基于改进图像多标签分类的古籍异体字部件识别研究*

郑澄璐[1,2]　王昊[1,2]　孟序阳[3]　杨思敏[1,2]
（1. 南京大学信息管理学院，南京 210023；2. 数据工程与知识服务江苏省高校重点实验室（南京大学），南京 210023；3. 中国人民大学信息资源管理学院，北京 100872）

摘要：针对古籍异体字多为图像存储，无法直接数字化检索的问题，本研究基于多标签分类的古籍异体字识别方法，构建ECA-HFF-DTA模型，包括增强型通道注意力（ECA）模块、多尺度特征融合分类头（HFF-Head）和动态阈值调整（DTA）机制3个模块，自动识别其构成部件并生成多标签分类结果。实验表明，模型在独立测试集上的F1值达到0.578 3，较原始模型提升显著，可以实现基于部件查询异体字的应用功能。但模型对复杂多嵌套结构异体字的识别效果仍有待提高。ECA-HFF-DTA模型提升了古籍异体字部件识别效果，为大规模古籍文献异体字的智能化整理与检索提供思路。

关键词：古籍；异体字识别；多标签分类
中图分类号：H087　DOI：10.3772/j.issn.1673-2286.2025.10.002
引文格式：郑澄璐，王昊，孟序阳，等. 基于改进图像多标签分类的古籍异体字部件识别研究[J]. 数字图书馆论坛，2025，21（10）：12-22.

古籍作为中华文明的重要载体，承载着丰富的历史文化信息与学术研究价值。然而，古籍文本中广泛存在的异体字（或称异形字）问题，已成为制约其数字化进程与深度利用的关键瓶颈。异体字指音义相同而形体不同的汉字变体，其成因复杂，既源于汉字初创时期造字方式的多样性，也与历代地域文化差异、书写材料演变及避讳制度等因素密切相关。据统计，《康熙字典》收录的4万余字中，异体字占比超过1/3；《汉语大字典》更附有11 900组异体字[1]。

这些异体字在现有计算机字符集中大多缺失，导致其在数字化过程中难以直接输入，往往以图像形式保存，严重阻碍了古籍文本的结构化处理与智能化检索。现有Unicode标准虽已收录大量异体字，但因字符优先于字形的设计原则，将异体字视为独立编码，导致字形关联断裂、检索效率低下[2]。尉迟治平[3]亦指出，字符集规模与古籍复杂信息的处理方法是制约计算机辅助汉语史研究的关键瓶颈。这一矛盾在古籍数字化中尤为凸显：一方面，异体字须保留原貌以维护文本的历史真实性[4]；另一方面，字符集冗余与检索障碍严重制约数据库的实用价值，异体字从整字识别转向部件分析已成为一大趋势。

当前古籍数字化实践主要聚焦于图像化与文本化两个层面。图像化技术已较为成熟，文本化工作仍面临字形处理标准的缺失与异体字识别的技术挑战[5]。传统光学字符识别（OCR）技术在处理异体字时，因字符集覆盖不足与字形变体复杂性，常面临误识别率高、集外

收稿日期：2025-07-30
*本研究得到国家自然科学基金面上项目“面向多模态非遗文化数据的特征挖掘及全知识图谱构建研究”（编号：72574098）、南京大学中央高校基本科研业务费专项资金资助项目AI for HASS“AI赋能中国古籍史书知识图谱构建研究”（编号：010814370338）、江苏青年社科英才项目、南京大学仲英青年学者项目资助。

字无法处理等问题。例如，《四库全书》电子版虽通过人工造字与规则匹配初步解决了部分异体字录入问题，但其字符集规模已超过10万，仍存在检索冗余与效率低下的缺陷[1]。此外，现有研究多依赖人工标注与专家经验，缺乏自动化潜力与可扩展性，难以应对海量古籍资源的处理需求[6]。因此，如何通过技术手段实现异体字的精准识别与结构化处理，已成为古籍数字化领域的核心议题。

本研究以部件为切入点，提出一种基于多标签分类的古籍异体字识别方法。在深度学习模型基础架构上引入增强型通道注意力（ECA）模块和多尺度特征融合分类头（HFF-Head）对异体字图像进行端到端分析，自动识别其构成部件并生成多标签分类结果。以部件为检索单元，可有效缓解字符集容量限制，提升古籍数据库的查全率与查准率。用户通过输入目标部件的组合，即可检索包含特定构形模式的所有异体字，极大增强了古籍资源的可访问性。

1　相关研究

1.1　古籍异体字识别研究进展

异体字是汉字发展过程中的重要现象，其界定与分类始终是文字学研究的基础问题。传统文献中，异体字常被定义为“音义全同而形体相异”的汉字[7]。李运富[8]进一步指出，异体字的内涵可从字样、字构、字用3个不同角度分别界定，其本质属性在于“功能相同”而非简单的“形体不同”。正如史建华[9]在《尔雅音义》异体字研究中所指出的，异体字的训诂用语与字形分类对理解汉字演变具有重要意义。尉迟治平[10]在电子《广韵》研究中发现，现有字符集（如CJK）仅能覆盖古籍异体字的有限子集，约15%的异体字因缺乏对应码点无法数字化，这直接影响了古籍文本的机器可读性。

传统异体字整理主要依托文献校勘与字典编纂。刘又辛[11]在《汉语大字典》修订中提出“字位择字”原则，强调以规范字为核心建立异体字谱系，通过考辨字形源流实现系统性整理。这种基于专家知识的整理模式虽具权威性，但面临效率瓶颈。据刘延玲[12]统计，《汉语大字典·异体字表》中约20%的条目存在误收现象，反映出人工整理的局限性。

数字化时代，异体字处理面临三重挑战。①字符集覆盖不足问题。ISO/IEC 10646标准虽已收录7万余个汉字，但古籍异体字存量远超此规模，章杰鑫等[13]指出，仅《广韵》中未被字符集收录的篆楷异体字占比达32%。②字形变体识别难题。古籍中异体字常伴随墨渍、残损、版刻风格差异（如宋体与楷体混排），传统OCR技术对此类变体的识别准确率不足60%[10]。③语义关联缺失。多数数字化方案仅关注字形映射，未构建异体字与正体字的语义网络，导致检索系统无法支持基于部件的联想查询[14]。

早期数字化尝试主要采用字符集扩展与字形关联技术。董月凯[15]分析GB 18030编码体系时指出，通过四字节编码虽可扩容至160万码位，但人工标注海量异体字的成本极高。微软“宋体-方正超大字符集”通过字形数据库关联异体字（如“赞”与“讚”），实现了部分异体字的可视化呈现，但受限于封闭系统架构，难以跨平台共享。近年来，深度学习技术为异体字识别提供了新路径。雷华等[16]在《汉语大字典》修订中尝试使用卷积神经网络（CNN）识别甲骨文异体字，实验结果显示，对结构规整字形的识别准确率达89%，但对笔画粘连或残缺样本的识别准确率仅为47%。

1.2　基于深度学习的图像多标签分类研究

早期多标签分类研究主要基于传统机器学习方法，可分为问题转换和算法自适应两类。问题转换方法将多标签任务拆解为多个二分类任务（如二元关联法[17]）或转化为多分类任务（如Label Powerset方法[18]），其局限性在于难以捕捉标签间的语义关联。算法自适应方法则直接扩展传统算法（如多标签决策树[19]、排名支持向量机[20]），通过修改损失函数或特征表示方式处理多标签数据，在高维标签空间仍面临计算复杂度高的问题。

深度学习技术的兴起为多标签分类提供了新思路。基于CNN的方法通过端到端特征学习显著提升了分类性能。例如，Wei等[21]提出的HCP框架通过假设池化策略整合多区域特征，在VOC2007数据集上取得90.9%的平均精度。薛丽霞等[22]提出基于DenseNet-121的特征提取框架，结合通道注意力机制动态调整特征响应权重，在Pascal VOC数据集上实现了90.4%的平均精度。然而，传统CNN模型对长距离依赖关系建模不足，难以有效处理标签间的复杂相关性。为此，研究者引入循环神经网络（RNN）与注意力机制，如Wang等[23]提出CNN-RNN架构，通过RNN建模标签序列依赖关系。

张辉宜等[24]采用ResNet-101作为主干网络，通过图注意力网络（GAT）建模标签共现关系，在MS-COCO数据集上将平均精度提升至78.2%。张晓瑞[25]引入四元数Gabor滤波进行纹理特征提取，结合CNN的双重优化策略，在VOC2007数据集上获得0.8以上的Kappa系数，表明传统特征提取方法与深度学习的融合潜力。近年来，提示学习与注意力机制的结合为多标签分类提供了新思路，如汪瑞等[26]提出的基于提示学习和交叉注意力的方法在部分标签场景下表现出优越性能。

针对类别不平衡和噪声标签问题，研究者提出多种优化方案。武红鑫等[27]系统分析了类别权重调整方法，通过逆频率加权缓解长尾分布影响，有效提升了低频部件的召回率。动态阈值调整（DTA）是另一个重要方向，不同于传统固定阈值法，张晓瑞[25]采用基于概率分布的自适应阈值，使得预测结果更贴合实际数据分布。非对称损失（ASL）函数[28]通过差异化调节正负样本权重，有效缓解了头部标签主导分类的问题。

综上所述，传统异体字识别方法由于字符集限制和人工标注成本高，并不能有效解决异体字查询与数字化问题。因此，异体字研究逐渐从整字识别转向部件分析，异体字部件自动化识别能够在有效减少人工成本的同时支持基于部件的联想查询。而深度学习驱动的多标签分类技术通过架构创新与损失函数优化，逐步解决了标签相关性建模、特征表达不足等核心问题。在古籍异体字识别场景中，基于深度学习，结合注意力机制、DTA及预训练模型微调的策略已展现出实用价值。

2　基于改进图像多标签分类的模型设计

2.1　ECA-HFF-DTA模型架构

针对古籍异体字部件识别任务中存在的部件语义关联复杂、类别分布不均衡、结构轻量化等问题，在EfficientNet-B0模型基础上进行多维度改进，提出ECA-HFF-DTA模型（见图1）。该模型进行了3个模块的改进，用ECA模块替换原有的压缩激励（SE）模块，引入了HFF-Head替换原始分类头，并加入DTA机制的决策优化策略。

模型基础架构是谷歌公司于2019年推出的EfficientNet模型。该模型重新定义了CNN的模型缩放策略，可满足图像分类领域对准确性和高效率的需求。EfficientNet-B0采用移动倒置瓶颈卷积（MBConv）模块作为基础单元，其结构包含扩展卷积、深度可分离卷积和SE模块3个核心组件。

EfficientNet-B0作为基础架构具有以下优势：相较于传统CNN模型，MBConv模块通过深度可分离卷积，在减少参数量的同时保持特征表达能力；复合缩放策略使模型能够自适应调整感受，适应古籍异体字中部件尺度差异显著的特点；在ImageNet数据集上的预训练权重提供良好的特征初始化参数，缓解小样本训练中的过拟合问题。

模块1在特征提取阶段进行了改进，引入了ECA模块，替代了传统的SE注意力机制。该模块采用一维卷积

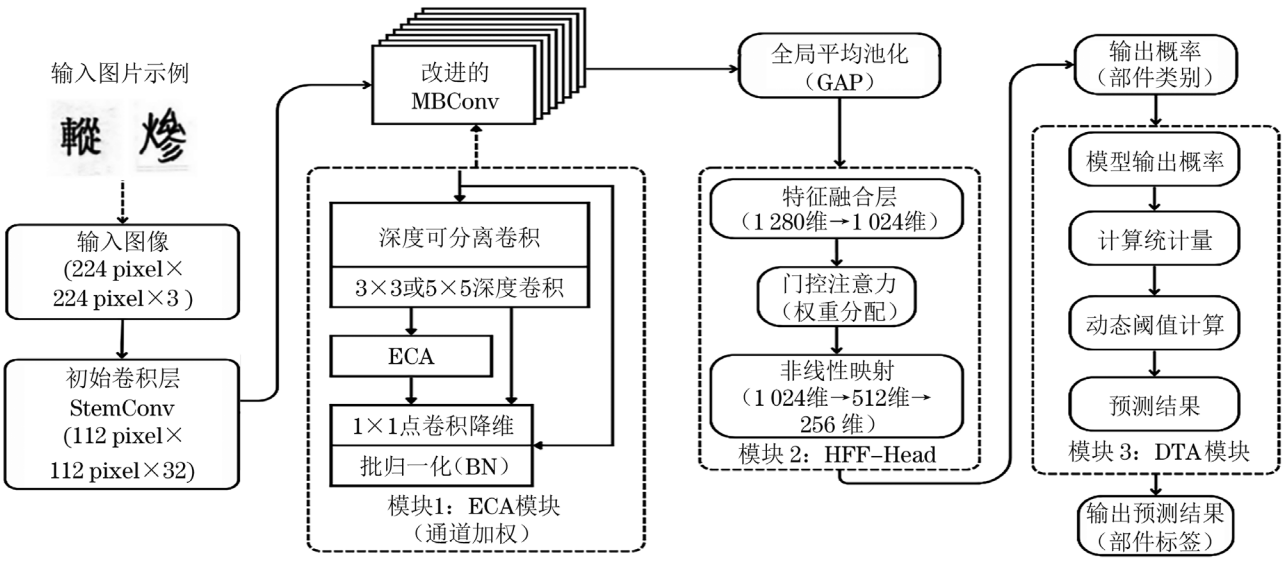

图1　ECA-HFF-DTA模型架构

实现局部跨通道交互，在显著降低参数量的同时保留完整的通道信息，有效增强了模型对异体字细微笔画差异（如“戶”与“户”的横笔倾斜度）的感知能力。

模块2提出了HFF-Head的改进策略。该结构突破传统单层全连接（FC）层的局限，采用三级特征融合框架，通过门控注意力机制实现特征权重的自适应分配，并经非线性映射网络进行高阶特征交互。这种多尺度融合机制使模型能够协同利用笔画纹理、部件轮廓和整体布局等多粒度信息，显著提升对复杂结构汉字的解析能力。

模块3实现了DTA机制。这种自适应策略使模型能够协调处理类别不平衡问题，同时保障部件预测的语义一致性，如避免“门”与“户”的互斥部件共存，显著增强系统对输入质量波动的鲁棒性。3项改进在结构上形成协同增强作用：ECA模块优化特征编码，HFF-Head强化特征融合，DTA机制完善决策逻辑，共同构建了端到端的异体字部件识别新范式。

2.2　数据集与数据预处理

本研究采用的数据集如表1所示。①基准部件库：包含317个标准汉字部件（含偏旁部首与独体字），经筛选后保留259个部件。独体字数据来源于中华人民共和国教育部、国家语言文字工作委员会于2009年发布的《语言文字规范 现代常用独体字规范》（GF 0013—2009）。该规范经国家语言文字工作委员会语言文字规范（标准）审定委员会审定，具有官方权威性。同时为保障样本数据平衡性，删去了出现频次低于15次的部件。②训练集和验证集：共包含8 106个复杂字样本，每个样本标注部件标签。③测试集：包含5 272个未标注复杂异体字，来源于某古籍数据公司整理的百衲本史书中包含的异体字，图2展示了部分异体字图像。为评估模型，从中随机抽取400个代表性样本进行精细人工标注，用于定量评估。

针对图像数据处理，采用系统性多模态数据处理

表1　数据集情况

数据集	样本数量/个
基准部件库	259
训练集	6 485
验证集	1 621
测试集	5 272

图2　部分异体字图像

框架，针对汉字部件识别任务的数据特性与模型需求，构建了包含数据清洗、几何校正、增强合成、特征归一化的四阶段预处理流程。

首先，通过自动化质量筛检系统对原始图像集进行清洗。基于Laplacian方差算法剔除模糊样本，利用连通域分析检测并修复断裂笔画，同时结合OCR校验修正标注错误，确保数据集的几何完整性与标签一致性。其次，针对汉字部件多尺度、多方向性特点，实施动态几何校正。采用各向异性插值算法将输入图像统一缩放至基准尺寸（224 pixel×224 pixel），并通过主成分分析（PCA）确定部件主方向后执行仿射变换，以消除书写风格差异导致的姿态偏移。

在数据增强阶段，采用分层次混合增强策略。基础层应用空间变换（随机裁剪、透视形变）与光度畸变（高斯噪声、对比度调节），模拟现实场景中的形变与光照变化；高级层引入汉字特异性增强模块，包括基于笔画骨架的弹性形变，模拟手写抖动效应以及部件遮挡（随机擦除区域占比10%~30%）。为解决类别不均衡问题，设计双通道采样机制，对低频部件实施条件生成对抗网络增强。

2.3　ECA模块设计

传统的通道注意力机制SE模块包含全局压缩和激励两个阶段。传统SE模块虽能建模通道相关性，对所有通道进行压缩和重新加权，但其FC层结构对局部空间信息的捕获能力有限，FC层的静态参数化交互模式忽略了相邻通道间的局部相关性，难以捕捉汉字部件中常见的连续性结构特征。

为此，引入如图3所示的ECA模块，以一维卷积替代FC层，增强局部特征交互。模块包含以下关键步骤。①特征压缩：对输入特征图$\boldsymbol{X}\in\mathbb{R}^{H\times W\times C}$实施全局平均池化（GAP），得到通道描述向量$\boldsymbol{z}\in\mathbb{R}^{C}$，计算方法如式（1）所示。②局部交互建模：对$\boldsymbol{z}$进行一维卷积操作（卷积核大小为$k$），捕获相邻通道间关联，输出$\tilde{\boldsymbol{z}}$，计算方法如式（2）所示。$k$根据通道数$C$自适应确定，计算方法如式（3）所示。③权重生成：通过Sigmoid函数生成通

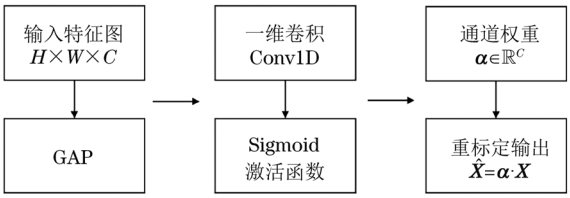

图3　ECA模块结构

道注意力权重$\boldsymbol{\alpha}\in\mathbb{R}^C$，计算方法如式（4）所示。④特征权重标定：将注意力权重与原始特征图逐通道相乘，得到增强后的特征图$\hat{\boldsymbol{X}}$，计算方法如式（5）所示。

$$z_C=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{X}_C(i,j) \tag{1}$$

$$\tilde{z}=\mathrm{Conv1D}(z;k) \tag{2}$$

$$k=\left|\frac{\mathrm{lb}(C)}{\gamma}+\frac{b}{\gamma}\right|_{\mathrm{odd}} \tag{3}$$

$$\boldsymbol{\alpha}=\mathrm{Sigmoid}(\tilde{z}) \tag{4}$$

$$\hat{\boldsymbol{X}}_C=\boldsymbol{\alpha}_C\cdot\boldsymbol{X}_C \tag{5}$$

式中：H、W、C为图像高度、宽度和通道数；γ=2，b=1，$|\cdot|_{\mathrm{odd}}$表示取最邻近奇数。

改进后的高效通道注意力模块通过轻量化动态交互机制实现了突破。结构优化方面，首先摒弃降维操作，直接对GAP输出的C维向量进行一维卷积处理，通过大小为k的卷积核捕获局部跨通道交互，在保留完整通道信息的同时，显著降低参数量；其次采用参数共享策略，同一卷积核作用于所有通道，增强模型对汉字部件空间分布一致性的建模能力。功能增强方面，ECA模块通过自适应卷积核选择机制，使得不同层级的特征图能根据通道数自主选择交互范围。浅层低维特征采用较小核（k=3）聚焦局部细节，深层高维特征使用较大核（k=9）捕捉全局关联；同时，一维卷积的滑动计算模式天然保留了通道间的空间连续性，尤其适用于汉字部件中常见的线性结构。此外，ECA模块摒弃了SE模块中的Sigmoid激活函数，直接对卷积输出进行L2归一化，避免极端权重值的出现，提升注意力权值的平滑性。相较于标准SE模块，ECA模块通过局部卷积替代FC操作，在降低参数量的同时增强相邻通道间的关联建模能力。

针对异体字部件识别任务，ECA模块对异体字部件识别的影响体现在3个方面：①局部感知特性可增强模型对部件形态细节的敏感性，例如“⺮”（竹字头）中竖笔长度差异，ECA模块能通过相邻通道的协同激活放大区分度；②轻量化结构降低了对训练数据量的依赖，在有限样本（如罕见异体字）下仍能稳定学习判别性通道权重；③无降维设计完整保留了部件间的空间拓扑关系，例如“門”字中左右部件的对称性特征可通过高维通道的完整交互得以强化，特别是在结构相似度高的困难样本识别上ECA模块更具优势。

2.4　HFF-Head模块设计

原始EfficientNet模型的分类头设计仅由GAP层和单个FC层构成，这种简化的设计存在一些缺陷。首先，单一的FC层难以充分挖掘主干网络提取的深层特征中蕴含的复杂语义信息。其次，直接将高维特征映射到分类空间，容易导致特征分布松散且判别性不足，特别是对具有多部件组合特性的汉字结构，难以捕捉部件间的空间关联。最重要的是，原始结构仅利用最终层的抽象特征，而忽略了浅中层特征中保留的细节信息（如笔画走向、部件连接等），这种单尺度特征处理方式严重制约了模型对汉字多粒度特征的融合能力。

基于此，设计如图4所示的HFF-Head模块结构，其构建过程如下。①多级特征提取：从主干网络的Stage 3~Stage 6提取4个层级特征图$\boldsymbol{F}_3$、$\boldsymbol{F}_4$、$\boldsymbol{F}_5$、$\boldsymbol{F}_6$，分辨率分别为输入图的1/8、1/16、1/32、1/32。②特征对齐与融合：将高层特征上采样至与$\boldsymbol{F}_3$相同的分辨率，拼接后得到融合特征$\boldsymbol{F}_{\mathrm{fuse}}$，计算方法如式（6）所示。③通道压缩：使用1×1卷积将通道数降至512，得到压缩特征$\boldsymbol{F}_{\mathrm{comp}}$，计算方法如式（7）所示。④非线性映射：构建一个包含BN层、ReLU激活函数与Dropout层的分层结构，将512维特征进一步映射到256维，增强特征的判别性，计算方法如式（8）所示。⑤分类器设计：最终通过FC层将256维特征映射到部件类别空间，并使用Sigmoid函数输出每个部件的概率，得到部件概率向量$\boldsymbol{p}\in\mathbb{R}^{N_{\mathrm{classes}}}$，计算方法如式（9）所示。该设计通过融合浅层高分辨率特征与深层语义特征，并经过多层次的特征

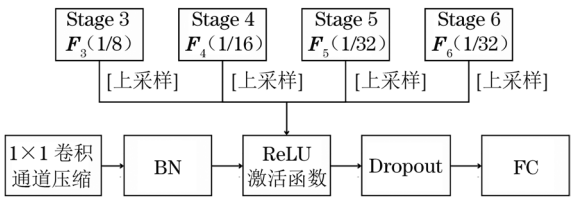

图4　HFF-Head模块结构

变换，显著提升对复杂异体字部件的识别能力。

$$\boldsymbol{F}_{\text{fuse}} = \text{Concat}\left[\boldsymbol{F}_3, \text{Up}\left(\boldsymbol{F}_4\right), \text{Up}\left(\boldsymbol{F}_5\right), \text{Up}\left(\boldsymbol{F}_6\right)\right] \quad (6)$$

$$\boldsymbol{F}_{\text{comp}} = \text{Conv}_{1\times1}\left(\boldsymbol{F}_{\text{fuse}}\right) \quad (7)$$

$$\boldsymbol{F}_{\text{feat}} = \text{Dropout}\left(\text{ReLU}\left(\text{BN}\left(\text{FC}_{512\to256}\left(\boldsymbol{F}_{\text{comp}}\right)\right)\right)\right) \quad (8)$$

$$\boldsymbol{p} = \text{Sigmoid}\left(\text{FC}\left(\boldsymbol{F}_{\text{feat}}\right)\right) \quad (9)$$

在HFF-Head的具体实现中，经过特征对齐与拼接后的融合特征通道数较高，采用1×1卷积将其压缩至512维。该维度的选择基于性能与复杂度的平衡：过低的维度（如256维）会削弱特征的表达能力，导致模型难以刻画复杂部件的细微差异；而过高的维度（如1 024维）则会显著增加后续FC层的参数量，易引发过拟合，且模型的计算开销与推理时间也会成倍增长。512维是计算机视觉领域中一个经验证能有效平衡表征能力和计算效率的通用设计。此外，在最终分类层之前，引入了Dropout层，其比率设置为0.5，能在训练过程中通过随机丢弃一半的神经元，有效破坏神经元间的复杂共适应性，从而强制网络学习更鲁棒的特征，是防止模型在中等规模数据集上过拟合的经典且有效的正则化策略。

HFF-Head通过以下机制实现特征融合：深层FC层对GAP层输出的全局语义特征进行高阶非线性变换，提取更具判别性的抽象特征；BN层在训练过程中动态学习特征分布的缩放偏移参数，自适应调整不同层次特征的贡献权重；Dropout层通过随机丢弃部分神经元，迫使网络在残差路径中建立多通道的特征关联，隐式实现跨层特征的组合优化。这种“特征细化-权重校准-随机强化”的级联处理，本质上构建了多阶段特征融合通道，使得浅层的局部细节特征、中层的结构特征（如部件轮廓）和深层的语义特征（如部件组合模式）能够通过梯度回传机制在反向传播过程中协同优化。

在异体字部件识别任务中，HFF-Head对异体字部件识别的影响体现为3个方面：①多层级特征融合机制可显著提升模型对微观差异的敏感性，例如“⺮”（竹字头）与“艹”（草字头）的区分，传统模型易因全局特征相似而混淆，而HFF-Head通过局部细节向量捕捉竹字头两短竖的平行度差异；②注意力权重学习使模型在面对高度相似的异体字（如“亏”与“亓”）时，能动态调整对关键层级特征的依赖，例如在光照不均的扫描图像中，提升中层轮廓特征权重以补偿局部噪声干扰；③深度非线性映射结合正则化，通过随机屏蔽神经元路径迫使网络构建冗余特征表达，从而增强对低质量输入（如古籍中破损字、手写字等）的泛化能力。

2.5　DTA模块设计

在传统的多标签分类任务中，固定阈值通常被广泛用于将模型输出的概率值二值化为类别标签。然而，在汉字异体字部件识别场景下，固定阈值策略存在显著局限性：①类别不平衡导致阈值敏感性差异，高频部件因训练充分往往输出概率较高，而低频部件易因概率偏低而被漏检；②输入质量波动引发分布偏移，固定阈值难以适应此类动态变化；③部件间语义关联性被忽视，如“門”与“户”作为互斥部件，其阈值决策应具备相关性，但固定阈值独立处理每个类别，可能产生逻辑矛盾。这些问题在复杂异体字识别中尤为突出。

针对上述缺陷，特别是古籍异体字部件固有的长尾分布问题（少数部件出现频繁，而大量部件出现频次少），DTA机制通过概率分布感知与语义关系建模实现了优化。该机制的核心目标是自适应地应对训练与推理过程中的数据分布变化，而非依赖一个对所有类别“一刀切”的固定阈值。

核心设计包含两个模块。①自适应基线阈值计算：基于当前批次的概率分布统计量动态调整基础阈值，使阈值随分布特性自动伸缩，如高置信度批次提升阈值以减少假阳性，低置信度批次降低阈值以避免假阴性。②语义相关性加权：构建部件共现矩阵，通过矩阵分解学习隐式语义关系，对互斥或强关联的部件对施加阈值补偿项，动态调整相邻类别阈值以强化逻辑一致性。

DTA模块结构如图5所示，其核心机制如下。①自适应基线阈值计算：设当前批次中所有样本对类别c的输出概率为$\{p_c^{(i)}\}_{i=1}^{B}$，计算其均值μ_c与标准差σ_c，方法如式（10）所示。基础阈值T_c^{base}计算方法见式（11）：α为频率权重因子，用于缓解类别不平衡；N_c为类别c的训练样本数，N_{total}为总样本数，λ为超参数（设为0.1）。②语义相关性加权：构建部件共现矩阵$\boldsymbol{M}\in\mathbb{R}^{N\times N}$，通过矩阵分解得到语义嵌入向量$\boldsymbol{e}_i$、$\boldsymbol{e}_j$，计算语义关联强度$s_{ij}$，计算方法如式（12）所示。③最终阈值确定：结合基础阈值与语义补偿β_c，得到最终动态阈值T_c，计算方法如式（13）所示，β为语义补偿系数（设为0.05）。该机制使模型能够自适应调整阈值，增强对低频部件的识别能力，并抑制对语义矛盾部件的共存预测。

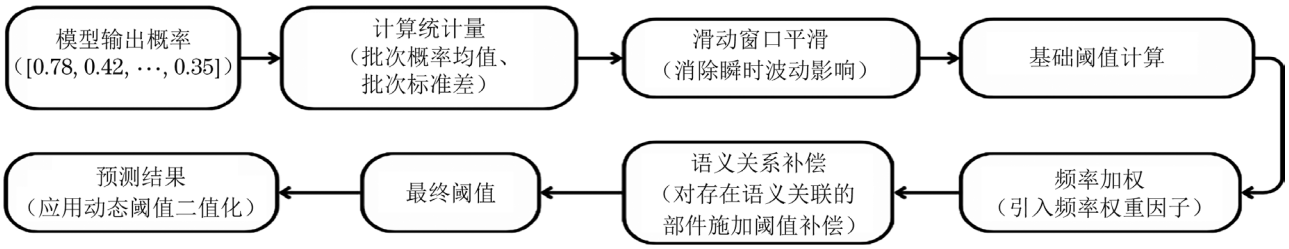

图5　DTA模块结构

$$\begin{cases}\mu_c = \dfrac{1}{B}\sum_{i=1}^{B} p_c^{(i)} \\ \sigma_c = \sqrt{\dfrac{1}{B}\sum_{i=1}^{B}(p_c^{(i)} - \mu_c)^2}\end{cases} \tag{10}$$

$$\begin{cases}T_c^{\text{base}} = \mu_c + \alpha \cdot \sigma_c \\ \alpha = \lambda \cdot \log\left(1 + \dfrac{N_{\text{total}}}{N_c}\right)\end{cases} \tag{11}$$

$$s_{ij} = \frac{\boldsymbol{e}_i^{\mathrm{T}} \boldsymbol{e}_j}{\|\boldsymbol{e}_i\|\|\boldsymbol{e}_j\|} \tag{12}$$

$$\begin{cases}T_c = T_c^{\text{base}} + \beta \cdot \beta_c \\ \beta_c = \dfrac{1}{N_c}\sum_{j \in N_c} s_{cj} \cdot p_j\end{cases} \tag{13}$$

在异体字部件识别中，DTA模块对异体字部件识别的影响体现为3个方面。①分布适应性：通过动态感知批次内概率分布，缓解图像质量退化（如古籍泛黄、污渍）导致的问题，确保关键部件不被遗漏。②类别平衡性：频率加权机制赋予低频异体字部件更灵活的激活阈值，可提高长尾部件的召回率。③语义逻辑性：基于共现关系的阈值补偿可抑制矛盾预测（如“门”与“户”同时激活），使复合结构的识别准确率提高。

3　ECA-HFF-DTA模型结果与分析

3.1　模型评估

为全面评估ECA-HFF-DTA模型的性能，在测试集中随机选取400个异体字样本进行人工标注，对该模型和EfficientNet-B0、EfficientNetV2-S、MnasNet、ShuffleNet 4个基准模型在测试集上的表现进行对比，并且为减少实验误差，取3次实验平均值，结果如表2所示。

由表2可知，提出的ECA-HFF-DTA模型在测试集上取得了最佳性能，其F1值达到0.578 3，与其他较前沿的模型比较，依然实现了不小的提升。这一结果充分证明了所提改进方案的有效性与先进性，验证了多尺度特征融合机制与轻量化注意力模块的协同优化效果，揭示了模型中3个核心改进模块在解决异体字部件识别关键难点，即同时实现高召回率与高精确率中所发挥的协同作用。从抽取测试集的异体字拆分结果（见表3）可以看出，本文提出的ECA-HFF-DTA模型相较EfficientNet-B0模型不仅显著提升了部件识别的准确性，还能识别出更多正确的结构细节，尤其在处理复杂异体字时，其性能改善更为显著。

表2　各模型在测试集上的评估对比

模型	精确率	召回率	F1值
EfficientNet-B0	0.255 4	0.497 8	0.337 5
EfficientNetV2-S	0.418 8	0.505 7	0.458 1
MnasNet	0.128 0	0.208 6	0.158 6
ShuffleNet	0.138 6	0.394 2	0.205 0
ECA-HFF-DTA	0.485 8	0.714 5	0.578 3

3.2　消融实验

消融实验结果见表4，结果表明，ECA模块、HFF-

表3　测试集部分异体字拆分结果对比

图像编号	图像	EfficientNet-B0模型	ECA-HFF-DTA模型
0001	騶	[‘艹’，‘虫’，‘巾’，‘也’，‘两’]	[‘心’，‘巾’，‘灬’，‘见’，‘也’]
1053	踈	[‘口’，‘木’，‘止’，‘巾’，‘束’，‘弟’]	[‘口’，‘木’，‘止’，‘束’]
2363	惸	[‘冖’，‘忄’，‘罒’，‘皿’，‘必’]	[‘冖’，‘忄’，‘工’，‘罒’]
3585	灩	[‘口’，‘艹’，‘讠’，‘亠’，‘勹’，‘羽’，‘幺’，‘母’]	[‘口’，‘艹’，‘土’，‘辶’，‘隹’，‘艹’]
4003	戄	[‘日’，‘土’，‘虫’，‘夂’，‘戈’，‘戊’，‘生’]	[‘口’，‘日’，‘土’，‘丷’，‘王’，‘夂’，‘夂’，‘立’，‘戈’]
5230	瑪	[‘王’，‘心’，‘灬’，‘雨’，‘非’，‘正’]	[‘灬’，‘正’，‘土’]

表4　消融实验结果

模型配置	精确率	召回率	F1值
EfficientNet-B0	0.255 4	0.497 8	0.337 5
EfficientNetV2-S	0.418 8	0.505 7	0.458 1
MnasNet	0.128 0	0.208 6	0.158 6
ShuffleNet	0.138 6	0.394 2	0.205 0
消融ECA模块	0.325 6	0.651 4	0.434 1
消融HFF-Head	0.365 2	0.636 7	0.464 1
消融DTA模块	0.398 5	0.665 2	0.498 4
ECA-HFF-DTA	0.485 8	0.714 5	0.578 3

Head和DTA机制均对模型性能提升具有显著贡献。

根据消融实验结果分析，所提出的3项改进，即ECA模块、HFF-Head与DTA模块均对模型性能提升产生了显著且互补的贡献。首先，从模块的单独作用来看，任意一个模块的缺失都会导致性能明显下降。例如，消融ECA模块后F1值降至0.434 1，消融HFF-Head后为0.464 1，消融DTA模块后为0.498 4，均显著低于ECA-HFF-DTA模型，这表明每个模块都是整体性能提升的重要一环。其次，3个模块之间存在明显的协同效应，单独使用某一模块或仅组合其中两者时，均未达到三者结合的最佳效果。这证明了提出的改进策略在整体上的有效性，通过模块间的有机结合，实现了识别性能的提升。

4　异体字部件识别应用分析

4.1　异体字部件识别结果分析

通过对5 272个异体字的系统化测试发现，模型在简单结构字符识别中表现优异，而在复杂嵌套结构识别中则面临挑战，这一现象揭示了字符结构复杂度与识别精度之间的内在关联规律。

对12个拆分部件较少、结构简单的异体字拆分结果进行展示（见表5）。可以发现，当异体字部件数量较少且空间拓扑关系明确时，如左右结构的图像3520拆分为[‘禾’，‘方’]、上下结构的图像2281拆分为[‘木’，‘火’]，模型识别准确率较好，能够准确识别出异体字的部件，基本无遗漏部件且识别出的错误部件极少。这得益于这些简单异体字大多为离散型结构、部件边界清晰，模型能有效捕捉局部特征和边界特征，预测效果较好。

模型在识别复杂异体字时的表现如表6所示，随着部件数量增加和嵌套层级加深，模型识别性能有所下降。可以观察出，在复杂结构的干扰下，模型虽然能够基本正确完成部件拆分，但同时产生了许多干扰项（如针对图像4852能够识别出异体字包含的[‘口’，‘又’，‘言’，‘了’，‘亠’]部件，但也识别出了12个其他部件），识别的准确率明显下降。这种衰减主要源于三重效应：①部件间遮挡导致边界模糊；②复合部件内部层级关系复杂化，如嵌套结构等导致部件边界难以识别；③笔画密度增加引发特征混淆。

通过对异体字预测结果的分析，可以得出以下规律。①结构复杂度与识别精度呈显著负相关。模型对简单结构异体字表现出较强的识别能力，这表明当部件空间位置明确、拓扑关系简单时，模型能有效捕捉部件边界特征。然而，对于复杂嵌套结构（部件数量≥6个），准确率显著降低，反映出模型对深层嵌套结构的解析能力不足。②部件粒度认知差异导致识

表5　简单结构异体字识别结果

图像编号	图像	拆分结果	图像编号	图像	拆分结果
0036		[‘口’，‘忄’，‘天’]	2281		[‘木’，‘火’]
0106		[‘忄’，‘刀’]	3164		[‘火’，‘羽’，‘彡’]
0857		[‘口’，‘又’，‘夂’]	3253		[‘口’，‘土’，‘夂’]
0986		[‘田’，‘耒’，‘巛’]	3299		[‘示’，‘水’]
1002		[‘亻’，‘女’]	3520		[‘禾’，‘方’]
2049		[‘彡’，‘耳’]	4110		[‘王’，‘人’，‘夂’]

表6　复杂结构异体字识别结果

图像编号	图像	拆分结果
0100		[‘口’，‘艹’，‘八’，‘二’，‘冂’，‘十’，‘田’，‘冖’，‘虫’，‘罒’，‘且’，‘中’，‘虍’，‘自’]
0130		[‘丷’，‘讠’，‘八’，‘二’，‘冂’，‘亠’，‘火’，‘心’，‘乂’，‘小’，‘灬’，‘羽’，‘川’，‘巛’]
0591		[‘口’，‘丷’，‘讠’，‘八’，‘二’，‘冂’，‘冖’，‘阝’，‘火’，‘心’，‘讠’，‘灬’，‘罒’，‘爫’，‘言’]
1287		[‘口’，‘丷’，‘讠’，‘八’，‘冂’，‘十’，‘亠’，‘田’，‘冖’，‘虫’，‘攵’，‘小’，‘豕’，‘糸’，‘母’]
2135		[‘口’，‘土’，‘丷’，‘八’，‘二’，‘冂’，‘十’，‘亠’，‘冖’，‘人’，‘士’，‘且’，‘中’，‘自’，‘言’，‘青’]
3339		[‘口’，‘艹’，‘土’，‘八’，‘十’，‘冖’，‘勹’，‘彡’，‘罒’，‘士’，‘且’，‘糸’，‘言’，‘贝’]
4493		[‘丷’，‘讠’，‘八’，‘二’，‘冂’，‘亠’，‘田’，‘冖’，‘小’，‘酉’，‘比’，‘糸’，‘豆’，‘川’，‘母’，‘曲’]
4852		[‘口’，‘讠’，‘二’，‘冂’，‘亠’，‘冖’，‘又’，‘厶’，‘刂’，‘讠’，‘子’，‘彡’，‘罒’，‘幺’，‘言’，‘矛’，‘了’]

别偏差。模型倾向于原子级拆分（如将“言”分解为“亠”“口”），而实际应用中更偏向整体拆分。这种认知差异在复杂字中尤为显著，说明模型仍须提高对部件功能层级的理解能力，实现视觉特征提取与语言学结构对齐。未来可引入结构约束机制（如部件空间关系图）和层级注意力机制，以提升对复合部件的整体感知能力。

4.2　基于部件查询异体字的模型应用分析

本研究提出的ECA-HFF-DTA模型及部件解构范式不仅在技术上实现了古籍异体字部件识别性能的提升，更在理论、方法与应用3个层面进行了系统性探索。

在理论方面，推动了古籍数字化从整字映射到部件解构的范式转移。传统范式受限于封闭字符集（如Unicode）的编码规则，试图为每一个异体字找到一个唯一的数字身份，这在理论上遇到了字符集容量有限、人工标注成本无限的根本性矛盾。本研究提出的部件解构范式，其核心理论创新在于将异体字视为由有限、标准化的部件通过空间关系组合而成的开放集合，这一转变具有以下学理意义。①解决了集外字的理论困境。通过将识别与检索的基本单元从整字降维至部件，使得任何异体字，无论其是否被字符集收录，只要其构成部件在基准库内，即可被识别和处理。②实现了文字学理论与计算机模型的认知对齐。汉字构形学认为，汉字是由基础部件按层次组合而成的。王宁[29]在汉语词源研究中强调，词源意义的探求需结合造词理据与历史文化背景，这与部件解构的认知逻辑具有内在一致性。本模型的多标签分类输出在形式上正是一种对构形成分的分解，使得计算机的识别过程更贴近文字学家对字形的分析和认知过程。

在方法方面，构建了针对异体字识别核心挑战的协同技术方案。本研究针对古籍异体字部件识别的特异性进行了一系列针对性设计，形成了协同增效的技术方案：通过ECA模块和HFF-Head的协同设计，模型实现了对异体字从微观笔画差异到宏观部件布局的全尺度特征捕捉。而DTA机制将类别不平衡问题与部件间的语义逻辑关系融入决策过程，不仅通过频率加权提升了对低频部件的召回率，更通过语义共现关系抑制了互斥部件的错误共存，显著提升了识别结果的合理性与准确性。

在应用方面，本研究为古籍智能化整理与检索提供了可落地的工具与功能。基于ECA-HFF-DTA模型可以实现精准部件级检索系统。如表7所示，用户可以通过指定一个或多个部件的组合，在古籍图像库中进行精确检索。例如，查询部件“虍”“力”“田”，可以唯一确定地找到异体字（编号3013）。这种组合定位的能力解决了以往因异体字字形生僻而无法输入、进而无法检索的根本性难题。

模型可以实现字形联想与知识发现，该系统支持不完整的部件查询。用户即使只记得一个字的某个部分，也能通过输入少量部件，检索到一系列相关的异体

表7　基于部件查询异体字结果

目标编号	异体字图像	查询部件	匹配异体字数量/个	结果编号
0505		‘艹’	552	
		‘艹’，‘角’	10	
		‘艹’，‘角’，‘刂’	1	0505
1188		‘火’	327	
		‘火’，‘彡’	5	
		‘火’，‘彡’，‘镸’	2	1188、0156
3013		‘虍’	147	
		‘虍’，‘力’	6	
		‘虍’，‘力’，‘田’	1	3013

字。这对于文字学研究和古籍整理极具价值：学者可以发现具有相同构形模式的异体字家族，从而分析字形的演变规律，或发现以往被忽略的字际关系。

模型可被集成到古籍扫描、图像处理的自动化流水线中，进行面向大规模数字化流水线的智能化处理。它能对海量古籍图像进行批量的、无需预知字符集的异体字部件标注，生成的结构化部件标签可以作为高级检索、知识图谱构建和数字化版本比对的基础数据，推动古籍文献从“图像库”向“结构化知识库”的升级。

5　结论与展望

本研究针对古籍数字化工程中的异体字部件识别难题，提出了一种基于改进EfficientNet的多标签分类方法，构建了ECA-HFF-DTA模型。通过引入ECA模块、HFF-Head和DTA机制，在保持计算效率的前提下，有效提升了汉字多部件识别的性能。在独立测试集上的实验表明，改进模型的F1值达到0.578 3，并优于包括EfficientNetV2-S在内的多个轻量级模型。本研究的创新模块（ECA、HFF-Head、DTA）具备通用性，未来可探索将其与更新、更强的基础网络结合，以期获得进一步的性能提升。

本研究不仅在技术上提出了ECA-HFF-DTA模型，更在方法论上实现了从整字映射到部件解构的范式转移。通过ECA、HFF-Head、DTA 3项核心技术的协同，有效应对了异体字部件识别中的尺度差异、形似、类别不平衡与逻辑约束等挑战。在此基础上，构建了基于部件组合的精准检索与联想查询系统，为突破古籍数字化中的字符集瓶颈、实现细粒度知识发现提供了可行的系统性解决方案，为文化遗产数字化提供了高效可靠的技术工具。未来可通过多模态知识融合与轻量化技术迭代，推动深度学习等技术在古籍整理、字库设计等领域的广泛应用，为中华传统文字文化的传承与创新作出贡献。

本研究也存在一些不足之处。尽管本研究构建的数据集在样本数量和类别上已具备一定规模，并采用了严格的数据清洗与增强策略，但仍存在一定的局限性，这主要源于古籍文献本身的复杂性。首先，在时间跨度上，数据集未能系统性地覆盖从先秦篆隶到明清楷书的所有重要汉字演变阶段。其次，在书写风格上，数据集以官方刻本的楷书、宋体为主，对于古籍中广泛存在的手写体、行书、草书等风格的异体字覆盖不足。这些局限性意味着，在面对书写风格差异巨大，或结构特征与训练集不同的异体字时，当前训练的ECA-HFF-DTA模型的泛化能力可能面临挑战。后续工作将通过引入更多元、跨朝代、多风格的古籍数据增强模型的普适性。

参考文献

[1] 王东海. 古文献数字语料库的异形字处理[J]. 语言文字应用，2005（4）：116-120.

[2] 高天俊. Unicode标准下古籍数字化的异体字处理[J]. 现代语文（语言研究版），2011（9）：114-115.

[3] 尉迟治平. 汉语信息处理和计算机辅助汉语史研究[J]. 语言研究，2004，24（3）：7-10.

[4] 尹琼. 古籍数字化中异体字认同问题的处理原则和标准：以法国国家图书馆藏刘凝批注本《说文长笺》为例[C]//中国文字发展论坛（2024）论文集. 安阳：中国文字博物馆，2024：96-106.

[5] 张健. 字料库视角下古籍数字化的字形处理标准[J]. 中国出版，2021（22）：55-59.

[6] 张叶. 古籍整理出版中异体字的编辑处理[J]. 文化产业，2024（25）：46-48.

[7] 裘锡圭. 文字学概要[M]. 北京：商务印书馆，1988.

[8] 李运富. 关于“异体字”的几个问题[J]. 语言文字应用，2006（1）：71-78.

[9] 史建华. 《尔雅音义》异体字研究[J]. 汉字文化，2025（1）：79-81.

[10] 尉迟治平. 电子古籍的异体字处理研究：以电子《广韵》为例[J]. 语言研究，2007，27（3）：118-122.

[11] 刘又辛. 谈谈假借字、异体字、古今字和本字[J]. 西南师范大学学报（人文社会科学版），1984，10（2）：42-48，64.

[12] 刘延玲. 近五十年来异体字研究与整理状况综述（上）[J]. 辞书

研究，2001（5）：35-44.

[13] 章杰鑫，潘悟云. 古籍数字化技术的新思路[J]. 语言研究，2014，34（1）：124-126.

[14] 张涌泉. 敦煌文书类化字研究[J]. 敦煌研究，1995（4）：71-79.

[15] 董月凯. 汉字编码论析[C]//中国语文现代化学会第七次学术会议暨换届大会论文集. 天津：天津师范大学国际教育交流学院，2006：117-126.

[16] 雷华，王祝英. 大国大典盛世丰碑：《汉语大字典》编纂、修订、出版历程侧记[J]. 辞书研究，2011（5）：57-67.

[17] BOUTELL M R，LUO J B，SHEN X P，et al. Learning multi-label scene classification[J]. Pattern Recognition，2004，37（9）：1757-1771.

[18] TSOUMAKAS G，KATAKIS I. Multi-label classification：an overview[M]//Data Warehousing and Mining. Geneva：IGI Global，2008：64-74.

[19] CLARE A，CLARE A，KING R D，et al. Knowledge discovery in multi-label phenotype data[C]//Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery. Heidelberg：Springer Berlin，2001：42-53.

[20] ELISSEEFF A，WESTON J. A kernel method for multi-labelled classification[M]//DIETTERICH T G，BECKER S，GHAHRAMANI Z. Advances in Neural Information Processing Systems 14. Cambridge：The MIT Press，2002：681-688.

[21] WEI Y C，XIA W，LIN M，et al. HCP：a flexible CNN framework for multi-label image classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2016，38（9）：1901-1907.

[22] 薛丽霞，江迪，汪荣贵，等. 融合注意力机制和语义关联性的多标签图像分类[J]. 光电工程，2019，46（9）：180468.

[23] WANG J，YANG Y，MAO J H，et al. CNN-RNN：a unified framework for multi-label image classification[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition（CVPR）. New York：IEEE Press，2016：2285-2294.

[24] 张辉宜，张进，黄俊. 基于图注意力网络的多标签图像分类模型[J]. 重庆工商大学学报（自然科学版），2022，39（1）：34-41.

[25] 张晓瑞. 基于卷积神经网络的多标签图像分类识别算法研究[J]. 通化师范学院学报，2022，43（2）：75-82.

[26] 汪瑞，武芳宇，张百灵. 基于提示学习和注意力机制的多标签图像分类方法[J]. 软件工程，2024，27（7）：42-46.

[27] 武红鑫，韩萌，陈志强，等. 监督和半监督学习下的多标签分类综述[J]. 计算机科学，2022，49（8）：12-25.

[28] RIDNIK T，BEN-BARUCH E，ZAMIR N，et al. Asymmetric loss for multi-label classification[C]//2021 IEEE/CVF International Conference on Computer Vision（ICCV）. New York：IEEE Press，2022：82-91.

[29] 王宁. 汉语词源的探求与阐释[J]. 中国社会科学，1995（2）：167-178.

作者简介

郑澄璐，女，硕士研究生，研究方向：数字人文。
王昊，男，博士，教授，博士生导师，通信作者，研究方向：数字人文、知识组织，E-mail：ywhaowang@nju.edu.cn。
孟序阳，男，博士研究生，研究方向：知识组织、数字人文。
杨思敏，女，博士研究生，研究方向：人文计算、信息分析。

Component Recognition of Variant Characters in Ancient Books Based on Improved Image Multi-Label Classification

ZHENG ChengLu[1,2] WANG Hao[1,2] MENG XuYang[3] YANG SiMin[1,2]
(1. School of Information Management, Nanjing University, Nanjing 210023, P. R. China; 2. Key Laboratory of Data Engineering and Knowledge Services in Jiangsu Provincial Universities (Nanjing University), Nanjing 210023, P. R. China; 3. School of Information Resources Management, Renmin University of China, Beijing 100872, P. R. China)

Abstract: In response to the problem that many variant characters in ancient books are stored in images and cannot be directly retrieved digitally, this study proposes an ECA-HFF-DTA model based on a multi-label classification method for recognizing variant characters in ancient books. It includes three modules: enhanced channel attention (ECA) module, hierarchical feature fusion classification head (HFF-Head), and dynamic threshold adjustment (DTA) mechanism, which automatically identify their constituent components and generate multi-label classification results. Experiments show that the model achieves an F1-score of 0.578 3 on the test set, demonstrating a significant improvement over the baseline, and can realize the application function of querying variant characters based on components. However, the recognition performance of the model for complex nested structures of variant characters still needs to be improved. The ECA-HFF-DTA model improves the component recognition effect of variant characters in ancient books, providing ideas for the intelligent sorting and retrieval of variant characters in large-scale ancient literature.

Keywords: Ancient Book; Variant Character Recognition; Multi-Label Classification

（责任编辑：常春）