kNViT: 基于深度学习的光刻热点精确检测方法

张 坤[1†]，郝文轩[1]，张立军[2]

(1. 河北科技大学 信息科学与工程学院，石家庄 050018；2. 苏州大学 轨道交通学院，江苏 苏州 215131)

摘 要: 随着集成电路 (IC) 工艺节点的不断缩小, 光刻版图中的热点问题对芯片性能和可靠性的影响日益显著. 鉴于传统基于深度学习的光刻热点检测方法难以满足先进集成电路制造对检测精度及模型泛化能力的要求, 提出一种基于深度学习的 k-Nearest Vision Transformer (kNViT) 模型, 用于光刻热点的精确检测. 所提出模型采用对比归一化模块 (CNM) 和 k-最近邻注意力模块 (kNAM) 来提升特征表示和识别精度. 同时, 利用光刻版图扩散模型 (PLDM) 生成图像, 增强了数据多样性. 此外, 提出电路特征感知损失函数 (CALF), 优化光刻版图扩散模型在预测噪声时的表现. 通过数据增强策略, 旋转和对比度调整, 进一步提升模型的泛化能力. 实验结果表明, kNViT 模型在多个光刻版图数据集上展现出高准确率的热点检测性能. 在 ICCAD 2012 数据集上, 平均召回率达 99.7%, 平均准确率 98%, 平均精确率 90.9%, F1 分数 95%. 研究表明, kNViT 模型可作为辅助检测工具, 有效提高检测准确性和效率, 具有工业设计应用潜力.

关键词: 光刻；热点；深度学习；k-Nearest Vision Transformer；光刻版图扩散模型；电路特征感知损失函数

中图分类号: TN305.7 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.0988

引用格式: 张坤, 郝文轩, 张立军. kNViT: 基于深度学习的光刻热点精确检测方法 [J]. 控制与决策.

kNViT: deep learning-based accurate detection method for photolithography hotspots

ZHANG Kun[1†]，*HAO Wen-xuan*[1]，*ZHANG Li-jun*[2]

(1. College of Information Science and Engineering，Hebei University of Science and Technology，Shijiazhuang 050018，China；2. College of Transportation Engineering，Soochow University，Suzhou 215131，China)

Abstract: With the continuous scaling of integrated circuit (IC) process nodes, the impact of hotspot issues in lithography layouts on chip performance and reliability has become increasingly significant. Traditional deep learning-based lithography hotspot detection methods struggle to meet the precision and model generalization requirements of advanced IC manufacturing. To address this, we propose a deep learning-based k-nearest vision Transformer (kNViT) model for precise detection of lithography hotspots. Our model employs a ContrastNorm module (CNM) and a k-nearest neighbor attention module (kNAM) to enhance feature representation and identification accuracy. Additionally, a photolithography layout diffusion model (PLDM) is employed to generate images, significantly improving data diversity. Furthermore, we introduce a circuit-aware loss function (CALF) to optimize the performance of the PLDM in predicting noise. Data augmentation strategies, including rotation and contrast adjustment, are applied to further enhance the model's generalization capability. Experimental results demonstrate that the kNViT model exhibits high-precision hotspot detection performance across multiple lithography layout datasets. On the ICCAD 2012 dataset, it achieves an average recall rate of 99.7%, accuracy of 98%, precision of 90.9%, and F1 score of 95%. The research indicates that the kNViT model can serve as an auxiliary detection tool, effectively improving detection accuracy and efficiency, with potential for industrial design applications.

Keywords: lithography；hotspot；deep learning；k-Nearest Vision Transformer；photolithography layout diffusion model；circuit-aware loss function

收稿日期: 2025-09-19；**录用日期:** 2026-03-10.

基金项目: 国家自然科学基金项目 (F011801)；石家庄市驻冀高校产学研合作项目 (241260054A).

责任编委: 褚菲.

[†]通信作者. E-mail: zhangkun@hebust.edu.cn.

0 引 言

随着集成电路 (Integrated Circuit, IC) 工艺节点的不断缩小, 光刻版图中纳米结构的关键尺寸已降至照明光波长以下, 导致显著的光衍射效应, 使硅片上曝光图形失真, 即光学邻近效应 (Optical Proximity Effect, OPE)[1]. OPE 直接影响集成电路的性能和可靠性, 是光刻工艺中的关键挑战. 业界通常采用光学邻近校正 (Optical Proximity Correction, OPC)[2-3] 技术补偿光学邻近效应, 但即使经过 OPC 优化, 版图中仍可能存在短路或开路等缺陷的区域, 即光刻热点 (hotspots), 而无缺陷区域则被称为“非热点” (non-hotspots). 热点区域通常表现为图案密度高、线条间距小或结构复杂, 导致曝光不均; 而热点可能引发开路和短路两种典型缺陷, 严重影响芯片功能.

传统热点检测方法主要有基于光刻仿真的方法[4] 和模式匹配方法[5], 前者精度高但耗时, 后者依赖热点库且难以检测未知热点. 近年来, 机器学习方法[6-7] 被引入, 但仍需大量标注数据且误报率较高. 深度学习模型如 RNN、CNN 等在多个领域表现优异[8-12], 并逐渐应用于热点检测. 吴清岳等[13] 改进 YOLOv5 提升了检测效率, 但复杂场景下仍有局限; JIANG 等[14] 提出二值化残差网络集成, 提高了准确性和效率; Liao 等[15] 采用迁移学习, 但模型泛化能力有待提升. 在深度学习架构设计方面, Transformer 与 CNN 的混合架构已在多个视觉任务中展现出优越性能. Yong 等[16] 针对机械臂抓取检测问题, 提出一种 Transformer-CNN 并行网络结构, 通过结合 CNN 的局部建模能力和 Transformer 的全局建模能力, 有效提升了跨模态特征的表征能力. 受此启发, 本文设计的 kNViT 模型同样采用 Transformer 架构, 并通过 CNM 和 kNAM 模块进一步增强对光刻版图热点区域的特征聚焦能力. 总体而言, 现有方法在检测精度和泛化能力上仍有不足.

针对上述挑战, 本文提出 kNViT(k-Nearest Vision Transformer) 模型, 主要贡献如下:

(1) 光刻版图扩散模型 (Photolithography Layout Diffusion Model, PLDM): 将扩散模型应用于光刻版图数据增强, 融入 28 nm 工艺节点的细线宽容限和 45 nm 相干核尺寸等物理约束, 生成高质量辅助样本, 解决数据多样性不足与类别不平衡问题.

(2) 电路特征感知损失函数 (Circuit-Aware Loss Function, CALF): 基于 IEEE Skeleton-EDT 量化线宽、间距及图案复杂度. 通过最小二乘预对齐和 softmax 生成动态权重, 优化噪声预测过程, 提升关键特征保真度.

(3) 对比归一化模块 (ContraNorm Module, CNM): 通过负样本加权方差和可学习温度系数, 在单样本内完成归一化, 主动放大关键通道差异, 增强对细微局部特征的敏感度.

(4) k-最近邻注意力模块 (kNNAttention Module, kNAM): 为每个查询挑选最相关的 k 个键进行注意力计算, 在聚焦热点区域的同时降低计算复杂度.

(5) CNM 与 kNAM 协同机制: 将 kNAM 输出的通道级峰值注意力以残差形式回传至 CNM, 动态调节温度参数, 形成注意力高亮到对比度增强再到特征再校准的自优化回路, 实现像素级检测与校正协同, 实验验证可减少线端回缩量 2.8 nm.

1 研究方法

本文提出的 kNViT 模型用于光刻版图热点检测, 其结构如图 1 所示. 模型处理流程分为四个阶段: 数据增强、特征提取、对比归一化和注意力聚焦分类.

在数据增强阶段 (a), 采用光刻版图扩散模型 (PLDM) 生成多样化的图像样本, 并实施旋转和对比度增强, 以扩充数据集并提高泛化能力. 特征提取阶段 (b), 将输入图像划分为图像块并添加位置编码和分类标记, 优化特征表示. 对比归一化模块 (CNM, c), 通过归一化消除特征尺度差异, 提升对细微局部特征的识别. k-最近邻注意力模块 (kNAM, d), 为每个查询筛选最相关的 k 个键, 最终将特征传递至分类层输出检测结果.

1.1 光刻版图扩散模型 (Photolithography Layout Diffusion Model, PLDM)

PLDM 模块对应图 1 流程阶段 (a), 通过前向扩散与反向去噪生成高质量辅助热点版图, 解决数据多样性不足与类别不平衡问题.

扩散模型在图像生成领域已取得显著进展[17-19], 但在光刻版图生成中仍面临样本多样性不足和质量提升的挑战. 为此, PLDM 在前向扩散过程中逐步添加噪声, 每一步条件分布如公式 (1) 所示:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I). \quad (1)$$

其中 $\alpha_t = 1 - \beta_t$ 在 [0.98,0.9999] 区间线性递增. 经 FID 实验校准, 该区间使初期噪声幅度小于 28 nm 细线宽容限, 末期形成 45 nm 相干核, 模拟光刻渐进失真.

为优化反向去噪中的特征保留, 提出了电路特征感知损失函数 (Circuit-Aware Loss Function, CALF). 采用 IEEE Skeleton-EDT 基线作为统一量化方式, 通过公式 (2) 统一量化线宽:

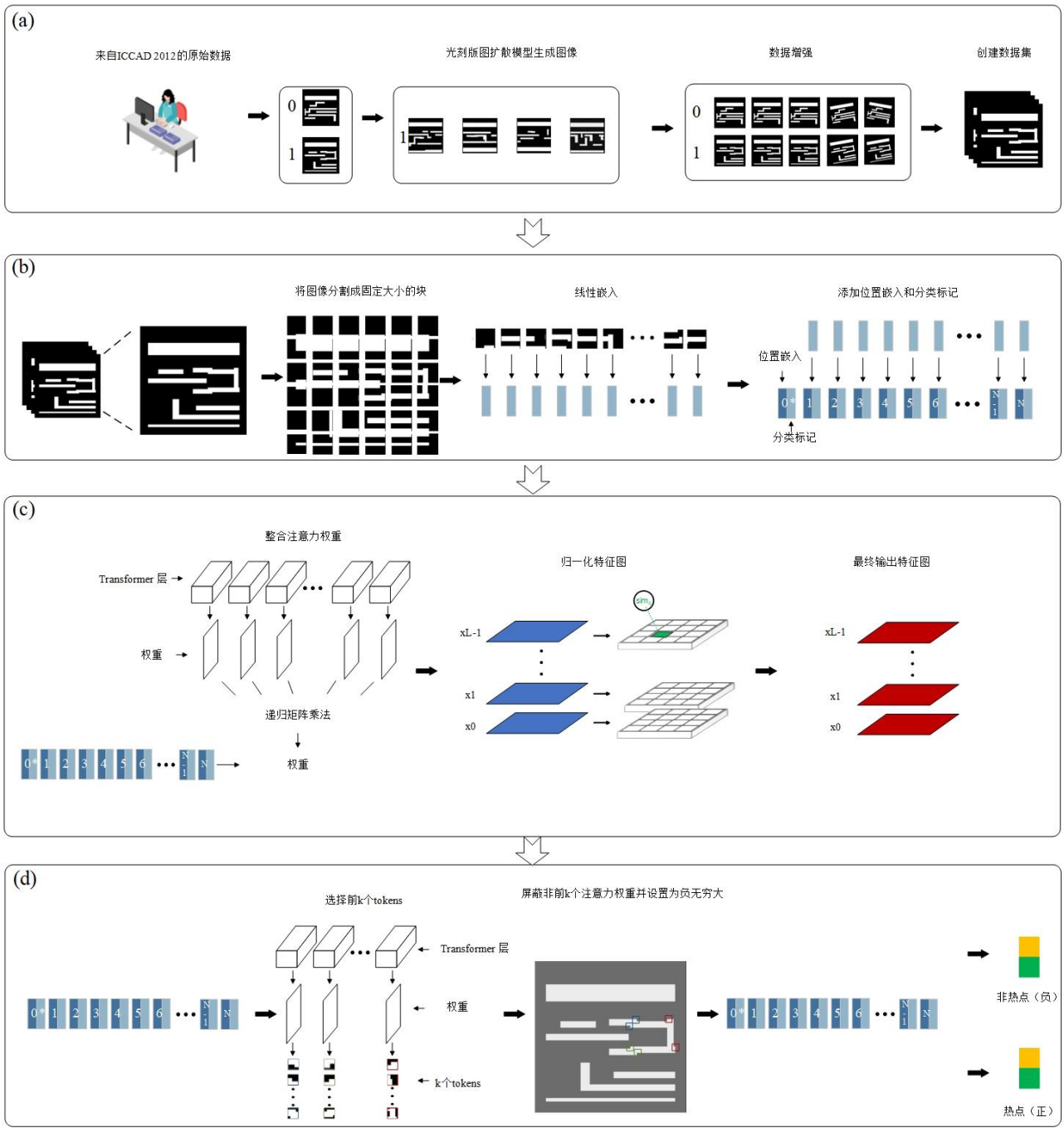


图1　kNViT 模型结构图

$$w = 2 \cdot \mathrm{mean}_{p \in S} \mathrm{EDT}(p). \tag{2}$$

通过公式 (3) 提取最小间距:

$$s = \min_{p \in \mathcal{L}} \mathrm{EDT}(p). \tag{3}$$

通过公式 (4) 计算布线复杂度:

$$c = \frac{|\mathcal{N}_{\mathrm{end}}| + |\mathcal{N}_{\mathrm{junc}}|}{\mathrm{Area}}. \tag{4}$$

将统一量化特征$\mathbf{f} = [w, s, c]^{\mathrm{T}}$送入公式 (5) 一次性最小二乘矩阵:

$$W = \mathrm{softmax}(Af + b). \tag{5}$$

其中A、b由训练集一次性最小二乘解析解确定, 且$\sum W_i = 1$. 该权重消除线宽以 nm 为单位、间距以 nm 为单位、复杂度为无量纲密度等尺度差异, 为后续模块提供可解释的初始权重.

表 1 实验表明, 仅采用线性预对齐加 softmax 归一化, 就使 F1 分数从固定权重基线的 0.892 提升至 0.903. 特征间高阶非线性依赖由后续 Transformer 编码器、CNM 和 kNAM 共同建模. 通过消融实验引入 CNM 与 kNAM 模块, 提高了模型的非线性处理能力, F1 分数进一步从 0.903 提升至 0.932, 验证了线

表1　ICCAD-3 基准统一量化基线与动态权重的性能比较

策略	F1	权重来源
统一量化	0.892	IEEE Skeleton-EDT基线
动态权重	0.903	统一尺度 + 线性层

性基础拟合与深度非线性建模分工协作架构的有效性.

CALF 可以通过公式 (6) 和 (7) 描述:

$$\mathcal{L}_{\text{SCALE}} = \frac{1}{N}\sum_{i=1}^{N} W_i \cdot \|\mathbf{E}_i - \epsilon_i\|_2^2, \tag{6}$$

$$W_i = \frac{w_i}{\frac{1}{N}\sum_j w_j}. \tag{7}$$

反向去噪通过迭代更新恢复样本. 通过公式 (8) 描述:

$$x_{t-1} = \frac{x_t - \sqrt{\beta_t}\cdot\epsilon_t}{\sqrt{1-\overline{\alpha}_t}}. \tag{8}$$

式中, x_{t-1}是去噪后的样本, x_t是当前步骤的噪声样本, $\overline{\alpha}_t$是累积噪声水平的补数.

通过迭代这个过程, 模型逐步学习如何从噪声数据中恢复出原始数据的特征. 每一步都使数据更接近原始分布, 直到最终恢复出清晰的、无噪声的数据样本.

PLDM 生成的图像如图 2 所示, 其 FID 值为 9.34, 质量与真实样本相近. 生成后经±10°旋转和±5% 对比度增强用于 kNViT 训练.

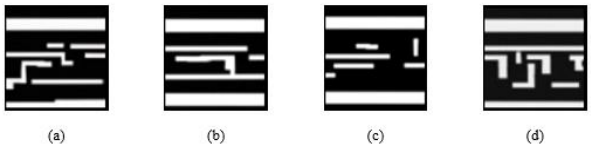


图2 PLDM 生成的辅助样本 (a) 真实热点版图; (b)(c)(d) 辅助热点版图

1.2 k-Nearest Vision Transformer (kNViT) 模型

kNViT 模块对应图 1 流程阶段 (b), 负责将 PLDM 增强样本转换为特征表示. 执行图像分块、线性投影与位置编码, 构建初始特征矩阵, 为后续 CNM 与 kNAM 提供输入.

与 ViT 相比, kNViT 在以下三个方面具有区别:

首先, 在处理流程上, ViT 采用图像分块、线性投影、位置编码、Transformer 编码器到分类输出的单向处理流程. 而 kNViT 在此基础上构建了基础编码、特征优化、注意力聚焦到闭环反馈的递进式处理流程.

其次, 在特征处理方式上, ViT 对所有图像块采用全局自注意力, 难以捕捉稀疏热点特征. kNViT 通过 CNM 负样本加权方差放大关键通道差异, 并通过 kNAM 为每个查询筛选最相关的 k 个键进行注意力计算, 在聚焦热点的同时降低复杂度.

第三, 在模块交互机制上, ViT 模块间单向传递, 缺乏反馈调节. kNViT 则建立闭环反馈机制, kNAM 输出的峰值注意力回传至 CNM 调节温度参数, 形成自适应优化回路.

kNViT 将输入图像划分成大小为$P\times P$的N个小块, 如公式 (9) 所示:

$$N = \frac{H}{P}\cdot\frac{W}{P}. \tag{9}$$

其中H、W为图像高、宽, 每个小块经线性映射得特征矩阵, 并加入位置编码, 加上位置编码后的初始特征矩阵如公式 (10) 所示:

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{E}\mathbf{x}^1_{\text{patch}}; \ldots; \mathbf{E}\mathbf{x}^N_{\text{patch}}] + \mathbf{E}_{\text{pos}}. \tag{10}$$

其中$\mathbf{x}_{\text{class}}$为分类令牌, $\mathbf{E}_{\text{pos}}$为位置编码. 初始特征矩阵输入 Transformer 编码器, 生成分类令牌和注意力权重矩阵, 供 CNM 和 kNAM 进一步优化特征. CNM 通过归一化消除特征尺度差异, 增强对边角圆化、线端缩短等细微特征的识别; kNAM 通过筛选关键特征, 提升对热点空间布局的感知.

1.3 对比归一化模块 (ContraNorm Module, CNM)

CNM 模块对应图 1 流程阶段 (c), 对初始特征矩阵进行优化. 基于对比学习视角, 通过负样本加权方差替代传统全局统计, 在单样本内完成归一化, 并引入可学习温度系数对通道执行对比式拉伸, 以放大关键通道差异、抑制过度平滑.

首先, 对输入特征图$\boldsymbol{X}$进行范数归一化, 如公式 11 所示:

$$x_n = \frac{X}{\sqrt{\sum_{j=1}^{D} X_{:,j}^2 + \epsilon}}. \tag{11}$$

其中, ϵ是一个小常数. 随后计算负样本加权和并进行对比式拉伸, 如公式 (12) 所示:

$$\text{ContraNorm}(X) = \text{LayerNorm}(X - \alpha\cdot(\frac{XX^{\mathrm{T}}}{\tau}\odot\text{softmax}(\frac{XX^{\mathrm{T}}}{\tau}))). \tag{12}$$

其中, $\odot$表示哈达玛积; α是可学习的缩放参数; τ是温度参数. 该机制使模型更关注特征间细微差异, 减少维度崩溃.

CNM 原理示意图如图 3 所示:

为验证 CNM 的抗噪性, 在 ICCAD-4 测试集上注入 σ=0.5 高斯白噪声, 对比不同归一化层的性能, 结果如表 2 所示. ContraNorm 加噪后准确率降幅仅 1.8%, 显著优于 LayerNorm(4.2%) 和 BatchNorm (4.9%), 且 F1 分数最高, 为 0.906, 表明其在保持特征稳定性方面具有优势.

CNM 通过负样本加权方差和对比式拉伸, 在单

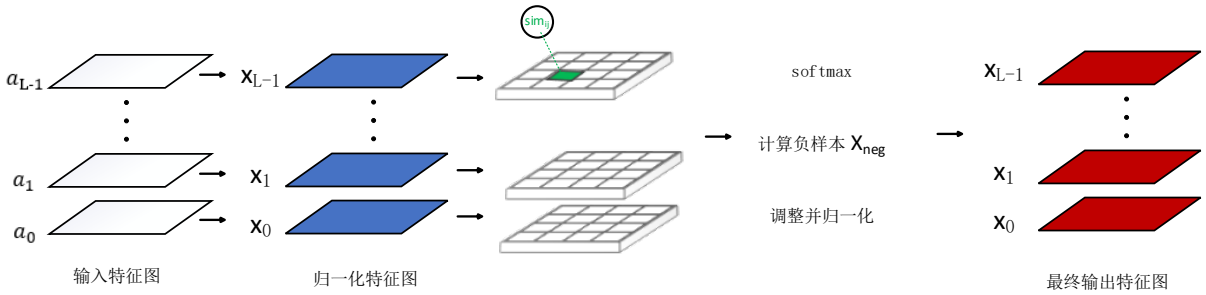


图3　CNM 原理示意图

表2　ICCAD-4 基准不同归一化层的性能对比

方法			召回率	准确率	精确率	加噪后准确率	抗噪准确率降幅	F1分数
LayerNorm	BatchNorm	ContraNorm						
√			0.977	0.979	0.835	0.937	0.042	0.900
	√		0.991	0.985	0.790	0.936	0.049	0.880
		√	0.993	0.984	0.833	0.966	0.018	0.906

样本内实现稳定归一化, 增强了对细微局部特征的敏感度和噪声鲁棒性. 优化后的特征矩阵将输入 kNAM 进行热点聚焦.

1.4　k-最近邻注意力模块 (kNNAttention Module, kNAM)

kNAM 模块对应图 1 流程阶段 (d), 接收 CNM 优化后的特征矩阵, 通过筛选最相关的 k 个键精准聚焦热点区域. 显著降低计算复杂度的同时提升检测精度.

传统的自注意力机制在处理复杂且稀疏的特征时存在以下局限性: 首先, 传统的自注意力机制需要计算每个查询与所有键的注意力权重, 这在大规模数据和复杂特征场景下会导致计算成本过高. 其次, 在复杂的工业环境中, 版图设计和成像条件存在显著差异, 传统的自注意力机制容易受到噪声特征的干扰, 导致误报率较高. 最后, 传统的自注意力机制虽然能够捕捉全局信息, 但在处理稀疏且复杂的热点特征时, 难以精准地聚焦于关键特征, 导致检测精度不足. kNAM 通过分析注意力权重, 为每个查询保留权重最高的 k 个键, 屏蔽非关键区域, 从而增强热点识别精度并降低复杂度.

输入特征X通过线性变换被映射为查询Q, 键K和值V, 如公式 (13) 所示:

$$Q = XW_Q, K = XW_K, V = XW_V \tag{13}$$

其中, W_Q, W_K和W_V是可学习的权重矩阵. 计算原始注意力矩阵并缩放, 如公式 (14) 所示:

$$\mathrm{A} = \frac{QK^{\mathrm{T}}}{\sqrt{d_k}} \tag{14}$$

接着, 应用函数选择每个查询最相关的 k 个键, 构建掩码矩阵以忽略非 top-k 的注意力权重. 最后, 对调整后的注意力矩阵执行 Softmax 归一化, 并加权求和 以获得输出特征.

如图 4 所示, kNAM 通过分析注意力权重矩阵$\boldsymbol{a}_i$识别出关键特征, 通过选择每个查询最相关的 k 个键来计算注意力, 从而增强模型的识别精度并优化特征表示. 通过减少每个查询需要考虑的键的数量, 降低了计算复杂度, 同时保留了最重要的信息, 以提高模型对关键特征的关注度.

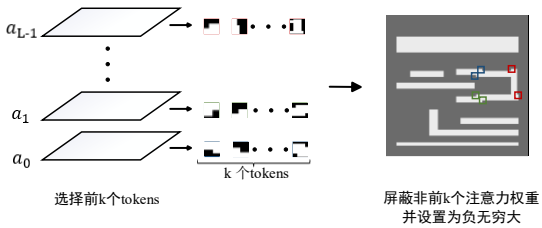


图4　kNAM 原理示意图

为确定最优 k 值, 在 ICCAD-2 上仅改变 k∈{5, 10,15,20}, 其余超参保持不变. 结果如表 3 所示, k=10 时 F1 达到 0.906.

表3　k 值敏感性对比

k	召回率	准确率	F1
5	0.971	0.978	0.904
10	0.993	0.984	0.906
15	0.992	0.983	0.905
20	0.990	0.981	0.903

注意力可视化如图 5 所示, 显示 k = 10 聚焦最尖锐, 后续实验统一采用 k = 10.

1.5　CNM 与 KNAM 的协同

CNM 与 KNAM 的协同模块贯穿图 1 流程阶段 (c) 特征优化与 (d) 热点聚焦检测阶段, 建立跨模块闭环协同机制. kNAM 输出的峰值注意力信号回

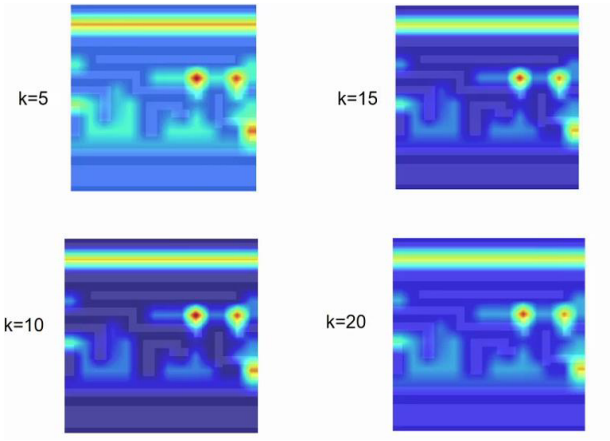


图5 kNAM 注意力权重可视化

传至 CNM, 动态调节温度系数, 形成注意力高亮到对比度增强再到特征再校准的自调节回路, 在像素级实现检测与校正协同.

CNM 对输入特征做通道维度温度缩放得到$\hat{X}$; kNAM 再以$\hat{X}$作为输入进行注意力聚焦, 并输出峰值注意力ω_c. CNM 温度参数更新如公式 (15) 所示:

$$T=T\odot(1+\lambda\cdot\omega_c). \tag{15}$$

式中ω_c表示通道在整张特征图中的最大注意力值, λ为可学习系数.

为验证协同逻辑, 在固定模型权重下仅改变残差系数$\lambda\in\{0, 0.1, 0.3, 0.5, 0.7\}$, 观测线端回缩量. 对 100 张版图施加高斯模糊, σ=2.5 px 以模拟光学扩散, 测量线端回缩量, 结果如表 4 所示. 当λ增至 0.3 时, 回缩量减少 2.8 nm, 性能最优; 继续增大 λ则改善饱和. 这表明 kNViT 的峰值注意力可映射为 OPC 校正窗口参数, 实现无需真实光刻机的检测与校正协同.

表4 不同λ下的回缩量变化

λ	ω	T	回缩/nm	拉伸情况
0	0.31	1.00	0	无反馈
0.1	0.31	0.91	−1.0	轻微拉伸
0.3	0.31	0.74	−2.8	最优拉伸
0.5	0.31	0.58	−2.7	过拉伸
0.7	0.31	0.44	−1.0	梯度饱和

2 分析与讨论

2.1 实验数据集

采用 ICCAD 2012 基准数据集, 包含 5 个 GDSII 格式版图, 其中 ICCAD-1 为 32 nm 工艺, ICCAD-2 至 ICCAD-5 为 28 nm 工艺. 数据划分为热点与非热点两类, 原始样本分布如表 5 所示.

为缓解类别不平衡, 按 1:5 比例对训练集热点样本进行数据增强 (采用 PLDM 生成辅助样本), 增强后分布如表 6 所示.

表5 ICCAD 2012 基准统计

基准	训练集		测试集	
	热点样本数量	无热点样本数量	热点样本数量	无热点样本数量
ICCAD1	99	340	226	4679
ICCAD2	174	5285	498	41298
ICCAD3	909	4643	1808	46333
ICCAD4	95	4452	177	31890
ICCAD5	26	2716	41	19327

表6 ICCAD 2012 数据增强后的数据集统计

基准	训练集		测试集	
	热点样本数量	无热点样本数量	热点样本数量	无热点样本数量
ICCAD1	68	340	226	4679
ICCAD2	1057	5285	498	41298
ICCAD3	929	4643	1808	46333
ICCAD4	890	4452	177	31890
ICCAD5	543	2716	41	19327

2.2 实验设置

kNViT 基于 PyTorch 框架, 采用 Vision Transformer (ViT) 作为骨干网络. 实验硬件为 11th Gen Intel(R) Core(TM) i7-11700K 处理器 (3.60 GHz)、80GB 的随机存取存储器 (RAM) 以及 NVIDIA 3080 图形处理单元, 显存容量 12GB. 训练 50 个 epoch, 批处理大小为 32. 优化算法采用随机梯度下降 (SGD), 初始学习率为 0.1. 验证损失不再显著下降时, 以 0.8 倍衰减以促进收敛. 根据验证集性能选取最优模型.

2.3 消融实验分析

为量化各模块贡献, 在 ICCAD-2 基准上以 ViT 为基线, 依次引入 PLDM 数据增强、CALF 损失函数、CNM 模块和 kNAM 模块, 结果如表 7 所示. PLDM 使 F1 提升 2.1%, 缓解类别不平衡; CALF 再提升 1.0%, 优化特征表示. CNM 与 kNAM 贡献最大. 单独引入 CNM 使 F1 提升 2.3%, 召回率达到 0.993, 增强细微局部特征敏感度. 而单独引入 kNAM 使 F1 分数提升 4.9%, 精确率增幅 7.1%, 大幅降低误报率. 二者协同后召回率达 1.000, F1 分数

表7 基于 ICCAD-2 的模块增量消融性能对比

方法				召回率	准确率	精确率	F1分数	运行时间/(h·mm^{-2})
PLDM	CALF	CNM	kNAM					
				0.951	0.953	0.771	0.852	0.450
√				0.967	0.968	0.796	0.873	0.614
√	√			0.977	0.979	0.805	0.883	0.683
√	√	√		0.993	0.984	0.833	0.906	0.734
√	√		√	0.995	0.981	0.876	0.932	0.612
√	√	√	√	1.000	0.986	0.872	0.932	0.725

稳定在 0.932, 运行时间 0.725 h·mm^{-2}, 验证了协同机制的有效性.

在明确各模块贡献的基础上, 鉴于 CNM 与 kNAM 在 ICCAD-2 基准上表现出显著的性能提升 (F1 分数分别提升 2.3% 和 4.9%), 本节进一步在 ICCAD 1–5 五个数据集上验证二者在不同工艺节点上的泛化能力. 所有配置均已包含 PLDM 数据增强和 CALF 损失函数 (即表 7 第三行配置为基础), 配置分别为: 添加对比归一化模块的 CNM 模型, 以及同时添加 CNM 和 kNAM 并形成闭环反馈机制的完整 kNViT 模型. 实验结果如表 8 所示.

表8　ICCAD1 ～ ICCAD15 基准上 CNM 与 kNAM 模块引入性能对比

数据集基准	方法			召回率	准确率	精确率	F1分数
	ViT	CNM	kNAM				
ICCAD1	√			0.984	0.953	0.968	0.976
	√	√		0.991	0.961	0.979	0.985
	√	√	√	**1.000**	**0.962**	**0.983**	**0.991**
ICCAD2	√			0.977	0.979	0.805	0.883
	√	√		0.993	0.984	0.833	0.906
	√	√	√	**1.000**	**0.986**	**0.872**	**0.932**
ICCAD3	√			0.980	0.974	0.838	0.903
	√	√		0.993	0.983	0.854	0.918
	√	√	√	**1.000**	**0.980**	**0.873**	**0.932**
ICCAD4	√			0.948	0.974	0.982	0.965
	√	√		0.959	0.986	0.991	0.975
	√	√	√	**0.991**	**0.991**	**0.985**	**0.988**
ICCAD5	√			0.957	0.953	0.772	0.855
	√	√		0.983	0.977	0.794	0.878
	√	√	√	**0.996**	**0.981**	**0.832**	**0.907**

从表 8 中可以看出, kNViT 模型在多个基准测试中的表现均优于其他模型. 特别是在 kNViT(即 ViT+CNM+kNAM) 和 ViT+CNM 这两个变体中, 精确率和召回率均接近 1.000, F1 分数达到了 0.878 以上, 表明 CNM 与 kNAM 的协同作用在不同工艺节点和数据集上均具有稳定的泛化能力.

2.4　kNViT 与先进模型性能对比分析

为了全面评估 kNViT 模型的性能, 进一步验证本节方法在热点检测领域的有效性, 将其与三种深度学习方法作为对比实验. Chen 等人融合了 Squeeze-and-Excitation(SE) 注意力机制和 Efficient Channel Attention(ECA) 机制, 提出了一种轻量级的光刻热点检测模型[20]. Lin 等人提出了一种改进的 YOLOv5 模型, 将空间注意力机制嵌入到 YOLOv5 模型的骨干网络中, 通过翻转策略进行数据增强, 解决了热点和非热点样本不平衡的问题[21]. Swin Transformer 作为一种新兴的 Transformer 架构, 也在光刻热点检测中展现出了良好的性能. 其实验结果如表 9 所示.

表9　kNViT 与先进模型性能对比分析

数据集基准	方法	召回率	准确率	精确率	F1分数
ICCAD1	[19]	0.971	0.954	0.976	0.973
	[20]	1.000	–	0.893	0.944
	Swin Trans	0.949	0.893	0.902	0.925
	kNViT	**1.000**	**0.962**	**0.983**	**0.991**
ICCAD2	[19]	0.993	0.989	0.893	0.940
	[20]	1.000	–	0.969	0.984
	Swin Trans	0.973	0.976	0.701	0.815
	kNViT	**1.000**	**0.986**	**0.872**	**0.932**
ICCAD3	[19]	0.953	0.963	0.861	0.905
	[20]	1.000	–	0.663	0.797
	Swin Trans	0.921	0.937	0.758	0.832
	kNViT	**1.000**	**0.980**	**0.873**	**0.932**
ICCAD4	[19]	0.997	0.992	0.927	0.961
	[20]	1.000	–	0.803	0.891
	Swin Trans	0.990	0.988	0.876	0.930
	kNViT	**0.991**	**0.991**	**0.985**	**0.988**
ICCAD5	[19]	1.000	0.991	0.956	0.978
	[20]	1.000	–	0.803	0.891
	Swin Trans	0.971	0.992	0.514	0.672
	kNViT	**0.996**	**0.981**	**0.832**	**0.907**
平均	[19]	0.983	0.978	0.923	0.951
	[20]	1.000	–	0.826	0.901
	Swin Trans	0.961	0.957	0.750	0.835
	kNViT	**0.997**	**0.980**	**0.909**	**0.950**

由表可见, kNViT 在多数数据集上取得最优或次优的 F1 分数, 平均召回率 0.997、F1 分数 0.950, 与文献 [20] 的 0.951 相当, 但召回率更高.

在特征提取方面, 文献 [20] 的 CNN 架构难以捕捉细微局部特征, 文献 [21] 的 YOLOv5 对细小热点聚焦不足, Swin Transformer 对复杂背景敏感; 而 kNViT 通过 CNM 放大关键通道差异、kNAM 聚焦热点, 有效提升了局部几何特征的识别精度.

在误报率控制上, 文献 [20] 易对噪声误报, 文献 [21] 在热点上误报较高, Swin Transformer 噪声敏感; kNViT 利用 CNM 和 kNAM 抑制干扰, 在精确率和 F1 上取得平衡.

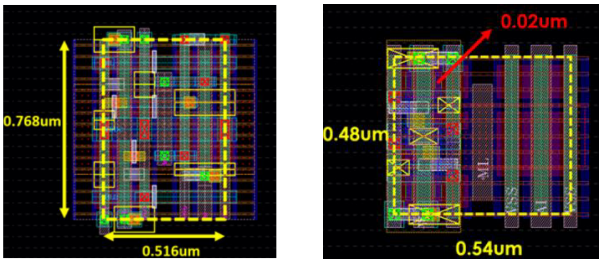
在部署可行性上, 文献 [21] 和 Swin Transformer 计算成本较高, kNViT 的稀疏注意力降低了复杂度, 更适合工业场景.

2.5　实验验证与模型应用

为评估 kNViT 的实用性, 选取 28 nm 和 32 nm 工艺版图各 20 张, 重点关注关键路径、高密度区域等易产生热点的位置. 模型对每张版图进行热点检测后, 针对预测为热点的区域提出优化建议 (如调整

线宽、增加间距等), 优化后再次检测.

以其中一张版图为例, 优化前模型预测为热点, 经版图调整后预测为非热点, 如图 6 所示.



(a) (b)

图6 优化前后的版图对比. (a) 优化前的版图; (b) 优化后的版图

本文对 28 nm 和 32 nm 工艺节点的版图进行了统计分析. 图 7 展示了 28 nm 和 32 nm 工艺节点在优化前后的热点数量对比. 在 28 nm 工艺节点下, 优化前的热点数量为 14 个, 优化后减少到 6 个, 热点数量减少了 57.1%. 在 32 nm 工艺节点下, 优化前的热点数量为 8 个, 优化后减少到 3 个, 热点数量减少了 62.5%. 实验证明 kNViT 能有效识别潜在热点, 辅助版图优化, 提升制造可靠性.

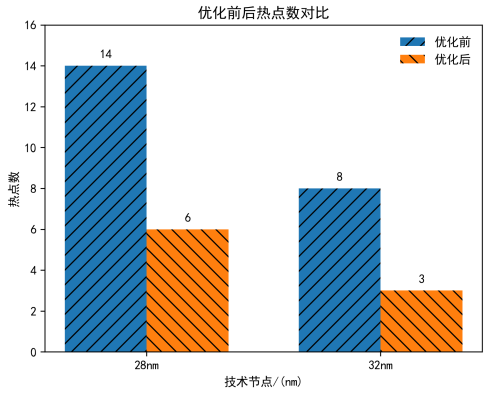


图7 14nm 和 28nm 工艺节点在优化前后的热点数量对比图

3 结 论

本文提出 kNViT 模型, 通过 CNM 和 kNAM 的协同设计, 有效提升了对光刻版图热点细微特征的识别能力与抗干扰性能, 并增强了模型的泛化能力. PLDM 生成样本的 FID 为 9.34; CNM 在 ICCAD-5 上将召回率由 0.957 提升至 0.983、F1 由 0.855 提升至 0.878; kNAM 在 ICCAD-2 上将精确率由 0.833 提升至 0.872、F1 由 0.906 提升至 0.932. 模型在工业部署中具有可行性, 其轻量化设计与并行架构可满足实时检测需求, 并能有效指导版图优化. 未来工作将聚焦于降低 FID 值、模型压缩、多任务学习及实时检测算法, 以进一步提升模型性能.

参考文献 (References)

[1] 郝芸芸, 董立松, 粟雅娟, 等. 基于模型的光学邻近效应修正应用技术 (特邀)[J]. 光学学报, 2025, 45(5): 0500001.
(Hao Y Y, Dong L S, Su Y J, et al. Model-based optical proximity correction application technology (invited)[J]. Acta Optica Sinica, 2025, 45(5): 0500001.)

[2] Wu R X, Dong L S, Wei Y Y. Method for optical proximity correction based on a vector imaging model[J]. Applied Optics, 2024, 63(10): 2719-2727.

[3] Ji J X, Wei B, Wen G J, et al. Edge-based near-field photolithography optical proximity effect correction technique[J]. Optical and Quantum Electronics, 2024, 56: 1812.

[4] Roseboom E, Rossman M, Chang F C, et al. Automated full-chip hotspot detection and removal flow for interconnect layers of cell-based designs[J]. Design for Manufacturability Through Design-Process Integration, 2007, 6521: 65210C.

[5] Chen K J, Chuang Y K, Yu B Y, et al. Minimizing cluster number with clip shifting in hotspot pattern classification[C]. Proceedings of the 54th Annual Design Automation Conference. Austin, 2017: 1-6.

[6] Kataoka G, Yamamoto M, Inagi M, et al. Feature vectors based on wire width and distance for lithography hotspot detection[J]. IPSJ Transactions on System and LSI Design Methodology, 2023, 16: 2-11.

[7] Liu Y J, Li X X, Pei B, et al. Towards smart scanning probe lithography: A framework accelerating nano-fabrication process with in-situ characterization via machine learning[J]. Microsystems & Nanoengineering, 2023, 9: 128.

[8] Brenner M, Weber E, Koppe G, et al. Learning Interpretable hierarchical dynamical systems models from time series data[C]. The 13th International Conference on Learning Representations. Singapore, 2025.

[9] Koresh E, Gross R D, Meir Y, et al. Unified CNNs and transformers underlying learning mechanism reveals multi-head attention modus vivendi[J]. Physica A: Statistical Mechanics and its Applications, 2025, 666: 130529.

[10] Schneider S, Antensteiner D, Soukup D, et al. Autoencoders — A comparative analysis in the realm of anomaly detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans, 2022: 1986-1992.

[11] Bai L, Cui L X, Hancock E, et al. HC-GAE: The hierarchical cluster-based graph auto-encoder for graph representation learning[C]. Advances in Neural Information Processing Systems 37. Vancouver, 2024: 127968-127986.

[12] Jabbar A, Li X, Omar B. A survey on generative adversarial networks: Variants, applications, and training[J]. ACM Computing Surveys, 2022, 54(8): 1-

49.

[13] 吴清岳, 刘佳敏, 张松, 等. 基于改进 Yolov5s 的光刻热点检测算法[J]. 激光与光电子学进展, 2023, 60(24): 251-259.
(Wu Q Y, Liu J M, Zhang S, et al. Lithography hotspot detection based on improved Yolov5s[J]. Laser & Optoelectronics Progress, 2023, 60(24): 251-259.)

[14] Jiang Y Y, Yang F, Yu B, et al. Efficient layout hotspot detection *via* binarized residual neural network ensemble[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2021, 40(7): 1476-1488.

[15] Liao L F, Li S K, Che Y Q, et al. Lithography hotspot detection method based on transfer learning using pre-trained deep convolutional neural network[J]. Applied Sciences, 2022, 12(4): 2192.

[16] 王勇, 李邑灵, 苗夺谦, 等. 基于 Transformer-CNN 混合架构的跨模态融合抓取检测[J]. 控制与决策, 2024, 39(11): 3607-3616.
(Wang Y, Li Y L, Miao D Q, et al. Cross-modal interaction fusion grasping detection based on Transformer-CNN hybrid architecture[J]. Control and Decision, 2024, 39(11): 3607-3616.)

[17] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis[C]. Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021). San Diego, 2021.

[18] Sadat S, Kansy M, Hilliges O, et al. No training, no problem: Rethinking classifier-free guidance for diffusion models[C]. The 13th International Conference on Learning Representations (ICLR 2025). Singapore, 2025.

[19] 齐时达. 基于布局控制的文本到图像扩散模型研究进展[J]. 计算机科学与应用, 2025, 15(4): 443-452.
(Qi S D. Research progress on text-to-Image diffusion models based on layout control[J]. Computer Science and Application, 2025, 15(4): 443-452.)

[20] Chen Y N, Li Y J, Wu B, et al. Lightweight hotspot detection model fusing SE and ECA mechanisms[J]. Micromachines, 2024, 15(10): 1217.

[21] Lin M, He W J, Liu J L, et al. An improved YOLOv5 model for lithographic hotspot detection[J]. Micromachines, 2025, 16(5): 568.

作者简介

张坤 (1982−), 女, 副教授, 博士, 主要研究方向为机器视觉、人工智能、三维图形学, E-mail: zhangkun@hebust.edu.cn;

郝文轩 (2000−), 女, 硕士生, 主要研究方向为集成电路 (IC) 制造中的光刻版图热点检测, E-mail: 3178127530@qq.com;

张立军 (1970−), 男, 研究员, 博士, 博士生导师, 主要研究方向为系统级芯片 (SOC) 设计及其方法, E-mail: zhanglijun@suda.edu.cn.