小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2024-0601
2026 年 2 月 第 2 期
Vol.47 No.2 2026

基于毫米波感知的皮革瑕疵分类方法

张 健,关灏文

(武汉大学 计算机学院,武汉 430072)
E-mail:guanhaowen@whu.edu.cn

摘 要: 皮革瑕疵分类是确保皮革产品质量的关键环节.传统的人工检测和图像处理方法受限于光照等环境因素,难以满足高效检测需求.近年来,深度学习特别是卷积神经网络(CNN)的应用提高了瑕疵检测的准确性和效率,但仍受到环境影响.毫米波雷达技术作为一种新兴的无损检测方法,因其强穿透性和不受光照等因素影响的特性而逐渐受到关注.文中提出了一种结合毫米波雷达与改进 Vision Transformer 模型的皮革瑕疵分类方法,利用毫米波雷达信号提取皮革瑕疵的时频特征,并通过深度学习模型进行分类,在自建数据集上达到了 95.62% 的准确率,相比经典的分类模型优势显著.

关 键 词: 毫米波雷达;皮革瑕疵分类;Vision Transformer 模型;迁移学习
中图分类号: TP393 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)02-0257-08

Leather Defect Classification Method Based on Millimeter-wave Sensing

ZHANG Jian, GUAN Haowen
(School of Computer Science, Wuhan University, Wuhan 430072, China)

Abstract: Leather defect classification plays a crucial role in ensuring the quality of leather products. Traditional methods, such as manual inspection and image processing, are often hindered by environmental factors like lighting, making it challenging to meet the demands for efficient detection. In recent years, the application of deep learning, particularly Convolutional Neural Network (CNN), has enhanced the accuracy and efficiency of defect detection, although environmental conditions still pose limitations. Millimeter-wave radar technology, as an emerging non-destructive testing method, has gained increasing attention due to its strong penetration capability and resistance to environmental factors such as lighting. This paper introduces a novel leather defect classification approach that integrates millimeter-wave radar with an improved Vision Transformer model. By leveraging millimeter-wave radar signals to extract the time-frequency features of leather defects and employing a deep learning model for classification, this method achieves an accuracy of 95.62% on a self-constructed dataset, showcasing significant advantages over traditional classification models.

Keywords: millimeter-wave radar; leather defect classification; Vision Transformer model; transfer learning

0 引 言

皮革,作为一种在服装、汽车内饰、家具制造等众多高端领域内占据举足轻重地位的材料,其外观品质无疑直接关乎最终产品的市场估价以及在白热化市场竞争中的核心竞争力.鉴于此关键性,瑕疵检测作为确保皮革产品质量的一道关键防线,其重要性愈发显著.而在皮革瑕疵检测的精密过程中,瑕疵分类扮演着举足轻重的角色.皮革瑕疵分类有助于构建一套标准化的质量检测体系,使得瑕疵能够得以规范化地识别与应对,进而大幅提升生产效率,并有效遏制不合格品的产生.通过细致的瑕疵分类,检测人员能够更为迅速地识别并处理各类缺陷,确保每一块皮革均能满足既定的质量标准.这一分类系统还赋予了生产商依据不同瑕疵类型采取恰当处理措施(如修复或降级)的能力,从而最大化资源利用效率.此外,瑕疵分类在资源分配方面亦发挥着重要作用,它使得生产商能够依据瑕疵等级对产品进行合理定价与市场定位,进而提升整体经济效益.同时,该分类系统还为企业制定科学的质量标准和保证提供了有力依据,确保所生产的产品能够精准满足消费者的多元化需求.更为关键的是,通过减少浪费、提高皮革原材料的使用效率,这一分类系统有力支持了企业的可持续发展目标.总而言之,皮革瑕疵分类不仅极大地优化了质量控制流程,更在促进资源最大化利用与提升市场竞争力方面发挥了积极作用.

传统上,瑕疵分类检测主要依赖于人工视觉检查或基础的图像处理技术.然而,这些方法不仅效率低下、劳动强度大,而且极易受到光照条件、天气状况等环境因素的干扰,从而难以满足现代生产环境对于高效生产和精确检测所提出的严格要求.

随着深度学习技术的发展,计算机视觉技术在皮革瑕疵分类检测领域的应用愈发广泛而深入.深度学习技术,尤其是卷积神经网络(Convolutional Neural Network,简称 CNN)等模型,能够从海量的图像数据中自动挖掘并学习特征,显著提升了皮革瑕疵分类检测的精确度和效率.目前,众多研究在深度学习视觉方法的应用探索上已取得了瞩目的成果.例如,一

收稿日期:2024-12-20 收修改稿日期:2025-02-27 基金项目:国家重点研发计划项目(2023YFB3106900)资助. 作者简介:张 健,男,1976 年生,博士,教授,博士生导师,研究方向为智能物联网、机器人操作系统;关灏文,男,2001 年生,硕士研究生,研究方向为智能感知.

些研究者巧妙地引入了迁移学习技术,通过对预训练模型在小规模数据集上进行精细调整,有效缓解了大规模标注数据缺失的难题. 此外,借助增强学习、生成对抗网络(Generative Adversarial Network,简称 GAN)等前沿技术,研究人员成功生成了高质量的瑕疵样本,为分类模型的进一步训练与验证提供了有力支持.

然而,计算机视觉技术在实际应用中仍面临一定的挑战,尤其是光照条件对其分类和检测效果具有显著影响. 在光照不均匀或存在强烈反射的环境中,分类性能可能会出现下滑. 同时,当背景复杂或皮革表面存在干扰因素时,也可能导致分类准确性的降低. 因此,如何在复杂多变的环境条件下保持检测的稳定性和准确性,仍是未来研究中亟待解决的关键问题.

近年来,毫米波雷达技术作为一种新兴的无损检测技术,正日益受到学术界与工业界的广泛关注. 该技术凭借其卓越的穿透性能及对多种材质的广泛适应性,能够在不破坏被测物体结构的前提下,精准地获取材料内部的详细信息,进而有效识别出材料内部的瑕疵. 相较于传统的光学检测技术,毫米波雷达在检测流程中展现出了显著的优越性,其检测过程不受光照条件、颜色差异及表面纹理等因素的干扰,能够在包括黑暗与强光在内的各种复杂环境中保持稳定的运行状态.

目前,毫米波雷达技术在皮革瑕疵分类和检测领域缺乏成熟的应用经验,且缺失相对应的雷达数据集. 为了填补此项技术的空白,本文创新性地引入了毫米波雷达技术,并设计了一种新型的、能有效克服光线干扰的瑕疵分类方法. 针对当前技术背景下皮革瑕疵雷达数据集匮乏的问题,本文利用商用毫米波雷达设备,成功构建了皮革瑕疵数据集,并运用了随机掩蔽策略,实现了对该数据集的有效增强. 鉴于传统分类模型在处理复杂瑕疵时所面临的局限性,本文采纳经过迁移学习和随机掩蔽策略改进的 Vision Transformer(简称 VIT)模型,以实现对瑕疵特征的高效分类处理.

综上所述,本文的主要贡献可归纳如下:

1)本文开创性地引入了毫米波雷达技术,以应用于皮革瑕疵的检测,并设计了一个能够有效抵御光线干扰的毫米波检测系统.

2)本文设计了一种依托于 Vision Transformer 模型的分类检测算法,并借助迁移学习技术和随机掩蔽策略对其进行了优化与提升. 在皮革瑕疵的分类任务中,该算法展现出了尤为显著的有效性,其性能表现令人瞩目.

3)本文利用商用现成的雷达设备实现了所提出的方法,并在自建的皮革瑕疵数据集上进行了充分的实验评估. 实验结果表明,本文所提出的方法在皮革瑕疵分类任务中表现出了较为出色的性能.

1 相关工作

1.1 光学方法

在皮革瑕疵分类领域,传统方法主要依托计算机视觉技术. 这类技术普遍涵盖基于特征的检测与分类算法,诸如边缘检测、纹理分析及颜色分割等[1]. 尽管这些传统方法在特定应用场景下能够有效识别瑕疵,但其性能却易受光照条件变化和背景噪声的干扰.

近年来,深度学习技术的蓬勃发展促使基于深度学习模型的光学成像方法成为研究的新热点. Liong 等人[2]巧妙地结合了预训练的 AlexNet 与支持向量机(Support Vector Machine,简称 SVM)技术,成功实现了对皮革瑕疵的三分类任务,具体涵盖了黑纹、皱纹以及无瑕疵三类,其最优分类性能卓越,达到了 94.67% 的准确率. 然而,遗憾的是,该研究受限于训练数据的规模,仅采用了 250 个缺陷样本与 125 个无缺陷样本,这无疑对模型的全面训练构成了一定挑战. 在 Liong 等人[2]工作的基础上,Gan 等人[3]进一步探索,巧妙地引入了生成对抗网络(GAN)来合成图像,以此策略性地扩充了原本有限的训练数据集. 通过这一创新方法,结合 AlexNet 与 SVM 进行三分类,最终实现了令人瞩目的 100% 准确率. 此外,Deng 等人[4]依托经过改进的 ResNet50 模型,针对划痕、烂面、空洞及针孔这 4 种皮革缺陷进行了深入且精确的分类处理. 其分类性能同样出色,平均分类准确率高达 94.6%.

然而,光学成像技术主要局限于皮革表面瑕疵的检测范畴,如裂纹、色差等显而易见的问题,而对于深层次的瑕疵或是皮革内部结构中的缺陷,其识别能力则显得不足. 此外,光学成像技术对于光照条件有着较高的敏感性,光源分布的不均匀性、反射光的干扰以及阴影效应等因素,均有可能削弱成像的质量,进而对瑕疵检测的精确度构成不利影响. 而针对具有高光泽度或反射性能较强的皮革材料,过度的光线反射往往会导致图像变得模糊不清,从而使得瑕疵的准确辨识变得困难重重. 该技术还极易受到皮革颜色以及外部环境的干扰,皮革自身的不同颜色以及环境中的灰尘、污渍、水汽等杂质有可能干扰光学传感器的正常运作,最终引发误报或漏报的现象.

1.2 传感器方法

此外,一些研究人员还致力于探索基于传感器的解决方案. Chen 等人[5]基于高光谱传感器,精心构建了 3 种结构模型用于分别处理不同的皮革瑕疵,分别是一维卷积神经网络(CNN)、二维 U-Net 以及三维 U-Net,旨在应用于蓝湿皮的瑕疵检测领域,并巧妙地运用高光谱目标检测技术有效抑制了背景干扰,成功研发出了一套针对蓝湿皮瑕疵(涵盖品牌缺失、表面破损、裂纹、蜱虫叮咬痕迹及抓痕等瑕疵)的检测方法. 其中,一维 CNN 擅长处理光谱信息,用于检测蜱咬痕迹. 二维 U-Net 模型则凭借其纳入空间信息的优势,用于处理品牌缺失问题. 而三维 U-Net 模型,通过综合考虑空间与光谱信息,用于针对烂面、裂纹以及抓痕等缺陷的检测. 然而,该工作仅限于运用高光谱传感技术与特定的深度学习模型,针对特定种类的瑕疵进行检测与分析,尚未实际拓展至多种瑕疵的分类应用中. Xu 等人[6]则利用触觉传感器,并依托 DeepLabv3 + 分类模型,创新性地融合了视觉特征与触觉特征,从而开发出一种新型皮革瑕疵检测与分类方法. 相较于单纯依赖光学手段,该方法的分类性能实现了显著提升,具体而言,其准确率由 59.6% 跃升至 69.5%,增幅达到 10%. Liu 等人[7]采用超声波传感技术,有效地揭示了皮革在光学检查中难以发现的瑕疵. Chudzik 等人[8]借助特殊的灯产生热激发脉冲,结合红外温度传感器检测天然鞣制生皮内部和外部的瑕疵. 值得注意的是,Liu 等人[7]与 Chudzik 等人[8]的研究目的并非在于证明其算法相较于传统处理算法具有优越性,而

是旨在探讨超声波传感技术或红外温度传感技术在皮革质量检测领域实现无损检测技术的工业应用潜力,其工作主要聚焦于对通过超声波传感或红外温度传感所提取的皮革瑕疵特征图像进行直观的视觉分析.

高光谱成像、触觉感知以及超声波传感方法均存在各自的局限性.高光谱成像技术主要依赖于表面反射光谱信息,因此易受光照条件变化的影响,且其数据处理流程复杂,穿透深度有限,难以有效探测皮革内部的瑕疵.触觉感知方法则需与皮革表面直接接触,这一特性限制了其应用场景,且容易受到表面污染物的影响,导致适应性不强.超声波传感方法虽然具备穿透皮革检测内部缺陷的能力,但其分辨率相对较低,且同样需要接触皮革表面,信号的反射易受表面形态的影响,故而检测精度有限.

1.3　毫米波雷达

毫米波雷达作为一种非接触式的无损检测技术,能够穿透大多数非金属材料,从而实现对微小瑕疵和内部结构的检测.同时,其高频特性使得毫米波雷达具备较高的分辨率,能够精确捕捉细微的变化.此外,毫米波雷达不依赖于光照条件,能够在光照变化大或环境干扰严重的环境下(如隧道、雾霾、雨天等),在不同颜色不同背景的场景下仍能有效运行,而光学成像方法的表现则难以达到理想的效果.相比于光学成像方法和其他传感技术,毫米波雷达在检测速度、精度和适应性方面具有明显优势.近年来,毫米波雷达在材料检测的应用受到广泛关注.本文的研究工作旨在首次探索并验证利用毫米波雷达技术进行皮革瑕疵检测的可行性,实现对各类皮革瑕疵的区分与检测,包括但不限于划痕、洞孔、折痕、针孔等常见瑕疵类型.

2　系统流程

本文旨在实现利用毫米波雷达技术对皮革瑕疵进行分类的目标.图 1 全面展示了该系统的整体架构:首先,系统通过毫米波雷达进行数据采集;随后,这些数据经由短时傅里叶变换(Short-Time Fourier Transform,简称 STFT)转换为时频谱图;最终,时频谱图被输入至改进的 Vision Transformer 模型中,以完成训练与分类任务.

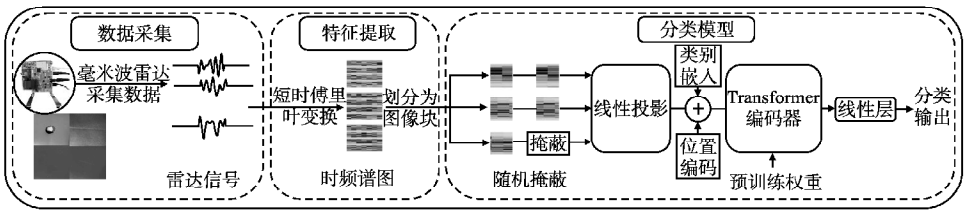

图 1　系统架构

Fig. 1　System architecture

2.1　数据采集和处理

2.1.1　*皮革数据采集*毫米波雷达技术是一种利用电磁波进行探测与成像的先进技术,其操作频段界定于 30 吉赫兹(GHz)至 300GHz 的毫米波频谱区间内.该技术机制涉及通过精密设计的发射器释放毫米波信号,这些信号在遭遇不同材质、形状及位置的物体时,会展现出独特的反射特性.随后,高灵敏度的接收器捕获这些经物体反射回来的毫米波信号,并运用复杂的信号处理算法对这些信号进行深入分析,旨在精确提取并解析出有关物体的各种信息,包括但不限于位置、速度及形状等[9].

毫米波雷达所发射的电磁波,在遇到皮革的不同部分时,会引发不同的反射与散射效应,特别是在瑕疵区域,这种效应表现得尤为显著.这些瑕疵因对皮革表层及内部结构造成各异的影响,从而使得反射信号在时间与频率维度上展现出相应的变化.皮革瑕疵所激发的反射信号,其本质通常呈现为非平稳状态,即其频谱特性会随时间变化而有所改变.尽管传统的纯频域分析技术(例如傅里叶变换),具备解析信号频率特性的能力,却难以揭示频率随时间变化的动态细节.与之相比,时频分析法则通过同步考察时间与频率两个维度,能够更为精准地捕捉非平稳信号的特征,能够深入挖掘更多富有价值的信息细节.例如,它能够揭示瑕疵所引发的反射波在时间维度上的滞后现象,以及瑕疵区域对毫米波频率响应所展现出的独特性质.在皮革中,尤其是那些微小的折痕、针孔等瑕疵,往往会诱发短暂且瞬时的反射信号,这些信号的频率特性在特定时刻可能会发生显著变化.而较大的洞孔、胶痕等瑕疵,则可能产生具有不同频率特性的持续信号.通过运用短时傅里叶变换(STFT)或小波变换等时频分析方法所生成的时频谱图,能够高效捕捉这些短暂且关键的事件,使得雷达系统能够实现对皮革瑕疵的精确检测.

本文采用毫米波雷达对皮革瑕疵样本进行数据采集.具体而言,雷达向皮革样本发射调频连续波(Frequency Modulated Continuous Wave,简称 FMCW)信号,并接收从皮革表面反射回来的响应信号.该雷达系统配备了 4 个信号接收器,每两个接收器之间的距离约为 3 毫米.在一次完整的信号发射与接收周期(即一个 chirp)中,每个接收器所收集的数据被视为一个原始数据样本.这些原始数据样本随后被送入数据处理模块,通过 STFT 技术进行时频特征的提取.

2.1.2　*时频特征提取*

短时傅里叶变换(STFT)是一种有效的分析工具,能够同时揭示信号在时间与频率两个维度上的详尽信息,因而被广泛应用于信号处理领域[10].鉴于此,本文采纳 STFT 技术来提取皮革数据中的时频特征.

汉宁窗函数具有较低的旁瓣泄漏的优势,能够显著抑制非必要频率成分对主频成分的干扰效应;同时,它巧妙地实现了频率分辨率与时间分辨率之间的优化平衡,相较于矩形窗等其他类型的窗函数,更能清晰地展现频谱特征,有效规避了频谱模糊现象.此外,汉宁窗函数在边缘部分的平滑过渡设计,极大地减弱了边界效应,进而减少了虚假频谱成分的产

生. 基于汉宁窗的上述显著优势,本文在短时傅里叶变换中择取了汉宁窗作为窗函数.

在特征提取流程中,本文首先采用汉宁窗函数对原始信号进行分段处理,将其切割成多个重叠的短时段. 随后,针对每个窗口内的信号片段,独立执行傅里叶变换,旨在获取该特定时段内的频谱细节. 紧接着,将这些时段对应的频谱结果按时间顺序逐一拼接,从而构建出一幅直观的时频谱图. 生成的时频谱图如图 2 所示,此图表的横轴代表时间轴,纵轴则对应频率轴,而图像中颜色的深浅则直观地映射出特定时间与频率点上信号的幅度或功率强度. 最终,这幅构建的时频谱图被选定为深度学习模型的输入特征集,用于后续的模型训练与预测. 其具体的数学表达式参见公式(1):

$$STFT\{x(t)\}(\tau,\Omega) = \int_{-\infty}^{\infty} x(t)\Omega(t-\tau)e^{-j\lambda t}dt \quad (1)$$

其中,$x(t)$是输入的信号;$\Omega(t-\tau)$是窗函数,通常是一个具有有限时长的函数,用于局部化信号的某一时刻 t 周围的信号(如汉宁窗、矩形窗等),本文选用的是汉宁窗;$e^{-j\lambda t}$是复指数,表示傅里叶变换中的频率成分;λ 是角频率,$\lambda = 2\pi f$,其中 f 是频率;t 是时刻,表示在该时刻计算傅里叶变换.

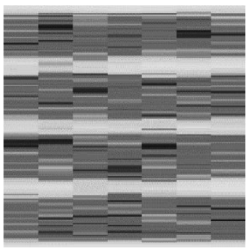

图 2 时频谱图

Fig. 2 Time-frequency spectrogram

2.2 分类模型

2.2.1 基础模型架构

图 3 展示了本文所采用的基础 Vision Transformer(VIT)模型架构[11]. 初始阶段,尺寸为 224 × 224 的图像数据被用作输入. 然而,若直接将此类图像展平转换为一维数据,其维度将达到50176(224 × 224),这一庞大的数据量对后续

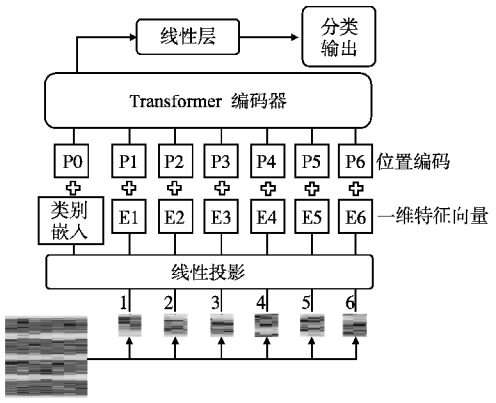

图 3 基础的 VIT 模型架构

Fig. 3 Architecture of base VIT model

在 Transformer 编码器中的处理构成了不便,故而采取了一种分割策略:将输入的图像切割为 14 × 14 个较小的图像块,每个图像块的大小为 16 × 16 像素. 随后,VIT 模型应用了一个线性投影层,该层的作用是将每个具有 3 个颜色通道的 16 × 16 像素图像块映射成长度为 768 的一维特征向量. 这一映射过程基于图像块的原始像素值(16 × 16 × 3),最终生成了 196 个(对应于 14 × 14 的图像块分割)这样的特征向量. 据此,生成了一个特征向量矩阵,其维度为 196 × 768.

此外,类似于 BERT 模型中的类别嵌入机制,模型在序列的起始位置引入了一个可训练的类别标记向量[12],旨在表征全局(即整个图像)的类别信息. 该向量的维度设定为 1 × 768,且初始值遵循标准正态分布的规律进行设定. 历经一系列 Transformer 编码器的处理后,此类别标记将包含整幅图像的信息,并最终应用于分类任务. 将上述类别标记与特征向量进行整合后,形成了一个新的矩阵,其维度为 197 × 768. 鉴于自注意力机制对于输入扰动的不变性,原始的 Transformer 架构难以捕捉输入序列的位序信息和空间结构信息. 因此,为了记忆输入的语序信息以及图像块间的位置关联,模型在 197 个一维特征向量中,各自融入了一个尺寸相同(1 × 768)的可训练位置编码,以协助模型精准捕捉图片的空间结构特征. 模型采用了与原始 Transformer 架构一致的一维位置编码方案[13]. 具体而言,位置编码向量是通过一种基于正弦和余弦的连续函数来生成的. 位置编码的维度与输入特征向量的维度保持一致,均为 768 维. 位置编码的计算参见公式(2)和公式(3):

$$P(k,2i) = sin\left(\frac{k}{n^{2i/d}}\right) \quad (2)$$

$$P(k,2i+1) = cos\left(\frac{k}{n^{2i/d}}\right) \quad (3)$$

其中 k 代表特征向量所对应的图像块的位置索引,其取值范围为 0 ~ 196(第 0 块特指类别标记);d 则为位置编码嵌入空间的维度,具体数值为 768 维;$P(k,j)$ 是一个位置函数,用于将输入序列中第 k 图像块的第 j 维度元素映射至位置矩阵 (k,j) 处;n 是一个用户定义的标量参数,默认设置为 10000[14];而 i 用于分别映射至奇偶列索引,其取值范围为 $0 \sim d/2$.

在编码过程中,奇偶索引分别采用正弦与余弦函数进行处理,旨在使不同维度的编码呈现出各异的周期变化特性. 这一设计有助于模型捕获并理解不同层级的位置信息. 随后,将所得位置编码与相应图像块的特征向量进行逐元素相加,从而生成每个图像块的最终特征向量,其维度为 197 × 768,这些向量将用于后续的模型训练过程. 位置编码可被视作增强图像块表征空间信息的一种手段,它赋予了模型学习并理解各个图像块在原始图像中相对位置的能力. Transformer 通过自注意力机制[14]自动地处理各类补丁嵌入. 在经过一系列 Transformer 的编码后,类别标记会包含整个图像的信息,最终应用于分类任务.

经过上述处理后生成的序列被送入 Transformer 模块中. 鉴于本模型专为分类任务而设计,因此本文仅采用了 Transformer 的编码器部分. 具体而言,本文使用原始的、未经过修改的 Transformer 编码器架构[15],编码器的嵌入维度为 768,有 12 层和 12 个头,编码器的结构如图 4 所示. 使用原始的架构优势在于:标准的模型架构的易于快速实现,且便于应用迁

移学习技术.

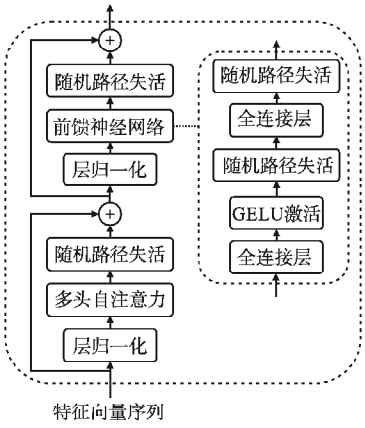

图 4　Transformer 编码器的结构

Fig. 4　Structure of the Transformer encoder

Transformer 编码器的每一层均分为多头注意力机制与前馈神经网络两部分,且在每一层的多头自注意力机制与前馈神经网络处理之前,都会预先实施层归一化技术. 层归一化作为一种专为神经网络设计的归一化手段,对于稳固训练流程、规避梯度消失及梯度爆炸等问题具有显著效用. 相较于传统的批量归一化技术,后者是在每个训练批次的样本集合上执行归一化操作,即依据批次维度施行归一化;层归一化则是在单个样本的特征维度上开展归一化工作,即依据特征维度进行归一化处理,故而不受批次规模变化的影响. 具体而言,针对某一层的输入数据,层归一化技术会逐一计算每个样本在特征维度上的均值与方差,并据此对样本实施归一化操作.

自注意力机制是一种专为序列数据设计的注意力机制,旨在构建序列中不同位置间的依赖关系模型. 此机制通过计算各位置与其他所有位置间的注意力权重,学习彼此间的关联性,进而生成蕴含上下文感知能力的表征. 而多头注意力机制则通过并行运用多个注意力头,进一步强化了自注意力的表征效能. 每个注意力头均配备独立的权重矩阵(涵盖查询、键和值),并生成相应的注意力权重矩阵. 经由对多个注意力头输出的加权求和,即可获得最终的多头注意力表征. 具体而言,多头注意力的计算流程始于输入包含查询(Q)、键(K)和值(V)的特征矩阵. 针对每个注意力头,首先计算注意力权重,即利用查询矩阵 Q 与键矩阵 K 求得注意力分数. 随后,对注意力分数进行缩放处理并施以 softmax 运算,以获得注意力权重. 再依据这些权重对值矩阵 V 进行加权求和,得出每个注意力头的注意力输出. 对于所有注意力头的输出,再次进行加权求和,即可得到多头注意力的最终表征.

此后,多头注意力的输出将输入至一个前馈神经网络中,其中每个表征均独立经过相同的前馈网络处理. 该前馈网络由两个全连接层构成:首个全连接层的维度由 768 扩展至 3072,并配备 GELU(Gaussian Error Linear Units)激活函数;第 2 个全连接层则将其映射回输入维度(768). 此过程旨在针对每个表征执行非线性变换,从而增强模型的表达能力.

在每个子层(即自注意力层与前馈网络层)之间,还会应用残差连接. 残差连接通过将每个子层的输入与输出相加,有助于网络在深层结构中维持信息的顺畅流动,有效避免梯度消失问题,并加速训练进程.

最终,自 Transformer 编码器的输出中提取出类别标记向量,并借助一个线性层完成分类任务.

2.2.2　迁移学习

相较于卷积神经网络(CNN),Transformer 通常需要更庞大的数据集进行训练,方能展现出超越 CNN 的性能. 然而,在实际应用中,如此大规模的数据集往往难以获取. 但考虑到图像与时频谱图在格式上存在相似性,众多的常用架构(包括 Vision Transformer)已备有现成的 ImageNet 训练模型,这些模型可直接应用于迁移学习,且已有研究成功实现了从图像分类任务到时频谱图分类任务的迁移学习[16]. 本文遵循这一原理,对现成的预训练 ViT 模型进行了适配,以融入所构建的模型中. 本文所采用的数据集为尺寸 224 × 224 像素的时频谱图. 尽管时频谱图与用于预训练 ViT 的图片数据在结构上具有一定的相似性,但二者之间并非完全吻合. 因此,在适应过程中,进行了一系列必要的调整.

首先,ViT 的预训练图片输入为三通道图像,而本文模型的输入为单通道时频谱图. 为解决这一问题,采取了将 ViT 补丁嵌入层的 3 个输入通道对应的权重进行平均化的策略,将单通道时频谱图平均分配为具有相同内容的 3 个通道,此举在保持信息完整性的同时,也显著提升了计算效率. 此外,还对输入的时频谱图进行了归一化处理,使数据集的平均值和标准偏差分别调整至 0 和 0.5. 其次,鉴于分类任务的本质差异,舍弃了基础 ViT 模型的最后一个分类层,并重新初始化了一个配备 softmax 激活函数的线性层,作为新的分类层.

本文采纳了在 ImageNet 数据集上预先训练的 VIT 模型. 具体而言,本文采纳了“google/vit-base-patch16-224-in21k”的预训练权重,该架构是在一个包含 21000 个类别的大型数据集(即 ImageNet-21k)上进行预训练的,其类别范围相较于标准 ImageNet(包含 1000 个类别)更为广泛,从而有助于模型学习到更为通用的视觉表示. 本文将该架构在大数据集上训练得到的 Transformer 权重作为本文模型中 Transformer 编码器的初始权重,以此加速模型收敛过程,并有效避免因数据集规模过小而引发的过拟合问题.

2.2.3　随机掩蔽

为了进一步提升模型的泛化能力,本文引入了随机掩蔽技术作为数据增强手段[17]. 在本文所采用的 VIT 模型中,训练时输入的时频谱图被均匀分割为 196 个 16 × 16 的图像块,这一设计为实施随机掩蔽提供了极大的便利. 具体而言,本文通过随机选取部分图像块,并使用零值对其进行填充覆盖,从而实现随机掩蔽. 本文在训练时选择屏蔽频谱图补丁而不是屏蔽整个时间帧,使模型仍然可以同时学习数据的时间和频率结构. 本文设定了 15% 的掩蔽率,即每个图像块有 15% 的概率被屏蔽. 经过随机掩蔽增强处理后的数据将继续被送入模型的后续部分进行训练.

3　实验与结果

3.1　数据集和实验设置

本文实验采用 TI 公司制造的 IWR1443BOOST 毫米波雷达评估板与 DCA1000 雷达数据采集板进行数据的采集工作.

该雷达模块能够发射 77 ~ 81 GHz 频段内的线性调频脉冲波，并具备 4 GHz 的宽带特性. 在实验设置中，雷达被配置为自顶向下的数据采集模式，皮革样本则水平放置于雷达正下方，并保持恒定的 0.5 米距离. 皮革样本选用人造磨砂麂皮. 皮革瑕疵样本如图 5 所示，共涵盖 6 类瑕疵，具体包括划痕、胶痕、折痕、针孔、洞孔以及正常(无瑕疵)样本. 雷达系统内置 4 个信号接收器，每个接收器在一次 chirp 信号周期内所收集的数据被视为一个独立的原始数据样本. 针对每个皮革样本，实验通过雷达获取了 10 份数据，总计包含 6 类瑕疵样本，每类瑕疵样本均采集了 400 份数据，因此，整个实验过程共收集了 2400 份数据样本. 在数据集划分方面，20% 的数据被指定为测试集，而剩余的 80% 则用作训练集. 为确保数据划分的合理性，同一皮革样本的 10 份数据被严格归入同一集合中，以此来防止训练模型在训练过程中接触到测试集数据，进而保证模型的泛化性能不受影响.

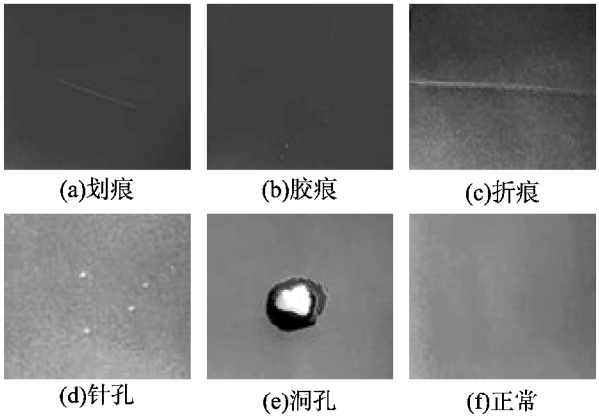

图 5 皮革瑕疵样本

Fig. 5 Leather defect samples

在模型实现方面，本文采用 PyTorch 框架，并选用“google/vit-base-patch16-224-in21k”作为 ImageNet 预训练模型. 损失函数选用二元交叉熵，优化器则选用 Adam，batch_size 设置为 16，训练轮数定为 30 轮. 为防止过拟合，本文在 Adam 优化器中加入了 L2 正则化参数 weight_decay，其值为 1×10^{-5}. 在训练过程中使用了动态学习率，初始学习率为 1×10^{-5}，训练 10 轮后，每隔 5 轮学习率减半. 评估指标则选用准确率(Acc)、精度(Precision)、召回率(Recall)和 F1 分数(F1-score)

3.2 对比实验

3.2.1 与经典模型的对比

为验证本文模型的性能，本小节选取了经典的深度学习分类模型进行性能对比. 以下是所选取的几种经典分类模型：1) ResNet50. 实验采用的是在 ImageNet 数据集上预先训练的 ResNet50 模型[18]. 此模型架构由 49 个卷积层及 1 个全连接层构成，这些层级被组织成多个残差块，并通过残差连接实现块间的相互关联；2) DenseNet201. 实验所应用的是在 ImageNet 数据集上预先训练的 DenseNet201 模型[19]. 该模型特色在于其包含的多个密集块，每个密集块内部由多层卷积层堆叠而成(不同密集块的层数各异)，这些卷积层通过密集连接的方式相互关联，即每一层的输入均包含了其前面所有层的输出信息；3) Inception V3. 实验选取了在 ImageNet 数据集上预先训练的 Inception V3 模型[20]. 此模型架构由多个 Inception 模块组合而成，每个 Inception 模块内嵌有多个并行分支，这些分支分别采用不同的卷积核大小进行卷积运算；4) LSTM. 实验中使用的 LSTM 模型的具体配置为包含两层隐藏层，每层均设有 512 个隐藏单元. 该 LSTM 模型的输出随后被送入一个全连接层和一个 Softmax 层，以实现分类功能. 所有模型均在本文自建的皮革瑕疵时频谱图数据库上进行测试，且超参数设置保持一致. 本小节记录了每个训练轮次的训练集准确率和平均精度，并确保所有模型均在收敛状态下完成训练. 实验结果如图 6 所示.

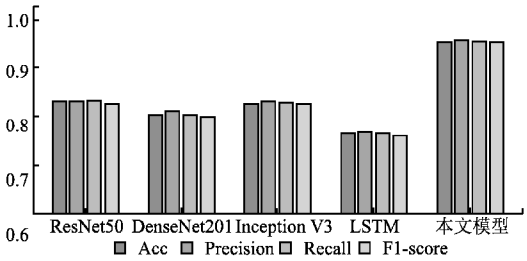

图 6 各个模型的性能

Fig. 6 Performance of each model

图 6 展示了各模型的各项性能指标对比. 从图中可以看出，本文模型在自建皮革瑕疵数据集上的性能明显优于其他所有对比模型. 具体而言，本文模型基于 Vision Transformer 网络进行了改进，其准确率达到了 95.62%，分别比 ResNet50、DenseNet201、Inception V3 和 LSTM 提高了 12.29%、15.2%、12.7% 和 18.95%，F1 分数则分别提高了 0.1264、0.1535、0.1271 和 0.1896. 这主要归因于以下 3 点：首先，Vision Transformer 网络在图像分类领域具有卓越的性能，其准确率等方面优于经典的卷积神经网络；其次，本文运用了迁移学习的策略，通过在 ImageNet 数据集上进行预训练，利用相似任务的参数更好地初始化模型，使模型直接继承了一般的特征提取能力；最后，通过加入随机掩蔽方法，本文引入了随机性，增强了模型的泛化能力，使其更有利于对测试集中未见样本的分类和推理.

3.2.2 与现有工作的对比

为了进一步验证本文所提模型的性能表现，本小节特选取近年来在皮革瑕疵分类领域表现卓越的研究成果，与本文工作进行对比分析. 鉴于高光谱传感技术[5]、超声波传感技术[7]以及红外温度传感技术[8]尚未真正应用于皮革瑕疵分类的实践操作中，且触觉传感技术[6]的准确率相对较低，仅为 69.5%，故而本小节仅与技术体系成熟且准确率较高的光学传感方法展开对比分析.

鉴于本文所采用的数据集与光学方法所依据的数据集存在差异，具体而言，本文运用了时频谱图，而光学方法则依赖于图片数据. 为此，专门利用光学摄像头捕获了光学方法所需的图片，构建了相应的图片数据集，以适应光学分析之需. 为了全面对比毫米波雷达传感技术与光学方法在抗光线干扰性能上的优劣，在一个不受日光干扰的会议室环境中，设计了 3 种不同的光照场景来采集数据集：弱光场景(仅依赖室内灯

光)、强光场景(室内灯光照明辅以直射皮革样本的台灯)以及黑暗场景(所有灯光均关闭). 在每个场景下,对每个样本进行了一次图片数据采集,共采集了 40 个样本,涵盖 6 种类别,因此每个场景获得了 240 份图片数据,总计 3 个场景则收集了 720 份图片数据. 随后,采用旋转、翻转、缩放、裁剪等多种技术手段对数据进行了扩充,每幅图片被扩充为 5 份,最终累积得到 3600 份图片数据. 3 种场景下的图片示例参见图 7.

(a)弱光　(b)强光　(c)黑暗

图 7　不同光照条件下的皮革样本

Fig. 7　Leather samples under different lighting conditions

与此同时,利用毫米波雷达在相同 3 种场景下分别采集了对应的数据,以构建对比数据库. 在每个场景下,对每个样本进行了 5 次采集,获取 5 份雷达数据,同样采集了 40 个样本,涵盖 6 种类别,故每个场景累计采集了 1200 份雷达数据,3 个场景共计 3600 份雷达数据. 这些雷达数据随后通过短时傅里叶变换(STFT)转换为相应的时频谱图,以供本文所提方法使用. 实验结果详见图 8、图 9 及图 10.

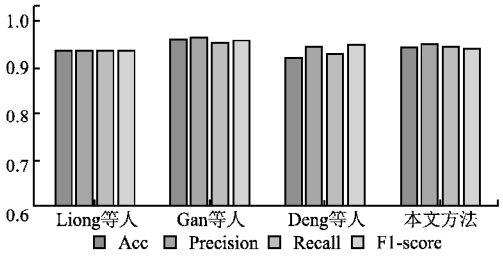

图 8　弱光场景下的性能对比

Fig. 8　Performance comparison under low light conditions

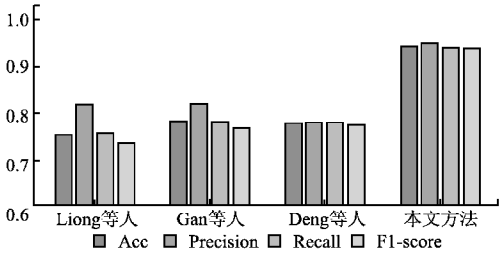

图 9　强光场景下的性能对比

Fig. 9　Performance comparison under strong light conditions

图 8、图 9 及图 10 分别详尽地呈现了本文所提方法与现有光学方法在弱光、强光及黑暗 3 种光照条件下对皮革瑕疵分类性能的比较结果. 具体而言,如图 8 所示,在弱光环境中,Liong 团队、Gan 团队、Deng 团队的光学方法以及本文方法的准确率依次为 93. 72%、96. 03%、92. 17% 及 94. 38%,其 F1 分数则分别为 0. 937、0. 9590、0. 9491 及 0. 9429. 由此可见,本文方法与各光学方法在弱光条件下的性能表现相当,均展现出优异水准. 进一步观察图 9,在强光场景下,Liong 团队、Gan 团队、Deng 团队的光学方法分别降至 75. 42%、78. 33%、77. 92%,其 F1 分数亦随之变化为 0. 7368、0. 7699、0. 7765,而本文方法的准确率和 F1 分数依然保持为 94. 17% 及 0. 9409. 显然,本文基于毫米波雷达传感的方法在强光环境下的性能显著优于所有对比的光学方法. 此优势可能源于光学方法易受光线条件影响,特别是在强光照射下,过度反射会掩盖皮革样本的瑕疵特征. 至于图10所展示的黑暗场景,Liong团队、Gan 团队、Deng 团队的光学方法以及本文方法的准确率分别为 41. 46%、57. 92%、50. 83% 及 94. 79%,F1 分数则分别为 0. 2879、0. 5543、0. 4554 及 0. 9472. 在此极端光照条件下,本文方法依然保持卓越表现,而光学方法则几乎失效. 这一现象有力证明了本文方法具备出色的抗光线干扰能力,能够在包括强光和黑暗在内的多种复杂环境中稳定工作,而现有光学方法则难以胜任.

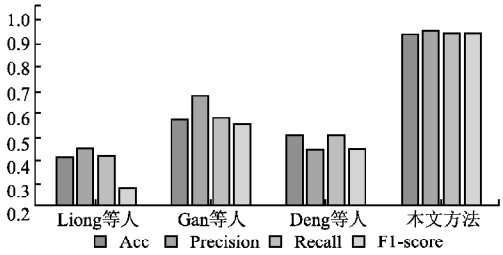

图 10　黑暗场景下的性能对比

Fig. 10　Performance comparison under strong dark conditions

3.3　消融实验

为进一步验证本文提出的改进点对基础模型性能提升的贡献,本小节进行了消融实验. 在同样训练 30 轮的情况下,消融实验结果如图 11 所示. "-预训练"表示不使用 ImageNet 预训练模型的权重,只使用随机的权重进行初始化;"-随机掩蔽"表示跳过随机掩蔽模块,将时频谱图数据分块后直接输入到模型中;"-预训练-随机掩蔽"表示使用基础的、未经改进的 VIT 模型,模型参数随机初始化,且不经过随机掩蔽模块.

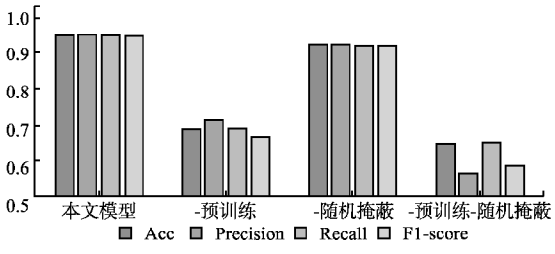

图 11　消融实验结果

Fig. 11　Ablation result

图 11 展示了在不同实验条件下进行的消融实验结果. 从该表中可以清晰地观察到,结合了预训练和随机掩蔽策略条件下的本文模型相较于其他条件下的模型展现出了最优的性能,这一结果进一步坚实地验证了本文所提出方法的有效性. 具体而言,相较于只使用随机掩蔽的模型,本文模型在效果上展现出了显著的优越性,其各项指标都取得了大幅度的提升,准确率、精度、召回率和 F1 分数分别提升了 26. 45%、0. 24、0. 2646和0. 2806,这有力地表明迁移学习不仅显著得增强了模型的推理能力,还加速了模型的收敛过程. 同样地,与仅使用预

训练的方法相比,本文模型也表现出了更优的性能,准确率、精度、召回率和 F1 分数分别提升了 2.91%、0.027、0.0292 和 0.0305,这进一步证实了随机掩蔽对于提升模型泛化能力的重要作用,使其在面对测试集中未见过的样本数据时,能够更好地进行推理和分类.综上所述,消融实验的结果充分且有力地证明了本文所提出的方法及其改进点的有效性和合理性.

3.4 掩蔽率实验

在随机掩蔽模块中,掩蔽率的设置对模型性能具有重要影响.掩蔽率过低可能导致过拟合,降低模型泛化能力;而掩蔽率过高则可能造成欠拟合,导致模型无法学习到足够多的信息.为探究随机掩蔽中最优的掩蔽率设置,本小节进行了不同掩蔽率设置的实验.实验结果如图 12 所示.

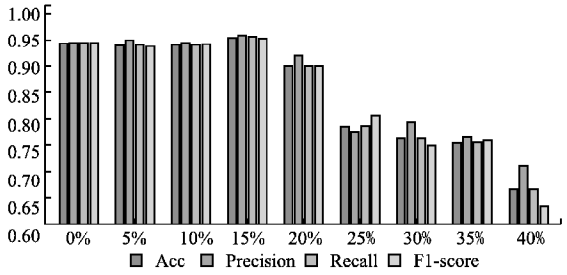

图 12 掩蔽率实验结果

Fig. 12 Mask rate result

根据图示信息,可以观察到,当掩蔽率从 0% 递增至 15% 时,模型性能呈现出逐步提升的趋势,这表明适度的掩蔽率对于提升模型的泛化能力具有积极作用.然而,自 15% 起,随着掩蔽率的继续增加,模型性能却开始急剧下滑,这揭示了过高的掩蔽率会阻碍模型获取足够的信息,进而对模型性能产生不利影响.综上所述,针对本文模型,最佳的掩蔽率被确定为 15%.

4 结 论

本文阐述了一种创新的抗光线干扰的皮革瑕疵分类技术,该技术巧妙融合了毫米波雷达技术与改进后的 Vision Transformer 模型.本文首先利用毫米波雷达技术采集皮革瑕疵相关数据,并借助短时傅里叶变换技术,将这些信号转化为时频谱图,进而构建了相应的时频谱图数据集.在此基础上,本文精心构建并训练了一个基于 Vision Transformer 的神经网络模型,旨在高效完成瑕疵分类任务,并通过迁移学习策略进一步优化了模型的性能.此外,本文还采用了随机掩蔽策略对数据集进行增强处理,从而显著提升了模型的泛化能力和整体性能.为验证本文方法的有效性,本文利用自建的包含 6 类皮革瑕疵的样本数据集进行了实验评估.实验结果显示,本文所提出的方法在皮革瑕疵分类方面展现出了极高的准确性与卓越的性能.然而,需注意的是,在实际工业生产环境中,皮革通常处于动态传输的带式输送线上,因此,将该方法拓展至动态场景下的皮革瑕疵分类任务,不仅具有重要意义,而且迫在眉睫.鉴于此,在未来的研究工作中,将继续深化这一领域的探索与研究,致力于实现更为精确且全面的皮革瑕疵检测与分类技术,为工业生产提供更加坚实的技术支撑与保障.

References:

[1] Chen Z, Xu D, Deng J, et al. Comparative study on deep-learning-based leather surface defect identification[J]. Measurement Science and Technology, 2023, 35(1): 015402.
[2] Liong S T, Zheng D, Huang Y C, et al. Leather defect classification and segmentation using deep learning architecture[J]. International Journal of Computer Integrated Manufacturing, 2020, 33(10-11): 1105-1117.
[3] Gan Y S, Liong S T, Zheng D, et al. Detection and localization of defects on natural leather surfaces[J]. Journal of Ambient Intelligence and Humanized Computing, 2023, 14(2): 1-15.
[4] Deng J, Liu J, Wu C, et al. A novel framework for classifying leather surface defects based on a parameter optimized residual network [J]. IEEE Access, 2020, 8: 192109-192118, doi: 10.1109/ACCESS.2020.3032164.
[5] Chen S Y, Cheng Y C, Yang W L, et al. Surface defect detection of wet-blue leather using hyperspectral imaging [J]. IEEE Access, 2021, 9: 127685-127702, doi: 10.1109/ACCESS.2021.3112133.
[6] Xu S, Xu H, Mao F, et al. Flexible material quality assessment based on visual-tactile fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 3386205, doi: 10.1109/TIM.2024.3386205.
[7] Liu C K, Latona N, Yoon S C. Evaluation of hides, wet blue and leather using airborne ultrasonics [J]. Journal of the American Leather Chemists Association, 2013, 108(4): 128-138.
[8] Chudzik S. Quality inspection of natural leather using non-destructive testing technique [J]. Quantitative InfraRed Thermography Journal, 2020, 17(4): 249-267.
[9] Wang F, Zeng X, Wu C, et al. mmHRV: contactless heart rate variability monitoring using millimeter-wave radio[J]. IEEE Internet of Things Journal, 2021, 8(22): 16623-16636.
[10] Cooley J W, Lewis P A W, Welch P D. The fast fourier transform and its applications[J]. IEEE Transactions on Education, 1969, 12(1): 27-34.
[11] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[C]//International Conference on Learning Representations, 2021: 1-21.
[12] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Conference of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies, 2019: 4171-4186.
[13] WANG Y, LI Y C, XU J W, et al. Crop disease recognition method based on improved vision transformer network[J]. Journal of Chinese Computer Systems, 2024, 45(4): 887-893.
[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. advances in neural information processing systems [J]. arXiv preprint arXiv: 1706.03762, 2017, 10: S0140525X16001837.
[15] Gong Y, Lai C I, Chung Y A, et al. Ssast: self-supervised audio spectrogram transformer[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2022: 10699-10709.
[16] Gong Y, Chung Y A, Glass J. Ast: audio spectrogram transformer [J]. arXiv preprint arXiv: 2104.01778, 2021.
[17] Chang H, Zhang H, Jiang L, et al. Maskgit: masked generative image transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 11315-11325.
[18] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
[19] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.
[20] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2818-2826.

附中文参考文献:

[13] 王 杨,李迎春,许佳炜,等.基于改进 Vision Transformer 网络的农作物病害识别方法[J].小型微型计算机系统,2024,45(4):887-893.