基于多模态与图注意力网络的表格结构识别算法

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 430070；2. 武汉理工大学信息工程学院，湖北 武汉 430070）

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实现非结构化数据转化为结构化数据的关键。表格形状、表头配置、跨行跨列单元格，以及分隔线样式复杂多样，尤其是缺乏边框的表格，对其识别既具有重要意义又极具挑战。针对无边框表格，提出了一种基于多模态与图注意力网络的表格结构识别模型——TableGatFormer。首先，引入多模态编码器，用于融合表格的视觉、文本、布局和边框特征。其次，提出基于文本位置分布的框线生成算法，以完善表格边框。接着，在编码-解码器架构中引入图注意力网络，使得节点在感知全局特征的同时，也能感知相邻节点的特征。最后，通过对比实验与消融实验验证了 TableGatFormer 模型在表格结构识别任务中的有效性。在 PubTabNet 公开数据集上的测试结果显示，TableGatFormer 模型在简单表格识别任务上的树编辑距离相似度（tree edit distance based similarity，TEDS）达 99.46%，在复杂表格上达到 96.65%，总体结果达到 98.18%。

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
中图分类号：TP391.4
文献标志码：A
doi：10.11959/j.issn.2096−3750.WLW25169

Table structure recognition via multimodality and graph attention networks

Xu Chongshan[1], Wu Lyu[2], Li Ning[1], Xu Ning[2], Pan Hao[1]
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 highly condensed and structured forms of information presentation, are widely found in various documents and images. Accurate understanding of table structure is regarded as the key to transforming unstructured data into structured data. Table shapes, header configurations, spanning cells, and separator line styles are complex and diverse, and table recognition is considered both highly significant and extremely challenging, especially for tables lacking borders. To address borderless tables, a table structure recognition model based on multimodal and graph attention network (GAN), named TableGatFormer, was proposed. First, a multimodal encoder was introduced to fuse visual, textual, layout, and border features of tables. Second, a border line generation algorithm based on text position distribution was proposed to refine table borders. Third, a GAN was introduced into the encoder-decoder architecture, enabling nodes to perceive both global features and features of adjacent nodes. Finally, the effectiveness of the TableGatFormer model in the task of borderless table structure recognition was verified through comparative experiments and ablation studies. The test results on the public PubTabNet dataset showed that the TableGatFormer model achieved a tree edit distance based similarity (TEDS) of 99.46% on simple table recognition tasks, 96.65% on complex tables, and 98.18% overall.

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

收稿日期：2025−11−07；**修回日期：**2026−09−11
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 Item: The National Natural Science Foundation of China (No. 92373102)

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。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，而表格结构识别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表格解析带来了诸多挑战。表格形状多样，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]。特别是无边框表格，由于其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。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，在斜线表头中，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，目前的识别方法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。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；文本对齐策略容易受到文本排版、字体大小等因素的影响，出现行列划分偏差。在 2023 国际计算机视觉与模式识别大会（IEEE/CVF Conference on Computer Vision and Pattern Recognition，CVPR）上提出的 VAST 模型[6]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，但在复杂场景中，如表格内容存在大量干扰信息、文本布局不规则等，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]。

针对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，本文提出了一种基于多模态与图注意力网络的表格结构识别模型——TableGatFormer，以实现更准确、更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，将编码–解码器架构中广泛使用的卷积神经网络（convolutional neural network，CNN）编码器更换为多模态编码器。多模态编码器融合了文本与框线区域图片信息、文本信息，以及对应的坐标信息，能够显著提高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，以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）在编码解码器架构中引入图注意力网络（graph attention network，GAN）模块。该模块能够使节点更好地感知邻近节点的特征，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 TableGatFormer 模型是一种基于编码—邻近感知—解码架构的多模态模型，它融合了表格图片、文本信息及布局信息，能够实现对表格结构的全面预测。通过对比实验证明了本文的方法在 PubTabNet 数据集上能达到最佳性能，消融实验也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，将文本块或检测到的单元格抽象为图节点，并将其对应的视觉特征作为初始节点表示。Qasim 等[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[9]设计了一种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[10]提出了一种配备了堆叠协作块的神经协作图机器（neural collaborative graph machine，NCGM）模型，该模型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的交互。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。孙俊等[11]提出了一种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，并引入矩阵分解头代替注意力机制重整全局特征。该模型利用图卷积网络有效感知单元格间的连接关系，从而辅助输出位置信息和逻辑邻接关系，其在多个数据集上的F1分数平均提升10.6%。此外，多模态图注意力网络，如Connectome GAT[12]，在结构化数据诊断中利用图注意力网络融合多模态特征，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　基于编码–解码器架构的表格结构识别方法

基于编码器–解码器（encoder-decoder）的表格结构识别方法利用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：（1）采用编码–解码器架构对表格元素的跨模态特征进行联合建模；（2）利用自回归解码生成结构化标记序列，如HTML标签、LaTeX代码。自Vaswani等[14]提出Transformer基础架构以来，Wang等[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，为解决跨行列合并这类复杂结构提供了新范式。阿里团队[16]提出的Master采用CNN-Transformer混合编码器，借助卷积神经网络提取单元格区域的高粒度视觉特征，并以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，该机制在PubTabNet、FinTabNet等基准数据集上实现了当时的SOTA性能。

当前研究进一步改进了编码–解码器框架。例如，Yang等[18]提出的局部关系感知Transformer网络LRATNet（local-relationship-aware Transformer network），利用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的预测性能。Xiao等[19]通过优化富文本文档的处理，提升了对复杂表格结构识别的准确性。

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[20]提出的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[21]提出了TaBERT，其在2 600万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[22]提出了UniTabNet，这是一种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—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[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，尽管现有方法各具优势，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，这一问题仍需重点关注。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，在计算过程中，全局注意力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TabGatFormer模型，用于解决复杂的表格结构识别任务。TabGatFormer模型巧妙地结合了GAT的局部感知能力与Transformer的全局建模优势。此外，为进一步提升模型的鲁棒性与泛化能力，在TabGatFormer模型中创新性地引入了多模态融合编码模块。

2　基于多模态与邻近感知的表格结构识别

TabGatFormer模型主要由多模态编码器、图注意力网络和解码器组成，如图1所示。TabGatFormer模型采用编码–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5个部分构成，分别为：视觉特征提取器、文本信息特征提取器、几何位置编码器、多模态特征融合器和Transformer编码层。

2.1.1　视觉特征提取器

视觉特征提取器本质上是一个ResNet-18网络。其第一层使用7×7卷积核（步长为2，填充为3）将输入通道从3扩展至64，经ReLU激活和3×3最大池化（步长为2，填充为1）后，输出56×56×64的特征图。随后依次经过四组残差块：第一组残差块（2个）保持通道数64不变，输出56×56×64特征图；第二组残差块（2个）通过步长为2的卷积将通道数从64提升至128，输出28×28×128特征图；第三组残差块（2个）将通道数从128提升至256，输出14×14×256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7×7×512特征图。经全局平均池化将7×7×512的特征图压缩为512维向量，再通过全连接层将512维特征映射至目标维度D_h，得到视觉特征$F_v \in \boldsymbol{R}^{B\cdot l_{max}\cdot D_h}$。其中，$B$为批大小（batch_size），$l_{max}$为序列最大长度（后续公式中出现的$B$、$l_{max}$均作此解释）。

2.1.2　文本信息特征提取器

本文构建了层次化文本表征管道。首先，对OCR识别文本T_{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$\boldsymbol{U}$。

$$\boldsymbol{U} = \mathrm{Tokenize}(T_{OCR}) \tag{1}$$

随后，经线性层$\mathrm{Linear_t}$进行特征空间对齐，并通过非线性激活函数ReLU与层归一化（layer normalization，LN）得到最终文本特征$F_t \in R^{B\cdot l_{max}\cdot D_h}$，文本特征$F_t$保留细粒度语义信息。

$$F_t = \mathrm{LN}(\mathrm{ReLU}(\mathrm{Linear_t}(\boldsymbol{U}))) \tag{2}$$

层归一化过程如式（3）所示。

$$\mathrm{LN}(x) = \gamma\cdot\frac{x-\mu}{\sqrt{\sigma^2+\varepsilon}} + \boldsymbol{\beta} \tag{3}$$

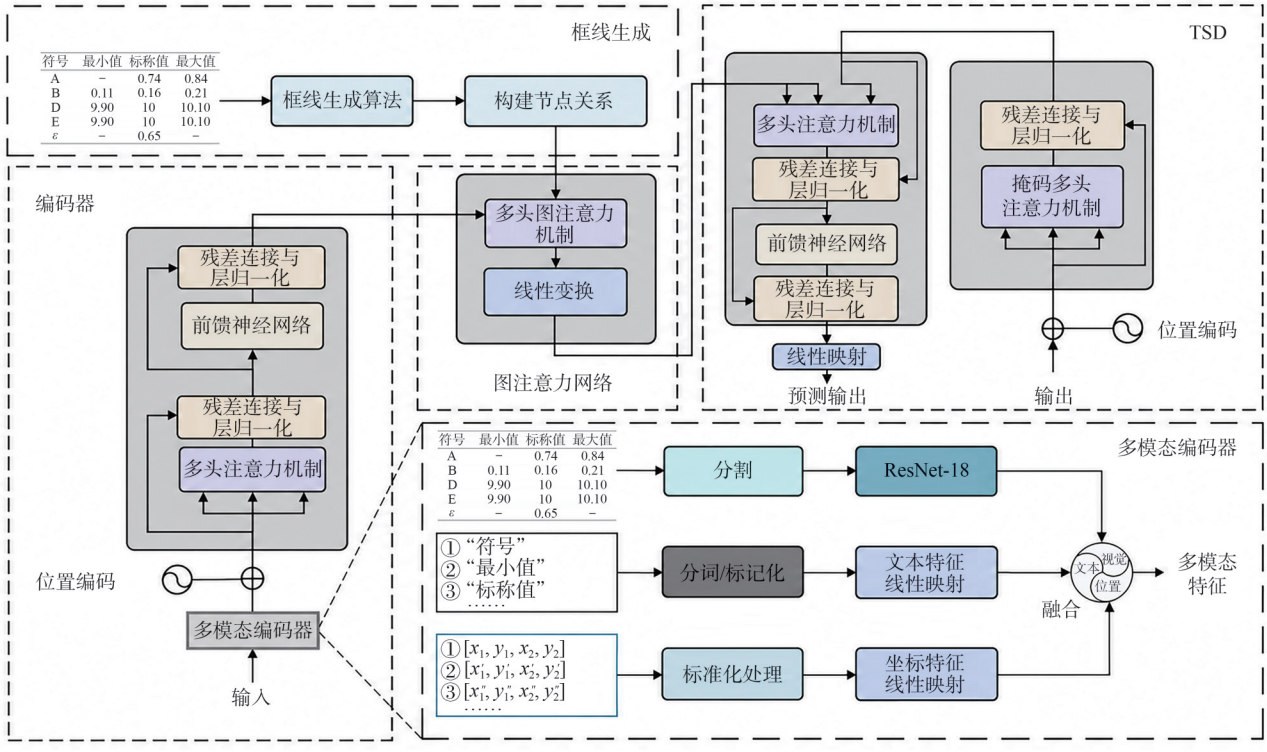


图1　TableGatFormer模型结构

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，设置为1×10^{-5}；$\boldsymbol{\gamma}$和$\boldsymbol{\beta}$为可学习参数向量

非线性激活函数ReLU过程如式（4）所示。

$$\mathrm{ReLU}(x)=\max(0,x) \tag{4}$$

2.1.3　几何位置编码器

本文设计了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。具体地，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$\left[\frac{x_1}{W},\frac{y_1}{H},\frac{x_2}{W},\frac{y_2}{H}\right]$，后文用Coord表示。其中，$W$、$H$分别为图像的宽和高。随后，通过线性层$\mathrm{Linear_C}$进行特征空间对齐，再经ReLU非线性激活函数与层归一化函数得到几何位置特征$F_{\mathrm{P}}\in\boldsymbol{R}^{B\cdot l_{\max}\cdot D_{\mathrm{h}}}$。

$$F_{\mathrm{P}}=\mathrm{LN}\left(\mathrm{ReLU}\left(\mathrm{Linear_C}(\mathrm{Coord})\right)\right) \tag{5}$$

该设计采用参数共享机制，使文本与框线在统一几何空间中对齐，从而增强拓扑关系的感知能力。

2.1.4　多模态特征融合

首先，将视觉特征F_{v}、文本特征F_{t}和几何位置特征F_{P}沿特征维度拼接，构建联合跨模态表征；随后，经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F_{enc}。该过程可形式化表示为式（6）。

$$F_{\mathrm{enc}}=\boldsymbol{W}_f\cdot\left[F_{\mathrm{v}}\|F_{\mathrm{t}}\|F_{\mathrm{P}}\right]+b_f \tag{6}$$

其中，$\boldsymbol{W}_f$为投影矩阵；b_f为偏移量；$F_{\mathrm{enc}}\in\boldsymbol{R}^{B\cdot l_{\max}\cdot D_{\mathrm{m}}}$，$D_{\mathrm{m}}$为输入编码器时模型的特征维度。

2.2　框线生成算法

表格框线通常与表格结构一起出现，是表格中用于分隔和界定单元格的线条。为使单元格节点能够更有效地感知邻近区域的几何特征，本文将表格框线的视觉特征和位置特征引入TableGatFormer模型的编码器中进行多模态融合。然而，从PubTabNet数据集提供的表格图片中提取框线坐标时，发现绝大多数表格是以类似三线表的形式存在的，图片中仅包含少量线条，缺乏完整的框线结构。若仅提取这些零星分布的原始线条，显然不能达到本文设计该模块的预期目标。为此，本文提出了一种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

输入：图像数组I、框线生成方向d（d=0表示竖直方向，d=1表示水平方向）

输出：生成的框线集合$\boldsymbol{A}$

I_{bin}= Binarization(I)

/*对输入图像I进行灰度化、二值化*/

$\boldsymbol{C}$=findContours(I_{bin})

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

$I_{\mathrm{mask}}=I_{\mathrm{bin}}$

/*生成二值化图像I_{bin}的副本I_{mask}便于进行掩码操作*/

for C_i in $\boldsymbol{C}$:

$x_1,y_1,x_2,y_2=C_i$

$I_{\mathrm{mask}}[y_1{:}y_2{:},x_1{:}x_2]=255$

/*依据轮廓坐标集合$\boldsymbol{C}$对I_{mask}中对应坐标区域执行掩码操作*/

$h=\mathrm{binImghist}(I_{\mathrm{mask}},d)$

/* 若$d=0$，沿X轴方向对I_{mask}进行投影运算，统计每个单位像素值在Y轴方向上的白色像素个数。反之，先沿Y轴投影，再沿X轴统计像素*/

$\boldsymbol{A}=\mathrm{ThresholdFiltering}(h)$

/* 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d的取值将阈值设置为0.1×图像长/宽）*/

Return $\boldsymbol{A}$ /*返回添加的框线集合$\boldsymbol{A}$*/

为提升框线生成算法的通用性，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。该过程可形式化为通用轮廓掩码算法，如算法2所示。优化前后投影对比如图2所示。

算法2　通用轮廓掩码算法

输入：轮廓的坐标集合$\boldsymbol{C}$、二值化图像副本I_{mask}

输出：掩码后的图像

for C_i in $\boldsymbol{C}$

$x_1,y_1,x_2,y_2=C_i$

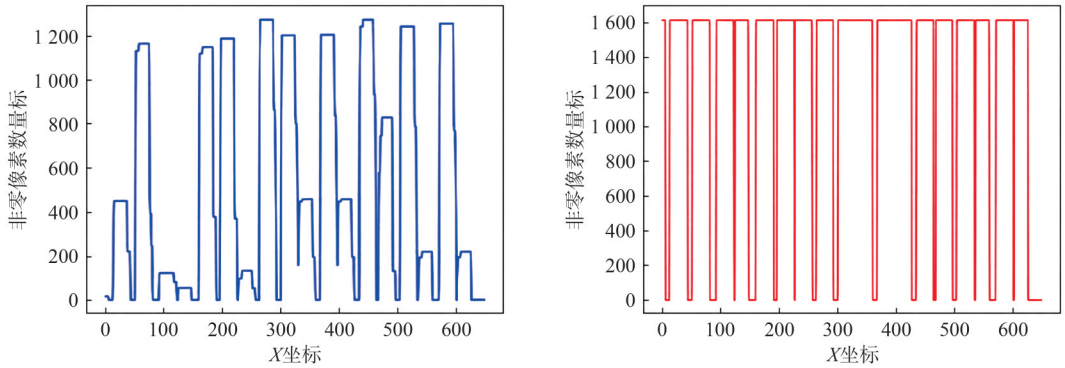
if $d==0$

$I_{\mathrm{mask}}[:,x_1{:}x_2]=255$

else

$I_{\mathrm{mask}}[y_1{:}y_2,:]=255$

/*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*/



（a）优化前　（b）优化后

图2　优化前后投影对比

采用框线生成算法对无边框表格图片添加框线，原始表格图片与算法生成图片对比如图3所示。

	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）原始表格图片

	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）经框线生成算法处理后表格图片

图3　原始表格图片与算法生成图片对比

2.3　图注意力网络架构

2.3.1　表格结构图定义

基于编码器输出的多模态特征及框线对表格图片的区域划分，本文构建了一个异构图结构，用于表征表格的拓扑布局与语义关联。

（1）节点特征。每个节点的特征h_i直接继承自编码器的输出特征Encoder(F_{enc})，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，它由N个编码层组成，每个编码层中包含一个多头自注意力（multi-head attention）机制与一个前馈神经网络（feedforward neural network，FNN），结构与Transformer[14]中编码器几乎相同。节点特征$h_i \in \boldsymbol{R}^{B\cdot l_{\max}\cdot D_m}$。

$$\{h_i\}^{l_{\max}} = \text{Encoder}(F_{\text{enc}}) \tag{7}$$

节点包括文本节点（前N_{t}项）和结构节点（后N_{l}项），且需要满足节点总数小于$l_{\max}$：

$$N_{\text{t}} + N_{\text{l}} \leqslant l_{\max} \tag{8}$$

（2）边（节点关系）。本文基于空间拓扑的边连接规则构建节点之间的关系。首先，利用水平框线和竖直框线将表格划分为互不重叠的矩形区域$\{G_k\}_{k=1}^{K}$，并记录每个区域G_k在4个方向的框线L_k。

$$L_k = \{L_{\text{top}}, L_{\text{bottom}}, L_{\text{left}}, L_{\text{right}}\} \tag{9}$$

其中，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4个方向垂直距离最短的框线。

构建区域G_k内所有文本节点关系：

$$\forall t_p, t_q \in G_k, (t_p, t_q) \in \varepsilon_{\text{TT}} \tag{10}$$

构建邻域文本节点关系：

$$\forall G_m \in Z_{\text{dir}}(G_k)$$
$$(t_p, t_q) \in \varepsilon_{TT}(t_p \in R_k, t_q \in G_m) \tag{11}$$

其中，$Z_{\text{dir}}(G_k)$表示G_k在指定方向（上、下、左、右）相邻的最近区域。

构建文本—结构节点关系：

$$\forall t_p \in G_k, (t_p, l_j) \in \varepsilon_{\text{TL}} \text{ if}, l_j \in L_k \tag{12}$$

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。

为满足GAT的输入条件，根据节点关系生成大小为$l_{\max}\times l_{\max}$的边邻接矩阵$\boldsymbol{E}$。

2.3.2　GAT

GAT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$\{h_i\}^{l_{\max}}$与边邻接矩阵$\boldsymbol{E}$。

多头注意力层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h_i通过一个共享的可学习线性

投影矩阵 $\boldsymbol{W} \in \boldsymbol{R}^{D_k \cdot D_m}$ 投影到低维空间，$\boldsymbol{z}_i \in \boldsymbol{R}^{B \cdot l_{\max} \cdot D_k}$。

$$\boldsymbol{z}_i = \boldsymbol{W} h_i \tag{13}$$

对于节点 i 和其邻居节点 j，可先拼接投影后的特征，再与可学习注意力向量 $\boldsymbol{a}$ 做点积的方式得到未归一化的注意力分数 $\boldsymbol{e}_{ij}$。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。

$$\boldsymbol{e}_{ij} = \text{LeakyReLU}\left(\boldsymbol{a}^{\mathrm{T}}\left(\boldsymbol{z}_i \| \boldsymbol{z}_j\right)\right) \tag{14}$$

其中，LeakyReLU的数学表达式如下：

$$\text{LeakyRelu} = \begin{cases} x, & x \geqslant 0 \\ \boldsymbol{\alpha} x, & x < 0 \end{cases} \tag{15}$$

其中，$\boldsymbol{\alpha}$在本文中取0.02。

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

$$\delta_{ij} = \text{Softmax}_k\left(\boldsymbol{e}_{ij}\right) \tag{16}$$

每个注意力头独立地对节点 i 的邻居特征进行加权聚合，拼接后得到更新的节点特征 $h_i' \in \boldsymbol{R}^{B \cdot l_{\max} \cdot D_k}$，$\delta_{ij}^k$、$\boldsymbol{z}_j^k$ 分别表示第 k 个注意力头中的注意力系数与投影向量。

$$h_i' = \|_{k=1}^{H}\left(\sum_{j \in \boldsymbol{Z}(i)} \delta_{ij}^k \boldsymbol{z}_j^k\right) \tag{17}$$

在特征聚合层中，将更新后的节点特征 h_i' 采用高斯误差线性单元GeLU激活函数引入非线性，以提升训练稳定性和性能，GeLU过程如式（18）所示。其中，$\boldsymbol{\Phi}(\cdot)$ 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_m。

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 h_i''，并将 h_i'' 作为解码器的输入，整个过程如式（20）所示。其中，$\boldsymbol{W}_0 \in \boldsymbol{R}^{D_m \cdot D_k}$ 为线性投影矩阵，确保输出维度不变。

$$\text{GeLU}(x) = x\boldsymbol{\Phi}(x) \tag{18}$$

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

$$h_i'' = \text{LN}\left(h_i + \boldsymbol{W}_0 \cdot \text{GeLU}\left(h_i'\right)\right) \tag{20}$$

2.4　解码器

表格结构解码器（table structure decoder，TSD）基于文献[25]提出的方式开展工作。TSD是由4个解码器层组成的Transformer解码器，每个解码器包含1个带掩码的多头自注意力（masked multi-head self-attention）机制、1个多头注意力机制，以及1个FNN。

在表格结构预测任务中，解码器的输出结果需要与真实标注（ground truth）形成精准的比对与优化反馈。因此，需先建立真实标注与解码器输出格式的统一标准，即对真实标注进行序列化处理。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PubTabNet数据集的结构特征，系统化筛选表格与相关的HTML标签（排除<b>，<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HTML标签与序列号的对应关系，见表1，为解码器的预测任务提供标准化输入。

表1　　HTML标签与序列号的对应关系

序列号	HTML标签	序列号	HTML标签
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>

3　实验

3.1　概述

在不同场景中使用的表格通常表现出不同的风格。为全面评估TableGatFormer模型的鲁棒性和泛化能力，本文在PubTabNet数据集[2]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[26]和TableBank[27]上评估TableGatFormer的性能。为评估TableGatFormer模型在表格结构识别方面的性能，使用树编辑距离相似度（tree-edit-

distance-based similarity，TEDS）比较预测值和实际值。

实验部分主要包括：（1）与当前性能最优（SOTA）方法进行定量对比的性能对比实验；（2）分别验证多模态编码器（MME）和图注意力网络（GAN）模块对模型性能贡献的消融实验。

3.2　实验环境设置

（1）实验设置

所有的实验均基于Pytorch 1.9.0+cu111，在4张V100-32G显卡上完成。

（2）数据预处理

TableGatFormer模型的输入包括表格图像、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数据提供的JSON文件中提取；框线坐标信息需要使用第2.2小节提到的框线生成算法，根据原始表格的结构类型生成对应的框线并得到它们的坐标，同时剔除文本与框线个数总和大于l_{max}的样本。

在预处理阶段，筛除了标注文件中单元格区域面积为零的异常样本。为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。预处理后的表格样式如图4所示，其中，实线矩形框为给定的标注区域，虚线矩形框为根据框线生成算法添加的空白标注区域。

（3）模型结构

编码器由1个多模态编码层和6个标准的Transformer编码层堆叠而成。基于文献[25]中的配置，本文将模型维度D_m设置为512，FFN的内部维度设置为1 024，多头自注意力机制的头数设置为4。编码器输出的序列特征（维度为$l_{max}\times D_m$）作为GAN中节点的初始特征。图的邻接关系由一个$l_{max}\times l_{max}$的邻接矩阵$\boldsymbol{E}$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TableGatFormer模型的所有线性层和注意力层后应用了Dropout机制，其丢弃率设置为0.5。

（4）模型训练

作为一个序列生成任务，本文采用交叉熵损失函数（CELoss）在PubTabNet数据集的训练集上对TableGatFormer模型进行预训练，优化器选用Adam。在训练过程中，前15轮训练将学习率设置为1×10^{-3}，B设置为24；后15轮训练将学习率降至1×10^{-4}，B减少到18，共开展30轮训练。

3.3　性能对比实验结果分析

为全面评估本文所提出TableGatFormer模型的表格结构识别性能，选取多种具有代表性的表格结构识别方法进行对比。除前文已介绍的方法外，本文进一步选取了以下具有代表性的模型进行比较：GTE[2]通过融合视觉上下文信息联合完成表格检测与单元格结构识别；WSTTableNet[28]采用弱监督学习策略进行表格结构识别；TableVLM[30]则通过多模态预训练融合表格图像与文本语义信息，以增强模型对复杂表格结构的表征能力。本文模型与上述模型在PubTabNet、FinTabNet和TableBank数据集上的TEDS指标对比结果见表2。从对比结果可以看出，不同模型在简单表格和复杂表格上的性能差异较为明显，说明复杂结构仍是表格结构识别中的主要难点。相比现有模型，TableGatFormer模型在多个数据集上取得了较为稳定的结果，尤其在全部表格上的TEDS指标表现出较好的竞争力，验证了本文所提出结构建模方法的有效性。

3.4　消融实验结果分析

为验证多模态编码器和图注意力网络这两个模块的有效性，在PubTabNet数据集的验证集上展开消融实验。

本文所设基准模型（Baseline）为基于Tableformer（其为多任务联合优化模型）的模型结构设计的单任务（表格结构识别）模型；编码器采用ResNet-18作为骨干网络，用于对输入的文本图像、框线图像进行编码，解码器则采用TSD。GatFormer在基准模型的基础上，进行图结构建模（参考第2.3小节），并引入GAN模块，其余网络配置与基准模

	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

图4　预处理后的表格样式

表2　4种模型在3个数据集上的实验效果

模型	数据集	TEDS		
		简单表格	复杂表格	全部表格
GTE	PubTabNet	—	—	93.01%
MTL-TabNet	PubTabNet	99.05%	**96.66%**	97.88%
Davar-Lab	PubTabNet	97.88%	94.78%	96.36%
Tableformer	PubTabNet	98.50%	95.00%	96.75%
UniTabNet	PubTabNet	—	—	97.50%
TableGatFormer	PubTabNet	**99.46%**	96.65%	**98.18%**
GTE	FinTabNet	—	—	87.14%
GTE（FT）	FinTabNet	—	—	91.02%
MTL-TabNet	FinTabNet	99.07%	**98.46%**	**98.79%**
Tableformer	FinTabNet	97.50%	96.00%	96.80%
WSTTableNet	FinTabNet	99.06%	98.33%	98.72%
TableGatFormer	FinTabNet	**99.34%**	97.45%	98.40%
Tableformer	TableBank	—	—	89.6%
TableVLM	TableBank	—	—	90.2%
TableGatFormer	TableBank	—	—	95.47%
Tablegat-Former	TableBank	—	—	**97.26%**

型一致。MEFormer在基准模型的基础上，将原编码器替换为多模态编码器，其余网络配置与基准模型保持一致。TableGatFormer为本文所提模型，其同时采用多模态编码器，并引入GAN模块，其余网络配置沿用基准模型的设计。为进一步验证本文提出的MME模块的有效性，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GAN模块。消融实验的模型设置见表3。

表3　消融实验的模型设置

模型	模型结构
Baseline	ResNet-18+TSD
GatFormer	ResNet-18+GAT+TSD
MEFormer	MME+TSD
TableGatFormer	MME+GAT+TSD
Tableformer	ResNet-18+编码-解码器
Tableformer+MME	MME+编码-解码器

消融实验结果见表4。表4数据表明，对比基准模型，GatFormer模型的TEDS指标提升了2.09%，表明GAN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模型的TEDS提升了2.98%，说明融合语义信息与位置特征的多模态编码器有助于模型更精确地理解表格结构。TableFormer模型TEDS指标的小幅提升（增长1.3%）亦从侧面印证了这一观点。TableGatFormer模型的TEDS指标提升幅度（7.03%）高于GatFormer模型与MEFormer模型的简单相加（2.09%+2.98%），揭示了多模态编码器与GAN模块间的互补效应：多模态编码器帮助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；GAN在此基础上，使模型在捕捉全局特征的同时，亦兼顾对邻近节点特征的精细感知。

表4　消融实验结果

模型	GAN	MME	TEDS
Baseline	✗	✗	91.75%
GatFormer	✓	✗	93.84%
MEFormer	✗	✓	94.73%
TableGatFormer	✓	✓	98.18%
Tableformer	✗	✗	96.26%
Tableformer+MME	✗	✓	97.59%

4　结束语

为高效识别表格的结构信息，本文提出了一种基于多模态与图注意力网络的表格结构识别模型——TableGatFormer。本文提出的TableGatFormer模型能够实现表格结构的有效重建，并可结合相应的后处理算法完成表格内容解析，为复杂场景下的表格理解与解析提供了一种可行方案。实验结果表明，该模型在多种表格场景下具有较好的适应能力，并在相关性能指标上取得了具有竞争力的结果。同时，消融实验进一步验证了各关键模块对模型性能提升的有效性。未来可围绕复杂表格结构、真实场景中的图像质量变化，以及模型轻量化与部署效率等开展研究，以提升模型在实际应用场景中的鲁棒性和泛化能力。

尽管本文所提TableGatFormer模型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，未来研究将在以下方面进行拓展：一是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TableGatFormer模型的普适性与鲁棒性。二是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，以实现端到端的表格任务解析，进而替代当前依赖后处理步骤辅助实现表格解析的方式。三是探索引入大语言模型（large language model，LLM）的智能体推理机制，借助LLM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，实现对复杂非常规表格结构的高效处理，进而推动表格结构理解与表格解析技术向智能化方向发展。

参考文献：

[1] Hou Q Y, Wang J. TABLET: table structure recognition using encoder-only Transformers[C]//Proceedings of the 19th International Conference. Cham: Springer, 2025: 253-278.

[2] Zhong X, ShafieiBavani E, Jimeno Yepes A. Image-based table recognition: data, model, and evaluation[C]//Proceedings of the 16th European Conference. Cham: Springer, 2020: 564-580.

[3] Yu C L, Li W B, Li W, et al. A survey for table recognition based on deep learning[J]. Neurocomputing, 2024, 600: 128154.

[4] Huang Y S, Lu N, Chen D P, et al. Improving table structure recognition with visual-alignment sequential coordinate modeling[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 11134-11143.

[5] Lin W H, Sun Z, Ma C X, et al. TSRFormer: table structure recognition with transformers[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM Press, 2022: 6473-6482.

[6] Salaheldin K 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: 314-318.

[8] Qasim S R, Mahmood H, Shafait F. Rethinking table recognition using graph neural networks[C]//Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR). Piscataway: IEEE Press, 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 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 4523-4532.

[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 based on 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[PP]. V3.arXiv(2018-02-04)[2025-01-26].arXiv:1710.10903.

[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, 26: 105200.

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[PP]. V7. arXiv(2023-08-03)[2025-03-26]. arXiv:1706.03762.

[15] Wang Z R, Dong H Y, 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. New York: ACM Press, 2021: 1780-1790.

[16] Lu N, Yu W W, Qi X B, 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[PP]. V2. arXiv(2023-03-29) [2025-06-27]. arXiv:2303.08648.

[18] Yang G J, Zhong D J, Xiong Y J, et al. LRATNet: local-relationship-aware Transformer network for table structure recognition[C]//Proceedings of the 30th International Conference. Cham: Springer, 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. Kerrville: ACL, 2020: 4320-4333.

[21] Yin P C, Neubig G, Yih W T, 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. Kerrville: ACL, 2020: 8413-8426.

[22] Zhang Z R, Liu S H, Hu P F, et al. UniTabNet: bridging vision and language models for enhanced table structure recognition[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024. Kerrville: ACL, 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[PP].V2. arXiv(2024-05-27)[2025-06-07]. arXiv:2403.04822.

[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 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 4604-4613.

[26] Zheng X Y, Burdick D, Popa L, et al. Global table extractor (GTE): a framework for joint table identification and cell structure recognition using visual context[C]//Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2021: 697-706.

[27] Li M H, Cui L, Huang S H, et al. TableBank: table benchmark for image-based table detection and recognition[C]//Proceedings of the Twelfth Language Resources and Evaluation Conference. Kerrville: ACL, 2020: 1918-1925.

[28] Jimeno Y A, Zhong P, Burdick D. ICDAR 2021 competition on scientific literature parsing[C]//Proceedings of the 16th International Conference. Cham: Springer, 2021: 605-617.

[29] Ly N, Takasu A, Nguyen P, et al. Rethinking image-based table recognition using weakly supervised methods[C]//Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods. ICPRAM, 2023: 872-880.

[30] Chen L Y, Huang C S, Zheng X Q, 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). Kerrville: ACL, 2023: 2437-2449.

[作者简介]



徐崇善（2001– ），男，武汉理工大学计算与人工智能学院硕士生，主要研究方向为人工智能、计算机视觉、表格解析等。



吴绿（1983– ），女，博士，武汉理工大学信息工程学院副教授，主要研究方向为基于AI的电子元器件参数提取。



李宁（1983– ），男，博士，武汉理工大学计算机与人工智能学院副教授，主要研究方向为工业软件、人工智能。



徐宁（1968– ），男，博士，武汉理工大学信息工程学院教授，主要研究方向为芯片布局布线。



潘昊（1964– ），男，博士，武汉理工大学计算机与人工智能学院教授，主要研究方向为人工智能、物联网技术、软件工程等。