

基于多模态与图注意力网络的表格结构识别算法研究

徐崇善¹, 吴绿², 李宁¹, 徐宁², 潘昊¹

(1. 武汉理工大学计算机与人工智能学院, 湖北 武汉 433000; 2. 武汉理工大学信息工程学院 湖北 武汉 433000)

摘要: 表格作为信息的高度浓缩和结构化呈现方式, 广泛存在于各类文档和图像中。准确理解表格结构是从非结构化数据向结构化数据转化的关键一步。由于表格形状、表头配置、多行单元格以及分隔线样式的复杂性和多样性, 尤其是在处理缺乏边框的表格时, 这项任务既重要又充满挑战。针对无边框表格, 提出了一种新的表格结构理解方法, 主要有以下几项关键创新: 首先, 为了提高模型的泛化能力, 引入了一种新的多模态编码器, 用于融合表格的视觉、文本、布局 and 边框特征。其次, 提出了基于文本位置分布的框线添加算法, 以完善表格边框。最后, 在编-解码器架构中创造性地引入了图注意力网络, 使得节点在感知全局节点特征的同时, 也能更好地感知相邻节点的特征。该算法显著提升了无边框表格结构识别的准确性, 相较于当前已有方法显著提高了树编辑距离相似度, 在 PubTabNet 公开数据集上进行了测试, 简单表格上达到 99.46%, 复杂表格上达到 96.65%, 无边框表格上达到了 98.34%。

关键词: 表格结构理解; 无边框表格; 多模态融合; 图注意力网络。

doi: 10.11959/j.issn.2096-3750.XXXX.

Table Structure Recognition via Multimodality and Graph Attention Networks

XU Chongshan¹, WU Lv², XU Ning¹, LI Ning², PAN Hao¹

1. School of Computer and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China

2. School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China

Abstract: Tables, as a highly condensed and structured form of information presentation, are ubiquitous across documents and images. Accurate table structure recognition is a critical step in converting unstructured visual data into machine-interpretable structured formats. This task remains particularly challenging due to the wide variability in table layouts — including irregular shapes, complex headers, spanning cells, and diverse ruling-line styles — with borderless tables posing the most formidable difficulties. To tackle borderless tables effectively, we propose a novel structure understanding framework featuring three key innovations. First, we introduce a new multimodal encoder that fuses visual, textual, layout, and ruling-line features, thereby significantly improving model generalization across diverse table appearances. Second, we develop an efficient border reconstruction algorithm based on the statistical distribution of text positions. By exploiting horizontal and vertical alignment patterns of cell contents, the method reliably infers missing ruling lines. Third, we incorporate a Graph Attention Network (GAT) module into the encoder-decoder architecture. This allows each table element to attend to both global table context and fine-grained local relationships with neighboring cells, yielding more coherent structure predictions. Experiments on the PubTabNet benchmark demonstrate the superiority of the proposed approach. It achieves state-of-the-art Tree-Edit-Distance-based Similarity (TEDS) scores: 99.46% on simple tables, 96.65% on complex tables, and notably 98.34% on borderless tables — representing substantial gains over prior methods.

Key words: Table structure understanding, borderless table, multimodal fusion, graph attention network

收稿日期: XXXX-XX-XX; 修回日期: XXXX-XX-XX

通信作者: 吴绿, wulv@whut.edu.cn

基金项目: 国家自然科学基金资助项目 (No. 92373102)

Foundation Items: The National Natural Science Foundation of China (No.92367302)

0 引言

表格作为信息时代重要的结构化数据载体,广泛存在于学术文献、财务报表、行政文档等各类场景中。其通过行列布局实现数据的规范化组织,有效提升信息检索与对比分析的效率,是文档自动化处理的关键步骤^[1]。表格结构识别(Table Structure Recognition, TSR)是指通过计算机技术自动分析表格图像或文档,提取表格的逻辑结构与内容信息,并将其转化为机器可理解、可编辑的结构化数据(如 Excel、JSON、HTML 等格式)的技术过程^[2]。

表格结构的复杂性给其解析带来了诸多挑战。表格在形状上具有多样性,从简单的规则表格到复杂的异形表格,形式各异;表头配置也极具灵活性,不同的表头设计对应着不同的数据逻辑;多行单元格合并的情况增加了单元格位置和内容关联的判断难度;分隔线样式差异,包括有线、无线、虚线等,也使得表格结构的识别变得复杂^[3]。特别是在无边框表格场景下,由于缺乏明确的框线作为结构标识,现有方法难以准确解析表格结构,导致解析效果不理想。这就迫切需要一种高效的表格结构识别技术,以满足日益增长的文档自动化处理需求^[4]。

当前该领域仍面临三大难点:其一,复杂表格结构的解析能力仍存在明显不足。例如,斜线表头中,文字沿斜线排列,传统算法难以准确提取表头信息并理解其与表格内容的对应关系;大面积合并单元格使得单元格的跨行跨列属性判断困难,目前的识别主要依赖手工设计的传统算法,这些算法通用性差,难以适应复杂多变的表格结构^[5]。

其二,现有识别方法多依赖线条检测或文本对齐策略。但在无框线表格场景中,由于缺乏明显的线条特征,线条检测方法往往失效;而文本对齐策略容易受到文本排版、字体大小等因素的影响,导致行列划分出现偏差。在 CVPR 2023 会议上提出的 VAST 模型^[6](Kasem et al., 2023)通过引入视觉对齐损失函数,在一定程度上提升了边界框预测精度,但在面对复杂场景时,如表格内容存在大量干扰信息、文本布局不规则等情况,其性能仍有较大的优化空间。

其三,图像噪声与形变干扰问题突出:扫描文

档中的模糊、歪斜或低对比度现象,会显著降低光学字符识别(Optical Character Recognition, OCR)与表格结构重建的准确性,而现有预处理技术对复杂背景的鲁棒性仍显不足^[7]。

本文重点聚焦于复杂表格结构解析和缺失框线表格识别精度这两个关键问题,旨在提出一种鲁棒的表格结构识别算法——TableGatFormer 模型,以实现更准确、高效的表格结构识别,提升表格在各类复杂场景下的解析效果。本文主要贡献如下。

(1) 为提高模型的泛化能力,将编解码架构中广泛使用的 CNN 编码器更换为多模态编码器。该多模态编码器融合了文本/框线区域图片信息、文本信息、文本/框线以及对应的坐标信息,能够充分利用多种模态的信息,显著提高了不同场景下表格结构的预测精度。

(2) 提出了基于文本位置分布的框线生成算法,通过分析表格中文字的位置分布规律,生成虚拟框线,并根据生成的框线对表格区域进行划分。以表格中的各种元素为节点,区域之间的关系构造边建立图结构,从而更好地表示表格的逻辑结构。

(3) 在编码-解码架构中引入了图注意力网络(Graph Attention Network, GAT)模块,该模块能够使节点更好地感知临近节点的特征,通过信息传递和特征融合,显著提高表格结构的预测精度。

(4) 本文提出的基于多模态与图注意力网络的表格结构识别模型(TableGatFormer, TGF)是一个基于编码-临近感知-解码架构的多模态模型,它可以融合表格图片、文本信息以及布局信息,全面地对表格结构进行预测。通过对比实验证明了本文的方法在 PubTabNet 中能达到最佳性能,消融实验也有效证明了本文方法的有效性。

1 相关工作

1.1 基于图神经网络的表格结构识别方法

基于图神经网络(Graph Neural Network, GNN)的表格结构识别方法,通过精准建模表格元素间的内在拓扑关系,这类方法的核心在于将文本块或检测到的单元格抽象为图节点,并利用其对应的视觉特征作为初始节点表示。Qasim 等人^[8]采用基于空间邻近度的启发式方法,计算节点中心点间的欧氏距离或区域边界框的交并比(Intersection over Union, IoU)来构建初始局部邻接关系;Li 等

人^[9]提出了更精细化的建模策略,他们设计了多模态图神经网络模型 TSRNet (Table Structure Recognition Network),该模型有效地融合了视觉特征和文本语义信息,以更全面地学习节点间潜在的结构关系。

近年来,图神经网络在表格结构识别中的应用取得了显著进展。Liu 等人^[10]提出了神经协作图机器 (Neural Collaborative Graph Machines, NCGM),该模型配备了堆叠协作块,能够以高效的层次化方式交替提取模态内上下文并建模模态间交互。与传统方法相比, NCGM 能够提供更稳健的表格元素模态内-间关系表示。孙俊和苟刚^[11]提出了融合图卷积网络的双分支识别网络模型,该模型以 ResNet+FPN 为主干网络,并创造性地引入矩阵分解头代替注意力机制重整全局特征。该模型通过图卷积网络有效感知单元格间连接关系,从而辅助输出位置信息和逻辑邻接关系,在多个数据集上 F1 指标平均提升 10.6%。此外,多模态图注意力网络如 Connectome GAT (CGAT) 在结构化数据诊断中通过图注意力网络^[12]融合多模态特征,提升了对复杂关系的捕捉能力,可扩展至表格结构识别^[13]。

1.2 基于编解码架构的表结构识别方法

基于编码器-解码器 (Encoder-Decoder) 的表格结构识别方法通过全局自注意力机制突破了传统序列建模的局部感知限制,并将表格结构解析重构为序列生成任务,其核心范式可归纳为: □ 基于编码器-解码器架构对表格元素的跨模态特征进行联合建模; □ 通过自回归解码生成结构化标记序列 (如 HTML 标签、LaTeX 代码)。自 Vaswani 等人^[14]提出 Transformer 基础架构以来, Wang 等人^[15]于 2021 年率先提出基于树形解码器的表格专用 Transformer 模型,该模型通过递归生成 HTML 标签树显式建模单元格的层次化隶属关系,首次为解决跨行列表合并等复杂结构提供新范式。Alibaba 团队^[16]提出的 Master 采用 CNN-Transformer 混合编码器,通过卷积神经网络提取单元格区域的高粒度视觉特征,并与 OCR 文本嵌入进行跨模态对齐,显著提升了模型对扫描文档表格的鲁棒性,充分体现了多模态特征融合的优势。IBM 团队^[17]提出的双流解码器架构引入结构-坐标解耦机制,分别使用结构解码器预测表格拓扑,坐标解码器回归单元格物理位置,该方法在 PubTabNet、FinTabNet 等基准数据集

上实现了当时的 SOTA 性能。

近期研究进一步改进了编码器-解码器框架。Yang 等人^[18]提出了局部关系感知 Transformer 网络 LRATNet (Local-Relationship-Aware Transformer Network),通过其核心的 LRAT 模块构建局部和全局信息之间的稳健关联,旨在通过探索不同的邻接关系提高逻辑位置预测性能。Xiao 等人^[19],通过优化富文本文档的处理,提升了复杂表格的准确性。

1.3 多模态融合与预训练技术

多模态融合与预训练技术的发展为表格结构识别带来了新的研究突破。Herzig 等人^[20]提出了 TAPAS,通过预训练实现了弱监督表格解析,该方法扩展了 BERT 架构以编码表格作为输入,并得益于从维基百科爬取的文本片段和表格的联合预训练进行初始化。Yin 等人^[21]提出了 TaBERT,其在 2600 万个表格及其英语上下文的大型语料库上进行训练,显著提升了语义解析性能。近期研究进一步探索了视觉和语言模型的融合。Zhang 等人^[22]提出了 UniTabNet,这是一个基于图像到文本模型的表格结构解析框架,它采用“分而治之”策略,利用图像到文本模型解耦表格单元格,并集成物理和逻辑解码器重建完整表格结构。该模型引入了视觉引导器和语言引导器,以分别提高模型对相关区域的关注和对表格图像中文本语义的理解能力。Peng 等人^[23]提出了 UniTable,它通过自监督预训练统一了表格识别的训练范式和目标,将像素级输入的简单性与自监督预训练的有效性和可扩展性相结合,在多个大型表格识别数据集上取得了 SOTA 性能。

表格结构识别领域的研究仍处于快速发展阶段,尽管现有各类方法均具备其独特优势,但仍存在亟待解决的挑战。基于 Transformer 的模型虽展现出优异的长程依赖建模能力,但其固有的全局注意力机制缺乏对局部拓扑结构的敏感度,这一不足仍需重点关注。表格单元格间的逻辑关联强烈依赖于局部邻域上下文信息;因此,全局注意力在计算过程中,可能因过度关注稀疏的长程关联,最终导致关键的局部结构特征被稀释。^[24]鉴于上述挑战并弥补现有方法的局限性,本文提出了一种新颖且高效的混合架构——基于图注意力网络增强的 Transformer 模型 TabGatformer,用于解决复杂的表格结构识别任务。TabGatformer 巧妙地结合了 GAT 局部感知能力与 Transformer 的全局建模优势;同时,

为进一步提升模型的鲁棒性与泛化能力, 本文在该模型中创新性地引入了多模态融合编码模块。

2 基于多模态与临近感知的表格结构识别

本文提出的模型主要由编码器、图注意力网络模块和解码器三个主要部分组成, 如图 1 所示。模型采用编码器-解码器架构, 并引入 GAT 模块处理表格单元格之间的空间关系。

2.1 多模态编码器

多模态编码器 (Multi-Modal Encoder, MME) 的输入包括三个部分: 坐标信息、文本信息以及按照坐标进行切分后的图片集 (图片大小为 56×56)。编码器由五个部分构成, 分别是视觉特征提取器, 文本信息特征提取器、几何位置编码器、多模态特征融合器以及 Transformer 编码层。

视觉特征提取器 本质上是一个 ResNet-18 网络, 第一层使用 7×7 卷积核 (步长 2, padding 3) 将输入通道从 3 扩展至 64, 经 ReLU 激活和 3×3 最大池化 (步长 2, padding 1) 后, 输出 $56 \times 56 \times 64$ 的特征图。第一组残差块 (2 个): 保持通道数 64 不变, 输出 $56 \times 56 \times 64$ 特征图: 通过步长 2 的卷积将通道数从 64 提升至 128, 输出 $28 \times 28 \times 128$ 特征图。第三组残差块 (2 个): 进一步将通道数从 128 提升至 256, 输出 $14 \times 14 \times 256$ 特征图。第四组残差块 (2 个): 最终将通道数从 256 提升至 512, 输出 $7 \times 7 \times 512$ 特征图。经全局平均池化将 $7 \times 7 \times 512$ 特征图压缩为 512 维向量, 再通过全连接层将 512 维特征映射至 d_h 维度, 得到视觉特征 $F_v \in R^{B \times \max_len \times d_h}$, 其中 B 为 batch_size, $\max_len$ 为序列最大长度, 后续公式中出现的 B 、 $\max_len$ 均作此解释。

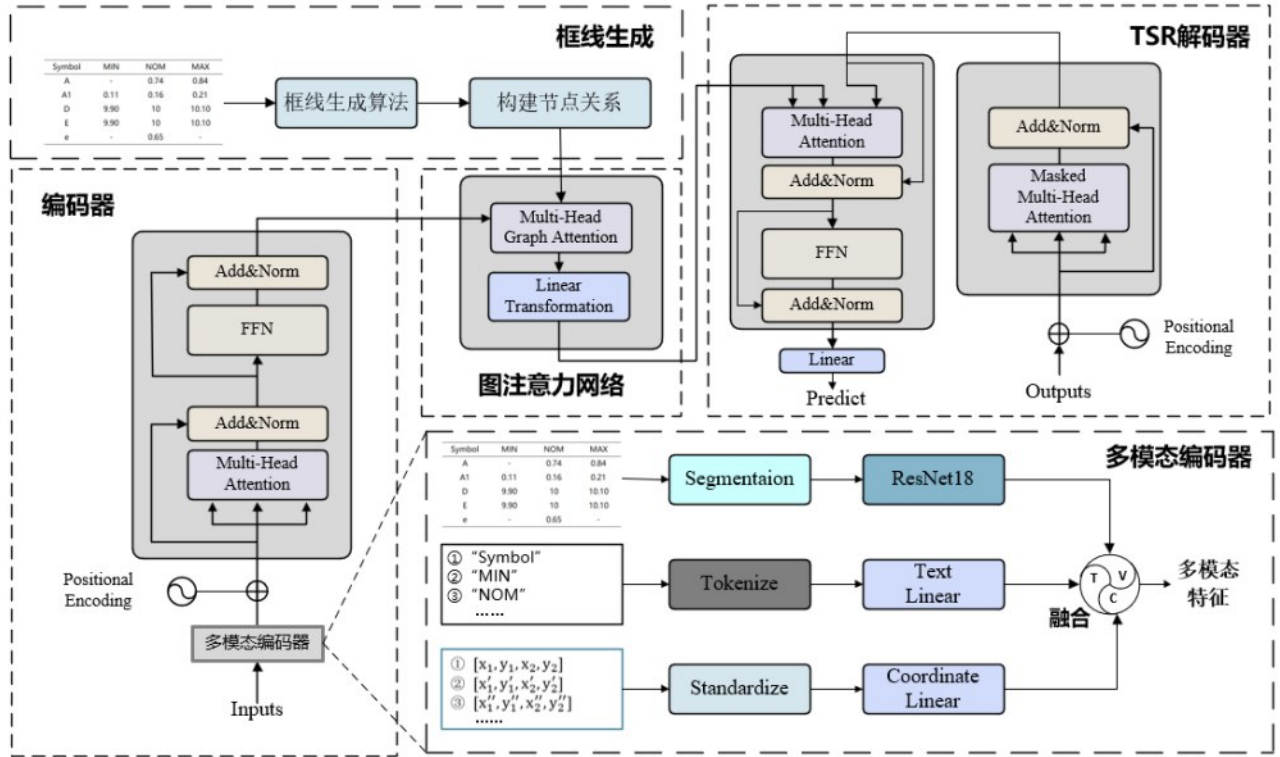


图 1 TableGatformer 模型结构

文本信息特征提取器 构建层次化文本表征管道, 首先对 OCR 识别文本 $Text_{OCR}$ 进行规范化分词处理, 通过预先收集的字典生成字符级嵌入向量 $Text_{emb}$ 。

$$Text_{emb} = \text{Tokenize}(Text_{OCR}) \quad (1)$$

随后经线性层 Linear_t 进行特征空间对齐, 通过非线性激活函数 ReLU 与层归一化 LayerNorm 得

到最终文本特征 $F_t \in R^{B \times \max_len \times d_h}$, 文本特征 F_t 保留细粒度语义信息。

$$F_t = \text{LayerNorm}(\text{ReLU}(\text{Linear}_t(Text_{emb}))) \quad (2)$$

其中层归一化 LayerNorm 过程如式 (3) 所示。其中 x 为输入特征。 μ 为 x 在其最后一个维度上的均值, σ^2 为 x 的方差, ε 为常数防止除零, 设置为 $1e-5$ 。 γ 和 β 为可学习参数向量。非线性激活函数

ReLU过程如式(4)所示。

$$\text{LayerNorm}(\mathbf{x}) = \gamma \cdot \frac{\mathbf{x} - \boldsymbol{\mu}}{\sqrt{\boldsymbol{\sigma}^2 + \varepsilon}} + \boldsymbol{\beta} \quad (3)$$

$$\text{ReLU}(\mathbf{x}) = \max(0, \mathbf{x}) \quad (4)$$

几何位置编码器 设计共享式坐标编码架构，对文本块与框线的几何信息进行联合建模：将原始坐标 (x_1, y_1, x_2, y_2) 转化为图像尺寸的相对坐标 *Coord* ($[\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}]$)，其中 W, H 分别为图像的宽和高，再通过线性层 Linear_c 进行特征空间对齐，然后通过ReLU非线性激活函数与层归一化函数得到几何位置特征 $\mathbf{F}_p \in R^{B * \max_len * d_h}$ 。

$$\mathbf{F}_p = \text{LayerNorm}(\text{ReLU}(\text{Linear}_c(\text{Coord}))) \quad (5)$$

该设计通过参数共享机制，使文本与框线在统一几何空间中对齐，增强拓扑关系感知能力。

多模态特征融合 将视觉特征 $\mathbf{F}_v$ 、文本特征 $\mathbf{F}_t$ 和几何位置特征 $\mathbf{F}_p$ 沿特征维度拼接，构建联合跨模态表征；随后通过线性变换将高维联合特征压缩至目标维度空间，得到表格的多模态特征 $\mathbf{F}_{enc}$ 。该过程可形式化表示为式(6)，其中 $\mathbf{W}_f$ 为投影矩阵， $\mathbf{b}_f$ 为偏移量， $\mathbf{F}_{enc} \in R^{B * \max_len * d_{model}}$ ， d_{model} 为输入编码器时特征的维度。

$$\mathbf{F}_{enc} = \mathbf{W}_f [\mathbf{F}_v || \mathbf{F}_t || \mathbf{F}_p] + \mathbf{b}_f \quad (6)$$

2.2 框线生成算法

表格框线通常与表格结构一起出现，指在表格中用于分隔和界定单元格的线条，为了使单元格节点能够更好的感知临近区域的几何特征，本文在模型中引入了表格框线的视觉特征和位置特征，在编码器中将这些特征进行融合。从PubTabNet提供的表格图片中提取框线坐标时，本文发现绝大多数表格都是类似三线表的形式存在的，图片中仅有少量的线条，并没有完整的框线结构，仅提取这些零星分布的线条显然不能达到本文设计这个模块的目的。因此本文提出了基于文本块位置分布的框线生成算法。具体算法如算法1所示。

算法1 框线生成算法伪代码

输入： 图像数组 Image_{np} 、框线生成方向 *direction* (0表示竖直方向，1表示水平方向)

输出： addLine 生成的框线集合

$$\mathbf{I}_{bin} = \text{Binarization}(\text{Image}_{np})$$

/*对输入图像 Image_{np} 进行灰度化、二值化*/

Contours = findContours($\mathbf{I}_{bin}$)

/*使用OpenCV库中的轮廓检测算法 findContours 提取二值化图像 $\mathbf{I}_{bin}$ 中所有轮廓的坐标集合*/

$$\mathbf{I}_{mask} = \mathbf{I}_{bin}$$

/*生成二值化图像 $\mathbf{I}_{bin}$ 的副本 $\mathbf{I}_{mask}$ 便于进行掩码操作*/

for C_i in Contours:

$$x_1, y_1, x_2, y_2 = C_i$$

$$\mathbf{I}_{mask}[y_1:y_2, x_1:x_2] = 255$$

/*依据轮廓坐标集合 Contours 对 $\mathbf{I}_{mask}$ 中对应坐标区域执行掩码操作*/

$$\text{hist} = \text{binImghist}(\mathbf{I}_{mask}, \text{direction})$$

/*若 *direction* = 0 对 $\mathbf{I}_{mask}$ 沿 x 轴方向进行投影运算，统计每个单位像素值 y 轴方向的白色像素个数。反之先沿 y 轴投影再沿 x 轴统计像素*/

$$\text{addLine} = \text{ThresholdFiltering}(\text{hist})$$

/*通过阈值滤波算法 ThresholdFiltering 确定待补充的横向框线与竖向框线的坐标 (根据 *direction* 的值将阈值设置为 $0.1 * \text{图像长/宽}$) */

Return addLine /*返回添加的框线集合*/

为提升算法的通用性，在遍历轮廓对图像执行掩码操作的过程中，依据上述投影方向 (水平/竖直) 分别对整行或整列区域执行掩码操作，并将 ThresholdFiltering 中的阈值设置为 1。优化后轮廓掩码操作流程如算法2所示，优化前后结果对比如图2所示。

算法2 通用轮廓掩码算法伪代码

输入： 轮廓坐标 Contours，二值化图像副本 $\mathbf{I}_{mask}$

输出： 掩码后的图像

for C_i in Contours:

$$x_1, y_1, x_2, y_2 = C_i$$

if *direction* = 0

$$\mathbf{I}_{mask}[:, x_1:x_2] = 255$$

else

$$\mathbf{I}_{mask}[y_1:y_2, :] = 255$$

/*生成竖直框线，对图像进行竖向范围掩码，生成水平框线，对图像进行横向范围掩码*/

对无边框表格图片添加框线后的效果如图3所示。

2.3 图注意力网络架构

表格结构图定义 基于编码器输出的特征以及

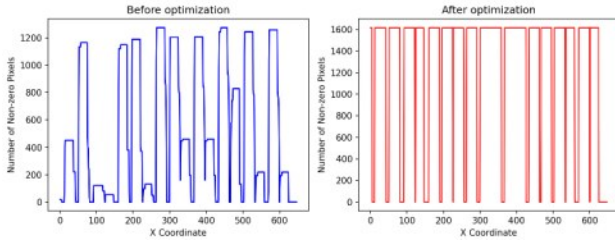


图2 优化前(左)后(右)投影对比结果

	range of error(°)	Mean squares	F-ratio	P	
		Screening	Tracking		
Side-bending	1-4.5	4.12	0.16	25.72	<0.001
Flexion-extension	1-4.5	1.2	0.29	4.21	0.012

a)

	range of error(°)		Mean squares		F-ratio	P
		Screening		Tracking		
Side-bending	1-4.5	4.12		0.16	25.72	<0.001
Flexion-extension	1-4.5	1.2		0.29	4.21	0.012

b)

图3 a)原始表格图片,b)中蓝色线条为根据框线生成算法添加的框线

框线对表格图片的区域划分构建异构图结构。

节点特征：每个节点的特征 $\mathbf{h}_i$ 直接继承自编码器的输出特征 $\text{Encoder}(\mathbf{F}_{enc})$ ， $\text{Encoder}()$ 为编码器处理过程，编码器结构如图 1 中编码器部分所示，由 N 个编码层组成，每个编码层中包含一个多头自注意力层与一个前馈神经网络组成，结构与 Transformer^[14] 中编码器几乎相同。节点特征 $\mathbf{h}_i \in \mathbf{R}^{B \times \max_len \times d_{model}}$ 。

$$\{\mathbf{h}_i\}^{\max_len} = \text{Encoder}(\mathbf{F}_{enc}) \quad (7)$$

节点分为文本节点（前 N_{text} 项）和结构节点（后 N_{line} 项），且需要满足节点总数小于 $\max_len$ ：

$$N_{test} + N_{line} \leq \max_len \quad (8)$$

边（节点关系）：基于空间拓扑的边连接规则构建节点之间的关系，首先利用水平和垂直框线将表格划分为互不重叠的矩形区域 $\{R_k\}_{k=1}^K$ ，对每个区域 R_k 记录其四个方向的框线 L_k 。

$$L_k = \{l_{top}, l_{bottom}, l_{left}, l_{right}\} \quad (9)$$

$l_{top}, l_{bottom}, l_{left}, l_{right}$ 分别表示距离区域 R_k 中心点上、下、左、右四个方向垂直距离最短的框线。

a) 构建区域 R_i 内所有文本节点关系

$$\forall t_p, t_q \in R_k, (t_p, t_q) \in \varepsilon_{TT} \quad (10)$$

b) 构建邻域文本节点关系

$$\begin{aligned} \forall R_m \in N_{dir}(R_k) \\ (t_p, t_q) \in \varepsilon_{TT}(t_p \in R_k, t_q \in R_m) \end{aligned} \quad (11)$$

其中 $N_{dir}(R_k)$ 表示 R_k 在指定方向（上下左右）相邻的最近区域。

c) 构建文本-结构节点关系

$$\forall t_p \in R_k, (t_p, l_j) \in \varepsilon_{TL} \text{ iff } l_j \in L_k \quad (12)$$

其中 ε_{TT} 表示文本节点之间的边， ε_{TL} 表示文本节点与结构节点之间的边， t_i 表示文本节点， l_i 表示结构节点。为了满足 GAT 的输入条件，根据节点关系生成大小为 $\max_len \times \max_len$ 的边邻接矩阵 Edge_index 。

GAT 网络采用两层级联的图注意力机制，在保持输入输出维度一致性的同时实现邻域信息融合。GAT 的输入为节点特征 $\{\mathbf{h}_i\}^{\max_len}$ 与边邻接矩阵 Edge_index 。

多头注意力层（如图 1 中 Multi-Head Graph Attention）中注意力头数设置为 2，每组独立学习节点间的关联权重。对于每个注意力头，每个节点的原始特征特征 $\mathbf{h}_i$ 通过一个共享的可学习线性投影矩阵 $\mathbf{W} \in \mathbf{R}^{d_i \times d_{model}}$ 投影到低维空间， $\mathbf{z}_i \in \mathbf{R}^{B \times \max_len \times d_k}$ 。

$$\mathbf{z}_i = \mathbf{W}\mathbf{h}_i \quad (13)$$

对于节点 i 和其邻居节点 j ，通过先拼接投影后的特征再与可学习注意力向量 $\mathbf{a}$ 做点积的形式得到未归一化的注意力分数 $\mathbf{e}_{ij}$ ，为了引入非线性并允许负值梯度回传，从而提升模型的表达能力和训练稳定性，对该分数应用 LeakyReLU 激活函数。

$$\mathbf{e}_{ij} = \text{LeakyRelu}(\mathbf{a}^T(\mathbf{z}_i \parallel \mathbf{z}_j)) \quad (14)$$

其中 LeakyReLU 定义如式 (15) 所示（本文中 $\alpha = 0.02$ ）：

$$\text{LeakyRelu} = \begin{cases} x & \text{if } x \geq 0 \\ \alpha x & \text{if } x < 0 \end{cases} \quad (15)$$

随后，为了使注意力系数在节点 i 的邻居集合 $N(i)$ 上形成有效的概率分布，对 $\mathbf{e}_{ij}$ 进行归一化，最终得到的注意力系数 α_{ij} ， α_{ij} 表示邻居节点 j 对中心节点 i 特征聚合的相对重要性，其值介于 0 和 1 之间，且对于节点 i 的所有邻居节点满足 $\sum_{j \in N(i)} \alpha_{ij} = 1$ 。

$$\alpha_{ij} = \text{Softmax}_k(\mathbf{e}_{ij}) \quad (16)$$

每个注意力头独立地对节点 i 的邻居特征进行加权聚合，最后进行拼接得到更新后的节点特征 $\mathbf{h}'_i \in \mathbf{R}^{B \times \max_len \times d_k}$ ， $\alpha_{ij}^k, \mathbf{z}_j^k$ 分别表示第 k 个注意力头中的注意力系数与投影向量。

$$\mathbf{h}_i' = \parallel_{k=1}^H (\sum_{j \in N(i)} \alpha_{ij}^k \mathbf{z}_j^k) \quad (17)$$

特征聚合层中，将更新后的节点特征 $\mathbf{h}_i'$ 采用高斯误差线性单元 GELU 激活函数引入非线性提升训练稳定性和性能（GELU 过程如式（18）所示， $\Phi(x)$ 为标准正态分布的累积分布函数），并通过全连接层映射回原始维度（ d_{model} ）。残差连接将聚合后的特征与原始节点特征相加，并应用层归一化稳定训练过程，最终输出与输入同维度的增强表征 $\mathbf{h}_i''$ ， $\mathbf{h}_i''$ 将作为解码器的输入。整个过程如式（20）所示。其中 $\mathbf{W}_0 \in R^{d_{model} \times d_k}$ 为线性投影矩阵，确保输出维度不变。

$$\text{GELU}(x) = x\Phi(x) \quad (18)$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt \quad (19)$$

$$\mathbf{h}_i'' = \text{LayerNorm}(\mathbf{h}_i + \mathbf{W}_0 \cdot \text{GELU}(\mathbf{h}_i')) \quad (20)$$

2.4 解码器

表格结构解码器（TSD，TableStructure Decoder）架构基于文献[25]提出的工作。TSD 是由 4 个解码器层组成的 Transformer 解码器，每个解码器包含一个带掩码的多头自注意力（Masked Multi-Head Self-Attention）、一个多头注意力（Multi-Head Attention）以及一个前馈神经网络。

在表格结构预测任务中，解码器的输出结果需与真实标注（GroundTruth）形成精准的比对与优化反馈，因此需先建立 GroundTruth 与解码器输出格式的统一标准——即对 GroundTruth 进行序列化处理。针对这一需求，本文提出了改进型结构化标签嵌入策略：在 Tableformer 标签嵌入方法的基础上，结合 PTN 数据集的结构特征，系统化筛选表格相关 HTML 标签（排除 ，<i>，<u> 等文本格式类标签，保留 <table>，<tr>，<td> 等布局相关基础标签）；对跨行 / 跨列单元格（spanning cells）采用拆分标记策略（拆分为 <td>，colspan=x/rowspan=x 及 >）；并引入 <SOS>、<EOS>、<PAD> 标签完善序列结构。上述处理最终形成如表 1 所示的 HTML 标签与序列号的对应关系，为解码器的预测任务提供标准化输入。

3 实验

3.1 概述

在不同的场景中使用的表格通常表现出不同的

表1 标签编码

0	<PAD>	12	</tbody>
1	<SOS>	13	colspan="2"
2	<thead>	14	colspan="7"
3	<tr>	15	colspan="3"
4	<td></td>	16	rowspan="2"
5	<td>	17	rowspan="3"
6	colspan="8"	18	colspan="9"
7	>	19	colspan="5"
8	</td>	20	colspan="4"
9	</tr>	21	colspan="6"
10	</thead>	22	colspan="10"
11	<tbody>	23	<EOS>

风格。为全面评估 TableGatFormer 模型的鲁棒性和泛化能力，本文在 PubTabNet（PTN）^[26] 的训练集上进行预训练，并在结构差异显著的两个公开基准数据集——FinTabNet（FTN）^[27] 和 TableBank（TB）^[28] 上评估了 TGF 的性能。为了评估本文的模型在预测表格结构识别方面的性能，本文使用了树编辑距离相似度（Tree-Edit-Distance-Based Similarity, TEDS）来比较预测值和实际值。

实验部分主要包括：1）性能对比实验：与当前最优（SOTA）方法进行定量对比；2）消融研究：分别验证多模态编码器（MME）和图注意力模块（GAT）对模型性能的贡献。

3.2 实验环境设置

实验设置 所有的实验均在 4 张 V100-32G 上完成，并基于 Pytorch 1.9.0+cu111 完成。

数据预处理 本文的模型输入包括表格图像、单元格内的文本信息以及对应的坐标信息。文本信息和文本坐标信息可以直接从 PubTabNet 提供的 json 文件中直接提取，框线坐标信息需要使用 2.2 中提及的框线生成算法，根据原始表格的表格结构类型生成对应的框线并得到它们的坐标，同时筛选文本与框线个数总和大于 max_len 的样本。在预处理阶段，本文筛选了标注文件中单元格区域面积为零的异常样本。为了使模型能够感知表格中的空白单元格并更好地理解邻域关系，本文将其视为空节点。这些节点的文本信息被设置为空嵌入，但保留其对应的几何坐标信息。预处理后的效果如图 4 所示。

模型结构 编码器由一个多模态编码层（详见 2.1 节）和 N=6 个标准的 Transformer 编码层堆叠而

	range of error(*)	Mean squares	F-ratio	P
		Screening	Tracking	
Side-bending	1-4.5	4.12	0.16	25.72
Flexion-extension	1-4.5	1.2	0.29	4.21
				0.001
				0.012

图4 绿色矩形框为给定的标注区域,红色矩形框为根据2.2中生成框线添加的空白标注区域

成。遵循[25]中的配置,本文将模型维度 d_{model} 设置为512,前馈神经网络(FFN)的内部维度设置为1024,多头自注意力机制的头数(heads)设置为4。编码器输出的序列特征(维度为 $\max_len \times d_{model}$,其中L为序列长度)被用作GAT中节点的初始特征。图的邻接关系由一个 $\max_len \times \max_len$ 的邻接矩阵Edge_index定义。解码器采用了与Transformer编码层相同的结构和参数配置。为了防止过拟合,本文在模型的所有线性层和注意力层后应用了Dropout,其丢弃率设置为0.5。

训练 作为一个序列生成任务,本文采用交叉熵损失函数(CELoss)来优化模型,并使用Adam优化器,在训练过程中,前15轮训练中本文将学习率设置为 $1e-3$,batch_size设置为24。然后,本文将学习率降低到 $1e-4$,batch_size减少到18,再进行15轮训练。

3.3 实验结果

为了验证本文所提方法的有效性,本文在PTN的训练集上进行了预训练,并在PTN的验证集、FTN验证集以及TB-latex验证集上进行了表格结构识别效果测试。实验结果如表2所示,与最先进的表格结构识别方法相比,本文的方法在简单表格的结构识别和总体检测效果上都实现了SOTA性能。同时本文的方法在PTN验证集上无框线表格结构的识别结果TEDS达到98.34。

3.4 消融实验结果分析

为了验证多模态编码器和图注意力网络这两个模块的有效性,设计了如表3所示的消融实验,并在PTN验证集上进行实验。

本文所设基准模型(Baseline)为基于Tableformer的模型结构设计的单任务(表格结构识别)模型(Tableformer为多任务联合优化模型);编码器采用ResNet18作为骨干网络(backbone),用于对输入的文本图像、框线图像进行编码,解码器则采用TableStructureDecoder。GatForme在Baseline的基础上,进行图结构建模(图的定义见2.3节),并引入GAT模块,其余网络配置保持与Baseline一

表2 模型在PTN(验证集)、FTN、TB三个数据集上的实验效果

Model	Dataset	TEDS-struc.(%)		
		Simple	Complex	All
GTE[27]	PTN	-	-	93.01
MTL-TabNet[17]	PTN	99.05	96.66	97.88
Davar-Lab[29]	PTN	97.88	94.78	96.36
Tableformer[25]	PTN	98.50	95.00	96.75
UniTabNet[22]	PTN	-	-	97.50
Our method	PTN	99.46	96.65	98.18
GTE[27]	FTN	-	-	87.14
GTE(FT)[27]	FTN	-	-	91.02
MTL-TabNet[17]	FTN	99.07	98.46	98.79
Tableformer[25]	FTN	97.50	96.00	96.80
WSTableNet[30]	FTN	99.06	98.33	98.72
Our method	FTN	99.34	97.45	98.40
Tableformer[25]	TB	-	-	89.6
TableVLM[31]	TB	-	-	90.2
Our method	TB	-	-	95.47
Our method(FT)	TB	-	-	97.26

致。MEFormerr在Baseline的基础上,将原编码器替换为多模态编码器,其余网络配置与Baseline保持一致。TableGatFormer为本文所提核心模型,其同时采用多模态编码器,并引入GAT模块,其余网络配置沿用Baseline的设计。为进一步验证本文提出的多模态编码器(MME)模块的有效性,我们在Tableformer中将骨干网络替换为多模态编码器,并开展了相应实验。尽管基准模型与Tableformer均采用ResNet-18作为骨干网络,但基准模型针对所有节点图像进行特征提取,而Tableformer则对整张图像进行整体特征提取;因此,受限于架构差异,无法在Tableformer中直接集成GAT模块。

表3 消融实验设置

Model	Model -Architecture
Baseline	ResNet18+Table Structure Decoder
GatFormer	ResNet18+GAT+Table Structure Decoder
MEFormer	MME+Table Structure Decoder
TableGatFormer	MME+GAT+TableStructure Decoder
Tableformer	ResNet-18 + Dual-Decoder
Tableformer+MME	MME + Dual-Decoder

实验结果如表4所示,GatFormer相对于基准模型的TEDS-struc指标提升了2.09%,这表明图注意力网络能够在模型感知全局特征的基础上,进一

步细化局部特征表示。MEFormer相对于基准模型的TEDS-struc指标提升了2.98%，这显示融合语义信息与位置特征的多模态表示有助于模型更精确地理解表格结构，在TableFormer上的小幅成效（指标提升1.3%）亦从侧面印证了这一观点。TableGatFormer的准确率提升幅度（7.03%）高于GatFormer与MEFormer的简单相加（2.09% + 2.98%），这揭示了多模态编码器与图注意力网络模块间的互补效应：多模态编码器使图节点突破单一模态限制，感知多元模态信息，从而提升模型的泛化能力和鲁棒性；图注意力网络在此基础上，使模型在捕捉全局特征的同时，亦兼顾对临近节点特征的精细感知。

表4 消融实验结果

Model	GAT	MME	TEDS-struc.(%)
Baseline	□	□	91.75
GatFormer	□	□	93.84
MEFormer	□	□	94.73
TableGatFormer	□	□	98.18
Tableformer	□	□	96.26
Tableformer+MME	□	□	97.59

4 结束语

在本文中，本文提出一种基于多模态、图注意力机制和编解码架构的模型——TableGatformer，旨在预测图像中表格的结构信息。该方法能够实现表格结构的重建，并可结合特定后处理算法完成表格解析任务。同时，该方法具备在真实场景下处理各类表格所需的良好多功能性。此外，本文方法在性能上显著优于当前所有最先进方法（SOTA methods）。最后，本文通过消融实验验证了所提模块的有效性。

尽管本研究已取得较为理想的实验效果，但仍存在进一步优化与拓展的空间。具体而言，未来工作将致力于在更多领域的公开表格数据集上开展验证实验，以进一步验证并提升本方法的普适性与鲁棒性。其次，本文将尝试将表格解析任务从单目标建模拓展为多目标建模，并引入新型损失函数开展多目标联合优化，以实现表格解析任务的端到端完成，替代当前依赖后处理步骤辅助实现表格解析的方式。再者，尽管当前大语言模型（Large Language Models, LLM）在表格结构识别任务中仍面

临诸多挑战，但智能推理能力始终是其核心优势。未来研究将探索引入LLM的智能体（Agent）推理机制，借助其优异的上下文理解能力与逻辑推理能力，对本文所提基于规则的框线生成策略及图结构构建方法进行优化，以期高效处理更为复杂的非常规表格结构，进而推动表格结构理解与表格解析技术向更高水平的智能化发展。

参考文献：

- [1] HOU Q, WANG J. TABLET: Table Structure Recognition Using Encoder-only Transformers[C]//International Conference on Document Analysis and Recognition. Cham: Springer Nature Switzerland, 2025: 253-278.
- [2] ZHONG X, SHAFIEIBAVANI E, JIMENO YEPES A. Image-based table recognition: data, model, and evaluation[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 564-580.
- [3] YU C, LI W, LI W, et al. A survey for table recognition based on deep learning[J]. Neurocomputing, 2024, 600 : 128154.
- [4] HUANG Y, LU N, CHEN D, et al. Improving table structure recognition with visual-alignment sequential coordinate modeling [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 11134-11143.
- [5] LIN W, SUN Z, MA C, et al. Tsrformer: Table structure recognition with transformers[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 6473-6482.
- [6] KASEM M, ABDALLAH A, BERENDEYEV A, et al. Deep learning for table detection and structure recognition: A survey[J]. ACM Computing Surveys, 2024, 56(12): 1-41.
- [7] SINGH A, BACCHUWAR K, BHASIN A. A survey of OCR applications[J]. International Journal of Machine Learning and Computing, 2012, 2(3): 314.
- [8] QASIM S R, MAHMOOD H, SHAFAIT F. Rethinking table recognition using graph neural networks[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019: 142-147.
- [9] LI X H, YIN F, DAI H S, et al. Table structure recognition and form parsing by end-to-end object detection and relation parsing [J]. Pattern Recognition, 2022, 132: 108946.
- [10] LIU H, LI X, LIU B, et al. Neural collaborative graph machines for table structure recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4533-4542.
- [11] 孙俊, 苟刚. 融合图卷积网络的表格图像结构识别模型[J]. 计算机工程与设计, 2024, 45(10): 3066-3073.
SUN J, GOU G. Table image structure recognition model fusing graph convolutional network[J]. Computer Engineering and Design, 2024, 45(10): 3066-3073.

- [12] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph Attention Networks[C]//International Conference on Learning Representations. 2018.
- [13] UDAYAKUMAR P, SUBHASHINI R. Connectome Graph Attention Network (CGAT): A Multimodal approach for enhanced classification of Alzheimer's disease[J]. Results in Engineering, 2025: 105200.
- [14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [15] WANG Z, DONG H, JIA R, et al. Tuta: Tree-based transformers for generally structured table pre-training[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 1780-1790.
- [16] Lu N, Yu W, Qi X, et al. Master: Multi-aspect non-local network for scene text recognition[J]. Pattern Recognition, 2021, 117: 107980.
- [17] LY N T, TAKASU A. An end-to-end multi-task learning model for image-based table recognition[C]//Document Analysis Systems. Cham: Springer Nature Switzerland, 2024: 150-163.
- [18] YANG G, ZHONG D, XIONG Y, et al. LRATNet: Local-Relationship-Aware Transformer Network for Table Structure Recognition[C]//International Conference on Multimedia Modeling. Cham: Springer Nature Switzerland, 2024: 507-520.
- [19] XIAO B, SIMSEK M, KANTARCI B, et al. Rethinking detection based table structure recognition for visually rich document images[J]. Expert Systems with Applications, 2025, 269: 126461.
- [20] HERZIG J, NOWAK P K, MÜLLER T, et al. TaPas: Weakly supervised table parsing via pre-training[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. 2020: 4320-4333.
- [21] YIN P, NEUBIG G, YIH W, et al. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8413-8426.
- [22] ZHANG Z, LIU S, HU P, et al. UniTabNet: Bridging Vision and Language Models for Enhanced Table Structure Recognition[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. Miami: Association for Computational Linguistics, 2024: 6131-6143.
- [23] PENG S Y, CHAKRAVARTHY A, LEE S, et al. UniTable: Towards a Unified Framework for Table Recognition via Self-Supervised Pretraining[C]//NeurIPS 2024 Third Table Representation Learning Workshop.
- [24] BADARO G, SAEED M, PAPOTTI P. Transformers for tabular data representation: A survey of models and applications[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 227-249.
- [25] NASSAR A, LIVATHINOS N, LYSAK M, et al. Tableformer: Table structure understanding with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4604-4613.
- [26] ZHONG X, SHAFIEIBAVANI E, JIMENO YEPES A. Image-based table recognition: data, model, and evaluation[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 564-580.
- [27] ZHENG X, WANG W, LIU W, et al. Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context[C]//2021 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2020: 697-706.
- [28] LI M, CUI L, HUANG S, et al. Tablebank: Table benchmark for image-based table detection and recognition[C]//Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020: 1918-1925.
- [29] JIMENO YEPES A, ZHONG P, BURDICK D. ICDAR 2021 competition on scientific literature parsing[C]//International conference on document analysis and recognition. Cham: Springer International Publishing, 2021: 605-617.
- [30] LY N T, TAKASU A, NGUYEN P, et al. Rethinking image-based table recognition using weakly supervised methods[C]//International Conference on Pattern Recognition Applications and Methods. Springer, 2023, 1774: 676-697.
- [31] CHEN L, HUANG C, ZHENG X, et al. TableVLM: multi-modal pre-training for table structure recognition[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 2437-2449.