现出显著的时间效率优势,平均构建时间相比VR复形降低了约82%,这表明Alpha复形更适合实时或大规模数据处理场景。

2)内存占用大幅降低。Alpha复形的内存消耗普遍低于VR复形,平均内存较VR复形降低了10.09%。这种差异源于Alpha基于德劳内三角剖分的几何约束,减少了冗余单形的生成,从而优化了内存利用率,尤其在密集或环形缺陷中更为显著。

3)拓扑结构更精简。Alpha复形的单纯形数量显著低于VR复形,证明VR复形在高维拓展过程中产生大量冗余结构。而Alpha复形通过几何约束剔除无效连接,避免了因数据维度膨胀而导致的额外计算与内存负担。以Ring类型缺陷为例,VR复形生成1 685 204个单纯形,而Alpha复形仅生成1 136个,二者相差约1 483倍。

4)特征提取能力更强大。VR复形的平均单纯形数量大约是Alpha复形的953倍,但是从这些拓扑结构中

提取的特征点却少于 Alpha 复形。这说明 Alpha 复形具有更好的几何形状的敏感性,能够更精确地捕捉缺陷的持久同调特征。

综上所述,Alpha 复形通过几何约束,在计算效率和特征质量之间取得了更好的平衡。相比 VR 复形,这些特性使其在处理大规模低维数据时更具优势,既能保持几何拓扑特征的完整性,又降低了冗余计算开销。

2　用 TDA 处理晶圆图并分析其拓扑特征

基于前述的拓扑数据分析理论基础,本章将详细阐述提取与处理晶圆图拓扑特征的具体步骤,并深入探讨晶圆图的拓扑学特征。通过对晶圆图进行拓扑分析,旨在揭示不同类型晶圆图缺陷所蕴含的内在拓扑结构,并从中提取具有区分性的拓扑特征,为后续的缺陷识别与分类提供理论依据。

2.1　晶圆图的处理流程

本文参考 Ko 等[12]提出的方法,将晶圆图拓扑特征提取与处理的具体步骤分为 3 步。

1)特征图构建。首先将原始晶圆图(RGB 格式)转换为 HSV 色彩空间以增强颜色与亮度区分度。再通过预设 HSV 阈值(黄色缺陷区域)生成二值掩码,提取目标像素坐标。最后利用 NumPy 库定位目标像素位置后,采用线性映射将其坐标归一化至[0,1]区间,形成标准化点云数据集。

2)持久同调计算并与 PD 生成。利用 GUDHI 库的 AlphaComplex 包来构建点云的几何拓扑结构,并通过 simplex tree 追踪拓扑特征随参数变化的演化过程。在这个过程中,计算每个拓扑特征的持久性,并以此来生成 PD。

3)PD 到 PI 的转换与向量化。计算 PD 上每个拓扑特征的零维和一维的持久性区间,并剔除无限值的零维区间,通过高斯核函数进行空间平滑分别生成零维与一维 PI,将二者展平并拼接为综合特征向量。

上述步骤的流程如图 3 所示,受图片大小限制,图 3 中只分别展平并拼接了部分的向量。

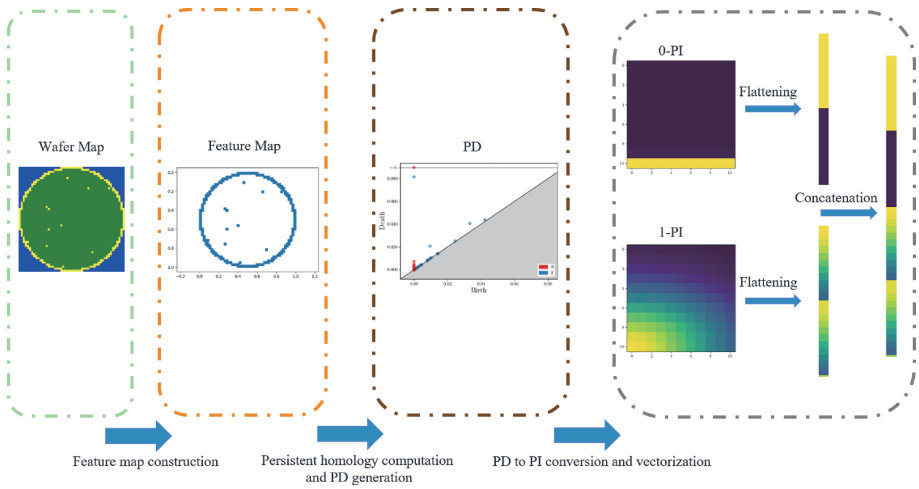


图 3　晶圆图拓扑特征提取与处理示意图

Fig. 3　Schematic diagram of topological feature extraction and processing for wafer maps

2.2　晶圆图拓扑特征分析

本文选取的模拟晶圆图数据集,包含聚集(Cluster)、密集(Dense)、稀疏(Sparse)、环(Ring)以及划痕(Scratch)5 种缺陷类型。

晶圆图的缺陷按产生机制和分布特征可以分为 3 类,随机缺陷、系统缺陷和组合缺陷。随机缺陷是无规律、随时间和空间变化而产生的,形成原因复杂多样。系统缺陷则是可重复出现的,通常表现出明显的聚集现象,可以通过缺陷分布和形貌类型识别工艺或仪器中的异常,例如光刻过程中掩膜位置的错位或蚀刻过程中的过度蚀刻。组合缺陷则是由随机缺陷和系统缺陷共同构成,表现出更复杂的特征。在本文的模拟数据集中,稀疏缺陷对应随机缺陷,密集缺陷对应系统缺陷,其余则被定义为组合缺陷。为了更清晰地分析晶圆图缺陷的拓扑特征,分析过程剔除了组合缺陷中包含的随机缺陷,从而聚焦于其系统性缺陷部分的拓扑结构。图 4 和 5 所示分别为 5 种缺陷类型的晶圆图空间分布及其基于 Alpha 复形的持久图。

由图 4 可见,每种缺陷都有明显的拓扑特征:

1)聚集缺陷表现为数据点在局部空间内的非均匀高密度分布,形成一个或多个连通分支。若聚集区域内部存在空洞,则可能产生一维同调群(H_1)。在 PD 上,该类型缺陷表现为大量的短寿命零维特征(对应 H_0 同调群),反映了局部高密度的连通分支。由于数据点的聚集,可

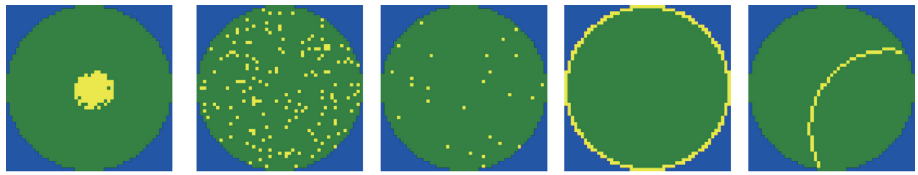


图 4　滤除随机缺陷点后的 5 种典型晶圆图缺陷类型

Fig. 4　Five typical wafer map defect types after filtering out random defect points

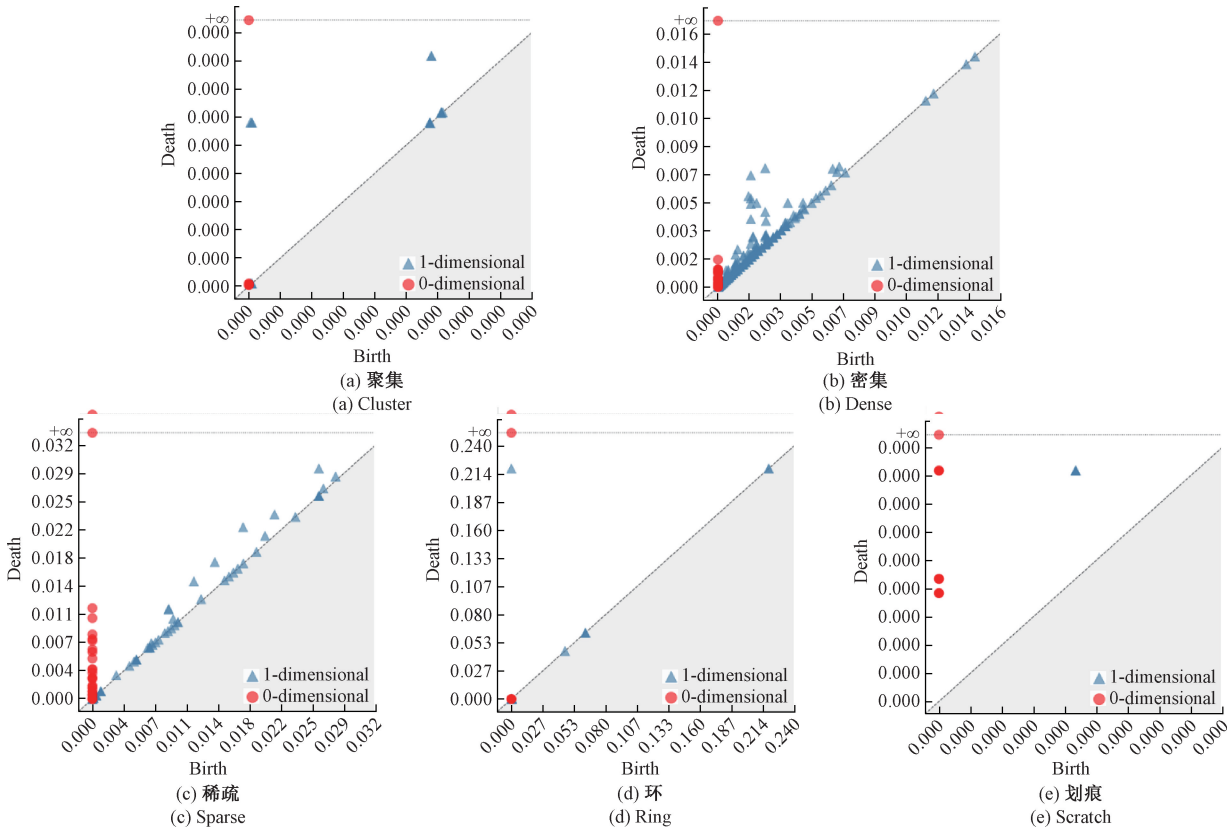


图 5　5 种典型晶圆图缺陷的持久图

Fig. 5　Persistence diagrams of five typical wafer map defects

能存在局部孔洞引起的一维同调特征，表现为 PD 上出生尺度较小的一维点。

2）密集缺陷的点密度低于聚集缺陷，点间距也相对较大。因此，需要更大的过滤尺度（filtration scale）才能形成连通分支，这在 PD 上体现为零维特征点的寿命普遍较长，表明连通分支的形成需要更大的尺度。由于数据点分散，可能存在更多由噪声或局部结构引起的一维同调特征，表现为 PD 上出生尺度较小的一维点。

3）稀疏区域由大量孤立点构成，点密度显著低于密集缺陷。与密集缺陷类似，形成连通分支需要更大的过滤尺度，因此，零维特征点的寿命较长。由于数据点总数较少，一维特征点的数量也相应减少。同时，由于点密度低，可能存在由噪声或偶然结构引起的长寿命一维点，其出生尺度相对较大。

4）环形缺陷最显著的特征是一维洞。在 PD 上，该类型缺陷表现为一个具有显著持久性的一维特征点，其生死时间差远大于其他特征点，表明该环形结构具有拓扑稳定性。其余特征点的寿命较短，接近对角线，可视为由噪声或局部扰动引起的拓扑噪声。

5）划痕可以被视为一个细长的、一维的结构。它可能表现为单个或多个连通分支，弧形划痕可能在特定尺度下形成一维洞。在 PD 上，零维特征点的数量较少，寿命较短，反映了划痕简单的连通特性。如果划痕形成闭合或近似闭合的路径，则可能存在一个出生尺度极小的一维特征点，指示环形结构。

分析表明，不同类型缺陷在 PD 上展现出独特的拓

扑特征。零维特征点的时空分布与点密度密切相关,而一维特征点的持久性暗示着真实环形结构与孔洞。这一发现为基于拓扑特征的晶圆图缺陷模式分类提供了理论依据,也体现了 TDA 的方法在可解释性上具有显著优势。

3　实验方案

3.1　数据集

WM-811K 数据集是专门用于晶圆图缺陷检测和分类任务的一个大型公开数据集。其凭借规模大、多样性高而被广泛应用于晶圆图缺陷模式分类领域[5]。但其缺陷类别定义粗略、数据陈旧以及部分单缺陷晶圆图中多种缺陷共存等问题,对利用 TDA 的方法提取晶圆图拓扑特征构成挑战。特别是,数据集中不同尺寸和比例的晶圆图使得固定坐标系下的拓扑特征提取变得复杂。例如,小尺寸晶圆图(如 20×20)中的密集型缺陷可能与大尺寸图(如 160×160)的聚集型缺陷具有相似拓扑特征。为了解决这一挑战并提高模型的泛化能力,采用模拟数据集成为一种可行的新方案[17-19]。借助模拟数据集,研究人员能够在统一标准下创建具备可控变量的数据集,进而更高效地开展缺陷分析与模型训练工作。

本文参考了 WM-811K 数据集中缺陷模式的特点,开发了一套晶圆图模拟数据集生成模型。其包含 5 种缺陷,分别是聚集、密集、稀疏、环以及划痕各 300 张,尺寸均为 510×510,图 6 所示为 5 种缺陷的示例,这些模拟缺陷的形态特征与真实晶圆图的缺陷模式高度吻合。

1)稀疏(Sparse)利用 numpy 库中的 random 函数随机选取缺陷点的数量。在给定的宽高范围内随机生成各缺陷点坐标(x, y)并保证这些点处于圆形区域内,最后在图上绘制符合条件的点。

$n_{\text{sparse}} \sim \text{Uniform}(10,60)$

2)密集(Dense)生成方式与稀疏缺陷相似,通过增加缺陷点数量来提高密度。

$n_{\text{dense}} \sim \text{Uniform}(120,240)$

3)聚集(Cluster):由聚集部分和随机部分构成。聚集部分使用 random 函数以正态分布随机采样中心点,结合随机半径和缺陷数量通过极坐标系生成缺陷簇。随机部分仿照稀疏缺陷生成缺陷点。

$n_{\text{random}} \sim \text{Uniform}(10,60)$

$n_{\text{cluster}} \sim \text{Uniform}(250,350)$

$r_{\text{cluster}} \sim \text{Uniform}(5,10)$

4)环(Ring)由环状部分与随机部分构成。环状部分在半径为 R 的晶圆边缘用 linspace 函数生成均匀分布的角度,再用黄色点绘制缺陷区域。随机部分同样仿照稀疏缺陷生成缺陷点。

$n_{\text{random}} \sim \text{Uniform}(10,60)$

$\theta_{\text{ring}} \sim \text{Uniform}(0,2\pi)$

$r = R$

5)划痕(Scratch)由划痕部分与随机部分构成。随机部分仿照稀疏缺陷生成缺陷点。划痕部分分为线性划痕与弧形划痕。线性划痕使用 random 函数在半径为 R 的晶圆区域内随机选择起点、方向和长度,使用 Bresenham 算法绘制直线。弧形划痕同样在晶圆区域内随机生成弧的中心点、半径、起始角度和角度跨度,然后,利用 linspace 函数采样弧上的点,绘制黄色弧形划痕。单条与多条(2~3 条)划痕的比例为 2∶1。

$n_{\text{random}} \sim \text{Uniform}(10,60)$

$\theta_{\text{line\_scratch}} \sim \text{Uniform}(0,2\pi)$

$d_{\text{line\_scratch}} \sim \text{Uniform}(R/3,2R)$

$r_{\text{arc\_scratch}} \sim \text{Uniform}(2R/3,3R/2)$

$start\_\theta_{\text{arc\_scratch}} \sim \text{Uniform}(0,2\pi)$

$span\_\theta_{\text{arc\_scratch}} \sim \text{Uniform}(\pi/3,\pi/2)$

3.2　特征提取方法对比与可视化验证

为了证明 TDA 的方法在晶圆图缺陷提取上具备独特优势,采用模拟数据集进行实验。首先将所提出的 TDA 方法与经典模型的特征提取部分进行对比。

1)基于 TDA 的方法。如第 2 节所述,先从晶圆图中提取缺陷点数据完成特征图构建;然后利用 Alpha 复形计算持久同调生成 PD;最后将 PD 转化为 PI 并向量化,PI 的分辨率设置为 11×11,零维特征带宽取 0.002,一维特征带宽取 0.001,权重函数选用双峰高斯混合模型权重函数,最后拼接不同维度特征向量再标准化即可得到包含了拓扑信息的 121 维特征向量。

2)基于 CNN 的方法。CNN 作为深度学习模型,在图像处理和计算机视觉任务中应用广泛。像 GoogLeNet[20] 用不同尺寸卷积核提取多尺度特征,ResNet[21] 引入残差连接,DenseNet[22] 采用密集连接机制,均展现出强大的特征提取能力。本文选用 121 层的 DenseNet121 进行对照实验,其密集连接结构可有效保留多尺度特征,在保持计算效率的同时具备较强的局部特征表达能力。

3)基于 Transformer 的方法。Transformer 采用自注意力机制学习全局上下文信息,克服了 CNN 局部感受野的限制。经典模型如 Vision Transformer[23] 将图像划分为序列化图像块并利用自注意力机制建模,DeiT(data-efficient image transformers)[24] 引入蒸馏机制,Swin Transformer[25] 采用分层结构和移动窗口自注意力。本文选用 Swin-T 结构的 Swin Transformer 进行对照实验,其分层窗口注意力机制可捕捉全局上下文信息,有效缓解了传统 CNN 的局部感受野限制。

4)融合了 CNN 和 Transformer 优势的方法。

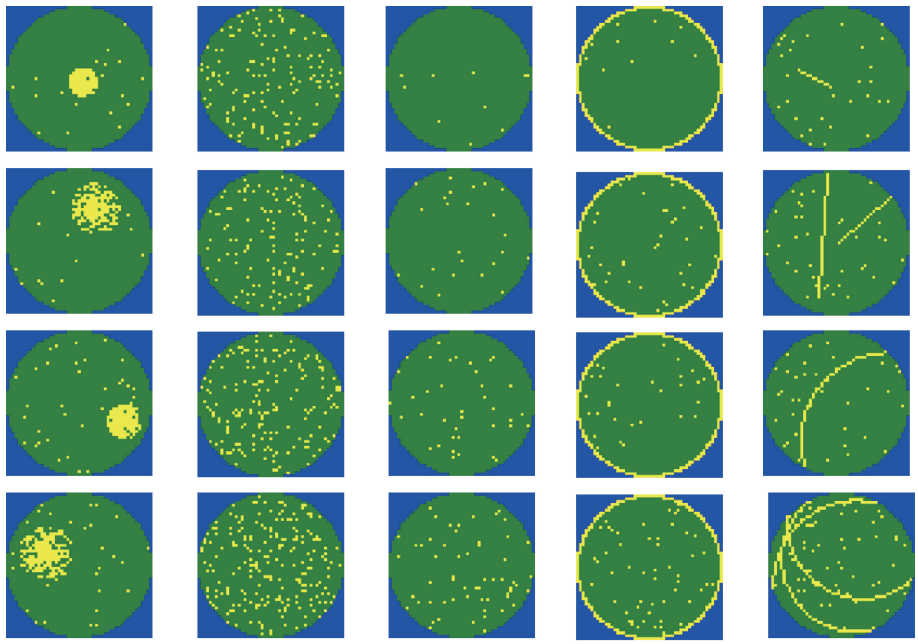


图 6　5 种晶圆图缺陷类型

Fig. 6　Five types of wafer map defects

ConvNeXt[26] 是 Meta AI 于 2022 年提出的卷积神经网络架构。它通过逐步“卷积化”Transformer 设计，并借鉴 ResNet 训练技巧，构建了纯卷积模型，在图像分类等任务中性能与 Transformer 相当甚至更优。本文选用 ConvNeXt-Base 的权重文件作为基准模型，在网络末端进行特征向量提取。

为了提高深度学习模型的泛化能力、降低计算资源需求，DenseNet，Swin Transformer 以及 ConvNeXt 采用了在 ImageNet 数据集上预训练的 V1 版本权重文件。

本文的所有实验均在配备 NVIDIA GeForce RTX 3060 Laptop GPU（用于 DenseNet121、Swin Transformer 和 ConvNeXt）和 Intel 12th Gen Core i7-12700H CPU（用于 TDA 方法）的平台上进行，实验流程如图 7 所示。对 4 种方法提取的特征向量进行了 t-SNE 降维和可视化，困惑度选择默认值 30，结果如图 8 所示。特征提取及 t-SNE 聚类性能指标对比如表 2 所示，本文实验选的聚类指标如下：

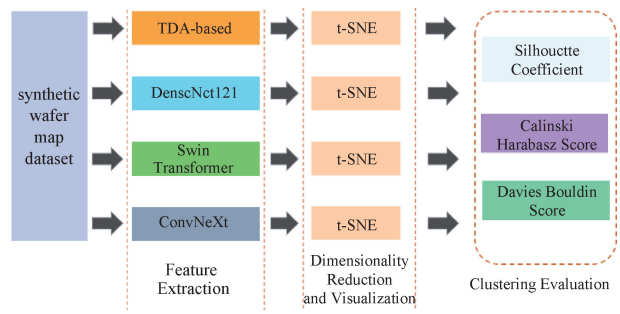


图 7　实验流程

Fig. 7　Experimental flowchart

表 2　不同方法的特征提取性能与 t-SNE 聚类效果评估指标对比

Table 2　Comparative evaluation of feature extraction performance and t-SNE clustering metrics across different methods

	特征向量维度	特征提取耗时/s	内存峰值/MB	Silhouette Coefficient	Calinski Harabasz Score	Davies Bouldin Score
TDA-based	**242**	**5.09**	**177.40**	**0.68**	**5 221.29**	**0.46**
DenseNet121	1 024	11.32	4 529.90	0.50	4 700.44	1.11
Swin Transformer	768	21.43	3 740.76	0.43	2 348.94	1.92
ConvNeXt	768	20.62	3 761.57	0.58	3 828.29	0.67

注：加粗表示最优值

1）轮廓系数（Silhouette Coefficient），该指标结合了内聚度和分离度，用于评估每个样本与其所在簇的紧密程度以及与其他簇的分离程度。其值范围为[-1，1]，值越接近 1 表示聚类效果越好。

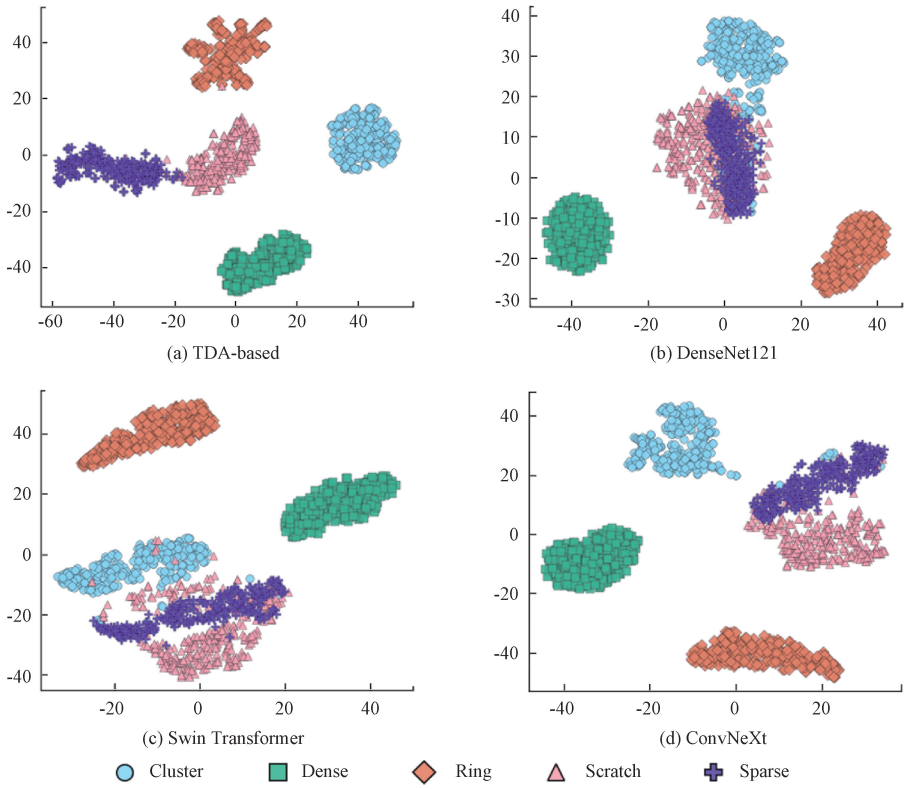


图 8　不同方法提取的晶圆图缺陷特征向量 t-SNE 可视化对比

Fig. 8　t-SNE visualization of feature vector across wafer map defect types using different methods

2) Calinski Harabasz 指数,该指数定义为类间离散度与类内离散度的比值。类间离散度越大、类内离散度越小,指数值越大,说明聚类效果越好。

3) Davies Bouldin 指数,该指数衡量每个簇与其他所有簇之间的平均相似度,值越小表示聚类效果越好。

实验结果表明,与其他 3 种模型相比,基于 TDA 的方法仅靠低维的特征向量就实现了最佳的聚类性能,在 3 种聚类指标上均排名第一。此外,TDA 方法在实验中显著降低了计算耗时和内存占用,相比于新兴的 ConvNeXt 模型,基于 TDA 的方法特征提取时间降低了约 75%,内存峰值仅为其 4.7%。这主要归因于 TDA 的无监督特征提取过程,它直接从输入数据计算拓扑特征,无需深度学习模型所需的训练过程(即使是预训练模型也需要加载权重和适配计算图),并避免了卷积和池化等复杂操作。这种仅依赖 CPU 的轻量级计算特性使得 TDA 更适合嵌入式设备或实时分析应用。尽管当前实验基于特定晶圆图数据集,但 TDA 方法的核心优势(拓扑敏感性和计算效率)表明其在不同工业检测场景中具有推广潜力。

3.3　模型性能指标

本研究采用 SVM 作为统一分类器,系统评估 TDA、DenseNet121、Swin Transformer 和 ConvNeXt 四类特征提取方法的分类性能。基于线性核 SVM(参数默认设置),通过五折交叉验证((5-fold cross-validation))获取准确率(Accuracy)、召回率(Recall)、F1 分数及混淆矩阵(图 9)。

上述指标,计算公式为:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{3}$$

式中:TP(True Positive)为正类且被预测为正类的样本数;TN(True Negative)为负类且被预测为负类的样本数;FP(False Positive)为实际为负类但被错误预测为正类的样本数;FN(False Negative)为实际为正类但被错误预测为负类的样本数。

实验流程实施以下控制:

1)将数据集均分 5 个数量相等的互斥子集,轮替选择一个作为验证集,其余 4 个子集合并作为训练集。

2)各特征提取方法共享相同数据划分与 SVM 参数。

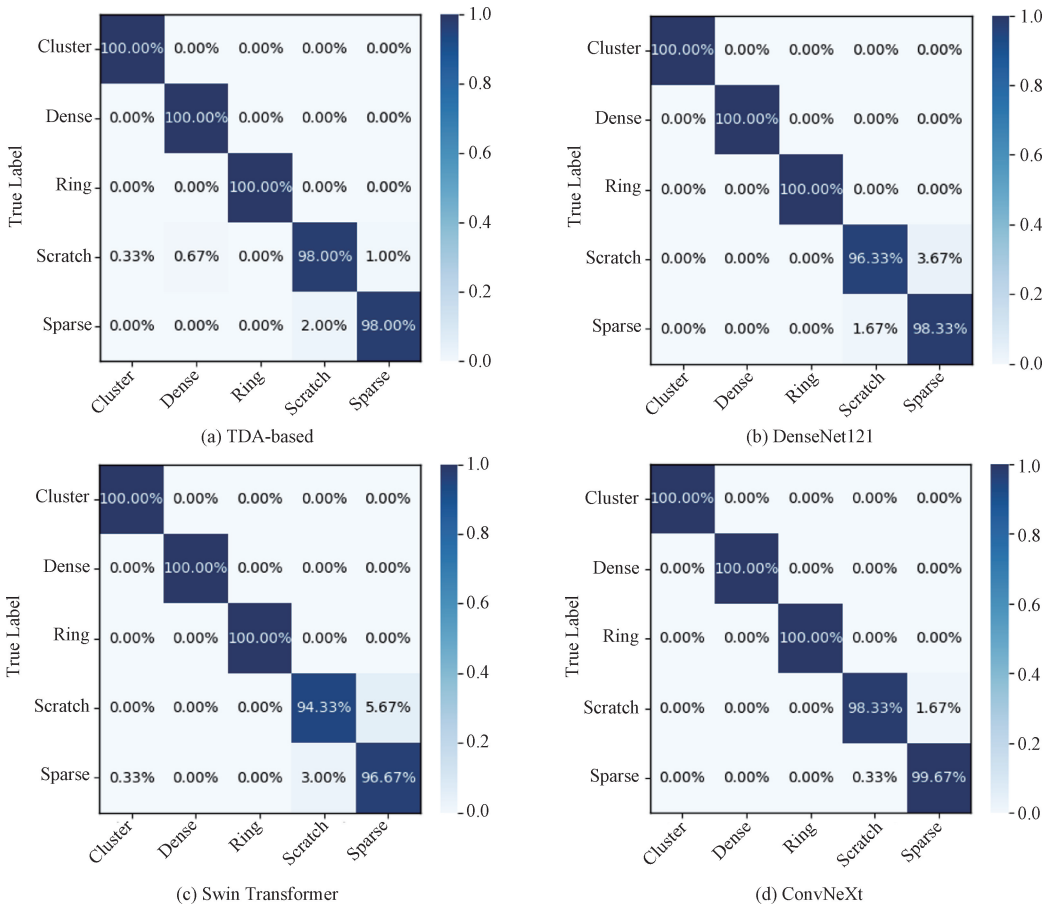
3)指标取 5 次均值以消除随机性,结果如表 3 所示。

从结果来看,新兴的 ConvNeXt 展现了最强分类性能,在交叉验证中实现平均准确率 0.996 0,其中两个 fold 达到完美分类(1.000 0),证实其作为前沿视觉模型的优

表 3　不同特征提取方法在 SVM 分类器下的性能评估

Table 3　Comparative performance evaluation of feature extraction methods using SVM classifier

Flod	TDA-based			DenseNet121			Swin Transformer			ConvNeXt		
	Accuracy	*F*1	*Recall*	*Accuracy*	*F*1	*Recall*	*Accuracy*	*F*1	*Recall*	*Accuracy*	*F*1	*Recall*
1	0. 987	0. 987	0. 987	0. 983	0. 983	0. 983	0. 973	0. 973	0. 973	1. 000	1. 000	1. 000
2	0. 997	0. 997	0. 997	0. 990	0. 990	0. 990	0. 990	0. 990	0. 990	1. 000	1. 000	1. 000
3	0. 997	0. 997	0. 997	0. 997	0. 997	0. 997	0. 987	0. 987	0. 987	0. 990	0. 990	0. 990
4	0. 987	0. 987	0. 987	0. 987	0. 987	0. 987	0. 987	0. 987	0. 987	0. 993	0. 993	0. 993
5	0. 993	0. 993	0. 993	0. 990	0. 990	0. 990	0. 973	0. 973	0. 973	0. 997	0. 997	0. 997
平均值	0. 992	0. 992	0. 992	0. 989	0. 989	0. 989	0. 982	0. 982	0. 982	0. 996	0. 996	0. 996



(a) TDA-based

(b) DenseNet121

(c) Swin Transformer

(d) ConvNeXt

图 9　混淆矩阵

Fig. 9　Confusion matrix

势。基于 TDA 的方法次之,在无监督特征基础上,取得 0. 992 0 平均准确率,相较监督学习的 DenseNet(0. 989 3)和 Swin Transformer(0. 982 0)更具竞争力,这验证了该特征提取方式能够更有效地捕捉数据中的关键信息,从而提高了模型的分类能力。进一步通过混淆矩阵(图 9)分析发现,基于 TDA 的方法与其他 3 种模型对密集、聚集与环形缺陷的识别准确率达到了 100%,这与其稳定的拓扑不变性密切相关。这 3 类缺陷在持久同调分析中呈现显著的零维/一维拓扑特征。在稀疏和划痕缺陷的识别上略逊一筹,这可能是因为划痕的特征多变且不易捕捉,尤其是短划痕的拓扑特征不够明显,对于线性划痕来说,没有弧形划痕较为显著的一维拓扑特征。此外,稀疏缺陷中的随机缺陷点形成的偶然结构同样可能产生一维特征点,这对划痕与稀疏缺陷的分类带来了挑战。

4　结　论

本文聚焦于改进一种利用 TDA 从晶圆图各类缺陷

模式中提取特征向量的新颖方法。与传统方法着重追求高分类精度不同,本文旨在探索 TDA 方法在晶圆图缺陷特征提取中的潜力,对比 TDA 与其他特征提取方法的优劣,以解决模型复杂、特征维度高以及缺陷模式与特征向量关系难以解释等问题。

为了应对上述挑战,本文改进了一种基于拓扑数据分析的新兴晶圆图缺陷模式分类方法。通过复形的优化,使得复形构建时间降低了约 82%、内存占用降低了 10.09%,通过构建更精简的拓扑结构获得了更强大的特征提取能力。特征提取可视化实验结果表明,该方法在聚类性能、特征提取时间以及峰值内存等指标上显著优于现有的经典方法,包括基于卷积神经网络的方法、基于 Transformer 的方法以及融合了 CNN 与 Transformer 优势的新兴模型 ConvNeXt,这充分证明了本文所提方法在晶圆图缺陷模式分类领域的优越性和应用潜力。尽管在加入分类器部分后,性能表现不如新兴模型 ConvNeXt,但依旧强于 DenseNet 和 Swin Transformer 等经典模型。值得注意的是,基于 TDA 的方法在特征提取效率(特征提取耗时较 ConvNeXt 模型提升约 3 倍)、资源消耗(内存峰值较 ConvNeXt 降低约 95%)方面展现出明显的优势,尤其在高速、低资源条件下具有广阔的应用前景。未来,结合深度学习与拓扑分析的多模态融合、扩展到大规模数据处理,将为 TDA 技术的应用开辟更广阔的空间,为多领域的智能分析提供新的解决方案。

参考文献

[1] 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.

[2] FAN M, WANG Q, VAN DER WAAL 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.

[3] 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.

[4] EZZAT A A, LIU S, HOCHBAUM D S, et al. A graph-theoretic approach for spatial filtering and its impact on mixed-type spatial pattern recognition in wafer bin maps[J]. IEEE Transactions on Semiconductor Manufacturing, 2021, 34(2): 194-206.

[5] 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.

[6] 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.

[7] 赵朗月, 吴一全. 基于机器视觉的表面缺陷检测方法研究进展[J]. 仪器仪表学报, 2022, 43(1): 198-219.
ZAHO L Y, WU Y Q. Research status and the prospect of PCB defect detection algorithm based on machine vision [J]. Chinese Journal of Scientific Instrument, 2022, 43(1): 198-219.

[8] 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.

[9] 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.

[10] 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.

[11] EDELSBRUNNER H, HARER J. Persistent homology-a survey [J]. Contemporary Mathematics, 2008, 453(26): 257-282.

[12] KO S, KOO D. A novel approach for wafer defect pattern classification based on topological data analysis [J]. Expert Systems with Applications, 2023, 231: 120765.

[13] CHAZAL F, MICHEL B. An introduction to topological data analysis: fundamental and practical aspects for data scientists[J]. Frontiers in Artificial Intelligence, 2021, 4: 667963.

[14] COHEN-STEINER D, EDELSBRUNNER H, HARER J. Stability of persistence diagrams [J]. Discrete & Computational Geometry, 2007, 37(1): 103-120.

[15] ADAMS H, EMERSON T, KIRBY M, et al. Persistence images: A stable vector representation of persistent homology[J]. Journal of Machine Learning Research, 2017, 18(8): 1-35.

[16] MARIA C, BOISSONNAT J D, GLISSE M, et al. The gudhi library: Simplicial complexes and persistent

homology[C]. International Congress on Mathematical Software. Springer, 2014: 167-174.
[17] 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, 2023, 46(6): 91-99.
[18] PARK S, YOU C. Deep convolutional generative adversarial networks-based data augmentation method for classifying class-imbalanced defect patterns in wafer bin map[J]. Applied Sciences, 2023, 13(9): 5507.
[19] YANG C J, CHEN Y H, HSIEH S Y. Enhanced wafer map defect pattern classification through stacking ensemble method and data augmentation integration[J]. The Journal of Supercomputing, 2025, 81(5): 1-31.
[20] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 1-9.
[21] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.
[22] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 4700-4708.
[23] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [C]. International Conference on Learning Representations, 2021.
[24] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention [C]. International Conference on Machine Learning, 2021: 10347-10357.
[25] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022.
[26] LIU Z, MAO H, WU C Y, et al. A convnet for the 2020s [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 11976-11986.

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.

丁家俊,2019 年于青岛理工大学获得学士学位,现为兰州理工大学硕士研究生,主要研究方向为拓扑数据分析与计算机视觉。
E-mail: happyjjding@ 163. com

Ding Jiajun received his B. Sc. degree from Qingdao University of Technology in 2019. Now he is a M. Sc. candidate in Lanzhou University of Technology. His main research interests include topological data analysis and computer vision.

董明月,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: 3270202883@ qq. com

Dong Mingyue 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.