DOI: 10.19650/j.cnki.cjsi.J2513981

大模型在工业缺陷检测领域的应用现状与展望*

何毓芬[1],何赟泽[1,2],尹 勇[1],邓堡元[1],王耀南[1,2]

(1.湖南大学电气与信息工程学院 长沙 410082; 2.湖南大学机器人视觉感知与控制技术国家工程研究中心 长沙 410082)

摘 要:工业缺陷检测是现代工业生产和运维的关键环节,是产品质量、生产效率、安全性的重要保障。大模型凭借其复杂逻辑推理能力与泛化能力成为了推动新一轮人工智能浪潮的关键引擎。大模型的涌现为工业缺陷检测提供了一个新范式,同时也带来了新机遇和新挑战。本综述总结了大模型在工业缺陷检测领域的应用现状。首先,对大模型的发展历程进行了系统梳理,并且详细介绍了大模型的核心技术,包括模型架构、多模态数据处理与预训练技术、微调技术、对齐技术和高效推理技术。接着,综述了基于传统机器学习和深度学习的缺陷检测方法,并与缺陷检测大模型进行对比,总结了各自的优点和局限性。然后,聚焦于工业缺陷检测领域,介绍了支撑大模型研究与性能评估的开源数据集和性能评价方法,并梳理了大模型目前的主要应用方向,即缺陷检测与定位、复杂场景与微小缺陷检测、小样本与零样本自适应检测、交互式缺陷分析与决策支持、缺陷数据生成与自动标注。最后,深入分析了工业缺陷检测大模型目前面临的数据质量与安全、高可靠性要求、成本限制与可持续发展、缺乏统一测评标准等挑战,并对其未来发展趋势进行了展望,旨在为大模型技术在工业缺陷检测领域的进一步发展和广泛应用提供有价值的参考和见解。

关键词:人工智能;大模型;工业缺陷检测;无损检测

中图分类号:TH89 TP391.41 **文献标识码**:A **国家标准学科分类代码**:520.20

Application status and prospects of large models in the field of industrial defect detection

He Yufen[1], He Yunze[1,2], Yin Yong[1], Deng Baoyuan[1], Wang Yaonan[1,2]

(*1. College of Electrical and Information Engineering, Hunan University, Changsha 410082, China; 2. National Engineering Research Center of Robot Visual Perception and Control Technology, Hunan University, Changsha 410082, China*)

Abstract: Industrial defect detection is a critical component of modern industrial production and operation, ensuring product quality, production efficiency, and safety. The complex logical reasoning and generalization capabilities of large models have positioned them as the critical force behind the new wave of artificial intelligence. With the emergence of large models, a new paradigm is established for industrial defect detection, bringing both fresh opportunities and challenges. This article provides a comprehensive review of the current application status of large models in the field of industrial defect detection. Firstly, the development process of large models is systematically combed, and the core technologies are introduced in detail, including model architecture, multimodal data processing, pre-training techniques, fine-tuning methods, alignment strategies and efficient reasoning mechanisms. Secondly, a survey of traditional methods based on machine learning and deep learning for industrial defect detection is provided, followed by a comparison with large model-based approaches and a summary of their respective strengths and limitations. Then, focusing on the industrial defect detection domain, the review introduces commonly used open-source datasets that support large model research and evaluation, as well as the performance evaluation methods of large models. Furthermore, it categorizes the current main application of large models into five directions, including defect detection and localization, defect detection in complex scenarios and micro-defect detection, few-shot and zero-shot adaptive detection, interactive defect analysis and decision support, and defect data generation with automatic annotation.

收稿日期:2025-04-28 Received Date: 2025-04-28

*基金项目:国家自然科学基金面上项目(52377009)、湖南省科技创新领军人才项目(2023RC1039)资助

Finally, this article thoroughly analyzes the challenges confronting large models in industrial defect detection, such as data quality and security, high-reliability requirements, cost constraints and sustainable development, and the lack of unified evaluation standards, while providing an outlook on their future trends. The review aims to provide valuable references and insights for the continued advancement and broader implementation of large models in industrial defect detection.

Keywords: artificial intelligence; large model; industrial defect detection; non-destructive testing

0　引　　言

工业缺陷检测是现代工业生产与运维中确保产品质量与安全的关键环节，旨在通过高效、精准的检测手段，识别产品在生产制造及使用过程中可能存在的缺陷，从而确保产品质量符合标准，提升安全性和客户满意度。精准的缺陷检测能有效避免有缺陷的产品进入市场或继续使用，提高产品可靠性和安全性，避免因产品缺陷而引发安全事故。同时，检测数据还能帮助企业排查设计、生产乃至使用环节中存在的问题，从而针对性地改进工艺参数与运维策略。因此，工业缺陷检测已成为智能制造领域不可或缺的基础技术之一[1]，被广泛应用于无人质检、过程监控等多个场景。然而传统检测方法难以满足当下工业缺陷检测对检测精度、速度、稳定性及自适应性等方面的高要求。人工检测效率低下且易受主观因素影响，漏检率高。传统的机器学习算法及深度学习算法虽然推动了缺陷检测技术的进步，但仍存在跨场景迁移能力不足等问题。面对现有挑战，大模型的火热发展为工业缺陷检测的智能化转型带来了新的机遇。

当前，关于大模型(large model，LM)的定义，在学术界与业界尚未形成明确的共识。一般认为，大模型，狭义上指大规模预训练语言模型(large language model，LLM)，广义上则涵盖了语言、声音、图像、视频等多种模态的大模型。“大”在此处是一个相对概念，而非一个绝对化的衡量标准。相较于参数量通常在数万至数亿区间的传统模型，大模型的参数量则至少在亿级以上，部分甚至攀升至万亿级的规模[2]。大模型以其规模性、涌现性及通用性等特性，在人工智能领域独树一帜[3]。大模型具备强大的理解和生成能力，能够从大规模数据中挖掘深层次特征，有效解决小样本训练与跨类别迁移等问题，大幅提高缺陷检测精度与效率。

现有文献对大模型在工业缺陷检测领域的应用缺乏详细、系统的综述。文献[4]聚焦于工业异常检测大模型典型方法，分为少样本学习和零样本学习两类方法展开介绍，缺乏大模型核心技术、大模型在工业缺陷检测领域的应用方向以及面临的挑战等相关内容。文献[5]从训练目标、模型结构和规模、模型性能以及未来发展方向等角度对比了大模型方法与其他缺陷检测方法，但缺少对大模型发展历程、应用方向的梳理介绍，对大模型核心技术的介绍也不够全面。因此本文将聚焦于工业缺陷检测领域，对大模型的发展历程、核心技术、现有应用方向、面临的挑战和未来发展方向进行系统梳理，以期推动大模型技术的持续进步和在工业检测领域应用的广泛拓展。

1　大模型简介

1.1　大模型发展历程

如图 1 所示，大模型的发展历程可以分为 5 个主要阶段。第 1 阶段为统计语言模型(statistical language model，SLM)，此类模型产生于基于统计方法的自然语言处理系统的研究中[6]，通常是根据词序列中若干个连续的上下文单词来预测下一个词的出现概率，其中最具代表性的是 N-gram 模型。该模型基于马尔可夫假设[7]，即一个词的出现概率仅依赖于其前面的 $n-1$ 个词，通过计算一系列词语的出现概率来预测或评估语句的合理性。统计语言模型计算简单，但其依赖语料库中的词频统计，又受限于数据稀疏性和长距离依赖问题，难以捕捉复杂的语言结构和语义信息，模型容量低，泛化能力弱。

Bengio 等[8]于 2003 年首次提出了神经概率语言模型，标志着进入第 2 阶段神经语言模型(neural language model，NLM)。此类模型通过多层人工神经网络处理大量文本数据，能够捕捉语言中的复杂模式和依赖关系，从而理解和生成人类语言。Mikolov 等[9]于 2013 年提出的 Word2Vec 是该阶段最重要的工作之一，它极大地推动了词嵌入技术的发展并简化了网络架构，为后续的神经语言模型研究奠定了基础。神经语言模型泛化能力较强，但需要大量的计算资源和数据进行训练，且模型的可解释性较差。

第 3 阶段为预训练语言模型(pre-trained language model，PLM)。2017 年，Google 发布基于自注意力机制的特征提取器 Transformer[10]，解决了早期模型在长程依赖性和顺序处理等方面的问题，迅速在自然语言处理领域占据主导地位，并广泛应用于图像处理等其他领域。预训练语言模型通常在大规模无标签文本数据集上进行无监督预训练以学习通用语言表示，然后针对具体下游任务进行适配微调，减少了对标注数据的依赖，实现“通用语言理解”。自此“预训练+微调”范式成为主流，模型规模和数据量大幅提升。

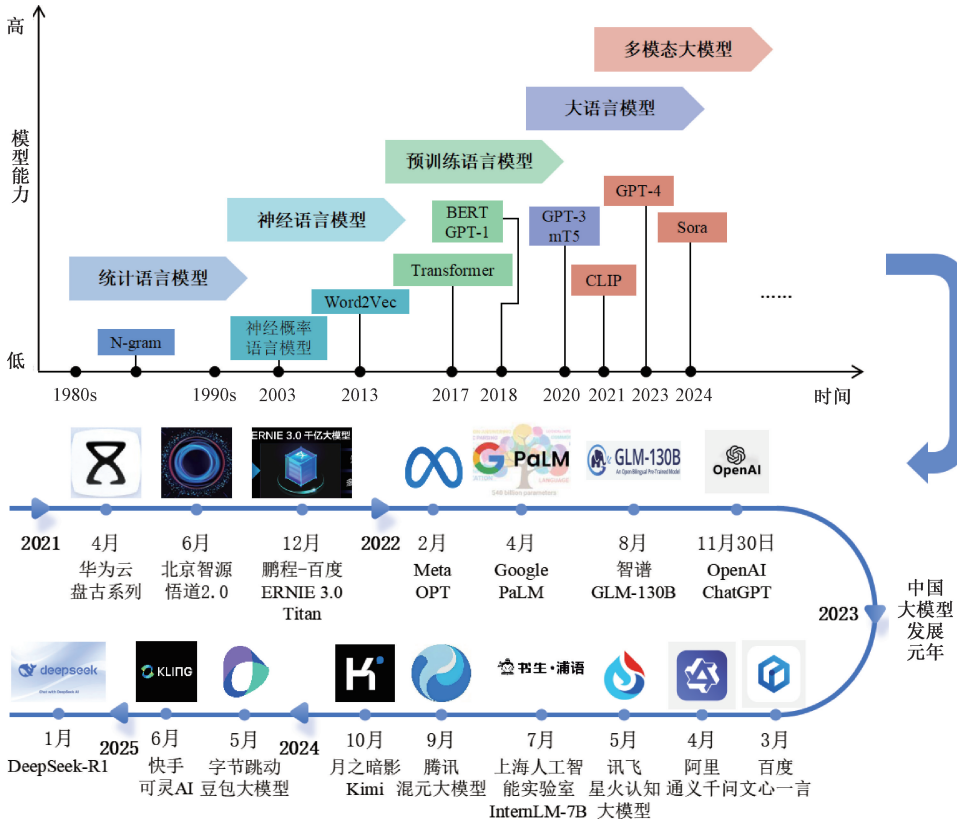

图 1 大模型发展历程

Fig. 1 The development process of large model

第 4 阶段为大语言模型 LLM,随着预训练语言模型的训练数据和参数的持续扩张,当参数规模超过一定水平时,模型展现出涌现能力,学界遂将此类模型命名为大语言模型。大语言模型通常具有数十亿甚至上千亿参数,通过在大规模文本数据上进行训练,能够生成、理解人类语言并进行交互,具有较强的语言生成和理解能力。2022 年 11 月,OpenAI 发布以 GPT-3.5 为基础的大语言模型应用 ChatGPT,迅速成为用户增长最快的互联网应用。ChatGPT 的全面爆火揭开了大模型时代的序幕,预示着 AI 迈向通用人工智能的新一轮冲刺[11]。

第 5 阶段为多模态大模型(large multimodal model, LMM),此类模型能够处理和理解多种类型的数据输入,如文本、图像、视频和音频等。通过多模态数据的融合和解释,能够实现更加符合人类认知方式的跨模态内容理解和生成。2023 年 3 月,OpenAI 推出多模态大模型 GPT-4,年底 Google 发布 Gemini,标志着“原生多模态”时代的到来[12]。2024 年 2 月,OpenAI 对外发布了 Sora 视频大模型,并于 12 月正式对用户开放。

如图 1 所示,在 OpenAI 发布 ChatGPT 之前,国内的部分研究机构已在大模型领域进行了投资和研发,并取得了若干成果。2023 年被视为中国大模型的发展元年,3 月,百度推出文心一言,成为国内首个基于大模型的人工智能产品。此后,腾讯、阿里、科大讯飞、快手及字节跳动等科技公司均顺应浪潮发布了自研大模型及衍生产品。

如今,大模型的热潮仍在持续,国内的 AI 大模型团队逐步扩展至视觉、决策等领域,甚至应用于蛋白质预测以及航天等领域的重大科学问题[13]。2025 年 1 月 20 日,DeepSeek 发布并开源了新一代推理模型 DeepSeek-R1,性能与 OpenAI 的 o1 正式版持平,缩小了中美顶级 AI 模型性能差距。

1.2 大模型核心技术

1)模型架构

当前较为流行的大模型基础架构沿用了自然语言处理领域最热门有效的 Transformer 架构,具体结构如图 2 所示。其关键创新在于引入了自注意力机制,使模型能够同时考虑输入序列中的所有位置,对序列中更重要的部分赋予更高的注意力权重,从而能够更好地捕捉语义关系,解决了长距离依赖问题,显著提升了训练速度以及序列建模的效果。此外,Transformer 中的自注意力机制被扩展为多个注意力头,每个头可以学习不同的注意力权重,从而能更好地捕捉不同类型的关系并有效缓解了

单一注意力机制可能出现的有效分辨率降低等问题。多头注意机制的并行计算模式提高了训练及推理的效率,使得模型能拥有更大的规模并处理更长的序列。如表 1 所示,基于经典的 Transformer 架构,衍生出 3 种主流的大模型框架:Encoder-Decoder、Encoder-Only、Decoder-Only,三者各有侧重,使用场景不同。

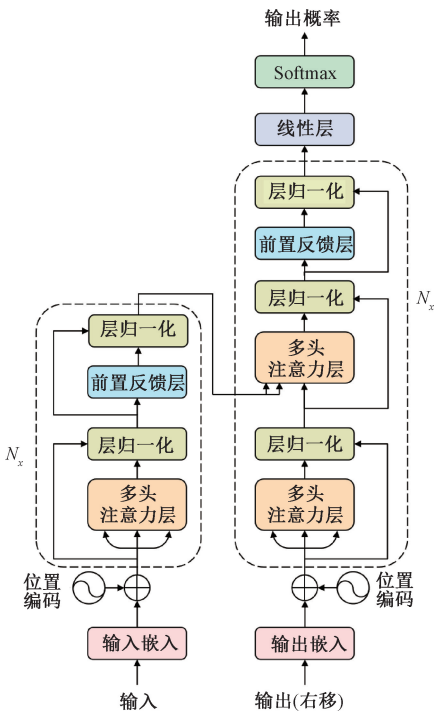

图 2　Transformer 架构

Fig. 2　The architecture of Transformer

表 1　大模型主流框架

Table 1　Mainstream framework of the large model

大模型框架	特点	适用场景	代表模型
Encoder-Decoder	结合了编码器和解码器的特点,更灵活	文本摘要、问答和对话建模等序列到序列任务	T5、Marian、BART 等
Encoder-Only	仅包含编码器部分,理解和处理输入数据,不生成新的内容	句子分类、命名实体识别、抽取式问题解答等任务	BERT 系列、GLM4、Electra 等
Decoder-Only	仅包含解码器部分,可以从输入的编码中生成相应的序列	适用于文本生成、机器翻译等需要生成序列的任务	GPT 系列、LLaMA、BLOOM 等

然而 Transformer 架构存在一定的局限性,随着序列长度的增加,注意力机制变慢,计算复杂度及对内存的需求也会增加。近年来,基于 Transformer 架构的混合专家系统(mixture of experts,MoE)逐渐成为大模型的主流架构。从早期的 Mixtral 到近期的 DeepSeek 再到 Qwen2.5-Max 以及 Llama 4,采用 MoE 架构的模型正不断涌现,并在性能表现上持续突破。如图 3 所示,MoE 架构主要由两部分组成,即稀疏 MoE 层和门控网络[14]。MoE 层包含若干专家模型,每个专家本身是一个独立的神经网络,具体结构可按需选取。门控网络根据输入样本的特征将其输送到特定专家,并根据预训练参数为专家的预测结果分配相应权重。最终输出结果由各专家网络输出加权得到。

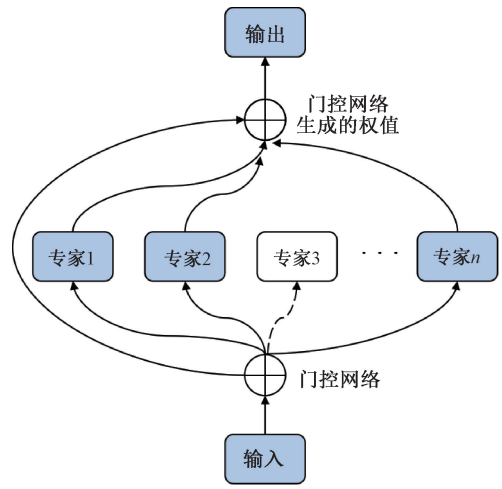

图 3　MoE 架构

Fig. 3　The architecture of MoE

MoE 架构凭借专业化的专家网络集成和动态门控设计[15],训练速度更快,效果更好,在相同参数下推理成本更低,扩展能力较强,故而模型能在保持计算成本不变的情况下增加参数量。且 MoE 架构在多任务学习中具备较优的性能,在工业场景中展现出巨大潜力,不同的“专家”网络可以专门处理不同类别的缺陷,或适配不同产线的产品,通过门控网络动态路由,实现在一个统一大模型内的高度专业化检测,显著提升模型容量和效率。

目前研究人员也正在积极探索可能取代或增强 Transformer 的全新架构,如 RetNet[16]、Mamba[17]、RWKV[18]、Hyena[19]等,以期提高模型效率,并降低计算复杂度和计算成本,从而应用于需要高速推理的边缘部署场景。

2)多模态数据处理与预训练技术

预训练是大模型的核心训练策略,模型以自监督的方式在大型语料库及预设的任务上进行预训练[20],以捕捉特定的属性和结构信息。预训练模型的通用特征可以

为下游任务提供充足的信息,并加快模型的收敛速度[21]。多模态数据(如文本、图像、音频、视频等)的来源各异,且表现形式和语义结构各不相同,使得模型难以直接处理并有效融合这些异构信息以辅助任务决策[22]。在工业缺陷检测领域,多模态数据在通用场景的基础上新增了红外热成像、超声波、X 光、高光谱、3D 点云等物理传感数据。这些模态的数据分布、分辨率、信息密度差异较大,对数据处理和预训练提出了更高要求。工业大模型的预训练通常包括两个阶段,首先在通用视觉-语言数据集上进行预训练以获得基础表征能力。随后在大规模未标注的工业图像与多模态数据上进行领域适应性预训练,让模型充分学习工业数据的纹理、结构、噪声分布。由于模型规模较大,参数较多,为提高训练效率,一般采用分布式训练模式。常见的分布式训练模式有两种[23]:(1)数据并行,即多个设备共享模型权重,同时处理不同数据;(2)模型并行,即将模型切分为多个部分,并分布到不同设备进行训练。

3)微调技术

微调是指在预训练模型的基础上[24],使用特定任务的数据集对模型进行训练,调整优化模型参数,使模型能够更好地适应特定的领域或场景。模型在预训练阶段已经学习了丰富的通用知识,在有限的标注数据训练下,微调能使模型更有效地学习特定任务的特征[25]。微调技术对于大模型在工业缺陷检测领域的应用至关重要。工业缺陷样本获取成本高,标注较为困难,使用微调技术,可利用较少的缺陷样本和计算资源,在预训练大模型基础上快速得到高精度、高泛化性的专业检测模型,极大地降低大模型的部署门槛和成本。全参数微调和参数高效微调均为常用的微调技术。

(1)全参数微调

全参数微调(full parameter fine-tuning, FPFT)的原理与预训练模型类似,但所有的参数均处于合适的初始值,从而能使用较少数据在初始值的基础上继续训练模型更新参数[26]。FPFT 方法能够充分利用模型的全部潜能,可以较快速地迁移学习[27],适用于下游任务和预训练模型之间存在较大差异的情况,或者任务需要模型具有高度灵活性和自适应能力的情况。但由于模型的每个参数都可能要根据新任务进行优化,需要更多的计算资源和时间,计算成本高。且当微调的训练数据有限,或者与原始预训练数据差异极大时,可能导致过拟合。

(2)参数高效微调

参数高效微调(parameter-efficient fine-tuning, PEFT)是一种针对大模型的微调技术,旨在减少微调过程中需要调整的参数,同时保持或提高模型的性能[28]。常见方法包括:

a. 基于重新参数化的方法

代表方法为低秩适配(low-rank adaptation, LoRA)及其各种变体[29],如动态低秩适配[30](dynamic low-rank adaptation, DyLoRA)、自适应低秩适配[31](adaptive low-rank adaptation, AdaLoRA)和量化低秩适配[32](quantized low-rank adaptation, QLoRA)等。LoRA 由 Hu 等[33]于 2021 年提出,旨在通过矩阵分解和低秩近似技术来减少微调过程中的计算和存储需求。其核心思想是在模型的关键层中添加小型、低秩的矩阵来调整模型的行为,而非改变整个模型的结构。LoRA 能够显著降低微调过程的计算资源损耗,且收敛迅速,效率高,这使得在单个消费级 GPU 上微调大模型成为可能,但由于低秩近似带来的信息损失,模型性能可能下降。

b. 基于 Adapter 的方法

适应微调(adapter tuning, AT)通过在预训练模型的特定位置添加适配器层,实现对特定任务的微调。此方法保留了预训练模型的原始权重,仅在需要适应新任务的地方进行小规模的参数调整,可达到与全参数微调相同的效果。在工业缺陷检测大模型中通过切换 Adapter 即可实现一个模型检测多种产品,简化了部署架构,但需要设计合适的适配器结构和训练策略,以确保其能够有效适应不同检测对象。

c. 软提示方法

代表方法为提示微调(prompt tuning, PT),在输入序列之前增加一些特定长度的特殊 token[34],如“以下内容是关于[缺陷类别]的问答”,通过训练确定这些特殊 token 的参数,增大生成期望序列的概率,从而引导模型生成更符合下游任务需求的输出。在微调过程中,模型自动学习提示规则,而预训练参数保持冻结[35]。在工业检测中,提示词可以是描述缺陷类型的文本,引导模型进行零样本或少样本检测,为处理未知缺陷提供了新范式。但提示微调的效果在很大程度上依赖于提示的设计质量,且对于逻辑推理、多跳问答等较复杂任务仍需结合其他参数微调方法。

4)对齐技术

大模型对齐技术是指通过一系列方法使模型的输出与人类价值观、任务目标或特定领域知识保持一致的过程[36]。在工业领域,对齐不仅要求模型的输出符合人类价值观,更要求其与领域专家知识、质量标准和生产工艺规范对齐。对齐技术主要基于强化学习算法[37],其中最具代表性的是基于人类反馈的强化学习(reinforcement learning from human feedback, RLHF)。OpenAI 于 2020 年首次将 RLHF 技术应用于大模型领域[38],其核心在于利用人类反馈来优化模型行为。利用 RLHF 技术让质量工程师对模型的检测结果(如漏检、误检)进行偏好排序,训练奖励模型,使模型的决策与工程师的经验对齐。然

而,RLHF 技术也面临许多挑战,如人类偏好的主观性、数据标注的高成本等。因此谷歌提出了基于人工智能反馈的强化学习(reinforcement learning from AI feedback, RLAIF),用 LLM 生成的反馈替代人类反馈,降低了数据成本,提高了训练效率[39]。大模型利用总结的工艺文档和质检标准自动生成反馈,构建奖励函数,使模型能够区分"可接受的工艺瑕疵"和"必须处理的严重缺陷",从而实现更加智能和符合生产实际的决策。但 RLAIF 技术也面临着大模型的"幻觉"等问题。

5)高效推理技术

工业场景对推理的实时性、可靠性和成本有较高要求,大模型因其庞大的参数规模和高计算需求,对部署环境的计算、存储和内存资源要求极高,严重限制了其在实际工业场景的部署与应用。而模型的高效推理技术则成为突破这一瓶颈,实现大模型在实际工程应用中广泛落地的关键技术。大模型高效推理技术是指通过优化方法和策略,提升大模型在推理过程中的性能和效率,使其能够更快速、准确地处理各种任务,具体可分为 3 个层次[40],即数据级优化、模型级优化、系统级优化。聚焦于工业缺陷检测领域,数据级优化包括图像预处理与感兴趣区域(region of interest, ROI)提取等方法,例如先对检测对象区域进行裁剪再送入模型分析,而非处理整张图像,大幅减少输入数据量;模型级优化主要借助剪枝、量化、蒸馏等压缩技术[41]对预训练模型进行精简,去除冗余参数或冗余计算,实现模型小型化与高效化;系统级优化则主要利用 TensorRT、OpenVINO 等推理加速库,针对工业级边缘计算硬件进行深度优化,实现计算图和内存调度优化,满足产线检测要求。

2　大模型与传统 AI 缺陷检测方法对比

缺陷是一个较为广泛的概念,在工业领域,缺陷通常是指产品、部件、材料或工艺过程中存在的不完美、瑕疵或不符合预期标准的部分。缺陷检测技术极具行业特点,任务场景分散,针对工业场景而言,是指使用机器学习、深度学习等方法,结合传感器、自动化软件等技术对产品表面或内部的缺陷进行检测,从而确保产品质量和安全。缺陷检测任务可进一步细化为缺陷识别、缺陷分类、缺陷定位、缺陷分割、缺陷量化等任务,即找出有缺陷的产品,对缺陷进行分类,定位缺陷的位置,给出缺陷的量化指标。然而工业场景具有任务高度分散、缺陷形态多变、缺陷标注数据稀缺等特点,对检测技术的泛化性、适应性及自动化程度提出了更高要求。为了应对这些挑战,工业缺陷检测的技术范式经历了从依赖人工特征工程与浅层分类器的传统机器学习,到能够端到端自动学习特征的深度学习的演进,如今已发展至以大模型技术为基础探索通用智能的新阶段。本章将系统回顾这一演进历程,重点综述基于传统机器学习与深度学习的缺陷检测方法,并通过与大模型方法的对比深入剖析其成就与固有局限性。

2.1　基于传统机器学习的缺陷检测方法

早期机器学习方法依赖人工特征工程与支持向量机等分类器,在缺陷检测领域有着广泛的应用。如表 2 所示,常用的机器学习算法包括支持向量机(support vector machine, SVM)、决策树、随机森林、遗传算法等。这些算法通过对已标注的样本数据进行学习,构建出能够对新样本进行分类或预测的模型。其学习过程包括数据预处理、数据特征提取、模型训练、模型测试、模型评估改进等部分。传统机器学习方法将缺陷检测问题建模为特征工程驱动下的分类任务,其核心思想在于利用人类的先验知识来设计和提取对缺陷敏感的特征,再交由分类器进行决策,性能上限高度依赖于研究者设计的特征对缺陷的表征能力。

文献[42]提出了一种基于多特征的 SVM 多分类缺陷检测方法,提取焊点图像的形状、纹理及方向梯度直方图特征,先用最优核函数的 SVM 对多锡、少锡、焊锡合适和漏焊 4 种类型进行检测,再对误检焊点用基于方向梯度直方图特征的 SVM 多分类算法进行二次检测分类,最终分类准确率可达 98.46%以上。文献[43]提出了一种基于 SVM 的绝缘子缺陷检测算法,通过提取暗通道、纹理等多特征并融合成特征矩阵,利用正负样本训练 SVM 分类器,最终检测出绝缘子缺陷位置。文献[44]提出了一种基于 SVM 和几何特征的纺织品缺陷检测方法,整体分类准确率达到 96.15%。

文献[45]提出了用 Boosting 算法结合决策树建立组合分类器来识别带钢表面缺陷的方法,对实际带钢表面缺陷数据集进行测试的准确率超过 90%。文献[46]提出了一种基于决策树的钢铁表面缺陷检测方法,通过结合局部二值特征提取、主成分降维和 Bagging 优化技术,显著提高了分类速度和准确性。

文献[47]提出了一种基于随机森林算法的分类模型,以实现纤维板表面大刨花、胶斑、杂物、油污等缺陷的快速、准确识别。文献[48]提出了一种基于随机森林算法的自动化荧光渗透检测方法,通过实验验证了其在航空组件缺陷检测中的有效性。

文献[49]提出一种基于改进遗传算法与二维最大熵的编织袋缺陷快速检测方法,能够快速选取图像分割的最佳阈值,提高分割速度与精度。文献[50]提出了一种基于遗传算法和神经网络的芯片缺陷检测方法,采用振动分析结合机器学习的非接触式检测方案,提升了检测效率与精度。

表 2 基于机器学习的缺陷检测方法

Table 2 Defect detection methods based on machine learning

类别	文献	方法	应用场景	效果
支持向量机	[42]	基于多特征的 SVM 多分类缺陷检测方法	印刷电路板焊点缺陷	最终分类准确率可达 98.46%以上
	[43]	基于 SVM 的多特征融合方法	绝缘子缺陷	计算量小、收集数据量少
	[44]	基于 SVM 和几何特征的缺陷检测方法	纺织品缺陷	整体分类准确率达到 96.15%
决策树	[45]	Boosting 算法结合决策树建立组合分类器	带钢表面缺陷	准确率达到了 90%以上
	[46]	结合局部二值特征提取、主成分降维和 Bagging 优化技术	钢铁表面缺陷	能够显著提高分类速度,同时保持较高的准确率
随机森林	[47]	基于随机森林算法的分类模型	纤维板表面缺陷	在生产线上的识别正确率达 97.8%
	[48]	基于随机森林算法的自动化荧光渗透检测方法	航空组件	缺陷正确识别率 76%,与二级检测员相当
遗传算法	[49]	基于改进遗传算法与二维最大熵的快速缺陷检测方法	编织袋缺陷	相比传统遗传算法分割时间降低了 7.4%
	[50]	基于遗传算法和 BP 神经网络的缺陷检测方法	倒装芯片	缺失、焊点开路焊点缺陷准确率达 100%

本节所述方法在特定条件下取得了良好效果,但其性能严重依赖于手工特征的质量,要求研究者既要有深厚的领域知识以设计特征,又要熟悉机器学习算法。这种特征工程范式泛化能力弱、效率低、处理海量数据的能力也较为有限,难以应对复杂多变且缺陷类型多样的工业场景。大模型沿用其“特征–分类”思想,并把“人工特征”升级为“预训练+提示”。

2.2 基于深度学习的缺陷检测方法

深度学习具有端到端特性,能够自动从大量原始数据中学习特征表示。在缺陷检测领域,大部分深度学习方法属于有监督的表征学习方法,可进一步细分为基于卷积神经网络(convolutional neural networks,CNN)的分类网络(如 AlexNet、DenseNet 等)、目标检测网络(如 Faster R-CNN、YOLO 系列等)、分割网络(如 Mask R-CNN、Unet 等)。其中分类网络主要用于识别缺陷,将完好样本和存在不同类型缺陷的样本分别归类;目标检测网络用于确定缺陷在图像中的具体位置信息;而分割网络将图像中的每个像素划分到不同的类别中,能够获得缺陷的几何信息,从而能进一步量化缺陷。在实际应用中,3 种网络可结合使用。深度学习的兴起标志着工业缺陷检测进入了“表征学习”时代,其核心优势在于摆脱了对人工特征工程的依赖。

基于深度学习的缺陷检测方法时间线如图 4 所示。

文献[51]提出了一种基于深度卷积神经网络的自动化缺陷检测模型,采用联合检测架构,结合全局分类与局部检测,兼顾缺陷分类和缺陷定位。文献[52]提出了一种基于 Mask R-CNN 的铸件 X 射线 DR 图像缺陷检测算法,首次将引导滤波增强与 Mask R-CNN 结合,能够同步输出目标框、类别和分割掩模,支持精细化缺陷分析。文献[53]提出了一种包含 2 个生成器和 4 个判别器的表面缺陷生成对抗网络(surface defect-generation adversarial network,SDGAN)框架,结合对抗损失与循环一致性损失,生成高质量、多样化的缺陷图像,为小样本工业场景下的深度学习模型训练提供了实用解决方案。文献[54]通过改进的条件生成对抗网络(condition generative adversarial nets,CGAN)生成高质量缺陷样本,结合深度 CNN 分类模型,有效解决了电子元件缺陷检测中的数据稀缺问题,为工业质检提供了新思路。

文献[55]对 Res-UNet 进行改进,使用 ResNet50 代替 ResNet18 作为编码模块,提升特征提取能力,并引入 DenseNet 结构,残差块间共享浅层特征,增强特征复用和深度延展能力。文献[56]提出了一种基于 Mobile-Unet 的高效织物缺陷分割方法,兼顾高精度与实时性,适合边缘设备部署,满足产线快速检测需求。文献[57]基于 YOLO-v4-tiny 框架,利用光致发光成像检测单晶硅太阳能电池中的微裂纹和暗斑缺陷,通过集成空间金字塔池

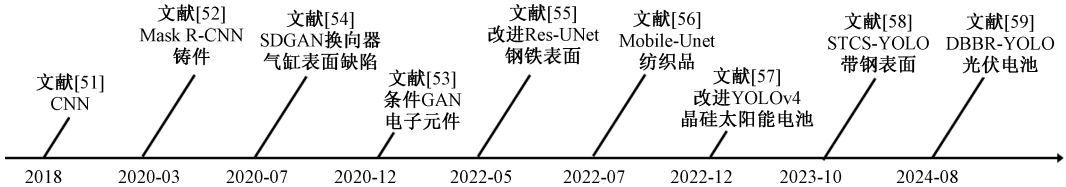

图 4 基于深度学习的缺陷检测方法时间线

Fig. 4 Timeline of defect detection methods based on deep learning

化模块,增强多尺度特征提取能力,在保持较高精度同时,提升模型推理速度。文献[58]对 YOLOv5 进行改进,提出了一种基于 STCS-YOLO 的带钢表面缺陷检测算法,平均精度提高了 3.9%。文献[59]将多样化分支模块(diverse branch block,DBB)引入至 YOLOv8n 网络,提出了一种基于 DBBR-YOLO 的光伏电池表面缺陷检测方法,mAP50 值达 93.1%。

深度学习方法极大地拓宽了工业缺陷检测的性能边界,但其本质上仍是针对封闭数据集和特定缺陷类别的感知“小模型”,存在通用性差、缺乏认知能力等局限。这些挑战催生了研究者对更具通用性、更高效范式的探索,即工业缺陷检测大模型。大模型保留深度学习的“端到端”思想,扩大网络参数,扩展数据模态,能够实现从“缺陷识别”到“决策支持”的转变。

2.3　对　比

随着人工智能技术的不断发展,大模型逐渐成为研究热点。大模型通常具有数十亿甚至数千亿参数,通过在大规模数据上进行预训练,学习到丰富的语言、图像等知识表示,具备较强的理解与生成能力。在工业缺陷检测领域,基于大模型的方法主要是利用大模型的预训练知识,通过微调或提示工程等方式,将其应用于具体的缺陷检测任务。

大模型与传统 AI 缺陷检测方法的优点和局限性对比如表 3 所示。大模型方法具备生成和跨模态学习能力,能够有效地结合语言信息和视觉信息,从而更精准地解析并描述图像中的复杂特征,适用于多任务和跨领域缺陷检测任务。虽然大模型在工业缺陷检测中展现出广阔的应用前景,但传统 AI 缺陷检测方法凭借其较小的参数规模和更高的计算效率,在特定的应用场景中,如资源受限环境,仍具有独特优势,不会被完全取代,大模型与传统 AI 缺陷检测方法将长期共存,协同演进。

表 3　大模型与传统 AI 缺陷检测方法对比

Table 3　Comparison between large model and traditional AI defect detection methods

方法	简介	优点	局限性
传统机器学习方法	针对特定应用场景,输入结构化数据,通过算法对数据进行建模。常见的算法包括支持向量机、决策树、随机森林、朴素贝叶斯等	模型相对简单,数据需求量小,计算资源需求低,训练及推理速度快,可解释性强,易部署到边缘设备,适合简单分类、回归任务	数据量过大时性能可能下降,依赖人工设计的特征或简单的自动特征提取方法,鲁棒性较差,难以适应复杂多变的场景,对于特征不明显或背景复杂的图像,检测效果较差
深度学习方法	深度学习方法通过构建多层神经网络自动学习数据中的特征,能够捕捉到图像中的局部特征和层次结构	可自动提取特征,鲁棒性较强,参数规模较小,计算效率较高,适合缺陷分类、缺陷分割等复杂任务,可用于资源受限的环境	数据需求量大,训练时间较长,可解释性差。难以提供关于缺陷的具体信息的详细描述,例如缺陷位置,类别,颜色和严重程度
多模态大模型	不针对特定场景,通用场景。预训练(特征学习)+微调(任务训练),具备通用泛化能力	具备跨模态学习和生成能力,能够处理复杂任务。可进行小样本和零样本学习,适合多任务多领域检测,更符合实际工业应用场景。能够提供关于缺陷的详细描述,并解释缺陷成因	模型参数量大,网络架构复杂,训练难度大,收敛速度较慢。对计算资源和数据量的要求极高,可解释性差,部署硬件要求高

3　工业缺陷检测行业大模型的应用现状

随着工业自动化的推进,利用先进的人工智能技术提升生产质量和效率已成为工业制造领域的重要课题。2024 年被视为中国行业大模型的发展元年,全国多个地区和部门紧跟时代潮流,积极响应国家战略部署,发布了一系列相关政策,以推动大模型技术在各行各业的深入发展与应用。重点支持在智能制造、生物医药、智能化教育教学、科技金融、数字政府等关键领域构建示范应用场景,打造标杆性大模型产品和服务。本章聚焦于工业缺陷检测行业,从支撑大模型研究与性能评估的开源数据集、大模型性能评价方法、现有应用方向 3 个角度,系统地梳理了大模型在该行业的应用现状。

3.1　支撑大模型研究与评估的关键开源数据集

支撑大模型研究与性能评估的关键开源数据集如表 4 所示,检测对象包括钢材、磁砖、印刷电路板、织物等多种工业产品,为验证大模型的领域泛化能力提供了基本条件。由于工业生产过程中,缺陷样本的获取与标注难度较大[60],且缺陷样本往往涉及到企业的核心技术,开源的缺

陷数据集样本量普遍偏小,最多不超过两万张。这样的数据量可满足小规模模型的训练需求,却难以直接用于大模型的预训练,但或可用于大模型的微调、少样本/零样本学习能力评估、以及特定下游任务性能基准测试等环节。

表 4 支撑大模型研究与性能评估的开源数据集

Table 4 Commonly used open-source datasets that support large model research and evaluation

数据集名称	检测对象	缺陷种类	样本数	主要特点	对大模型研究的支撑性
NEU surface defect database	钢材	6 种典型表面缺陷(龟裂、斑块、夹杂、麻点、压入氧化皮及划痕)	1 800	经典、类别均衡	小样本学习、分类任务基准
Severstal-steel-defect	钢材	4 种(麻点、疤痕、斑点、人字纹)	6 666	标注为分割掩模	缺陷分割任务基准
Magnetic tile surface defect[61]	磁砖	5 种(气孔、裂纹、磨损、断裂、凹凸不平)	1 344	背景干净,缺陷特征明显	轻量化模型、边缘部署测试
PCB Defects[62]	印刷电路板	6 种(缺失孔、鼠咬、开路、短路、毛刺、多余铜)	1 386	背景高度结构化	小样本学习、高精度检测
Aitex fabric image database[63]	织物	12 种(毛球、割边等)	245	织物纹理复杂	复杂背景下的异常发现
Defect Spectrum[64]	多种工业对象	125 种(划痕、裂纹、变形、装配错误等)	5 438	缺陷种类丰富	验证模型泛化能力
MVTec-AD[65]	15 种纹理和物品	70 种(划痕、裂纹、凹陷、斑点等)	5 354	学术界基准, 提供 pixel-level 标注	评估 MLLM 的 精细定位与描述能力
VisA[66]	12 种工业对象	67 种(划痕、破损、凹陷、缺失等)	10 821	规模较大,物体种类多	微调、多物体类别泛化测试
MMAD[67]	38 类工业产品	244 种(污染、缺失、磨损、装配错误等)	8 366	多模态问答基准,语义标注丰富	全面评估 MLLM 的感知、推理、报告生成能力

MVTec-AD 和 VisA 数据集因其高质量的像素级标注和多样性,是工业视觉异常检测领域评估大模型视觉定位与分割能力的有效性基准。MMAD 数据集是由南方科技大学和腾讯优图实验室等机构联合创建的首个用于工业异常检测的多模态大语言模型综合基准,代表了大模型时代缺陷检测开源数据集的发展方向。MMAD 数据集包含 8 366 张工业图像,涵盖 38 类产品和 244 种缺陷类型,生成了 39 672 道多选问题,形成了工业领域最全面的多模态大模型能力测评基准。问题覆盖七大核心子任务的测评,即异常识别、缺陷分类、缺陷定位、缺陷描述、缺陷分析、产品分类以及产品分析。数据集的创建过程使用了 GPT-4V 生成丰富的语义标注,并通过人工审核确保问题和选项的合理性与准确性。MMAD 数据集主要应用于工业质量检测领域,旨在评估和提升多模态大模型在工业异常检测任务中的性能,解决传统方法在灵活性和详细报告生成方面的不足。

3.2 性能评价方法

性能评价是衡量大模型在工业缺陷检测领域表现的重要依据,涵盖从计算效率到检测精度等多个维度,以确保系统在实际应用中的高效性和可靠性。对于大模型的性能评价,目前主要有两种方式:一种是自动评价,另一种是人工评价[3]。自动评价是深度学习领域的一种广泛使用的评估方法,一般基于客观指标,即模型预测结果与实际值或参考值之间的量化差异来评估模型性能。工业缺陷检测领域的常用客观指标包括准确率、平均精度均值、推理延迟等。人工评价通常基于评估者个人感受、专业判断等主观指标来评估模型性能。常用的主观指标包括相关度、完整度、专业性、连贯性、有效性及图文一致性等。客观指标不受人的主观感受或偏见影响,能够提供客观准确的模型性能评估,但难以全面地评估模型输出内容。主观指标更加符合实际应用场景,能够提供更为全面的评价,但受评估者的主观性影响,如经验、偏好、知识等。在实际应用中,需根据实际情况选择合适的评价方式,必要时可综合使用两种方式。

3.3 现有应用方向

在工业缺陷检测领域,现有研究大都是针对特定对象如钢铁[68-71]、电子[72-75]、汽车[76-79]、风机叶片[80-81]、印刷品[82-85]等,及特定任务,如表面瑕疵识别[86-87]、焊缝检测[88-89]、内部分层探查[90-91]等,并设计特定算法和模型来适应具体的缺陷类型和生产环境。大模型凭借其大数据驱动、多模态融合和强泛化性等优势,理论上能够提供更通用的解决方案。然而,两者之间存在固有矛盾:一方面,工业场景下的缺陷检测通常面临数据稀缺、标注困难

等问题,难以满足大模型对海量标注数据的需求;另一方面,大模型的复杂性和资源消耗与工业场景对实时性和高效性的要求存在冲突。因此,基于大模型的工业缺陷检测方法仍处于探索阶段,相关成果较为有限。如表 5 所示,经过系统调研后,本文认为现有研究主要聚焦于以下 5 个方向:缺陷检测与定位、复杂场景与微小缺陷检测、小样本与零样本自适应检测、交互式缺陷分析与决策支持、缺陷数据生成与自动标注。

表 5　大模型在工业缺陷检测领域的应用方向(案例)

Table 5　The application directions of large model in the field of industrial defect detection (case studies)

应用方向	文献来源/发布者	模型	应用场景	功能效果
缺陷检测与定位	[92]	AnomalyGPT	通用工业场景	在无需手动设置阈值的情况下判断图像中是否含有异常部分并指出异常位置
	哈工大重庆研究院	自研多模态大模型	半导体、注塑件/金属件、柔性材料等工业产品	集缺陷检测、缺陷定位、缺陷分类三大核心功能为一体,在多种业务场景测试中实现了“零样本”工业缺陷检测
	[93]	PSAM	增材制造过程监控	在训练样本数量仅为 50 的情况下,PSAM 表现出了良好的分割性能,平均交并比可达到 65.02%
复杂场景与微小缺陷检测	[94]	基于 V-MoE 视觉大模型的轨交视觉大模型	复杂隧道壁潜在风险和缺陷检测	能精准识别隧道壁渗水、材料脱落和霉变,准确率达 95%;对更细小裂纹的识别率为 78%;在光线不均匀或有阴影的环境下,模型表现稳定
	[95]	视觉大模型 SAM	电机磁瓦和钢轨	在面对包括小缺陷实例和其他干扰杂质的不同实例缺陷图像时,能够有效地分割出各个实例
小样本与零样本自适应检测	联想研究院	自研工业质检基础大模型	屏幕、螺丝等质检场景	仅需少量正常样本即可快速建模以识别产品异常,在某些场景中甚至无需训练即可直接进行质检推理;具备泛化能力和成熟的工具套件,能快速适应多种产品和生产线
	[96]	基于 CLIP 模型的 PromptAD 方法	通用工业场景	能够在只有少量正常样本的情况下学习有效的提示,减少了对大量标注数据的依赖,适用于实际工业场景中数据有限的情况
交互式缺陷分析与决策支持	[97]	Myriad(基于 MiniGPT-4)	通用工业场景	能提供详细的缺陷描述,包括缺陷的位置、类别、颜色和严重程度等信息;且具备多轮对话能力,可以与用户进行交互,提供更深入的缺陷分析和解释
	[98]	CLAD	机械故障、材料缺陷或工艺偏差等工业场景	通过引入上下文推理模块,CLAD 能够为检测到的缺陷生成文本解释,从而为模型决策提供有价值见解,提高了检测过程的可解释性
缺陷数据生成与自动标注	高视科技	GoMind-LVM	屏幕、半导体等工业产品	能在 1 h 内,从一张样本图像生成上万张仿真缺陷图像
	华院计算	钢铁大模型	钢铁缺陷检测	能够自动生成多模态数据,不仅提高了模型的准确性,还有效缓解了长尾分布对模型性能的负面影响
	凌云光	LuserLVM 工业领域通用视觉大模型	锂电极片涂布、PVD 手机中框外观缺陷检测等	针对不同行业和应用场景进行了优化。在缺陷生成、辅助标注和缺陷提示等方面,展现出了卓越的性能,大幅提升工业质检的效率和精度

1)缺陷检测与定位

工业生产中,缺陷的精确检测与定位是保障产品质量的核心环节。大模型凭借其强大的视觉语言理解能力和深度学习能力,能够对工业图像进行分析,判断是否存

在异常,并精确地定位缺陷区域的像素位置。同时,它还能对不同类型的缺陷进行分类和分级,为产品质量评估提供依据。具体应用案例为:

2023 年 8 月,中科视语联合中国科学院自动化研究所推出了工业异常检测大模型 AnomalyGPT[92]。该模型利用大模型的强大语义理解能力,通过设计的图像解码器和提示嵌入微调方法,使大模型能够充分理解工业场景图像。且能在无需手动设置阈值的情况下判断图像中是否含有异常部分并指出异常位置,在 MVTec-AD 数据集上,实现了 86.1% 的准确率,94.1% 的图像级曲线下面积(area under curve,AUC)和 95.3% 的像素级 AUC,推动了大模型在工业领域的落地应用。2023 年 9 月,哈工大重庆研究院工业视觉研究中心(https://app.cqrb.cn/html/2023-09-18/1531883_pc.html)成功将前沿 AI 大模型技术转化落地,自主研发了一套面向工业缺陷检测的多模态大模型,并将其集成到了该中心研发的视觉检测系列化设备中。该模型融合了缺陷检测、缺陷定位、缺陷分类三大核心功能,在半导体、注塑件/金属件、柔性材料等工业产品的测试中实现了“零样本”工业缺陷检测。这一创新为解决工业质检场景下需求分散、换产频繁、缺陷类型多样等痛点问题,提供了一种新的解决思路。2024 年 5 月,西安交通大学机械工程学院航空发动机研究所[93]提出了一种基于视觉大模型的激光粉末床熔融铺粉缺陷检测方法,在训练样本数量仅为 50 的情况下,模型表现出了良好的分割性能,平均交并比可达到 65.02%,相较于 Deeplab v3 和 Unet 分别提升了 8.52% 和 5.31%,展示了视觉大模型在增材制造过程监控中的应用价值和潜力。

2)复杂场景与微小缺陷检测

在复杂工业场景中,如多材质混合、强背景干扰等,以及面对微小缺陷时,检测难度会显著增加。大模型凭借其强大的特征提取和分析能力,能够在复杂环境下和微小尺度上精准识别缺陷,有效提升了检测的精度和可靠性,为高质量生产提供了重要保障。具体应用案例为:

上海点泽智能科技有限公司基于 V-MoE 视觉大模型开发了轨交视觉大模型。与传统的机器视觉相比,在复杂隧道壁潜在风险识别和缺陷检测方面,该方法的识别类型和识别精度均大幅增加[94]。文献[95]将视觉大模型分割一切模型(segment anything model,SAM)应用于工业材料表面缺陷检测任务,通过对图像编码器和掩码解码器进行微调,在电机磁瓦和钢轨两个缺陷数据集上进行实验评估,对微小缺陷的检测效果验证了其在面对复杂微小缺陷时的泛化能力,为工业场景提供了一种通用高效的新型缺陷检测方案。

3)小样本与零样本自适应检测

在实际的工业生产中,获取并标注大量样本往往成本高昂且耗时费力。大模型的小样本与少样本学习能力则有效解决了这一问题。通过上下文学习、生成模拟缺陷数据等方法,大模型仅需少量正常样本即可快速适应新场景,实现精准的缺陷检测,大大降低了数据采集和标注成本,提高了生产效率。具体应用案例为:

2024 年 10 月,联想研究院人工智能实验室(https://mbrand.lenovo.com.cn/brand/ppn03004.html)推出了一个基于自研工业质检基础大模型的边缘大脑人工智能小样本终身学习质检平台。该平台仅需少量正常样本即可快速建模以识别产品异常,在某些场景中甚至无需训练即可直接进行质检推理。同时,该方案具备泛化能力和成熟的工具套件,使其能快速适应多种产品和生产线,提供精准高效的质检,显著提升检测效率并降低人工成本与误检率。文献[96]提出了基于对比语言图像预训练模型(contrastive language-image pre-training,CLIP)的少样本异常检测的提示学习方法 PromptAD,通过语义拼接将正常提示转换为异常提示构建负样本,并引入显式异常边界控制正常与异常提示特征间距,实现了仅用正常样本进行少样本异常检测。

4)交互式缺陷分析与决策支持

在工业缺陷检测过程中,人机协作的效率和透明度直接影响着检测结果的准确性和可靠性。支持多轮对话的大模型缺陷检测系统允许用户通过自然语言提问,动态调整检测重点。例如,操作员可针对可疑区域提问,如“第 3 张图中右侧边缘是否存在裂纹”,模型实时生成带定位标记的响应及描述,并输出置信度评分,支持交互式参与,使用户能够根据需求和所提供的答案提出后续问题。这种交互模式提升了检测透明度和人机协作效率。具体应用案例为:

2023 年 10 月底,哈尔滨工业大学左旺孟团队提出了利用视觉专家进行工业异常检测的大型多模态模型 Myriad[97]。该模型通过应用视觉专家进行工业异常检测,不仅能够进行准确的缺陷检测,还能提供详细的描述,包括缺陷的位置、类别、颜色和严重程度等信息,且具备多轮对话能力,可以与用户进行交互,提供更深入的异常分析和解释。文献[98]提出了一种通过对比交叉模态训练的视觉-语言缺陷检测新方法。通过引入上下文推理模块,该方法能够为检测到的缺陷生成文本解释,提高检测过程的可解释性。

5)缺陷数据生成与自动标注

在实际工业场景中,正常产品的比例往往远高于缺陷产品,导致收集到的缺陷样本数量有限。生成式大模型可基于语义描述及已有缺陷样本,自动合成多样化的异常图像并标注缺陷类型与区域,有效缓解了数据稀缺问题。具体应用案例为:

2024 年 7 月,高视科技(http://govion.cn/newsinfo/7338104.html)发布 GoMind-LVM 大模型,通过分层开发

方法,构建了基础大模型以及针对特定行业应用的工业缺陷生成模型、工业缺陷辅助标注模型和工业缺陷检测模型,能在 1 h 内,从一张样本图像生成上万张仿真缺陷图像。华院计算(https://www.163.com/dy/article/JIGEEBVU05564K6P.html)推出的钢铁大模型融合了数据与知识方法,基于认知智能引擎平台底层能力,结合多年钢铁冶金行业模型训练能力与多模态大模型技术,能够自动生成多模态数据,不仅提高了模型的准确性,还有效缓解了长尾分布对模型性能的负面影响。凌云光(https://www.lusterinc.com)推出的 LuserLVM 工业领域通用视觉大模型,通过分层设计,既满足了基础大模型的通用性需求,又针对不同行业和应用场景进行了优化。在缺陷生成、辅助标注和缺陷提示等方面,该模型展现出了卓越的性能,能大幅提升工业质检的效率和精度,而无需依赖大量的算力,从而能快速部署在锂电等 10 多种工业视觉检测环境中。目前已在锂电极片涂布缺陷检测、物理气相沉积(physical vapor deposition,PVD)手机中框外观缺陷检测和显示屏屏幕模组外观检测等场景中取得显著成效。

4　面临的挑战与未来发展趋势

4.1　面临的挑战

1)数据质量与安全

在实际的工业生产环境中,传感器采集的数据可能会受到各种因素的干扰,导致数据中存在噪声和误差。这些噪声和误差会影响大模型的训练效果和预测准确性,使得模型难以准确地检测出缺陷。工业大模型应用中,高质量数据是基础,但目前大多数情况下高质量数据的供给不足,这直接影响了模型的训练效果和应用性能[99]。且大模型在训练过程中很少接触到工业缺陷检测领域的专业知识,使得其对某些产品的缺陷类型和异常模式理解不足。目前数据的开放共享机制不完善,限制了高质量数据资源的有效利用和共享。

大模型依赖于海量数据进行训练与微调,数据获取、处理和使用等环节都面临着隐私泄露的风险。工业缺陷检测通常涉及大量敏感数据,如生产参数、设备状态、材料特性、产品设计细节、客户信息等。这些数据一旦泄露,可能导致严重的技术泄密和商业损失。且大模型普遍依赖于公有云环境提供服务,会引发使用机构对其私有缺陷数据及敏感数据安全的担忧[2]。如 2023 年,三星电子在引入 ChatGPT 后的短短 20 天内,就遭遇 3 起内部数据泄露事件,其中两起与半导体设备程序代码有关,另一起与内部会议记录有关。

2)高可靠性要求

在工业缺陷检测中,任何微小的错误或不确定性都可能导致严重的后果,如产品质量下降、生产效率降低甚至安全事故。而大模型的不可解释的“黑盒”特性及固有的训练数据问题,使其易出现“幻觉”,输出结果存在严重的可信性问题。大模型的“幻觉”问题是伴随着人工智能底层技术路径与生俱来的,和创新能力一体两面,难以单凭技术手段彻底消除。且在大规模生产中,需要对大量的产品进行缺陷检测,大模型需要在批量检测中保持高度的一致性,确保每个产品的检测结果都具有可靠性和可比性。但大模型的概率化输出特性与工业场景的确定性需求存在矛盾。幻觉检测、知识编辑、训练数据质量增强、知识增强等方法,一定程度上缓解了多模态大模型的部分可信性问题,但多模态大模型仍然存在严重的事实性错误问题。

3)成本限制与可持续发展

虽然 AI 大模型在工业缺陷检测领域的应用能够带来长期效益,但初期的投入成本也相对较高,大模型的训练和部署都有高资源、高显存、高存储需求。且目前缺少对大模型落地应用的统一效益评估标准,短期内无法看到明显的回报,应用成本过高,增加了企业的经济压力[11]。同时,大模型的训练和部署需要大量算力支持,能耗和碳排放问题突出,环境影响不容忽视[100]。《2024 年人工智能指数报告》指出,训练 GPT-3 参数量级的大模型耗电量堪比数百次跨美飞行,ChatGPT 的日常运行能耗则相当于约 19 000 户家庭的日用电量[101]。此外,电力消耗产生的大量热量使得用于设备冷却的水资源消耗也同样惊人。OpenAl CEO 山姆·奥特曼坦言人工智能行业正面临能源危机,下一代生成式 AI 系统能耗可能超出系统承载极限。因此,如何在成本与效益之间找到平衡点,实现可持续发展,是大模型落地应用必须面对的问题。

4)缺乏统一测评标准

缺乏统一测评标准是大模型在工业缺陷检测领域规模化落地的关键瓶颈之一,当前各行业、企业甚至项目间采用的评估体系存在较大差异,导致模型性能难以横向对比与迭代优化。一方面,现有通用指标,如准确率、召回率等,往往过于笼统,无法精准反映工业场景的核心需求,如微小缺陷检出率、极端工况稳定性等,而自定义指标又缺乏普适性;另一方面,测试数据集的质量与覆盖度参差不齐,部分企业依赖少量内部数据评估,忽视了长尾分布与动态环境干扰的挑战,而公开数据集又难以涵盖所有代表性工业场景。此外,评估流程的标准化程度不足,从数据预处理、测试环境配置到结果统计方法均存在显著差异,导致同一模型在不同测评体系下表现波动巨大,这不仅增加了模型对比选择的难度,也阻碍了跨企业技术协作与生态共建。为解决这一问题,急需构建涵盖“数据-指标-流程”三位一体的工业级测评标准体系,通过引入领域专家共识、物理约束验证与动态环境模拟,确保评估结果的可信度与可比性。

4.2 发展趋势

1)多模态工业质检数据

针对工业缺陷检测这一具体应用场景,目前的多模态大模型多基于常见的图像、文本、视频等模态,局限于表面缺陷检测。而工业缺陷远不止表面纹理异常,随着传感器技术的不断进步,未来将会融合红外热成像、X 射线、超声、高光谱、3D 点云以及电磁等多物理场传感数据,构建对缺陷的立体化、深层次感知能力。例如,超声成像能精准探测内部气泡与裂纹[102-103];红外热成像可通过热异常定位金属裂纹和复合材料内部缺陷[104-105];3D 点云能精确量化缺陷的深度与体积[106-107];高光谱能揭示缺陷与正常区域的内在物理化学差异[108-109]。多模态融合能有效克服单一传感器在特定环境下的感知局限,显著提高检测的准确性和鲁棒性。

2)大小模型协同的缺陷检测范式

大模型凭借其强大的泛化能力和处理复杂任务的优势,在工业缺陷检测领域展现出巨大潜力[110]。尽管大模型发展迅猛,但在工业缺陷检测领域的实际应用中,其高计算成本、高延迟和"黑箱"特性使其无法替代针对特定场景和任务的高精度、高实时性小模型。展望未来,二者将演化出新的工业级协同范式,即以云端大模型为知识底座、边缘侧小模型为区域协调器、终端微型模型为感知执行单元。具体而言,云端大模型利用海量工业数据进行预训练,提炼通用缺陷表征、建立异常模式识别能力,并通过在线蒸馏与数据生成,赋能下游小模型;边缘侧模型则针对特定产线、任务进行轻量化微调,实现快速适配与推理;终端模型则嵌入检测设备,保障实时检测。这种分工协作的范式,能够同时兼顾大模型的泛化能力与小模型的效率与可靠性,是推动大模型在工业领域落地的可行路径。

3)增强模型可解释性与人机协作

工业领域高度强调过程控制与质量追溯,因此在大模型驱动的工业缺陷检测中,增强模型的可解释性不仅是技术需求,更是工业应用落地的关键。包括大模型在内的深度学习模型常被视为"黑箱",其决策过程缺乏透明性,难以满足工业场景对可靠性和可追溯性的严格要求。通过引入行业知识增强模型可解释性已成为目前的主流方法。增强可解释性不仅可以提升工业用户对大模型的信任度,还能推动检测系统从"结果输出"向"决策支持"演进。将模型决策与生产参数进行关联分析,帮助工程师追溯缺陷产生的工艺环节与设备根源,实现从"质量检测"到"工艺优化"的赋能跃迁。

4)建立工业缺陷检测大模型统一评价框架

目前大模型在工业缺陷检测领域缺乏统一评价标准,导致模型性能难以横向对比,且依赖的通用指标难以全面衡量其在真实工业环境下的综合性能。未来需构建融合工业先验知识的跨模态、跨任务、跨场景的统一评价框架,提升评测结果在实际应用中的参考价值。从而更好地评估工业缺陷检测大模型的能力和局限性,推动模型优化和技术创新,提高工业缺陷检测中大模型的应用效果和可靠性。同时,统一的评价框架也有利于不同模型之间的比较和选择,为工业企业在实际应用中选择合适的大模型提供参考,为大模型在工业缺陷检测等领域的可持续发展提供有力支持。

5)具身智能驱动的主动式缺陷检测系统

工业缺陷检测方法正经历从传统被动静态分析向主动感知交互的具身智能的范式变革。"具身智能"是指将人工智能系统与物理环境中的机器人或传感器等紧密结合,使系统能够通过感知、行动和认知来与环境互动,从而显著提升系统的自主性和适应性。具身智能通过构建"感知-决策-执行"闭环的检测物理实体,如搭载多模态传感器的机械臂、移动机器人、无人机等,突破单一识别功能,适用于复杂结构件、大型设备,如风机叶片、船体等传统固定式检测难以应对的场景。大模型驱动的具身智能可实现缺陷定位、分类、修复的全流程自动化,大幅减少人工干预,提高检测效率,推动工业质检从"算法辅助"迈向"自主进化"的新阶段,为智能制造提供具备实体交互能力的新方案。

5 结 论

工业缺陷检测是智能制造的关键环节之一,大模型的兴起无疑为工业缺陷检测领域带来了新的变革并注入了新的活力。本文对大模型的发展历程、聚焦于工业缺陷检测的核心技术进行了系统梳理。并将大模型与传统 AI 缺陷检测方法进行对比,归纳了各自的优点和局限性。进一步着眼于工业缺陷检测领域,对支撑大模型研究与性能评估的关键开源数据集、大模型性能评价指标及现有应用方向进行了详细介绍。然而,工业缺陷检测大模型在推动行业变革的同时,也面临着数据质量与安全、高可靠性要求、成本限制与可持续发展、缺乏统一测评标准等多重挑战。这些挑战不仅关乎大模型自身的健康发展,更直接影响到其在工业缺陷检测领域中的广泛应用。因此,必须在技术创新的同时,加强法规建设与审查,确保大模型在合法、合规、安全的轨道上稳步前行。展望未来,大模型在工业缺陷检测领域的应用呈现出利用多模态工业质检数据、大小模型协同缺陷检测、增强模型可解释性与人机协作能力、建立工业缺陷检测大模型统一评价框架、具身智能驱动的主动式缺陷检测系统等多元化发展趋势。大模型将在工业缺陷检测领域持续发挥变革性作用,并成为驱动产业升级与社会进步的核心技术力量。

参考文献

[1] 罗东亮, 蔡雨萱, 杨子豪, 等. 工业缺陷检测深度学习方法综述[J]. 中国科学:信息科学, 2022, 52(6): 1002-1039.
LUO D L, CAI Y X, YANG Z H, et al. Survey on industrial defect detection with deep learning [J]. Scientia Sinica (Informationis), 2022, 52(6): 1002-1039.

[2] 腾讯研究院大模型研究课题组. AI 行业大模型培育新质生产力[J]. 企业管理, 2024(7): 56-63.
Tencent Research Institute Large Model Research Group. AI Industry-specific large models cultivate new quality productivity[J]. Enterprise Management, 2024(7): 56-63.

[3] 刘学博, 户保田, 陈科海, 等. 大模型关键技术与未来发展方向——从 ChatGPT 谈起[J]. 中国科学基金, 2023, 37(5): 758-766.
LIU X B, HU B T, CHEN K H, et al. Key technologies and future development direction of large language models: Insights from ChatGPT[J]. Bulletin of National Natural Science Foundation of China, 2023, 37(5): 758-766.

[4] 闫奕樸, 刘桂雄, 邢星奥. 工业异常检测大模型方法研究进展[J]. 中国测试, 2025, 51(1): 1-10, 23.
YAN Y P, LIU G X, XING X AO. Progress of research on large visual language model methods for industrial anomaly detection [J]. China Measurement & Test, 2025, 51(1): 1-10, 23.

[5] YANG T L, CHANG L Y, YAN J D, et al. A survey on foundation-model-based industrial defect detection [J]. ArXiv preprint arXiv: 2502.19106, 2025.

[6] 邢永康, 马少平. 统计语言模型综述[J]. 计算机科学, 2003, 30(9): 22-26.
XING Y K, MA SH P. A survey on statistical language models[J]. Computer Science, 2003, 30(9): 22-26.

[7] 尹陈, 吴敏. N-gram 模型综述[J]. 计算机系统应用, 2018, 27(10): 33-38.
YIN CH, WU M. Survey on N-gram model [J]. Computer Systems & Applications, 2018, 27(10): 33-38.

[8] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.

[9] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]. 27th International Conference on Neural Information Processing Systems, 2013: 3111-3119.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]. The 31st International Conference on Neural Information Processing Systems, 2017: 6000-6010.

[11] 腾讯研究院. 向 AI 而行,共筑新质生产力——行业大模型调研报告[R]. 北京: 腾讯研究院, 2024.
Tencent Research Institute. Towards AI, building new productive forces together: Industry large model research report[R]. Beijing: Tencent Research Institute, 2024.

[12] 何友, 刘瑜, 李耀文, 等. 多源信息融合发展及展望[J]. 航空学报, 2025, 46(6): 29-54.
HE Y, LIU Y, LI Y W, et al. Development and prospects of multisource information fusion [J]. Acta Aeronautica et Astronautica Sinica, 2025, 46(6): 29-54.

[13] 张乾君. AI 大模型发展综述[J]. 通信技术, 2023, 56(3): 255-262.
ZHANG Q J. Review of AI large model development[J]. Communication Technology, 2023, 56(3): 255-262.

[14] 秦小林, 古徐, 李弟诚, 等. 大语言模型综述与展望[J]. 计算机应用, 2025, 45(3): 685-696.
QIN X L, GU X, LI D CH, et al. Survey and prospect of large language models [J]. Journal of Computer Applications, 2025, 45(3): 685-696.

[15] 史宏志, 赵健, 赵雅倩, 等. 大模型时代的混合专家系统优化综述[J]. 计算机研究与发展, 2025, 62(5): 1164-1189.
SHI H ZH, ZHAO J, ZHAO Y Q, et al. Survey on system optimization for mixture of experts in the era of large models [J]. Journal of Computer Research and Development, 2025, 62(5): 1164-1189.

[16] SUN Y T, DONG L, HUANG SH H, et al. Retentive network: A successor to Transformer for large language models[J]. ArXiv preprint arXiv: 2307.08621, 2023.

[17] GU A, DAO T. Mamba: Linear-time sequence modeling with selective state spaces[J]. ArXiv preprint arXiv: 2312.00752, 2023.

[18] PENG B, ALCAIDE E, ANTHONY Q, et al. RWKV: Reinventing RNNs for the Transformer era[J]. ArXiv preprint arXiv: 2305.13048, 2023.

[19] POLI M, MASSAROLI S, NGUYEN E, et al. Hyena hierarchy: Towards larger convolutional language models[C]. The 40th International Conference on

Machine Learning, 2023: 28043-28078.

[20] NAVEED H, KHAN A U, QIU S, et al. A comprehensive overview of large language models[J]. ArXiv preprint arXiv: 2307.06435, 2023.

[21] PROTTASHA N J, MAHMUD A, SOBUJ M S I, et al. Parameter-efficient fine-tuning of large language models using semantic knowledge tuning[J]. Scientific Reports, 2024, 14(1): 30667.

[22] 李超烽, 顾瑞春, 李卓仑. 基于对比学习和交叉注意力的多模态联邦学习优化方法[J/OL]. 计算机工程与应用, 1-29[2025-06-20].
LI CH F, GU R CH, LI ZH L. Multi-modal federated learning optimization method based on contrastive learning and cross-attention[J/OL]. Computer Engineering and Applications, 1-29[2025-06-20].

[23] NARAYANAN D, SHOEYBI M, CASPER J, et al. Efficient large-scale language model training on GPU clusters using Megatron-LM[C]. International Conference for High Performance Computing, Networking, Storage and Analysis, 2021: 1-15.

[24] 吕正浩, 咸鹤群. 基于鲁棒分区水印的深度学习模型保护方法[J/OL]. 计算机科学, 1-13[2025-06-20].
LYU ZH H, XIAN H Q. Deep learning model protection method based on robust partitioned watermarking[J/OL]. Computer Science, 1-13[2025-06-20].

[25] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. ArXiv preprint arXiv: 1503.02531, 2015.

[26] 张钦彤, 王昱超, 王鹤羲, 等. 大语言模型微调技术的研究综述[J]. 计算机工程与应用, 2024, 60(17): 17-33.
ZHANG Q T, WANG Y CH, WANG H X, et al. Comprehensive review of large language model fine-tuning[J]. Computer Engineering and Applications, 2024, 60(17): 17-33.

[27] TANWISUTH K, ZHANG SH J, ZHENG H J, et al. POUF: Prompt-oriented unsupervised fine-tuning for large pre-trained models[C]. 40th International Conference on Machine Learning, 2023: 33816-33832.

[28] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[C]. 36th International Conference on Machine Learning, 2019: 2790-2799.

[29] 秦董洪, 李政韬, 白凤波, 等. 大语言模型参数高效微调技术综述[J]. 计算机工程与应用, 2025, 61(16): 38-63.
QIN D H, LI ZH T, BAI F B, et al. A review of parameter-efficient fine-tuning technology for large language models[J]. Computer Engineering and Applications, 2025, 61(16): 38-63.

[30] VALIPOUR M, REZAGHOLIZADEH M, KOBYZEV I, et al. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation[C]. 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023: 3274-3287.

[31] ZHANG Q R, CHEN M SH, BUKHARIN A, et al. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning[J]. ArXiv preprint arXiv: 2303.10512, 2023.

[32] DETTMERS T, PAGNONI A, HOLTZMAN A, et al. QLoRA: Efficient finetuning of quantized LLMs[J]. Advances in Neural Information Processing Systems, 2023, 36: 10088-10115.

[33] HU E J, SHEN Y L, WALLIS P, et al. LoRA: Low-rank adaptation of large language models[J]. ArXiv preprint arXiv: 2106.09685, 2021.

[34] 张雨轩, 黄诚, 柳蓉, 等. 结合提示词微调的智能合约漏洞检测方法[J]. 信息网络安全, 2025, 25(4): 664-673.
ZHANG Y X, HUANG CH, LIU R, et al. Smart contract vulnerability detection method combining prompt tuning[J]. Netinfo Security, 2025, 25(4): 664-673.

[35] 吴春志, 赵玉龙, 刘鑫, 等. 大语言模型微调方法研究综述[J]. 中文信息学报, 2025, 39(2): 1-26.
WU CH ZH, ZHAO Y L, LIU X, et al. Fine-tuning methods for large language models: A survey[J]. Journal of Chinese Information Processing, 2025, 39(2): 1-26.

[36] 张钰莹, 云静, 刘雪颖, 等. 基于反馈的大语言模型内容与行为对齐方法综述[J/OL]. 计算机工程与应用, 1-37[2025-06-20].
ZHANG Y Y, YUN J, LIU X Y, et al. A survey of feedback-based content and behavior alignment methods for large language model[J/OL]. Computer Engineering and Applications, 1-37[2025-06-20].

[37] 韩炳涛, 刘涛. 大模型关键技术与应用[J]. 中兴通讯技术, 2024, 30(2): 76-88.
HAN B T, LIU T. Key technologies and applications of large models[J]. ZTE Technology Journal, 2024, 30(2): 76-88.

[38] STIENNON N, OUYANG L, WU J, et al. Learning to summarize from human feedback[C]. 34th International

Conference on Neural Information Processing Systems, 2020: 3008-3021.

[39] LEE H, PHATALE S, MANSOOR H, et al. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback[J]. ArXiv preprint arXiv: 2309.00267, 2023.

[40] ZHOU Z X, NING X F, HONG K, et al. A survey on efficient inference for large language models[J]. ArXiv preprint arXiv: 2404.14294, 2024.

[41] 张浩然, 李君, 邢立宁, 等. 大模型与智能优化算法集成研究综述[J/OL]. 控制与决策, 1-20[2025-06-20].
ZHANG H R, LI J, XING L N, et al. A research review on the integration of large models and intelligent optimization algorithms[J/OL]. Control and Decision, 1-20[2025-06-20].

[42] 陈寿宏, 赵爽, 马峻, 等. 基于多特征的 SVM 多分类 PCB 焊点缺陷检测方法[J]. 激光杂志, 2019, 40(6): 21-26.
CHEN SH H, ZHAO SH, MA J, et al. Solder joint defect detection method of SVM multi-classification PCB based on multi-feature [J]. Laser Journal, 2019, 40(6): 21-26.

[43] 王照, 葛馨远, 饶毅. 基于 SVM 多特征融合的绝缘子缺陷检测算法研究[J]. 自动化技术与应用, 2024, 43(5): 83-88.
WANG ZH, GE X Y, RAO Y. Research on insulator defect detection algorithm based on SVM multi-feature fusion[J]. Techniques of Automation and Applications, 2024, 43(5): 83-88.

[44] ABDELLAH H, AHMED R, SLIMANE O. Defect detection and identification in textile fabric by SVM method [J]. IOSR Journal of Engineering, 2014, 4(12): 69-77.

[45] 杨水山, 何永辉, 赵万生. Boosting 优化决策树的带钢表面缺陷识别技术[J]. 红外与激光工程, 2010, 39(5): 954-958.
YANG SH SH, HE Y H, ZHAO W SH. Strip steel surface defect recognition based on Boosting optimized decision[J]. Infrared and Laser Engineering, 2010, 39(5): 954-958.

[46] MADDEH M, AYOUNI S, ALYAHYA S, et al. Decision tree-based design defects detection[J]. IEEE Access, 2021, 9: 71606-71614.

[47] 刘传泽, 王霄, 陈龙现, 等. 基于随机森林算法的纤维板表面缺陷识别[J]. 林业科学, 2018, 54(11): 121-126.
LIU CH Z, WANG X, CHEN L X, et al. Surface defect recognition of fiberboard based on random forest[J]. Scientia Silvae Sinicae, 2018, 54(11): 121-126.

[48] SHIPWAY N J, BARDEN T J, HUTHWAITE P, et al. Automated defect detection for fluorescent penetrant inspection using random forest [J]. NDT & E International, 2019, 101: 113-123.

[49] 黄露, 夏军勇, 吴庆华, 等. 基于遗传算法与二维最大熵的编织袋缺陷检测[J]. 轻工机械, 2021, 39(5): 69-73,78.
HUANG L, XIA J Y, WU Q H, et al. Woven bag defect detection based on genetic algorithm and two-dimensional maximum entropy[J]. Light Industry Machinery, 2021, 39(5): 69-73,78.

[50] SU L, SHI T L, DU L, et al. Genetic algorithms for defect detection of flip chips [J]. Microelectronics Reliability, 2015, 55(1): 213-220.

[51] WANG T, CHEN Y, QIAO M N, et al. A fast and robust convolutional neural network-based defect detection model in product quality control [J]. The International Journal of Advanced Manufacturing Technology, 2018, 94(9/12): 3465-3471.

[52] 蔡彪, 沈宽, 付金磊, 等. 基于 Mask R-CNN 的铸件 X 射线 DR 图像缺陷检测研究[J]. 仪器仪表学报, 2020, 41(3): 61-69.
CAI B, SHEN K, FU J L, et al. Research on defect detection of X-ray DR images of casting based on Mask R-CNN[J]. Chinese Journal of Scientific Instrument, 2020, 41(3): 61-69.

[53] NIU SH L, LI B, WANG X G, et al. Defect image sample generation with GAN for improving defect recognition [J]. IEEE Transactions on Automation Science and Engineering, 2020, 17(3): 1611-1622.

[54] 段宣尧, 陈雪云, 许韬. 基于条件 GAN 的电子元件缺陷检测研究[J]. 计算机应用研究, 2020, 37(S2): 395-397.
DUAN X Y, CHEN X Y, XU T. Conditional GAN-based defect detection for electronic components [J]. Application Research of Computers, 2020, 37(S2): 395-397.

[55] 李原, 李燕君, 刘进超, 等. 基于改进 Res-UNet 网络的钢铁表面缺陷图像分割研究[J]. 电子与信息学报, 2022, 44(5): 1513-1520.
LI Y, LI Y J, LIU J CH, et al. Research on segmentation of steel surface defect images based on

improved Res-UNet network[J]. Journal of Electronics & Information Technology, 2022, 44(5): 1513-1520.

[56] JING J F, WANG ZH, RÄTSCH M, et al. Mobile-Unet: An efficient convolutional neural network for fabric defect detection[J]. Textile Research Journal, 2022, 92(1/2): 30-42.

[57] BINOMAIRAH A, ABDULLAH A, KHOO B E, et al. Detection of microcracks and dark spots in monocrystalline PERC cells using photoluminescene imaging and YOLO-based CNN with spatial pyramid pooling[J]. EPJ Photovoltaics, 2022, 13: 27.

[58] 周亚罗, 武献超, 刘文广, 等. 基于 STCS-YOLO 的带钢表面缺陷检测算法[J]. 中国冶金, 2023, 33(12): 128-138.
ZHOU Y L, WU X CH, LIU W G, et al. Defect detection algorithm of strip surface based on STCS-YOLO[J]. China Metallurgy, 2023, 33(12): 128-138.

[59] 刘义艳, 郝婷楠, 贺晨, 等. 基于 DBBR-YOLO 的光伏电池表面缺陷检测[J]. 图学学报, 2024, 45(5): 913-921.
LIU Y Y, HAO T N, HE CH, et al. Photovoltaic cell surface defect detection based on DBBR-YOLO[J]. Journal of Graphics, 2024, 45(5): 913-921.

[60] 孙林, 张子洋, 贾坤昊, 等. 工业缺陷图像生成:非线性重构与多级滤波优化[J/OL]. 计算机工程与应用, 1-16[2025-06-20].
SUN L, ZHANG Z Y, JIA K H, et al. Industrial defect image generation: nonlinear reconstruction and multi-level filtering optimization[J/OL]. Computer Engineering and Applications, 1-16[2025-06-20].

[61] HUANG Y B, QIU C Y, GUO Y, et al. Surface defect saliency of magnetic tile[C]. 2018 IEEE 14th International Conference on Automation Science and Engineering, 2018: 612-617.

[62] HUANG W B, WEI P, ZHANG M H, et al. HRIPCB: A challenging dataset for PCB defects detection and classification[J]. The Journal of Engineering, 2020, 2020(13): 303-309.

[63] SILVESTRE-BLANES J, ALBERO-ALBERO T, MIRALLES I, et al. A public fabric database for defect detection methods and results[J]. Autex Research Journal, 2019, 19(4): 363-374.

[64] YANG SH, CHEN ZH F, CHEN P G, et al. Defect Spectrum: A granular look of large-scale defect datasets with rich semantics[C]. Computer Vision-ECCV 2024, 2024: 187-203.

[65] BERGMANN P, FAUSER M, SATTLEGGER D, et al. MVTec AD — a comprehensive real-world dataset for unsupervised anomaly detection[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 9584-9592.

[66] ZOU Y, JEONG J, PEMULA L, et al. SPot-the-difference self-supervised pre-training for anomaly detection and segmentation[C]. European Conference on Computer Vision, 2022: 392-408.

[67] JIANG X, LI J, DENG H Q, et al. MMAD: A comprehensive benchmark for multimodal large language models in industrial anomaly detection[J]. ArXiv preprint arXiv: 2410.09453, 2024.

[68] 唐湘龙, 石兰娟, 陶利民, 等. GCA-YOLO:一种改进的钢铁表面缺陷检测算法[J]. 杭州师范大学学报(自然科学版), 2025, 24(1): 9-18.
TANG X L, SHI L J, TAO L M, et al. GCA-YOLO: An improved steel surface defect detection algorithm[J]. Journal of Hangzhou Normal University (Natural Science Edition), 2025, 24(1): 9-18.

[69] 张瑞芳, 伏铭强, 程小辉. 基于 YOLOv5s 的钢铁表面缺陷检测算法[J]. 科学技术与工程, 2024, 24(23): 9980-9988.
ZHANG R F, FU M Q, CHEN X H. Steel surface defect detection algorithm based on YOLOv5s[J]. Science Technology and Engineering, 2024, 24(23): 9980-9988.

[70] 崔克彬, 焦静颐. 基于 MCB-FAH-YOLOv8 的钢材表面缺陷检测算法[J]. 图学学报, 2024, 45(1): 112-125.
CUI K B, JIAO J Y. Steel surface defect detection algorithm based on MCB-FAH-YOLOv8[J]. Journal of Graphics, 2024, 45(1): 112-125.

[71] 李跃, 王子铭, 李鑫林, 等. 带钢表面缺陷检测方法研究进展[J]. 钢铁研究学报, 2023, 35(8): 950-962.
LI Y, WANG Z M, LI X L, et al. Research progress on surface defect detection methods of strip steel[J]. Journal of Iron and Steel Research, 2023, 35(8): 950-962.

[72] 范博淦, 王淑青, 陈开元. 基于改进 YOLOv8 的印刷电路板缺陷检测模型[J]. 现代电子技术, 2025, 48(11): 144-150.
FAN B G, WANG SH Q, CHEN K Y. Printed circuit board defect detection model based on improved YOLOv8[J]. Modern Electronic Technique, 2025,

48(11): 144-150.

[73] 吴葛, 朱宇凡, 贾泽宁. 改进 YOLO11 的 PCB 表面缺陷检测方法[J]. 电子测量技术, 2025, 48(14): 136-145.
WU G, ZHU Y F, JIA Z N. Improved PCB surface defect detection method based on YOLO11 [J]. Electronic Measurement Technology, 2025, 48(14): 136-145.

[74] 殷旭鹏, 赵兴强. YOLOv11-MAS:一种高效的 PCB 缺陷检测算法[J]. 计算机工程与应用, 2025, 61(17): 102-111.
DUAN X P, ZHAO X Q. Improved YOLOv11-based algorithm for PCB defect detection [J]. Computer Engineering and Applications, 2025, 61(17): 102-111.

[75] 胡志强, 吴一全. 基于机器视觉的半导体晶圆缺陷检测方法综述[J]. 中国图象图形学报, 2025, 30(1): 25-50.
HU ZH Q, WU Y Q. Survey of semiconductor wafer defect detection method based on machine vision [J]. Journal of Image and Graphics, 2025, 30(1): 25-50.

[76] 曾治霖, 瞿昊, 杜正春. 基于深度学习和生成对抗网络的发动机缸体表面缺陷检测方法[J]. 机械工程学报, 2025, 61(2): 46-55.
ZEN ZH L, QU H, DU ZH CH. An engine cylinder surface defect detection algorithm based on the YOLOv5 network and Pix2Pix model [J]. Journal of Mechanical Engineering, 2025, 61(2): 46-55.

[77] 袁海兵, 杨奕洋, 赵凤胜, 等. 改进 YOLOv8 的汽车齿轮齿面缺陷检测研究[J]. 仪表技术与传感器, 2024(12): 91-99,106.
YUAN H B, YANG Y Y, ZHAO F SH, et al. Study on defect detection of automotive gear tooth surface of improved YOLOv8 [J]. Instrument Technique and Sensor, 2024(12): 91-99,106.

[78] 周晓龙, 刘常杰. 基于改进 YOLOv5 的车辆焊缝气孔缺陷检测方法[J]. 激光与光电子学进展, 2025, 62(4): 129-135.
ZHOU X L, LIU CH J. Defect detection method for vehicle weld porosity based on improved YOLOv5 [J]. Laser & Optoelectronics Progress, 2025, 62(4): 129-135.

[79] 叶娜, 周学良, 张映锋, 等. 改进 YOLOv8 的汽车制动器装配缺陷检测算法[J]. 组合机床与自动化加工技术, 2025(4): 140-145,151.
YE N, ZHOU X L, ZHANG Y F, et al. Automotive brake assembly defect detection on improved YOLOv8 [J]. Modular Machine Tool & Automatic Manufacturing Technique, 2025(4): 140-145,151.

[80] 王洪金, 刘香怡, 何赟泽, 等. 基于改进 NCC 算法的大尺寸原位风机叶片可见光图像拼接[J]. 电子测量与仪器学报, 2024, 38(7): 1-12.
WANG H J, LIU X Y, HE Y Z, et al. Visible image stitching of large in-situ wind turbine blade based on improved NCC [J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(7): 1-12.

[81] 彭一誉, 何赟泽, 虞俊锋. 风机叶片内部缺陷日光激励动态热成像方法研究[J]. 电子测量与仪器学报, 2024, 38(1): 64-71.
PENG Y Y, HE Y Z, YU J F. Study on the method of daylight-excited thermal imaging of internal defects in wind turbine blades [J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(1): 64-71.

[82] 张德海, 祝志逢, 李艳芹, 等. 基于机器视觉的二维图像质量缺陷检测研究进展[J]. 包装工程, 2023, 44(23): 198-207.
ZHANG D H, ZHU ZH F, LI Y Q, et al. Research progress of two-dimensional image quality defect detection based on machine vision [J]. Packaging Engineering, 2023, 44(23): 198-207.

[83] 梁承权, 吕德深, 陆晓. 一种基于改进 Faster RCNN 的易拉罐印刷缺陷检测方法[J]. 印刷与数字媒体技术研究, 2023(6): 22-29.
LIANG CH Q, LYU D SH, LU X. A method for detecting the printing defects of easy open can based on improved Faster RCNN [J]. Printing and Digital Media Technology Study, 2023(6): 22-29.

[84] 刘海文, 郑元林, 钟崇军, 等. 基于改进 YOLOv5l 的印刷品缺陷检测[J]. 激光与光电子学进展, 2024, 61(10): 228-235.
LIU H W, ZHENG Y L, ZHONG CH J, et al. Defect detection of printed matter based on improved YOLOv5l [J]. Laser & Optoelectronics Progress, 2024, 61(10): 228-235.

[85] 张小雪, 孙帮勇, 谭家海, 等. 印刷品缺陷检测方法综述[J]. 印刷与数字媒体技术研究, 2025(2): 1-10.
ZHANG X X, SUN B Y, TAN J H, et al. A review of defect detection methods for printed products [J]. Printing and Digital Media Technology Study, 2025(2): 1-10.

[86] 孙小栋, 朱启兵, 徐华伟, 等. 用于超纤革表面瑕疵

识别的 MFL_YOLOv8 算法[J]. 光学 精密工程, 2025, 33(2): 311-323.
SUN X D, ZHU Q B, XU H W, et al. MFL_YOLOv8 algorithm for surface defect detection of microfiber leather[J]. Optics and Precision Engineering, 2025, 33(2): 311-323.

[87] 于光许, 张富宇. 基于改进 Res-UNet 网络的织物瑕疵图像识别方法[J]. 毛纺科技, 2024, 52(7): 100-106.
YU G X, ZHANG F Y. Image recognition method for fabrics defects based on improved Res-UNet network[J]. Wool Textile Journal, 2024, 52(7): 100-106.

[88] 陈利琼, 梅后金, 胡洪宣, 等. 基于改进 Faster R-CNN 的焊缝缺陷检测方法[J]. 科学技术与工程, 2025, 25(5): 2027-2033.
CHEN L Q, MEI H J, HU H X, et al. Weld defect detection based on improved Faster R-CNN method[J]. Science Technology and Engineering, 2025, 25(5): 2027-2033.

[89] 吴磊, 储钰昆, 杨洪刚, 等. 基于 YOLOv7TS 的铝合金焊缝 DR 图像缺陷检测技术[J]. 中国激光, 2024, 51(20): 20-29.
WU L, CHU Y K, YANG H G, et al. Aluminum alloy weld DR image defect detection technology based on YOLOv7TS[J]. Chinese Journal of Lasers, 2024, 51(20): 20-29.

[90] 周建民, 陈超, 涂文兵, 等. 红外热波技术、有限元与 SVM 相结合的复合材料分层缺陷检测方法[J]. 仪器仪表学报, 2020, 41(3): 29-38.
ZHOU J M, CHEN CH, TU W B, et al. Composite layer defect detection method based on infrared heat wave technology,finite element and SVM[J]. Chinese Journal of Scientific Instrument, 2020, 41(3): 29-38.

[91] 张海兵, 杜百强. 相控阵超声检测技术在碳纤维结构分层缺陷检测中的试验[J]. 无损检测, 2020, 42(4): 46-49, 55.
ZHANG H B, DU B Q. Experiment on phased array ultrasonic inspection technology of delamination damage ofcarbon fiber structure[J]. Nondestructive Testing, 2020, 42(4): 46-49, 55.

[92] GU ZH P, ZHU B K, ZHU G B, et al. AnomalyGPT: Detecting industrial anomalies using large vision-language models[J]. 38th AAAI Conference on Artificial Intelligence, 2024, 38(3): 1932-1940.

[93] 谭鲲鹏, 唐甲锋, 赵志斌, 等. 基于视觉大模型的激光粉末床熔融铺粉缺陷检测[J]. 中国激光, 2024, 51(10): 1002319.
TAN K P, TANG J F, ZHAO ZH B, et al. Powder spreading defect detection in laser powder bed fusion based on large vision model[J]. Chinese Journal of Lasers, 2024, 51(10): 1002319.

[94] 方爱国. 视觉大模型在轨交隧道缺陷检测中的应用研究[J]. 中国科技纵横, 2024(10): 27-30.
FANG AI G. Application of large vision model for defect detection in orbital tunnel[J]. China Science & Technology Overview, 2024(10): 27-30.

[95] 黄亚如, 钟剑斌, 成颖怡. 基于 SAM 视觉大模型的工业材料表面缺陷检测[J]. 智能制造, 2024(6): 129-134.
HUANG Y R, ZHONG J B, CHENG Y Y. Industrial material surface defect detection based on the SAM vision large model[J]. Intelligent Manufacturing, 2024(6): 129-134.

[96] LI X F, ZHANG ZH ZH, TAN X, et al. PromptAD: Learning prompts with only normal samples for few-shot anomaly detection[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 16848-16858.

[97] LI Y Z, WANG H L, YUAN SH H, et al. Myriad: Large multimodal model by applying vision experts for industrial anomaly detection[J]. ArXiv preprint arXiv: 2310.19070, 2023.

[98] QIAN K, SUN T Y, WANG W H. Exploring large vision-language models for robust and efficient industrial anomaly detection[J]. ArXiv preprint arXiv: 2412.00890, 2024.

[99] 彭家德. 大模型产品技术能力与高质量数据缺一不可[J]. 数字经济, 2024(7): 90-91.
PENG J D. The technical capabilities of large-scale model products and high-quality data are both indispensable[J]. Digital Economy, 2024(7): 90-91.

[100] 王林. 生成式人工智能大模型观察[J]. 上海信息化, 2023(8): 42-45.
WANG L. Observation on generative artificial intelligence large models[J]. Shanghai Informatization, 2023(8): 42-45.

[101] STANFORD HAI. Artificial intelligence index report 2024[R]. Stanford HAI, 2024.

[102] 赵宣翔, 杨昊, 曹雯, 等. 基于超声检测的室温硫化硅橡胶涂层内部微小缺陷识别方法研究[J]. 电工电能新技术, 2022, 41(6): 55-63.
ZHAO X X, YANG H, CAO W, et al. Research on

identification method of internal micro-defects in room temperature vulcanized silicone rubber coating based on ultrasonic testing[J]. Advanced Technology of Electrical Engineering and Energy, 2022, 41(6): 55-63.

[103] 罗朝莉, 朱冰, 王波, 等. 铝板表面裂纹的激光超声检测与信号处理研究[J]. 电子测量与仪器学报, 2023, 37(10): 41-52.
LUO ZH L, ZHU B, WANG B, et al. Research on laser ultrasonic testing and signal processing of surface cracks in aluminum plate[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(10): 41-52.

[104] 何赟泽, 陈琦, 王洪金, 等. 激光热成像无损检测研究进展(特邀)[J]. 红外与激光工程, 2024, 53(7): 51-66.
HE Y Z, CHEN Q, WANG H J, et al. Research progress of laser thermography non-destructive testing (invited)[J]. Infrared and Laser Engineering, 2024, 53(7): 51-66.

[105] 邓海明, 邓堡元, 王洪金, 等. 卤素灯阵列激励的红外热成像涂层内部缺陷检测[J]. 红外与激光工程, 2025, 54(8): 84-95.
DENG H M, DENG B Y, WANG H J, et al. Infrared thermal imaging coating debonding defect detection excited by halogen lamp array[J]. Infrared and Laser Engineering, 2025, 54(8): 84-95.

[106] 吴昆鹏, 王少聪, 苏成. 基于 3D 点云的钢管表面缺陷检测系统[J]. 轧钢, 2024, 41(3):113-118,126.
WU K P, WANG SH C, SU CH, et al. Surface defect detection system based on 3D point cloud for steel pipe[J]. Steel Rolling, 2024, 41(3):113-118,126.

[107] 肖苏华, 乔明娟, 赖南英, 等. 基于 3D 视觉的风电塔筒焊缝检测系统设计[J]. 电子测量与仪器学报, 2022, 36(2): 122-130.
XIAO S H, QIAO M J, LAI N Y, et al. Design of wind turbine tower weld detection system based on 3D vision[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(2): 122-130.

[108] 魏永超, 蔡双, 岳雨琛, 等. 基于高光谱成像的航空叶片缺陷检测[J/OL]. 激光杂志, 1-7[2025-08-25].
WEI Y CH, CAI SH, YUE Y CH, et al. Aircraft blade defect detection based on hyperspectral imaging[J/OL]. Laser Journal, 1-7[2025-08-25].

[109] 田勇, 周曾鹏, 田劲东, 等. 结合高光谱和机器学习的无线充电金属异物检测[J]. 电子测量与仪器学报, 2022, 36(8): 238-247.
TIAN Y, ZHOU Z P, TIAN J D, et al. Metal object detection in wireless charging systems combining hyperspectral imaging and machine learning[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(8): 238-247.

[110] 腾讯研究院. 工业大模型应用报告[R]. 北京: 腾讯研究院, 2024.
Tencent Research Institute. Industrial large model application report [R]. Beijing: Tencent Research Institute, 2024.

作者简介

何毓芬,2024 年于武汉大学获得学士学位,现为湖南大学博士研究生,主要研究方向为红外无损检测和智能检测算法。
E-mail:yfhe@ hnu. edu. cn

He Yufen received her B. Sc. degree from Wuhan University in 2024. She is currently a Ph. D. candidate at Hunan University. Her main research interests include infrared non-destructive testing and intelligent detection algorithms.

何赟泽(通信作者),2006 年于西安交通大学获得学士学位,2008 年于国防科技大学获得硕士学位,2012 年于国防科学技术大学获得博士学位,现任湖南大学教授,主要研究方向为嵌入式人工智能与边缘计算、红外热成像与机器视觉。
E-mail:yhe@ vip. 163. com

He Yunze (Corresponding author) received his B. Sc. degree from Xi′an Jiaotong University in 2006, his M. Sc. and Ph. D. degrees both from University of Defense Science and Technology in 2008 and 2012, respectively. He is currently a professor at Hunan University. His main research interests include embedded artificial intelligence and edge computing, infrared thermal imaging and machine vision.