# 融合图像与文本识别的民国纺织文献数字化修复研究

吴迪
西安工程大学

摘要：民国时期纺织文献是研究中国近代纺织工业发展的珍贵史料，但因年代久远普遍存在纸张老化、墨迹褪色、污渍破损等退化问题，严重影响文献的可读性和保存价值。本研究提出一种融合深度学习图像增强与改进OCR文本识别的数字化修复方法，通过构建基于生成对抗网络的图像修复模型、自适应增强算法和优化的文本识别系统，实现对退化文献的高质量数字化处理。实验结果表明，该方法在图像质量提升和文本识别准确率方面均优于传统方法，PSNR值提升12.3%，文本识别准确率达到94.7%，为民国纺织文献的数字化保护提供了有效的技术方案。

关键词：民国纺织文献；图像增强；深度学习；OCR识别；数字化修

## 一、引言

民国时期（1912-1949）是中国纺织工业从传统向现代转型的关键阶段，该时期产生的纺织技术文献、行业报告、企业档案等历史资料，详细记录了纺织工艺改良、机器引进、产业布局等重要信息，对研究中国近代工业史、经济史具有不可替代的价值。然而，经过近百年的时间侵蚀，这些文献普遍面临严重的物理退化问题：纸张因酸化而脆化泛黄，墨迹因氧化而褪色模糊，污渍、霉斑、虫蛀、破损等病害广泛存在，部分文献甚至面临消失的危险。

传统的文献保护方法主要依靠物理修复和扫描存档，但物理修复成本高昂、耗时长且需要专业技师，而简单的扫描存档虽能保留图像信息，却难以解决图像质量差、文本难以识别的问题，无法满足数字化时代学术研究和知识传播的需求。近年来，随着计算机视觉和人工智能技术的快速发展，基于深度学习的图像处理和文本识别技术为历史文献的数字化修复提供了新的解决方案。特别是生成对抗网络（GAN）在图像修复领域的突破，以及深度学习OCR模型在复杂场景文本识别中的优异表现，使得对退化文献进行智能化、自动化的数字修复成为可能。

本研究针对民国纺织文献的退化特征，提出一种融合图像增强与文本识别的数字化修复框架，通过深度学习技术实现对退化图像的智能修复和高精度文本提取，为民国纺织文献的数字化保护和利用提供技术支撑，同时为其他类型历史文献的数字化修复提供借鉴。

## 二、民国纺织文献退化特征与技术需求分析

### 2.1 退化特征分类

通过对收藏于国家图书馆、上海图书馆等机构的200余份民国纺织文献样本进行调研分析，归纳出以下典型退化特征：

（1）纸张老化退化：民国时期多采用机制木浆纸，酸性物质导致纸张纤维断裂，表现为纸张泛黄、脆化，背景呈现不均匀的褐色或黄色，严重影响图像对比度。统计显示，85%以上的文献存在明显纸张老化现象。

（2）墨迹褪色模糊：铁胆墨等传统墨水因氧化作用导致字迹颜色变淡、边缘模糊，与背景对比度降低。部分文献的文字几乎难以辨认，给文本识别带来极大挑战。

（3）污渍与噪声干扰：长期保存过程中产生的霉斑、水渍、油污等污渍广泛分布于文献表面，形成大小不一的暗色斑点。此外，扫描过程引入的椒盐噪声也影响图像质量。

（4）物理破损缺失：虫蛀、撕裂、缺角等物理损伤导致文献信息缺失，部分区域出现孔洞或边缘残缺，需要进行图像修补。

（5）版面复杂多样：民国纺织文献包含竖排繁体文字、手绘图表、印章标注等多种元素，版面布局复杂，增加了文本定位和识别的难度。

### 2.2 技术需求分析

针对上述退化特征，数字化修复系统需要满足以下技术要求：一是能够有效去除背景噪声和污渍，恢复纸张原有的白色或浅色背景；二是增强文字与背景的对比度，使模糊褪色的文字清晰可辨；三是修复破损区域，重建缺失的图像信息；四是准确识别竖排繁体文字，处理复杂版面结构；五是保持原始文献的历史特征，避免过度处理导致信息失真。这些需求构成了本研究技术框架设计的基础。

## 三、融合图像增强与文本识别的技术框架

### 3.1 总体架构设计

本研究提出的数字化修复框架采用模块化设计，包括预处理模块、图像增强模块、文本识别模块和后处理优化模块四个核心部分，各模块功能如下：

预处理模块负责对扫描获得的原始图像进行初步处理，包括几何校正、去噪和归一化。采用透视变换算法校正扫描过程中产生的倾斜和变形，使用改进的中值滤波去除椒盐噪声，并将图像统一调整为标准分辨率（300 DPI），为后续处理奠定基础。

图像增强模块是系统的核心，采用基于生成对抗网络（GAN）的深度学习模型，实现对退化图像的智能修复。该模块分为两个子网络：生成网络负责学习从退化图像到高质量图像的映射关系，重建破损区域、去除污渍、增强对比度；判别网络则评估生成图像的真实性，通过对抗训练不断优化生成效果。同时引入自适应增强算法，针对不同退化程度的文献动态调整处理参数。文本识别模块基于改进的CRNN（卷积循环神经网络）模型，实现端到端的文本检测与识别。该模块首先利用文本检测网络定位文献中的文字区域，包括竖排、横排、印章等不同类型的文本；然后通过识别网络将文本图像转换为文字序列。针对繁体字和古旧字形，构建专门的字符集和训练数据集，显著提升识别准确率。

后处理优化模块对识别结果进行校正和验证。利用语言模型对识别文本进行上下文分析，纠正明显的识别错误；结合纺织专业词库进行术语匹配，提高专业名词的识别准确性；最终输出结构化的文本数据和修复后的高清图像。

3.2 关键技术实现

（1）基于生成对抗网络的图像修复

本研究设计的GAN模型采用U-Net架构作为生成器，该架构通过编码器-解码器结构实现多尺度特征提取，跳跃连接机制有效保留图像细节信息。编码器由5层卷积层组成，逐步提取从低级到高级的图像特征；解码器通过上采样和卷积操作重建图像，跳跃连接将编码器的特征图传递到解码器对应层，避免信息损失。

判别器采用PatchGAN结构，将图像划分为多个局部区域分别判别真伪，相比全局判别能更好地关注图像细节。损失函数设计为对抗损失、内容损失和感知损失的加权组合：对抗损失促使生成图像逼近真实分布；内容损失（L1范数）约束像素级相似度；感知损失基于预训练VGG网络提取的高层特征，保证语义一致性。训练过程采用交替优化策略，使用Adam优化器，学习率设为0.0002，batch size为16。

（2）自适应图像增强算法

考虑到不同文献的退化程度差异显著，本研究提出自适应增强策略。首先通过图像质量评估模型（基于无参考图像质量评价方法）自动分析输入图像的退化类型和严重程度，将文献分为轻度、中度、重度退化三个等级。然后根据评估结果动态调整GAN模型的生成强度和后续增强参数。

对于轻度退化文献，采用温和的对比度增强和去噪处理；中度退化文献增加污渍去除和边缘锐化；重度退化文献则启用完整的GAN修复流程，并加强局部细节重建。自适应策略通过引入反馈机制，根据文本识别模块的初步识别效果自动微调参数，形成闭环优化。

（3）改进的深度学习OCR模型

文本识别模块采用CRNN架构，由卷积层、循环层和转录层三部分组成。卷积层使用ResNet-34作为骨干网络，提取文本图像的视觉特征，生成特征序列；循环层采用双向LSTM（长短期记忆网络），捕捉序列中的上下文依赖关系；转录层使用CTC（Connectionist Temporal Classification）损失函数，实现不定长序列的对齐和解码。

文本识别模块采用CRNN架构，由卷积层、循环层和转录层三部分组成。卷积层使用ResNet-34作为骨干网络，提取文本图像的视觉特征，生成特征序列；循环层采用双向LSTM（长短期记忆网络），捕捉序列中的上下文依赖关系；转录层使用CTC（Connectionist Temporal Classification）损失函数，实现不定长序列的对齐和解码。

（4）文本校正与验证机制

为进一步提高识别准确性，建立基于规则和统计的双重验证机制。规则验证基于民国文献的语言特点和版面规律，检查识别结果是否符合繁体字语法、标点使用习惯等；统计验证利用N-gram语言模型，计算文本序列的概率，识别低概率组合（可能的识别错误），并结合字形相似度进行候选字替换。对于置信度较低的识别结果，系统自动标注并提示人工复核，实现人机协同的质量控制。

四、实验与结果分析

4.1 实验设置

数据集构建：从国家图书馆、上海图书馆收集民国纺织文献扫描件500份，按7:2:1比例划分为训练集、验证集和测试集。同时构建合成退化数据集，通过在现代清晰文献图像上模拟老化、污渍、噪声等效果，生成3000张训练样本，解决真实退化文献样本不足的问题。评价指标：图像增强效果采用峰值信噪比（PSNR）、结构相似度（SSIM）和视觉质量评分进行评估；文本识别准确率采用字符准确率（Character Accuracy）和字符串准确率（String Accuracy）衡量；综合性能通过F-Measure指标反映。

对比方法：选择传统图像增强方法（直方图均衡化、双边滤波）、经典OCR系统（Tesseract、ABBYY FineReader）以及近期深度学习方法（pix2pix、DocUNet）作为基线进行对比。

4.2 实验结果

图像增强效果评估：在测试集上，本研究方法的平均PSNR值达到28.45 dB，相比传统直方图均衡化方法（25.32 dB）提升12.3%，比pix2pix（26.78 dB）提升6.2%。SSIM指标达到0.912，表明修复图像与参考图像在结构上高度相似。视觉质量评分由专业人员进行盲测，采用5分制，本方法平均得分4.3分，明显优于其他方法（传统方法3.1分，pix2pix 3.8分）。针对不同退化类型的处理效果，实验表明本方法对纸张老化和墨迹褪色的修复效果最佳，PSNR提升幅度分别达到14.2%和15.8%；对污渍去除也有较好表现，但对严重破损缺失区域的重建仍存在一定挑战，部分细节恢复不够理想。

文本识别准确率评估：在经过图像增强的测试集上，改进OCR模型的字符准确率达到94.7%，字符串准确率达到87.3%，分别比Tesseract OCR（78.2%，62.5%）提升21.1%和39.7%，比ABBYY FineReader（85.6%，75.8%）提升10.6%和15.2%。对于竖排繁体文字，本方法准确率达到93.2%，显著优于通用OCR系统。处理效率评估：在配置NVIDIA RTX 3090 GPU的工作站上，单页文献（A4大小，300 DPI）的完整处理时间约为8.5秒，其中图像增强5.2秒，文本识别3.3秒，满足批量数字化的效率要求。

4.3 案例分析

选取一份1935年《中国纺织建设月刊》作为典型案例进行分析。该文献存在严重纸张泛黄（背景灰度值仅180）、墨迹褪色（文字与背景对比度低至1.8）以及多处水渍污染。传统方法处理后虽然对比度有所改善，但噪声放大明显，部分文字仍难以辨认，OCR识别准确率仅68.5%。

采用本研究方法后，图像背景恢复为接近白色（灰度值230），文字清晰度显著提升（对比度达到5.2），水渍污染基本去除，视觉效果接近原始印刷质量。文本识别准确率提升至92.8%，专业术语如"纺锭产能""双丝光工艺""经纬密度"等均被正确识别，为后续的文献数字化利用奠定了基础。

五、结论与展望

本研究针对民国纺织文献的退化特征和数字化保护需求，提出了一种融合深度学习图像增强与改进OCR文本识别的数字化修复方法。通过构建基于生成对抗网络的图像修复模型、自适应增强算法和专门优化的文本识别系统，实现了对退化文献的高质量数字化处理。实验结果表明，该方法在图像质量提升和文本识别准确率方面均显著优于传统方法，为民国纺织文献的数字化保护提供了有效的技术方案。

然而，本研究仍存在一些不足：一是对于极度破损和信息严重缺失的文献，现有模型的修复效果有限；二是训练数据集规模相对较小，模型的泛化能力仍有提升空间；三是系统尚未充分考虑文献中的图表、印章等非文本元素的处理。

未来研究可以从以下方向展开：首先，引入Transformer等更先进的深度学习架构，提升模型对复杂退化模式的学习能力；其次，扩大数据集规模，收集更多不同类型、不同退化程度的历史文献样本，增强模型的泛化性；再次，开发针对图表、印章等特殊元素的专门识别模块，实现文献信息的全面数字化；最后，探索多模态学习方法，结合文献的文本、图像、元数据等多维信息，构建更加智能的历史文献数字化修复与知识提取系统。本研究的技术框架和方法也可推广应用于其他类型历史文献的数字化保护，为文化遗产的数字化传承提供技术支撑。

参考文献

[1] SikuBERT与SikuRoBERTa：面向数字人文的《四库全书》预训练模型构建及应用研究

[2] 一种双判别器GAN的古彝文字符修复方法[J]. 陈善雄;朱世宇;熊海灵;赵富佳;王定旺;刘云.自动化学报,2022(03)

[3] 数字化背景下纸质图书与电子图书的对比分析[J]. 张悦洁.产业与科技论坛,2019(15)

[4] 图像修复的CDD模型新算法[J]. 王军锋;裴艳侠;王涛.计算机系统应用,2016(08)

[5] 改进优先级的分步匹配图像修复算法[J]. 朱晓临;王传奇;范承凯.图学学报,2015(03)