DOI: 10.13382/j.jemi.B2205613

基于深度学习的高精度晶圆缺陷检测方法研究*

史浩琛 金致远 唐文婧 王 静 蒋 楷 夏 伟
(济南大学物理科学与技术学院 济南 250022)

摘 要:为了解决半导体制造领域缺陷检测中出现的检测效率低、错误率高、结果不稳定、成像精度低下导致无法精确地检测出不同种类的缺陷等问题,本文利用定制的 CCD 工业相机搭配高倍率的光学显微镜采集晶圆表面的扫描图像,结合改进的 YOLOv4 算法,实现了基于深度学习的高精度晶圆缺陷检测方法。实验表明,对于碳化硅晶圆缺陷,提出的方法模型可以识别各种复杂条件下的不同种类缺陷,具有良好的鲁棒性。对缺陷的平均识别精度达到 99.24%,相较于 YOLOv4-Tiny 和原 YOLOv4 分别提升 10.08%和 1.92%。对缺陷的平均每图识别时间达到 0.028 3 s,相较于基于 Halcon 软件方法和 OpenCV 模板匹配方法分别提升 93.42%和 90.52%,优于其他常规的晶圆缺陷检测方法,已实现在自主设计的验证系统和应用平台上稳定运行。

关键词:深度学习;晶圆缺陷检测;碳化硅晶圆;YOLOv4
中图分类号:TP391.41 **文献标识码**:A **国家标准学科分类代码**:510.4050

Research on high precision wafer defect detection based on deep learning

Shi Haochen Jin Zhiyuan Tang Wenjing Wang Jing Jiang Kai Xia Wei
(School of Physics and Technology, University of Jinan, Jinan 250022, China)

Abstract: In order to solve the semiconductor manufacturing defect detection with low efficiency, the error rate is high, the result is not stable, imaging accuracy is low and cannot accurately detect the problem such as different kinds of defects. In this paper, by using a custom CCD industrial camera with a high ratio of optical microscope scan images on the surface of the wafer, combined with the improved YOLOv4 algorithm, a high precision wafer defect detection method based on deep learning is implemented. Experimental results show that the proposed model can identify different kinds of silicon carbide wafer defects under various complex conditions and has good robustness. The average accuracy of defect identification is 99.24%, which is about 10.08% and 1.92% higher than that of YOLOV4-Tiny and original YOLOv4, respectively. Compared with the Halcon-based method and OpenCV template matching method, the average recognition time of defects per graph reaches 0.028 3 s, which is about 93.42% and 90.52% higher than other conventional wafer defect detection methods and has realized stable operation in independently designed verification systems and application platform.

Keywords: deep learning; wafer defect detection; silicon carbide wafer; YOLOv4

0 引 言

半导体产业的发展与我国各个高精尖科技领域的进步息息相关。本文以第 3 代半导体材料碳化硅为主要研究对象,在碳化硅晶体生长的过程当中,晶体缺陷的产生会导致最后的器件性能受到很大的影响,抑制缺陷的产生成为晶体生长过程中最重要的工作,如何快速识别并区分不同种类的碳化硅晶圆缺陷就成为了非常关键的环节。

由于原子绝对严格按照晶格的周期性排列的晶体是不存在的,故在实际中,晶体都会或多或少存在不同程度的缺陷[1]。在晶体生长、化学气相沉积等过程都可能使晶圆表面产生缺陷。在晶圆制造的缺陷类型中,晶圆表

收稿日期:2022-06-22 Received Date: 2022-06-22
*基金项目:国家自然科学基金(62005094)、山东省自然科学基金(ZR2021MF128)、济南市引进创新团队项目(2018GXRC011)、山东省工业技术研究院协同创新中心共建项目(CXZX2019007)资助

面冗余物、晶体缺陷等属于较普遍的问题。本文研究的对象主要是碳化硅晶圆的晶体缺陷,分为基平面位错、穿透型螺位错、穿透型刃位错和微管 4 类进行检测。

目前晶圆缺陷检测技术通过人工检测或基于机器视觉检测的两种方式进行。人工检测存在工作效率低下,检测精度得不到保障等问题。基于机器视觉的检测利用轮廓提取、裁剪和形态学变换的方法处理后,通过模板匹配进行检测。Tsai 等[2]使用傅里叶变化重构图像来检测缺陷,但仅能检测高频部分特定形态的缺陷。Liu 等[3]引入光谱作差和模板匹配结合的方法检测缺陷,但仅能检测边缘,并且无法处理遮挡或光照造成的一系列干扰问题。基于机器视觉的检测方法可以采用 Halcon 软件或 OpenCV 库很好的实现,但普遍存在易受环境影响、检测范围小、检测结果不稳定且检测速度慢等问题。这些都使得深度学习和图像处理结合的检测方法成为了未来的趋势。基于以上的分析,本文针对碳化硅晶圆不同的缺陷类型特点,提出基于深度学习的高精度晶圆缺陷检测方法。

基于深度学习的目标检测算法在计算机视觉领域占有十分关键的地位,范围十分广泛,如目标跟踪、交通检测、人脸识别、实时监控、无人驾驶等[4]。Redmon 等[5]提出了 YOLO(you only look once)算法研究,并开辟了目标检测方法的新思路,YOLO 的核心思想就是利用整张图作为网络的输入,把目标检测转变成一个包含类别信息的空间位置回归问题[6]。YOLO 做到了极快的检测速度,能够很好的满足工业作业环境下,实时和快速的应用需求,成为了目前常用的目标检测算法。

本文提出的方法在检测方案上,首先通过定制的 CCD 工业相机和高倍率的光学显微镜,配合自主搭建可编程的电动载物滑台,采集晶圆表面的扫描图像发送给计算机。接着,利用 SIFT[7]算法对采集到的高精度显微图像进行特征点提取,利用 FLANN[8]和 RANSAC[9]算法进行图像的拼接与融合,合成输出一张高精度的全局晶圆缺陷图像。最后,将全局图像输入训练好的基于 YOLOv4[10]算法的卷积神经网络进行检测,并通过基于 QT 框架搭建的可视化程序输出最终检测结果。

与之前的工作相比,本文的主要贡献包括如下 3 个方面:

1)将深度学习与晶圆检测结合,相比较与人工检测和传统基于机器视觉的图像识别等的方法,检测速度快,识别准确度高,成本低,工作效率高,能极大程度的帮助科研人员快速分析晶圆缺陷问题,生成相应的改进工艺方案。

2)结合不同算法对图像进行特征点的提取,拼接与融合,获得的图像更为精细,解决高倍显微镜下无法看到全局高精度图像的情况,能更好地应对多种情况的晶圆缺陷检测。

3)针对碳化硅晶圆特性,对 YOLOv4 算法进行了改进,在图像输入网络前添加了预处理模块提升速度、在骨干网络添加了注意力机制模块、使用了改进的 K-Means 聚类算法、使用了 Swish 激活函数以及解耦合 YOLO-Head 检测头,使得神经网络算法能充分提取晶圆缺陷的多尺度特征,检测精度更准,误判更低,能更好的满足行业需求。

1 图像数据的预处理

由于图像在采集的过程中存在部分重合区域,需要对图像进行特征点匹配后合成全局图像的预处理操作。通过 SIFT 算法可以实现特征点的提取,全部提取完毕后,得到的每一个 SIFT 特征区域都会具有位置、尺度、方向 3 个信息。最后用一组向量把极值点描述起来,从而形成了一个描述符,使得描述符不能因为其他条件的改变而发生变化。生成的 SIFT 描述子是一个具有独特性和唯一性的向量,能够有效的提升正确匹配特征点的概率。将完成特征点提取的图像进行匹配操作,本文基于图像处理的拼接融合主要通过 FLANN、RANSAC 等算法,采用图像粗匹配和图像精匹配相结合的方法进行。

2 基于深度学习的晶圆缺陷检测算法

本文使用改进后的 YOLOv4 算法对晶圆缺陷进行检测,与标准 VOC 和 COCO 数据集相比,碳化硅晶圆的缺陷普遍存在背景复杂、干扰多、缺陷尺度变化不一、目标旋转、模糊等一系列问题,因此针对碳化硅晶圆缺陷的应用场景提出了几种改进方法。改进 YOLOv4 的结构图如图 1 所示。

2.1 骨干网络的改进

1)图像预处理输入模块改进

本文在将图像输入骨干网络的卷积层之间,加入了一个预处理模块,类似于 Swim Transformer[11]模型中的 Patch Partition 模块,图像输入模块后,进行分片处理,通过相似于邻近下采样的方法,在一张图片中分别对相邻间隔的像素点取值,并堆叠与图像相邻的 4 个区域,把 W、H 维度的信息集中到 C 通道空间,输入通道增大 4 倍,宽高减少 1/4,拼接出来的图片从原先的 3 通道变成了 12 个通道,即[H, W, C]→[H/4, W/4, C]。

加入该模型的主要思路是为了在高分辨率图中,通过周期性的提取像素点重构到低分辨率图像,在进行堆叠后增加每个点的感受野,从而降低对原始信号的损失,以便于合理的减小运算工作量,提高计算速度。

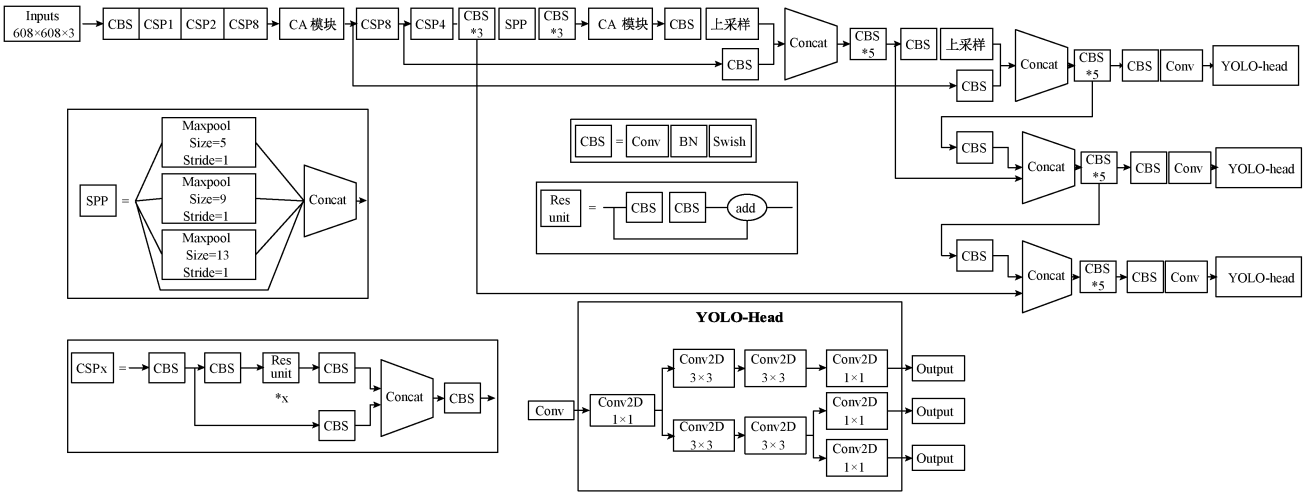


图 1　改进 YOLOv4 的结构图

Fig. 1　Structure diagram of improved YOLOv4

2)注意力机制模块改进

注意力机制在深度学习中应用广泛,注意力机制主要是起到对图像的关键信息进行提取的作用。简单来说,将神经网络的注意力不放在背景上,更多的是放在前景上的时候,神经网络模型能更好的学习到深层次的重要信息。

注意力机制的加入,会在一定程度上对神经网络带来一定的计算量。许多神经网络使用如 SE(squeeze-and-excitation)[12] 模块、CBAM(convolutional block attention module)[13]模块等机制。但 SE 一般只考虑内部的通道信息,并不重视空间位置,但是位置信息对于生成空间选择性特征图是很重要的。而 CBAM 可以看做只是在 SE 的基础上加入了空间注意力的改进模块。

本文结合实际应用场景,改进了注意力机制,使用一种名为 CA(coordinate attention)[14] 的模块。CA 是新提出的一种模块,使用两个一维的全局池化方法把 X 和 Y 两种不同方向通道的输入信息分别集合为两种相互独立的方向感知特征图。它不仅考虑了通道关系,还考虑了特征空间的位置信息。具体而言,就是将聚合后的特征图分别编码为注意力图像,每个注意力图像都沿一个空间的方向获取输入特征图的距离关系,因而位置信息可以被存储于生成的注意力图像中[14]。CA 模块整体结构图如图 2 所示。

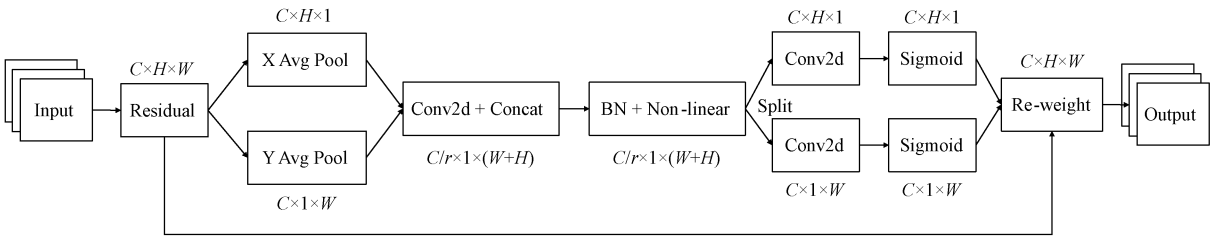


图 2　CA 模块结构图

Fig. 2　Structure of the CA module

在 CA 中,将全局池化操作进行分解,可以转化为一对一的特征编码操作,输入一个给定的特征张量 x_c,其第 c 通道的 squeeze 步长公式如下:

$$z_c = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j) \tag{1}$$

其中,z_c 是与第 c 个通道相关联的输出,输入的 x_c 直接来自一个固定内核大小的卷积层,可以看作是局部描述符的集合。H 为通道垂直高度,W 为通道水平宽度,通过尺寸为$(H,1)$或$(1,W)$的池化内核依次按照水平坐标系和垂直坐标系方向对各个通道进行了编码,高度为 h 的第 c 通道输出可以表示为:

$$z_c^h(h) = \frac{1}{W}\sum_{0\leqslant j<W} x_c(h,i) \tag{2}$$

同理,宽度为 w 的第 c 通道输出可以表示为:

$$z_c^w(w) = \frac{1}{H}\sum_{0\leqslant j<H} x_c(j,w) \tag{3}$$

接着进行第 2 个转换，也称为 CA 生成。该转换将之前的变换进行连接操作，使用 1×1 的卷积变换函数 F_1 进行，公式如下：

$$f = \delta(F_1([z^h, z^w])) \tag{4}$$

$$g^h = \sigma(F_h(f^h)) \tag{5}$$

$$g^w = \sigma(F_w(f^w)) \tag{6}$$

其中，f 表示在水平方向和垂直方向进行编码的特征映射，δ 表示非线性激活函数，$[z^h, z^w]$ 表示沿空间维数的连接操作。F_h 为垂直方向的 1×1 卷积变换函数，F_w 为水平方向的 1×1 卷积变换函数，f^h、f^w 表示 F_1 沿空间维数分解后的张量。g^h、g^w 表示把 f^h 和 f^w 转化为一个相同通道数的张量。σ 表示 sigmoid 激活函数。

最后，CA block 的输出 Y 表达式为：

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{7}$$

其中，y_c 表示第 c 通道输出 Y 方向的信息，x_c 表示第 c 通道给定的特征张量，$g_c{}^h$，$g_c{}^w$ 表示把第 c 通道的 f^h 和 f^w 转化为一个相同通道数的张量。

这样做的优点主要有：可以捕获跨通道的信息，从而使得模型更精准地定位并辨识目标区域；CA 模块灵活且轻量，可以很方便接入到经典网络中提高性能；对于有密集预测任务的预训练的模型来说，会有非常明显的性能提升。

本文分别在位于 SPP 模块之前，CSPlayer 之后和骨干网络末端进行上采样之前的部位嵌入 CA 注意力机制模块，嵌入模块后的改进结构图如图 3 所示。

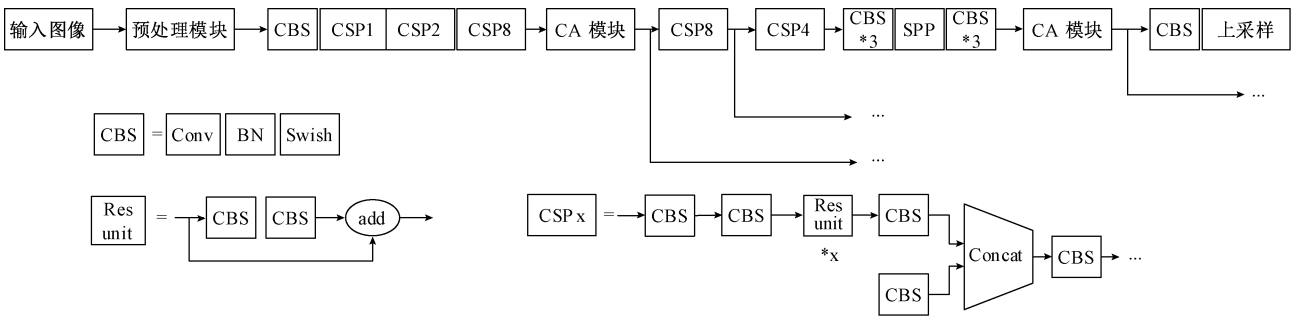


图 3 嵌入 CA 模块后的改进骨干网络结构

Fig. 3 Improved backbone network structure after CA module is embedded

2.2 激活函数改进

在原 YOLOv4 中，Backbone 部分使用了 Mish[15] 激活函数，Neck 部分使用了 Leaky ReLU[16] 激活函数。本文对激活函数进行了改进，上述两个部分的激活函数全部替换为 Swish[17] 激活函数进行计算，Swish 激活函数公式如下：

$$Swish(x) = x \times Sigmoid(\beta x) \tag{8}$$

其中，$Sigmoid(x) = 1/(1+e^{-x})$，Swish 激活函数可以用在慢速训练期间，由于激活函数饱和使得神经网络性能大幅度降低，进而产生梯度逐步减少最后趋于 0 的情形，并且 Swish 函数处处可导，连续光滑，平滑度在训练的优化与泛化上起着很大影响[17]。

与 Mish 和 Leaky ReLU 相比，Swish 在性能上稍弱于 Mish，强于 Leaky ReLU，在计算时间上强于 Mish，弱于 Leaky ReLU[17]。故本文在综合权衡性能和计算时间的基础上，选择使用 Swish 替换原有的激活函数。Swish 激活函数及其一阶导函数的图像如图 4 和 5 所示。

与此同时，本文也将空间金字塔池化结构（SPP）[18] 中的卷积层激活函数替换为 Swish，并且将 SPP 从 FPN 特征金字塔中剥离，融合到了 Backbone 的残差块（Resblock_body）当中，使 SPP 完全成为骨干特征提取网络中的一部分。

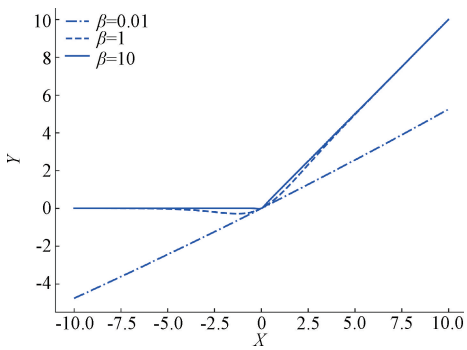


图 4 Swish 激活函数图像

Fig. 4 Swish activation function image

2.3 聚类算法改进

Anchor Box 的选取会直接对网络模型的训练效果产生影响，聚类分析中常采取 K-means[19] 聚类算法来对 Anchor Box 进行筛选。

进行 K-means 聚类算法计算之前，需采用欧氏距离平方作为样本之间的距离，即 $d(x_i, x_j) = \| x_i - x_j \|_2{}^2$，其中 x_i、x_j 表示样本。在进行 K-means 聚类时，实际上就是求解最优化问题，计算表达式如下：

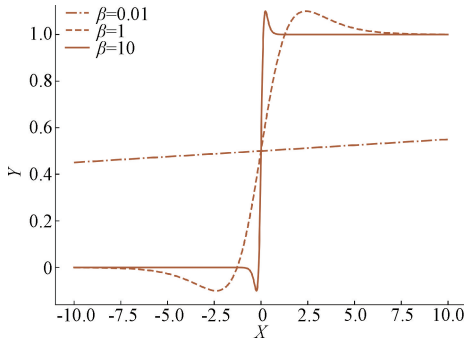


图 5　Swish 激活函数一阶导函数图像

Fig. 5　Image of the first derivative function of Swish activation function

$$\min E = \sum_{i=1}^{k} \sum_{x \in c_i} \| x - \boldsymbol{u}_i \|_2^2 \tag{9}$$

其中，k 表示要划分的集合个数，C_i 表示第 i 簇，x 表示样本，$\boldsymbol{u}_i$ 表示第 i 个簇的均值向量。进行 K-means++[20] 聚类时，寻找在未选中点与已选中点的最远距离点，其表达式为：

$$X_i = \operatorname{argmax}_{i=0}^{n}(\min_{j=1}^{m} \| x_i - c_j \|^2) \tag{10}$$

其中，x_i 表示第 i 簇的样本，c_j 表示第 j 个聚类的聚类中心，n 表示点数，m 表示被选中心数。

本文使用了一种改进的 K-means 聚类算法，该算法实际是在 K-means 和 K-means++的基础上进行了改进，对于之前算法中出现的一系列缺陷进行了解决，改进的 K-means 聚类算法主要的实现思路是：首先，随机抽取样本设定为候选的聚类中心。接着，计算各个样本到候选中心的距离，再根据概率选择样本替换聚类中心。按步骤循环 5 次后，可以得到比预设定大一些的候选集。最后，计算每个候选质心的分布程度。进行了上述主要步骤之后，再在候选质心的集合上执行有权重的 K-means++和 K-means 算法。

2.4　检测头改进

在原 YOLOv4 使用的 YOLO-Head 中，检测头是耦合是在一起将分类任务和回归任务一同实现的，实质上是分别进行 3×3 和 1×1 的卷积，3×3 卷积用于特征整合，1×1 卷积用于通道数调整[21]。然而，分类任务和回归任务在目标检测中是存在冲突关系的，因此，通过解耦合检测头，并将其代替原有的检测头能够较好地解决这个问题。将检测头解耦无疑会增加运算复杂度，但均在可接受的范围内。检测头解耦后能较好地加快模型的收敛速度、提高检测精度[21]。

本文先利用一个 1×1 的卷积进行降维处理后，接着分为两个 3×3 的卷积分支进行特征转化，分别进行分类任务和回归任务，分类任务的分支用于判断每一个特征点所包含的物体种类，通过一个 1×1 的卷积进行通道数的转换。回归任务的分支又会分为两个分支，分别用于判断每一个特征点的回归参数和判断每一个特征点是否包含物体，通过两个 1×1 的卷积进行通道数的转换。最后将得到的 3 个预测结果堆叠整合后输出，有效的避免了任务之间的冲突。改进后的检测头结构图如图 6 所示。

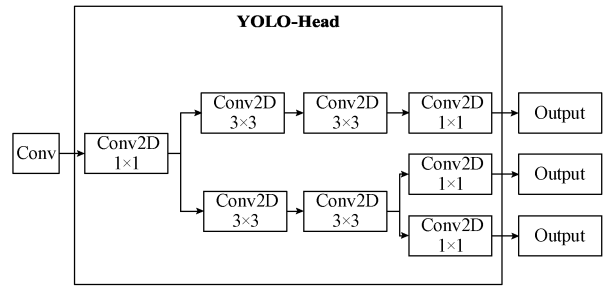


图 6　改进后的 YOLO-Head 模块结构图

Fig. 6　Structure diagram of improved YOLO-Head module

3　实验与分析

3.1　实验环境

本文所使用的实验硬件环境：图像采集平台为奥林巴斯 BX53(LED)显微镜，中央处理器(CPU)为 Inter(R) Core(TM) i7-12700KF @ 3.6 GHz，显卡(GPU)为 NVIDIA GeForce RTX 3080 12 GB，运行内存 32 GB。

本文所使用的实验软件环境：操作系统为 Windows 1164 位，深度学习框架为 Darknet 和 Pytorch，开发语言为 Python 和 C++，开发环境为 Microsoft Visual Studio 2017 和 Pycharm，其他软件环境为 Python 3.9、OpenCV 4.5、CUDA 11.5、cuDNN 8.3.2、CMake 3.21.3。

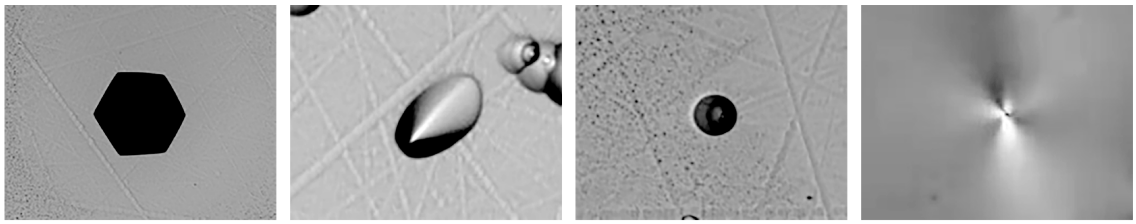
3.2　数据集

由于碳化硅晶圆缺陷图像的数据集比较特殊，针对性强，没有可以在公开渠道能够获取的数据集，故本文实验中使用的数据集是自主建立的，使用高精度显微镜在不同放大倍数下，采集了大量不同种类的碳化硅晶圆缺陷，主要分为基平面位错、穿透型螺位错、穿透型刃位错和微管 4 个大类，数据集按照 VOC 2007 的经典结构进行部署，进行数据增强和扩充处理后，共计 2 160 张晶圆缺陷图片数据，将其按照训练集：测试集：验证集 = 6：2：2 的比例进行数据集的划分。各种缺陷种类如图 7 所示。数据集的划分和各缺陷种类的数量如表 1 和 2 所示。

表 1　数据集的划分

Table 1　Division of dataset

训练集	测试集	验证集	合计
1 296	432	432	2 160



(a) 穿透型螺位错 (a) Penetrating screw dislocation
(b) 基平面位错 (b) Base plane dislocation
(c) 穿透型刃位错 (c) Penetrating edge dislocation
(d) 微管 (d) Micropipe

图 7 4 种碳化硅晶圆缺陷类型

Fig. 7 Four silicon carbide wafer defect types

数据集中各个种类目标的数量如表 2 所示。其中，基平面位错、穿透型螺位错、穿透型刃位错的数量较多，微管的数量较少。

表 2 数据集中各缺陷种类的数量

Table 2 Number of defect types in the dataset

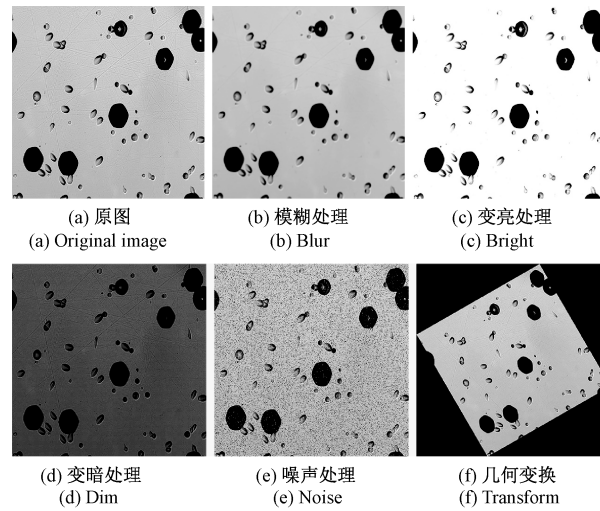
晶圆缺陷类型	目标数量
穿透型螺位错	12 730
基平面位错	23 930
穿透型刃位错	45 050
微管	21 060

3.3 数据增强及参数设置

虽然上文中自主建立的数据集已经具有了一定的数量规模，但相比于大型数据集，如 VOC 和 COCO 等，数据的数量和丰富程度还是有所欠缺的。为了在不同情况下都能保证模型的精度和鲁棒性，以及避免大量相似图像输入训练后造成的过拟合问题，对自主建立数据集的图像数据进行了数据的扩充与增强。主要采用的方式有：翻转增强，包括将图像进行水平、垂直和镜像翻转；旋转增强，包括多个角度旋转扩充；缩放增强，包括图像的放大或缩小，并按原始尺寸进行裁剪；高斯噪声增强，可以有效减弱高频特征失真的影响；亮度和对比度增强。数据增强的部分示例如图 8 所示。

参数设置方面，基于原始 YOLOv4 的权重模型在 Darknet 框架下进行训练，实验设置的输入图片大小为 608×608，batch 大小为 64，subdivisions 大小为 32，class 设置为 4，filters 设置为(4+5)×3=27，max_batches 设置为 6 000 次，steps 步长分别取最大迭代次数的 80%和 90%，同时使用 cutmix 和 mosaic 等数据增强，提升网络训练效果。

为了方便算法改进，将 YOLOv4 算法在 TensorFlow2.0 框架上进行复现，并加入第 2 节所述改进模块的相关内容，在此基础上进行改进模型的训练，实验设置的输入图片大小为 640×640，总 epoch 为 200，训练分为两个阶段，分别是冻结阶段和解冻阶段。在冻结阶段，不启用模型的骨干，特征提取网络不发生改变，仅对



(a) 原图 (a) Original image
(b) 模糊处理 (b) Blur
(c) 变亮处理 (c) Bright
(d) 变暗处理 (d) Dim
(e) 噪声处理 (e) Noise
(f) 几何变换 (f) Transform

图 8 数据增强的部分示例

Fig. 8 Shows a partial example of data enhancement

网络进行微调，此时执行前 50 次 epoch，batch_size 为 8。在解冻阶段，启用模型的骨干，特征提取网络发生改变，网络所有的参数都会发生改变，此时接着执行 epoch 到 200，batch_size 为 4。在训练过程中，使用 SGD 优化器，模型的最大学习率设置为 0.01，为了防止过拟合现象，权值衰减设置为 5×10^{-4}，同时使用 mixup 和 mosaic 等数据增强。

3.4 数据增强及参数设置

1) 评价指标

针对本文所使用的基于改进 YOLOv4 算法的卷积神经网络模型进行性能评价，采用目标检测模型的经典参数评估指标，计算公式如下：

$$P = \frac{T_P}{T_P + F_P} \tag{11}$$

$$R = \frac{T_P}{T_P + F_N} \tag{12}$$

$$AP = \int_0^1 P(R)\,\mathrm{d}R \tag{13}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{14}$$

其中，P 表示精确率，R 表示召回率，AP 表示单类别平均精度，mAP 表示各类别平均精度，N 表示检测的缺陷类别数目，T_P 表示预测正确的样本数量，即 IoU（交并比）>0.5 的检测框数量，F_P 表示将错误样本预测为正确样本的数量，即 IoU<=0.5 的检测框，F_N 将正确样本预测为错误样本的数量，即没有检测到的 Ground Truth 的数量，IoU=0。

2）训练结果评价

模型训练的总损失和各类别平均精确度如图所示。使用 NVIDIA RTX 3080 型号的 GPU 经过 200 个 epoch 后，损失函数趋于收敛状态，波动变化的幅度在较小的范围内，精确率（Precision）和召回率（Recall）都达到 99%以上，AP 和 mAP 到达 98%以上，表明 epoch、batch_size、steps、weight_decay 和学习率等参数设置合理，模型的最终训练效果较好，精确度高。

训练模型的 Loss 曲线图、AP 曲线图、精确率曲线图以及召回率曲线图如图 9、10、11 和 12 所示。

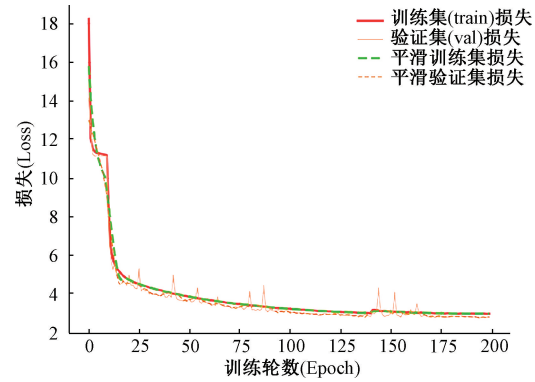


图 9　训练模型的 Loss 曲线

Fig. 9　Loss curve of the training model

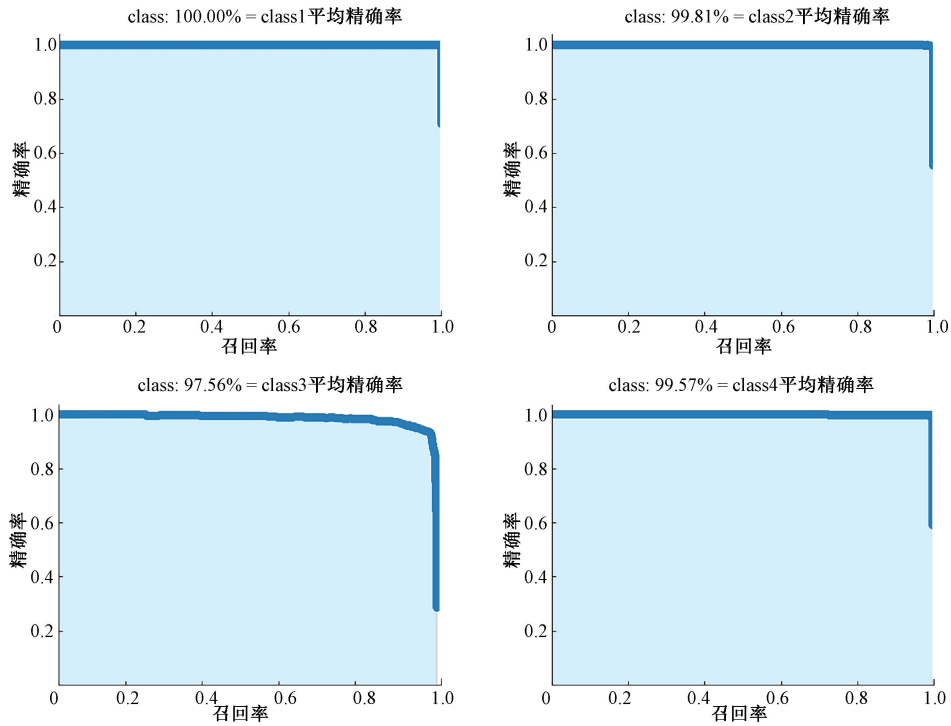


图 10　训练模型的 AP 曲线

Fig. 10　AP curve of the training model

3）检测结果评价

晶圆缺陷检测任务流程图如图 13 所示。

在完成训练后，将采集到的图片手动输入模型进行测试，如图 14~17 所示，在不同情况下对 4 种不同的缺陷进行检测，仅有少量的缺陷被漏检，模型总体识别率高，效果好，能满足实际的检测需求。

为了方便操作，本文将训练好的改进模型封装到独立软件，使用 QT 框架搭建出可视化操作界面，并结合 CCD 工业相机的控制操作，对图像采集后进行拼接融合与检测，最后输出检测结果，实现在软件端的一体化工作流程。基于 QT 搭建的晶圆缺陷检测识别软件系统如图 18 所示。

本文将自建数据集放入不同的检测算法模型进行了对比测试，包括轻量级 YOLOv4-Tiny 算法，原 YOLOv4 算

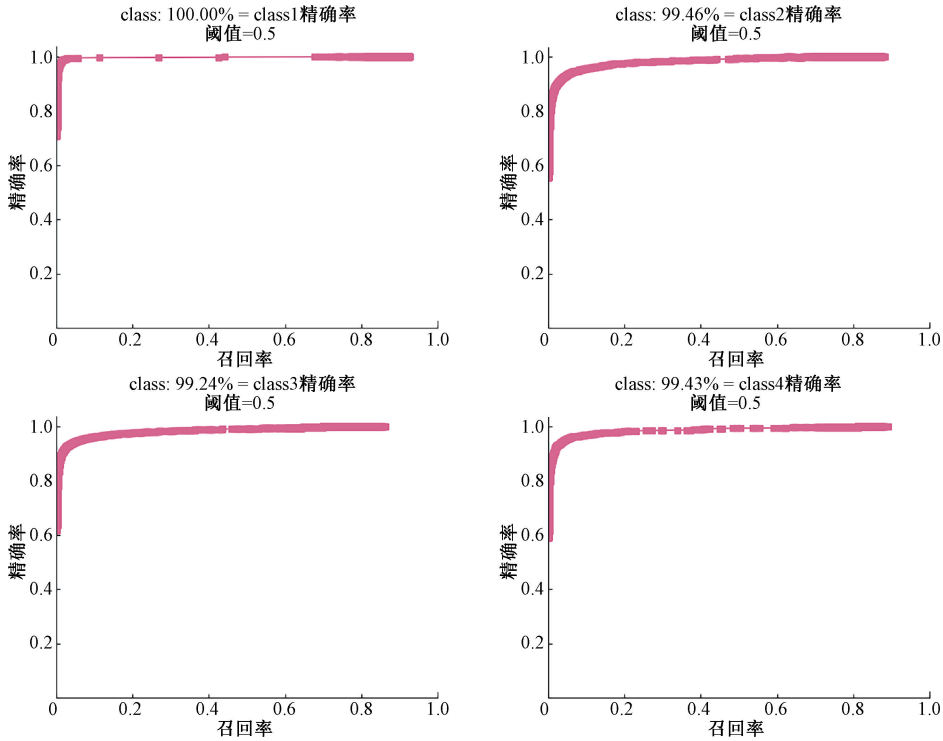


图 11 训练模型的精确率曲线

Fig. 11 Accuracy curve of training model

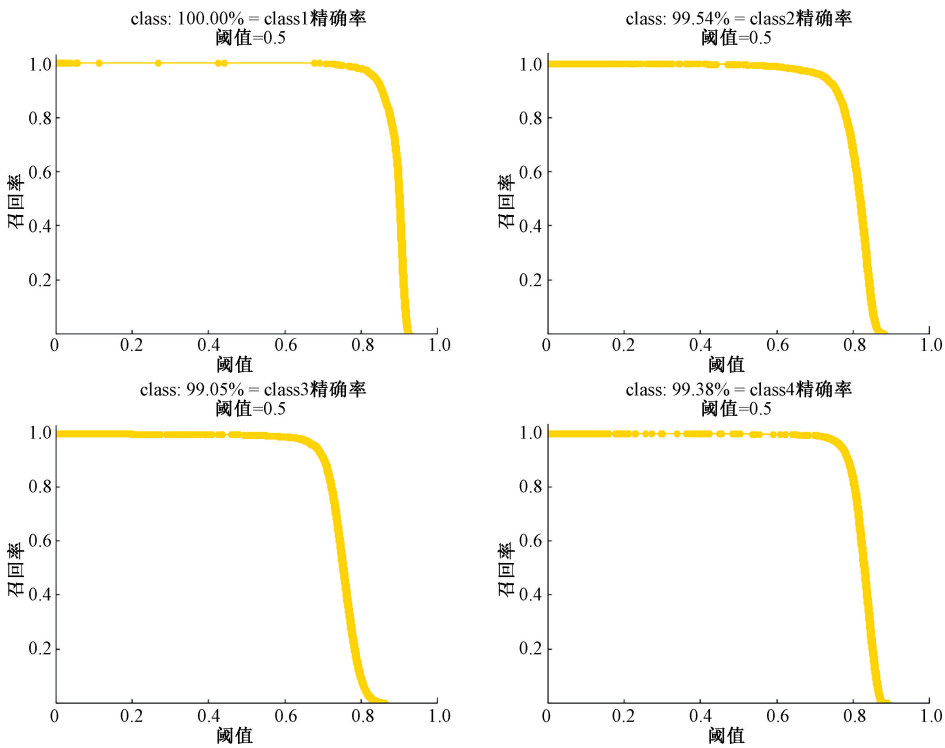


图 12 训练模型的召回率曲线

Fig. 12 Recall rate curve of training model

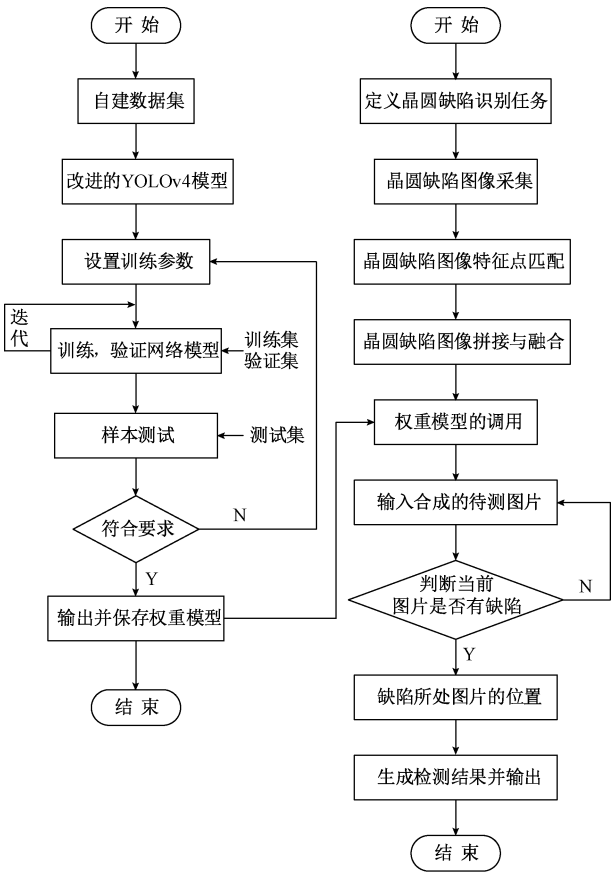


图 13　晶圆缺陷检测任务流程

Fig. 13　Flow chart of wafer defect detection task

法和本文所提出的方法，对比实验结果如表 3 和 4 所示。结果表明，本文提出的方法模型精确率达到 99. 53%，召回率达到 99. 49%，mAP 达到 99. 24%，相较于轻量级 YOLOv4-Tiny 算法分别提升 3. 88%、3. 73%、10. 08%，相较于原 YOLOv4 算法分别提升 2. 53%、1. 69%、1. 92%，有效地提升了碳化硅晶圆缺陷检测的识别率。平均晶圆缺陷检测每图的用时 0. 028 3 s，相较于基于 Halcon 软件方法提升 93. 42%，相较于 OpenCV 模板匹配方法提升 90. 52%，有效地缩减了碳化硅晶圆缺陷的检测时间。

表 3　改进 YOLOv4 算法检测性能对比结果

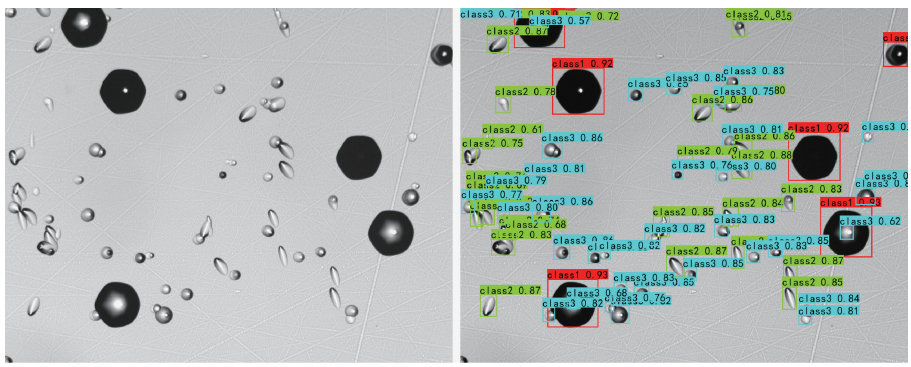
Table 3　Comparison results of detection performance of the improved YOLOv4 algorithm (%)

检测算法	精确率	召回率	mAP
YOLOv4-Tiny	95. 65	95. 76	89. 16
原 YOLOv4	97. 00	97. 80	97. 32
本文方法	99. 53	99. 49	99. 24

表 4　晶圆缺陷检测算法检测时间对比结果

Table 4　Wafer defect detection algorithm detection time comparison results

检测算法	平均每图用时/s
基于 Halcon 软件方法	0. 430 0
OpenCV 模板匹配方法	0. 298 5
本文方法	0. 028 3



(a) 测试图片　(b) 测试结果

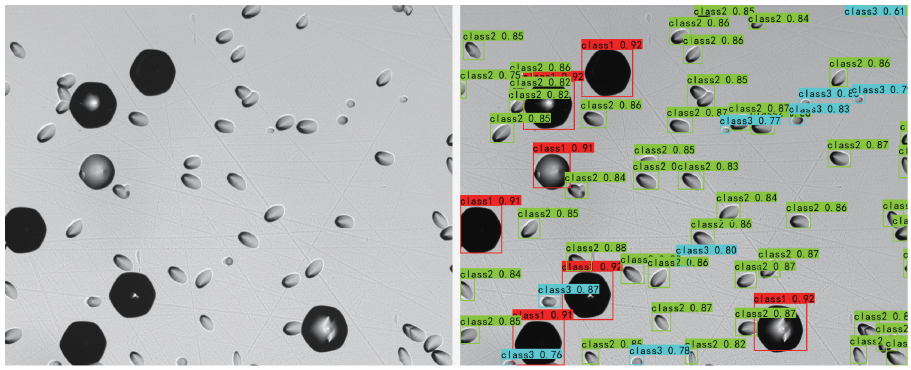
(a) Test image　(b) Test results

图 14　晶圆缺陷检测效果图 1

Fig. 14　Wafer defect inspection effect figure 1

4　结　论

本文提出一种基于深度学习与图像处理的高精度晶圆缺陷检测方法，结合深度学习和数字图像处理方法，实现了对高倍显微镜下采集的高精度晶圆缺陷图像进行快速定位和检测。通过对碳化硅晶圆缺陷的数据进行图像增强与扩张，自主标注构建晶圆缺陷的数据集，解决图像数量偏少的问题。通过图像特征点提取与匹配算法，将采集到的局部细节图像进行全局合成。并且，针对不同晶圆缺陷的特点，在原始 YOLOv4 的算法基础上，对算法进行了如下改进：加入了预处理模块提速，引入了

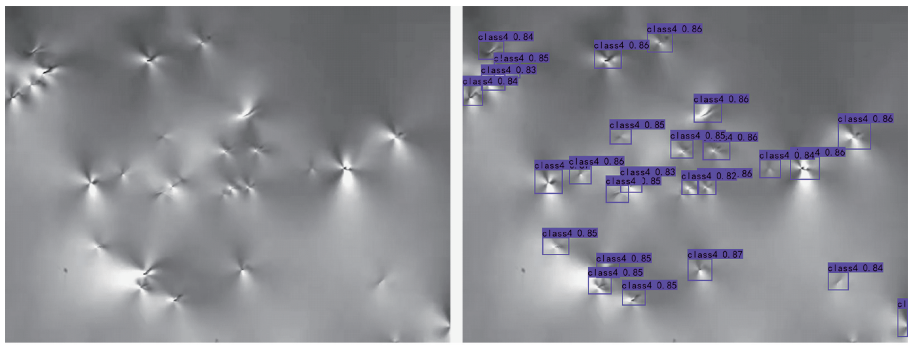


(a) 测试图片
(a) Test image

(b) 测试结果
(b) Test results

图 15 晶圆缺陷检测效果图 2

Fig. 15 Wafer defect inspection effect figure 2



(a) 测试图片
(a) Test image

(b) 测试结果
(b) Test results

图 16 晶圆缺陷检测效果图 3

Fig. 16 Wafer defect inspection effect figure 3

(a) 测试图片
(a) Test image

(b) 测试结果
(b) Test results

图 17 晶圆缺陷检测效果图 4

Fig. 17 Wafer defect inspection effect figure 4

Coordinate Attention 注意力机制模块改进骨干网络、对用于 Anchor Box 筛选的 K-means 聚类算法进行了改进、使用了 Swish 激活函数以及解耦合 YOLO-head 检测头。这些改进增强了算法的特征提取能力和检测的精度,在实验测试中,模型的平均识别精度达到 99.24%,相较于轻量级 YOLOv4 提升约 10.08%,相较于原 YOLOv4 提升约 2%。平均晶圆缺陷检测每图的用时约 0.028 3 s,相较于基于机器视觉的检测方法提升达 90%以上。实验结果表

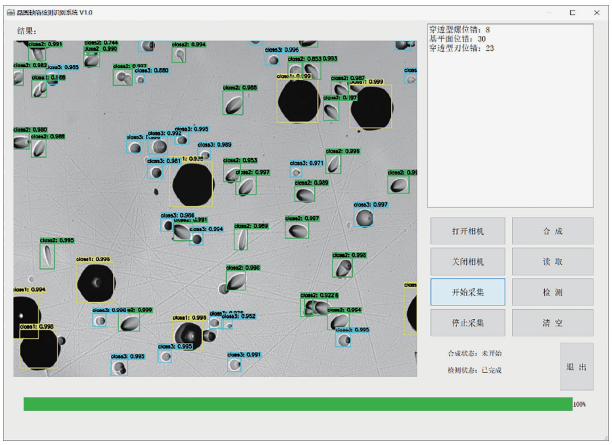


图 18　晶圆缺陷检测识别系统界面图

Fig. 18　Interface diagram of wafer defect detection and recognition system

明，本文算法具有检测速度快、精度高、针对性强等特点，能满足实际缺陷检测需求，有效地对晶圆缺陷检测技术提供了参考。

本文的提出的检测方法仍有改进空间，在现有实验结果的基础上如何进一步提升采集图片的拼接融合速度，以及由于算法存在的数据类别较少，后期需要增加更多的晶圆缺陷种类，提升模型的泛化性能以便其能适用于多种丰富的场景是接下来待解决的问题。

参考文献

[1] 王佳楠. 助熔剂法生长碳化硅晶体研究[D]. 天津：天津理工大学，2022.
WANG J N. Study on silicon carbide crystal growth by flux method [D]. Tianjin: Tianjin University of Technology, 2022.

[2] TSAI D M, WU S C, LI W C. Defect detection of solar cells in electroluminescence images using Fourier image reconstruction[J]. Solar Energy Materials and Solar Cells, 2012, 99: 250-262.

[3] LIU H, ZHOU W, KUANG Q, et al. Defect detection of IC wafer based on spectral subtraction [J]. IEEE Transactions on Semiconductor Manufacturing, 2010, 23(1): 141-147.

[4] 徐晓光，李海. 多尺度特征在 YOLO 算法中的应用研究[J]. 电子测量与仪器学报，2021，35(6)：96-101.
XU X G, LI H. Application of multi-scale features in YOLO algorithm[J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(6): 96-101.

[5] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[6] 周志锋，万旺根，王旭智. 基于 YOLOv3 框架改进的目标检测[J]. 电子测量技术，2020，43(18)：102-106.
ZHOU ZH F, WAN W G, WANG X ZH. Improved target detection based on YOLOv3 framework [J]. Electronic Measurement Technology, 2020, 43(18): 102-106.

[7] KARAMI E, PRASAD S, SHEHATA M. Image matching using SIFT, SURF, BRIEF and ORB: Performance comparison for distorted images[J]. arXiv preprint arXiv: 1710. 02726, 2017.

[8] 徐明，刁燕. 基于 SURF 算子与 FLANN 搜索的图像匹配方法研究[J]. 现代计算机，2020(14)：49-52，57.
XU M, DIAO Y. Research on image matching method based on SURF operator and FLANN search [J]. Modern Computer, 2020(14): 49-52, 57.

[9] LI H, QIN J, XIANG X, et al. An efficient image matching algorithm based on adaptive threshold and RANSAC[J]. IEEE Access, 2018, 6: 66963-66971.

[10] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv: 10934, 2020.

[11] 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.

[12] 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.

[13] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3-19.

[14] 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.

[15] MISRA D. Mish: A self regularized non-monotonic activation function [J]. arXiv preprint arXiv: 08681, 2019.

[16] MAAS A L, HANNUN A Y, NG A Y. Rectifier nonlinearities improve neural network acoustic models[J]. Proc. icml, 2013, 30(1): 3.

[17] RAMACHANDRAN P, ZOPH B, LE Q V. Searching for activation functions [J]. arXiv preprint arXiv: 05941, 2017.

[18] HE K, ZHANG X, REN S, et al. Spatial pyramid

pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis Machine Intelligence, 2015, 37(9): 1904-1916.

[19] DAS C, BOSE S, CHATTOPADHYAY M, et al. A novel distance based modified K-means clustering algorithm for estimation of missing values in micro-array gene expression data [J]. International Journal of Information Technology Management Information System, 2014, 5(3): 1-13.

[20] MAKARYCHEV K, REDDY A, SHAN L. Improved guarantees for K-means++ and K-means++ parallel[J]. Advances in Neural Information Processing Systems, 2020, 33: 16142-16152.

[21] GE Z, LIU S, WANG F, et al. YOLOx: Exceeding YOLO series in 2021 [J]. arXiv preprint arXiv: 2107.08430, 2021.

作者简介

史浩琛,现为济南大学物理科学与技术学院研究生,主要研究方向为计算机视觉、深度学习、光学工程。

E-mail: 517325506@qq.com

Shi Haochen is now a M. Sc. candidate at University of Jinan. His main research interests include computer vision, deep learning and optical engineering.

夏伟(通信作者),现为济南大学教授、博士生导师,主要研究方向为半导体激光器件及其应用、光学工程。

E-mail: sps_xiaw@ujn.edu.cn

Xia Wei (Corresponding author) is now a professor of University of Jinan. His main research interests include semiconductor laser device and application, optical engineering.