

面向产线AI质检的少样本评测方法研究和验证

焦若丹, 高东辉, 黄艳华, 刘硕, 段宣翡, 王蕊, 刘伟东

引用本文

焦若丹, 高东辉, 黄艳华, 刘硕, 段宣翡, 王蕊, 刘伟东. [面向产线AI质检的少样本评测方法研究和验证](#)[J]. 计算机科学, 2024, 51(6A): 230700086-8.

JIAO Ruodan, GAO Donghui, HUANG Yanhua, LIU Shuo, DUAN Xuanfei, WANG Rui, LIU Weidong. [Study and Verification on Few-shot Evaluation Methods for AI-based Quality Inspection in Production Lines](#) [J]. Computer Science, 2024, 51(6A): 230700086-8.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于BERT和CNN的药物不良反应个案报道文献分类方法](#)

Literature Classification of Individual Reports of Adverse Drug Reactions Based on BERT and CNN
计算机科学, 2024, 51(6A): 230400049-6. <https://doi.org/10.11896/jsjcx.230400049>

[DUWe:动态未知词嵌入方法在Web异常检测中的应用](#)

DUWe:Dynamic Unknown Word Embedding Approach for Web Anomaly Detection
计算机科学, 2024, 51(6A): 230300191-5. <https://doi.org/10.11896/jsjcx.230300191>

[WiCare:一种非接触式的老人如厕跌倒监测模型](#)

WiCare:Non-contact Fall Monitoring Model for Elderly in Toilet
计算机科学, 2024, 51(6A): 230700044-8. <https://doi.org/10.11896/jsjcx.230700044>

[深度学习驱动下IaaS云运维异常检测算法的研究进展](#)

Research Progress of Anomaly Detection in IaaS Cloud Operation Driven by Deep Learning
计算机科学, 2024, 51(6A): 230400016-8. <https://doi.org/10.11896/jsjcx.230400016>

[基于注意力的多尺度蒸馏异常检测](#)

Attention-based Multi-scale Distillation Anomaly Detection
计算机科学, 2024, 51(6A): 230300223-11. <https://doi.org/10.11896/jsjcx.230300223>

面向产线 AI 质检的少样本评测方法研究和验证

焦若丹 高东辉 黄艳华 刘硕 段宣翡 王蕊 刘伟东

中国移动通信有限公司研究院 北京 100053

(jiaoruodan@chinamobile.com)

摘要 随着工业 4.0 时代的到来,制造业与人工智能的深度融合已成为了业界的重要发展趋势,工业质检是其中的重要突破口,但目前业界缺乏对工业质检产品进行评估的标准方法,各质检产品的性能往往不透明,不利于优化迭代及规模推广。针对上述问题,提出了一种面向工业产线应用需求的 AI 工业质检算法评测方法,可在工业领域样本数量少且不均衡的情况下,面向产线落地应用需求,对 AI 工业质检产品及其竞品进行对标评估。该评测方法通过交叉验证法构建数据集,从而避免数据集规模小且不均衡导致的评测结果波动较大的问题,通过灰盒测试法避免数据集来源单一导致的评测结果不客观问题,并根据产线实际生产需求制定相关评估指标,可真实反映产线应用场景下质检产品检测性能。将上述方法应用于光伏电池片 EL 检测产品的对标评测中进行验证,结果表明该评测方法具备可行性,且能客观反映各产品的真实性能。最后,基于对评测结果的分析对比,为 AI 工业质检产品的优化提供了一些建议。

关键词: AI 工业质检;深度学习;目标检测;缺陷检测;评测方法;光伏电池 EL 检测

中图分类号 TP391

Study and Verification on Few-shot Evaluation Methods for AI-based Quality Inspection in Production Lines

JIAO Ruodan, GAO Donghui, HUANG Yanhua, LIU Shuo, DUAN Xuanfei, WANG Rui and LIU Weidong

China Mobile Research Institute, Beijing 100053, China

Abstract With the advent of industry 4.0, the deep integration of manufacturing industry with artificial intelligence(AI) has become an important development trend. Industrial quality inspection has emerged as a significant breakthrough point. However, there is currently a lack of standardized methods for evaluating industrial quality inspection products in the industry. The performance of various quality inspection products is often opaque, making it difficult to optimize and scale up. In response to this situation, this paper proposes an AI-based industrial quality inspection algorithm evaluation method, which is suitable for the application needs of production lines in the industrial field. This method can evaluate AI-based industrial quality inspection products and their competitors in situations where the sample size is small and imbalanced. The evaluation method constructs a data set through cross-validation to avoid the problem of large evaluation result fluctuations caused by small and imbalanced data sets. It also uses gray box testing to avoid the subjectivity in evaluation results caused by a single source of data. Furthermore, it formulates relevant evaluation indicators based on the actual production needs of the production line, which can truly reflect the detection performance of quality inspection products in the production line application scenario. The proposed method is validated through benchmark evaluation of EL testing products for photovoltaic cells, demonstrating its feasibility and its ability to objectively reflect the true performance of various products. Finally, based on the analysis and comparison of the evaluation results, some suggestions are provided for the optimization of AI-based industrial quality inspection products.

Keywords Industrial quality inspection by AI, Deep learning, Object detection, Defect detection, Evaluation methods, Photovoltaic cell EL inspection

1 引言

随着深度学习、大模型等技术的快速发展, AI 工业质检在质检效率及准确率方面的优势逐渐凸显,其市场规模呈现高速增长态势^[1]。从质检需求角度出发, AI 工业质检主要进行分类、定位、分割等任务,多采用目标检测或实例分割模型或算法,目前,已有很多研究人员对可适用于工业质检的各种模型和算法进行了探索和研究,并从准确率等多种指标对性能进行了比较^[2]。如将 CNN 应用于 PCB 板^[3]、光伏电池^[4]、金属^[5]等物体的图像分类,采用 Faster R-CNN^[6]、融合多尺

度特征^[7]、DenseNet^[8]、YOLO^[9-11]等方法实现钢表面^[12]、LED^[13]、桥梁^[14-16]等产品的缺陷检测,基于 FCN^[17]、U-Net^[18]等实现像素级表面缺陷检测^[19-20],以及实现细小水泥裂纹检测^[21]等。但大多研究都是从学术角度对各种网络结构及其性能指标进行比较,缺乏贴近真实产线环境,从工业界厂家需求角度对各算法进行评测的相关研究。

本文提出了一种面向工业界需求的 AI 工业质检算法评测方法,可在工业领域样本数量少且不均衡的情况下,面向产线落地应用需求,对工业质检算法及其竞品进行准确、公正的对标评估。本文第 2 节对工业质检场景及过程进行简要介

绍;第3节对当前学术界及工业界的评价指标进行说明,并提出一种面向工业产线应用的 AI 工业质检算法评测方法;第4节将评测方法应用于光伏电池片 EL 缺陷检测算法对标评测中以验证其有效性,并对评测结果进行分析;最后总结全文并为 AI 工业质检算法优化提出了一些建议。

2 工业质检介绍

2.1 工业质检场景

工业质检场景具有多样性、复杂性、随机性等特点。以光伏电池 EL 缺陷检测为例,常见缺陷包括隐裂、黑团、黑片、亮片、失效、虚焊等,其中常见的隐裂缺陷又有线性隐裂、树状隐裂、叉状隐裂等多种类型,如图 1 所示。此外,电池组件的原材料、版型、成片方式、切割方式等方面的差异(如图 2—图 5 所示),也会给特征提取带来更大的挑战。传统工业质检过程通常面临着人力成本高昂且准确性不足的问题,因此, AI 工业质检成为应用热点。

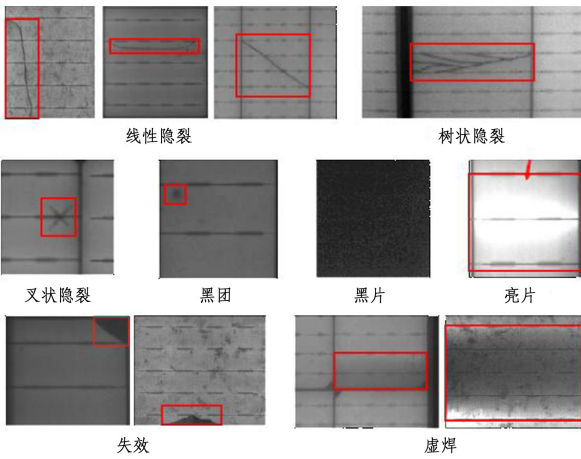
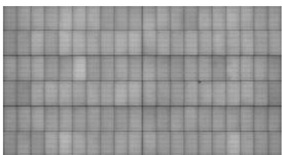
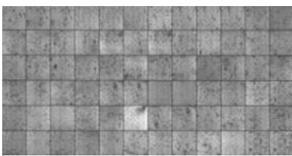


图 1 光伏电池组件 EL 缺陷

Fig. 1 EL Defects in photovoltaic cell components



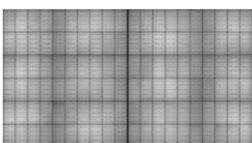
(a) 单晶电池组件



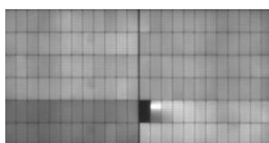
(b) 多晶电池组件

图 2 不同原材料电池组件

Fig. 2 Battery components made by different materials



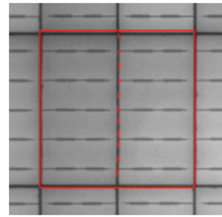
(a) 120 版型



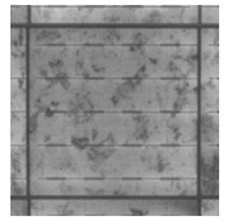
(b) 144 版型

图 3 不同版型电池组件

Fig. 3 Battery components with different modules



(a) 半片



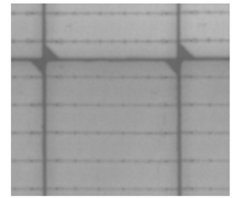
(b) 全片

图 4 不同成片方式电池组件

Fig. 4 Battery components made by different assembly methods



(a) 直角电池片



(b) 半角电池片

图 5 不同切割方式电池组件

Fig. 5 Battery components made by different cutting methods

目前 AI 工业质检主要应用于消费电子、汽车及零部件、钢铁冶金以及新能源等行业,并在向服装、装备制造、包装印刷、家装等行业扩展,应用场景包括板材缺陷检测、汽车轮毂检测、传送带异物检测、纺织瑕疵检测以及废钢定级等,如图 6—图 10 所示。其中,图 6—图 9 为目标检测类,图 10 为像素级实例分割类任务。

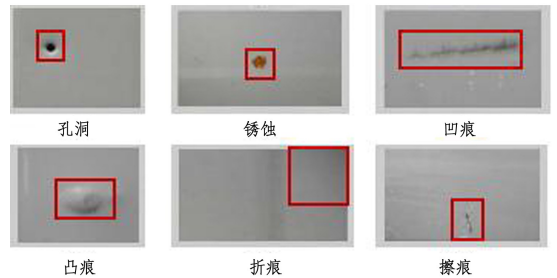


图 6 板材缺陷检测

Fig. 6 Board defect detection



图 7 汽车轮毂检测

Fig. 7 Car wheel hub inspection



图 8 传送带托辊及异物检测

Fig. 8 Conveyor belt roller and foreign substance detection

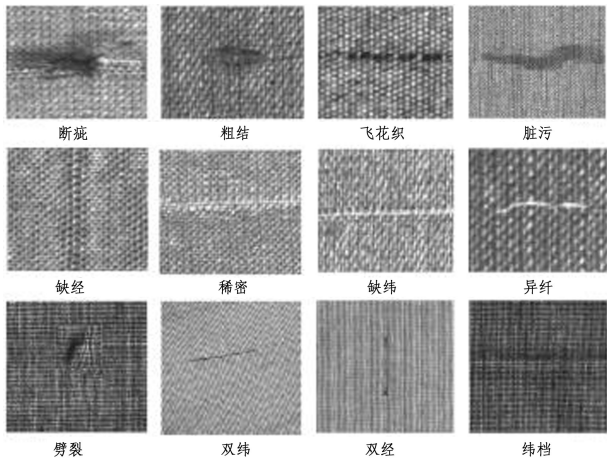


图 9 纺织瑕疵检测
Fig. 9 Textile flaw detection



图 10 废钢定级
Fig. 10 Scrap steel grading

2.2 工业质检过程

工业质检是产线生产过程中的一个重要环节,通常处于生产和出厂两个环节中间,用于确保产品质量,如图 11 所示。工业质检水平与前后环节中两个重要指标紧密相关,即生产良品率和出厂良品率。良品是指符合质量要求的产品,生产良品率是指良品数量在生产产品总量中的占比,出厂良品率是指良品数量在出厂产品总量中的占比。

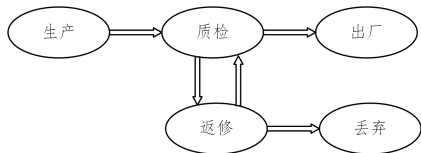


图 11 工业质检过程示意图
Fig. 11 Diagram of industrial quality inspection process

3 AI 工业质检评测技术

3.1 评价指标

在产线中, AI 工业质检任务大多为目标检测任务或实例分割任务,已有很多研究人员对该类任务进行研究,并提出了多项评测指标,但实际上工业界与产业界关注的重点不同,评测指标也有所区别,本节以光伏电池 EL 缺陷检测为例说明二者的区别。

3.1.1 学术界评价指标

目前学术界对目标检测和实例分割的评价指标主要分为精度评价指标和速度评价指标两类。常用的精度评价指标包括精确率(Precision)、召回率(Recall)、AP、mAP 等,常用的速度评价指标包括 FPS 等。

在讨论这些评价指标之前,我们首先要理解混淆矩阵和 IoU 这两个基础概念,它们是评价指标的基础。

混淆矩阵是一个二维矩阵,其行和列分别表示预测值和真值类别,通过比较预测样本真值和样本预测值,将预测结果划分为 TP(True Positive)、FN(False Negative)、FP(False Positive)、TN(True Negative) 4 类,如表 1 所列。

表 1 混淆矩阵
Table 1 Confusion matrix

	预测为正样本	预测为负样本
真值为正样本	TP	FN
真值为负样本	FP	TN

目标检测和实例分割任务通常使用 IoU 来评估预测结果。IoU(Intersection over Union,交并比)为预测框与真值框的交集面积与并集面积之比,用于衡量预测值与标注真值之间的差距。当预测框的分类标签正确,并且预测框与真值框的交并比(IoU)超过设定阈值时,认为该次预测是 TP。如果预测框的类别标签正确,或标签正确但交并比未超过设定阈值,那么该预测被视为 FP。

精确率和召回率的定义均基于上述概念。精确率是指预测正确的预测框在所有预测框中的占比,公式如下:

$$Precision = TP / (TP + FP)$$

召回率是指预测正确的预测框在所有真值框中的占比,公式如下:

$$Recall = TP / (TP + FN)$$

AP 和 mAP 是综合考虑精确率和召回率的评价指标。AP(Average Precision)是平均精度,用于评估单个类别的预测效果,mAP 是对所有类别 AP 进行平均后的数值,用于评估整体预测效果。AP 计算过程为:1)将本类别预测结果按照置信度由高到低排列,然后自上而下依次将当前置信度设定为阈值;2)将置信度阈值之下的数据舍弃不计入计算,以置信度阈值之上的数据计算精确率和召回率;3)将召回率分布划分为多个等大区,对每个召回率区间取最高精确率,并对所有区间的最高精确率求平均,即为 AP。对所有类别的 AP 取平均值,即为 mAP。

另外还有一种 AP 和 mAP 计算方式,这种方法在计算 AP 时会考虑多个 IoU 阈值,取各阈值下的 AP 值作为最终的 AP,然后对所有类别计算所得 AP 值取平均值,即为 mAP。

FPS 是指在同一硬件条件下,每秒处理的图片数量,用于评估模型的实时性。

3.1.2 工业界评价指标

工业界对质检产品性能进行评估时,往往是面向实际生产需求进行评估,更关注产线整体的出厂良品率,一般以漏检率、误检率作为评价指标,但不同厂家在不同场景下对漏检率、误检率可能有所不同,以下为其中的 3 种:

1) 计算方式 1:

$$\text{漏检率} = \text{漏检电池片数} / \text{总检测电池片数}$$

$$\text{误检率} = \text{误判电池片数} / \text{总检测电池片数}$$

2) 计算方式 2:

$$\text{漏检率} = \text{漏检电池组件数} / \text{总检测电池组件数}$$

$$\text{误检率} = \text{误判电池组件数} / \text{总检测电池组件数}$$

3) 计算方式 3:

$$\text{漏检率} = \text{未检测出的缺陷数} / \text{总缺陷数}$$

$$\text{误检率} = \text{误判电池片数} / \text{总检测电池片数}$$

其中,误判电池片数可以定义为电池片实际并无缺陷但

被判定包含缺陷的电池片数,也可以定义为存在将非缺陷位置检测为有缺陷位置这种误判的电池片数量。

除上述 3 种方式外,还存在许多其他的计算方式,各厂家的评测指标也并不完全一致。另外,从计算公式可以看出,上述方式均涉及到总检测电池数,因此,最终计算得出的漏检率、误检率等指标,实际上均与产线生产良品率密切相关,不能完全准确地反映产品的真实性能。

速度指标一般与厂家生产模式相关,对于实时检测的情况,检测速度需满足产线生产速度,不应成为生产瓶颈。

此外,在产线应用中,模型大小也是一个重要的衡量因素,因为它直接影响到计算资源的需求、系统的实时性、存储空间的占用,以及能耗和网络传输速度等,还可能会影响到生产效率和成本,所以在实际应用中需要特别注意。但由于该因素涉及范围较广,本文不对此详细讨论。

3.2 评测方法

目前工业界仍缺乏对质检产品性能进行评测的统一指标和标准方法,同类质检产品间很难进行对比评估,不利于产品的进一步优化以及在业界的规模推广。本节仍以光伏电池 EL 缺陷检测为例,对工业质检产品进行标准评测的问题及难点进行分析,并提出一种通用的 AI 工业质检产品评测方法。

如 3.1.2 节所述,不同厂家的评价指标定义往往存在偏差,且大多与产线良品率相关,无法客观反映质检产品的真实性能水平,因此,通过直接对比各质检产品宣称的性能指标值得出的结论存在不够客观的问题,仍需要通过真实测试进行对比。

真实测试的最佳方法是将各质检产品部署于光伏厂家,收集多个产线的大量生产数据,并进行实时检测,检测完毕后使用该厂家的指标计算公式进行结果统计,并对各质检产品进行对比。但显然这需要光伏厂家的充分配合,实际上基本不具备可行性。同时,由于工业数据往往涉及商业机密,为敏感数据,光伏厂家并不会提供大量产线数据用于离场测试,可用于测试的数据集规模往往较小,那么,如何在小规模数据集的情况下尽量保证测试结果的准确性就是真实测试面临的第一个问题。

第二个问题是数据来源,最佳方法是从多个光伏厂家采集数据并混合在一起作为数据集。但如前所述,工业数据为敏感数据,实际可获取并用于测试的数据一般只能为单一来源,且往往与某待测质检产品相关,比如为某产品调优及训练时使用的数据集,则此时各质检产品的评测可能存在公平问题,评测结果不够客观和准确。

第三个问题是评价指标,如 3.1.2 所述,以检测电池总数作为分母的评测指标实际上与数据良品率相关性较大,无法准确反映出质检产品的检测性能。

针对上述 3 个问题,本文分别提出相应解决方案。首先,关于指标定义问题,我们仍以学术界评价指标为基准,同时结合厂家需求,分别针对单缺陷和所有缺陷定义误报率、检出率,具体如下:

$$\text{单缺陷误报率} = \text{单缺陷错误缺陷检出数} / \text{单缺陷检出总数} = 1 - \text{单缺陷精确率}$$

$$\text{单缺陷检出率} = \text{单缺陷正确缺陷检出数} / \text{单缺陷总数} = \text{单缺陷召回率}$$

$$\text{总体误报率} = \text{错误缺陷检出数} / \text{缺陷检出总数}$$

$$\text{总体检出率} = \text{正确缺陷检出数} / \text{缺陷总数}$$

接下来针对第二个数据来源问题,我们采取灰盒测试方法进行评测,在评测阶段增加模型训练过程,避免因个别待测产品针对数据集进行特殊调参导致评测不够公平,同时,也增加了对产品泛化性的测试。灰盒测试法如表 2 所列。

表 2 灰盒测试法
Table 2 Grey box testing

1. 每个质检产品的评测过程包括训练阶段和推理阶段;
2. 在训练阶段,部署待测质检产品初始模型及训练环境,加载训练数据集,开始训练;
3. 模型训练结束后,按照规则获取推理模型,并进入推理阶段;
4. 在推理阶段,部署推理模型及推理环境,加载与训练阶段不同的数据作为测试集,开始推理;
5. 推理结束后,根据评价指标统计测试结果;
6. 若需进行多轮次测试,则重复执行步骤 1-5,最后,取每轮次测试的平均值作为最终测试结果。

最后针对第一个数据集规模过小的问题,我们通过交叉验证方法构建数据集,充分利用数据集的每部分数据特征,尽量降低数据集规模小及缺陷数据随机分布对测试结果的影响,避免测试结果随机性过大。交叉验证法如表 3 所列。

表 3 交叉验证法
Table 3 Cross-validation method

1. 将随机打乱后的数据集等分为 K 份数据集(可根据具体质检产品设置 K 值,建议从区间[3,10]中选择),并分别编号为 1,2,3,...,K 号;
2. 取第 1 号数据集作为测试数据集,并取剩余 K-1 份数据集作为训练数据集,作为第一轮次测试的数据集;
3. 同步骤 2,依次取 2-K 号数据集作为测试集,并取剩余 K-1 份数据集作为训练集,分别构建 K-1 轮次测试的数据集;
4. 使用 K 轮次数据集进行测试后,取 K 轮次测试结果平均值作为最终测试结果。

综合上述方法即为工业质检通用评测方法,首先采集现场数据,然后使用交叉验证法构建 K 轮次评测数据集,接着分别使用 K 轮次评测数据集完成 K 轮次训练和推理,并统计各轮次误报率及检出率,最后计算各轮次平均值作为该产品测试结果,并对比各产品评测结果,具体过程如图 12 所示。

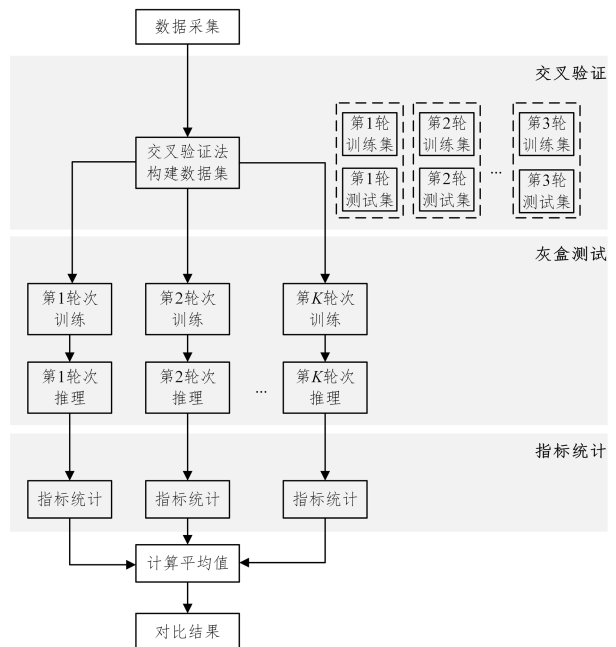


图 12 工业质检缺陷检测评测方法

Fig. 12 Evaluation method for industrial quality inspection defect detection

4 评测方法验证及分析

本节将描述缺陷检测评测方法在光伏电池 EL 检测评测中的应用验证情况,并对评测结果进行分析。验证时,共分两组对 A,B,A plus,B plus 这 4 款产品进行评测,其中,A plus 是对 A 优化后的产品,B plus 是对 B 优化后的产品,A 和 B 为第一组,A plus 和 B plus 为第二组。为进行对比分析,我们在第一组评测中,未采用交叉验证法构建数据集。

4.1 不采用交叉验证的评测方法验证

第一组评测中,采用 3.2 节所述的评测方法,评测 A 和 B 两款产品对单晶硅及多晶硅电池组件中隐裂、失效、亮片、黑

团 4 种缺陷的检测性能,但在构建数据集不采用交叉验证法,而是采用随机划分的方法,即对总数据集进行 3 轮次随机划分,每轮次将总数据集按照固定比例随机划分为训练集和测试集,最终得到 3 组评测数据集。

本次评测数据集包含 120 和 144 两种型号电池组件,半片、全片两种切片方式电池,直角、圆角两种切割方式电池,数据集规格相对复杂,在进行数据集随机划分时,会分别针对每一类按照比例随机划分训练集和测试集,保证训练和推理阶段各类数据分布均衡。表 4 列出了总数据集缺陷分布情况,表 5 列出了 3 轮次训练集缺陷分布情况,表 6 列出了 3 轮次测试集缺陷分布情况。

表 4 总数据集缺陷分布

Table 4 Overall dataset defect distribution

电池片类型	电池组件数	有瑕疵 电池组件数	电池片数	隐裂框数	失效框数	亮片框数	黑团框数
单晶	665	665	89664	1486	424	401	543
多晶	350	350	43800	1342	251	147	306

表 5 训练集缺陷分布

Table 5 Training set defect distribution

电池片 类型	轮次	训练集 电池组件数	训练集 电池片数	训练集 隐裂框数	训练集 失效框数	训练集 亮片框数	训练集 黑团框数
单晶	1	464	62616	1040	288	267	380
	2	464	62616	1025	294	261	352
	3	464	62616	882	283	284	391
多晶	1	245	30648	992	192	106	215
	2	245	30648	942	195	112	218
	3	245	30648	834	186	101	207

表 6 测试集缺陷分布

Table 6 Test set defect distribution

电池片 类型	轮次	测试集 电池组件数	测试集 电池片数	测试集 隐裂框数	测试集 失效框数	测试集 亮片框数	测试集 黑团框数
单晶	1	201	27048	446	136	134	163
	2	201	27048	461	130	140	191
	3	201	27048	604	141	117	152
多晶	1	105	13152	350	59	41	91
	2	105	13152	400	56	35	88
	3	105	13152	508	65	46	99

本组评测误报率不超过 20% 情况下各类缺陷的检出率,误报率、检出率定义见 3.2 节,A 和 B 两产品测试结果如表 7、表 8 所列。由测试结果可见:

1) 在单晶电池检测中,产品 A 对隐裂、失效的检测效果更好,产品 B 对亮片、黑团的检测效果更好;在多晶电池检测中,产品 B 对 4 种缺陷的检测效果均更好。

2) 针对同一种缺陷,产品 A 的误报率波动非常大,且经常无法满足误报率不超过 20% 的前提条件。

表 7 第一组单晶电池评测结果

Table 7 Evaluation results of the first group of monocrystalline solar cells

	(%)			
瑕疵 类型	产品 A 检出率	产品 A 误报率	产品 B 检出率	产品 B 误报率
隐裂	96.34	19.93	55.34	19.48
失效	95.57	1.26	90.93	19.39
亮片	97.70	25.80	98.20	19.15
黑团	88.14	6.06	90.34	19.33

表 8 第一组多晶电池评测结果

Table 8 Evaluation results of the first group of polycrystalline solar cells

	(%)			
瑕疵 类型	产品 A 检出率	产品 A 误报率	产品 B 检出率	产品 B 误报率
隐裂	94.11	52.31	68.15	19.42
失效	95.60	10.76	97.87	19.26
亮片	97.37	29.04	96.20	18.74
黑团	90.19	8.96	96.78	19.21

3) 4 种缺陷中,产品 A 和 B 对隐裂的检测效果最差,两个产品对隐裂的检测结果均存在波动较大的情况。其中,对多晶电池的检测误差尤其显著,产品 A 误报率较高,产品 B 检出率较低,如图 13、图 14 所示。

4) A 和 B 两个产品对不同缺陷的检测性能各有优劣,从目前指标及结果无法评判哪个产品性能更优。

实际上,A 和 B 两款产品采用的技术路线不同,产品 A 采用传统机器视觉算法与深度学习相结合的技术进行质检,产品 B 采用纯深度学习技术进行质检,因此,产品 A 无法通

过参数动态调整检测误报率,从用户角度而言不够灵活。另外,从检测结果来看,在误报率不超过 20% 的前提下,产品 A 的综合性能要弱于产品 B。从上述测试结果可以看出,对于场景相对复杂、面向产线应用的质检产品,深度学习的特征提取能力要优于人工的特征提取能力,采用深度学习可能会获得更好的泛化性及灵活性。

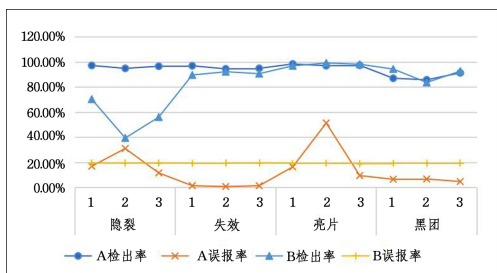


图 13 第一组单晶电池评测结果

Fig. 13 Evaluation results of the first group of monocrystalline solar cells

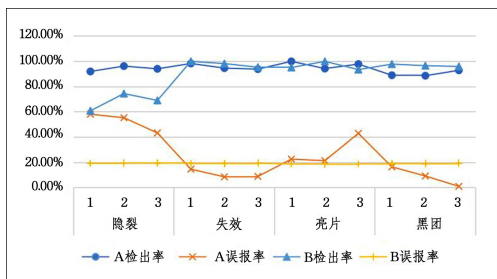


图 14 第一组多晶电池评测结果

Fig. 14 Evaluation results of the first group of polycrystalline solar cells

观察三轮次评测数据集可以发现,第三轮评测中隐裂缺陷在训练集及测试集中的分布与其他轮次有一定差距,这可能是导致隐裂缺陷检测结果波动较大的原因之一。

另外,由表 4 可见,实际数据集中各缺陷并非均匀分布,

其占比差距可能较大。对于光伏厂家,最关注的是总体的出厂良品率,而不是某种缺陷的良品率,若某种缺陷在产线中占比较少,即使检测效果相对较差,也并不会影响产品的整体检测性能。因此,评估质检产品检测性能时,应以所有缺陷的总体检出率及误报率为准进行评估。

4.2 采用交叉验证的评测方法验证

第二组评测中,采用 3.2 节所述的评测方法,评测 A plus 和 B plus 两款产品对单晶硅电池组件中虚焊、线性隐裂、树状隐裂、交叉隐裂 4 种缺陷的检测性能。A plus 为 A 优化后的产品,将传统机器视觉与深度学习结合的算法全部调整为深度学习算法,可通过调整置信度阈值动态调整误报率。B plus 为 B 优化后的产品,主要算法保持不变,主要优化方向为对新增缺陷的识别适配。

表 9 为总数据集缺陷分布情况,表 10 为 4 轮次训练集缺陷分布情况,表 11 为 4 轮次测试集缺陷分布情况。

如 4.1 节所述,考虑到不同缺陷在总数据集中占比不同,除统计每种缺陷的检出率及误报率外,本组评测中增加了对总体缺陷检出率、总体缺陷误报率的统计,以更准确地反映检测性能。

本组实验评测误报率不超过 5% 情况下各类缺陷的检出率,误报率、检出率定义见 3.2 节。经测试,A plus 和 B plus 两产品测试结果如表 12 所列,由测试结果可见:

1) 产品 A plus 在各种缺陷识别中表现比较稳定,检出率均在 90% 左右。

2) 产品 B plus 对交叉隐裂的检测性能很差,检出率较低,平均在 30% 左右,且波动较大,对虚焊的检测性能较好,检出率在 98% 以上。

3) 从总体来看,A plus 产品的总体检出率总体误报率,优于 B plus 产品。

4) 从评测结果看,本组评测中,未出现两个产品对某缺陷的检测结果均波动较大的情况,且相比于第一组单晶电池评测,总体检出率、误报率相对稳定,如图 15、图 16 所示。

表 9 总数据集缺陷分布

Table 9 Overall dataset defect distribution

电池组件数	有瑕疵 电池组件数	电池片数	虚焊 框数	线性隐裂 框数	树状隐裂 框数	交叉隐裂 框数
10007	6048	1429704	2754	1453	1140	2254

表 10 训练集缺陷分布

Table 10 Training set defect distribution

轮次	训练集 电池组件数	训练集 电池片数	训练集 虚焊框数	训练集 线性隐裂框数	训练集 树状隐裂框数	训练集 交叉隐裂框数
1	7506	1072344	2060	1093	848	1616
2	7505	1072272	2063	1095	860	1746
3	7505	1072272	2071	1089	861	1713
4	7505	1072224	2068	1082	851	1687

表 11 测试集缺陷分布

Table 11 Test set defect distribution

轮次	测试集 电池组件数	测试集 电池片数	测试集 虚焊框数	测试集 线性隐裂框数	测试集 树状隐裂框数	测试集 交叉隐裂框数
1	2501	357360	694	360	292	638
2	2502	357432	691	358	280	508
3	2502	357432	683	364	279	541
4	2502	357480	686	371	289	567

表 12 第二组评测结果

Table 12 Evaluation results of the second group of (%)

瑕疵类型	A plus 检出率	B plus 检出率	A plus 误报率	B plus 误报率
虚焊	92.09	98.62	1.21	4.43
线性隐裂	92.43	80.00	1.89	4.36
树状隐裂	89.73	79.10	2.84	4.25
交叉隐裂	91.67	30.96	4.11	4.35
总体	91.66	72.10	2.48	4.38

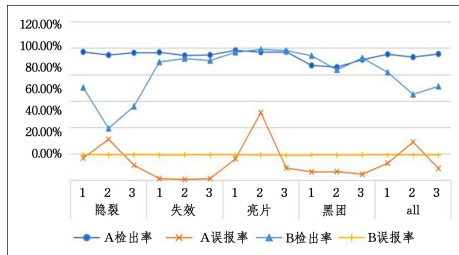


图 15 第一组单晶评测结果(增加总体检出率及总体误报率)

Fig. 15 Evaluation results of the first set of single crystal

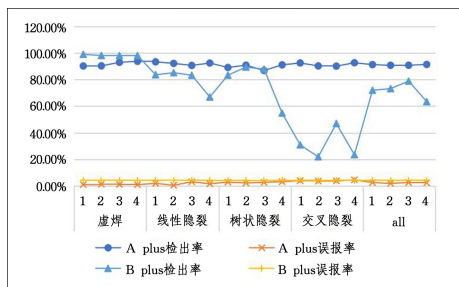


图 16 第二组评测结果

Fig. 16 Evaluation results of the second group

为更直观地比较两组评测的稳定性,我们分别针对两个产品计算两组检测的总体检出率及总体误报率方差均值。以 $N_j(k)$ 表示产品 k 第 j 组评测的测试轮次数量, $P_{ji}(k)$ 表示产品 k 第 j 组评测第 i 轮次的检出率, $\mu_{pj}(k)$ 表示产品 k 第 j 组评测检出率均值, $R_{ji}(k)$ 表示产品 k 第 j 组评测第 i 轮次的误报率, $\mu_{rj}(k)$ 表示产品 k 第 j 组评测误报率均值, $\sigma_{pj}^2(k)$ 表示产品 k 第 j 组评测的检出率方差, $\sigma_{rj}^2(k)$ 表示产品 k 第 j 组评测的误报率方差, (其中, $j=1$ 或 $2, k$ 为 A 或 B)。计算公式如下:

$$\sigma_{pj}^2(k) = \frac{\sum_{1 \leq i \leq N_j} (p_{ji}(k) - \mu_{pj}(k))^2}{N_j}, j=1 \text{ 或 } 2, k=A \text{ 或 } B$$

$$\sigma_{rj}^2(k) = \frac{\sum_{1 \leq i \leq N_j} (r_{ji}(k) - \mu_{rj}(k))^2}{N_j}, j=1 \text{ 或 } 2, k=A \text{ 或 } B$$

$$\sigma_{p1}^2 = \frac{\sigma_{p1}^2(A) + \sigma_{p1}^2(B)}{2}$$

$$\sigma_{p2}^2 = \frac{\sigma_{p2}^2(A) + \sigma_{p2}^2(B)}{2}$$

$$\sigma_{r1}^2 = \frac{\sigma_{r1}^2(A) + \sigma_{r1}^2(B)}{2}$$

$$\sigma_{r2}^2 = \frac{\sigma_{r2}^2(A) + \sigma_{r2}^2(B)}{2}$$

根据表 13、表 14 中的详细数据,可计算两组评测检出率方差均值 σ_{p1}^2 和 σ_{p2}^2 。两组评测误报率方差均值 σ_{r1}^2 和 σ_{r2}^2 如表 15 所列。由结果可见,第二组评测检出率及误报率方差均值均小于第一组,即第二组评测中各轮次检测结果波动性更

小、稳定性更好,可知第二组评测的数据集划分更均衡。相比随机划分法,采用交叉验证法构建数据集更客观。

表 13 第一组总体检出率及总体误报率

Table 13 Overall detection rate and false alarm rate of the first group

产品	轮次	总体检出率/%	总体误报率/%
A 单晶	1	95.56	13.13
	2	93.38	29.19
	3	95.76	9.17
	AVG	94.90	17.16
B 单晶	1	81.91	19.37
	2	65.18	19.33
	3	71.30	19.40
	AVG	72.80	19.37

表 14 第二组总体检出率及总体误报率

Table 14 Overall detection rate and false alarm rate of the second group

产品	轮次	总体检出率/%	总体误报率/%
A plus	1	91.53	2.63
	2	91.02	2.05
	3	91.06	2.63
	4	93.05	2.63
AVG	91.66	2.48	
B plus	1	72.23	4.40
	2	73.43	4.39
	3	79.11	4.40
	4	63.62	4.32
AVG	72.10	4.38	

表 15 检出率及误报率方差均值

Table 15 Mean variance of detection rate and false alarm rate

指标	含义	数值
σ_{p1}^2	第一组评测单晶电池检出率方差均值	0.002445568
σ_{p2}^2	第二组评测检出率方差均值	0.001569436
σ_{r1}^2	第一组评测单晶电池误报率方差均值	0.003748597
σ_{r2}^2	第二组评测误报率方差均值	3.21125×10^{-6}

结束语 本文针对 AI 工业质检产品提出了一种面向工业界产线应用的通用评测方法,并应用于光伏电池 EL 缺陷检测对标评测中。经验证,该方法可客观反映参测产品的真实性能情况。通过对多款产品的评测结果分析可以发现,在复杂和随机性强的工业质检场景下,神经网络的自动特征提取能力往往依赖于人工规则的特征提取方法,深度学习算法的检测性能及泛化能力往往要优于传统机器视觉算法;数据集对质检产品的优化至关重要,数据分布不均可能会对质检产品性能造成较大影响。因此,对于面向产线应用的质检产品,可从算法和数据集两个维度进行优化提升。在算法上,建议优先采用深度学习算法;在数据集上,应着重提高数据集的质量,并尽量保持评测数据集与实际产线数据分布一致,以实现更优的检测性能。

鉴于实际工业生产线中应用场景的多样性和复杂性,本文所提出的评估指标集仍需要进一步扩展和深化。在后续研究中,我们将深入探索更加广泛的应用领域和产线场景,继续完善评估指标集,持续优化评估方法,以更好地满足工业产线应用对 AI 工业质检产品的评估需求。

参考文献

[1] 国家工业信息安全发展研究中心. AI 工业质检应用发展白皮书. [R/OL]. <https://www.zhangqiaokeyan.com/academic>

journal-cn_ceocio-china_thesis/02012102861837.html.

- [2] LUO D L, CAI Y X, YANG Z H, et al. Survey on industrial defect detection with deep learning[J]. *Sci SinInform*, 2022, 52: 1002-1039.
- [3] DENG Y S, LUO A C, DAI M J. Building an automatic defect verification system using deep neural network for PCB defect classification[C]// *Proceedings of the 4th International Conference on Frontiers of Signal Processing(ICFSP)*. 2018:145-149.
- [4] DEITSCH S, CHRISTLEIN V, BERGER S, et al. Automatic classification of defective photovoltaic module cells in electroluminescence images[J]. *Sol Energy*, 2019, 185:455-468.
- [5] ZHANG Z, WEN G, CHEN S. Weld image deep learning-based on-line defects detection using convolutional neural networks for Al alloy in robotic arc welding[J]. *J Manuf Processes*, 2019, 45: 208-216.
- [6] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]// *Proceedings of Advances in Neural Information Processing Systems*. 2015:91-99.
- [7] HE Y, SONG K, MENG Q, et al. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features[J]. *IEEE Trans Instrum Meas*, 2020, 69:1493-1504.
- [8] 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*. 2017: 4700-4708.
- [9] REDMON J, FARHADI A. Yolov3: an incremental improvement[J]. *arXiv:1804.02767*101255, 2018.
- [10] 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.
- [11] REDMON J, FARHADI A. Yolo9000: better, faster, stronger [C]// *CVPR*. 2017.
- [12] LI J, SU Z, GENG J, et al. Real-time detection of steel strip surface defects based on improved YOLO detection network[J]. *IFAC-Papers OnLine*, 2018, 51:76-81.
- [13] CHEN S H, TSAI C C. SMD LED chips defect detection using a YOLOv3-dense model[J]. *Adv Eng Inf*, 2021, 47:101255.
- [14] CHA Y J, CHOI W, SUH G, et al. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types[J]. *Comput-Aided Civil Infrastruct Eng*, 2018, 33:731-747.
- [15] TAO X, ZHANG D, WANG Z, et al. Detection of power line insulator defects using aerial images analyzed with convolutional neural networks[J]. *IEEE Trans Syst Man Cybern Syst*, 2020, 50:1486-1498.
- [16] ZHANG C, CHANG C, JAMSHIDI M. Concrete bridge surface damage detection using a single-stage detector[J]. *Comput-Aided Civil Infrastruct Eng*, 2020, 35:389-409.
- [17] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015:3431-3440.
- [18] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]// *Proceedings of International Conference on Medical Image Computing and Computer-assisted Intervention*. Cham: Springer, 2015: 234-241.
- [19] QIU L, WU X, YU Z. A high-efficiency fully convolutional networks for pixel-wise surface defect detection[J]. *IEEE Access*, 2019, 7:15884-15893.
- [20] HUANG Y, QIU C, YUAN K. Surface defect saliency of magnetic tile[J]. *Vis Comput*, 2020, 36:85-96.
- [21] XIE Y, ZHU F, FU Y. Main-secondary network for defect segmentation of textured surface images[C]// *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020:3531-3540.



JIAO Ruodan, born in 1989, master, intermediate engineer. Her main research interests include artificial intelligence, vertical industry and other fields evaluation technology.



GAO Donghui, born in 1980, Ph.D, principal researcher, senior Engineer. His main research interests include the evaluation of artificial intelligence and information communication technologies, and the development of assessment systems.