DOI: 10.19650/j.cnki.cjsi.J2514531

视觉 Transformer 在表面缺陷检测中的应用：研究进展与挑战

杨 洋,吴一全
(南京航空航天大学电子信息工程学院 南京 211106)

摘 要:传统卷积神经网络(CNN)受限于局部卷积操作,难以有效建模长程依赖关系;相比之下,视觉 Transformer 通过自注意力等机制实现了对全局依赖关系的显式建模。在表面缺陷检测任务中,尤其是在背景纹理复杂、缺陷形态多变等检测场景中,展现出优于 CNN 的检测性能。围绕视觉 Transformer 在表面缺陷检测中的技术优势与应用方法、面临的关键挑战及应对策略两大维度,综述了近年来国内外基于视觉 Transformer 的表面缺陷检测研究进展与挑战,为视觉 Transformer 在表面缺陷检测中的应用提供了理论依据与方法支撑。首先,阐释了表面缺陷检测的基本定义,归纳了该领域的技术特征与主要瓶颈。其次,深入剖析了视觉 Transformer 在缺陷检测任务中所具备的技术优势及其在实际应用中存在的关键挑战。然后,结合视觉 Transformer 的技术优势,重点分析了视觉 Transformer 在表面缺陷检测任务中的典型应用方向,包括应对复杂纹理背景干扰、实现多模态信息融合、基于模块化思想的局部-全局特征信息融合等应用场景。随后,探讨了视觉 Transformer 在面对表面缺陷检测任务中存在的样本量稀缺、模型计算复杂度高与实时性不足、训练效率低下以及小目标缺陷检测性能差等关键挑战时,所采用的主要优化策略与应对方法。最后,围绕迁移学习驱动的预训练视觉大模型构建、视觉 Transformer 与多模态的深度融合等方向,对视觉 Transformer 在表面缺陷检测领域的发展趋势进行了展望。

关键词: 视觉 Transformer;缺陷检测;技术优势与典型应用;关键挑战与应对策略
中图分类号: TP391.41 TH89 TP389.1 **文献标识码**: A **国家标准学科分类代码**: 510.4050 520.2060

Applications of vision Transformer in surface defect detection: Research progress and challenges

Yang Yang, Wu Yiquan
(*College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China*)

Abstract: Convolutional neural network (CNN) have been limited in their ability to effectively model long-range dependencies due to their localized convolution operations. In contrast, vision Transformer achieves explicit modeling of global dependencies through mechanisms such as self-attention. In surface defect detection tasks, especially in scenarios with complex background textures or diverse defect morphologies, vision Transformer shows superior performance compared with CNN. This article provides a comprehensive review of recent domestic and international research progress and challenges in surface defect detection based on vision Transformer, focusing on two dimensions: The technical advantages and application methodologies, as well as key challenges and corresponding strategies. Firstly, the fundamental definition of surface defect detection is elucidated, and the technical characteristics and main challenges in this field are summarized. Secondly, the technical advantages and key challenges of the vision Transformer in the context of defect detection are analyzed. Subsequently, leveraging the technical strengths of vision Transformer, typical applications in surface defect detection tasks are examined in detail, including handling complex texture background interference, achieving multimodal information fusion, and integrating local-global feature information based on a modular design approach. Subsequently, the article discusses the main optimization strategies and solutions adopted by vision Transformer to address key challenges in surface defect detection, such as scarce sample data, high computational complexity, insufficient real-time performance, low training efficiency, and poor performance in detecting small targets. Finally, future research directions and development trends of vision Transformer in the field of surface defect detection are prospected, such as the development of transfer learning-based pre-trained models and their advanced fusion with

收稿日期:2025-09-24 Received Date: 2025-09-24

multimodal methodologies, among others.

Keywords: vision Transformer; defect detection; technical advantages and typical applications; key challenges and response strategies

0 引 言

表面缺陷检测是工业质量控制的核心环节,其对检测算法的精度、鲁棒性及效率要求极高[1]。然而,表面缺陷检测任务面临着缺陷形态多变、复杂背景干扰、小目标缺陷难识别以及样本量稀缺等独特挑战,这些挑战使得传统图像处理方法在特征表达能力与泛化性能上存在局限,而基于卷积神经网络(convolutional neural network, CNN)的深度学习模型也因局部感受野和归纳偏差等固有约束,在长距离依赖建模与多尺度缺陷适应方面表现不足。近年来,视觉 Transformer 技术[2-3]凭借其强大的全局特征提取能力与长距离依赖建模优势,被广泛应用于表面缺陷检测任务中,展现出显著潜力。为此,系统梳理基于视觉 Transformer 的缺陷检测方法的应用现状及其面临的关键挑战,对该领域的进一步发展具有重要意义,具体而言,体现在两个方面:

1) 学术研究层面

视觉 Transformer 在计算机视觉领域中,尤其是缺陷检测任务中的应用,近年来呈现出爆发式增长,涌现出大量围绕全局-局部特征融合机制、注意力机制优化、层级网络结构设计、跨尺度特征融合、端到端的目标检测网络以及基于视觉 Transformer 多模态融合等应用场景的改进架构与方法。系统归纳并总结相关方法,不仅有助于厘清视觉 Transformer 在缺陷检测领域中的技术发展脉络与核心创新点,还能揭示现有方法在面对小目标检测精度低、训练效率不高及样本量稀缺等挑战时所表现出的特征提取能力不足与易过拟合等局限性,从而为缺陷检测领域的后续研究与应用指明发展方向,并进一步推动其基础理论模型的创新与发展。

2) 工程实践方面

一方面,缺陷检测场景对算法的准确性、鲁棒性和实时性均有极高要求;另一方面,复杂背景干扰、样本量稀缺及缺陷目标的跨尺度差异,又对检测任务构成显著挑战。视觉 Transformer 凭借其长距离依赖建模能力,在全局特征提取方面表现出色,有助于提升检测性能,但其实际应用仍面临诸多限制。因此,系统梳理并深入分析现有视觉 Transformer 的技术应用场景与挑战应对策略,可为工程实践在模型选型、部署与优化环节提供扎实的工程依据与科学的决策支持。例如,针对如细微裂纹、易与背景混淆的低对比度瑕疵等特定缺陷类型,基于视觉 Transformer 及其相关方法来选择或开发最合适的模型架构,从而平衡检测精度与推理速度,并推动相关技术在实际生产线中的落地应用,从而最终提升相关行业质量控制的自动化与智能化水平。

国内外学者基于视觉 Transformer 思想围绕表面缺陷检测任务进行了大量研究,然而专注于该领域的相关综述研究却较为贫乏。现有相关综述主要从通用计算机视觉任务或跨领域应用的角度展开探讨:例如,一方面,相关学者[4-9]聚焦于视觉 Transformer 在人员识别、自动驾驶及音视频等通用任务上的应用,主要研究了围绕分类、检测、分割等基础模型所构建的通用框架及相关改进策略;另一方面,也有工作探讨了视觉 Transformer 在跨领域场景——如医学图像分析[10]及广义异常检测[11]——中的潜力与挑战。区别于上述相关综述文献,本研究聚焦于视觉 Transformer 在表面缺陷检测中的应用,重点围绕其技术优势与方法应用、关键挑战及解决方案两个核心维度进行系统梳理,以推动深度学习算法在缺陷检测领域不断向前发展。文章区别于其他综述主要体现在两个方面:

1) 文章的研究对象聚焦于计算机视觉细分领域中的关键方向——表面缺陷检测。相关研究与总结工作,将为视觉 Transformer 及其各类衍生方法在表面缺陷检测领域的应用实践与理论研究提供更有效的优化思路,进而推动表面缺陷检测技术向更稳定、更成熟的方向发展。

2) 文章从表面缺陷检测的实际需求与挑战出发,首先,辨析了其与通用目标检测任务的区别,以凸显表面缺陷检测特有的难点与复杂性。其次,在分析视觉 Transformer 的核心原理基础上,总结了视觉 Transformer 在缺陷检测应用领域中的相关技术优势与关键挑战。然后,一方面,基于其技术优势,分别对复杂纹理干扰、多模态融合以及模块化思想应用等场景中相关视觉 Transformer 的应用方法进行了总结;另一方面,针对视觉 Transformer 在缺陷检测任务中面临的样本量稀缺、模型实时性不足与复杂度高、模型训练效率偏低及小目标检测精度不足等核心技术瓶颈,深入剖析了对应的解决策略。通过上述研究,系统总结了视觉 Transformer 在缺陷检测任务中的应用,为该领域的发展提供了关键理论支撑与实践参考。

1 表面缺陷检测任务分析

表面缺陷检测主要指的是针对产品表面可能出现的瑕疵进行识别与检测等操作。区别于通用目标检测,表面缺陷检测在应对复杂工况干扰、工艺性特征影响、小目标检测、瑕疵间尺度差异、小样本学习、复杂背景干扰以

及类别不平衡等问题上面临更大挑战。其中,工艺性特征所带来的问题尤为特殊:由于产品制造工艺环节的某些系统性偏差,可能导致某类缺陷呈现批次性爆发,甚至在同一位置反复出现,进一步增加了检测的难度与不确定性。冯夫健等[12]分析认为铆钉尺寸小,且存在金属反光现象,复杂背景下铆钉及镦头形态特征难以识别,而不同铆接缺陷类型之间差异小,缺陷特征不明显,相同铆接缺陷类型之间差异大,从而对缺陷分类造成困难。李刚等[13]指出在输电线路相关设备缺陷检测场景中,一方面各类螺栓缺陷样本数量存在不平衡,进而无法提供充足的样本保证模型训练,严重限制了深度学习模型的学习能力;另一方面输电线路螺栓缺陷检测具有特殊性,螺栓所处的自然背景复杂多变,不同金具上的螺栓所需的紧固元件构成也各不相同。张乃雪等[14]总结出在工业产品表面缺陷检测中,获取大量经过专家标注且涵盖多样缺陷类型的样本通常十分困难,导致训练集规模极为有限。此外,工业产品表面缺陷往往尺度较小、对比度低、表面特征不明显,这些因素共同构成了缺陷检测任务中的重要挑战。又如印刷电路板(printed circuit board, PCB)、计算机主板等电子元器件表面缺陷检测,陈俊英等[15]分析认为此类场景的元器件种类多样且空间分布复杂,致使缺陷检测面临诸多挑战:缺陷目标特征微弱、不同缺陷类型间相似性高、以及缺陷尺度跨度大等,同时,实际工业生产环境下的缺陷样本相对稀缺,这些都使得诸如此类电子元器件的缺陷检测成为工业缺陷检测领域中的一项复杂难题。综上所述,产品表面缺陷检测存在着与通用目标检测明显的区别点与挑战具体为:

1) 核心任务差异

通常而言,通用目标检测主要针对人员、车辆、障碍物、信号灯等常规对象的检测、识别与计数。与之不同,表面缺陷检测则专注于物体表面的瑕疵识别,其重点在于检测出与正常模式不一致的异常特征。因此,通用目标检测侧重于对各种预定义目标进行判别;而缺陷检测更侧重于“辨识异常”,即在整体中找出细微的异常区域。

2) 关注的语义层次不同

缺陷检测关注的焦点在于识别产品表面的局部特征异常,即在整体结构中定位并辨识出局部异常区域;而通用目标检测则更侧重于对目标的语义类别与状态进行理解与判断,其关注点在于整体层面的高级语义信息。

3) 从数据集角度

常规的目标识别属于大数据集,如 ImageNet 数据集等。缺陷检测属于小数据集。且缺陷样本极难提取与收集,其样本量远远小于正常样本。

4) 从场景应用方面

缺陷检测一般用于工业检测、农业检测等,通用目标检测主要用于自动驾驶、安防监控、零售分析等。

5) 从技术要求层面

缺陷检测因为瑕疵的小目标与微弱特征需求,往往需要更高分辨率的工业相机,甚至是多传感器协同的多模态技术。

6) 从技术挑战角度

表面缺陷检测任务除了兼具通用目标检测场景下面临的目标遮挡、尺度变化以及目标多样性等挑战外,表面缺陷检测还存在着光照不均、复杂背景纹理干扰、单模态特征目标难拍难检、实时性检测要求高、样本量稀缺等关键挑战。

正是因为产品表面缺陷检测的各种检测挑战性,决定了不论是传统图像处理还是基于 CNN 的深度学习方法都无法实现高效检测,尤其是针对复杂背景干扰或缺陷特征微弱等检测场景。而视觉 Transformer 作为深度学习算法当前较为先进的技术手段之一,借助其优秀的全局特征提取能力,将会推动基于机器视觉技术的表面缺陷检测领域不断向前发展。

2　视觉 Transformer 原理及其在表面缺陷检测中的优势与挑战分析

在大规模样本数据的支撑下,视觉 Transformer 模型(vision Transformer, ViT)和检测 Transformer 模型(detection Transformer, DETR)等基于 Transformer 的深度学习网络凭借输入数据序列化、位置编码以及自注意力机制等 3 大核心机制,突破了传统 CNN 的局部建模局限,成为了计算机视觉领域的新范式。这些核心机制使其在表面缺陷检测任务中展现出独特技术优势,同时也面临诸多关键挑战。

2.1　视觉 Transformer 原理分析

视觉 Transformer 核心思想在于:以序列化输入为前提、以位置编码为空间信息补充、以自注意力机制为核心,具体为:

1) 输入数据序列化

输入数据序列化是二维图像数据适配 Transformer 架构的关键前提。由于 Transformer 思想最初设计时以一维序列为核心处理对象,而图像数据具备二维空间结构,输入数据序列化通过将图像划分为规则图像块,并将其展平为向量序列,从而将二维图像数据转换为模型可处理的序列格式。这一处理在保留局部结构信息的同时,为后续的位置编码与自注意力机制提供了必要的输入形式。正是这一序列化步骤,使 Transformer 架构得以在长距离依赖建模、多模态融合支撑、多尺度层级特征表示以及高度的可扩展性等方面发挥着重要的支撑作用。

2) 位置编码

位置编码的作用在于弥补 Transformer 自身缺乏空间

感知的局限性。由于图像语义高度依赖图像块的二维空间拓扑与相对关系,若在缺乏位置信息时,自注意力仅能通过内容相似性计算权重,无法区分内容相似但空间位置不同的图像块。通过为序列中的每个元素注入可学习或固定的位置信号,模型能够感知其在原始图像中的二维空间位置信息,从而让模型不仅可以维持空间拓扑结构,还可以学习图像块间的相对距离与邻接关系,从而理解缺陷目标之间的相对布局关系。

3) 视觉 Transformer 编码器

以自注意力机制为核心的 Transformer 编码器则承担特征抽象与融合的关键角色。其通过全局交互为每个单元生成全新的、富含依赖关系的特征表示。其核心目标超越了传统注意力机制的"聚焦局部"思想,实现了对全局依赖关系的显式建模。该机制弱化了局部性、平移不变性等强归纳偏置,转而强调对全局关系与动态权重分配的学习能力。这一特性使其能够有效理解图像中复杂的结构性布局,在缺陷检测等需要高度细化全局上下文理解的任务中表现出显著优势。

2.2 视觉 Transformer 在表面缺陷检测中的技术优势

基于上述对视觉 Transformer 原理与核心架构的分析,视觉 Transformer 技术展现出了超越 CNN 的显著优势,其优势可主要归因于自注意力机制、多层编码器结构、序列化及位置编码等方面的协同作用。具体而言,这些优势体现在 3 个方面,即:

1) 长距离依赖建模驱动下的全局特征提取能力

视觉 Transformer 通过自注意力机制可建立全局区域的关联关系,能够有效捕捉非连续、分布零散的缺陷特征。例如,在检测瓷砖表面断续分布的浅裂纹时,传统 CNN 受限于局部感受野,难以整合跨区域的缺陷信息,而 Transformer 能够建立全局上下文联系,完整勾勒缺陷形态。此外,在面对缺陷与背景纹理高度相似的场景时,该机制也可通过全局上下文捕捉细微差异,提升小目标缺陷检出率。

2) 为多模态融合提供关键支撑

视觉 Transformer 架构天然支持多模态数据处理,因其以序列作为统一输入形式,自注意力机制本身对输入模态类型不做特定假设,只需将图像、文本等不同模态信息嵌入为向量序列并添加相应位置编码,即可实现跨模态交互与融合,为视觉-语言联合建模等任务奠定基础。

3) 视觉 Transformer 模块化思想

一方面,视觉 Transformer 的相关核心基础组件均被设计为高度标准化且接口统一的构建单元,这种标准化的设计使其具备优异的可堆叠性与可替换性,为后续模型的构建与演化提供了技术基础;另一方面,视觉 Transformer 为适配视觉任务需求,演化出以层次化划分与功能解耦为核心的模块化架构范式[16]。进一步地,移位窗口 Transformer(Swin Transformer)[17]、金字塔视觉 Transformer(pyramid vision Transformer, PVT)[18]及数据高效的图像 Transformer(data-efficient image Transformers, DeiT)[19]等代表性视觉 Transformer 变体模型,通过引入窗口划分、移位操作、金字塔下采样等方法,持续强化架构对视觉信息的归纳偏置能力,使整个模型或其子模块可作为即插即用的骨干网络,无缝嵌入检测、分割等复杂视觉模型,实现与下游任务的高效协同。

2.3 视觉 Transformer 在表面缺陷检测中的关键挑战

视觉 Transformer 虽然在各类视觉任务中表现出卓越性能,但是在实际工程落地中仍面临显著瓶颈:数据驱动型的本质导致了其在小样本场景下泛化能力骤降;高计算复杂度的约束决定了其难以适配边缘计算设备;缺乏空间归纳偏置使其对低分辨率图像的细节理解精度不足;而局部特征提取能力的欠缺又制约了其在细粒度识别任务中的表现,这些局限性共同限制了视觉 Transformer 在缺陷检测领域的广泛应用。基于此,国内外学者结合视觉 Transformer 思想的优势与不足,并在借鉴了部分优秀的 CNN 算法模型思想的基础上,研究设计了多种基于视觉 Transformer 思想的改进方法,如下一代视觉 Transformer(next-generation vision Transformer, Next-Vit)[20]、Swin Transformer、可变形 DETR(deformable DETR)[21]、瓶颈结构 Transformer 网络(bottleneck Transformer network, BoTNet)[22]、实时检测 Transformer(real-time detection Transformer, RT-DETR)[23-24]等,这些方法有效缓解了原始 ViT 和 DETR 的诸多局限,从而使得 Transformer 思想在目标检测、图像分割[25-26]、实时视觉感知乃至移动端推理等特定场景中实现了更为广泛的应用。然而,这些改进方法在表面缺陷检测任务中仍面临部分关键因素的制约,即:1)在表面缺陷检测中,正样本极少且形态多变,尽管 BoTNet 等尝试引入卷积先验,但模型仍难以从稀缺且不平衡的样本中充分学习泛化特征,导致其在实际应用中稳定性不足;2)模型轻量化的困境源于编码器堆叠与自注意力的大量参数,即使如 Swin Transformer 等通过局部窗口划分降低了计算量,但其复杂的窗口移位机制与相对位置编码仍增加了部署的复杂性,难以满足工业场景对实时性的严苛要求;3)训练效率低下与过拟合风险源于 Transformer 架构缺乏归纳偏置,严重依赖大规模数据;4)小目标检测精度不足的根源在于自注意力机制固有的计算复杂度与序列长度的平方次关系,迫使模型必须对高分辨率输入进行下采样或使用较大图像块尺寸,致使微小缺陷的细节信息在序列化过程中丢失。因此,尽管 ViT 与 DETR 及其改进模型在通用基准上不断提升性能,但其在资源受限、高精度要求的缺陷检测领域中的应用,仍受限于上述由基本架构特性所带来的根本性挑战。

2.4　小结

本章重点分析了视觉 Transformer 的在缺陷检测中的技术优势及其面临的关键挑战。从而为后续基于视觉 Transformer 技术驱动下的缺陷检测应用方法与关键挑战的应对措施分析奠定基础。

3　视觉 Transformer 技术优势驱动下的缺陷检测应用方法

基于前述分析,视觉 Transformer 凭借其特有的自注意力机制与架构设计,在多项核心技术能力上展现出显著优势,这些优势为其在表面缺陷检测任务中的成功应用奠定了坚实基础。尤其在应对背景纹理噪声敏感、结构复杂且缺陷形态多变等场景中,视觉 Transformer 能够有效捕捉全局上下文信息,抑制复杂背景纹理干扰,显著提升对复杂形态缺陷和低对比度缺陷的识别能力。此外,其在多模态融合方面的优势,支持可见光、红外、深度信息等多种传感数据的统一建模与互补增强,为复杂工况下的缺陷检测提供了新的解决路径,进一步推动了高精度、高鲁棒性视觉检测系统的发展。

3.1　基于视觉 Transformer 的复杂纹理表面缺陷检测方法

复杂纹理背景干扰作为缺陷检测领域核心技术挑战之一,一直是国内外基于视觉算法进行缺陷检测研究的重点,如瓷砖表面缺陷检测、印制电路板(printed circuit board,PCB)表面缺陷检测、织物表面缺陷检测以及木材表面缺陷检测等领域。在复杂纹理背景的干扰下,基于手工经验特征的传统图像处理算法往往表现受限,检测性能不佳。而基于数据驱动的深度学习方法能够自适应地学习鲁棒的判别性特征,为复杂纹理背景干扰场景下的缺陷检测提供了更为有效的解决方案。虽然基于 CNN 的算法模型在复杂纹理背景下的缺陷检测中已经展现出高效的局部特征提取能力,但同时由于其局部感受野限制而难以区分纹理背景与缺陷的细微差异,从而导致高相似度干扰下的误检率升高,相比之下,Transformer 基于多头自注意力机制通过全局信息提取来构建模型长程依赖关系,从而既可以实现抑制纹理干扰,也可以显著提升复杂背景下的缺陷识别鲁棒性。

1) 基于 CNN-Transformer 协同模块的特征增强

基于混合单元的特征提取增强方法,其思想在于将 CNN 的局部特征提取能力与 Transformer 的全局建模优势在基础网络单元层面进行深度融合,从而构建新型特征提取模块,以提升对复杂纹理背景下缺陷的捕获能力。该维度是解决复杂纹理背景干扰下的“单个特征提取单元能力不足”的问题,本质是单元组件级的能力重构,以“提升特征提取单元的基础性能”为目标。

如陈俊英等[15]提出了一种并行残差特征提取网络,将部分卷积的局部特征提取能力与视觉 Transformer 的全局建模优势相融合,显著提升了模型对长距离依赖关系和细节特征的捕获效率,为后续特征融合奠定了高质量的特征基础。Yu 等[27]针对传统 CNN 局部感受野限制导致的误检漏检问题,通过引入 MobileViTv3 模块,将视觉 Transformer 的全局上下文理解能力与 CNN 的局部特征提取能力相结合,在实现模型轻量化的同时,有效增强了复杂纹理背景下细微缺陷的检出能力。这些方法在单元层面的创新与协同有效解决了传统 CNN 感受野有限和 Transformer 局部细节不足的问题,在 PCB 主板、瓷砖、地板等检测领域中取得了较好的指标性能。

2) 基于视觉 Transformer 为主干的网络优化

该方法以 Swin Transformer 等基于视觉 Transformer 的主干网络为基础,充分利用其全局注意力机制带来的长距离依赖建模能力,同时针对缺陷检测中特有的不规则形态、边缘模糊、多尺度等难点,通过引入注意力机制、可变形卷积等模块进行针对性增强,以实现检测性能的精准提升。

如图 1 所示,Liu 等[28]聚焦非周期复杂纹理背景下的非规则目标检测,以 Swin Transformer 网络为核心构建了一种 U 型网络,其分层移位窗口自注意力机制有效区分了背景纹理与真实缺陷。此外,为强化不规则缺陷感知与泛化能力,引入了集成键值丢弃(DropKey)策略的 Res-Drop Swin Transformer 块来平滑注意力权重分布,并设计注意力引导的可变形卷积模块(attention-based deformable convolution, ABDC),通过自适应感受野特性来精准捕捉缺陷边缘与形状特征,形成全局与局部的功能互补,在织物缺陷数据集上表现优异。

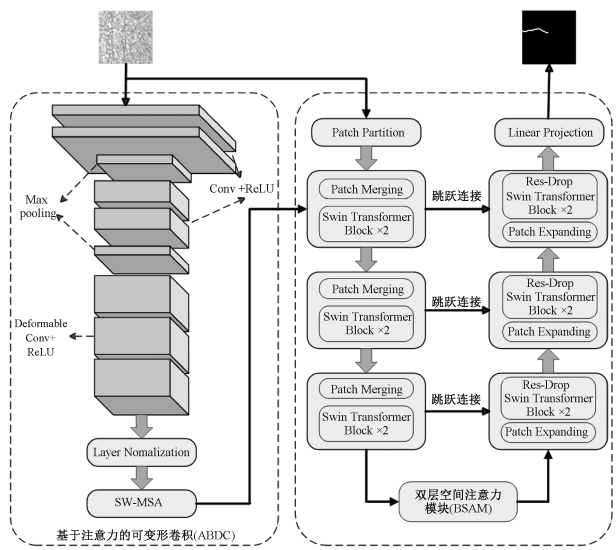

图 1　ARDSwin-Unet 网络总体架构

Fig. 1　The overall architecture of ARDSwin-Unet

Yao 等[29]采用 DETR 的改进版 DINO 作为基础框架,利用其对比去噪训练机制增强鲁棒性,同时以 Swin Transformer 替换 ResNet-50 主干,并引入可变形注意力与多尺度训练,提升了复杂背景下的检测适配性。这些增强策略使 Transformer 主干网络在保持全局上下文理解的同时,强化了对局部特征的适应性,在织物、色织物等检测任务中表现出色。

3) CNN 与 Transformer 双分支并行架构与特征融合机制

该维度的核心是解决“架构级的效率优化”的问题,通过构建 CNN 与 Transformer 双分支并行架构,一方面使得 CNN 分支专注于局部细节、边缘等精细特征的高效提取;另一方面,Transformer 分支则聚焦于全局上下文、长距离依赖关系的建模,再通过设计精细化的特征融合机制来实现两类特征的协同优化,从而充分发挥各自优势。如 Yang 等[30]提出了一种双分支检测架构:编码器提取多尺度特征后,高层特征输入 Transformer 模块建立全局依赖,再通过特征融合机制将全局语义与 CNN 提取的局部细节协同优化,在织物数据集上实现了增强对齐度量值(enhanced-alignment measure, E-measure)为 97.19%以及检测速度每秒帧数(frame per second, FPS)达到 27。Qu 等[31]结合 ResNet-50 与 Swin Transformer 来构建双主干特征提取分支,提出了一种融合全局上下文与局部细节特征的混合网络(U-Swin Transformer & multi-residual network, U-SMR),其中,ResNet-50 发挥 CNN 局部特征提取优势,Swin Transformer 承担全局长距离依赖建模任务,后续通过一种轻量级的双分支金字塔模块(dual-branch pyramid module, DBPM)实现了多尺度特征均衡提取,形成“局部-全局”的协同感知体系,有效提升了复杂纹理背景下的缺陷检测性能。双分支并行架构方法旨在解决单一模型在复杂纹理背景下全局信息分析不足或局部特征提取弱的问题,通过双分支的并行处理实现特征互补。

4) 基于 Transformer 自编码器的残差缺陷检测方法

基于重构与残差分析的缺陷定位方法[32],并不直接进行缺陷识别或分类,而是利用 Transformer 的自编码器结构学习无缺陷样本的模板,通过重构图像与输入图像的残差分析来定位缺陷。这种方法避免了复杂背景对直接检测的干扰,利用 Transformer 的全局特征提取能力实现更准确的图像重构,从而突出缺陷区域。

上述 4 个维度主要基于视觉 Transformer 解决复杂纹理背景干扰下缺陷检测任务的技术演进与研究方法,分别从“组件创新-主干强化-架构设计-范式突破”等方面组成的技术链进行了分析。

3.2 多模态与视觉 Transformer 在表面缺陷缺陷检测中的协同应用

一方面,传统的单模态数据一直存在着特征维度单一、任务适应性不足等局限性问题,如最常见的可见光单模态成像场景极为依赖光学影响,易受光照不均、表面反光干扰等因素影响。再比如 X 射线、红外线等不可见光数据虽然可实现穿透检测,但分辨率受限且无法获取表面纹理特征,所以单模态场景下,单一数据源难以覆盖复杂缺陷类型的多尺度特征表达。

另一方面,在缺陷检测领域,由于待检测产品的材质、环境以及检测需求的不同,导致需要检测融合的数据模态特征多种多样,除了常见的可见光数据模态,还包括文本特征数据[33-34]、红外热成像[35-36]、三维点云、超声波及声信号等多模态数据,各模态通过物理关联性或交互式语义补充形成互补表征[37]。多模态技术通过异构数据的时空对齐与特征互补,可构建多层次缺陷表征体系[38],例如热成像揭示材料应力集中区域,声学特征可捕捉内部应力特征变化,多源信息融合可显著提升检测系统的容错性与泛化能力。得益于 Transformer 技术在自注意力、并行计算以及长距离依赖建模能力上的独特优势,使得 Transformer 技术在处理多模态融合[39]、高效率计算以及处理模态间跨模态长距离依赖问题上表现得更加高效。

Feng 等[34]提出了一种基于图像-文本双模态融合的深度学习检测方法,其核心创新在于引入了视觉 Transformer 架构来实现对图像区域特征与文本语义特征的交互对齐与深度融合。Luo 等[40]则聚焦可见光图像和近红外图像模态融合,基于当前柑橘表面缺陷检测中单模态视觉方法存在信息局限,以及传统多模态融合方法多为简单通道拼接、难以有效挖掘跨模态互补特征的技术瓶颈,通过将视觉 Transformer 思想应用于多模态融合,构建了全新的实时多模态复合域注意力融合模型检测方法(real-time detection Transformer multimodal compound domain attention fusion, RT-DETR-MCDAF):一方面在时域分支中采用差异化特征提取策略来分别处理可见光与近红外图像,并利用空间注意力块(spatial attention block, SAB)来增强跨模态语义特征的空间交互;另一方面在频域分支中引入傅里叶变换与自驱动通道注意力(self-attention-driven channel attention, SCA)机制,在频率空间中强化缺陷相关的关键频谱成分。该双路径融合机制充分发挥了 Transformer 全局上下文建模能力与注意力机制的特征筛选优势,实现了多模态信息在深层语义层面的自适应融合。实验表明,该模型在精确率、召回率及平均精度均值(mean accuracy precision, mAP)等关键指标上显著优于传统融合模型与单模态方法,为农业产品表面缺陷检测中的多模态感知问题提供了有效的解决方案。

3.3 模块化思想驱动下的检测方法

视觉 Transformer 的模块化技术优势,为构建高性能

的缺陷检测模型提供了强大而灵活的架构基础。其核心价值在于能够作为一种通用的“即插即用”式构建单元,系统性地嵌入各类检测、分类及分割等检测框架的不同层级中,显著增强模型的特征表示与跨域迁移能力。具体而言,其既可以采用如 Swin Transformer 等性能更强的主干网络替代 VGG、ResNet 等传统卷积主干网络,并借助其分层设计与自注意力机制,实现能够同时捕获长程依赖与多尺度局部特征的分层表示;也可将 Swin 块或编码器组件等独立的视觉 Transformer 模块嵌入 CNN 特征提取路径中,在卷积生成的细粒度局部特征基础上,自适应地建模与增强全局上下文信息,从而在深层语义层面实现局部细节与全局信息的有效互补与协同。此外,视觉 Transformer 的模块化设计还为其在不同任务与领域间的结构复用提供了可能:同一预训练模块可经微调适配于多种缺陷类型与成像环境,显著提升了模型泛化能力与部署效率[41-42],即:

1) 局部嵌入或融合

视觉 Transformer 以模块化的方式局部嵌入或融合到分类、检测及分割等 CNN 网络中,能够有效实现局部特征与全局表示的互补与协同,增强模型表征能力。同时,这种即插即用式的模块化集成策略,可有效适配缺陷检测领域普遍存在的小样本数据场景,有助于缓解过拟合问题,并促进模型的轻量化,具有较高的实用性与可扩展性。查健等[43]将视觉 Transformer 模块融入主干网络,替换网络末端的原有跨阶段局部三卷积模块(cross stage partial network with 3 convolutions, C3)为融合 Transformer 的 C3 模块(C3 with Transformer, C3TR)组件,该模块充分结合了 CNN 局部细节特征提取优势与 Transformer 多头自注意力机制,强化了刨花板小目标缺陷的特征提取能力。

2) 主干网络替换融合与并行协同检测

视觉 Transformer 作为强大的特征提取引擎,其价值不仅限于局部增强,更体现在对 CNN 主干网络的全局性替代或并行协同检测上。这类方法从架构层面彻底革新了特征学习范式,一方面主干网络替换通过 Transformer 的全局注意力机制彻底克服了 CNN 在长程依赖建模上的固有局限;另一方面协同并行检测机制则构建了“全局-局部”算法框架,实现了两种范式的深层互补与融合,从而强化了多尺度缺陷捕捉与复杂背景适应力,是推动检测性能实现质变的关键路径。Li 等[44]针对 YOLOX 原有主干网络(CSPDarknet)应对复杂探地雷达(ground penetrating radar, GPR)图像背景存在全局上下文建模能力不足的问题,通过引入 Swin Transformer 网络并替换 YOLOX 模型原有主干特征网络,利用其全局注意力机制与多尺度特征融合能力,来增强对 GPR 图像中缺陷特征的全局感知与判别能力,提升了对复杂缺陷特征的提取性能。如图 2 所示,李季桐等[45]提出了一种基于层次化多尺度特征融合的金属缺陷分类模型(hierarchical multi-scale feature fusion, HMFF),该模型融合了 Swin Transformer 与 ConvNeXt 的特征提取互补的模块化优势:基于 Swin Transformer 支线网络实现全局特征的高效提取,而通过 ConvNeXt 的 CNN 支线来完成局部特征提取,形成了“全局-局部”双分支并行协同模块架构。

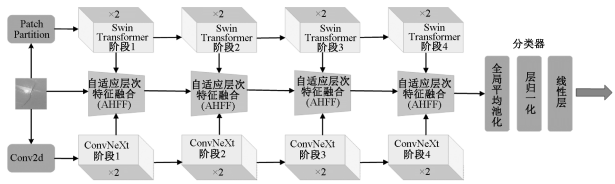

图 2　HMFF 网络结构

Fig. 2　HMFF network architecture

如表 1 所示,总结了视觉 Transformer 基于模块化思想与 CNN 网络进行局部模块或主干网络并行融合检测的相关应用方法。

3.4　小结

本章围绕视觉 Transformer 在长距离依赖建模驱动下的全局特征提取能力、多模态融合优势及模块化设计等这 3 个方面技术优势,从缺陷检测任务的实际应用需求为出发点,详细分析了视觉 Transformer 近年来在复杂纹理背景抑制、基于多模态融合的缺陷检测以及基于 CNN-Transformer 在模块化设计思想驱动下的联合应用。本章针对视觉 Transformer 技术优势驱动下的多任务应用方法的总结,为实现高精度缺陷检测任务提供了有效解决方案。

4　视觉 Transformer 应对表面缺陷检测中关键挑战的方法

为突破表面缺陷检测任务存在样本量稀缺、实时性不足、训练效率低以及小目标检测精度低等核心瓶颈,国内外学者从小样本学习与数据增强技术、计算效率提升、部署加速优化以及微小缺陷感知能力增强等多个维度提出了相应的解决策略。本章旨在对此进行系统性梳理与总结,将重点围绕小样本学习、模型推理的加速策略、复杂模型轻量化方法、模型训练效率的优化途径以及小目标特征增强机制等几个关键方向,深入剖析当前的解决方案,旨在为基于视觉 Transformer 的缺陷检测模型优化提供明确的设计思路与技术路径。

4.1　小样本学习与数据增强技术

一方面,小样本问题作为缺陷检测领域最主要的挑战之一,样本数量欠缺易造成数据类别不平衡而难以有效覆盖数据分布的多样性与复杂性,从而可能引发算

表 1 视觉 Transformer 基于模块化思想与 CNN 的联合应用方法

Table 1 The combined application methods of vision Transformer based on modular design and CNN

文献	缺陷检测场景	模块化思想	模块化方法	贡献
[27]	瓷砖检测	局部嵌入	将 MobileViTv3 嵌入 YOLOv5 的主干网络末端	*mAP*@0.5 提升 1.1%
[43]	木材检测	局部替换	是在主干网络的末端,用结合了 Transformer 编码器的 C3TR 模块替换了原有的一个 C3 模块	*mAP*@0.5 提升 4.6%
[44]	道路内部结构	主干网络替换	将原始 YOLOX 的主干网络 CSPDarknet 替换为 Swin Transformer Tiny(Swin-T)	*mAP*@0.5 提升 10.18%
[45]	金属表面	并行检测	将 Swin Transformer 网络作为全局特征提取分支,与基于 CNN 的 ConvNeXt 主干网络提取分支并行检测	在多个金属表面数据集上实现精度与效率的平衡
[46]	绝缘子检测	局部替换	Swin Transformer 块嵌入到原有的主干网络中的 C2f 模块中,构成 C2fSTR,并对整个网络的 C2f 模块进行替换	*mAP* 提升 2.1%
[47]	输电线路	局部嵌入	Transformer 注意力机制被并联地嵌入到 YOLOv8 的骨干网络中	在多环境下,*mAP* 均有提升
[48]	绝缘子检测	主干网络替换	使用 Swin Transformer 作为主干网络,替代原有的 CNN 结构	*AP* 和召回率均有稳定提升
[49]	钢铁表面	主干网络替换	将 Swin Transformer 作为主干网络整体替换掉 CNN 骨干,并集成到 Faster R-CNN 检测框架中	*mAP* 达到 81.1%,显著高于其他模型
[50]	带钢表面	局部嵌入 局部融合	(1)在主干网络末端嵌入 Swin Transformer 模块; (2)在特征融合层中引入了基于 Transformer 的 CGT 模块	平均交并比达 75.16%,提升 7.83%,*F*1 提升 6.3%。
[51]	电池检测	局部替换	将 Swin Transformer 块嵌入 YOLOv8 的主干网络末端,并与最后 1 个 C2f 模块融合,构建 C2f-ST 模块。	*mAP*@0.5 从 86.7% 提升至 86.9%
[52]	PCB 检测	局部嵌入	将 Swin Transformer 块作为增强模块,分别嵌入到了主干网络的末端和颈部网络的特征金字塔路径中。	*mAP* 提升 1.35%,召回率提升 0.2%
[53]	木材检测	局部替换/ 局部嵌入	Transformer Encoder 嵌入到主干网络中,替换末端 C3 模块,形成新的 C3TF 模块;Swin Transformer 块被嵌入到检测头部分,形成新的检测头 Shead、Mhead、Lhead	引入视觉 Transformer 思想,并进一步结合坐标注意力,使得 *mAP* 提升 3.1%
[54]	织物检测	主干替换	使用 Swin Transformer 作为主干网络,替代 YOLOv5 原有的 CSPDarknet	*mAP* 提升了 4.4%
[55]	变电站设备检测	局部替换	在 YOLOv5 的主干网络末端的 CSP 网络中用 Transformer block 替换原有的 Bottleneck 结构,形成新的 CSP 结构	精确率和 *mAP* 有所提升
[56]	钢板检测	局部嵌入	在 MobileNetv3 主干网络末端嵌入 Transformer 编码模块	*mAP* 提升约 4.85%
[57]	热轧钢带	局部替换	在 Neck 部分的 C3 模块中嵌入 Swin-Transformer 块,形成 C3STR 模块,并替换了 Neck 中的 4 个 C3 模块	*mAP* 提升 4.1%
[58]	多任务 (钢铁、PCB)	主干网络替换	以 Faster R-CNN 为基础框架,将 Faster R-CNN 中的 CNN 骨干网络整体替换为 Swin Transformer	多任务场景下引入 Swin Transformer 后性能提升
[59]	金属表面	并行融合	设计了双分支特征提取模块:分支 1 基于卷积的局部特征提取;分支 2 基于 Swin-Transformer 块并融合 CNN	*mAP*@0.5 提高了 1.3%

法模型召回率低、漏检率高、泛化能力差及过拟合等问题,为此国内外学者围绕小样本学习从增量学习、度量学习[60]、元学习[61]等多种方法进行了广泛研究;另一方面,视觉 Transformer 高性能的背后正是需要超大规模的样本量的支持,即使随着视觉 Transformer 技术的不断演化而发展了诸如 Swin Transformer、DeiT 以及 RT-DETR 等变体方法而使得视觉 Transformer 技术在小数据集上发挥高性能成为可能,但依然难以撼动基于 CNN 技术的相关分类或检测算法模型在小数据集场景下的缺陷检测核心应用地位。所以如何使得具备全局特征提取能力的视觉

Transformer 技术能够在小样本场景下高效应用,一直是国内外学者研究的热点。结合视觉 Transformer 技术的相关研究成果,可以从 3 个方面来探讨视觉 Transformer 在小样本场景下的应用关键技术:基于数据或特征的增强策略、基于模型方法改进的增强策略、基于知识迁移与先验信息引入策略等,即:

1) 基于数据或特征的增强策略

小样本场景的核心问题不仅在于数据量不足,更在于数据的信息密度与泛化能力不足。其本质体现在:有限样本难以覆盖真实数据分布,导致模型无法学习到丰富完整的缺陷特征,从而在未知样本上泛化能力显著下降。解决小样本场景最直接高效的方式就是针对现有数据进行扩充增强或引入其他模态数据以增强数据的泛化表示能力。

(1) 单模态数据扩充增强

基于现有数据的扩充增强,直接快速有效的方式就是采用如旋转、裁剪等几何变换、颜色扰动等常规数据增强方法来提升原有样本多样性。Xi 等[48]通过常规数据增强与合成缺陷样本等操作,将原有数据集规模从 2 002 张扩展至 10 223 张,显著增加了缺陷样本多样性,同时引入可变形感受野模块(deformable receptive field block, DRFB)模块并通过用 Swin Transformer 主干网络来替换原有 CNN 网络等改进方法,从而使得缺陷检测的平均精度从 81.0% 提升至 87.1%。Wang 等[62]采用了相关数据增强操作——包括亮度、对比度、饱和度及色调调整、多角度旋转、镜像翻转及高斯噪声注入等多种数据增强方法,模拟了真实环境中的成像变化,从而将原始数据集扩增了 14 倍,有效提升了数据的多样性和规模。

(2) 基于多模态技术的多特征数据增强

虽然单模态数据增强可以快速实现数据量翻倍增加,但由于其只是通过旋转、裁剪等常规操作来对原始数据进行的一种变换,本质上是一种泛化性能的“表面体现”,难以真正覆盖真实场景的多特征泛化与语义变化,而且不当使用相关样本增强方法反而会引发过拟合问题,甚至因为裁剪不当导致关键特征信息被破坏。基于多模态技术的多维特征数据增强,其价值体现在不会破坏原有数据集特征信息与语义前提下,显著增强其他维度特征信息互补能力,从而实现与原有单模态数据结合以生成更合理、多样化且富含语义信息的多维度数据。同时,在基于多模态进行多特征互补增强的同时,可以进一步针对多模态数据进行增强操作,由于多模态数据源的异构性,虽然难以统一应用传统图像数据增强的方式对多模态数据进行增强,但是可以从两点进行:一方面可以将部分模态数据转化为二维图像数据从而采用图像旋转、缩放、裁剪等传统图像数据增强方式,如 Liu 等[63]就通过 Gramian 角求和场(Gramian angular summation field, GASF)将一维时间序列特征转换为二维图像;另一方面也可以针对不同模态的数据特征进行针对性增强,如通过同义词替换、随机插入、删除、交换、掩码以及噪声注入等方式来对文本模态数据进行增强[64-65]。值得注意的是,尤其在样本量稀缺场景下,基于多模态技术进行数据融合,需要充分考虑到各模态数据在不同样本上的质量分布的均匀性与模态间信息动态变化的可靠性,避免信息融合过程受噪声干扰或模态缺失等因素的影响而无法充分发挥融合优势。

2) 基于模型方法的改进策略

(1) 模型改进优化

在小样本学习场景下,如何有效引入视觉 Transformer 以突破样本量稀缺问题对模型表征能力的制约,国内外学者基于此并围绕视觉 Transformer 相关改进方法进行了一系列研究。一方面,基于 ViT 或 DETR 等现有通用视觉 Transformer 架构进行改进,如 Swin Transformer、MobileViT、RT-DETR 以及 Deformable-DETR 等,尽管这些改进并非直接针对小样本场景设计,但其技术创新间接提升了视觉 Transformer 在小样本场景下的检测效率与鲁棒性:例如,Swin Transformer 通过引入多尺度感知与局部窗口注意力计算来优化特征提取;MobileViT 结合深度可分离卷积来降低模型复杂度,适配了小样本场景下的轻量化需求;Deformable-DETR 借助可变形注意力机制来减少全局注意力的计算冗余,实现了对关键区域的精准聚焦;RT-DETR 则通过 CNN 与 Transformer 的混合结构设计,平衡了特征捕捉能力与计算成本;另一方面,国内外学者专门针对小样本场景对相关方法进行了改进,如 Liu 等[28]针对原有 Swin Transformer 网络在织物缺陷检测任务的小样本场景下存在过拟合的问题,通过在 Swin Transformer 架构中引入了 Res-Drop Swin Transformer 块,该模块集成了残差连接与 DropKey 正则化策略:残差连接确保了深层网络中的梯度有效传播,缓解了训练退化问题;而 DropKey 机制则通过随机丢弃注意力计算中的键值,惩罚过于集中的注意力分布,从而促进模型捕获更全面的全局上下文信息,显著增强了模型的泛化能力。Huang 等[66]考虑到交叉 Transformer(cross-Transformer)在工业表面缺陷检测任务中面临的小样本挑战,提出了一种稀疏交叉 Transformer 网络(sparse cross-Transformer network),一方面以交叉 Transformer 为骨干,利用跨注意力机制来聚合查询集与支持集的关键特征,强化了小样本场景下有限标注样本与待检测样本的特征关联;另一方面通过引入残差层来保留高低层特征信息,避免了数据稀缺导致的特征提取不充分,同时将挤压-激励模块集成于编码器之后,通过动态学习并重校准特征通道的权重,自适应地强化对缺陷敏感特征通道的响应,同时抑制无关背景干扰,从而显

著提升了模型对关键特征的聚焦与利用效率。Deng 等[67]考虑到小样本场景下类别不平衡问题,通过水平和垂直方向的空间-通道注意力融合来动态分配同类样本的权重,并通过改进损失函数(modified cost function, MCF)显著提升了小样本场景下模型抗噪性和跨任务泛化性能。

(2) 无监督学习

在小样本视觉任务中,核心挑战在于缺陷样本量稀缺,而无缺陷的正常样本则易于获取且样本量充足。因此,如何通过有效机制充分挖掘并利用大量正常样本中蕴含的结构、纹理等先验信息,成为推动视觉 Transformer 在该场景下成功应用的关键策略之一。基于无监督学习的缺陷检测方法可以通过仅依赖无缺陷样本来实现异常检测,从而显著降低对缺陷样本的依赖。无监督学习的核心思想是学习正常数据的分布或特征模式,并通过检测与正常模式的偏差来识别缺陷。基于无监督学习思想的典型方法包括:图像重构与残差分析、特征提取与特征建模、基于生成对抗网络以及基于自编码器等方法。黄媛媛等[32]针对复杂花纹色织物缺陷检测场景中缺陷目标样本数量少且种类不平衡的问题,提出一种基于 U 型 Swin Transformer 自编码器的方法来应对复杂花纹色织物缺陷检测。通过无监督重构正常样本的分布,同时利用残差分析定位异常区域,并通过对无缺陷样本叠加高斯噪声,让模型学习“从带噪正常样本中重构无噪正常样本”,而无需缺陷样本的标注信息,从而规避了缺陷样本量稀缺和类别不平衡问题,本质上将缺陷检测转化为分布差异分析,突破了传统方法对缺陷样本的强依赖性。Zhao 等[68]针对金属表面缺陷小样本检测的困局,提出了一种基于 Transformer 与多尺度掩码特征融合的金属表面缺陷检测方法,将 ViT 嵌入生成对抗网络,仅使用正常样本且通过多尺度掩码遮挡输入图像,从而生成对抗网络来重建正常图像,一方面通过掩码重建策略避免了对缺陷样本的依赖,另一方面则利用 ViT 的全局注意力机制与多尺度掩码的局部遮挡策略相结合来覆盖不同尺寸的潜在缺陷,从而提升了模型对小样本异常模式的敏感性。邓凯丽等[69]通过引入缺陷拟合模块、基于 Transformer 的图像全局修复模块以及端到端检测等优化技术,从而构建了一个改进的掩码自编码器(masked autoencoder, MAE)无监督缺陷检测框架,仅需依赖正常样本即可训练,核心在于通过“合成缺陷来增强自监督掩码重建”替代对真实缺陷样本的依赖,无需任何缺陷标注或监督信号。

3) 基于知识迁移与先验信息引入策略

数据扩充增强策略是通过引入样本变换与生成方法来直接解决小样本场景下数据多样性不足的本质问题;算法模型增强策略则聚焦于改进算法架构与训练机制以提升模型在小样本条件下的内在适应能力。相比较前两者,知识迁移与先验信息引入则是降低算法模型对数据的依赖性,从而提升系统对小样本场景的鲁棒性能。即通过复用预训练知识、注入领域规则或引入先验信息等方式,从底层逻辑降低模型对样本数量的绝对依赖。

(1) 知识迁移

视觉 Transformer 技术在大规模数据集支持下可以展现出超越 CNN 的性能优势,所以在小样本检测场景中,可以充分利用预训练 Transformer 大模型,基于大模型通用领域知识 + 特定检测领域小样本,从而实现算法模型对小样本的利用。Chen 等[70]以通用分割大模型(segment anything model, SAM)为预训练基础,在保留 SAM 通用视觉知识的前提下,通过低秩适配技术向 SAM 大模型中注入织物缺陷的领域知识,即通过基于 SAM 的 ViT 架构,在 SAM 的 Transformer 自注意力模块中引入可训练的织物缺陷相关参数,使模型同时融合自然图像的通用特征和织物缺陷的专有特征,从而解决了织物缺陷分割中的数据稀缺与泛化难题。

(2) 先验信息引入

先验信息不仅仅局限于经验知识或语义先验信息,还包括如正则化、结构化以及多模态数据特征等,都可以被认为是某种形式的先验信息。常规算法模型之所以需要借助于大规模数据,一方面是无法预知目标特征子空间,只能在大规模数据集中以某几种统计的规则或方法进行全域搜索;另一方面由于算法模型的参数量超过了数据集的支撑度就会导致模型在小样本场景中倾向于噪声学习从而造成过拟合问题。针对上述问题,先验信息的引入不仅可以通过注入领域经验来缩小样本搜索子空间和限定学习区域以减轻因为搜索域的问题而产生的对数据集规模的依赖程度,而且还可以通过隐式约束与添加显式正则项等方法来抑制因为小样本造成的过度拟合噪声问题。如图 3 所示,李刚等[13]针对当前输电线路巡检中的螺栓缺陷小样本检测场景,利用视觉-知识注意力模块将螺栓图像的视觉特征与螺栓先验知识有机融合,获得螺栓对应的增强视觉特征,并将增强视觉特征送入基于 Transformer 思想的 DETR 模型框架中,来对螺栓目标进行识别与分类。引入的先验知识能够引导模型即使在缺乏足够样本量的情况下也能进行有效的推理,例如,螺栓的正常状态和缺陷状态具有明确的装配规则与结构性定义,先验知识通过这种规则对目标进行约束,使得模型能够根据已知的信息推测未见过的缺陷类型。这样,即便缺少某些类别的样本,模型也能基于已有的知识进行合理预测,从而减少对大量数据的依赖。

基于数据变换与多模态技术的数据扩充增强、基于视觉 Transformer 算法模型改进与无监督思想引入以及基于知识迁移等策略分别从数据、模型与知识维度形成

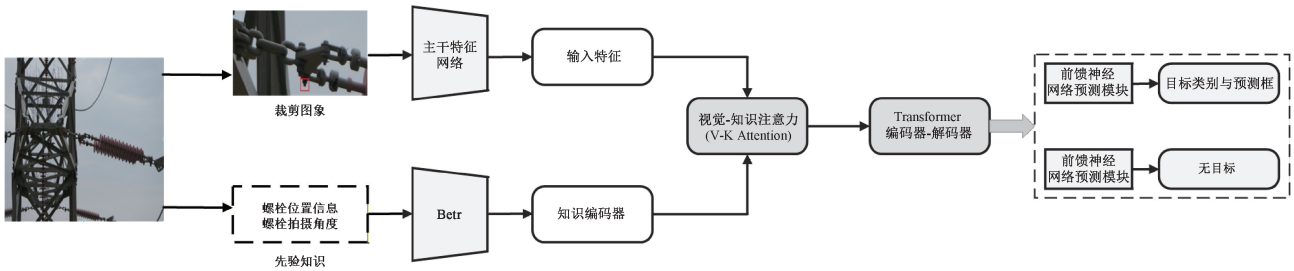

图 3　基于 DETR 与先验知识融合的螺栓缺陷检测框架

Fig. 3　Bolt defect detection framework based on fusion of DETR and prior knowledge

互补,共同构建了基于 Transformer 技术在小样本缺陷检测场景学习的系统性解决方案。

4.2　模型轻量化与加速推理

针对视觉 Transformer 在轻量化部署与实时性不足方面存在的参数量庞大、计算成本高、难以适配车载/无人机/相关边缘设备以及推理速度慢等挑战,国内外学者从结构重设计、注意力机制优化、模型压缩与硬件适配等多个维度提出了系统性改进方法,通过采用轻量级骨干网络、设计高效注意力模块、引入模型剪枝与量化技术、融合 CNN 与 Transformer 的混合架构,以及使用重参数化、知识蒸馏等技术,在显著降低模型参数量、计算复杂度和推理时间的同时,有效保持了甚至提升了模型在不同缺陷检测场景中的精度与鲁棒性,实现了检测性能与部署效率的平衡。

1) 基于 CNN-Transformer 混合架构的轻量化

基于 CNN-Transformer 混合架构的轻量化策略核心思想在于因考虑到如 DETR 等相关视觉 Transformer 复杂网络所固有的计算复杂度大、模型参数量大等瓶颈,从而尽量避免直接采用相关计算密集的纯 Transformer 模型,替代以 MobileNetv3、RepVGG 等轻量级卷积神经网络作为主干基础,仅在网络深层或特定模块中引入经过简化的 Transformer 组件,即以轻量级 CNN 为主体、Transformer 为补充的混合模型。从而以最低的计算代价获取全局上下文信息,实现检测性能和计算开销的平衡。

张林等[56]以轻量化 MobileNetv3 为基础来构建主干网络,仅在网络末端低分辨率特征图中融入 Transformer 编码模块,避免了 Transformer 在高分辨率特征处理中的大量计算,同时结合协调注意力瓶颈模块(coordinate attention bottleneck, CA-Bneck)来增强特征表示,网络整体主要基于 CNN 进行特征提取,Transformer 模块仅作为深层全局特征补充,使模型在钢材缺陷数据集上 mAP 达 80.46%,检测速度 FPS 达 20.1,实现了精度与效率的平衡。又如 Duanmu 等[71]针对实木地板纹理与颜色分类任务中的轻量化与实时性需求,提出了一种基于重参数化技术的 CNN-Transformer 混合模型(re-parameterized CNN-Transformer hybrid model based on MobileViT, Rep-MobileViT),该方法以 CNN 结构为核心,仅在多特征融合环节构建重参数化多特征融合模块(re-parameterized multi-feature fusion MobileViT, RepMFF-MobileViT),通过引入简化的 Transformer 相关特征融合策略,用 1×1 卷积替代高开销的 3×3 卷积,在减少参数量的同时实现了局部与全局特征的结合,最终在参数量减少 30.53% 的情况下仍能保持 94.98% 的分类准确率。

2) 针对 Transformer 核心模块的轻量化设计

针对 Transformer 核心模块的轻量化设计聚焦于视觉 Transformer 思想自身核心模块的算法级优化,旨在通过重构其核心组件——尤其是自注意力机制——以解决其固有的计算复杂性与参数冗余问题,主要可归纳为 3 类:

(1)注意力机制的结构稀疏化,即通过设计结构化稀疏注意力机制,如可变形注意力[72]、条带注意力[73]、空间缩减注意力[74]以及分流窗口机制[75],有效限制每个查询与键的交互范围,或通过降维减少参与计算的 Key-Value 对数量,从而降低注意力计算复杂度;

(2)计算单元与过程的优化,即在模块内部集成或改用更高效的计算单元,例如引入局部增强模块或采用深度可分离卷积[76]来替代部分全连接操作,以减少矩阵运算量,此外,也可以通过引入如随机丢弃部分 Key 的正则化方法[77],或利用全局载体令牌限制交互[78],从而进一步优化计算过程并增强模型泛化能力;

(3)复杂组件的轻量级替代,即采用计算成本更低的架构替代标准 Transformer 内部组件,例如利用 CNN 模块进行特征嵌入或使用由多层感知机(multilayer perceptron, MLP)层构成的解码器,在减少参数量的同时,缓解了因特征压缩而导致的信息丢失问题。上述针对视觉 Transformer 核心模块的根本性重构,在提升其全局建模能力的前提下,显著降低了模型的计算负担与参数量,同时实现了检测性能的提升。

3) 其他轻量化方法

除了上述两种主要轻量化方法,还包括其他一些轻量化措施。如以后处理加速、损失函数优化、模型压缩、

高效特征融合等多个环节为出发点,并进一步通过多技术协同与互补,从而实现检测框架整体的轻量化与加速。

如表 2 所示,总结了国内外面向轻量化与加速推理挑战的视觉 Transformer 优化策略与方法。

表 2 面向轻量化与加速推理挑战的视觉 Transformer 优化策略

Table 2 The optimization strategies for vision Transformers towards the challenges of lightweight design and efficient inference

文献	方法	数据集与规模	贡献	局限性
[31]	采用双分支金字塔模块替代复杂卷积结构,结合递归多级残差参数共享,通过混合连接减少信息损失	ZJU-Leaper-18855 HKU-Fabric-3633	参数量显著减少, ZJU 数据集上推理时间为 43 ms	小缺陷易漏检
[53]	移除部分 C3 模块,引入 C3TF、BiFPN、Group Convolution,降低参数量	自建数据集-3606	参数量从 7.1 M 降至 6.0 M,*FPS* 为 74.6,*mAP* 提升至 84.2%	对细裂纹和类似纹理的缺陷检测效果待提升
[56]	融合 MobileNetv3、Transformer 编码模块和 CA-Bneck,构建轻量主干网络	NEU-DET-1800	*mAP* 提升 5.96%,检测速度 *FPS* 达 20.1,参数量为 71.9 M	模型参数量仍较大,实时性有待进一步提升
[71]	引入 RepAIRB 模块进行重参数化结构、使用 1×1 卷积替代 3×3 卷积减少参数	自建数据集-13951	计算量减少 42.31%、参数量减少 30.53%、推理速度达 20.94 ms/单张	尽管速度达到实时要求,但其分类精度提升有限
[72]	引入 GSConv、Res-GSCSP 模块,结合动态可变形卷积与注意力机制,实现轻量化	GRDDC2020-14569	计算量为 50.2 G、参数量为 19.6 M、检测速度 *FPS* 为 42.3	对复杂路况下细小的裂缝检测效果不佳
[74]	全 MLP 层轻量解码器、Criss-Cross 注意力模块	Type-I RSDD-67 Type-II RSDD-128 自建 RSD-111	显著降低计算复杂度,提升了推理效率	对高斯噪声敏感
[75]	用分流大小窗口替代移位窗口,引入局部连接模块	滚动轴承数据集 Bearing500-600	计算量降低 0.6 G,参数量减少 3 M,吞吐量略有下降	大窗口设计增加内存与计算负担,收敛波动较大
[76]	采用轻量化自注意力模块与深度可分离卷积,减少参数与计算量	自建数据集-3436	计算量为 164 G,参数量 51 M,*FPS* 达 14	推理速度提升有限
[77]	参数量化、知识蒸馏、轻量化主干网络(ResNet50)及注意力计算正则化方法	SFID-13000	模型参数量降至 9 M,推理速度 *FPS* 提升至 42,速度较基线提升明显	检测精度有所下降
[78]	引入部分卷积(PConv)减少计算冗余,结合层级注意力机制降低 ViT 计算复杂度	PKU-Market-PCB-693 Deep PCB PCBA	检测速度 *FPS* 达到 25	计算复杂度仍相对较高
[79]	使用轻量级 RES15 骨干网络,采用多尺度特征图与可学习位置编码网络	自建数据集-2730	推理速度比 DETR 快近 3 倍	检测精度有下降
[80]	采用 iRMB 模块减少参数量、添加 P2 检测头优化细节处理、损失函数优化	自建数据集-9828	参数量降低至 30.7 M,计算量为 164.7 G	

4.3 高效训练策略与方法

为解决引入视觉 Transformer 思想后而产生的诸如训练过程收敛缓慢、计算开销大、易过拟合及易陷入局部最优解等效率低下问题,相关学者通过引入低秩适应、部分参数微调、Token 合并与剪枝、渐进式归一化替换、预训练权重迁移、轻量级结构设计以及引入先验特征增强等方法,分别从参数效率优化、训练策略创新及利用预训练网络与先验知识等方面提出了相关改进方法,显著降低了模型训练的计算与内存开销,加速了收敛过程,并有效缓解了过拟合与局部最优问题,大幅提高了训练效率。

1) 参数高效微调策略

参数高效微调策略的核心在于冻结预训练模型的大部分参数,仅通过增量更新少量额外参数或选择性优化关键模块,以大幅降低训练时的计算和存储开销,同时保持模型性能。Chen 等[70] 通过应用低秩适应技术(low-

rank adaptation, LoRA)并冻结 99%以上的预训练参数,在仅训练模型总参数 0.18%~0.72%的情况下,在跨数据集和小样本场景中实现了高效微调,提升了训练效率。Zhang 等[81]在电力设备异常检测中,通过改进 ViT 模型并采用仅更新注意力层参数的策略,将需优化的参数量限制在模型总参数的 30%以内,在保持精度的同时节省了 10%的显存并使训练速度提升了 10%。

2) 结构重参数化与动态计算

结构重参数化与动态计算的核心思想是在训练阶段采用复杂结构以保障性能,而在推理阶段转换为简化结构,或在训练过程中动态调整计算方式以减少计算量,从而平衡训练稳定性与推理效率,有效降低了计算复杂度和内存开销。Duanmu 等[71]通过渐进式重参数化批归一化方法(progressive re-parameterized batch-norm, PRepBN),在训练初期以层归一化为主维持稳定性,随训练步数增加逐步过渡至批归一化,最终在推理阶段实现批归一化与相邻线性层的合并,避免了直接替换可能引起的训练崩溃,提升了计算效率。

3) 利用预训练与先验知识

利用预训练与先验知识策略强调借助大规模预训练模型或手工设计的外部先验知识,来为模型提供强大的特征基础,降低了学习低维特征的难度,从而加速收敛、避免局部最优和过拟合。Xu 等[82]通过在模型输入端嵌入参数固定的 Gabor 滤波器组,为织物图像缺陷检测提供预处理后的纹理特征先验,降低了模型在训练初期理解复杂背景的难度,省去了底层特征学习所需的大量迭代,为加速早期收敛奠定了基础。Zhu 等[83]在钢材缺陷检测中,通过引入迁移学习策略,使用 ImageNet-1K 预训练权重初始化线性平移窗口 Transformer(LSwin Transformer)网络,有效缓解了随机初始化导致的收敛缓慢和过拟合问题,使模型在钢材缺陷数据集上仅需约 10 个训练周期即可收敛,相比随机初始化的 42 个训练周期大幅缩短。

如表 3 所示,总结了视觉 Transformer 面临训练效率低下的问题所提出的解决方案与相关方法。

表 3　视觉 Transformer 高效训练策略与方法

Table 3　The methods and strategies of vision Transformer to achieve efficient training

文献	方法	数据集与规模	贡献	局限性
[33]	LoRA 低秩适应技术和 Q-Former 跨模态对齐模块	自建数据集-4400	显著降低训练参数量,提升训练速度	基线模型规模受限,参数配置需优化
[68]	Token Merging (ToMe) 剪枝与合并策略	自建数据集-2778 MVTec AD-5354	训练时间平均减少了 10%,每轮训练迭代处理数据量平均减少了 8%	计算资源需求仍然较高
[70]	参数高效微调、低秩适应(LoRA)、领域知识注入;损失函数优化	Fabric Stain-394 ATTEX-185 自建数据集-1515	可训练参数占比低至 0.18%~0.72% 显著降低训练时的显存占用、计算量	-
[71]	渐进式重参数化批归一化(PRepBN)	自建数据集-13951	PRepBN 提升训练稳定性,无额外计算成本,避免了训练崩溃问题	增加了模型训练前的设计复杂性和调优成本
[81]	hMLP 块嵌入结构(分层特征映射)、参数微调	自建数据集-120	需训练参数量降低约 30%;节省显存约 10%;训练速度提升约 10%	模型复杂度增加,仅微调注意力层可能影响泛化性
[82]	前置特征增强(Gabor 滤波器)、损失函数优化(FDL)、非完全对称 U 型结构	Cropped ATTEX、Colored Fabric-835	训练曲线初期缓增期消失,收敛速度显著加快	过于依赖特定领域先验知识
[83]	预训练与迁移学习	NEU-1800 GC10-DET-2294	仅需约 10 个训练周期即可收敛	计算复杂度高、依赖预训练模型

4.4　小目标特征增强机制

针对视觉 Transformer 在小目标检测中的不足,国内外学者从损失函数优化、多尺度特征融合以及注意力机制优化等角度来对相关视觉 Transformer 模型进行了针对性改进,包括引入超分辨率重建、局部感知增强以及动态注意力机制等策略,显著提升了小目标特征的表征能力与检测精度。

1)损失函数与匹配策略优化

损失函数与匹配策略优化方法聚焦于通过改进损失函数和优化目标匹配过程:一方面对现有损失函数进行创新与组合;另一方面也可以对目标与预测之间的匹配过程进行优化,从而直接提升小目标边界框的回归精度和训练稳定性。例如,张乃雪等[14]通过结合 Smooth-L1 损失与广义交并比(generalized intersection over union, GIOU)损失的优势,优化了边界框的回归精度,实现了

小目标缺陷如 class3 的曲线下面积(area under curve, AUC)指标提升了 1.8%,且 class5 达 100%。Zhou 等[75]在损失函数中采用了基于二元交叉熵的掩码损失(mask loss, Lmask),通过像素级误差计算进一步提升了小目标边缘的定位精度。

2)多尺度特征融合

通过构建或利用多尺度特征图,来同时捕获全局上下文信息和局部细节,核心是为了解决小目标特征在骨干网络下采样过程中易丢失的固有问题。基于多尺度特征融合策略来解决视觉 Transformer 在小目标检测中的不足,可以从两方面着手。一方面,可以在网络中显式地构建多尺度特征模块。如 Ge 等[79]针对 DETR 因特征多次下采样导致小缺陷细节丢失的问题,构建了多尺度特征融合网络,通过采用 16×16、32×32、64×64 这 3 种尺度的特征图,并利用大尺寸特征图来捕捉局部细粒度信息,弥补了小目标细节丢失问题;另一方面,则可以直接选用或设计具有层次化多尺度特征表征能力的骨干网络,如 Yao 等[29]将 DINO 的骨干网络替换为 Swin Transformer,正是基于其层次化多尺度特征表征能力来扩大感受野并适配不同尺度的小目标。Liu 等[84]则通过引入 ResNeXt 网络作为骨干,强化特征复用性以保障低维特征的有效传递,减少了小目标特征的信息丢失。

3)注意力与特征提取机制优化

通过对注意力机制或相关卷积模块等网络核心模块进行改进,以增强对局部特征的感知能力,可以解决 Transformer 架构在视觉任务中可能存在的局部信息提取不足、感受野有限等问题,从而提升小目标的检测能力。

(1)增强局部感知与空间特征提取能力

针对 Transformer 局部特征感知能力弱的缺点,相关学者通过引入可变形卷积或空洞卷积等操作,来增强网络细节信息获取能力。Zhou 等[75]引入了基于深度卷积的局部连接模块,从而增强了相邻窗口间的信息交互,弥补了因窗口分割造成的特征断裂。如图 4 所示,Dong 等[76]设计了增强局部感知单元(enhanced local perception unit, ELPU),通过多分支不同膨胀率的空洞卷积来扩大空间特征感受野,强化了局部空间特征提取。Hu 等[85]在 Swin Transformer 中通过引入可变形卷积块,有效提取了细长缺陷的局部特征。

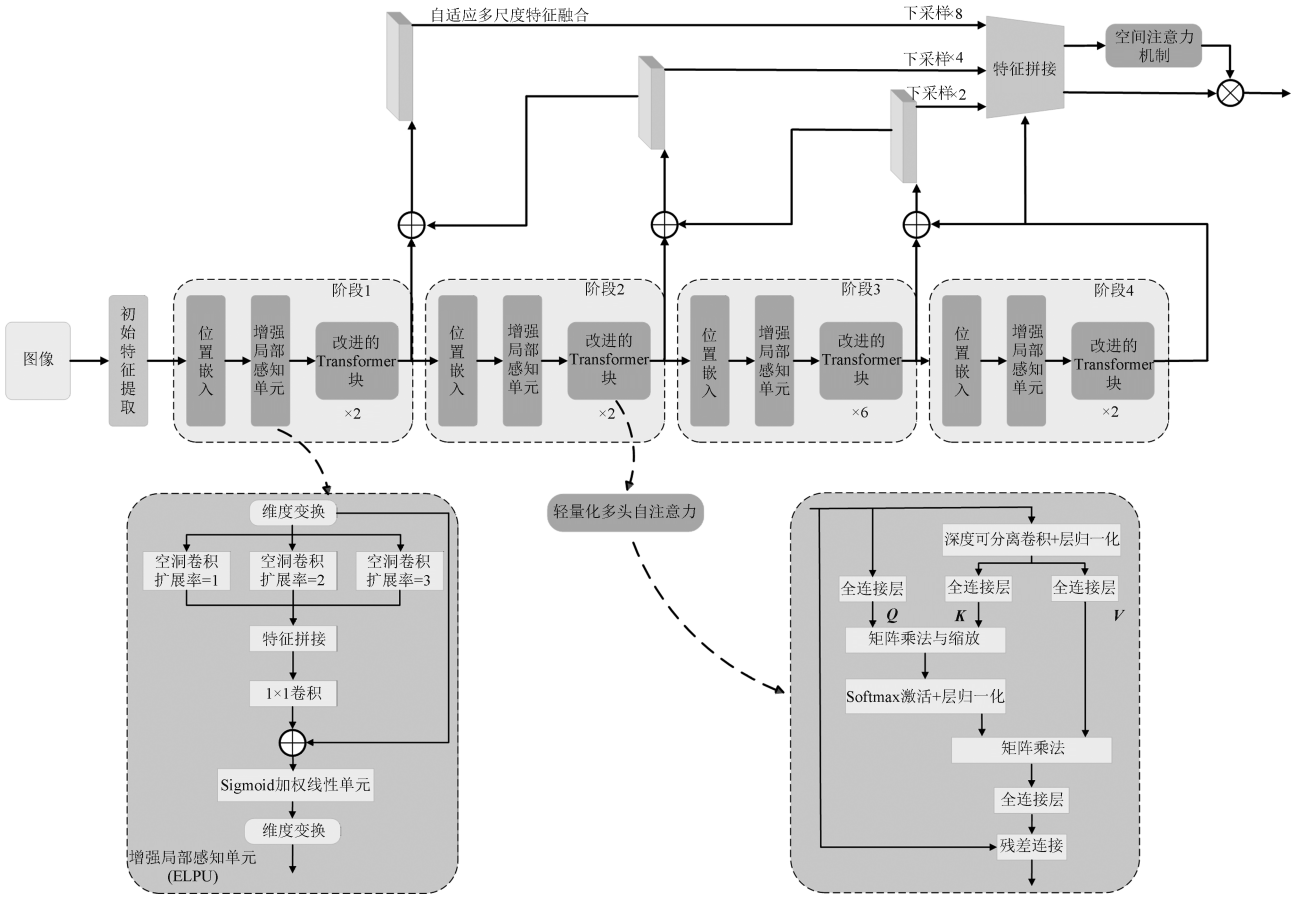

图 4 改进的 Swin Transformer 主干特征提取网络

Fig. 4 Improved Swin Transformer backbone

(2)针对自注意力机制的优化

直接对视觉 Transformer 自注意力机制进行优化,可以使其更适应小目标检测。Yao 等[29]在 Swin Transformer 中引入可变形注意力机制以适配小目标形状与尺度。李鹏等[73]将自注意力机制改进为条带内和条带间注意力块,在水平与垂直方向上提取特征以平衡局部与全局信息。Zhou 等[75]提出分流窗口机制,使不同注意力头能并行捕捉细粒度和粗粒度特征,从而打破了固定尺度感受野限制。

(3)优化位置编码

位置信息对小目标定位至关重要。Ge 等[79]设计自学习位置编码网络来替代手动编码,适配了小目标位置学习。Ding 等[86]则通过将对数空间连续位置偏置(logarithmic spatial continuous position bias, Log-CPB)引入 Swin Transformer,使其能动态适应不同窗口尺寸,从而提升了小目标感知能力。

如表 4 所示,总结了视觉 Transformer 针对小目标检测挑战下的相关应对方法。

表 4　视觉 Transformer 针对小目标检测挑战下的应对方法

Table 4　The strategies of vision Transformer for small object detection challenges

文献	方法	数据集与规模	贡献	局限性
[14]	全局特征学习、损失函数融合、匹配策略优化	DAGM 2007-6800 KolektorSDD-399	class3 的 *AUC* 提升 1.8%,class5 提升 3.8%	依赖特定数据集,泛化性能可能不足
[29]	超分辨率重建、数据增强、替换主干为 Swin Transfomer 并引入可变形注意力	阿里天池织物数据集-9576	*mAP*@0.5 提升 20.8%;“Knot”提升 18.5%,“Broken warp”提升 33%	超分辨率重建对硬件性能要求高,训练时间较长
[73]	条带注意力模块、引入对比学习损失	CPLID、自建数据集、挑战赛-1941	YOLOv5/v7 对绝缘子和自爆缺陷的 *mAP* 分别提升 37.6% 和 38.7%	模型复杂度有所增加;模型泛化性能待验证
[75]	分流大小窗口机制替代移位窗口,引入局部连接模块	滚动轴承数据集 Bearing500-600	在 Bearing500 数据集上,边界框平均检测精度提升了 1.1%~3.4%,掩码平均检测精度提升了 0.9%~1.6%	模型轻量化性能不足,复杂背景下检测精度下降
[76]	局部感知增强(ELPU)、自适应多尺度特征融合(AMF)	自建数据集-3436	小目标检测的平均精度达 23%,优于相关主流模型	模型泛化性能待验证
[79]	多尺度特征融合、自学习位置编码、损失函数优化	自建数据集-2730	在活节、死节小缺陷的平均精度分别提升 0.6%、0.5%	未较好地平衡检测精度与速度
[84]	引入 ResNeXt 网络提取特征,使用可变形注意力机制提升精度,FFN 预测边界框	自建数据集-1361	整体平均精度为 54.1%,小目标的检测平均精度为 35.4%	模型计算量大,对硬件要求高;小目标精度仍较低
[85]	引入可变形卷积块	自建数据集-5474 TILDA-3000	精确率提升 3.69%,召回率提升 3.12%	模型复杂度较高
[86]	引入对数空间连续位置偏置	自建数据集-500	小目标的 *mAP* 提升了 20.8%	可能会增加计算开销
[87]	引入位置-尺度约束机制,基于强化学习驱动的自适应门控单元来调整解码层	自建数据集-5691	*mAP* 提升 11.7%,减震器、均压环等平均精度分别提升 8.5% 与 4.4%	极端小目标与密集小目标检测精度仍不足
[88]	集成多尺度主干网络、自注意力上采样模块(SAU)和 IDIoU 损失函数	IFDD-1600 CPLID-848 AI-TOD-28,036	在 IFDD 上平均精度提升 7.47%,在 CPLID 上准确率达 99.2%,在 AI-TOD 上 *mAP* 达 18.13%	模型在不同电网结构和环境下的泛化能力及实时检测效率仍需进一步验证
[89]	双分支主干网络融合、全局与局部增强注意力机制设计、检测头改进	自建数据集-3700	*mAP* 达 93.10%,小目标相关缺陷(如裂纹精度达 92.22%)精度显著提升	模型计算复杂度与参数量增加,实时性略有下降

4.5　小结

本章围绕视觉 Transformer 在表面缺陷检测应用中面临的若干关键挑战——包括样本量稀缺、模型轻量化需求、训练效率低下以及小目标检测等问题,系统梳理了近年来国内外学者在上述 4 个方面的代表性解决策略,旨在为视觉 Transformer 在该领域的进一步应用与研究提供理论支持与方法参考。

5 未来研究展望

1）基于迁移学习的预训练视觉大模型构建

在自然语言处理领域，DeepSeek 等大模型已基本实现了统一架构下的多任务泛化能力。然而，在计算机视觉领域，尤其是缺陷检测场景中，由于缺陷形态多样、背景复杂多变，基于 Transformer 架构来构建统一的视觉大模型仍面临较大挑战。尽管如此，可结合缺陷检测具体应用领域特点，通过构建面向缺陷检测子领域的预训练大模型来实现有效迁移[90]，例如关键场景迁移或关键类别迁移，从而实现面向缺陷检测某个细分领域下的多任务视觉检测大模型算法：(1) 基于视觉 Transformer 技术的关键场景迁移。如针对板材类检测场景，可以结合木地板、瓷砖、大理石等缺陷检测特征与数据，以实现针对板材类产品的基于视觉 Transformer 技术的预训练大模型，从而在实际子场景应用时进行迁移学习，并在具体任务上进行模型微调；(2) 基于视觉 Transformer 技术的关键缺陷类别迁移。如针对裂纹大类缺陷，考虑到多任务或场景下的裂纹形态和特征的相似性，可以针对裂纹大类研究设计裂纹这一关键缺陷类别的检测大模型，并基于该大模型算法就具体检测任务进行迁移或微调应用。

2）CNN + 视觉 Transformer 协同应用趋势

尽管 CNN-Transformer 协同应用通过结合 CNN 的局部特征提取与视觉 Transformer 的全局上下文建模能力，在缺陷检测中展现出了卓越性能，但其规模化成熟应用仍面临严峻挑战。核心问题在于模型复杂度高而导致的计算与部署瓶颈，以及融合模块的位置与方式缺乏理论指导，从而造成检测性能的不确定与泛化性不足等问题。展望未来，该领域的发展一方面应持续聚焦轻量化设计，通过开发高效注意力机制与模型压缩技术以适配工程应用实时检测需求；另一方面，需探索更加成熟的 CNN-Transformer 协同融合架构，如考虑基于神经架构搜索的自动化融合策略[91-92]，以规避人工设计的局限性，从而实现精度与效率的均衡，推动其从学术验证走向多行业工程应用落地。

3）视觉 Transformer 与多模态技术的深度融合

实际上多模态技术并非新兴概念，其思想在学术界已存在多年。然而，近几年来，多模态技术之所以在学术界被广泛关注，主要得益于 Transformer 架构的诞生，Transformer 提供了强大的跨模态注意力交互机制，可以动态地、自适应地学习不同模态信息之间的对应关系和重要性权重，这取代了过去基于简单拼接或加权平均的融合方式。正如 3.2 节与 4.1 节中所阐述的，在缺陷检测领域，多模态技术通过融合红外、激光点云数据以及超声波、热成像等不同传感器数据，解决了许多传统单一视觉难以解决的难题。但目前应用还处于起步阶段，仍面临数据对齐、模型复杂等诸多挑战。在未来工作中，应着力探索视觉 Transformer 与多模态技术的深度融合机制。通过充分发挥视觉 Transformer 的全局建模能力与跨模态注意力的交互机制，可进一步突破复杂应用场景中缺陷检测的现有局限[93-94]，推动该领域向自适应、高精度、轻量化以及可解释方向发展。

4）基于 DETR 的缺陷目标检测算法的发展

尽管 DETR 等基于视觉 Transformer 的目标检测网络，目前仍存在训练收敛速度慢、对大规模标注数据依赖性强、计算成本高且实时性差等局限，从而导致其难以直接应用于实时性要求高、数据稀缺的缺陷检测场景，应用成熟度也因此不及 ViT、Swin Transformer 等分类网络。然而，相比较传统 CNN 目标检测网络，DETR 也同时展现出显著潜力：一方面，它摒弃了传统网络中手工设计的锚框与非极大值抑制后处理步骤，从根本上简化了目标检测的整体流程，减少了人工调参依赖与流程冗余；另一方面，依托强大的自注意力机制，DETR 能够更高效地捕捉图像中缺陷区域的全局关联信息，提升了复杂场景下缺陷检测的准确性与鲁棒性。基于上述针对 DETR 当前存在的挑战与潜力分析，为推动其在表面缺陷检测领域的进一步发展，未来的研究与应用应重点聚焦于以下 3 个方向：(1) 通过模型压缩、知识蒸馏[95]及高效注意力机制[96]等轻量化技术，着力优化其计算效率和收敛性能，以满足在边缘设备上的实时检测需求；(2) 探索与自监督及小样本学习的深度融合[97]，通过在大规模无缺陷图像上预训练以学习通用表征，大幅降低对高成本标注数据的依赖；(3) 加强与多模态数据融合及可解释性分析能力的结合，从而构建更可靠、自适应的高精度目标检测系统。

6 结 论

传统卷积神经网络方法受限于局部感受野的归纳偏置，在缺陷检测任务中存在长程依赖建模能力较弱、复杂背景纹理下缺陷特征提取能力不足等核心瓶颈。而视觉 Transformer 通过其全局自注意力机制，能够有效建模像素之间的全局依赖关系，从而显著增强对复杂背景下缺陷目标的感知与识别性能。本研究以视觉 Transformer 技术为核心，将提高缺陷检测性能、解决该领域痛点作为研究目标，系统梳理了近年来国内外基于视觉 Transformer 的缺陷检测方法的研究进展与应用实践。重点分析了视觉 Transformer 中序列化处理、位置编码、编码器架构以及多头自注意力机制等核心模块的原理，并在此基础上阐述了该技术在缺陷检测领域中的优势与面临的关键挑战。进一步地，本研究一方面围绕复杂纹理背景抑制、多

模态数据协同分析、以及基于模块化思想的 CNN-Transformer 融合模型等典型应用场景,探讨了视觉 Transformer 技术优势驱动下的相关方法应用与发展;另一方面,针对样本量稀缺、模型轻量化需求、训练效率低下、小目标检测等实际应用难题,总结了近年来提出的多层次多维度的解决方案,为视觉 Transformer 在相关实时检测环境中的落地应用提供了理论依据与方法支撑。综上所述,文章不仅解析了视觉 Transformer 的基础机理,突出了其核心优势与典型应用方向,还针对关键挑战提出了应对策略,旨在为基于深度学习的缺陷检测研究提供有价值的参考。

参考文献

[1] 赵朗月, 吴一全. 基于机器视觉的表面缺陷检测方法研究进展[J]. 仪器仪表学报, 2022, 43(1): 198-219.
ZHAO L Y, WU Y Q. Research progress of surface defect detection methods based on machine vision[J]. Chinese Journal of Scientific Instrument, 2022, 43(1): 198-219.

[2] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. ArXiv preprint arXiv: 2010.11929, 2020.

[3] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with Transformers[C]. Computer Vision-ECCV 2020, 2020: 213-229.

[4] 田永林, 王雨桐, 王建功, 等. 视觉 Transformer 研究的关键问题: 现状及展望[J]. 自动化学报, 2022, 48(4): 957-979.
TIAN Y L, WANG Y T, WANG J G, et al. Key problems and progress of vision Transformers: The state of the art and prospects[J]. Acta Automatica Sinica, 2022, 48(4): 957-979.

[5] LI Y, MIAO N P, MA L D, et al. Transformer for object detection: Review and benchmark[J]. Engineering Applications of Artificial Intelligence, 2023, 126: 107021.

[6] 李文生, 张菁, 卓力, 等. 基于 Transformer 的视觉分割技术进展[J]. 计算机学报, 2024, 47(12): 2760-2782.
LI W SH, ZHANG J, ZHUO L, et al. Overview of Transformer-based visual segmentation techniques[J]. Chinese Journal of Computers, 2024, 47(12): 2760-2782.

[7] 石争浩, 李成建, 周亮, 等. Transformer 驱动的图像分类研究进展[J]. 中国图象图形学报, 2023, 28(9): 2661-2692.
SHI ZH H, LI CH J, ZHOU L, et al. Survey on Transformer for image classification[J]. Journal of Image and Graphics, 2023, 28(9): 2661-2692.

[8] 周丽娟,毛嘉宁. 视觉 Transformer 识别任务研究综述[J]. 中国图象图形学报, 2023, 28(10): 2969-3003.
ZHOU L J, MAO J N. Vision Transformer-based recognition tasks: A critical review[J]. Journal of Image and Graphics, 2023, 28(10): 2969-3003.

[9] ISLAM K. Recent advances in vision Transformer: A survey and outlook of recent work[J]. ArXiv preprint arXiv:2203.01536, 2022.

[10] TAKAHASHI S, SAKAGUCHI Y, KOUNU N, et al. Comparison of vision Transformers and convolutional neural networks in medical image analysis: A systematic review[J]. Journal of Medical Systems, 2024, 48(1): 84.

[11] MA M R, HAN L SH, ZHOU CH J, et al. Research and application of transformer based anomaly detection model: A literature review[J]. ArXiv preprint arXiv: 2402.08975, 2024.

[12] 冯夫健, 罗太维, 谭棉, 等. 基于自注意特征融合的钢材表面小目标缺陷检测[J]. 电子测量技术, 2024, 47(19): 172-180.
FENG F J, LUO T W, TAN M, et al. Defect detection of small targets on steel surface based on self-attention feature fusion[J]. Electronic Measurement Technology, 2024, 47(19): 172-180.

[13] 李刚, 张运涛, 汪文凯, 等. 采用 DETR 与先验知识融合的输电线路螺栓缺陷检测方法[J]. 图学学报, 2023, 44(3): 438-447.
LI G, ZHANG Y T, WANG W K, et al. Defect detection method of transmission line bolts based on DETR and prior knowledge fusion[J]. Journal of Graphics, 2023, 44(3): 438-447.

[14] 张乃雪, 钟羽中, 赵涛, 等. 基于 Smooth-DETR 的产品表面小尺寸缺陷检测算法[J]. 计算机应用研究, 2022, 39(8):2520-2525.
ZHANG N X, ZHONG Y ZH, ZHAO T, et al. Detection method for small-size surface defects based on Smooth-DETR[J]. Application Research of Computers, 2022, 39(8): 2520-2525.

[15] 陈俊英, 李朝阳, 黄汉涛, 等. 并行特征提取和渐进特征融合的计算机主板装配缺陷检测[J]. 光学 精密工程, 2024, 32(10): 1622-1637.
CHEN J Y, LI CH Y, HUANG H T, et al. Computer motherboard assembly defect detection using parallel feature extraction and progressive feature fusion[J]. Optics and Precision Engineering, 2024, 32(10): 1622-1637.

[16] 王琦,张涛,徐超炜, 等. 多尺度注意力融合与视觉Transformer方法优化的电阻抗层析成像深度学习方法[J]. 仪器仪表学报, 2024,45(7): 52-63.
WANG Q, ZHANG T, XU CH W, et al. Optimized learning method for electrical impedance tomography with multi-scale attention fusion and vision Transformer[J]. Chinese Journal of Scientific Instrument, 2024, 45(7): 52-63.

[17] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]. 2021 IEEE/CVF International Conference on Computer Vision, 2021: 9992-10002.

[18] WANG W H, XIE EN Z, LI X, et al. Pyramid vision Transformer: A versatile backbone for dense prediction without convolutions[C]. 2021 IEEE/CVF International Conference on Computer Vision, 2021: 548-558.

[19] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image Transformers & distillation through attention[C]. 38th International Conference on Machine Learning, 2021: 10337-10347.

[20] LI J SH, XIA X, LI W, et al. Next-vit: Next generation vision Transformer for efficient deployment in realistic industrial scenarios[J]. ArXiv preprint arXiv: 2207.05501, 2022.

[21] ZHU X ZH, SU W J, LU L W, et al. Deformable detr: Deformable transformers for end-to-end object detection[J]. ArXiv preprint arXiv:2010.04159, 2020.

[22] SRINIVAS A, LIN T Y, PARMAR N, et al. Bottleneck Transformers for visual recognition[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 16514-16524.

[23] ZHAO Y, LYU W Y, XU SH L, et al. Detrs beat yolos on real-time object detection[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 16965-16974.

[24] FANG H, LIN S, HU J W, et al. CPF-DETR: An end-to-end DETR model for detecting complex patterned fabric defects[J]. Fibers and Polymers, 2025, 26(1): 369-382.

[25] GUO J, LI T CH, DU B G. Segmentation head networks with harnessing self-attention and Transformer for insulator surface defect detection[J]. Applied Sciences, 2023, 13(16): 9109.

[26] HUANG J CH, WU Y, ZHOU X F, et al. Multi-scale adaptive prototype Transformer network for few-shot strip steel surface defect segmentation[J]. IEEE Transactions on Instrumentation and Measurement, 2025(74): 5016514.

[27] YU X L, YU Q CH, MU Q Y, et al. MCAW-YOLO: An efficient detection model for ceramic tile surface defects[J]. Applied Sciences, 2023, 13(21): 12057.

[28] LIU CH M, DUAN SH Y, PANG H B. ARDSwin-Unet: A variant swin transformer for fabric defect detection[C]. 2024 IEEE International Symposium on Digital Home, 2024: 13-18.

[29] YAO L, CHEN ZH Q, WAN Y. Research on fabric defect detection technology based on RDN-LTE and improved DINO[C]. Advances in Computer Graphics, 2023: 137-148.

[30] YANG R M, GUO N, TIAN B, et al. Fabric defect detection via saliency model based on adjacent context coordination and Transformer[J]. Journal of Engineered Fibers and Fabrics, 2024, 19: 1-13.

[31] QU H, DI L, LIANG J ZH, et al. U-SMR: U-SwinT & multi-residual network for fabric defect detection[J]. Engineering Applications of Artificial Intelligence, 2023, 126: 107094.

[32] 黄媛媛, 熊文博, 张宏伟, 等. 基于U型Swin Transformer自编码器的色织物缺陷检测[J]. 激光与光电子学进展, 2023, 60(12): 303-310.
HUANG Y Y, XIONG W B, ZHANG H W, et al. Yarn-dyed fabric defect detection based on U-shaped Swin Transformer auto-encoder[J]. Laser & Optoelectronics Progress, 2023, 60(12): 303-310.

[33] WANG Q Q, ZHANG J, DU J M, et al. A fine-tuned multimodal large model for power defect image-text question-answering[J]. Signal, Image and Video Processing, 2024, 18(12): 9191-9203.

[34] FENG B, XIA X, ZHANG L, et al. An image-text multimodal fusion of deep learning for detecting insulator defects[J]. International Journal of Parallel, Emergent and Distributed Systems, 2025: 2469512.

[35] LI N SH, WU R B, LI H F, et al. M2FNet: Multimodal fusion network for airport runway subsurface defect detection using GPR data[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5108816.

[36] LI W L, YANYACHI A, SUN T X, et al. Multimodal characterization of coating defects in graphite electrodes for lithium-ion batteries[J]. Journal of the Electrochemical Society, 2025, 172(8): 080523.

[37] MA M M, REN J, ZHAO L, et al. Are multimodal Transformers robust to missing modality?[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 18156-18165.

[38] CABRAL S, KLIMENKA M, BADEMOSI F, et al. A contactless multi-modal sensing approach for material assessment and recovery in building deconstruction[J]. Sustainability, 2025,17(2): 585.

[39] YAN J J, LIU Y F, SUN J J, et al. Cross modal Transformer: Towards fast and robust 3D object detection[C]. 2023 IEEE/CVF International Conference on Computer Vision, 2023: 18222-18232.

[40] LUO J X, YANG ZH W, CAO Y, et al. RT-DETR-MCDAF: Multimodal fusion of visible light and near-infrared images for citrus surface defect detection in the compound domain[J]. Agriculture, 2025, 15(6): 630.

[41] GE K, WANG C, GUO Y T, et al. Fine-tuning vision foundation model for crack segmentation in civil infrastructures[J]. Construction and Building Materials, 2024, 431: 136573.

[42] VASAN V, VENKATESH N V, VAITHIYANATHAN S, et al. Detection and classification of surface defects on hot-rolled steel using vision Transformers[J]. Heliyou, 2024, 10(19): 38498.

[43] 查健, 陈先中, 王文财, 等. 基于改进的 YOLOv5s 刨花板表面小目标缺陷检测算法[J]. 计算机工程与应用, 2024, 60(17): 158-166.
ZHA J, CHEN X ZH, WANG W C, et al. Small defect detection algorithm of particle board surface based on improved YOLOv5s[J]. Computer Engineering and Applications, 2024, 60(17): 158-166.

[44] LI L J, YANG L, HAO ZH Y, et al. Road sub-surface defect detection based on gprMax forward simulation-sample generation and Swin Transformer-YOLOX[J]. Frontiers of Structural and Civil Engineering, 2024, 18(3): 334-349.

[45] 李季桐, 刘杰, 杨娜, 等. 基于层次化多尺度特征融合的金属缺陷分类模型[J]. 仪器仪表学报, 2025, 46(3): 206-218.
LI J T, LIU J, YANG N, et al. Enhanced hierarchical multi-scale feature fusion model for metal defect classification[J]. Chinese Journal of Scientific Instrument, 2025, 46(3): 206-218.

[46] HE ZH D, YANG W B, LIU Y J, et al. Insulator defect detection based on YOLOv8s-SwinT[J]. Information, 2024, 15(4): 206.

[47] 裴少通, 张行远, 胡晨龙, 等. 基于 ER-YOLO 算法的跨环境输电线路缺陷识别方法[J]. 电工技术学报, 2024, 39(9): 2825-2840.
PEI SH T, ZHANG H Y, HU CH L, et al. The defect detection method for cross-environment power transmission line based on ER-YOLO algorithm[J]. Transactions of China Electrotechnical Society, 2024, 39(9): 2825-2840.

[48] XI Y, ZHOU K, MENG L W, et al. Transmission line insulator defect detection based on Swin Transformer and context[J]. Machine Intelligence Research, 2023, 20(5): 729-740.

[49] TANG B, SONG Z K, SUN W, et al. An end-to-end steel surface defect detection approach via Swin Transformer[J]. IET Image Processing, 2023, 17(5): 1334-1345.

[50] LIU H, CHEN CH, HU R K, et al. CGTD-net: Channel-wise global Transformer-based dual-branch network for industrial strip steel surface defect detection[J]. IEEE Sensors Journal, 2024, 24(4): 4863-4873.

[51] 陈俊生, 陈沂蒙, 刘明杰, 等. 基于 MFES-YOLOV8n 的光伏电池缺陷检测方法[J]. 仪器仪表学报, 2025, 46(6): 251-262.
CHEN J SH, CHEN Y M, LIU M J, et al. A defect detection method for photovoltaic cells based on MFES-YOLOV8n[J]. Chinese Journal of Scientific Instrument, 2025,46(6): 251-262.

[52] 毛坤, 朱学军, 赖惠鸽, 等. 多尺度融合优化的印刷

电路板缺陷检测算法[J]. 激光与光电子学进展, 2025, 62(10): 141-152.
MAO K, ZHU X J, LAI H G, et al. Multi-scale fusion optimization algorithm for printed circuit board defect detection[J]. Laser & Optoelectronics Progress, 2025, 62(10): 141-152.

[53] HAN S Y, JIANG X T, WU ZH Y. An improved YOLOv5 algorithm for wood defect detection based on attention[J]. IEEE Access, 2023, 11: 71800-71810.

[54] LIN G J, LIU K Y, XIA X K, et al. An efficient and intelligent detection method for fabric defects based on improved YOLOv5[J]. Sensors, 2022, 23(1): 97.

[55] 徐忠锴, 刘艳玲, 盛晓娟, 等. 基于改进 YOLOv5 的变电站典型缺陷自动检测算法[J]. 系统仿真学报, 2024, 36(11): 2604-2615.
XU ZH K, LIU Y L, SHENG X J, et al. Automatic detection algorithm for typical defects of substation based on improved YOLOv5[J]. Journal of System Simulation, 2024, 36(11): 2604-2615.

[56] 张林, 谢刚, 谢新林, 等. 融合 MobileNetv3 与 Transformer 的钢板缺陷实时检测算法[J]. 计算机集成制造系统, 2023, 29(12): 3951-3963.
ZHANG L, XIE G, XIE X L, et al. Real-time detection algorithm of steel plate defects integrating MobileNetv3 and Transformer[J]. Computer Integrated Manufacturing Systems, 2023, 29(12): 3951-3963.

[57] WANG Q Y, DONG H B, HUANG H Y. Swin-Transformer-YOLOv5 for lightweight hot-rolled steel strips surface defect detection algorithm[J]. Plos One, 2024, 19(1): 292082.

[58] MIA M S, LI CH B. STD2: Swin Transformer-based defect detector for surface anomaly detection[J]. IEEE Transactions on Instrumentation and Measurement, 2025(74): 5002715.

[59] 陈俊英, 黄汉涛, 李朝阳. 特征增强和度量优化的钢材表面缺陷检测[J]. 激光与光电子学进展, 2024, 61(24): 92-101.
CHEN J Y, HUANG H T, LI CH Y. Feature enhancement and metric optimization for defect detection on steel surface[J]. Laser & Optoelectronics Progress, 2024, 61(24): 92-101.

[60] 于重重, 萨良兵, 马先钦, 等. 基于度量学习的小样本零器件表面缺陷检测[J]. 仪器仪表学报, 2020, 41(7): 214-223.
YU CH CH, SA L B, MA X Q, et al. Few-shot parts surface defect detection based on the metric learning[J]. Chinese Journal of Scientific Instrument, 2020, 41(7): 214-223.

[61] 陈哲煊, 高雪莲, 宋佳宇, 等. 融合双编码与元学习的小样本输电线异物检测[J]. 仪器仪表学报, 2025, 46(3): 193-205.
CHEN ZH X, GAO X L, SONG J Y, et al. Small sample foreign body detection in power lines based on double coding and meta-learning[J]. Chinese Journal of Scientific Instrument, 2025, 46(3): 193-205.

[62] WANG H X, NGUYEN T H, NGUYEN T N, et al. PD-TR: End-to-end plant diseases detection using a Transformer[J]. Computers and Electronics in Agriculture, 2024, 224: 109123.

[63] LIU J K, LONG X Y, JIANG CH, et al. Multi-feature vision Transformer for automatic defect detection and quantification in composites using thermography[J]. NDT & E International, 2024, 143: 103033.

[64] PARK D S, CHAN W, ZHANG Y, et al. Specaugment: A simple data augmentation method for automatic speech recognition[J]. ArXiv preprint arXiv: 1904.08779, 2019.

[65] GENG X Y, LIU H, LEE L, et al. Multimodal masked autoencoders learn transferable representations[J]. ArXiv preprint arXiv:2205.14204, 2022.

[66] HUANG X H, LI Y, BAO Y Q, et al. Sparse cross-Transformer network for surface defect detection[J]. Scientific Reports, 2024, 14(1): 24731.

[67] DENG F Y, HUANG Z H, HAO R J, et al. An inception Transformer-based weighted prototype network for few-shot defect recognition of wheelset bearing[J]. Journal of Computational Design and Engineering, 2025, 12(3): 36-50.

[68] ZHAO L, ZHENG Y, PENG T, et al. Metal surface defect detection based on a Transformer with multi-scale mask feature fusion[J]. Sensors, 2023, 23(23): 9381.

[69] 邓凯丽, 魏伟波, 潘振宽. 改进掩码自编码器的工业缺陷检测方法[J]. 计算机应用, 2024, 44(8): 2595-2603.
DENG K L, WEI W B, PAN ZH K. Industrial defect detection method with improved masked autoencoder[J].

Journal of Computer Applications, 2024, 44(8): 2595-2603.

[70] CHEN ZH W, WONG W K, ZHONG Z F, et al. Effective transfer of pretrained large visual model for fabric defect segmentation via specifc knowledge injection[J]. ArXiv preprint arXiv:2306.16186 2023.

[71] DUANMU A, XUE SH, LI ZH Y, et al. Rep-MobileViT: Texture and color classification of solid wood floors based on a Re-parameterized CNN-Transformer hybrid model[J]. IEEE Access, 2025, 13: 39950-39963.

[72] 章东平,何数技,魏杨悦,等. 基于动态可变形卷积的轻量化道路缺陷检测方法[J/OL]. 计算机辅助设计与图形学学报,1-12[2024-10-29].
ZHANG D P, HE SH J, WEI Y Y, et al. Lightweight road defect detection method based on dynamic deformable convolution[J/OL]. Journal of Computer-Aided Design & Computer Graphics, 1-12[2024-10-29].

[73] 李鹏, 常乐, 覃发富, 等. 基于改进 Transformer 结构的电力绝缘子运动模糊图像复原网络[J]. 电网技术, 2025, 49(6): 2623-2631.
LI P, CHANG L, QIN F F, et al. Motion blurred image restoration of power insulators based on improved Transformer structure[J]. Power System Technology, 2025, 49(6): 2623-2631.

[74] GUO F, LIU J, QIAN Y, et al. Rail surface defect detection using a Transformer-based network[J]. Journal of Industrial Information Integration, 2024, 38: 100584.

[75] ZHOU X, REN ZH H, ZHANG Y CH, et al. A shunted-swin Transformer for surface defect detection in roller bearings[J]. Measurement, 2024, 238: 115283.

[76] DONG K, SHEN Q B, WANG CH Y, et al. Improved Swin Transformer-based defect detection method for transmission line patrol inspection images[J]. Evolutionary Intelligence, 2024, 17(1): 549-558.

[77] LIU X L, RAO ZH Y, ZHANG Y X, et al. UAVs images based real-time insulator defect detection with Transformer deep learning[C]. 2023 IEEE International Conference on Robotics and Biomimetics, 2023: 1-6.

[78] 陈俊英, 李朝阳, 席月芸, 等. ViT 和注意力融合的类别不均衡 PCB 缺陷检测方法[J]. 仪器仪表学报, 2024, 45(4): 294-306.
CHEN J Y, LI CH Y, XI Y Y, et al. ViT and attention fusion for class-imbalanced PCB defect detection[J]. Chinese Journal of Scientific Instrument, 2024, 45(4): 294-306.

[79] GE Y, JIANG D P, SUN L P. Wood veneer defect detection based on multiscale DETR with position encoder net[J]. Sensors, 2023, 23(10): 4837.

[80] SU Z B, SHAO Y L, LI P F, et al. Improved RT-DETR network for high-quality defect detection on digital printing fabric[J]. Journal of Natural Fibers, 2025, 22(1): 2476634.

[81] ZHANG X W, WANG Y L, QIN R F, et al. Improved ViT multimodal anomaly detection for electricity equipment[C]. 2023 IEEE 10th International Forum on Electrical Engineering and Automation, 2023:444-447.

[82] XU H T, LIU CH M, DUAN SH Y, et al. A fabric defect segmentation model based on improved Swin-Unet with gabor filter[J]. Applied Sciences, 2023, 13(20): 11386.

[83] ZHU W, ZHANG H, ZHANG CH, et al. Surface defect detection and classification of steel using an efficient Swin Transformer[J]. Advanced Engineering Informatics, 2023, 57: 102061.

[84] LIU Q L, MENG H, ZHAO R N, et al. Green apple detector based on optimized deformable detection Transformer[J]. Agriculture, 2024, 15(1): 75.

[85] HU Y, JIANG G. Weft-knitted fabric defect classification based on a Swin Transformer deformable convolutional network[J]. Textile Research Journal, 2023, 93(9/10): 2409-2420.

[86] DING ZH G, FU F CH, ZHENG J SH, et al. Intelligent wood inspection approach utilizing enhanced Swin Transformer[J]. IEEE Access, 2024, 12: 16794-16804.

[87] LI W T, TONG Q Q, GU J Q, et al. A self-adjusting Transformer network for detecting transmission line defects[J]. Neural Computing and Applications, 2024, 36(9): 4467-4484.

[88] LI D, YANG P F, ZOU Y T. Optimizing insulator defect detection with improved DETR models[J]. Mathematics, 2024, 12(10): 1507.

[89] 苏俊, 唐潮龙, 刘智权, 等. 基于全局与局部特征提取增强的光伏板缺陷检测算法[J]. 计算机工程与应用, 2025, 61(12): 299-310.

SU J, TANG CH L, LIU ZH Q, et al. Photovoltaic panel defect detection algorithm enhanced by global and local feature extraction[J]. Computer Engineering and Applications, 2025, 61(12): 299-310.

[90] JEONG M, YANG M, JEONG J. Hybrid-DC: A hybrid framework using ResNet-50 and vision Transformer for steel surface defect classification in the rolling process[J]. Electronics, 2024, 13(22): 4467.

[91] YAN C X, CHANG X J, LI ZH H, et al. Masked distillation advances self-supervised Transformer architecture search[C]. The Twelfth International Conference on Learning Representations, 2024.

[92] HO S T, VO T V, EBRAHIMKHANI S, et al. Vision Transformer neural architecture search for out-of-distribution generalization: Benchmark and insights[J]. ArXiv preprint arXiv:2501.03782, 2025.

[93] 金宇锋, 陶重犇. 基于 Transformer 的融合信息增强 3D 目标检测算法[J]. 仪器仪表学报, 2023, 44(12): 297-306.
JIN Y F, TAO CH B. Fusion information enhanced method based on Transformer for 3D object detection[J]. Chinese Journal of Scientific Instrument, 2023, 44(12): 297-306.

[94] 李学钊, 王伟, 薛冰. 基于梯度算子和注意力的多模态融合目标检测[J]. 仪器仪表学报, 2024, 45(11): 224-232.
LI X ZH, WANG W, XUE B. Multi-modal fusion object detection based on gradient operator and attention[J]. Chinese Journal of Scientific Instrument, 2024, 45(11): 224-232.

[95] WANG Y, LI X, WENG SH ZH, et al. KD-DETR: Knowledge distillation for detection Transformer with consistent distillation points sampling[C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 16016-16025.

[96] CHEN Q, SU X B, ZHANG X Y, et al. LW-DETR: A Transformer replacement to yolo for real-time detection[J]. ArXiv preprint arXiv:2406.03459, 2024.

[97] SHANGGUAN Z Y, HUAI L, LIU T, et al. Decoupled DETR for few-shot object detection[C]. Computer Vision-ACCV 2024, 2024: 158-174.

作者简介

杨洋,2015 年于江苏科技大学获得硕士学位、现为南京航空航天大学博士研究生,主要研究方向为图像处理与机器视觉。
E-mail:1716481354@qq.com

Yang Yang received his M. Sc. degree from Jiangsu University of Science and Technology in 2015. He is currently a Ph. D. candidate at Nanjing University of Aeronautics and Astronautics. His main research interests include image processing and machine vision.

吴一全(通信作者),1998 年于南京航空航天大学获得博士学位,现为南京航空航天大学教授、博士生导师,主要研究方向为视觉检测与图像测量、遥感图像处理与理解、视频处理与智能分析。
E-mail:nuaaimage@163.com

Wu Yiquan (Corresponding author) received his Ph. D. degree from Nanjing University of Aeronautics and Astronautics in 1998. He is currently a professor and Ph. D. advisor at Nanjing University of Aeronautics and Astronautics. His main research interests include visual detection and image measurement, remote sensing image processing and understanding, video processing and intelligent analysis.