

TFLOP：基于布局指针机制的表格结构识别框架

摘要

Minsoo Khang 和 Teakgyu Hong

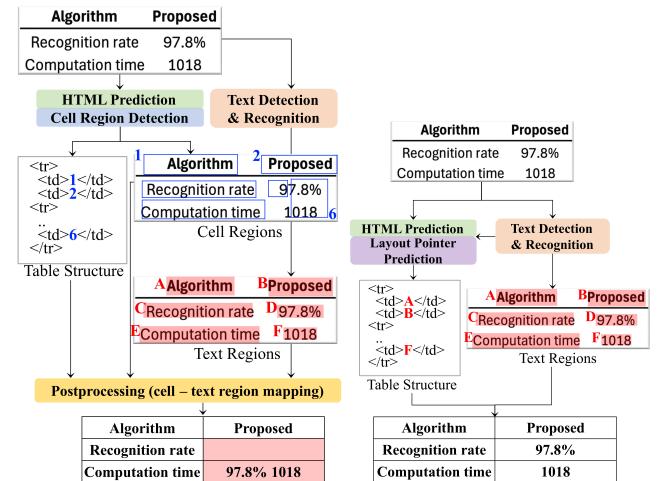
Upstage AI, 韩国

{mkhang, tghong}@upstage.ai

表格结构识别 (Table Structure Recognition, TSR) 旨在将表格图像转换为机器可读的格式 (如HTML)，以促进信息检索等应用。近期的工作通过识别HTML标签和文本区域来解决这一问题，后者用于从表格文档中提取文本。然而，这些工作在将文本映射到识别的文本区域时存在对齐问题。本文提出了一种新的TSR框架，称为TFLOP (带有布局指针机制的TSR框架)，该框架将传统的文本区域预测和匹配问题重新表述为直接的文本区域指向问题。具体而言，TFLOP利用文本区域信息同时识别表格的结构标签及其对齐的文本区域，无需进行区域预测和对齐。TFLOP避免了需要精细校准后处理的额外文本区域匹配阶段。此外，TFLOP采用了跨度感知的对比监督，以增强复杂结构表格中的指向机制。因此，TFLOP在多个基准测试 (如PubTabNet、FinTabNet和SynthTabNet) 中达到了最先进的性能。在我们的广泛实验中，TFLOP不仅表现出竞争性的性能，而且在带有水印或非英语领域的工业文档TSR场景中也展示了有前景的结果。我们的工作源代码公开发布于：<https://github.com/UpstageAI/TFLOP>。

1 Introduction

表格在各种文档 (如商务文件、学术论文) 中广泛使用，因其紧凑且高效的表示方式。然而，这种紧凑的表示方式对直接的机器解析提出了重大挑战。表格结构识别 (Table Structure Recognition, TSR) 旨在将表格图像数字化为机器可读的格式 (如HTML)，表示其结构和文本，从而支持各种下游应用，如信息检索或表格问答。



(a) Dual decoder framework

(b) TFLOP

图1：两种TSR框架的概述。双重解码器识别表格单元区域及其HTML结构，需要进一步进行单元格和文本区域映射以获得最终输出。相比之下，TFLOP利用文本区域信息并直接识别带有相应文本区域关系的HTML结构。

TSR通常包括在构建完整表格结构之前预测两组结构：逻辑结构和物理结构 [Huang et al., 2023]。逻辑结构表示表格单元格的语义组织和关系信息，通常以HTML或LaTeX的形式表示。另一方面，物理结构表示表格单元格的布局信息，如它们的边界框。

最近的研究采用了图像到文本的方法，其中逻辑结构和物理结构都被预测，后者以前者为条件。首先将物理结构 (单元格边界框) 映射到表格的文本区域，这些文本区域是通过OCR引擎或通过PDF解析获得的，然后将匹配的文本与逻辑结构结合，形成完整的表格。尽管在逻辑结构预测方面表现出色，但这些框架常常遇到对齐问题，即由于表格文本区域与预测的单元格边界框之间对齐不完善，导致错误的文本被匹配。这种方法在文本匹配过程中需要精细校准的后处理。

区域以获得满意的结果。本工作TFLOP旨在通过利用布局指针机制直接在框架中处理表格文本区域，从而消除基于启发式的边界框匹配需求。TFLOP将原始的边界框预测问题重新定义为边界框指向问题。具体而言，它不是在逻辑结构条件下预测单元格边界框，而是通过指针机制预测边界框与逻辑序列之间的关联。TFLOP的指针机制不仅解决了对齐问题，还消除了基于启发式的边界框匹配需求。除了对齐问题外，识别具有行或列跨度的表格（即复杂表格）的结构是TSR的关键挑战之一。利用我们框架的灵活性，TFLOP在处理表格文本区域时采用跨度感知的对比监督，以提高对复杂表格的识别。基于提出的指针机制和跨度感知的对比监督，TFLOP在流行的TSR基准测试中达到了最先进的性能。在本工作中，我们超越了基准数据集，从工业角度探索了TFLOP的多功能性。我们进行了广泛的实验，并展示了TFLOP不仅具有竞争性能，而且在处理工业文档TSR场景（如带水印的文档或甚至是非英语表格）时也具有多功能性，尽管它仅在英语表格上进行了训练。我们工作的主要贡献如下：

- 我们提出了一种新颖的TSR框架，结合了布局指针机制，不仅解决了文本区域对齐问题，还消除了将文本区域映射到预测单元格边界框时所需的后期处理步骤。
- 我们还引入了跨度感知的对比监督机制在我们的框架中。这种监督机制增强了模型识别涉及行或列跨度的复杂表格结构的能力。
- TFLOP在多个流行的TSR基准测试中达到了最先进的性能。
- 除了基准测试性能外，TFLOP在处理工业TSR文档场景时也展现了竞争力和多功能性，例如带有水印的表格或非英语领域的表格。

2相关工作

TSR方法涵盖了处理表格的逻辑和物理结构的不同变体。这些方法大致可以分为两类：基于检测的方法和图像到文本的方法。

2.1 基于检测的交通标志识别方法

基于检测的TSR是常见的识别表格结构的方法之一，它利用检测到的表格特征，如分隔线或单元格级特征。这些方法通常先进行物理结构的理解，然后再推理相应的逻辑结构以完成TSR。

基于网格的方法代表了利用网格表示的检测表格特征的方法。早期的工作[Schreiber et al., 2017; Paliwal et al., 2019]通过基于分割的方法检测行和列的掩码，然后将它们聚合以形成表格结构。SPLERGE [Tensmeyer et al., 2019]提出了一个分割-合并的流水线，首先检测与表格匹配的网格结构，然后合并相邻单元格以处理跨行和跨列的条目。后续工作在此网格表示的基础上进行了改进，例如TRUST [Guo et al., 2022]提出了基于查询的分割和基于顶点的合并模块，以改进跨单元格的预测，而SEM [Zhang et al., 2022]提出了在表格网格生成中聚合视觉和文本特征。RobusTabNet [Ma et al., 2023]提出了一种空间CNN模块，在预测单元格网格检测前的分隔线时改进了物理结构推理。后续工作TSRFormer [Lin et al., 2022]将线预测任务重新定义为回归问题，而不是通过基于两阶段DETR [Carion et al., 2020]的方法进行图像分割。最近的工作GridFormer [Lyu et al., 2023]提出了一种新方法，直接从表格图像中预测表格网格的顶点和边（逻辑结构）。

基于单元格的方法是另一种基于检测的方法，首先检测单元格级别的特征（物理结构），然后对单元格之间的关系（逻辑结构）进行分类，以形成完整的表格结构。一些代表性工作包括TabStructNet [Raja et al., 2020]和FLAG-Net [Liu et al., 2021a]，它们是端到端的框架，利用DGCNN架构 [Wang et al., 2019b]来建模检测到的单元格级别特征之间的关系。最近，Hetero-TSR [Liu et al., 2022]提出了NCGM，旨在改进处理复杂TSR场景时的跨模态协作。

2.2 基于图像到文本的TSR方法

图像到文本方法将TSR任务重新定义为图像到序列的翻译任务，其中表格结构被表示为一个文本序列（例如HTML、LaTeX等）。近期方法通常首先预测表格的逻辑结构，然后在此基础上进行物理结构预测。这两个预测结果随后被整合以形成完整的表格结构。早期的图像到文本TSR工作直接生成完整的表格结构，例如[Deng et al., 2019]提出的基于LSTM的table-to-LaTeX框架。相比之下，近期的工作转向了基于Transformer的序列生成模型，这些模型分别生成逻辑和物理结构，然后再将其整合以形成完整的表格结构。这类工作的显著例子包括[Ye et al., 2021; Nassar et al., 2022]，他们提出了图像编码器双解码器（IEDD）方法。在这些工作中，在编码表格图像后，一个解码器首先预测HTML结构标签（逻辑结构），而另一个解码器则在这些标签的基础上生成单元格的边界框（物理结构）。这些单元格边界框随后通过OCR引擎或PDF解析映射到表格的文本区域，最后与逻辑结构整合。

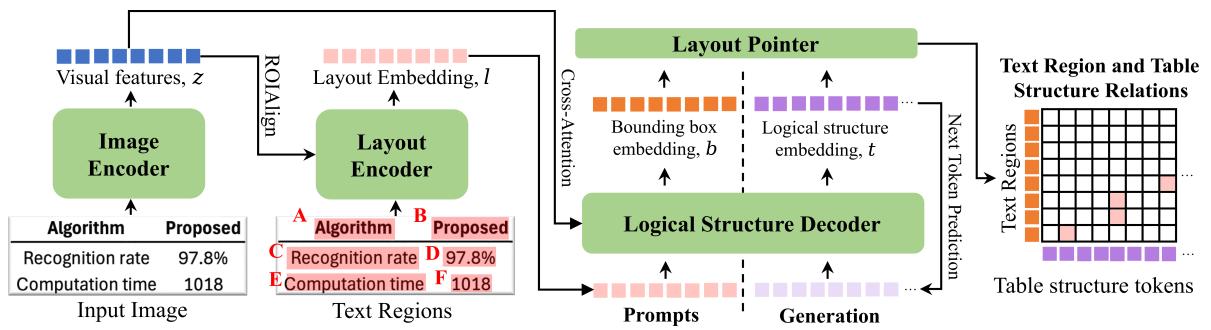


图2：TFLOP概览示意图。给定一个表格图像及其文本区域边界框，图像和布局编码器输出视觉特征和布局嵌入。逻辑结构解码器接收这些特征，自回归生成表格结构标记（标签），同时通过布局指针预测文本区域边界框与表格数据标签之间的关联。这些关联和表格标签被聚合以生成完整的表格结构。

逻辑结构以完成HTML序列。后续工作提出了不同的方法来改进双解码器框架。VAST [Huang et al., 2023] 提出了视觉对齐损失，通过在解码阶段强制加入详细的视觉信息来改进物理结构的预测。同时，DRCC [Shen et al., 2023] 提出了一种半自回归方法，减少了逻辑和物理结构生成中的错误累积效应。尽管这两项工作的贡献确实改进了结构预测，但它们都面临着边界框对齐的固有问题。在映射预测的单元格边界框以进行文本检索时，边界框与表格文本区域之间的对齐错误可能导致表格结构错误。因此，基于物理结构预测的框架容易受到边界框对齐问题的影响，需要启发式后处理才能获得满意的结果。

3Method

3.1 总体架构

TFLOP 包括四个模块：图像编码器、布局编码器、逻辑结构解码器和布局指针。我们的框架接收一个表格图像及其对应文本区域，这些文本区域可以是单元格级别的标注，也可以通过现成的OCR引擎获取。TFLOP首先使用图像编码器从表格图像中提取视觉特征，同时使用布局编码器嵌入文本区域的边界框。生成的视觉特征和布局嵌入随后由逻辑结构解码器处理。视觉特征通过交叉注意力机制提供给解码器，而布局嵌入则作为上下文提示用于生成逻辑结构序列。在自回归生成逻辑结构（例如HTML标签）的基础上，解码器的最后一个隐藏状态进一步由布局指针模块处理，该模块将预测的表格数据标签（例如HTML的标签，OTSL的C标签）与相应的文本区域关联，以形成完整的表格结构。TFLOP架构如图2所示。

3.2 图像编码器

受Donut架构[Kim et al., 2022]的启发，我们采用Swin Transformer[Liu et al., 2021b]作为TFLOP的图像编码器。所有表格图像均被预处理为固定分辨率，并嵌入为视觉特征， $\{z_i | z_i \in R^d, 1 \leq i \leq P\}$ ，其中P是图像块的数量，d是潜在向量的维度。

3.3 布局编码器

布局编码器由多个MLP模块组成，这些模块嵌入了文本区域的边界框及其对应的2x2 ROIAlign [He et al., 2017]应用于视觉特征 $\{z_i\}$ 。这些嵌入被聚合以形成布局嵌入 $\{l_j | l_j \in R^d, 1 \leq j \leq B\}$ ，其中B是布局嵌入的上下文长度。

3.4 逻辑结构解码器

逻辑结构解码器根据视觉特征 $\{z_i\}$ 和布局嵌入 $\{l_j\}$ 生成表格标签序列。TFLOP采用BART [Lewis et al., 2019]架构，并遵循与Donut [Kim et al., 2022]相似的配置。TFLOP的解码器输出一个序列 $\{y_k | y_k \in R^v, 1 \leq k \leq T\}$ ，其中T是表格标签的总数，v是标记词汇表的大小。交叉熵损失Lcls用于监督解码器的标签分类。先前工作的解码器[Shen et al., 2023; Huang et al., 2023; Nassar et al., 2022]生成逻辑结构序列，通常采用HTML格式。尽管HTML表示的序列较长，但由于其灵活性和广泛覆盖的表格布局，仍被广泛使用。为了减少其长序列长度，[Huang et al., 2023; Ye et al., 2021]合并了特定标签（例如）。TFLOP通过生成OTSL标签序列[Lysak et al., 2023]实现了类似效果，这些序列与目标HTML序列具有1对1的映射关系。

3.5 布局指针

除了生成一系列表格标签外，解码器的最后一个隐藏状态特征， $\{h_i | h_i \in R^d, 1 \leq i \leq N\}$ ，被用于我们的布局指针模块中。N是边界框数量（B）和表格标签数量（T）的总和。

具体来说，特征序列 $\{hi\}_{i=1}^N$ 首先被分割成两个子序列： $\{bj\}_{j=1}^B$ and $\{tk\}_{k=1}^T$. $\{bj\}_{j=1}^B$ 是一个固定长度为B的序列，代表边界框的最后一个隐藏状态特征。另一方面， $\{tk\}_{k=1}^T$ 是一个长度为T的序列，代表预测的表格标签的最后一个隐藏状态特征。这两个特征序列随后通过线性变换（公式1）被投影为 $\{\bar{b}_j\}$ 和 $\{\bar{t}_k\}$ 。在表格标签特征 $\{\bar{tk}\}$ 中，我们定义那些对应于表格数据标签的索引为集合D。布局指针监督随后按照公式2应用。

$$\bar{b}_j = \text{proj}_b(b_j), \quad \bar{t}_k = \text{proj}_t(t_k)$$

$$\mathcal{L}_{ptr} = -\frac{1}{B} \sum_{j=1}^B \log\left(\frac{\exp(\bar{b}_j \cdot \bar{t}_{k^*}/\tau)}{\sum_{k' \in D} \exp(\bar{b}_j \cdot \bar{t}_{k'}/\tau)}\right)$$

L_{ptr} 表示布局指针监督的损失，其中 \bar{b}_j 表示第 j^{th} 个边界框的投影特征， k^* 是与第 j^{th} 个边界框对应的表格数据标签的索引。 \cdot 和 τ 分别表示点积和温度超参数。值得注意的是，边界框与表格标签之间可能存在一对一或多对一的关系，因为一个表格单元格内可能有一个或多个文本边界框。因此，在公式2中， L_{ptr} 通过对每个边界框计算负对数似然，然后取其算术平均值来计算。还应注意，表格数据标签可能没有任何对应的边界框（即空表格单元格）。为了确保对所有表格数据标签提供指针监督，对那些没有任何对应边界框的标签应用了单独的损失监督 $L_{empty_{ptr}}$ ，如下所示：

$$\mathcal{L}_{empty_{ptr}} = -\frac{1}{|D|} \sum_{k' \in D} \text{BCE}(\sigma(\bar{b}_0 \cdot \bar{t}_{k'}), I(k'))$$

\bar{b}_0 是对专门用于空表格数据标签的特殊嵌入的线性投影。 $\sigma()$ 和 $\text{BCE}()$ 分别表示sigmoid激活函数和二元交叉熵，而 $I(k')$ 表示二元标签，指示 k' 数据标签是否为空。

3.6 跨度感知对比监督

为了更好地处理复杂的表格结构（包含rowspan或colspan），TFLOP在边界框嵌入 $\{bj\}$ 之间采用了跨度的对比监督，以提升其对表格布局的理解。尽管先前的研究在行和列两个方向上对表格元素提供了对比监督，但TFLOP在此基础上进一步引入了跨度感知的调整。给定一个 j^{th} 边界框嵌入 bj ，首先对其进行投影。

使用线性层来形成 \hat{b}_j （公式4），然后在公式5中评估其跨度感知的对比损失。

$$\hat{b}_j = \text{proj}_s(b_j)$$

$$\mathcal{L}_{contr,j} = -\frac{1}{\sum_{p \in P(j)} c_p(j)} \sum_{p \in P(j)} c_p(j) \log\left(\frac{\exp(\hat{b}_j \cdot \hat{b}_p/\tau)}{\sum_{a \in A(j)} \exp(\hat{b}_j \cdot \hat{b}_a/\tau)}\right)$$

Dopant	SnO ₂ -d (110) in A					
	Urea	Ammonia	Chemically Doped	Impregnated Powders	Urea	Ammonia
Undoped	3.35853	3.35492	-	-	3.36927	3.3682
Cu	-	-	3.3732	3.36927	3.3682	3.3927
Pt	-	-	3.39422	3.38839	3.3801	3.35427
Pd	-	-	3.41855	3.40501	3.40697	3.35706
Target	Partial-Positive	□) =	□) =	□) =	□) =	□) =
Positive	Negative	□) =	□) =	□) =	□) =	□) =

图3：涉及多跨结构的范围感知对比监督的示例可视化。在上面的按列对比监督示例中，对于给定的边界框（i，粉色），正样本（P(i)）是那些具有完全重叠（绿色）或部分重叠（橙色）的样本，而其余的（红色）为负样本。

$L_{contr,j}$ 表示第 j^{th} 个边界框的跨度感知对比损失。此公式适用于行跨度和列跨度的监督。这里， $A(j)$ 表示除第 j^{th} 个边界框外的所有边界框， $P(j)$ 表示 $A(j)$ 中所有正样本的边界框（即与第 j^{th} 个边界框处于同一行或同一列）， \hat{b}_p 和 \hat{b}_b 分别表示 $P(j)$ 和 $A(j)$ 的投影边界框嵌入。上述公式类似于监督对比损失[Khosla et al., 2020]，除了跨度系数 $c_p(j)$ 。跨度系数 $c_p(j)$ 表示第 j^{th} 个边界框与 p 之间的接近程度，基于两个边界框之间的跨度重叠。例如，参考图3，在列向对比监督中，第 j^{th} 个边界框（粉色）与正样本边界框（绿色或黄色）之间的跨度系数可以表示为：

$$c_p(j) = \frac{\text{overlap}(p, j)}{\text{span}(p) \times \text{span}(j)}$$

这里， $\text{span}()$ 表示给定边界框的跨度计数（行或列），而 $\text{overlap}(x, y)$ 表示边界框 x 和 y 之间的重叠单元格数量。例如，图3中“Ammonia”的边界框相对于“Chemically Doped”的边界框的跨度系数为 $1/(2 \times 1)$ 。值得注意的是，当跨度系数设置为常数值1（即均匀对比监督）时， $L_{contr,j}$ 简化为标准的监督对比损失公式[Khosla et al., 2020]。

3.7 损失函数

TFLOP的训练目标由标签分类损失、布局指针损失和跨度感知对比损失组成。标签分类损失（ L_{cls} ）通过表格标签预测的负对数似然进行评估，而布局指针损失是 L_{ptr} 和 $L_{empty_{ptr}}$ 的线性组合。跨度感知对比损失也是 $L_{row_{contr,j}}$ 和 $L_{col_{contr,j}}$ 的线性组合，分别表示第 j^{th} 个边界框的行向和列向对比损失。

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{ptr} + \lambda_3 \mathcal{L}_{empty_{ptr}} \quad (7)$$

$$+ \lambda_4 \frac{1}{B} \sum_{j=1}^B \mathcal{L}_{row_{contr,j}} + \lambda_5 \frac{1}{B} \sum_{j=1}^B \mathcal{L}_{col_{contr,j}}$$

Methods	PubTabNet .Val		PubTabNet .Test	
	TEDS-S	TEDS	TEDS-S	TEDS
TableMaster [2021]			96.32	
LGPMA [2021]	96.7	94.6		
TableFormer [2022]	97.5		96.75	93.60
VAST [2023]			97.23	96.31
RobusTabNet [2023]	97.0			
DRCC [2023]	98.9	97.8		
TFLOP BASE	98.1	97.8	98.25	96.42
TFLOP FULL	98.3	98.0	98.38	96.66

表1：TEDS-Struct (TEDS-S) 和TEDS在PubTab-Net验证集和测试集上的评估结果。

4 Experiments

4.1 Datasets

为了验证我们框架的有效性，我们在三个流行的TSR基准数据集上进行了实验：PubTabNet [Zhong et al., 2020]、FinTabNet [Zheng et al., 2021] 和 SynthTabNet [Nassar et al., 2022]。

PubTabNet 是大规模表格识别 (TSR) 数据集之一，包含从科学文章中提取的表格的 HTML 标注。它由 500,777 张训练图像和 9,115 张验证图像组成。随后发布了包含 9,064 张图像的标注测试数据集，本研究报告了 TFLOP 在验证和测试数据集上的 TSR 性能。需要注意的是，对于 PubTabNet 测试数据集，未提供单元格级别的标注（即文本区域边界框），而是使用了现成的 OCR 引擎来获取这些标注。

FinTabNet 是一个由财务报告中的单页 PDF 文档组成的流行 TSR 基准。该数据集包含从文档中提取的 112,887 个表格及其单元格级别的标注。FinTabNet 有助于评估 TFLOP 在表格中的性能，其中文本区域不是通过 OCR 引擎获取的（即不受 OCR 相关噪声影响，类似于 PDF 解析）。

SynthTabNet 由 [Nassar et al., 2022] 引入，作为一个不仅规模大而且表格外观和内容多样化的基准数据集。SynthTabNet 由 600,000 张不同风格的表格图像组成，并提供类似于 FinTabNet 的单元格级别标注。

4.2 实验设置

在TFLOP的训练中，输入图像分辨率在所有基准数据集上均设置为 768×768 。输出序列长度 N 固定为 1,376，以确保足够的布局嵌入长度和表格标签的生成。框架的特征维度 d 设置为 1,024，损失公式 (Equation 7) 的超参数为：
 $1 = 2 = 3 = 1$ 和 $4 = 5 = 0.5$ 。温度值 τ 设置为 0.1。所有实验在 $4 \times$ A100 GPU 上进行，训练步数为 250K。

Methods	FinTabNet		SynthTabNet	
	TEDS-S	TEDS	TEDS-S	TEDS
TableFormer [2022]	96.80		96.70	
GridFormer [2023]	98.63			
VAST [2023]	98.63	98.21		
DRCC [2023]			98.70	
TFLOP BASE	99.43	99.22	99.42	99.34
TFLOP FULL	99.56	99.45	99.42	99.40

表2：TEDS-Struct / 在FinTabNet和SynthTabNet上的TEDS。

4.3 评估指标

为了评估TFLOP的性能，我们采用了基于树编辑距离的相似度，即TEDS [Zhong et al., 2020]，以及TEDS-Struct [Huang et al., 2023; Nassar et al., 2022]，分别计算预测的HTML表格结构与真实HTML表格结构在包含和不包含表格文本内容情况下的TEDS得分。

$$\text{TEDS}(T_{pr}, T_{gt}) = 1 - \frac{\text{EditDist}(T_{pr}, T_{gt})}{\max(|T_{pr}|, |T_{gt}|)}$$

在公式8中，T和|T|分别表示HTML结构和T中的节点数量，而EditDist()表示HTML结构之间的树编辑距离。

4.4 Results

我们使用三个流行数据集对TFLOP进行了基准测试，结果如表1和表2所示。在所有基准测试中，我们不仅报告了TFLOPFULL的结果，还报告了TFLOPBASE的结果，以便更好地评估我们的布局指针机制的有效性。TFLOPBASE与TFLOPFULL的区别在于缺少图像ROIAlign和跨度感知的对比监督。表1的结果显示，TFLOP在PubTabNet的验证和测试集的完整表格结构识别中优于先前的工作。为了确保与仅在验证集上报告结果的先前工作进行公平比较，我们在PubTabNet的验证数据集上进行了评估。对于测试数据集，由于未提供单元级注释，我们使用PSENet [Wang et al., 2019a] 和 Master [Lu et al., 2021] 获取文本区域边界框注释，类似于先前的工作 [Ye et al., 2021; Guo et al., 2022; Huang et al., 2023]，以确保公平比较。TFLOP在PubTabNet测试数据集上的最先进性能清楚地展示了我们的框架在使用现成OCR引擎导出的文本区域时的有效性。图4中的可视化展示了TFLOP识别具有复杂结构的表格的能力，例如层次化的行跨度（顶部）和层次化的列跨度（底部）。表2的结果也证实了TFLOP在各种先前工作中表现出的优越性能，通过在FinTabNet和SynthTabNet上实现最先进的识别结果。FinTabNet是从财务报告中提取的表格，TFLOP的最先进性能（99.45 TEDS）在工业应用中具有重要意义，因为在这些应用中容错率极低。另一方面，SynthTabNet包含各种风格的表格结构，TFLOP的最先进性能（99.40

Advanced, hormone naïve		GnRH agonist	1989			
Histrelin acetate	GnRH agonist	1989				
Triptorelin pamoate	GnRH agonist	1991				
Abarelix	GnRH antagonist	2003				
Degarelix	GnRH antagonist	2008				
Bicalutamide	AR antagonist	1995				
Flutamide	AR antagonist	1989				
Castrate-resistant		Abirateron acetate	CYP17 inhibitor 2011 ^b 2012 ^a			
Area(Wales)		Enzalutamide	AR antagonist 2012 ^b 2014 ^a			
Location and Violence Strata						
North		Urban	Town/Fringe			
4	2	High	Low	High	Low	
2	1	8	6	3	10	
...	
Location and Violence Strata						
Urban		Town/Fringe	Rural			
...	High	Low	High	Low	High	Low
...	8	6	3	10	1	2
...

图4：从生成的HTML序列构建的表格的可视化，附有相应的表格图像（为提高可读性重新创建）以供参考。TFLOP成功构建了具有复杂结构的表格，例如层次化的行跨度（顶部）或层次化的列跨度（底部）。

Methods	Simple	Complex	All
TFLOP _{BASE}	97.92	94.85	96.42
TFLOP _{BASE + I}	+0.0	+0.1	+0.0
TFLOP _{BASE + I + U}	+0.0	+0.1	+0.0
TFLOP _{BASE + I + S}	+0.1	+0.3	+0.2
TFLOP _{FULL}	98.06	95.20	96.66

表3：在PubTabNet测试数据集上对I (ImageROI)、U (Uniform contrastive) 和S (Span-aware contrastive) 进行消融实验的TEDS (%) 结果。注意，TFLOP_{BASE + I + S} 和TFLOP_{FULL} 是等效的。

TEDS) 清楚地表明，该框架并不局限于特定的表格格式或风格。在表1和表2中，可以注意到TFLOP在TEDS-Struct指标（仅限HTML表格标签）方面也取得了显著的改进。我们认为这是我们框架中布局嵌入的一个副作用。提供布局嵌入对于我们框架的布局指针机制至关重要，因为它作为从生成的表格标签中指向的目标。顺便提一下，这种布局嵌入也可能提高框架对表格布局的理解，从而改进表格标签的生成，正如TFLOP的TEDS-Struct结果所示。除了在基准数据集上达到最先进的TEDS得分外，还值得注意的是我们框架的TEDS-Struct和TEDS得分之间的差距，相比于先前的工作。虽然TEDS指标评估了完整表格结构的准确性，但TEDS-Struct和TEDS之间的差距间接表明了边界框错位（对于先前的工作）的重要性，或者

6	2005-2006	State	Alzheimer's Disease		Unspec. dementia		Total dementia	
			Count	Percent	Count	Percent	Count	Percent
18	Washington	19	6,669	20.48	23	4,956	24	36
...
41	1999-2000	Washington	3,342	53	41	...	4,014	49
...
...	8,213	100

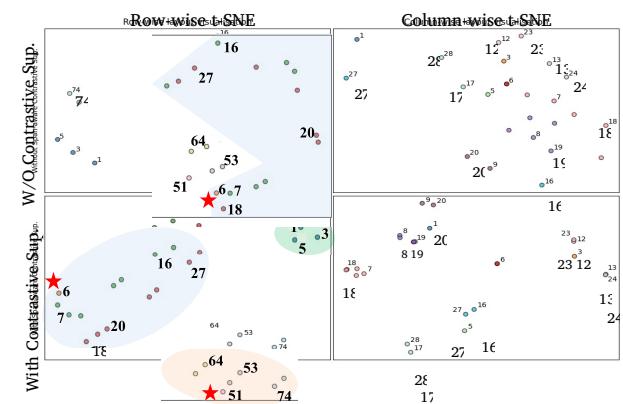


图5：行和列的t-SNE可视化展示的边界框嵌入。PubTabNet表格图像（顶部，表格已重新绘制以提高可读性）和25个边界框被抽样用于可视化。填充颜色代表不同的行跨度组，而边框颜色代表不同的列跨度组，这些组被抽样用于可视化。t-SNE图中的颜色与上方的表格相匹配，跨越多行/列的方框用红色星号标记。

布局指针机制的重要性（针对TFLOP）。除了PubTabNet测试数据集外，该数据集的TEDS指标还受到PSENet [Wang et al., 2019a]和Master [Lu et al., 2021]的OCR错误影响，TFLOP在剩余的基准测试中始终实现了TEDS-Struct和TEDS之间最小的差距（例如，在FinTabNet中为0.11 vs 0.42）。这清楚地展示了布局指针机制在解决先前工作面临的边界框对齐问题上的有效性。

4.5 消融研究

除了布局指针机制外，我们还通过比较表3中的TFLOP_{BASE}和TFLOP_{FULL}，分析了框架中其他组件的有效性。首先，对于图像ROI对齐，与[Huang et al., 2023; Shen et al., 2023]一致，从表3可以明显看出，将ROI对齐的视觉特征融入布局嵌入中，对于识别表格结构也有益于我们的框架。其次，表3显示，与其他方法配置相比，跨度感知的对比监督带来了明显的性能提升。对于结构复杂的表格，这种性能提升尤为显著，表明我们的跨度感知对比监督有助于框架提升对具有行或列跨度的表格的识别能力。这一点也可以在包围盒嵌入空间的t-SNE可视化（图5）中观察到，其中嵌入明显地结构化为行或列跨度的集群。

5 TFLOP 多功能性

除了在时间序列回归 (TSR) 方面的出色表现外，我们还进一步探索了我们的框架在工业应用中常见的两种场景中的多功能性：表格数据



图6：带有“草稿版本”水印的FinTabNet图像展示了在TSR中使用双解码器框架处理水印的挑战。蓝色和绿色方框分别表示文本区域和水印区域，红色表示一个样本预测。

Methods	IOU	TEDS-Struct (%)	TEDS (%)
TableMaster		82.18	72.83
Gold _{greedy}	0.0		96.45
Gold _{selective}	0.5		98.16
TFLOP FULL		99.54	99.41

表4：在嵌入水印的FinTabNet数据集上的TSR性能。

水印和非英语文本。

5.1 水印传输成功率 (TSR)

与基准数据集中的表格不同，实际工业文档中的表格通常包含水印等不需要的文本。这些不需要的文本如果不准确过滤，可能会导致错误的表格结构识别。基于双解码器框架的先前工作在处理带有水印的表格时并不理想，因为它们需要复杂的边界框匹配启发式方法来正确区分所需文本区域边界框和水印边界框（如图6所示）。相反，我们的工作TFLOP具有多功能性，可以在布局指向之前训练忽略这些水印边界框。为此，我们首先通过将水印绘制到FinTabNet [Zheng et al., 2021]数据集中，准备了一个带有水印的表格数据集。然后，我们使用这个数据集训练TFLOP，仅需要一个两层的多层感知机 (MLP) 和二元交叉熵损失函数的轻微增加。简而言之，在预测边界框和表格标签之间的指针关联之前，训练一个二元分类器来过滤水印边界框。更多细节可以在补充材料中找到。在表4中，我们将TFLOP在带有水印的数据集上的表格结构识别 (TSR) 与TableMaster及假设逻辑结构无误的黄金注释变体进行了比较。Goldgreedy通过包含与文本边界框有任何IOU的水印来构建完整的表格结构，而Goldselective则通过0.5的IOU阈值过滤水印。表4显示了令人鼓舞的结果，TFLOP通过简单的两层MLP过滤掉了大部分水印文本，展示了我们框架的多功能性。

5.2 跨语言TSR

工业TSR应用的另一个重要方面在于其在非英语表格上的表现。由于数据可用性有限，非英语表格上的TSR是一个具有挑战性的任务。尽管TFLOP仅在英语表格上进行了训练，我们仍考察了其在非英语表格上执行TSR的多功能性。

Methods	TEDS (%)		QA Acc. (%)
	Simple	Complex	
Image-only			56.00
GPT-4V	79.43	68.39	78.86
TableMaster	89.96	83.94	82.86
TFLOP	95.76	89.41	92.00

表5：韩语表格的TSR和QA表现。

训练于英语表格。为此，我们自标注了30张韩语表格图像（15张简单表格 & 15张复杂表格），这些图像提取自真实的韩语财务报告，包括表格图像的HTML序列和单元格级别的标注。我们将我们的框架与TableMaster [Ye et al., 2021] 和 GPT-4V [OpenAI, 2023] 进行了基准测试。需要注意的是，TFLOP和TableMaster在评估韩语表格数据集之前，都是在PubTabNet数据集上进行训练的。表5的结果显示了TFLOP在跨语言方面的多功能性，通过显著的优势在简单和复杂的韩语表格上持续超越TableMaster和GPT-4V。为了更好地理解表5中TSR结果的工业应用意义，我们在生成的表格结构基础上进行了额外的问答 (QA) 评估。在评估中，我们构建了175个独特的问答对，其中问题需要对提供的表格有清晰的理解才能准确回答。在评估中，问题和生成的表格结构 (HTML) 连同表格图像一起提供给GPT-4V，然后将其输出与答案标签进行比较。表5中的每个175个答案都经过手动评估，以显示QA的准确性。表5中的QA准确性结果不仅显示了HTML序列对于GPT-4V在非英语领域进行表格QA的重要性，还突显了TEDS评分的差异如何在跨语言设置中转化为表格QA的实际工业应用。数据集的详细信息和定性结果可在补充材料中找到。

6 Conclusion

在这项工作中，我们提出了TFLOP，这是一个基于布局指针机制和跨度感知对比监督的表格结构识别 (TSR) 框架，它不仅解决了边界框对齐问题，还能准确识别复杂结构的表格，无需精细校准的后处理。凭借这些特点，TFLOP在三个流行的TSR基准测试中达到了新的最先进性能。除了强大的TSR性能外，TFLOP还在工业应用场景中展现了显著的多功能性和有前景的表现，特别是在带有水印或非英语领域的文档表格中。

致谢

我们衷心感谢Upstage的同事们，特别是Sungrae Park，感谢他们在整个研究过程中提供的深刻讨论、坚定支持和鼓励。

参考文献

- [Carion 等, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, 和 Sergey Zagoruyko. 使用 transformers 进行端到端对象检测。在欧洲计算机视觉会议 , 页码 213–229。Springer, 2020。
- [Deng 等, 2019] Yuntian Deng, David Rosenberg, 和 Gideon Mann. 端到端神经科学表格识别的挑战。在 2019 年国际文档分析与识别会议 (ICDAR) , 页码 894–901。IEEE, 2019。
- [Guo 等, 2022] Zengyuan Guo, Yuechen Yu, Pengyuan Lv, Chengquan Zhang, Haojie Li, Zhihui Wang, Kun Yao, Jingtuo Liu, 和 Jingdong Wang. TRUST: 使用基于分割的 transformers 进行准确且端到端的表格结构识别。arXiv 预印本 arXiv:2208.14687, 2022。
- [He 等, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, 和 Ross Girshick. Mask R-CNN。在 IEEE 国际计算机视觉会议论文集 , 页码 2961–2969, 2017。
- [Huang 等, 2023] Yongshuai Huang, Ning Lu, Dapeng Chen, Yibo Li, Zecheng Xie, Shenggao Zhu, Liangcai Gao, 和 Wei Peng. 通过视觉对齐的顺序坐标建模改进表格结构识别。在 IEEE/CVF 计算机视觉与模式识别会议论文集 , 页码 11134–11143, 2023。
- [Khosla 等, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, 和 Dilip Krishnan. 监督对比学习。神经信息处理系统进展 , 33:18661–18673, 2020。
- [Kim 等, 2022] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, JinYeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, 和 Seunghyun Park. 无 OCR 的文档理解 transformer。在欧洲计算机视觉会议 , 页码 498–517。Springer, 2022。
- [Lewis 等, 2019] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, 和 Luke Zettlemoyer. BART: 用于自然语言生成、翻译和理解的序列到序列预训练的去噪。arXiv 预印本 arXiv:1910.13461, 2019。
- [Lin 等, 2022] Weihong Lin, Zheng Sun, Chixiang Ma, Mingze Li, Jiawei Wang, Lei Sun, 和 Qiang Huo. TSRFormer: 使用 transformers 进行表格结构识别。在第 30 届 ACM 国际多媒体会议论文集 , 页码 6473–6482, 2022。
- [Liu 等, 2021a] Hao Liu, Xin Li, Bing Liu, Deqiang Jiang, Yinsong Liu, Bo Ren, 和 Rongrong Ji. 展示、阅读和推理 : 使用灵活上下文聚合器的表格结构识别。在第 29 届 ACM 国际多媒体会议论文集 , 页码 1084–1092, 2021。
- [Liu et al., 2021b] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, 和 Baining Guo. Swin transformer: 使用移动窗口的分层视觉transformer。在 IEEE/CVF 国际计算机视觉会议论文集 , 第 10012–10022 页 , 2021年。
- [Liu et al., 2022] Hao Liu, Xin Li, Bing Liu, Deqiang Jiang, Yinsong Liu, 和 Bo Ren. 神经协同图机器用于表格结构识别。在 IEEE/CVF 计算机视觉与模式识别会议论文集 , 第 4533–4542 页 , 2022年。
- [Lu et al., 2021] Ning Lu, Wenwen Yu, Xianbiao Qi, Yihao Chen, Ping Gong, Rong Xiao, 和 Xiang Bai. Master: 多方面非局部网络用于场景文本识别。模式识别 , 117:107980 , 2021年。
- [Lysak et al., 2023] Maksym Lysak, Ahmed Nassar, Nikolaos Livathinos, Christoph Auer, 和 Peter Staar. 优化的表格分词用于表格结构识别。arXiv 预印本 arXiv:2305.03393 , 2023年。
- [Lyu et al., 2023] Pengyuan Lyu, Weihong Ma, Hongyi Wang, Yuechen Yu, Chengquan Zhang, Kun Yao, Yang Xue, 和 Jingdong Wang. Gridformer: 通过网格预测实现精确的表格结构识别。在第 31 届 ACM 国际多媒体会议论文集 , 第 7747–7757 页 , 2023年。
- [Ma et al., 2023] Chixiang Ma, Weihong Lin, Lei Sun, 和 Qiang Huo. 从异构文档图像中进行鲁棒的表格检测和结构识别。模式识别 , 133:109006 , 2023年。
- [Nassar et al., 2022] Ahmed Nassar, Nikolaos Livathinos, Maksym Lysak, 和 Peter Staar. Tableformer: 使用 transformer 理解表格结构。在 IEEE/CVF 计算机视觉与模式识别会议论文集 , 第 4614–4623 页 , 2022年。
- [OpenAI, 2023] OpenAI. Gpt-4v(ision) 系统卡片。2023 年。
- [Paliwal et al., 2019] Shubham Singh Paliwal, D Vishwanath, Rohit Rahul, Monika Sharma, 和 Lovekesh Vig. Tablenet: 用于从扫描文档图像中进行端到端表格检测和表格数据提取的深度学习模型。在 2019 年国际文档分析与识别会议论文集 , 第 128–133 页。IEEE , 2019年。
- [Qiao et al., 2021] Liang Qiao, Zaisheng Li, Zhanzhan Cheng, Peng Zhang, Shiliang Pu, Yi Niu, Wenqi Ren, Wenming Tan, 和 Fei Wu. Lgpma: 使用局部和全局金字塔掩码对齐的复杂表格结构识别。在国际文档分析与识别会议论文集 , 第 99–114 页。Springer , 2021年。
- [Raja et al., 2020] Sachin Raja, Ajoy Mondal, 和 CV Jawahar. 使用自上而下和自下而上线索的表格结构识别。在计算机视觉-ECCV 2020: 第 16 届欧洲会议 , 格拉斯哥 , 英国 , 2020 年 8 月 23–28 日 , 第 XXVIII 部分 , 第 16 卷 , 第 70–86 页。Springer , 2020 年。

[Schreiber et al., 2017] Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, 和 Sheraz Ahmed. Deepdesrt: 文档图像中表格检测和结构识别的深度学习。在2017年第14届IAPR国际文档分析与识别会议(ICDAR)上, 第1卷, 第1162–1167页。IEEE, 2017.

[Shen et al., 2023] Huawei Shen, Xiang Gao, Jin Wei, Liang Qiao, Yu Zhou, Qiang Li, 和 Zhanzhan Cheng. 分隔行并征服单元格 : 面向大型表格的结构识别。在第三十二届国际人工智能联合会议论文集上, 第1369–1377页, 2023.

[Tensmeyer et al., 2019] Chris Tensmeyer, Vlad I Morariu, Brian Price, Scott Cohen, 和 Tony Martinez. 表格结构分解的深度分割与合并。在2019年国际文档分析与识别会议(ICDAR)上, 第114–121页。IEEE, 2019.

[Wang et al., 2019a] Wenhui Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, 和 Shuai Shao. 形状鲁棒的文本检测与渐进尺度扩展网络。在IEEE/CVF计算机视觉与模式识别会议论文集上, 第9336–9345页, 2019.

[Wang et al., 2019b] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, 和 Justin M Solomon. 点云学习上的动态图卷积网络。ACM Transactions on Graphics (tog), 38(5):1–12, 2019.

[Ye et al., 2021] Jiaquan Ye, Xianbiao Qi, Yelin He, Yihao Chen, Dengyi Gu, Peng Gao, 和 Rong Xiao. Pingan-vcgroup在ICDAR 2021科学文献解析任务B中的解决方案 : 表格识别为HTML。arXiv预印本arXiv:2105.01848, 2021.

[Zhang et al., 2022] Zhenrong Zhang, Jianshu Zhang, Jun Du, 和 Fengren Wang. 分割、嵌入与合并 : 一种精确的表格结构识别器。模式识别, 126:108565, 2022.

[Zheng et al., 2021] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, 和 Nancy Xin Ru Wang. 全局表格提取器(GTE) : 一种联合表格识别与单元格结构识别的视觉上下文框架。在IEEE/CVF冬季计算机视觉应用会议上, 第697–706页, 2021.

[Zhong et al., 2020] Xu Zhong, Elaheh Shafiei Bavani, 和 Antonio Jimeno Yepes. 基于图像的表格识别 : 数据、模型与评估。在欧洲计算机视觉会议上, 第564–580页。Springer, 2020.