

基于改进 YOLO11n 的芯片微小缺陷检测方法

汪涛¹

WANG Tao

摘要

针对芯片表面微小缺陷尺寸小、边界弱、背景复杂,易出现误检和漏检等问题,文章提出一种基于改进 YOLO11n 的轻量化检测方法。该方法引入特征互补感知结构、SPPF-LSKA 模块以及 MPDIoU 与 NWD 组合损失函数,以提升模型对芯片污点、芯片破损和芯片裂纹等微小缺陷的检测能力。实验结果表明,改进模型的精确率较原 YOLO11n 提升了 2%,mAP 提升了 1%,参数量下降约 64%,FLOPs 为 6.4×10^9 ,帧率达到 119 帧/s。与原 YOLO11n 相比,所提方法在提高检测精度的同时降低了模型规模,能够满足芯片表面微小缺陷实时检测需求。

关键词

芯片表面缺陷检测; YOLO11n; 微小目标检测; 轻量化

doi: 10.3969/j.issn.1672-9528.2026.07.032

0 引言

芯片表面质量直接关系到器件性能和运行可靠性。随着半导体制造与封装工艺不断向精细化方向发展,表面缺陷检测已成为芯片质量控制中的关键环节^[1]。但在实际检测过程中,芯片表面缺陷往往尺寸较小,且容易受边界模糊、对比度较低以及背景纹理复杂等因素影响。传统人工检测方法效率较低,主观性较强;常规机器视觉方法又较依赖人工经验和手工特征设计,在面对微小缺陷时容易出现误检、漏检等问题^[2]。因此,研究一种能够兼顾检测精度、检测速度和模型规模的芯片表面微小缺陷检测方法,对提升芯片检测效率和产品质量具有实际意义。

近年来,深度学习目标检测算法在工业缺陷检测领域得到了广泛应用,主要包括 R-CNN 系列两阶段检测模型^[3],以及 SSD、YOLO 系列等一阶段检测模型^[4-5]。其中,两阶段模型通常具有较高的检测精度,但网络结构相对复杂,推理速度较慢;一阶段模型结构更加简洁,检测效率较高,更适合实时检测场景。在芯片表面缺陷检测方面,Wang 等^[6]提出了结合空间注意力机制的 YOLOV4-SA 方法,用于芯片表面微小缺陷检测,并构建了包含真实样本和合成样本的芯片表面缺陷数据集。然而,在轻量化部署和实际工业检测环境中,微小缺陷仍容易受到弱边界、复杂纹理和噪声干扰的影响,模型检测稳定性仍有进一步提升空间。

围绕芯片表面缺陷检测任务,已有研究主要从注意力机制、特征融合、损失函数优化和模型轻量化等方面对 YOLO 系列算法进行改进。杨桂华等^[7]基于 YOLOX 构建 YOLOX-

PRO 网络,在主干网络中引入 CA 注意力模块,并采用 Var-iFocal 损失函数,提高了 QFN 芯片缺陷检测的精度与速度。栾方军等^[8]针对芯片表面微小缺陷易误检、漏检的问题,对 YOLOv8 进行改进,引入轻量级 C2f_WT 模块、多尺度边缘增强模块、空间通道特征融合模块和动态上采样模块,增强了模型对缺陷边缘和局部纹理信息的表达能力。孙佩月等^[9]以 YOLO11n 为基础,引入轻量级自适应提取卷积、DySample 上采样模块和多维协作注意力机制,在降低模型计算成本的同时提升了检测性能。上述研究表明,YOLO 系列模型在芯片表面缺陷检测中具有较好的应用基础,但在微小缺陷特征提取、复杂背景干扰抑制以及边界框定位稳定性等方面仍存在优化空间。

基于以上分析,芯片表面微小缺陷检测本质上属于复杂背景下的小目标检测问题。模型不仅需要保持较高的检测精度,还应满足实时检测和轻量化部署的要求。为此,本文在 YOLO11n 基础上进行改进,提出一种面向芯片表面微小缺陷检测的轻量化目标检测算法。该算法主要从网络结构、注意力机制和损失函数三个方面进行优化:首先,在网络结构中引入由特征互补映射模块和多核感知单元组成的特征互补感知结构,并基于该结构对特征提取与检测分支进行调整,使原 YOLO11n 的三检测头结构简化为双检测头结构;其次,在特征融合部分,将原 SPPF 模块替换为 SPPF-LSKA 模块,以增强模型对关键缺陷区域的特征响应;最后,在边界框回归阶段引入 MPDIoU+NWD 组合损失函数,以提高微小缺陷定位的稳定性。通过上述改进,模型在检测精度、参数量和推理效率之间取得了较好平衡,能够更好地满足芯片表面微小缺陷实时检测的应用需求。

1. 武汉轻工大学 湖北武汉 430048

1 芯片表面微小缺陷检测模型构建

针对芯片表面微小缺陷检测任务, 本文以 YOLO11n 为基础构建轻量化目标检测算法。为增强模型对芯片污点、破损和裂纹等微小缺陷的识别能力, 在网络结构中引入由特征互补映射模块 (FCM) 和多核感知单元 (MKP) 组成的特征互补感知结构, 通过多分支特征提取与互补融合增强模型对微小缺陷局部细节的表达能, 并基于该结构对检测分支进行调整, 将原 YOLO11n 的三检测头结构简化为双检测头结构, 去除面向大目标的检测分支, 从而降低模型参数量和计算复杂度; 在特征融合部分, 将原 SPPF 模块改进为 SPPF-LSKA 模块, 并引入 LSKA^[10] 注意力机制, 以扩大模型的有效感受野, 增强关键缺陷区域的特征响应; 在边界框回归阶段, 引入 MPDIoU 与 NWD 组合损失函数, 以减弱微小缺陷位置偏移对回归结果的影响, 提高边界框定位稳定性。改进后 YOLO11n 的网络结构如图 1 所示。

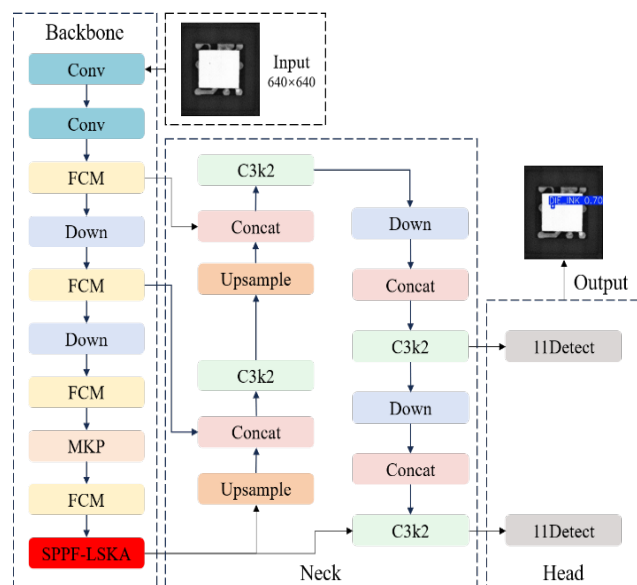


图 1 改进后 YOLO11n 网络结构

1.1 特征互补感知结构

特征互补感知结构由特征互补映射模块 (FCM) 和多核感知单元 (MKP) 组成, 其结构如图 2 所示。FCM 首先对输入特征进行通道划分, 形成不同特征分支。其中, 3×3 卷积分支主要用于提取局部纹理和边缘细节; 1×1 卷积分支则用于调整通道信息并保留有效原始特征。随后, 模块引入通道注意力和空间注意力, 分别强化缺陷相关通道响应和关键空间位置特征, 并通过 1×1 卷积对多分支特征进行融合, 从而增强模型对微小缺陷细节的表达能

力。MKP 采用 DW-Conv 与 PW-Conv 交替连接的方式, 在减少参数量和计算量的同时扩大特征感受范围, 使网络能够充分地捕捉芯片污点、破损和裂纹等小目标缺陷的细节信息。整体来看, 该结构在提升特征表达能力的同时兼顾了轻量化需求, 有助于提高模型检测精度和推理效率。由于本文检测对象主要为芯片表面微小缺陷, 目标尺度整体较小, 因此保留浅层与中层特征检测分支, 去除面向大目标的检测分支, 以降低冗余计算量。该设计能够在减少参数量的同时保留对微小缺陷边缘、纹理和局部形态的表征能力。

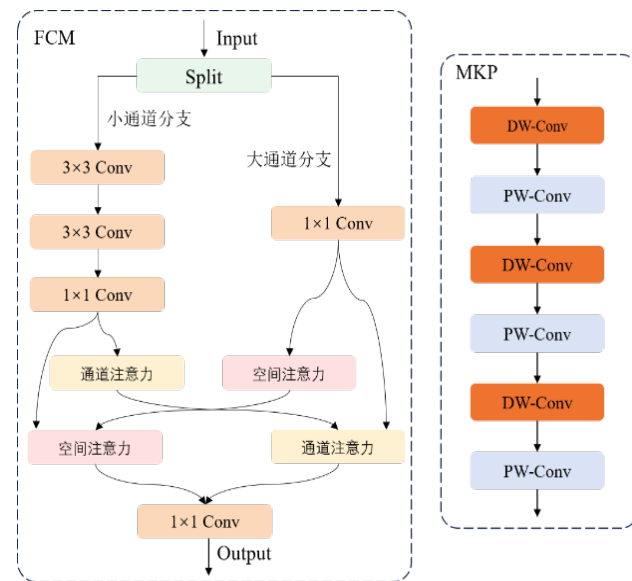


图 2 特征互补感知结构图

1.2 SPPF-LSKA 模块

SPPF-LSKA 模块是在原 SPPF 结构的基础上引入大核可分离注意力机制形成的, 整体结构如图 3 所示。输入特征首先经过卷积完成通道调整, 随后通过连续最大池化获得不同尺度的上下文信息, 并利用特征拼接实现多尺度融合。在此基础上, 引入 LSKA 注意力机制对融合后的特征进行加权, 使网络能够更集中地响应芯片缺陷所在区域, 减弱表面纹理和背景噪声等无关信息的干扰。相比原 SPPF 模块, SPPF-LSKA 在保留多尺度特征提取能力的同时进一步扩大了

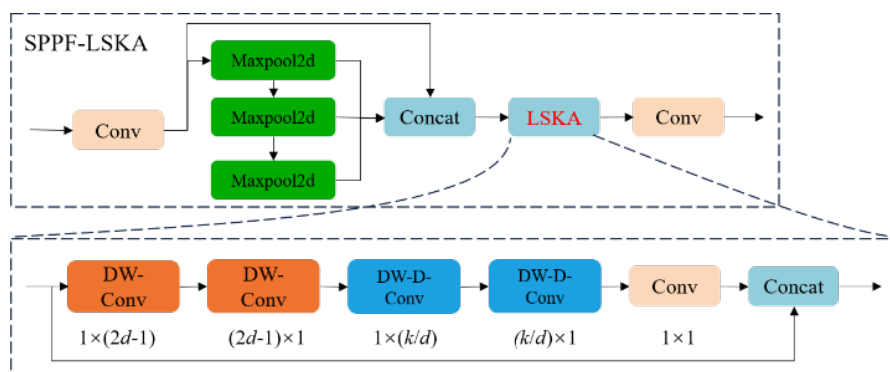


图 3 SPPF-LSKA 结构图

有效感受野，增强了局部细节与全局空间信息的表达能力，有助于提高模型对芯片污点、破损和裂纹等微小缺陷的检测精度和稳定性。

1.3 损失函数融合改进

为提高模型对芯片表面微小缺陷的定位精度，本文对 YOLO11n 原有检测头的边界框回归损失函数进行改进，将其替换为 MPDIoU+NWD 组合损失。改进后的检测头结构如图 4 所示，其中分类分支仍采用原有分类损失，用于完成芯片污点、芯片破损和芯片裂纹等缺陷类别判别；边界框回归分支采用 MPDIoU+NWD 损失，用于优化预测框与真实框之间的位置匹配关系。该改进不改变检测头的基本输出结构，能够在保持模型轻量化的同时，增强模型对微小缺陷区域的定位能力。

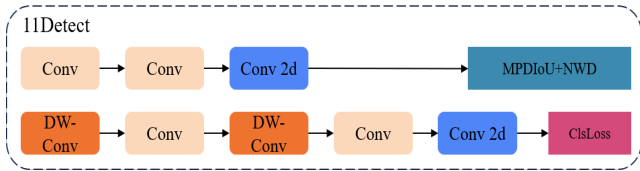


图 4 MPDIoU+NWD 损失函数结构图

MPDIoU Loss 与小目标 IoU 变化关系如图 5 所示。MPDIoU Loss 在传统 IoU 的基础上引入最小点距离约束，不仅考虑预测框与真实框之间的重叠面积，还进一步约束二者对应角点之间的位置偏差。其计算公式为：

$$L_{\text{MPDIoU}} = 1 - \text{MPDIoU} \quad (1)$$

$$\text{MPDIoU} = (\text{pre} \cap \text{gt}) / (\text{pre} \cup \text{gt}) - d_1^2 / (w^2 + h^2) - d_2^2 / (w^2 + h^2) \quad (2)$$

$$d_1^2 = (x_1^{\text{pre}} - x_1^{\text{gt}})^2 + (y_1^{\text{pre}} - y_1^{\text{gt}})^2 \quad (3)$$

$$d_2^2 = (x_2^{\text{pre}} - x_2^{\text{gt}})^2 + (y_2^{\text{pre}} - y_2^{\text{gt}})^2 \quad (4)$$

式中：pre 表示预测框，gt 表示真实框， d_1 和 d_2 分别表示预测框与真实框对应角点之间的距离， w 和 h 表示外接矩形的宽和高。

通过角点距离约束，MPDIoU 能够增强边界框回归过程中的位置约束能力，使预测框更加准确地靠近真实缺陷区域。

但芯片表面缺陷通常属于微小目标。当预测框与真实框发生相同程度的位置偏移时，大目标框的重叠面积变化较小，IoU 下降不明显；而小目标由于自身面积较小，轻微偏移就会导致重叠区域明显减少，甚至造成 IoU 急剧下降。这说明 IoU 对小目标位置偏差较为敏感，容易造成损失值波动，从而影响模型训练稳定性。

为缓解上述问题，本文进一步引入 NWD 损失。NWD 将边界框建模为二维高斯分布，并通过分布距离衡量预测框与真实框之间的相似程度，其计算公式为：

$$W_2^2(N_{\text{pre}}, N_{\text{gt}}) = \left\| \begin{bmatrix} x_c^{\text{pre}}, y_c^{\text{pre}}, w_{\text{pre}}/2, h_{\text{pre}}/2 \end{bmatrix}^T - \begin{bmatrix} x_c^{\text{gt}}, y_c^{\text{gt}}, w_{\text{gt}}/2, h_{\text{gt}}/2 \end{bmatrix}^T \right\|_2^2 \quad (5)$$

$$\text{NWD}(N_{\text{pre}}, N_{\text{gt}}) = \exp \left(-\frac{\sqrt{W_2^2(N_{\text{pre}}, N_{\text{gt}})}}{C} \right) \quad (6)$$

$$L_{\text{NWD}} = 1 - \text{NWD}(N_{\text{pre}}, N_{\text{gt}}) \quad (7)$$

式中： x_c 、 y_c 分别表示边界框中心点的横、纵坐标， w 和 h 分别表示边界框宽度和高度， C 表示归一化常数。与 IoU 类损失相比，NWD 对微小目标的轻微位置偏移更加平滑，能够缓解小目标因少量偏移导致 IoU 剧烈变化的问题。

最终，本文将 MPDIoU Loss 与 NWD Loss 进行加权融合，组合损失函数表示为：

$$\text{Loss} = \alpha L_{\text{MPDIoU}} + (1 - \alpha) L_{\text{NWD}} \quad (8)$$

式中： α 为权重系数，用于平衡 MPDIoU 和 NWD 在边界框回归中的作用。MPDIoU 能够强化预测框与真实框之间的几何位置约束，NWD 能够提高微小目标回归过程的稳定性。二者结合后，既提升了边界框定位精度，又缓解了小目标 IoU 波动问题，更适用于芯片污点、芯片破损和芯片裂纹等微小缺陷检测任务。

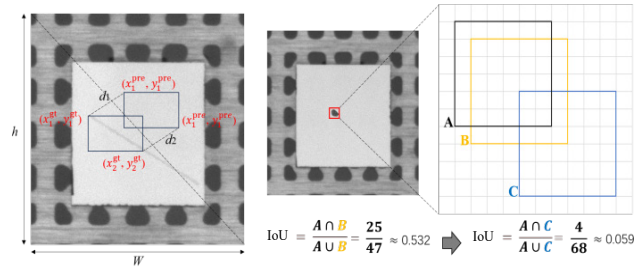


图 5 MPDIoU Loss 与小目标 IoU 变化分析图

2 实验及结果分析

2.1 实验环境与数据

实验在 Ubuntu 22.04.3 LTS 系统下进行，平台配置为 AMD EPYC 7502 32-Core 处理器、32 GB 内存和 NVIDIA A10G 显卡，显存容量为 24 GB。模型训练过程中采用的主要超参数如表 1 所示。

表 1 训练超参数设置

名称	参数
初始学习率	0.01
学习率动量	0.937
权重衰减系数	0.000 5
网络优化器	SGD
批次大小	16
迭代轮次	200

本文采用 Zhu 等^[11]构建的公开芯片表面缺陷数据集作为实验数据来源。该数据集共包含 10 761 张图像,涵盖真实无缺陷样本、真实缺陷样本以及合成缺陷样本。结合本文研究内容,从中选取芯片污点、芯片破损和芯片裂纹三类微小缺陷作为检测对象。由于这三类缺陷普遍具有尺寸小、形态不规则等特点,且检测过程容易受到表面纹理、成像噪声和复杂背景的干扰,因此本文对筛选后的图像进行了数据增强处理,包括高斯噪声、椒盐噪声、随机旋转以及亮度和对比度调整等。经过增强后,样本数量由 1 490 张扩展至 5 960 张,并按照 7:2:1 的比例划分为训练集、验证集和测试集。三类芯片表面微小缺陷的典型样例如图 6 所示。

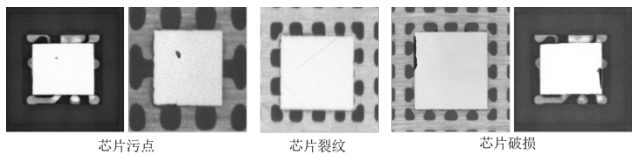


图 6 芯片表面缺陷图

2.2 评价指标

为保证评价结果的可比性,本文在相同实验条件下对改进前后的模型进行测试,并选取精确率、召回率、mAP 和帧率作为主要评价指标。其中,精确率用于反映模型检测结果的准确性,召回率用于衡量模型对目标缺陷的检出能力,mAP 用于综合评价模型的整体检测性能,帧率则用于表征模型的检测速度。此外,本文还结合参数量和 FLOPs 对模型的轻量化程度进行分析。

2.3 对比实验

为验证本文方法的有效性,在相同训练参数、输入尺寸和实验环境下,将改进后 YOLO11n 与 YOLO10n、YOLO11n、YOLO12n 和 YOLO13n 等轻量级目标检测模型进行对比,如表 2 所示。

由表 2 可知,改进后 YOLO11n 的精确率达到 95.4%,mAP 达到 94.7%,整体性能优于其他对比模型。与原 YOLO11n 相比,改进模型在精确率和 mAP 上均有所提升,召回率保持一致;同时,参数量由 2.58×10^6 降至 0.93×10^6 ,FLOPs 为

6.4×10^9 ,模型规模进一步减小。虽然每秒帧数相比原模型有所下降,但仍达到 119,能够满足实时检测需求。综合来看,改进模型在保持检测效率的同时,提高了检测准确性,并降低了模型复杂度。

2.4 消融实验

由表 3 可知,A、B、C 分别表示特征互补感知结构、SPPF-LSKA 模块和 MPDIoU+NWD 组合损失函数。单独加入特征互补感知结构后,模型参数量由 2.58×10^6 降至 0.93×10^6 ,轻量化效果较为明显,但精确率、召回率和 mAP 略有下降,说明该结构在压缩模型规模的同时,对特征表达仍需进一步补充。继续引入 SPPF-LSKA 模块后,精确率和 mAP 分别提升至 94.0% 和 94.1%,检测性能开始优于原 YOLO11n。进一步加入 MPDIoU+NWD 组合损失函数后,模型精确率达到 95.4%,mAP 达到 94.7%,均为各组实验中的最高值。虽然最终模型的帧率相比原模型有所下降,但仍保持在 119,能够满足实时检测需求。整体来看,三个改进模块分别在模型轻量化、特征增强和定位优化方面发挥作用,组合后模型取得了更好的综合检测效果。

2.5 实验效果可视化比较

如图 7~8 所示,原模型的热力图响应范围较为分散,部分注意力落在芯片边缘、背景纹理等无关区域,对真实缺陷位置的关注不够集中。相比之下,改进后模型的高响应区域主要分布在红色标注框附近,与芯片污点、破损和裂纹等缺陷位置更加吻合。从检测结果来看,原模型在部分样本中存在误检、重复检测和漏检现象,而改进后模型的检测框与目

表 2 不同模型的实验结果

模型	精确率 /%	召回率 /%	mAP /%	参数量 / 10^6	FLOPs / 10^9	FPS / (帧 $\cdot$ s $^{-1}$)
YOLO-10n	93.0	86.8	93.5	2.27	6.5	108
YOLO-12n	93.8	87.6	93.4	2.51	5.8	126
YOLO-13n	94.0	87.8	94.0	2.45	6.2	138
YOLO-11n	93.4	88.4	93.7	2.58	6.3	131
改进后 YOLO	95.4	88.4	94.7	0.93	6.4	119

表 3 模型消融实验

模型	精确率 /%	召回率 /%	mAP /%	参数量 / 10^6	FLOPs / 10^9	FPS / (帧 $\cdot$ s $^{-1}$)
YOLO11n	93.4	88.4	93.7	2.58	6.3	131
YOLO11n+A	92.8	87.9	93.1	0.81	6.0	106
YOLO11n+A+B	94.0	88.5	94.1	0.93	6.4	119
YOLO11n+A+B+C	95.4	88.4	94.7	0.93	6.4	119

标区域匹配程度更高,置信度也有所提升。由此可见,改进模型能够更有效地抑制背景干扰,减少误检和漏检,对微小缺陷具有更稳定的检测效果。

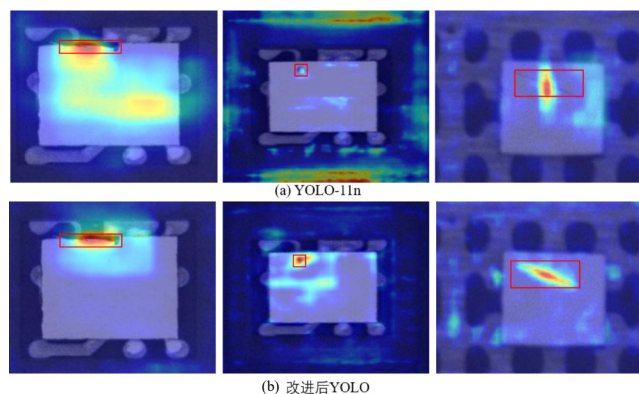


图7 热力图对比

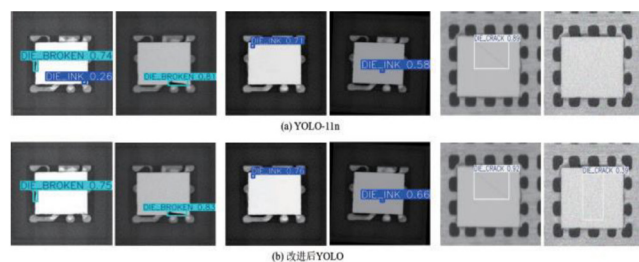


图8 检测结果图对比

3 结论

本文针对芯片表面微小缺陷检测任务提出了一种基于改进YOLO11n的轻量化目标检测算法。该算法引入特征互补感知结构、SPPF-LSKA模块、双检测头结构以及MPDI-oU+NWD组合损失函数,以提升模型对芯片污点、芯片破损和芯片裂纹等微小缺陷的检测能力。实验结果表明,改进模型的精确率由93.4%提升至95.4%,mAP由93.7%提升至94.7%,参数量由 2.58×10^6 降至 0.93×10^6 ,每秒帧数达到119,能够满足实时检测需求。总体来看,本文方法在检测精度和模型轻量化方面取得了较好平衡。后续可从困难样本增强、类别不平衡优化和实际工业场景部署等方面进一步改进。

参考文献:

- [1] 王新宇,蒋三新.芯片缺陷检测综述[J].现代制造技术与装备,2022,58(5):94-98.
- [2] 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.
- [3] Ren Shaoqing, He Kaiming, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149.

- [4] Liu Wei, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C]//Computer Vision-ECCV 2016. Berlin:Springer,2016: 21-37.
- [5] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2016: 779-788.
- [6] Wang Shuo, Wang Hongyu, Yang Fan, et al. Attention-based deep learning for chip-surface-defect detection[J]. The International Journal of Advanced Manufacturing Technology, 2022, 121: 1957-1971.
- [7] 杨桂华,吴振生,杨子康.基于YOLOX的QFN芯片表面缺陷检测技术研究[J].传感器与微系统,2025,44(3): 46-49.
- [8] 栾方军,马明乐,袁帅.改进YOLOv8的芯片表面微小缺陷检测算法[J/OL].计算机测量与控制,1-12[2026-07-21]. <https://link.cnki.net/urlid/11.4762.TP.20260310.1039.002>.
- [9] 孙佩月,黄娟,顾寄南,等.基于YOLO11n的轻量化芯片表面缺陷检测方法[J].半导体技术,2025,50(7):707-713.
- [10] Lau K W, Po L M, Ur Rehman Y A. Large separable kernel attention: rethinking the large kernel attention design in CNN[J]. Expert Systems with Applications, 2024, 236: 121352.
- [11] Zhu Xiaoming, Wang Shuo, Su Junyu, et al. High-speed and accurate cascade detection method for chip surface defects[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 2506612.

【作者简介】

汪涛(2002—),男,湖北荆州人,硕士,研究方向:深度学习、目标检测。

(收稿日期:2026-06-22 修回日期:2026-07-21)