智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

历史文档布局分析研究

彭阳, 王德军, 孟博, 吴余龙, 胡宗华

引用本文:
彭阳, 王德军, 孟博, 等. 历史文档布局分析研究[J]. 智能系统学报, 2026, 21(3): 727-738.
PENG Yang, WANG Dejun, MENG Bo, et al. Historical document layout analysis[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(3): 727-738.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202501011

DOI: 10.11992/tis.202501011
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20260326.1755.002

历史文档布局分析研究

彭阳[1], 王德军[1], 孟博[1], 吴余龙[2], 胡宗华[2]
（1. 中南民族大学 计算机科学学院, 湖北 武汉 430074; 2. 武汉力龙信息科技股份有限公司, 湖北 武汉 430015）

摘　要: 文档布局分析是将扫描得到的页面图像转换为可搜索的全部文本, 然而, 该研究主要集中在结构化和半结构化文档领域, 历史文档图像质量差、结构混乱, 比常规结构化文档更具挑战性。为解决上述问题, 本文在现有数据集上新增标注, 构建融合文本内容、文档图像及空间特征的多模态网络; 进一步地, 通过结合语义相似度与空间邻接关系, 设计关系预测网络来预测阅读顺序, 获得最终的布局分析结果。本文方法在 LA-READ 和 LA-FCR 数据集上的 mAP 值分别达到 92.4% 和 85.6%, 总排序准确率 (Acc) 与现有方法的最好结果相比分别提升 3.5% 和 2.7%, 实验结果表明, 所提方法有效提升了历史文档布局分析任务的效果。

关键词: 文档理解; 机器学习; 历史文档; 页面结构; 布局分析; 阅读顺序; 空间特征; 语义相似度
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2026)03−0727−12

中文引用格式: 彭阳, 王德军, 孟博, 等. 历史文档布局分析研究 [J]. 智能系统学报, 2026, 21(3): 727–738.
英文引用格式: PENG Yang, WANG Dejun, MENG Bo, et al. Historical document layout analysis[J]. CAAI transactions on intelligent systems, 2026, 21(3): 727–738.

Historical document layout analysis

PENG Yang[1], WANG Dejun[1], MENG Bo[1], WU Yulong[2], HU Zonghua[2]
(1. College of Computer Science, South-Central Minzu University, Wuhan 430074, China; 2. Wuhan Lilosoft Co., Ltd, Wuhan 430015, China)

Abstract: Document layout analysis converts scanned page images into fully searchable text, yet existing research primarily focuses on structured and semi-structured documents. Compared with conventional structured documents, historical documents pose greater challenges due to poor image quality and irregular layouts. To address these issues, we augmented annotations on existing datasets and constructed a multimodal network that integrates textual content, document images, and spatial features. Furthermore, we designed a relation prediction network that combines semantic similarity with spatial adjacency to determine reading order, yielding the final layout analysis result. Our method achieves mean average precision (mAP) scores of 92.4% and 85.6% on the LA-READ and LA-FCR datasets, respectively, with total ordering accuracy (Acc) surpassing previous state-of-the-art results by 3.5% and 2.7%. Experimental results demonstrate the effectiveness of our approach for historical document layout analysis.

Keywords: document understanding; machine learning; historical documents; page structure; layout analysis; reading order; spatial features; semantic similarity

将纸质版的历史文档利用文档图像识别与理解等新技术进行高效率的数字化处理, 可以方便后续历史文献的内容保护、智能搜索、语义理解、知识发现等任务[1]。文档图像的分析主要有两个过程: 文本识别和文档布局分析。第一个过程主要定位文档中的文本并获得其转录内容。第二个过程通过返回边界框和类别来检测文档的布局位置, 获得文档页面结构的描述性表示。这两个过程是互补的, 如果不能按照逻辑顺序对文本内容进行排序, 那么获得的转录本可能毫无意义。因此, 识别不同区域并排列阅读顺序对于将文本转换为有用的信息至关重要。现有的文档布局分析主要集中在印刷文件上, 对历史手写文档的关注要少得多, 且现有的历史文档图像大多仅有转录

收稿日期: 2025−01−08.　**网络出版日期**: 2026−03−27.
基金项目: 国家重点研发计划项目 (2020YFC1522900); 湖北省科技创新人才计划项目 (2023DJC094); 民族语言智能分析与安全治理教育部重点实验室开放课题 (ORP-2024-04).
通信作者: 王德军. E-mail: dejun@scuec.edu.cn.

后的文本内容，缺少逻辑布局标注和阅读顺序标注，导致对历史文档的布局分析难度较大。

为了更好地解决上述问题，本文通过融合文档的图像、文本和空间特征，提出了一种基于多模态和关系预测网络的布局分析模型，主要贡献包括：构建历史文档布局分析数据集，在原有标注基础上补充逻辑结构和阅读顺序信息，并设计多模态关系预测网络，通过融合图像、文本和空间特征，建模区域实体间的语义相似度和空间邻接关系，从而实现对历史文档布局分析和阅读顺序的准确预测。所提数据集和模型代码均已公开：https://github.com/PyTc-PengYang/MRNM_layout_analysis。

1　相关工作

文档结构分析可分为物理布局分析和逻辑结构分析[2]。早期的文档结构分析方法主要基于启发式规则或语法分析[3]。在过去的十年中，越来越多的研究同时进行物理布局分析和逻辑角色分类，该任务也称为页面对象检测[4]。除了检测页面对象之外，许多研究还深入研究了文档中组件之间的阅读顺序。

1.1　布局分析方法

根据使用的模态特征数量，文档布局分析方法可分为单模态、双模态和多模态方法[5]。文档布局分析任务中的模态主要分为视觉模态、文本模态和空间模态，来自不同模态的特征信息可以帮助更好地分析文档布局[6]。

1.1.1　单模态方法

单模态方法依赖单一模态 (文本或视觉) 进行文档分析。例如 Afzal 等[7] 提出基于 CNN(convolutional neural network) 的文档图像分类方法，Davis 等[8] 则利用视觉特征来进行关键信息提取任务。然而，单模态方法在处理复杂文档时存在局限性，难以同时捕捉视觉和语义信息。

1.1.2　双模态方法

双模态方法通过结合文本与视觉或空间模态，来提升文档分析效果。例如 Bakkali 等[9] 结合视觉和文本模态，利用监督对比学习提升文档分类性能。BROS(bertrelying on spatiality) 模型[10] 通过结合文本的空间布局信息和语义信息，在关键信息提取任务中表现优异。

1.1.3　多模态方法

多模态方法整合文本、视觉、空间等多种模态来全面捕捉文档信息。例如，Huang 等[11] 提出基于 Transformer 的多模态预训练模型，结合文本、图像、空间模态，适应多种文档理解任务。Li 等[12] 引入文本、空间、图像及辅助模态，进一步增强文档表示能力。多模态方法在复杂文档中表现优异，但计算复杂度高且依赖大量标注数据。

1.2　阅读顺序预测方法

阅读顺序预测 (reading order prediction, ROP) 是针对文档布局的特殊任务。在视觉文档理解领域中，下游任务的性能可通过利用适当的阅读顺序信息而得到提高[13]。阅读顺序预测方法可分为两类。

1.2.1　基于规则的方法

这类方法主要依赖人工设定的规则或启发式算法来推测文档的阅读顺序，适用于结构规范的文档。如 XY-cut[14] 方法通过在水平方向和垂直方向上反复切割，构建出一棵 XY-cut 树，遍历该树可获得布局元素的排列顺序。Ferilli 等[15] 基于人类阅读行为的一般假设策略来识别文档页面组件的阅读顺序。然而文档的阅读顺序并不总是从上到下和从左到右的，这限制了基于规则的方法的有效性。

1.2.2　基于数据驱动的方法

这类方法利用统计学习或神经网络，从数据中自动学习阅读顺序模式。Li 等[16] 提出了一个带有 GCN(graph convolutional network) 布局编码器的指针网络来预测阅读顺序。Quirós 等[17] 提出将阅读顺序问题转化为文本基线集合的排序问题，通过排列或二元顺序关系来定义阅读顺序。马伟洪等[18] 基于图神经网络和语言模型来完成中文古籍文档的阅读顺序检测。Zhang 等[19] 将任务建模为在令牌图中预测路径，判断文本令牌之间是否存在连接关系。Qiao 等[20] 提出布局感知位置嵌入模块，该模块可生成灵活且自适应的位置映射，并引入新的损失函数用于优化模型。

综上所述，目前针对手写历史文档同时进行布局分析和阅读顺序研究的成果较少，因此本文首先建立包含页面布局和阅读顺序的数据集，利用文本内容和文档图像之间的相关性，联合使用多种特征来划分文本区域，并结合图神经网络建立不同区域实体之间的联系，以此完成布局分析任务和阅读顺序预测。

2　模型结构

本文提出一种基于多模态和关系预测网络的布局分析模型 MRNM(multi-modal features and relation prediction network)。方法主要分为两个阶

段, 第 1 阶段提取不同特征并融合构成多模态特征, 并基于 Transformer Encoder 实现初步布局检测; 第 2 阶段通过 GNN(graph neural network) 图模型建立关系矩阵来检测阅读顺序。整体流程如图 1 所示, 其中主要包含 3 个部分: 数据集构造、多模态网络、关系预测网络。

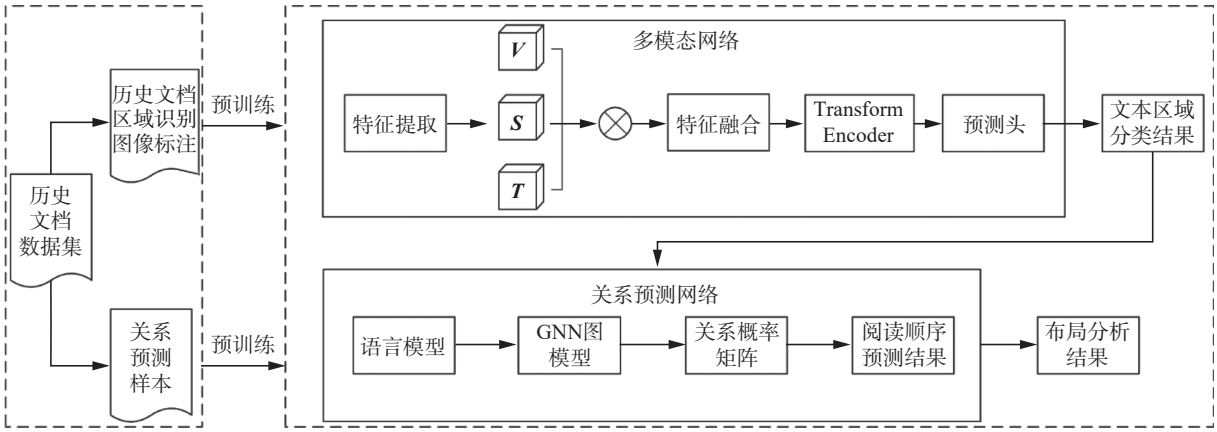


图 1　方法流程

Fig. 1　Method flow chart

2.1　多模态网络

本文从 3 个维度分别提取特征: 利用卷积神经网络和多尺度特征融合技术获取视觉特征, 基于 BERT(bidirectional encoder representations from transformers) 预训练语言模型并结合注意力融合机制提取文本特征, 通过引入新的坐标变换方式得到空间特征。针对文本行和文本段, 采用分层嵌入结合特定注意力机制进行特征融合。最后利用区域候选网络获得文本区域检测框, 完成布局分析任务。本文提出的多模态网络如图 2 所示。

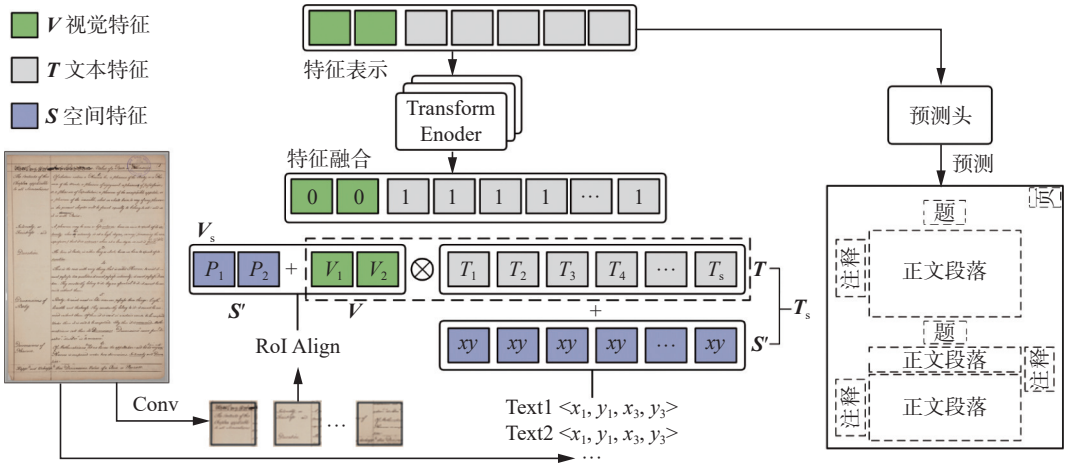


图 2　多模态网络示意

Fig. 2　Schematic of multimodal network

2.1.1　视觉嵌入

将历史文档图像调整为 $H\times W$ 大小, 设图像用 $\boldsymbol{I}\in\mathbf{R}^{C\times W\times H}$ 表示, 其中 C、W 和 H 分别表示图像通道数、图像的宽度和高度。本文基于 CNN 卷积神经网络来提取视觉特征, 使用多尺度特征融合来增强模型的感知能力[21]。卷积后会形成不同层次的特征映射, 它们的特征图大小不同, C_3 分辨率最高, C_4 和 C_5 逐层下采样, 通常缩小为前一层的一半, 直接融合这些特征图会因分辨率不一致导致空间上的信息失配。本文通过上采样, 首先将特征映射 C_4 和 C_5 的尺寸调整为 C_3 的大小, 然后将它们连接起来送到 3×3 的卷积层中, 生成一个具有 256 个通道的特征映射 C_f, 根据 C_f 的文本区域 $r_i=(x_{1i},y_{1i},x_{2i},y_{2i})$, 使用 RoIAlign 算法从中提取对应的特征, 其中 (x_{1i},y_{1i}) 和 (x_{2i},y_{2i}) 分别表示左上角和右下角的坐标。最终的视觉嵌入表示为

$$\boldsymbol{V}_{r_i}=\mathrm{LN}(\mathrm{ReLU}(\mathrm{FC}(\mathrm{RoIAlign}(C_\mathrm{f},r_i))))$$

式中: FC 为包含 1024 个节点的全连接层 (fully connected layer), ReLU 为非线性激活函数(rectified linear unit), LN 为层归一化 (layer normalization)。

2.1.2　文本嵌入

根据文本区域是否跨行的原则, 将文本区域

r_i 分为两种类型：文本行 t^L 和文本段 t^P，文本段 t^P 中包含 n 个文本行 $t^L(n\geqslant 2)$，如“标题”是典型的文本行，“段落”大多属于文本段。本文依据 Wang 等[22] 的策略，使用基于 Sentence-BERT[23] 的预训练语言模型来提取每个文本行 t_i^L 的文本嵌入 $\boldsymbol{T}_{t_i}^L$，通过注意力融合机制对文本段 t^P 中的所有文本行嵌入 $\boldsymbol{T}_{t_i}^L$ 进行加权求和，获得文本段的文本嵌入 $\boldsymbol{T}_{t_n}^P$。

对于文本行 t^L 的文本嵌入 $\boldsymbol{T}_{t_i}^L$，按照从左上到右下的顺序读取文档图像中的所有文本行并将其序列化为一维序列，利用 Sentence-BERT 中的分词器将每个文本行转换为一个子词标记序列 (sub-word token sequence)，然后将其输入到预训练的 Sentence-BERT 模型中获得每个 token 的嵌入。计算每个文本行中所有 token 的嵌入向量的平均值，得到该文本行的整体文本嵌入，最后输入到一个包含 1 024 个节点的全连接层，使其与视觉嵌入的维度相同。最终文本行 t_i^L 的文本嵌入表示为

$$\boldsymbol{T}_{t_i}^L=\mathrm{LN}\left(\mathrm{ReLU}\left(\mathrm{FC}\left(\frac{1}{m}\sum_{j=1}^{m}\boldsymbol{E}\left(\boldsymbol{x}_j\right)\right)\right)\right)$$

式中：m 是文本行 t_i^L 中 token 的数量，$\boldsymbol{E}(\boldsymbol{x}_j)$ 是文本行中第 j 个 token 的嵌入向量 (token embedding)。

对于文本段 t^P 的文本嵌入 $\boldsymbol{T}_{t_n}^P$，假设文本段 t_n^P 中包含的文本行为 $[t_{n_1}\ t_{n_2}\ \cdots\ t_{n_k}]$，首先使用两个全连接层计算文本段中的每个文本行的注意力得分 $a_{t_{n_j}}$，表示为

$$a_{t_{n_j}}=\mathrm{FC}_1(\tanh(\mathrm{FC}_2(\boldsymbol{T}_{t_{n_j}}^L)))$$

式中：t_n 表示第 n 个文本段，j 表示文本段中第 j 个文本行，FC_1 和 FC_2 分别是具有 1 024 个节点和 1 个节点的全连接层，tanh 为非线性激活函数。FC_2 层将输入特征转换为中间表示，便于捕捉文本行的高阶语义特征，FC_1 层将中间表示转换为一个注意力得分，表示当前文本行的重要性。之后通过 Softmax 函数将所有文本行的注意力得分归一化为一个概率分布 (注意力权重 $w_{t_{n_j}}$)，确保所有权重的总和为 1，表示为

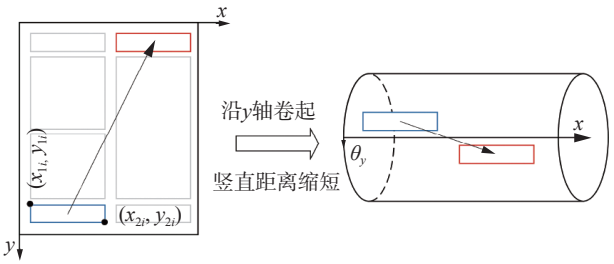
$$w_{t_{n_j}}=\frac{\exp(a_{t_{n_j}})}{\sum_{j=1}^{k}\exp(a_{t_{n_j}})}$$

式中：k 表示文本段 t_n 中的文本行数量。最后利用注意力权重对所有文本行嵌入进行加权求和，得到文本段 t_n^P 的文本嵌入 $\boldsymbol{T}_{t_n}^P$，最终表示为

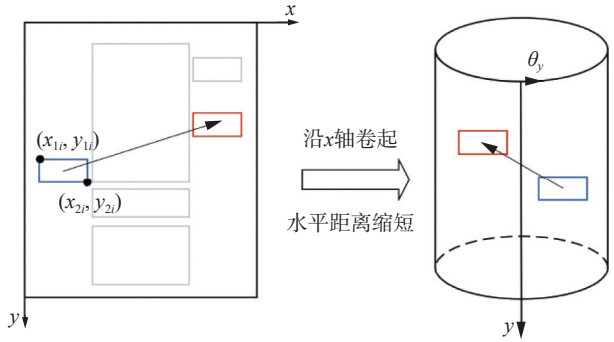
$$\boldsymbol{T}_{t_n}^P=\sum_{j=1}^{k}w_{t_{n_j}}\boldsymbol{T}_{t_{n_j}}^L$$

2.1.3　空间位置嵌入

传统的空间特征提取依赖于原始二维空间中的边界框的绝对坐标，通常距离更近的文本区域在语义上可能更相关，但历史手写文档图像中经常表现出复杂的空间关系，偏离了这种模式[20]。因此本文引入了一个新的空间位置嵌入，假设页面图像是一个双列文档，若沿着 y 轴卷起页面，形成类似于一个圆柱体，这样可减少上下位置文本区域的间隔，如图 3(a) 所示，两个文本区域在竖直线上的距离被缩短；同样地，沿着 x 轴卷起可减少左右两边文本区域的间隔，如图 3(b) 所示，两个文本区域在水平线上的距离被缩短。



(a) 沿着y轴卷起



(b) 沿着x轴卷起

图 3　沿着坐标轴卷起页面的示意

Fig. 3　Rolling up the page along the coordinate axes

具体地，对于大小为 $H\times W$ 的文档图像，以左上角为原点，其原始二维空间中的区域边界框坐标 $r_i=(x_{1i},y_{1i},x_{2i},y_{2i})$，借助极坐标与弧长公式可将其进行坐标变换，变换后的坐标为

$$r_i'=\left(x_{1i}',y_{1i}',x_{2i}',y_{2i}'\right)=\left(\frac{W}{2\pi}\cos\frac{2\pi x_{\mathrm{mid}_i}}{W},\frac{W}{2\pi}\sin\frac{2\pi x_{\mathrm{mid}_i}}{W},\right.$$
$$\left.\frac{H}{2\pi}\cos\frac{2\pi y_{\mathrm{mid}_i}}{H},\frac{H}{2\pi}\sin\frac{2\pi y_{\mathrm{mid}_i}}{H}\right)$$

式中：$(x_{\mathrm{mid}_i},y_{\mathrm{mid}_i})$ 是区域边界框 r_i 的中点坐标。通过以上变换进行坐标映射，可达到类似于滚动页面的操作。将原始位置信息和转换后的位置信息进行编码获得新的空间位置嵌入，表示为

$$\boldsymbol{S}_{r_i}=\mathrm{LN}\left(\mathrm{ReLU}\left(\mathrm{MLP}\left(r_i,r_i',\frac{w_i}{W},\frac{h_i}{H}\right)\right)\right)$$

式中：w_i 和 h_i 分别表示区域边界框 r_i 的高度和宽度，MLP 为多层感知机 (multilayer perceptron)，由 2 个全连接层和 1 024 个节点组成。

2.1.4 特征融合

卷积神经网络的目标是将从高分辨率浅层特征到低分辨率深层特征的多个卷积层进行整合, 深层特征具有较丰富的语义信息, 而浅层特征对上下文的依赖较少, 更关注局部细节。为了更有效地整合视觉、文本和空间特征, 本文针对不同层级的特征, 采用分层嵌入[24]并结合特定的注意力机制进行多模态特征融合, 来增强对历史手稿的检测能力。

具体地, 对于每个文本区域 r_i, 首先提取其视觉嵌入 $\boldsymbol{V}_{r_i}$、文本嵌入 $\boldsymbol{T}_{r_i}$ 和空间嵌入 $\boldsymbol{S}_{r_i}$, 然后将 3 个模态的特征在通道维度上进行拼接, 经过全连接层压缩到统一的维数 (1 024 维), 形成其初始多模态表示 $\boldsymbol{F}_{r_i}$:

$$\boldsymbol{F}_{r_i}=\mathrm{FC}(\mathrm{Concatt}(\boldsymbol{V}_{r_i},\boldsymbol{T}_{r_i},\boldsymbol{S}_{r_i}))$$

式中 FC 是具有 1 024 个节点的全连接层。

对于文本行 t^L, 本文采用通道注意力机制 (channel attention module, CAM) 来处理文本行特征。对输入特征图 $\boldsymbol{F}_{r_i}^{L}\in\mathbf{R}^{C\times W\times H}$ 进行全局最大池化和全局平均池化操作, 分别获得通道信息的两个统计特征, 再分别送入一个共享的多层感知机 (MLP) 中, 最后将输出结果相加, 获得通道注意力权重矩阵 $\boldsymbol{M}_{\mathrm{C}}$, 表示为

$$\boldsymbol{M}_{\mathrm{C}}=\sigma(\mathrm{MLP}(\mathrm{AvgPool}(\boldsymbol{F}_{r_i}^{L}))+\mathrm{MLP}(\mathrm{MaxPool}(\boldsymbol{F}_{r_i}^{L})))$$

式中: σ 表示 Sigmoid 激活函数, 再使用通道权重矩阵 $\boldsymbol{M}_{\mathrm{C}}$ 对原始特征图进行加权, 获得通道增强后的特征:

$$\boldsymbol{F}_{r_i}^{L\cdot\mathrm{CAM}}=\boldsymbol{F}_{r_i}^{L}\cdot\boldsymbol{M}_{C}$$

最后将增强后的特征输入到 Transformer Encoder 中进行全局特征表示建模, 最终的文本行区域多模态表示 $\tilde{\boldsymbol{F}}_{r_i}^{L}$:

$$\tilde{\boldsymbol{F}}_{r_i}^{L}=\text{Transformer Encoder}(\boldsymbol{F}_{r_i}^{L\cdot\mathrm{CAM}})$$

对于文本段 t^P, 本文采用空间注意力机制 (spatial attention module, SAM) 来处理文本段特征。对输入特征图 $\boldsymbol{F}_{r_i}^{P}\in\boldsymbol{R}^{C\times W\times H}$ 进行通道维度的全局最大池化和全局平均池化, 再按照通道拼接, 最后通过卷积得到空间注意力权重矩阵 $\boldsymbol{M}_{\mathrm{S}}$, 表示为

$$\boldsymbol{M}_{\mathrm{S}}=\sigma(\mathrm{Conv}(\mathrm{Concat}[\mathrm{AvgPool}(\boldsymbol{F}_{r_i}^{P}),\mathrm{MaxPool}(\boldsymbol{F}_{r_i}^{P})]))$$

再使用空间权重矩阵 $\boldsymbol{M}_{\mathrm{S}}$ 对原始特征图进行加权, 获得空间增强后的特征:

$$\boldsymbol{F}_{r_i}^{P\cdot\mathrm{SAM}}=\boldsymbol{F}_{r_i}^{P}\cdot\boldsymbol{M}_{\mathrm{S}}$$

最后将增强后的特征输入到 Transformer Encoder 中, 最终的文本段多模态表示 $\tilde{\boldsymbol{F}}_{r_i}^{P}$:

$$\tilde{\boldsymbol{F}}_{r_i}^{P}=\text{Transformer Encoder}(\boldsymbol{F}_{r_i}^{P\cdot\mathrm{SAM}})$$

为了节省计算量, 本文使用 1 层 Transformer Encoder, 其中头数设置为 12, 隐藏状态维数设置为 768, 前馈网络维数设置为 2 048, 输出维数与隐藏状态维数保持一致为 768。

为了增强文本、视觉和空间特征的相互作用, 本文采用对比学习损失 (contrastive loss), 对同一文本区域的各模态特征向量进行对齐:

$$\boldsymbol{L}_{\mathrm{cont}}=\sum_{i=1}^{N}\left[\|\boldsymbol{T}_i-\boldsymbol{V}_i\|_2^2+\|\boldsymbol{T}_i-\boldsymbol{S}_i\|_2^2+\|\boldsymbol{V}_i-\boldsymbol{S}_i\|_2^2\right]$$

式中: $\boldsymbol{T}_i$、$\boldsymbol{V}_i$、$\boldsymbol{S}_i$ 分别表示文本、视觉和空间特征。同时为了防止过拟合, 对特征向量添加正则化约束, 本文使用 L2 正则化约束:

$$\boldsymbol{L}_{\mathrm{regular}}=\lambda\sum_{i=1}^{N}\left[\|\boldsymbol{F}_i\|_2^2\right]$$

式中: λ 是正则化权重, $\boldsymbol{F}_i$ 为多模态特征。

2.1.5 文本区域识别

本文利用区域候选网络 (region proposal network, RPN) 来识别和分类不同的文本区域, 在 RPN 网络中采用 3 种锚框比例 (1∶1、1∶2、2∶1) 处理不同大小的文档组件, 采样数为 256, 正样本比例为 0.5, 阈值设置为 0.5。在模型后处理阶段, 采用非极大值抑制 (non maximum suppression, NMS) 算法过滤掉 IoU 超过 0.5 的重叠锚框和分数小于 0.05 的预测锚框。对预测的所属类别的识别结果, 本文使用交叉熵损失 (cross-entropy, CE) 对模型进行优化:

$$\boldsymbol{L}_{\mathrm{class}}=-\sum_{i=1}^{N}y_i\log(\hat{y}_i)$$

式中: y_i 是真实类别标签的独热编码, $\hat{y}_i$ 是模型预测的类别概率。

2.2 关系预测网络

本文通过文本区域识别的结果, 基于 GNN 图神经网络, 采用将每个历史文档表示为图的建模策略来完成 ROP 任务, 将每个独立的不同类别的文本区域视为由其属性组成的特征节点, 利用其空间和语义信息构造边来表示文本区域之间的顺序关系。本文提出的关系预测网络如图 4 所示。

2.2.1 图表示文档

1) 节点的表示。利用前文提取到的多模态特征, 再叠加类别标签, 可构成图的节点特征, 表示为

$$\boldsymbol{E}_u=\mathrm{Concat}(\boldsymbol{F}_u,\boldsymbol{C}_u)$$

$$\boldsymbol{E}_u=\mathrm{Concat}(\boldsymbol{F}_u,\boldsymbol{C}_u)$$

式中: $\boldsymbol{F}_u$ 和 $\boldsymbol{C}_u$ 分别为第 u 个文本区域的多模态特

征和类别信息$\boldsymbol{E}_u$。之后通过使用 GNN 重构文档来学习节点表示，表示为

$$\boldsymbol{h}_u^G=\mathrm{GNN}(\boldsymbol{E}_u)$$

关于$\boldsymbol{h}_u^G$的详细信息在第 2.2.2 节介绍。

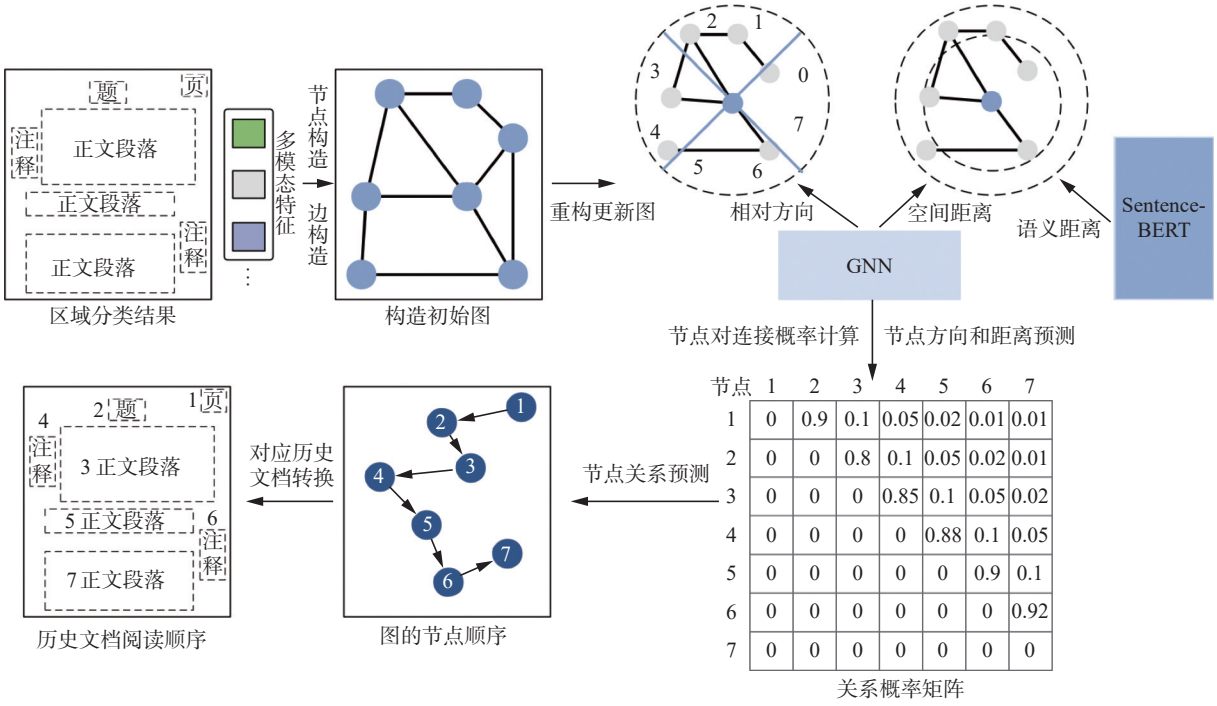


图 4　关系预测网络示意

Fig. 4　Schematic of relational prediction network

2) 边的表示。本文基于文本相似度和 D-LoS 方法[25]重建文本区域之间的距离和方向。将文本区域视作节点，D-LoS 方法在源节点周围的 360°方向上，水平划分出 8 个离散的 45°扇区。

具体地，两个节点 u 和 v 的边表示由相对距离 e_{dis} 和相对方向 e_{dir} 构成。两个节点之间的相对距离 e_{dis} 包括空间距离 $d_{u,v}^S$ 和语义距离 $d_{u,v}^T$。空间距离使用节点边界框中心点之间的欧几里得距离计算：

$$d_{u,v}^S=\sqrt{(x_u-x_v)^2+(y_u-y_v)^2}+\sqrt{(x'_u-x'_v)^2+(y'_u-y'_v)^2}$$

式中：(x,y)为原始二维坐标系上的坐标，(x',y')为第 2.1.3 节中变换后的坐标。对于节点之间的语义距离 $d_{u,v}^T$，首先计算节点文本嵌入的余弦相似度：

$$\mathrm{Sim}(u,v)=\frac{\boldsymbol{T}_u\cdot\boldsymbol{T}_v}{\|\boldsymbol{T}_u\|\cdot\|\boldsymbol{T}_v\|}$$

再转化为语义距离：

$$d_{u,v}^T=1-\mathrm{Sim}(u,v)$$

对空间距离和语义距离直接加权融合会导致某一项对最终结果的影响过大[26]，因此需进行放缩

$$e_{\mathrm{dis}}=\alpha\cdot\log(d_{u,v}^S+1)+\beta\cdot d_{u,v}^T$$

式中：$\log(d_{u,v}^S+1)$为平滑处理后的空间距离；α和β为权重系数，本文分别设置为 0.7 和 0.3。通过相对距离 e_{dis} 构建加权邻接矩阵 $\boldsymbol{A}$：

$$A_{u,v}=\begin{cases}\exp(-e_{\mathrm{dis}}),\text{如果节点}v\text{和}u\text{有边连接}\\0,\ \text{其他}\end{cases}$$

式中：$A_{u,v}$是邻接矩阵$\boldsymbol{A}$中节点 u 和 v 的边权重，两个节点之间的相对方向 $e_{\mathrm{dir}}\in\{0,1,\cdots,7\}$，边$(u,v)$的最终特征表示为

$$\boldsymbol{e}_{u,v}=[e_{\mathrm{dis}},e_{\mathrm{dir}}]$$

式中：e_{dis}是平滑后的相对距离，e_{dir}是离散的 8 个方向类别。

2.2.2　阅读顺序预测

GNN 的输入是编码的节点表示，通过 GNN 的消息传递机制进行更新。本文使用加权邻接矩阵$\boldsymbol{A}$和节点特征$\boldsymbol{E}_u$来生成节点表示$\boldsymbol{h}_u^G$：

$$\boldsymbol{h}_u^{G,(l+1)}=\mathrm{ReLU}(W^{(L)}\boldsymbol{h}_u^{G,(l)}+\sum_{v\in N(u)}A_{u,v}\cdot W^{(L)}\boldsymbol{h}_u^{G,(l)})$$

式中：$W^{(L)}$为第 l 层的权重，$N(u)$为节点 u 的邻居集合，经过多层消息传递后获得节点的最终表示$\boldsymbol{h}_u^G$。

本文在 Wang 等[25]的基础上，新提出关系概率预测任务，在 3 个任务上联合训练 GNN，分别是距离预测、方向预测和关系概率预测。对于距离预测，定义一个回归头，通过两个节点向量的点积得到标量值，并使用线性激活

$$\hat{d}_{u,v}=\mathrm{Linear}((\boldsymbol{h}_u^G)^{\mathrm{T}}\times\boldsymbol{h}_v^G)$$

使用均方误差损失 (mean squared error, MSE) 进行距离回归

$$\boldsymbol{L}_{\text{dis}}=\boldsymbol{L}^{\text{MSE}}\left(\hat{d}_{u,v},d_{u,v}\right)$$

对于方向预测, 定义一个分类头 $\hat{y}_{u,v}$, 根据两个节点之间的元素乘积为每个边缘分配 8 个方向之一。

$$\hat{y}_{u,v}=\sigma((\boldsymbol{h}_u^G\odot\boldsymbol{h}_v^G)\times W)$$

式中: $\boldsymbol{h}_u^G\odot\boldsymbol{h}_u^G$是两个节点之间的元素乘积, W是乘积向量的可学习权重, σ是非线性激活函数。使用交叉熵损失 (CE) 进行方向分类:

$$\boldsymbol{L}_{\text{dir}}=\boldsymbol{L}^{\text{CE}}\left(\hat{y}_{u,v},y_{u,v}\right)$$

对于关系概率预测, 通过节点表示和边特征生成最终的关系概率矩阵 $\boldsymbol{P}$, 其中 $\hat{p}_{u,v}$ 表示节点 u 和 v 之间的关系概率。首先结合节点表示和边特征生成边的特征表示 $\boldsymbol{f}_{u,v}$:

$$\boldsymbol{f}_{u,v}=\text{Concat}(\boldsymbol{h}_u^G,\boldsymbol{h}_v^G,\boldsymbol{e}_{u,v})$$

使用全连接层来生成关系概率 $\hat{p}_{u,v}$:

$$\hat{p}_{u,v}=\sigma(W_P\cdot\boldsymbol{f}_{u,v})$$

式中: W_P为全连接层的可学习权重, σ为 Sigmoid 激活函数, 将结果映射到开区间 (0,1)。使用二元交叉熵损失 (binary cross-entropy, BCE) 进行关系概率预测:

$$\boldsymbol{L}_{\text{rela}}=\boldsymbol{L}^{\text{BCE}}\left(\hat{p}_{u,v},p_{u,v}\right)$$

最终的阅读顺序预测损失为

$$\boldsymbol{L}_{\text{ROP}}=(\lambda_1\cdot\boldsymbol{L}_{\text{dis}}+\lambda_2\cdot\boldsymbol{L}_{\text{dir}}+\lambda_3\cdot\boldsymbol{L}_{\text{rela}})\cdot(1-r_{u,v})$$

式中: λ_1、λ_2、λ_3是 3 个可调的超参数, $r_{u,v}$是相对距离 e_{dis} 的标准化。对所有的节点对计算 $\hat{p}_{u,v}$, 生成完整的关系概率矩阵 $\boldsymbol{P}$:

$$\boldsymbol{P}=\begin{pmatrix}\hat{p}_{1,1} & \cdots & \hat{p}_{1,N}\\ \vdots & \ddots & \vdots\\ \hat{p}_{N,1} & \cdots & \hat{p}_{N,N}\end{pmatrix}$$

式中 N 是节点总数。利用生成的关系概率矩阵, 找到一条从起始节点到结束节点、包含所有节点的路径, 使得路径中所有边的概率乘积最大化, 该路径即为预测的文档阅读顺序。此要求对应于经典的图遍历搜索问题, 可通过各种现有算法有效解决。

3　实验与结果分析

3.1　数据集

近代历史手稿是历史文档的重要组成部分, 其与古代古籍和现代数字文档存在显著差异。图 5 为古籍与近代历史手稿的样例图, 可以看到古籍的结构相对固定, 有较为规范的章节划分和版式设计, 文字排列整齐, 而近代历史手稿的版面结构多变, 没有标准化格式, 存在非线性阅读顺序。

图 5　古籍 (上) 与近代历史手稿 (下) 的样例

Fig. 5　Sample drawings of ancient books (top) and modern historical manuscripts (bottom)

本文在 READ 2016 和 FCR_500 数据集上进行人工标注并进行扩充, 数据集详情统计如表 1 所示。

表 1　READ 2016 数据集和 FCR_500 数据集统计

Table 1　statistics of the READ 2016 dataset and the FCR_500 dataset

数据集	级别	训练集	验证集	测试集
READ 2016	文本行	8 367	1 043	1 140
	段落	1 602	182	199
	页面	350	50	50
FCR_500	文本行	22 477	6 505	3 195
	段落	3 594	1 023	486
	页面	173	44	25
	双页面	177	56	25

READ 2016 数据集由 1470—1805 年的会议纪要中选取的 450 页图像组成, 其文字排版并不规则, 行间距与字间距不均, 部分页面带有手写注释与修改痕迹; 纸张老化痕迹明显, 可看到渗透的墨水。FCR_500 数据集是由 19 世纪芬兰的地区法院记录中的 500 页图像组成, 页面形式涵盖单页和双页, 布局与排版富于变化。

3.1.1　历史文档区域识别图像数据集

区域识别图像数据集构造如图 6 所示, 左侧为原图像标签内容, 包括: 1) 文本行边框坐标; 2) 文本行转录内容; 3) 文本行顺序索引。

本文在原数据集基础上, 新增标签内容, 如图 6 右侧所示, 新增标签为: 4) 文本区域边框坐标, 文本区域通常由多个文本行组成, 其边框坐标有助于模型学习区分相邻区域; 5) 文本区域转录内容, 整合多行文本的语义信息, 通过语言模

型对区域转录内容进行编码，增强模型对语义关联的理解；6) 文本区域类型，标注每个文本区域的功能类别 (如正文、标题、注释等)，明确区域的语义角色。

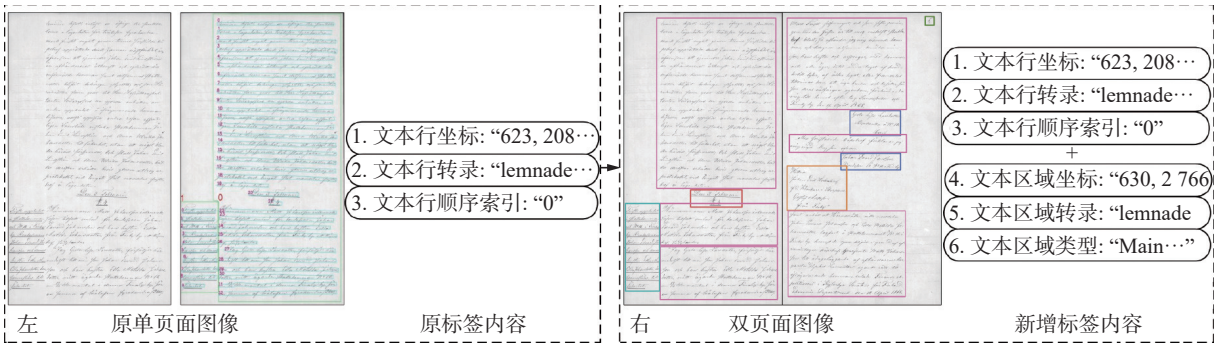


图 6　历史文档区域识别图像数据集构造

Fig. 6　Construction of historical document region recognition image dataset

3.1.2　关系预测样本数据集

关系预测样本数据集在上述区域识别图像数据集标签的基础上，新增了阅读顺序标签。原图像标签内容包括：1) 文本区域边框坐标；2) 文本区域转录内容；3) 文本区域类型。新增标签为：4) 文本区域顺序索引，为每个文本区域分配一个唯一的顺序编号，表示其在文档中的标准阅读顺序，将阅读顺序预测任务转化为序列生成问题。

将经过处理后的 READ 2016 数据集和 FCR_500 数据集分别称为 LA-READ 数据集和 LA-FCR 数据集，则 LA-READ 数据集和 LA-FCR 数据集的类别统计如表 2 所示。

表 2　数据集类别统计

Table 2　Category statistics of dataset

数据集	类别	数量
LA-READ	页面(P)	450
	页码(N)	450
	标题(T)	351
	正文(B)	918
	注释(A)	523
LA-FCR	页面(P)	500
	页码(N)	341
	标题(T)	853
	正文(B)	1 020
	注释(A)	773
	表格(TA)	263

3.2　模型训练

本文方法使用 PyTorch 框架实现，所有训练和实验均在 NVIDIA GTX 3090 GPU、Intel i9-12900K CPU 和 64 GB 内存中进行。历史文档区域识别模型以 ResNet50 网络为基础。使用公开的大型布局分析数据集对视觉模型进行预训练，使其具有初步的文档区域识别能力；文本特征提取器的参数使用预训练的 BROSBERT 模型进行初始化。模型使用 AdamW 优化器进行训练，图像批量大小设置为 8，动量参数设置为 0.99，初始学习率为 0.000 1。

3.3　评价指标

为了评价模型在文档区域识别数据集上的性能，本文选取边界框的类别和平均精度均值 mAP@IoU[0.50:0.95]，即交并比 (intersection over union, IoU) 阈值为 0.50~0.95 时的平均精度来作为评价指标，计算公式为

$$m_{\mathrm{AP}}=\frac{\sum_{i=1}^{C} I_{\mathrm{AP}}^{i}}{C}$$

式中：AP(average precision) 为精确率 P(Precision) 和召回率 R(Recall) 的平均精度，C 为文档布局分析任务中的类别数量。

阅读顺序的评价指标，本文遵循 Li 等[16] 的做法，使用总排序准确率 (total order accuracy, Acc) 和 BLEU 分数 (bilingual evaluation understudy score)[27] 来评估阅读顺序预测任务的性能。Acc 计算公式为

$$\mathrm{Acc}=\frac{\text{预测正确顺序的元素数}}{\text{总元素数}}\times 100\%$$

本文采用 BLEU-2 作为 BLEU 指标。设预测序列为 C，真实序列为 R，c 为预测序列 C 的长度，r 为真实序列 R 的长度，首先计算 2-gram 精度：

$$p_2=\frac{\mathrm{Count}_{\mathrm{clip}}(\text{2-gram},C,R)}{\mathrm{Count}_{\text{2-gram}}(C)}$$

式中：$\mathrm{Count}_{\text{2-gram}}(C)$为预测序列 C 中所有 2-gram 的数量，$\mathrm{Count}_{\mathrm{clip}}(\text{2-gram},C,R)$为预测序列 C 中在真实

序列 R 里能匹配到的 2-gram 的数量。再计算惩罚因子 (brevity penalty, BP):

$$I_{\mathrm{BP}}=\begin{cases}1, & 若 c \geqslant r\\ \mathrm{e}^{1-\frac{r}{c}}, & 若 c<r\end{cases}$$

最后计算 BLEU-2 分数:

$$I_{\mathrm{BLEU\text{-}2}}=I_{\mathrm{BP}}\times\sqrt{p_2}$$

3.4　实验结果

本文挑选多个方法与本文方法进行对比。单模态方法: MaskRCNN[28](mask region-based convolutional neural network), 是经典的两阶段目标检测框架, 代表传统视觉模态方法的性能基线; YOLOv8 (you only look once, YOLO), YOLOv8 作为广泛使用的一阶段目标检测模型, 可代表高效视觉模型的性能上限。双模态方法: BROS[10](BERT relying on spatiality), 通过结合文本和空间布局特征, 在表单理解任务中表现突出, 可作为文本−空间双模态方法的代表。多模态方法: LayoutLMv3[29] (layout-aware Transformer for document image understanding, LayoutLM), LayoutLMv3 模型统一使用 Transformer 来处理文本、图像和空间坐标, 所有模态直接交互, 代表多模态方法的最新进展。以上模型在 LA-READ 数据集和 LA-FCR 数据集的实验结果如表 3 所示。

表 3　各模型的文本区域识别实验结果

Table 3　experimental results of text region recognition for each model　%

模型	模态	mAP	
		LA-READ	LA-FCR
MaskRCNN[28]	视觉	84.2	77.4
YOLOv8		88.6	81.7
BROS[10]	文本+空间	83.3	76.6
LayoutLMv3[29]	视觉+文本+空间	90.3	83.6
MRNM(本文)		92.4	85.6

通过表 3 可以看出, 本文提出的 MRNM 模型在 LA-READ 数据集和 LA-FCR 数据集上的整体性能分别达到了 92.4% 和 85.6%, 和现有模型的最好结果相比, 识别精度分别提升了 2.1% 和 2.0%。加入文本和空间特征后, 模型在两个数据集上的准确率均有所提高, 说明在复杂场景下的文档布局分析任务中, 随着视觉特征提取能力的加强和文本及空间特征的加入, 模型的识别准确率逐步上升。

为了进一步体现所提出的 MRNM 模型对历史文档图像中每类标签元素的识别能力, 本文还针对 LA-FCR 数据集中的不同类别的元素进行识别对比实验, 实验结果如表 4 所示。

表 4　不同模型对各类元素识别的 mAP 值

Table 4　mAP values for identifying various elements by different models　%

模型	页码 (N)	标题 (T)	正文 (B)	注释 (A)	表格 (TA)
MaskRCNN	75.9	71.0	78.3	80.6	82.0
YOLOv8	81.7	77.5	82.4	83.2	83.9
Uni-modal Approaches	80.5	76.1	80.7	82.9	83.6
BROS	75.9	70.2	77.6	79.6	81.1
LayoutLMv3	84.0	79.7	83.2	85.4	86.2
MRNM(本文)	85.8	83.2	86.6	85.9	87.3

关于 ROP 任务的实验结果, 选择 3 种方法与本文提出的方法进行比较, 分别为: XY-Cut[14], GCN-PN(graph convolutional network-pointer network)[16] 和 Layoutreader[30]。XY-Cut 方法是经典的基于规则的阅读顺序方法; GCN-PN 是基于图卷积和指针网络的端到端模型; Layoutreader 是基于 Transformer 的序列到序列模型, 通过文本和布局特征来预测阅读顺序。以上方法为研究 ROP 任务的 3 类主流技术路线。实验结果如表 5 所示。

表 5　各方法关于阅读顺序预测的实验结果

Table 5　Experimental results of each method on reading order prediction

方法	Acc/%		BLEU-2/%	
	LA-READ	LA-FCR	LA-READ	LA-FCR
XY-Cut[14]	58.2	51.8	62.6	55.2
GCN-PN[16]	69.1	67.4	71.7	70.9
Layoutreader[30]	73.6	76.5	75.4	77.5
MRNM(本文)	77.1	79.2	78.1	80.4

通过表 5 可以看出, 基于规则的 XY-Cut 方法不及其他基于深度学习的方法, 说明了在 ROP 任务中面对具有复杂布局的历史文档时, 基于规则的解决方法存在一定局限性。而本文所提出的方法在 LA-READ 数据集和 LA-FCR 数据集上关于 ROP 任务的预测结果, 总排序准确率 Acc 分别达到了 77.1% 和 79.2%, BLEU-2 分数分别达到了 78.1% 和 80.4%, 与上述表现最好的模型相比, 在两个指标上的预测准确率分别提升了 3.1% 和 2.8% 左右, 证明了利用图神经网络和语言模型来进行历史文档 ROP 任务的可行性和有效性。Layoutreader 表现不佳的主要原因是历史文档中具有大量

长文本段，基于序列的生成模型难以充分应对这种场景。

3.5　消融实验

为了验证分层设计和注意力机制在文本区域识别任务中的有效性，构建一个 Baseline 基础模型，将文本区域的所有特征看作单一层级的特征，取消文本行和文本段的分层设计，对所有文本区域统一应用通道注意力机制 (channel attention mechanism, CAM) 或空间注意力机制 (spatial attention mechanism, SAM)，详细对比结果如表 6 所示。

表 6　关于文本区域识别的消融实验结果

Table 6　Results of ablation experiments on text region recognition　%

方法	+分层	+CAM	+SAM	+All	mAP				
					页码(N)	标题(T)	正文(B)	注释(A)	表格(TA)
Baseline	—	—	—	—	84.48	81.27	85.61	84.60	83.25
+分层	√	—	—	—	84.92	82.05	86.14	85.32	86.79
+CAM	√	√	—	—	85.53	82.81	85.79	85.57	86.83
+SAM	√	—	√	—	85.27	82.63	86.32	85.61	87.14
+All	√	√	√	√	85.76	83.18	86.58	85.94	87.26

通过表 6 可以看出，经过分层设计的模型可以更好地区分出不同类别的文本区域。在文本行区域添加 CAM 机制后，在识别页码和标题这类较小的文本区域时，识别准确率获得了明显的提高，CAM 能够有效地强调关键特征，从而帮助提高小目标的识别准确率。而在文本段区域添加 SAM 后，通过其中的可变性卷积可以捕获更多的空间上下文信息，从而帮助识别正文、注释和表格这类上下文对象。在最后一组消融实验中，通过对提出的方法进行整合，最终与 Baseline 基础模型相比，在历史文档的文本区域识别任务中平均识别准确率提高了 1.9% 左右。

为了验证 GNN 和语义相似度的有效性，进行消融实验，详细对比结果如表 7 所示。

表 7　关于阅读顺序预测的消融实验结果

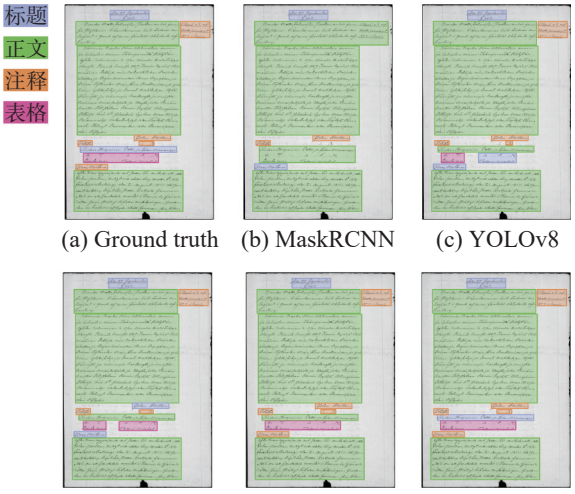
Table 7　Results of ablation experiments on reading order prediction　%

方法	+距离	+坐标转换	+语义	+GNN	Acc	BLEU-2
+距离	√	—	—	—	65.6	68.3
+坐标转换	√	√	—	—	67.2	69.7
+语义	—	—	√	—	39.5	42.3
+语义 (距离)	√	√	√	—	68.4	71.1
+GNN (距离)	√	√	—	√	75.7	77.2
+All	√	√	√	√	78.1	79.6

通过表 7 可以看出，采用几何信息进行阅读顺序预测具有一定的效果，且叠加语义信息能提高准确率。将提出的所有方法进行整合，在本文数据集上执行 ROP 任务的准确率获得较大提升。

3.6　可视化分析

为了更好地展示 MRNM 的性能，本文进行了结果可视化对比。从图 7 可以看出，本文提出的 MRNM 模型能更精准地识别文档中的不同区域，分类结果更准确，接近真实标注。



(a) Ground truth　(b) MaskRCNN　(c) YOLOv8

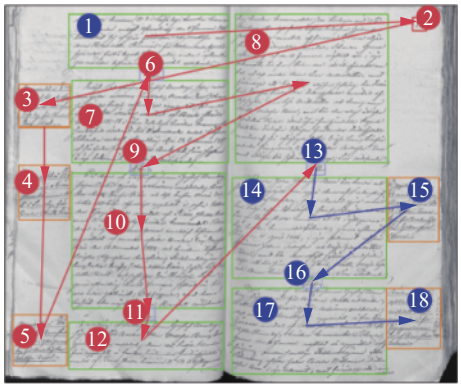
(d) BROS　(e) LayoutLMv3　(f) MRNM (本文)

图 7　不同模型的布局分析效果比较

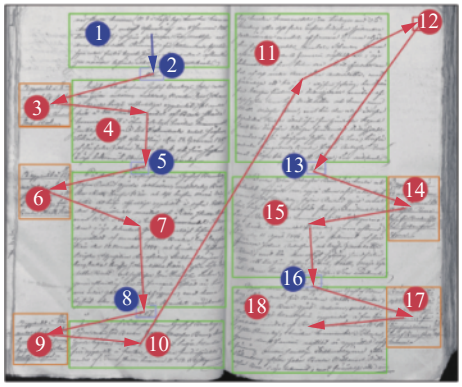
Fig. 7　Comparisons of layout analysis effects of different models

不同模型在 LA-FCR 数据集上的阅读顺序预测结果如图 8 所示，红色表示错误结果，蓝色表示正确结果。从图中可以看到，基于规则的 XY-Cut 方法存在较大的局限性，通常仅适用于具有简单文本布局的场景。GCN-PN 方法在处理文档时表现出的性能不佳，因为其处理规则主要针对分布较规整的电子文档。而基于序列的 Layoutreader 生成模型难以充分处理具有大量文本实例的场景。本文提出的模型在面对具有复杂布局和大量

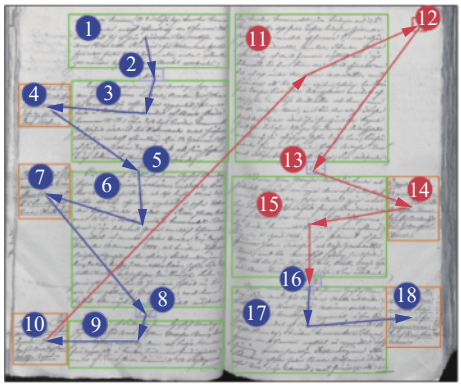
文本的近代历史手稿时, 能较好地预测出符合阅读顺序的文本序列。



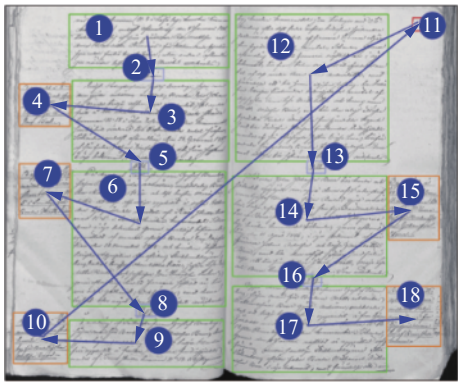
(a) XY-cut



(b) GCN-PN



(c) LayoutReader



(d) MRNM (本文)

图 8　不同模型的阅读顺序预测效果比较

Fig. 8　Comparisons of layout analysis effects of different models

4　结束语

本文针对历史文档布局分析任务中数据集缺乏、识别准确率较低的问题, 标注并构造了历史文档数据集, 提出了基于多模态和关系预测网络的布局分析模型 MRNM。多模态网络通过分层设计和注意力机制, 分别对文本行和文本段的特征进行优化, 增强了模型对局部细节和全局上下文信息的捕捉能力; 关系预测网络结合语义相似度和空间邻接关系, 利用图神经网络建模文本区域间的复杂关系, 提升了阅读顺序预测的准确性。未来的工作将尝试对更多类型的文档进行布局分析任务, 进一步提升模型的泛化能力。

参考文献:

[1] 王军, 杨海峥, 刘石, 等. 系列笔谈之一: 智能时代古典文献学的机遇与挑战[J]. 数字人文, 2022(2): 108−132.
WANG Jun, YANG Haizheng, LIU Shi, et al. The first of a series of pen talks: opportunities and challenges of classical philology in the intelligent age[J]. Digital humanities, 2022(2): 108−132.

[2] 刘成林, 金连文, 白翔, 等. 文档智能分析与识别前沿: 回顾与展望[J]. 中国图象图形学报, 2023, 28(8): 2223−2252.
LIU Chenglin, JIN Lianwen, BAI Xiang, et al. Frontiers of intelligent document analysis and recognition: review and prospects[J]. Journal of image and graphics, 2023, 28(8): 2223−2252.

[3] TANG Y Y, LEE S W, SUEN C Y. Automatic document processing: a survey[J]. Pattern recognition, 1996, 29(12): 1931−1952.

[4] GAO Liangcai, YI Xiaohan, JIANG Zhuoren, et al. ICDAR2017 competition on page object detection[C]//2017 14th IAPR International Conference on Document Analysis and Recognition. Kyoto: IEEE, 2017: 1417−1422.

[5] 王尚荣. 文档布局分析的多模态学习方法研究与实现[D]. 北京: 北京邮电大学, 2023.
WANG Shangrong. Research and implementation of multimodal learning method for document layout analysis[D]. Beijing: Beijing University of Posts and Telecommunications, 2023.

[6] SASSIOUI A, BENOUINI R, EL OUARGUI Y, et al. Visually-rich document understanding: concepts, taxonomy and challenges[C]//2023 10th International Conference on Wireless Networks and Mobile Communications. Istanbul: IEEE, 2023: 1−7.

[7] AFZAL M Z, KÖLSCH A, AHMED S, et al. Cutting the error by half: investigation of very deep CNN and advanced training strategies for document image classification[C]//2017 14th IAPR International Conference on Document Analysis and Recognition. Kyoto: IEEE, 2017: 883−888.

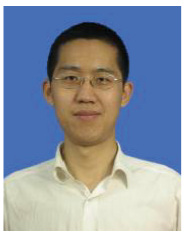
[8] DAVIS B, MORSE B, PRICE B, et al. End-to-end document recognition and understanding with dessurt[M]//Computer Vision–ECCV 2022 Workshops. Cham: Springer Nature Switzerland, 2023: 280−296.

[9] BAKKALI S, MING Zuheng, COUSTATY M, et al. VL-CDoC: vision-language contrastive pre-training model for

cross-modal document classification[J]. Pattern recognition, 2023, 139: 109419.
[10] HONG T, KIM D, JI M, et al. BROS: a pre-trained language model focusing on text and layout for better key information extraction from documents[EB/OL]. (2021−09−10)[2024−01−01]. https://arxiv.org/abs/2108.04539.
[11] HUANG Yupan, LYUengchao, CUI Lei, et al. LayoutLMv3: pre-training for document AI with unified text and image masking[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022: 4083−4091.
[12] LI Xin, ZHENG Yan, HU Yiqing, et al. Relational representation learning in visually-rich documents[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022: 4614−4624.
[13] ZHANG Chong, TU Yi ZHAO Yixi, et al. Modeling layout reading order as ordering relations for visually-rich document Uunderstanding[EB/OL]. (2024−09−29)[2025−01−01]. https://arxiv.org/abs/2409.19672.
[14] AIELLO M, SMEULDERS A M W. Bidimensional relations for reading order detection[M]//EPRINTS-BOOK-TITLE. University of Groningen, Johann Bernoulli Institute for Mathematics and Computer Science, 2003.
[15] FERILLI S, GRIECO D, REDAVID D, et al. Abstract argumentation for reading order detection[C]//Proceedings of the 2014 ACM Symposium on Document Engineering. Fort Collins: ACM, 2014: 45−48.
[16] LI Liangcheng, GAO Feiyu, BU Jiajun, et al. An end-to-end OCR text re-organization sequence learning for rich-text detail image comprehension[M]//Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 85−100.
[17] QUIRÓS L, VIDAL E. Reading order detection on handwritten documents[J]. Neural computing and applications, 2022, 34(12): 9593−9611.
[18] 马伟洪. 面向古籍文档分析的文字检测识别与阅读顺序理解[D]. 广州: 华南理工大学, 2022.
MA Weihong. Character detection, recognition and reading order comprehension for historical document analysis [D]. Guangzhou: South China University of Technology, 2022.
[19] ZHANG Chong, GUO Ya, TU Yi, et al. Reading order matters: Information extraction from visually-rich documents by token path prediction[EB/OL]. (2023−10−17)[2025−01−01]. https://arxiv.org/abs/2310.11016.
[20] QIAO Liang, LI Can, CHENG Zhanzhan, et al. Reading order detection in visually-rich documents with multi-modal layout-aware relation prediction[J]. Pattern recognition, 2024, 150: 110314.
[21] 温绍杰, 吴瑞刚, 冯超文, 等. 基于 Transformer 的多模态级联文档布局分析网络[J]. 浙江大学学报 (工学版), 2024, 58(2): 317−324, 369.
WEN Shaojie, WU Ruigang, FENG Chaowen, et al. Multimodal cascaded document layout analysis network based on Transformer[J]. Journal of Zhejiang University (engineering science), 2024, 58(2): 317−324, 369.
[22] WANG Jiawei, HU Kai, ZHONG Zhuoyao, et al. Detect-order-construct: a tree construction based approach for hierarchical document structure analysis[J]. Pattern recognit, 2017, 156: 110836.
[23] REIMERS N, GUREVYCH I. SENTENCe-BERT: sentence embeddings using siamese BERT-Networks[EB/OL]. (2019−08−27)[2025−01−01]. https://arxiv.org/abs/1908.10084.
[24] XU Canhui, LI Yuteng, SHI Cao, et al. HiM: hierarchical multimodal network for document layout analysis[J]. Applied intelligence, 2023, 53(20): 24314−24326.
[25] WANG Dongsheng, MA Zhiqiang, NOURBAKHSH A, et al. DocGraphLM: documental graph language model for information extraction[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei: ACM, 2023: 1944−1948.
[26] 徐捷, 邵玉斌, 杜庆治, 等. 结合混合特征提取与深度学习的长文本语义相似度计算[J]. 计算机工程与科学, 2024, 46(8): 1513−1520.
XU Jie, SHAO Yubin, DU Qingzhi, et al. Long text semantic similarity calculation combining hybrid feature extraction and deep learning[J]. Computer engineering & science, 2024, 46(8): 1513−1520.
[27] PAPINENI K. BLEU: a method for automatic evaluation of MT[R]. New York: IBM, 2001.
[28] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2980−2988.
[29] HUANG Yupan, LYU ngchao, CUI Lei, et al. LayoutLMv3: pre-training for document AI with unified text and image masking[C]//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022: 4083−4091.
[30] WANG Z, XU Y, CUI L, et al. Layoutreader: Pre-training of text and layout for reading order detection[EB/OL]. (2021−08−26)[2025−01−01]. https://arxiv.org/abs/2108.11591.

作者简介：

彭阳，硕士研究生，主要研究方向为目标检测、人工智能。E-mail: pengyang@mail.scuec.edu.cn。



王德军，副教授，博士，主要研究方向为人工智能、信息安全、大数据处理。获得湖北省科技进步二等奖 2 项。发表学术论文 13 篇。获国家发明授权专利 5 项。E-mail: dejun@scuec.edu.cn。

孟博，教授，博士，主要研究方向为人工智能安全与隐私、网络与系统安全。主持和参与国家和省部级科研项目 10 余项，横向课题 10 余项；获湖北省科技进步二等奖和一等奖各 1 项。发表学术论文 60 余篇；在科学出版社出版专著 3 部；申请国家发明专利 40 余项，其中获授权 20 余项；获软件著作权 20 余项。E-mail: mengscuec@gmail.com。