doi:10.3772/j.issn.2095-915x.2025.04.010

RRO-ChatData 的中文期刊版面分析方法

王朝阳　李超　郭红梅

中国科学技术信息研究所 北京 100038

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 大语言模型技术在图书馆智慧服务的创新应用，不仅对期刊语料提出了更高的要求，其技术本身也具备参与学术论文数字化加工的能力。以中文期刊为例对学术论文全文语料优化与文献关键信息提取方法进行了探究。[方法 / 过程] 提出一种 RRO-ChatData 方法，通过阅读顺序还原方法对不同版面结构的中文期刊论文 PDF 文件按照人类阅读顺序进行复原，并针对中文期刊资源特点，基于大语言模型技术构建 ChatData 特征识别方法进行关键信息识别。[局限] 由于资助信息内容相对复杂且表述方式不固定，因此该方法对于资助信息的识别准确率较低。[结果 / 结论] RRO-ChatData 方法能够准确对学术论文进行阅读顺序还原与内容提取，有效提升中文期刊全文语料的质量与数字化加工过程中关键信息提取的效率。

关键词：大语言模型；阅读顺序还原；学术论文；关键信息识别；ChatData

中图分类号：G250.7；G35

RRO ChatData Chinese Journal Layout Analysis Method

WANG Zhaoyang　LI Chao　GUO Hongmei

Institute of Scientific and Technical Information of China, Beijing 100038, China

Abstract: [Objective/Significance] The innovative application of large language modeling technology in intelligent library services not only puts forward higher requirements for journal corpus, but also has the ability to participate in the digital processing of academic papers. This article takes Chinese journals as an example to explore the optimization of full-text corpus and the extraction of key information from academic papers. [Methods/Processes] This article proposes an RRO-ChatData method, which uses the Restoration of Reading Order (RRO) method to restore PDF files of Chinese journal articles with different layout structures according to human reading order. Based on the characteristics of Chinese journal resources, a ChatData feature recognition method is constructed using large language model technology for key information recognition. [Limitations] Due to the relatively complex content and unstable expression of funding information, the accuracy of this method to identify funding

基金项目　中国科学技术信息研究所创新研究基金青年项目“AI 驱动下专业图书馆资源服务一体化创新模式研究”（QN2025-10）；国家重点研发计划项目“科技文献内容深度挖掘及智能分析关键技术和软件”（2022YFF0711900）。

作者简介　王朝阳（1996-），硕士，助理研究员，主要研究方向为信息资源建设、智能化科技情报挖掘；李超（1988-），学士，工程师，主要研究方向为软件工程，E-mail：chaoli1346@foxmail.com；郭红梅（1985-），博士，副研究馆员，主要研究方向为文本挖掘、机构规范文档构建和资源聚合分析。

引用格式　王朝阳 , 李超 , 郭红梅 . RRO-ChatData 的中文期刊版面分析方法 [J]. 情报工程 , 2025, 11(4): 115-127.

doi:10.3772/j.issn.2095-915x.2025.04.010

information is relatively low. [Results/Conclusions] The RRO ChatData method can accurately restore the reading order and extract content of academic papers, effectively improving the quality of Chinese journal full-text corpus and the efficiency of key information extraction in digital processing.

Keywords: Large Language Model; Restoration of Reading Order; Academic Papers; Key Information Identification; ChatData

引言

版面分析技术通过检测和定位非结构化文档或文档图像中的标题、段落、表格、插图等页面元素，实现对其分类与结构化处理[1]，是中文期刊数字化加工的核心环节。传统图书馆的论文数字化过程依赖人工操作，在经费紧缩的背景下，效率低、成本高的问题成为制约发展的瓶颈[2]，亟须探索更高效、经济的数字化加工方法以满足需求。当前中文期刊论文的正文排版形式多样且复杂，分栏布局、图文混排[3]等设计导致机器自动解析时易出现段落错位、文本拼接错误等问题。2023 年，以 ChatGPT 为代表的一系列大语言模型的出现标志着自然语言处理技术进入了一个新的阶段，它们推动着问答服务、数据分析、信息查询与知识组织等服务的智慧化发展，并促进着图书馆知识服务的迭代与创新[4]，若资源数字化加工精度不足，文本噪声或错误将直接影响后续智能服务的可靠性。同时，大语言模型技术在中文期刊知识服务中的创新应用，不仅对期刊语料提出了更高的要求，其技术本身强大的推理能力也能够参与科技文献加工。本文基于段落还原方法与大语言模型对 YOLOv8 模型进行优化，提出一种基于 RRO-ChatData 的中文期刊版面分析方法，通过阅读顺序还原（Restoration of Reading Order，RRO）技术对 YOLOv8 模型识别后的期刊论文按照人类阅读顺序进行优化，最大程度还原文章结构与段落关系，并针对中文期刊论文资源特点，利用大语言模型推理能力构建 ChatData 中文期刊关键信息识别方法，识别并提取论文中的目标实体，提高中文期刊数字化加工的质量与效率。

1 相关研究

中文期刊文献多数以便携文档格式（Portable Document Format，PDF）的文档格式进行存储[5]。传统光学字符识别（Optical Character Recognition，OCR）可以实现文档的文本化[6-7]，通常采用自上而下和自底向上的策略[8]，前者将文档图像逐步分割成较小的区域，而后者将较小的元素合并成较大的图像区域。早期研究主要采用基于启发式规则的 OCR 技术，这类方法依赖手工特征工程，其低维特征表示难以应对复杂版面布局。尽管在标准文档处理中表现稳定，但在学术论文这类包含多栏排版、数学公式和混合语义结构的场景中，其识别准确率显著下降。伴随人工智能技术的不断发展，卷积神经网络（Convolutional Neural Networks，CNN）的引入显著提升了版面分析的准确性以及特征提取能力，在通用文档图像处理中取得

进展[9-14]。然而 CNN 有限的感受使其难以捕获全局上下文信息，且存在运算速度较慢并依赖大量标注数据等问题，为了解决这些缺陷，研究者们正在不断探索新的 CNN 架构和算法，但依旧很难突破其固有的局限性[15]。迁移学习的不断进步与预训练模型的共享和开放很好地解决了这一问题，迁移学习能够显著提高新任务的性能，并减少对新任务数据的依赖[16]。越来越多的学者使用 YOLO 系列模型参与到版面分析的相关研究[17-18]，并取得了显著的成果，但在进行版面分析时依旧存在段落拼接错误的情况，导致上下文语义关系损失。如今，大语言模型凭借其强大的推理能力[19]，可以在 OCR 识别中根据特定的业务需求，生成符合要求的规则化结果，从而更好地满足实际应用的需求。微软的 LayoutLMv3 是文档智能化领域首个不需要预训练主干网络的多模态模型[20-21]，从而简化了模型结构，提高了运行效率，主要面向针对合同、票据等商务文件的版面识别。目前国内外相关研究，技术迭代与优化主要聚焦多模态特征识别提取、不规则布局处理以及性能优化等方向，对复杂版面的段落重组的研究比较少，对于判断段落合并的条件仍采用固定阈值的垂直间距规则及水平间距规则，对于不同版面结构的文档难以兼容，无法最大程度保留上下文语义关系，并且缺乏基于大语言模型的科研文献特征提取方法。

2 RRO-ChatData 方法

本文提出一种基于 RRO-ChatData 的中文期刊论文阅读顺序还原与关键信息识别方法，该方法适用于不同格式以及排版结构的中文期刊论文，主要包含以下流程，如图 1 所示。

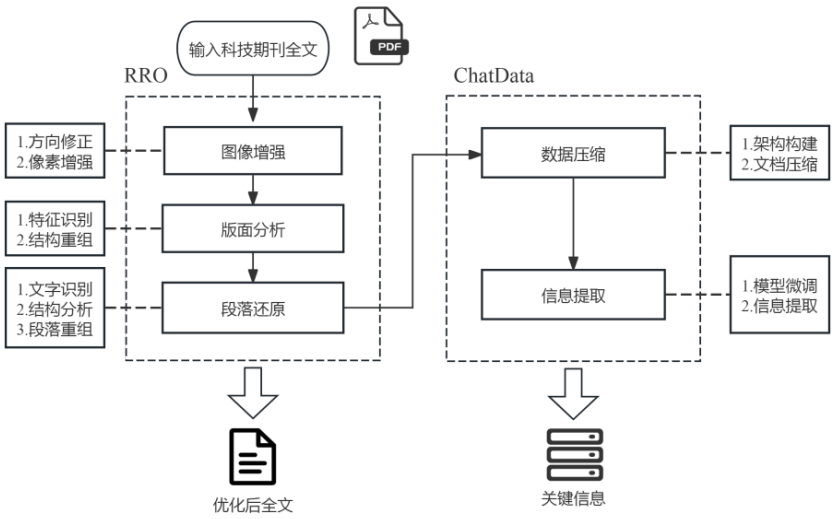


图 1　RRO-ChatData 方法流程图

（1）图像增强。首先对全文 PDF 进行预处理，转化为增强图像。

（2）版面分析。通过 YOLOv8 模型对预处理后的图像进行区域块识别，并进行分类。

（3）段落重组。通过总结归纳中文期刊段落结构特征，使用 RRO 方法对版面分析结果按

doi:10.3772/j.issn.2095-915x.2025.04.010

照人类阅读习惯对论文正文段落进行优化。

（4）数据压缩。对科技文献资源特点进行总结，形成中文期刊文献主体架构，并按照该架构进行数据压缩，提高大模型对于关键信息识别的准确性与计算速度，减少算力消耗并减轻模型幻觉。

（5）信息提取。依托（4）中压缩后的数据设计提示工程（Prompt Engineering），并进行微调，调用通义千问开源大模型进行关键信息提取。

3 RRO 阅读顺序还原方法

中文期刊论文中的文本布局和格式往往较为复杂，包括多种字体、字号、段落缩进、列表等。OCR 系统在处理这些复杂布局时，可能无法准确识别段落之间的分隔符（如换行符、段落标记等），在面对双栏等复杂排版格式时无法自动判断，会导致段落识别错误，不利于大语言模型进行理解。RRO 方法首先将 PDF 文件转换为 JPG 格式图片，并进行图像增强处理以提升识别效果。利用 YOLOv8 算法对文献中的文字、图片、公式、表格等区域进行标注和分类，通过深度学习和图像处理技术精确识别并优化边界框，根据识别对象的 X 坐标进行聚类分析，识别文献的版面布局，并据此恢复内容的正常阅读顺序。相较于固定阈值的垂直间距规则及水平间距规则合并方法，RRO 利用段落特征（如首行缩进、行对齐、行间距等）构建边界距离矩阵，采用动态阈值的方法应对不同的版面特征并确认段落关系，从而实现准确的段落合并，输出按照人类阅读顺序排列且段落划分准确的中文期刊论文 MarkDown 文件。

3.1 数据增强

首先对学术论文全文 PDF 文件进行分类，包括结构化 PDF 文件与非结构化 PDF 文件。结构化 PDF 文件具备完整的物理结构与逻辑结构 [22]，存储的文本可通过计算机直接读取，因此无需进行数据增强。对于图片形式 PDF 文件，首先将 PDF 文件统一转化为 JPG 格式，检测图片尺寸长宽比，并将图片统一旋转至正常阅读方向。使用 OPENCV[23] 开源库中的高斯函数对图像进行平滑处理，然后采用直方图均衡化（Histogram Equalization）函数通过拉伸图像的强度分布来提高图像的对比度并输出图像，完成数据增强 [24]。

3.2 版面分析

本文选用 YOLOv8 算法对中文期刊文献的各个区域进行标注与分类，涵盖文字、图片、公式、表格等多种类型。作为一阶段检测算法的最新代表，YOLOv8 继承了 YOLO 系列高效的特点，它省去了预先生成候选区域的步骤，直接通过深度卷积神经网络对整个输入图像进行特征提取，一次性输出目标的类别和位置信息，因此具有更快的检测速度，广泛应用于实时性要求较高的领域 [25]。首先，将待检测的中文期刊文献图像输入到已经完成预训练和参数调优的 YOLOv8 模型中。模型利用深度卷积神经网络对输入图像进行特征提取，生成一系列特征图，这些特征图捕捉了图像在不同尺度和位置上的纹理、颜色等特征信息。接下来，通过每个网格单元预测多个边界框的方式，即图

像被划分为 a × a 个网格单元，每个网格单元预测 B 个边界框。每个边界框包含 5 个预测值，分别是物体的中心坐标 (x, y)、高度 h、宽度 w 以及置信度 confidence。然后，对这些初步检测到的边界框进行非极大值抑制处理，根据置信度分数选择最佳的边界框，并删除与最佳边界框重叠程度较高的其他边界框，优化边界框的位置和置信度，提高检测的准确性和鲁棒性。最终，通过边界框的分类和置信度计算，输出检测结果，包括目标的位置、类别和置信度，识别类别包括：页眉、页脚、标题、正文、图片、图注、表格、表注等，根据 YOLOv8 算法处理后的文献区域切分结果如图 2 所示。

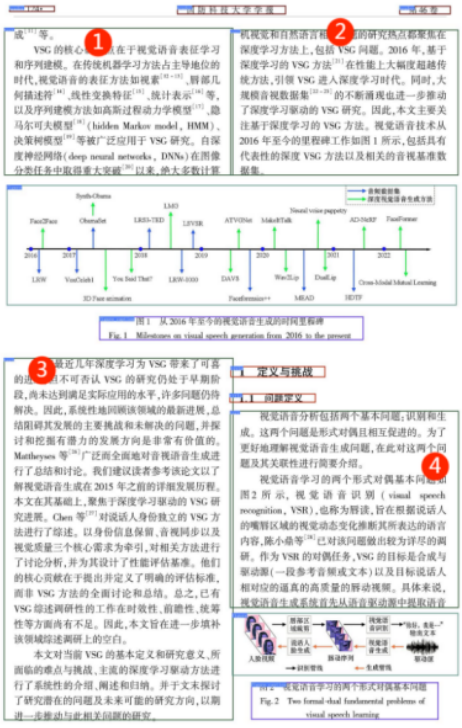


图 2　版面分析结果

3.3　结构重组

针对 3.2 中识别对象中心位置的 X 坐标进行 K-means 聚类，如果大部分对象的 X 坐标集中在同一区域内，且这些边界框的宽度接近或等于图像宽度，判定为单栏布局。如果边界框的 X 坐标分布在多个区域内，每个区域内识别对象的边界框宽度接近或相等，且宽度小于图像宽度的一半，或者存在多个明显的区域，则判定为多栏布局。人类的阅读顺序通常是从左到右、从上到下。在双栏排版中，读者会首先阅读左侧栏目的内容，然后移动到右侧栏目继续阅读。这种顺序有助于保持阅读的连贯性和流畅性。针对图像内识别对象按照 X 坐标聚类结果进行版面结构识别，并按照 Y 坐标自上而下重新排列并输出重组结果，按照人类阅读顺序复原。为更加准确地对段落进行还原，结构重组过程中会将文档的页眉页脚进行隐藏，在后续的论文关键信息识别环节中会进行恢复。

3.4　段落还原

使用 OPENCV 开源库中 OCR 工具对按照人类阅读顺序重组的中文期刊图像进行文字识别，逐行对文字划定边界框线，识别结果如图 3 所示。

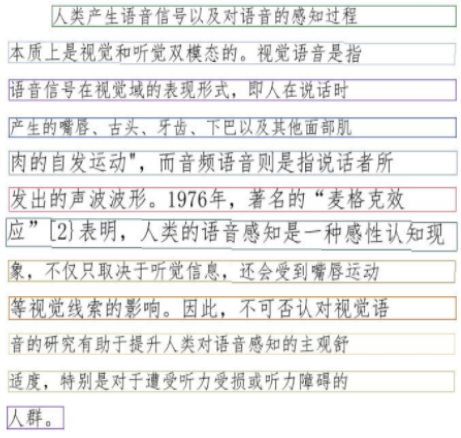


图 3　文本行切分结果

doi:10.3772/j.issn.2095-915x.2025.04.010

对于中文期刊论文而言，文档中的文本行均由若干字符组成，每个字符间的间距很小，不会大于字符宽度，段落中行与行的间隙也基本相同。采用二维直角坐标系来定位识别结果中各行文本的边界位置，设水平方向上有 m 个文本块，构建中文期刊文本块矩阵 S。

$$S=\begin{pmatrix} S_0 \\ S_1 \\ \cdots \\ S_m \end{pmatrix} \quad (1)$$

每一个文本块 S_m 的边界坐标为 $[A(X_m, Y_m), B(X_m, Y'_m), C(X'_m, Y_m), D(X'_m, Y'_m)]$，如图 4 所示。

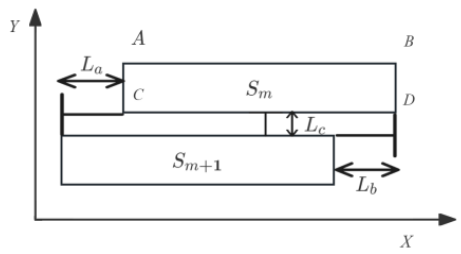


图 4　文本行边界示意图

计算每一个文本块 S_m 与 Y 轴负方向相邻文本块 S_{m+1} 的首距离 L_a，段尾距离 L_b 以及行间距离 L_c。

$$L_a=X_m-X_{m+1} \quad (2)$$

$$L_b=X'_m-X'_{m+1} \quad (3)$$

$$L_c=Y'_m-Y_{m+1} \quad (4)$$

根据计算结果生成每一个文本块的边界距离集合 $L_{n,m}(S_m, L_a, L_b, L_c)$，并生成边界距离矩阵 L。

$$L=\begin{pmatrix} L_0 \\ L_1 \\ \cdots \\ L_m \end{pmatrix} \quad (5)$$

中文期刊论文一般分为多个段落，每个段落的特征较为明显 [26]，如：首行缩进固定字符；除第一行外，段落中各行行首对齐；除最后一行外，各行行尾对齐；最后一行的行尾小于等于其上一行的行尾；段落内的行间距基本相同。根据边界距离矩阵 L 中段首距离 L_a，段尾距离 L_b 以及行间距离 L_c 的计算结果判断上下两段文本行的相对位置。为减小边界框线标注产生的误差，取段落中 L_a、L_b 以及 L_c 的最大值作为段落中两行首尾均对齐时的段首距离阈值 L_{Ea}，段尾距离阈值 L_{Eb} 以及行间距离阈值 L_{Ec}。通过对比距离阈值与各文本块的边界距离，当 $L_c>L_{Ec}$ 时，判定上下两行分为不同段落；当 $L_c<L_{Ec}$ 时，判定上下两行属于相同段落，上下两段文本行可能形成的结构如图 5 所示。

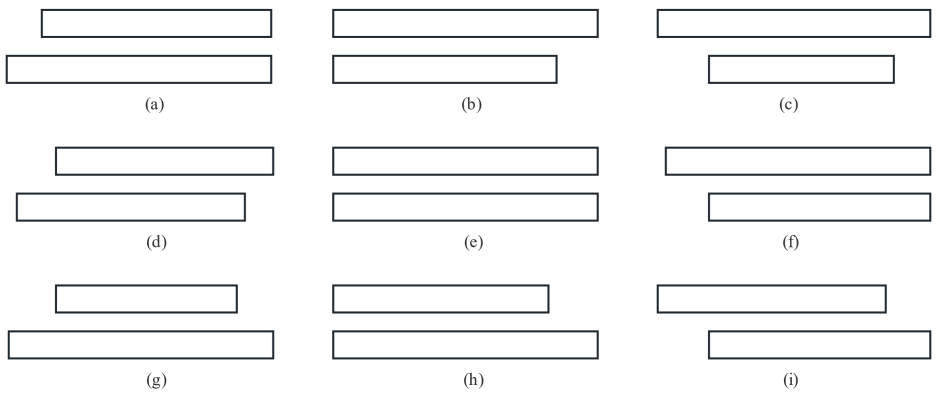


图 5　文本行相对位置

图 5 中的 9 种结构可以分成两类：

（1）分成不同段落

呈现结构（c）、（f）、（g）、（h）、（i）的上下文本行可以分成不同的段落，其中结构（c）、（f）、（i）中 $L_a < (-L_{Ea})$，即第二行的文本对于第一行文本有缩进的现象；结构（g）中 $L_a > L_{Ea}$ 且 $L_b < (-L_{Eb})$，第一行明显比第二行短，首尾均不对齐；结构（h）中 $L_b < (-L_{Eb})$，第一行行尾小于第二行行尾；以上结构的两个文本行判定为不同段落。

（2）判定同一段落

结构（a）、（b）、（d）、（e）可以判定为同一段落，其中结构（a）、（d）中 $L_a > L_{Ea}$，首行有明显缩进；结构（b）中 $L_b > 0$，第二行行尾小于第一行行尾，判定为段落尾部；结构（e）中 $|L_a| < L_{Ea}$ 且 $|L_b| < L_{Eb}$，上下文本行首尾相齐；以上结构的两个文本行判定为同一段落。

段落复原后将复原结果转化为 MARKDOWN 格式以及 JSON 格式，前者为科技文献大语言模型提供优质语料储备，后者便于后续使用 ChatData 关键信息抽取方法进行处理。

4 ChatData 关键信息抽取方法

传统的论文关键信息抽取方法确实存在泛化能力受限的问题，它们往往依赖于预先设定的复杂规则来识别并提取论文中的关键信息。然而，随着学术论文格式和内容的多样化，这些固定的规则很难适应所有的情况，一旦论文的格式或内容稍有变化，就可能需要重新设计或调整规则，极大地增加了工作的复杂性和成本。ChatData 依托大模型的文本理解能力，在提示工程的帮助下对特定区域内的文本内容进行语义理解，即使面对不同排版格式的论文也能够准确识别并抽取关键信息。

4.1 大语言模型文本理解原理

大语言模型在处理文章内容方面的能力得益于多种训练技术与方法的结合，包括预训练、指令微调（Instruction Fine-Tuning，IFT）、有监督微调（Supervised Fine-Tuning，SFT）、思维链（Chain-of-Thought，CoT）以及基于人类反馈的强化学习（Reinforcement Learning with Human Feedback，RLHF）等。此外，模型参数的增加和数据集的扩大也起到了关键作用。这些进步使得大语言模型能够遵循人类指令完成相关任务，从而推动了信息抽取任务范式从 Fine-Tuning 向 Zero/Few-Shot 的转变。在使用大语言模型完成具体任务时，得益于这些先进的训练技术，通常不再需要频繁地更改模型参数。这使得大语言模型具有更高的灵活性和通用性，不再局限于特定的任务，而是可以实现多种任务共用一个模型。核心意义即接收 X 作为输入，去预测 Y，其中 X_{Prompt} 表示具体任务的提示指令，这是用户对模型的具体要求（如“总结以下文章”或“提取人物关系”），直接影响模型的输出方向。当 X_{Prompt}=“请提取机构名称”时，模型会聚焦于识别组织实体；若 P=“请分析情感倾向”，则转向情感分类任务。$X_{Context}$ 表示提取信息的文本。可以看出大语言模型的生成结果不仅与输入的文本有关，而且受到提示工程的影响，其结果由二者共同决定；另外大语言模型自身的语言能力也会影响生成结果。

doi:10.3772/j.issn.2095-915x.2025.04.010

$$P(Y \mid X_{\mathrm{Pr}ompt} + X_{Context}) = \sum_{i} \log P(y_i \mid X_{\mathrm{Pr}ompt} + X_{Context}) \quad (6)$$

4.2 数据压缩

大语言模型在生成文本或进行自然语言处理任务时，有时会表现出一种被称为“幻觉”的现象。这种现象指的是模型生成了与事实不符、逻辑上不合理或完全虚构的信息，其问题通常源于数据、训练、推理等因素[27]。在处理中文期刊论文这类高度专业化和结构化的文本时，模型需跟踪并记忆大量信息及上下文，这增加了生成文本时出现幻觉的风险，这种幻觉现象可能导致关键信息的提取出现错误，影响信息的准确性和可靠性。本文分析常见的学术论文的文章结构、关键信息特征以及各关键要素所在位置，构建学术论文关键信息框架，并根据框架结构对 3.4 中输出的 JSON 格式全文文件进行数据压缩。本文分析常见的学术论文结构以及主流文献发现系统实际应用案例，后归纳得到学术论文关键信息框架，以中文期刊为例，主要包含以下 3 类核心信息。

4.2.1 论文描述信息

（1）论文标题，能够简洁、恰当地反映文章的特定内容，包括中文标题及非中文标题。中文标题通常位于论文首页的开头位置，以单独段落呈现。非中文标题并非必然出现，当存在时通常位于首页非中文关键信息段落内或尾页非中文关键信息段落内。

（2）论文摘要，提供文献内容梗概，简明确切地记述论文的主要内容，包括中文摘要及非中文摘要。中文摘要通常以“摘要”或“[摘要]”等开头，以单独段落呈现。非中文摘要并非必然出现，当存在时通常位于首页或尾页，以“Abstract”或“[Abstract]”等开头，以单独段落呈现。

（3）关键词，反映文献特征内容，便于文献检索，包括中文关键词及非中文关键词。中文关键词通常以“关键词”或“[关键词]”等开头，以单独段落呈现。非中文关键词并非必然出现，当存在时通常位于首页或尾页，以“Keywords”或“[Keywords]”等开头，以单独段落呈现。

4.2.2 论文作者信息

（1）作者姓名，论文主题内容的构思、具体研究工作的执行及撰稿执笔等提供主要贡献的人员。中文作者姓名通常在中文标题下方，单独成段落。非中文作者姓名并非必然出现，当存在时通常位于首页或尾页，在非中文标题下方。

（2）作者机构，通常指的是作者在撰写论文时所隶属的大学、实验室或研究中心等机构。中文作者机构通常位于中文作者姓名下方，单独成段落，并通过序号与中文作者姓名对应。非中文作者机构并非必然出现，当存在时通常位于首页或尾页，在非中文作者姓名下方。

4.2.3 论文资助信息

资助信息，通常指的是在学术论文中提到的，为该论文的研究工作提供经费或其他形式支持的机构或基金的信息，通常位于首页的页脚中。

在上述过程中为保证段落还原的准确性去除了页眉页脚，现需对页眉页脚进行还原，以确保信息的完整。中文期刊论文的关键信息全部集中在论文的首页与尾页，对 2.3 中输出的

JSON 格式全文文件进行压缩，仅保留页面标识为首页与尾页的信息，通过明确这些关键页面和内容区域可以大幅度减少需要处理的数据量以及文本长度，从而提升信息提取的效率并减少大语言模型产生的幻觉。

4.3 提示工程模板设计

在大语言模型中，提示工程是指设计和编写提示文本，以引导模型生成符合特定要求的语言输出。提示工程可以包括选择合适的词汇、语法、上下文和主题等元素，以及使用不同的技巧和策略来影响模型的生成行为和结果。提示工程模板设计应尽可能准确且易于解释，本文提示工程模板构成如式（7）及图 6 所示。

$$Prompt = Description + Input + Requirements + Task \quad (7)$$

图 6 提示工程模板样例

其中 *Description* 为任务说明，在使用中固定，包含模型角色以及主要任务的描述；*Input* 为输入，包含待抽取的文本变量 {context} 以及输出注意事项；*Requirements* 为输出要求，在使用中固定，包含输出的注意事项与格式要求；*Task* 为需要执行的具体任务，与中文期刊论文关键信息框架中 6 类核心信息对应，包含输出指令以及与 6 类核心信息对应的 JSON Schema 样例。以论文作者信息抽取为例，*Description* 中将模型角色设定为“信息抽取专家，具备强大的观察力和信息识别能力”，主要任务为“你的任务是从这段文本中抽取关键信息，并严格按照指定的格式输出为 JSON”；*Input* 中文本变量 {context} 为 4.2 中经过压缩后的 JSON 格式中文期刊数据，输出注意事项设定为“OCR 的文字识别结果可能包含错误，如长句子换行被切断、不合理的分词、文字被错误合并等。因此，你需要结合上下文语义进行综合判断，确保抽取的信息准确无误”；*Requirements* 中输出格式设定为“严格按照 JSON 格式输出抽取结果 {schema}”，异常处理设定为“如果文本内容中无法找到相关信息，请直接返回‘NONE’”；Task 设定为“请根据以上说明，执行信息提取任务，并返回符合要求的 JSON 格式输出”，并给出与作者信息提取任务对应的 JSON Schema 样例。

doi:10.3772/j.issn.2095-915x.2025.04.010

5 实验与结果分析

为了验证本文 RRO-ChatData 方法的有效性，随机抽取 200 篇中文期刊论文作为实验样本，全部为扫描型 PDF 全文，实验样本均包含论文标题、论文作者、论文摘要、关键词、资助信息以及出版信息等关键元素。通过人工处理方式对实验样本数据进行阅读顺序还原与关键信息抽取，经校对确认无误后作为实验对照组。实验环境配置如表 1 所示。

表 1 实验环境配置

名称	配置信息
操作系统	Ubuntu 22.04
开发语言	Python:3.10.16
CPU	Intel(R) Xeon(R) 4214
GPU	Geforce RTX 3090 (24G)
内存	256
框架	Cuda 12.4
大语言模型	Qwen2.5-7B-Instruct (128K)

5.1 阅读顺序还原效果

使用对比实验的方法，通过人工校验的方式判断 RRO-ChatData 方法对 YOLOV8 模型识别结果的阅读顺序还原方法效果，实验结果如表 2 所示。经统计，200 份实验样本中双栏排版格式占比最高，共计 138 篇，占比 69%，通栏共计 45 篇，占比 22.5%，三栏共计 17 篇，占比 8.5%。实验样本共涉及 5329 个段落，其中跨页段落共计 413 段，跨图、表及公式段落总计 226 段。经检验，RRO-ChatData 方法将段落识别的准确率整体提高 12.7%，跨页段落识别准确率在通栏场景提升 29.7%，表明算法对分页断行、页眉页脚干扰等问题的处理有效性；针对占比 69% 的双栏排版核心场景，总段落准确率提升 14.2%，跨图段落识别准确率提升幅度最高达 22.3%，有效解决双栏文本绕图排版的识别难题。针对复杂场景，跨表段落识别准确率在三栏场景提升 22.7%，跨公式段落在双栏场景提升 14.6%，验证方法对非文本元素的抗干扰能力；经人工复核，错误主要集中于复合跨元素段落（同时跨越表格与公式的段落）和高密度数学公式，导致行间关系判定出现错误，但总体而言，RRO-ChatData 方法能够有效地按照人类阅读顺序对中文期刊论文进行段落还原，还原效果如图 7、图 8 所示。

表 2 段落还原对比实验结果

识别方法	段落类型	总段落准确率	跨页段落准确率	跨图段落准确率	跨表段落准确率	跨公式段落准确率
原始识别结果	通栏	70.2%	55.5%	68.3%	65.1%	63.8%
	双栏	71.1%	68.1%	60.7%	57.2%	55.4%
	三栏	68.5%	62.2%	53.2%	50.8%	47.1%
	全部	70.3%	58.9%	65.3%	62.1%	60.5%
RRO-ChatData 优化结果	通栏	82.5%(+12.3%)	85.2%(+29.7%)	81.4%(+13.1%)	80.0%(+14.9%)	71.5%(+7.7%)
	双栏	85.3%(+14.2%)	88.5%(+20.4%)	83.0%(+22.3%)	76.5%(+19.3%)	70.0%(+14.6%)
	三栏	80.5%(+12.0%)	79.2%(+17.0%)	72.5%(+19.3%)	73.5%(+22.7%)	68.2%(+21.1%)
	全部	83.0%(+12.7%)	85.4%(+26.5%)	81.0%(+15.7%)	78.7%(+16.6%)	70.9%(+10.4%)

图 7　期刊论文全文 PDF

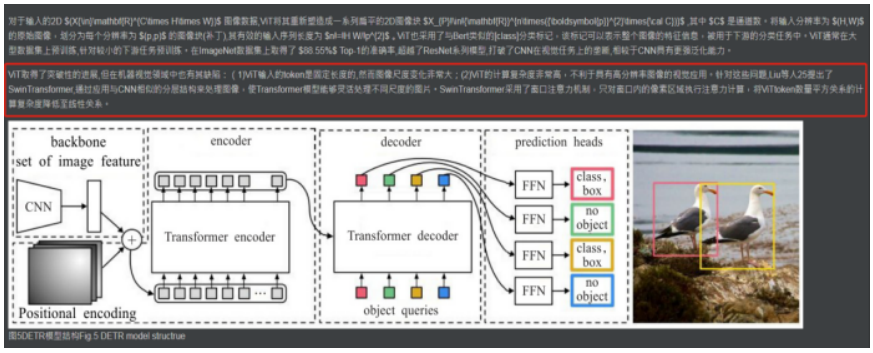


图 8　段落还原后论文 Markdown 截图

5.2　关键信息抽取效果

使用准确率来判断“关键信息抽取”的效果，准确率的计算方式为 6 类核心信息正确提取数量占所有提取数量的百分比，待抽取文本内容如图 9 所示。实验结果表明，RRO-ChatData 方法相较于 LayoutLMv3 方法，在各类信息的提取准确率上均有显著提升，如表 3 所示：中文标题准确率提升至 99.5%（+8.0%），几乎能够完全准确地提取中文论文标题。中文摘要准确率达到 95.0%（+5.0%），在中文摘要提取上的准确性有了很大提高。中文关键词准确率提升至 95.5%（+17.5%），相比 LayoutLMv3 方法有较大提高，论文描述信息及作者信息的识别结果如图 10、图 11 所示。在资助信息提取方面，虽然其内容复杂且表述不固定，但 RRO-ChatData 方法仍取得了 68.5%（+48%）的准确率，显示出该方法在处理复杂信息提取任务上的强大能力，资助信息识别结果如图 12 所示。整体而言，RRO-ChatData 方法能够更有效地提取中文期刊论文中 PDF 的关键信息，尤其在处理复杂信息如资助信息时，相比 LayoutLMv3 方法具有明显优势。

doi:10.3772/j.issn.2095-915x.2025.04.010

◎热点与综述◎

Transformer在计算机视觉领域的研究综述

李 翔,张 涛,张 哲,魏宏杨,钱育蓉
新疆大学 软件学院,乌鲁木齐 830002

摘 要: Transformer是一种基于自注意力机制的深度神经网络。近几年,基于Transformer的模型已成为计算机视觉领域的热门研究方向,其结构也在不断改进和扩展,比如局部注意力机制、全金塔结构等。通过对基于Transformer结构改进的视觉模型,分别从性能优化和结构改进两个方面进行综述和总结;也对比分析了Transformer和CNN各自结构的优缺点,并介绍了一种新型的CNN+Transformer的混合结构;最后,对Transformer在计算机视觉上的发展进行总结和展望。

关键词: Transformer;卷积神经网络(CNN);混合结构;计算机视觉;深度学习
文献标志码: A **中图分类号:** TP391 doi:10.3778/j.issn.1002-8331.2204-0207

Survey of Transformer Research in Computer Vision

LI Xiang, ZHANG Tao, ZHANG Zhe, WEI Hongyang, QIAN Yurong
College of Software, Xinjiang University, Urumqi 830002, China

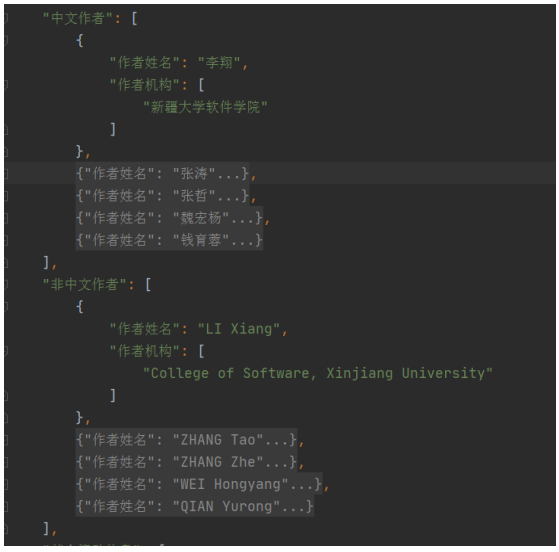
Abstract: Transformer is a deep neural network based on self-attention mechanism. In recent years, Transformer-based models have become a hot research direction in the field of computer vision, and their structures are constantly being improved and expanded, such as local attention mechanisms, pyramid structures, and so on. Through the improved vision model based on Transformer structure, the performance optimization and structure improvement are reviewed and summarized respectively. In addition, the advantages and disadvantages of the respective structures of the Transformer and convolutional neural network (CNN) are compared and analyzed, and a new hybrid structure of CNN+Transformer is introduced. Finally, the development of Transformer in computer vision is summarized and prospected.

Key words: Transformer; convolutional neural network(CNN); hybrid structure; computer vision; deep learning

图 9 待抽取文本内容

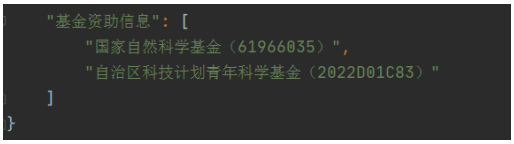

```
{
    "中文文章题名": "Transformer在计算机视觉领域的研究综述",
    "非中文文章题名": "Survey of Transformer Research in Computer Vision",
    "非中文文章题名语种": "英语",
    "中文摘要": "Transformer是一种基于自注意力机制的深度神经网络。近几年,基于Transformer的模型已成为计算机
    "中文关键词": [
        "Transformer",
        "卷积神经网络(CNN)",
        "混合结构",
        "计算机视觉",
        "深度学习"
    ],
    "非中文摘要": "Transformer is a deep neural network based on self-attention mechanism. In r
    "非中文摘要语种": "英语",
    "非中文关键词": [
        "Transformer",
        "convolutional neural network( CNN)",
        "hybrid structure",
        "computer vision",
        "deep learning"
    ],
    "非中文关键词语种": "英语",
```


图 10 论文描述信息识别结果




```
    "中文作者": [
        {
            "作者姓名": "李翔",
            "作者机构": [
                "新疆大学软件学院"
            ]
        },
        {"作者姓名": "张涛"...},
        {"作者姓名": "张哲"...},
        {"作者姓名": "魏宏杨"...},
        {"作者姓名": "钱育蓉"...}
    ],
    "非中文作者": [
        {
            "作者姓名": "LI Xiang",
            "作者机构": [
                "College of Software, Xinjiang University"
            ]
        },
        {"作者姓名": "ZHANG Tao"...},
        {"作者姓名": "ZHANG Zhe"...},
        {"作者姓名": "WEI Hongyang"...},
        {"作者姓名": "QIAN Yurong"...}
    ],
```


图 11 作者信息识别结果




```
    "基金资助信息": [
        "国家自然科学基金(61966035)",
        "自治区科技计划青年科学基金(2022D01C83)"
    ]
}
```


图 12 资助信息识别结果

表 3 关键信息识别准确率

提取方法	字段名称	描述信息						作者信息				资助信息
		中文标题	非中文标题	中文摘要	非中文摘要	中文关键词	非中文关键词	中文作者姓名	非中文作者姓名	中文作者机构	非中文作者机构	资助信息
LayoutLMv3	准确率	91.5%	88.5%	90.0%	87.5%	77.0%	74.5%	82.8%	85.6%	62.2%	68.1%	20.5%
RRO-ChatData	准确率	99.5% (+8.0%)	99.5% (+7.0%)	95.0% (+5.0%)	95.0% (+7.5%)	95.5% (+17.5%)	92.6% (+18.1%)	98.8% (+16.0%)	96.5% (+10.9%)	84.4% (+22.2%)	81.2% (+13.1%)	68.5% (+48%)

6 总结

本文提出一种基于 RRO-ChatData 的中文期刊关键信息识别方法，前者通过阅读顺序还原技术对 YOLOv8 模型识别后的中文期刊文献全文进行段落重组，为大语言模型储备按照人类阅读顺序复原的全文语料；后者针对中文期刊资源特点，利用大语言模型推理能力构建 ChatData 特征识别方法，并使用该方法对中文期刊全文进行关键信息识别。经过实验验证，基于 RRO-ChatData 的中文期刊关键信息识别方法能够相对准确地对中文期刊进行全文段落顺序还原与关键内容提取，目前该方法已经在国家工程技术数字图书馆试点应用并取得初步成效，能够有效提升图书馆单位数据加工效率，并为问答式检索、智能综述等应用场景提供更加优质的语料。

参 考 文 献

[1] 张洪红 . 基于注意力机制的文档图像版面分析算

法 [D]. 青岛：青岛科技大学 , 2023.
[2] 戴璐 . 美国亨廷顿图书馆经费筹措策略及其启示 [J]. 河北科技图苑 , 2020, 33(3): 80-83.
[3] 蒋鸿标 . 学术期刊排版十法 [J]. 中国期刊研究 , 2003, 14(2): 229-230.
[4] 郭利敏 , 付雅明 . 以大语言模型构建智慧图书馆：框架和未来 [J]. 图书馆杂志 , 2023, 42(11): 22-30, 133.
[5] 何彦青 , 陈光云 , 兰天 , 等 . 基于 Rule-Faster-RCNN 的多语科技论文 PDF 文档结构框架元素解析 [J]. 情报科学 , 2023, 41(4): 51-61.
[6] 周楠 , 欧阳鑫玉 . 卷积神经网络发展 [J]. 辽宁科技大学学报 , 2021, 44(5): 349-356.
[7] 王腾阳 , 赵小丹 , 胡林 . 基于词性标注规则的马铃薯文献信息抽取方法 [J]. 科学技术与工程 , 2023, 23(27): 11562-11569.
[8] PEÑA A, MORALES A, FIERREZ J, et al. Continuous document layout analysis: human-in-the-loop AI-based data curation, database, and evaluation in the domain of public affairs[J]. Information Fusion, 2024(108): 102398.
[9] 陈善雄 , 韩旭 , 林小渝 , 等 . 基于 MSER 和 CNN 的彝文古籍文献的字符检测方法 [J]. 华南理工大学学报（自然科学版）, 2020, 48(6): 123-133.
[10] 尹潇伟 . 基于深度学习的表格结构与文本识别研究 [D]. 青岛：青岛大学 , 2023.
[11] HE D, COHEN S, PRICE B, et al. Multi-scale multi-task FCN for semantic page segmentation and table detection[C]// 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017: 254-261.
[12] SIMONYAN K, ZISSERMAN A.Very deep convolutional networks for large-scale image recognition[J]. arxiv preprint arxiv:1409.1556, 2014.
[13] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.
[14] ZHANG H, XU C, SHI C, et al. HSCA-Net: a hybrid spatial-channel attention network in multiscale feature pyramid for document layout analysis[J]. Journal of Artificial Intelligence and Technology, 2023, 3(1): 10-17.
[15] LIU R, LEHMAN J, MOLINO P, et al. An intriguing failing of convolutional neural networks and the CoordConv solution[J]. arxiv preprint arxiv:1807.03247, 2018.
[16] HAN X, ZHANG Z, DING N, et al. Pre-trained models: past, present and future[J]. AI Open, 2021(2): 225-250.
[17] REDMON J, FARHADI A. YOLOv3: an incremental improvement[J]. arxiv preprint arxiv:1804.02767, 2018.
[18] ZHAO Y, LV W, XU S. DETRs Beat YOLOs on real-time object detection[J]. arxiv preprint arxiv:2304.08069, 2023.
[19] 罗文 , 王厚峰 . 大语言模型评测综述 [J]. 中文信息学报 , 2024, 38(1): 1-23.
[20] SHEN Z, ZHANG R, DELL M, et al. LayoutParser: a unified toolkit for deep learning based document image analysis[J]. arxiv preprint arxiv:2103.15348, 2021.
[21] HUANG Y, LV T, CUI L, et al. LayoutLMv3: pre-training for document AI with unified text and image masking[J]. arxiv preprint arxiv:2204.08387, 2022.
[22] 周忆莲 . 学术论文 PDF 结构解析技术的研究 [D]. 长沙：湖南大学 , 2020.
[23] 潘定 , 梁倬骞 , 温秋华 , 等 . 一种基于 opencv 库的文本版面识别方法：CN201911059127. X[P]. 2020-03-17.
[24] 王红霞 , 武甲礼 , 陈德山 . 分离复杂背景下的文档图像二值化方法 [J]. 中国图象图形学报 , 2023, 28(7): 2011-2025.
[25] 刘瑞锦 , 何章鸣 . 基于 YOLOv8 的卫星遥感图像快速目标检测方法 [J]. 空间控制技术与应用 , 2023, 49(5): 89-97.
[26] 王玲 . 链编码的获取和文档图像的版面分析 [D]. 上海：华东师范大学 , 2008.
[27] 时宗彬 , 朱丽雅 , 乐小虬 . 基于本地大语言模型和提示工程的材料信息抽取方法研究 [J]. 数据分析与知识发现 , 2024, 8(7): 23-31.