文档智能：数据集、模型和应用

崔磊，徐毅恒，吕腾超，韦福如

微软亚洲研究院
{lecu, t-yihengxu, tengchaolv, fuwei}@microsoft.com

摘要

文档智能是指通过计算机进行自动阅读、理解以及分析商业文档的过程，是自然语言处理和计算机视觉交叉领域的一个重要研究方向。近年来，深度学习技术的普及极大地推动了文档智能领域的发展，以文档版面分析、文档信息抽取、文档视觉问答以及文档图像分类等为代表的文档智能任务都有显著的性能提升。本文对于早期基于启发式规则的文档分析技术、基于统计机器学习的算法、以及近年来基于深度学习和预训练的方法进行简要介绍，并展望了文档智能技术的未来发展方向。

关键词： 文档智能；文档版面分析；文档信息抽取；文档视觉问答；文档图像分类；深度学习；多模态预训练

Document AI: Benchmarks, Models and Applications

Lei CUI, Yiheng XU, Tengchao LV, Furu WEI

Microsoft Research Asia
{lecu, t-yihengxu, tengchaolv, fuwei}@microsoft.com

Abstract

Document AI, or Document Intelligence, is a relatively new research topic that refers to the techniques to automatically read, understand and analyze business documents. It is an important research direction for the interdisciplinary of natural language processing and computer vision. In recent years, the popularity of deep learning technology has greatly advanced the development of Document AI tasks, such as document layout analysis, document information extraction, document visual question answering, and document image classification etc. This paper briefly introduces the early-stage heuristic rule-based document analysis, statistical machine learning based algorithms, as well as the deep learning based approaches especially the pre-training approaches. Finally, we also look into the future direction of Document AI.

Keywords: Document AI , document layout analysis , document information extraction , document visual question answering , document image classification , deep learning , multimodal pre-training

1 文档智能

文档智能（Document AI, or Document Intelligence）是近年来一项蓬勃发展的研究课题和实际的工业界需求，主要是指对于网页、数字文档或扫描文档所包含的文本以及丰富的排版格式等信息，通过人工智能技术进行理解、分类、提取以及信息归纳的过程。由于布局和格式的多样性、低质量的扫描文档图像以及模板结构的复杂性，文档智能是一项非常具有挑战性的任务并获得相关领域的广泛关注。随着数字化进程的加快，文档、图像等载体的结构化分析和内容提取成为关乎企业数字化转型成败的关键一环，自动、精准、快速的信息处理对于生产力的提升至关重要。以商业文档为例，不仅包含了公司内外部事务的处理细节和知识沉淀，还有大量行业相关的实体和数字信息。人工提取这些信息既耗时费力且精度低，而且可复用性也不高，因此，文档智能技术应运而生。文档智能技术深层次地结合了人工智能和人类智能，在金融、医疗、保险、能源、物流等多个行业都有不同类型的应用。例如：在金融领域，它可以实现财报分析和智能决策分析，为企业战略的制定和投资决策提供科学、系统的数据支撑；在医疗领域，它可以实现病例的数字化，提高诊断的精准度，并通过分析医学文献和病例的关联性，定位潜在的治疗方案。在财务领域，它可以实现发票和采购单的自动化信息提取，将大量无结构化文档进行自动结构化转换，并支撑大量下游业务场景，节省大量人工处理时间开销。

在过去的30年中，文档智能的发展大致经历了三个阶段，从简单的规则启发式方法逐渐进化至神经网络的方法。90年代初期，研究人员大多使用基于启发式规则的方法进行文档的理解与分析，通过人工观察文档的布局信息，总结归纳一些处理规则，对固定布局信息的文档进行处理。然而，传统基于规则的方法往往需要较大的人力成本，而且这些人工总结的规则可扩展性不强，因此研究人员开始采用基于统计学习的方法。2000年以来，随着机器学习技术的发展和进步，基于大规模标注数据驱动的机器学习模型成为了文档智能的主流方法，它通过人工设计的特征模板，利用有监督学习的方式在标注数据中学习不同特征的权重，以此来理解、分析文档的内容和布局。然而，虽然传统的文档理解和分析技术基于人工定制的规则或少量标注数据进行学习，这些方法虽然能够带来一定程度的性能提升，但由于定制规则和可学习的样本数量不足，其通用性往往不尽如人意，而且针对不同类别文档的分析迁移成本较高，这距离文档智能技术的实用化和产业化还有相当一段距离。近年来，随着深度学习技术的发展，以及大量无标注电子文档的积累，文档分析与识别技术进入了一个全新的时代。图1所表示的是在当前深度学习框架下文档智能技术的基本框架，其中不同类型的文档通过内容提取工具（HTML/XML抽取、PDF解析器、光学字符识别OCR等）将文本内容、位置布局信息和视觉图像信息组织起来，利用大规模预训练的深度神经网络进行分析，最终完成各项下游应用任务，包括文档版面分析、文档信息抽取、文档视觉问答以及文档图像分类等。深度学习技术的出现，特别是以卷积神经网络（CNN）、图神经网络（GNN）以及Transformer架构 (Vaswani et al., 2017)为代表预训练技术的出现，彻底改变了传统机器学习需要大量人工标注数据的前提，更多地依赖大量无标注数据进行自监督学习，进而通过“预训练-参数调优”模式来解决文档智能相关的应用任务，取得了显著性突破。

尽管深度学习极大地提高了文档智能技术的准确性，但是在实际应用中仍然有很多问题亟待解决。首先，受限于当前大规模预训练模型输入长度的限制，文档智能预训练模型通常需要将文档截断为几个部分分别输入模型进行处理，这对于复杂长文档的多页跨页处理带来了极大的挑战。其次，由于实际场景中的扫描文档图像质量参差不齐，特别是人工标注的训练数据往往质量较高，而业务场景的文档图像由于扫描设备的清晰度、纸张褶皱和摆放位置的随意性，导致了性能不佳，因而需要利用更多数据增强技术来帮助现有模型提升性能。此外，当前文档智能各项任务通常是独立训练的，不同任务之间的关联性还未被有效的利用，例如文档信息抽取和文档视觉问答有某些共性的语义表示，可以利用多任务学习框架更好的解决这类问题。最后，基于预训练的文档智能模型在实际应用中也遇到了计算资源和训练样本不足的问题，探索基于小模型的深度学习架构和模型压缩技术，以及少样本学习（few-shot learning）和零样本学习（zero-shot learning）技术也是当前重要的研究方向，并具有很大的实用价值。

接下来，我们首先将介绍当前主流的文档智能模型框架、任务和数据集，随后将分别重点介绍早期基于启发式规则的文档分析技术、基于传统统计机器学习的算法模型、以及近年来基于深度学习，特别是基于多模态预训练技术的文档智能模型和算法，最后我们将展望文档智能技术的未来发展方向。

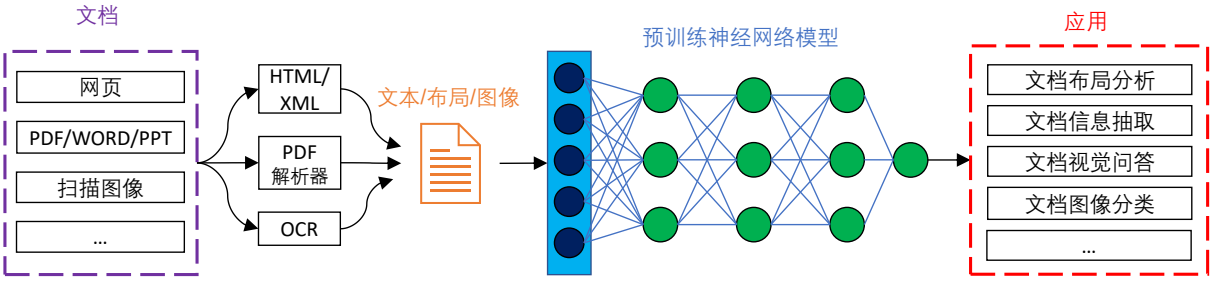


Figure 1: 基于深度学习的文档智能技术框架

2 主流文档智能技术模型框架、任务及数据集

2.1 基于卷积神经网络架构的文档版面分析模型

近年来，卷积神经网络在计算机视觉领域取得了巨大的成功，特别是基于大规模标注数据集ImageNet和COCO的有监督预训练模型ResNet (He et al., 2015)在图像分类、物体检测以及场景分割任务上都带来了极大的性能提升。具体来讲，随着多阶段检测Faster R-CNN (Ren et al., 2016)和Mask R-CNN (He et al., 2018)等模型以及单阶段检测模型SSD (Liu et al., 2016)和YOLO (Redmon and Farhadi, 2018)的普及，物体识别在计算机视觉中几乎成为了已解决问题。文档版面分析本质上可以看作一种文档图像的物体检测任务，文档中的标题、段落、表格、插图等基本单元就是需要检测和识别的物体。(Yang et al., 2017a)将文档版面分析看作是一个像素级分割任务，并尝试利用卷积神经网络进行像素分类取得很好的效果。(Schreiber et al., 2017)首次利用Faster R-CNN模型应用于文档版面分析中的表格识别任务，如图 2所示，在ICDAR 2013 (Göbel et al., 2013)表格识别数据集取得了SOTA的结果。然而，文档版面分析虽然是一个经典的文档智能任务，但是多年来一直受限于较小的数据集规模，仅仅套用经典计算机视觉预训练模型依然是不够的。随着大规模弱监督文档版面分析数据集PubLayNet (Zhong et al., 2019b)、PubTabNet (Zhong et al., 2019a)、TableBank (Li et al., 2020a)以及DocBank (Li et al., 2020b)数据集的出现，研究人员可以对不同的计算机视觉模型和算法进行更为深入的比较和分析，进一步推动了文档版面分析技术的发展。

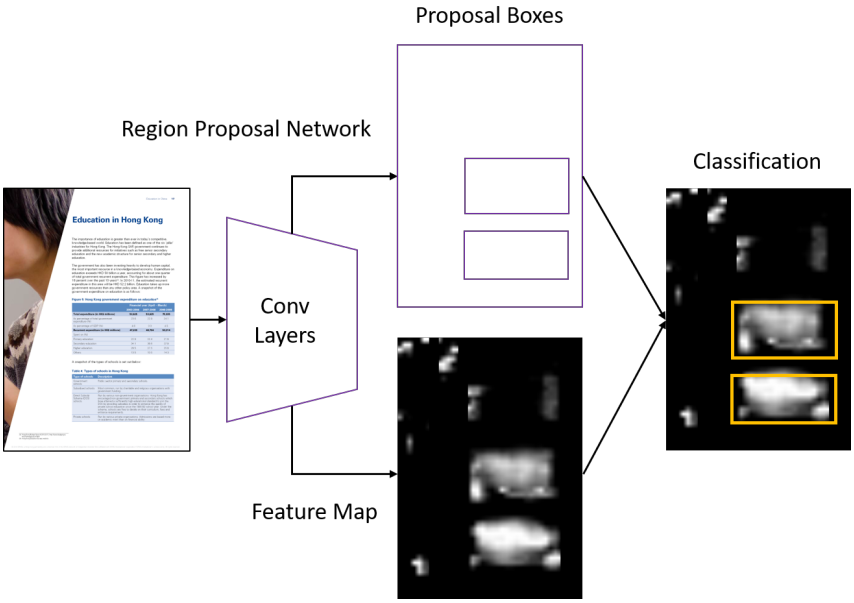
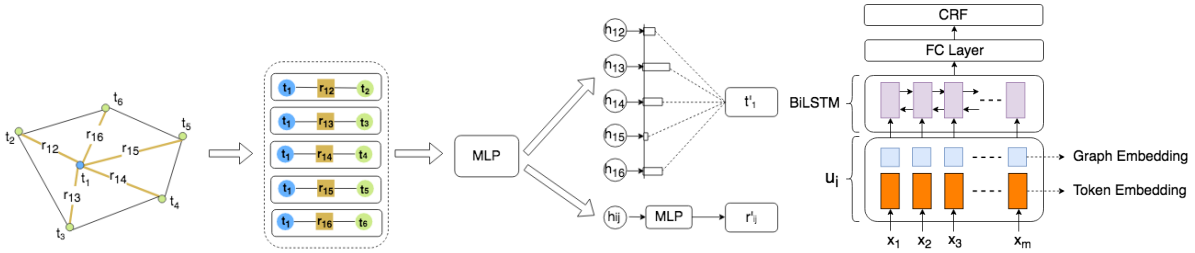


Figure 2: 基于卷积神经网络Faster R-CNN的文档版面分析模型

2.2 基于图神经网络架构的文档信息抽取模型

信息抽取是从非结构化文本中提取结构化信息的过程，其作为一个经典和基础的自然语言



(a) (b)

Figure 3: 基于图神经网络架构的文档信息抽取模型

处理问题已经得到广泛研究。传统的信息抽取聚焦于从纯文本中提取实体与关系信息，却较少有对视觉富文本的研究。视觉富文本数据是指语义结构不仅由本文内容决定，也与排版、表格结构、字体等视觉元素有关的文本数据。视觉富文本数据在生活中随处可见，例如收据、证件、保险单等。(Liu et al., 2019)提出利用图卷积神经网络对视觉富文本数据进行建模，如图3所示。每张图片经过OCR系统后会得到一组文本块，每个文本块包含其在图片中的坐标信息与文本内容。这项工作将这一组文本块构成全连接有向图，即每个文本块构成一个节点，每个节点都与其他所有节点有连接。节点的初始特征由文本块的文本内容通过Bi-LSTM编码得到。边的初始特征为邻居文本块与当前文本块的相对坐标与长宽信息，该特征使用当前文本块的高度进行归一化，具有仿射不变性。与其他图卷积模型仅在节点上进行卷积不同，这项工作更加关注在信息抽取中“个体-关系-个体”的三元信息更加重要，所以在“节点-边-节点”的三元特征组上进行卷积。除此之外，还引入了自注意力机制，让网络在全连接有向图构成的所有有向三元组中挑选更加值得注意的信息，并加权聚合特征。初始的节点特征与边特征经过多层卷积后得到节点与边的高层表征。

这项工作在两份真实商业数据上测试了所提出方法的效果，分别为增值税发票（VATI，固定版式，3000张）和国际采购收据（IPR，非固定版式，1500张）。使用了两个Baseline，Baseline I为对每个文本块的文本内容独立做BiLSTM+CRF解码，Baseline II为将所有文本块的文本内容进行“从左到右、从上到下”的顺序拼接后，对拼接文本整体做BiLSTM+CRF解码。实验表明，基于图卷积的模型在Basline的基础上都有明显提升，其中在仅依靠文本信息就可以抽取的字段（如日期）上与Baseline持平，而在需要依靠视觉信息做判断的字段（如价格、税额）上有较大提升。此外，实验显示，视觉信息起主要作用，增加了语义相近文本的区分度。文本信息也对视觉信息起到一定的辅助作用。自注意力机制在固定版式数据上基本没有帮助，但是在非固定版式数据上有一定提升。

2.3 基于Transformer架构的通用文档理解预训练模型

很多情况下，文档中文字的位置关系蕴含着丰富的语义信息。例如，表单通常是以键值对（key-value pair）的形式展示的。通常情况下，键值对的排布通常是左右或者上下形式，并且有特殊的类型关系。类似地，在表格文档中，表格中的文字通常是网格状排列，并且表头一般出现在第一列或第一行。通过预训练，这些与文本天然对齐的位置信息可以为下游的信息抽取任务提供更丰富的语义信息。对于富文本文档，除了文字本身的位置关系之外，文字格式所呈现的视觉信息同样可以帮助下游任务。对文本级（token-level）任务来说，文字大小，是否倾斜，是否加粗，以及字体等富文本格式能够体现相应的语义。通常来说，表单键值对的键位（key）通常会以加粗的形式给出。对于一般文档来说，文章的标题通常会放大加粗呈现，特殊概念名词会以斜体呈现等。对文档级（document-level）任务来说，整体的文档图像能提供全局的结构信息。例如个人简历的整体文档结构与科学文献的文档结构是有明显的视觉差异的。这些模态对齐的富文本格式所展现的视觉特征可以通过视觉模型抽取，结合到预训练阶段，从而有效地帮助下游任务。

为了利用上述信息，我们提出了通用文档预训练模型LayoutLM (Xu et al., 2020)，如图4所示。在现有的预训练模型基础上添加2-D Position Embedding和Image Embedding两种新的Embedding 层，这样一来可以有效地结合文档结构和视觉信息。具体来讲，根据OCR获得的文本Bounding Box，我们能获取文本在文档中的具体位置。将对应坐标转化为虚拟坐

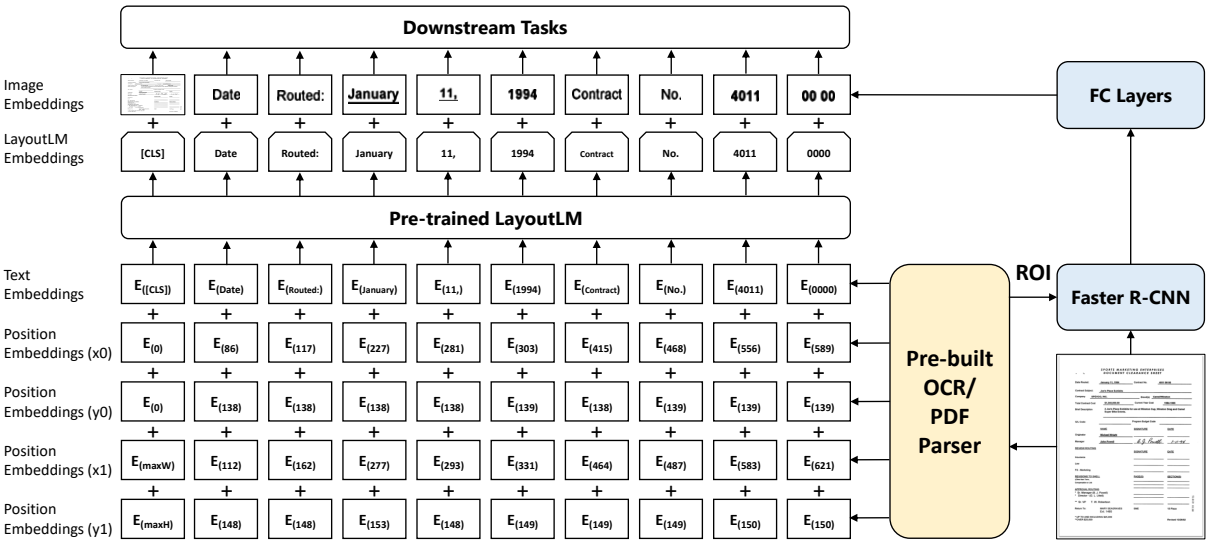


Figure 4: 基于Transformer架构的通用文档理解预训练模型LayoutLM

标之后，我们计算该坐标对应在x、y、w、h四个Embedding子层的表示，最终的2-D Position Embedding为四个子层的Embedding之和。在Image Embedding部分，我们将每个文本相应的Bounding Box当作Faster R-CNN中的候选框（Proposal），从而提取对应的局部特征。特殊地，由于[CLS]符号用于表示整个输入文本的语义，我们同样使用整张文档图像作为该位置的Image Embedding，从而保持模态对齐。

在预训练阶段，我们针对LayoutLM的特点提出两个自监督预训练任务：

1. Masked Visual-Language Model（MVLM，遮罩式视觉语言模型）：大量实验已经证明MLM能够在预训练阶段有效地进行自监督学习。我们在此基础上进行了修改：在遮盖（Mask）当前词之后，保留对应的2-D Position Embedding暗示，让模型预测对应的词。在这种方法下，模型根据已有的上下文和对应的视觉暗示预测被遮罩的词，从而让模型更好地学习文本位置和文本语义的模态对齐关系。

2. Multi-label Document Classification (MDC，多标签文档分类)：MLM能够有效的表示词级别的信息，但是对于文档级的表示，我们需要文档级的预训练任务来引入更高层的语义信息。在预训练阶段我们使用的IIT-CDIP数据集为每个文档提供了多标签的文档类型标注，我们引入MDC多标签文档分类任务。该任务使得模型可以利用这些监督信号去聚合相应的文档类别，并捕捉文档类型信息，从而获得更有效的高层语义表示。

实验结果表明，我们在预训练中引入的结构和视觉信息，能够有效地迁移到下游任务中。最终在多个下游任务中都取得了显著的准确率提升。与传统的基于卷积神经网络和图神经网络模型不同，通用文档智能预训练模型的优势在于可以支持不同类型的下游应用。

2.4 文档智能主流任务和数据集

文档智能涉及了自动阅读、理解和分析文档的相关技术，在实际场景的应用中主要包括四大类任务，分别是：

- **文档版面分析**：是指对文档版面内的图像、文本、表格信息和位置关系所进行的自动分析、识别和理解的过程。

- **文档信息抽取**：是指从文档中大量非结构化内容抽取实体及其关系的技术，与传统的纯文本信息抽取不同，文档的构建使得文字由一维的顺序排列变为二维的空间排列，因此文本信息、视觉信息和位置信息在文档信息抽取中都是极为重要的影响因素。

- **文档视觉问答**：是指给定文档图像数据，利用OCR技术或其他文字提取技术自动识别影像资料后，通过判断所识别文字的内在逻辑，回答关于图片的自然语言问题。

任务	数据集	支持语言	论文/链接
文档版面分析	ICDAR 2013	英文	(Göbel et al., 2013)
	ICDAR 2019	英文	(Gao et al., 2019)
	ICDAR 2021	英文	(Yepes et al., 2021)
	UNLV	英文	(Shahab et al., 2010)
	Marmot	中文/英文	(Fang et al., 2012)
	PubTabNet	英文	(Zhong et al., 2019a)
	PubLayNet	英文	(Zhong et al., 2019b)
	TableBank	英文	(Li et al., 2020a)
	DocBank	英文	(Li et al., 2020b)
	TNCR	英文	(Abdallah et al., 2021)
	TabLeX	英文	(Desai et al., 2021)
	IIIT-AR-13K	英文	(Mondal et al., 2020)
	ReadingBank	英文	https://aka.ms/readingbank
文档信息抽取	FUNSD	英文	(Guillaume Jaume, 2019)
	SROIE	英文	(Huang et al., 2019)
	CORD	英文	(Park et al., 2019)
	EATEN	中文	(Guo et al., 2019)
	EPHOIE	中文	(Wang et al., 2021)
	Deepform	英文	(Stray and Svetlichnaya, 2020)
	Kleister	英文	(Stanisławek et al., 2021)
	XFUND	中文/日文/西班牙文/法文/意大利文/德文/葡萄牙文	(Xu et al., 2021b)
文档视觉问答	DocVQA	英文	(Mathew et al., 2021b)
	InfographicsVQA	英文	(Mathew et al., 2021a)
	VisualMRC	英文	(Tanaka et al., 2021)
	保险文本视觉问答	中文	https://bit.ly/3602Vow
文档图像分类	Tobacco-3482	英文	(Kumar et al., 2014)
	RVL-CDIP	英文	(Harley et al., 2015)

Table 1: 文档智能领域主流任务（文档版面分析、文档信息抽取、文档视觉问答、文档图像分类）开源数据集

- **文档图像分类**：是指针对文档图像进行分析识别从而归类的过程。

对于这四种主要的文档智能任务，学术界和工业界也开源了大量相关的基准数据集，如表1所示。这也极大地推动了相关领域的研究人员构建新的算法模型，特别是当前基于深度神经网络的模型在这些任务上都有不俗的表现。接下来，我们将分别详细介绍在过去不同时期的经典模型和算法，包括基于启发式规则的文档分析技术、基于统计机器学习的文档分析技术和基于深度学习的通用文档智能模型，为大家提供参考。

3 基于启发式规则的文档分析技术

采用启发式规则的文档分析技术大致可分为自顶向下、自底向上和混合模式三种方式。自顶向下方式将文档图片作为整体逐步将其划分为不同区域。以递归方式进行切割，直至区域分割至预定义的标准，通常为块或列。自底向上以像素或组件为基本元素单位，对基本元素进行分组、合并以形成更大的同质区域。自顶向下方式在特定格式下的文档中能够更快、更高效地分析文档。而自底向上虽需要耗费更多的计算时间，但通用型更强，可覆盖更多不同布局类型的文档。混合方式则将其两者相结合以尝试产生更好的效果。

本节从自顶向下和自底向上两种角度出发，介绍基于Projection Profile、Image Smearing、Connected Components等方式的文档分析技术。

3.1 Projection Profile

Projection Profile作为一种自顶向下的分析方式被广泛应用于文档分析。(Nagy and Seth, 1984)使用Projection Profile中的X-Y切割算法对文档进行切割，这一方式适用于具有固定文本区域和行距的结构化文本，但该方式对边界噪声敏感且无法在倾斜的文本上提供良好性能，对文档质量要求较高。(Bar-Yosef et al., 2009)使用自适应局部投影方式计算文档的倾斜度，以尝试消除文本倾斜导致的性能下降，实验证明模型在倾斜和弯曲文本上得到了较为准确的结果。此外还有很多X-Y切割算法的变体被提出以解决现存的缺陷，(O'Gorman, 1993)将X-Y切割算法扩展至使用组件边界框的投影，(Sylwester and Seth, 1995)使用了编辑成本评估指标以指导模型进行分割，均在一定程度上提高了模型的性能。

Projection Profile分析算法适用于结构化文本，尤其是曼哈顿（Manhatten）布局文档。在布局复杂、文本倾斜或包含边界噪声的文档上可能无法展现出良好的性能。

3.2 Image Smearing

Image Smearing分析法指从一个位置向四周渗透，逐渐扩展至所有同质区域，以此确定页面当中的一个区域。(Wong et al., 1982)采用自顶向下策略，使用游长平滑算法（Run-length Smoothing Algorithm，RLSA）判断同质区域。将图像二值化后，像素值0表示背景，1为前景，当0周围的0数目小于指定阈值C时，该位置的0修改为1，游长平滑算法通过这一操作将距离相近的前景内容合并为整体。这种方式可以逐步将字符合并为单词，单词合并为文本行，继而将范围不断延伸至整个同质区域。(Fisher et al., 1990)在此基础上对其进行进一步改进，增加了除噪、倾斜矫正等预处理，此外游长平滑算法的阈值C修改为依据动态算法进行调整，进一步提升模型的适应能力。(Esposito et al., 1990)采用了类似的方法，但操作对象由像素改为了字符框。(Shi and Govindaraju, 2004)则是对图片中的每一个位置像素进行扩展，得到一个新的灰度图，随后进行抽取，在手写字体、文本倾斜等情况下仍能表现出良好的性能。

3.3 Connected Components

Connected Components分析法作为一种自底向上的技术，推测最小粒度元素之间的关系，用于寻找同质区域，最终将区域分类为不同属性。(Fisher et al., 1990)采用Connected Components技术，找到每个组件的K近邻（K Nearest Neighbors，KNN）组件，通过互相之间的位置、角度等关系来推断当前区域属性。(Saitoh et al., 1993)判断并根据文档的倾角将文字合并成线，继而将线合并为区域，随后将其分类为不同的属性。(Kise et al., 1998)同样尝试解决文本的倾斜问题，作者采用了近似面积Voronoi图（Approximated Area Voronoi Diagram）来获得区域的候选边界，这一操作对于任意倾角的区域有效。但由于计算过程中需要估计字符间距和行内间距，因此当文档中包含大字体及宽字间距等情况时，模型并不能发挥出良好性能。此外(Bukhari et al., 2010)也尝试在使用Connected Components的基础上使用AutoMLP以便寻找分类器最佳参数，进一步提升性能。

3.4 其他方法

除上文所述外，还有一些其他的启发式规则方法，例如，(Baird et al., 1990)采用自顶向下的方式按空白将文档进行切割划分区域。(Xiao and Yan, 2003)使用了Delaunay Triangulation算法进行文档分析，(Bukhari et al., 2009)在此基础上将其应用于书写随意的手写文档。此外还有一些混合算法，(Okamoto and Takahashi, 1993)通过分隔符和空白来切割块，在每个块中进一步将内部组件合并为文本行。(Smith, 2009)将文档分析分成了两部分，首先使用自底向上的方式来定位制表符，借助于制表符推断列布局。随后在列布局上采用自顶向下方式来推断结构和文本顺序。

4 基于统计机器学习的文档分析技术

传统的文档分析过程通常分为两阶段：1.将文档图片切割，得到多个不同候选区域。2.对区域进行属性分类，将其判别为文本、图像等规定类。基于机器学习的方案也通常从这两个角度入手，部分工作尝试使用机器学习算法参与文档的切割，其余则尝试在已生成的区域上构造特征使用机器学习算法对区域进行分类。此外由于统计机器学习技术带来的性能上的提升，较多基于统计机器学习的方法在表格检测任务中被尝试使用，因表格检测作为文档分析的一个重要

子任务，本节也会对其进行一些介绍。因此与前文基于技术角度的阐述方式不同的是，从下文开始将会从文档分析中的不同任务角度来对其发展情况做出介绍。

4.1 文档切割

在文档的切割过程中，(Baechler and Ingold, 2011)结合x-y裁剪算法，使用逻辑斯蒂回归对文档进行切割，丢弃空白部分。在得到相应区域后，实验比较了K近邻、逻辑斯蒂回归（Logistic Regression，LR）和最大熵马尔可夫模型（Maximum Entropy Markov Models，MEMM）等算法作为分类器的性能优劣，实验表明最大熵马尔可夫模型和逻辑斯蒂回归在属性分类任务上可以展现出较好的性能。(Esposito et al., 2008)在文档分割过程中进一步加强机器学习算法在其中的参与程度。在自底向上的过程中，从字母到单词到文本行逐渐合并的过程中使用了一种基于内核的算法 (Dietterich et al., 1997)，并将结果转换成xml结构存储。之后使用文档组织算法（Document Organization Composer，DOC）对文档进行分析。(Wu et al., 2008)则致力于文字同时存在两种阅读顺序的问题，此前的算法均假定文字只有一种书写方向，但遇到诸如汉语或日语等可以水平或者垂直方向书写的文字时无法正常地工作。该算法将文档分割分为四个步骤用于判断并处理文本，并使用了支持向量机以决定是否执行步骤。

4.2 区域分类

在区域属性分类问题上，大量工作主要致力于尝试不同机器学习算法作为分类器输出结果。其中(Wei et al., 2013)实验比较了支持向量机、多层感知机（Multi-Layer Perceptron，MLP）和高斯混合模型（Gaussian Mixture Models，GMM）几种机器学习算法作为分类器时的性能优劣，实验结果表明支持向量机和多层感知机在区域属性上的分类性能明显优于高斯混合模型。(Bukhari et al., 2012)手动构造了多个特征，对区域抽取相应特征后使用AutoMLP算法进行分类，在阿拉伯语数据集中得到了95%的分割准确率。(Baechler and Ingold, 2011)在文档分割上做了进一步改进，使用了金字塔形算法，在中世纪手稿上进行了三个不同级别的分析，最后使用动态多层感知机（Dynamic Multi-Layer Perceptron，DMLP）作为分类器。

4.3 表格检测

除上述方式之外，基于统计机器学习技术在表格识别领域存在大量研究。(Wang et al., 2000; Wangt et al., 2001; Wang et al., 2002)使用了二叉树对文档进行自上而下分析查找表格候选区，继而根据区域特征确定最终表格区域。(Pinto et al., 2003)则使用了条件随机场在HTML页面中抽取表格区域，并确定表格中的标题、子标题等内容。(e Silva, 2009)使用隐马尔可夫（Hidden Markov Models，HMMs）抽取表格区域。(Chen and Lopresti, 2011)在手写文档中检索表格区域，并使用支持向量机识别其中的文字区域，随后依据文本行确定表格所在位置。(Kasar et al., 2013)同样使用了支持向量机技术。首先识别图中水平和竖直的垂直线，随后使用支持向量机对每条线的属性进行分类，判断该线条是否属于表格。(Barlas et al., 2014)使用多层感知机对文档中的connected component进行分类，判断其是否为文本。(Bansal et al., 2014)使用leptonica库 (Bloomberg, 1991)对文档进行分割，随后对每一个区域构造包含周围环境信息的特征。使用Fix-point model (Li et al., 2013)对每一个区域进行分类，用以识别文档中的表格区域。它使得模型在分类过程中不再孤立地对其进行分类，而是学习区域相互之间的关系。(Rashid et al., 2017)采用了与前一份工作相同的思路，但将操作粒度缩小为单词级别，对每一个词进行分类，之后使用AutoMLP来判断该词是否属于表格。

5 基于深度学习的文档智能技术

近年来，深度学习方法已经成为许多机器学习问题的解决范式。在许多研究领域，深度学习方法被证明是十分有效的。最近，预训练模型的流行也进一步发掘了深度神经网络的性能。而文档智能领域的发展也体现出同样的趋势。在本节中我们将现存的模型分为针对特定任务的深度学习模型和支持多种下游任务的通用预训练模型两个章节进行介绍。

5.1 针对特定任务的深度学习模型

5.1.1 文档版面分析

文档版面分析包含两个主要的子任务：文档视觉结构分析和文档语义结构分析 (Binmakhashen and Mahmoud, 2019)。文档视觉分析的主要目的是检测文档结构并确定其同类区域的边界。而文档语义结构分析是需要为这些检测到的区域标记具体的文档类别，如标题、段落、表格等。PubLayNet (Zhong et al., 2019b)是一个大规模的文档版面分析数据集，通过自动解析PubMed的XML文件构建了超过360,000个文档图片。DocBank (Li et al., 2020b)通过arXiv网站的PDF文件和LaTeX文件的对应关系自动构建了一个可扩展的文档版面分析数据集，同时支持对基于文本的方法和基于图像的方法进行评测。IIIT-AR-13K (Mondal et al., 2020)提供了13,000的人工标注的文档图片用于版面分析。

在章节2.2中，我们介绍了将较为经典的卷积神经网络应用在文档版面分析领域的工作 (He et al., 2015; Ren et al., 2016; He et al., 2018; Liu et al., 2016; Redmon and Farhadi, 2018; Yang et al., 2017a; Schreiber et al., 2017)，但随着对文档版面分析的性能要求逐渐提高，越来越多的科研工作针对文档这一领域对目标检测算法进行了针对性的改进。(Yang et al., 2017b)将文档语义结构分析任务视为一个逐像素的分类问题。他们提出了一个同时考虑视觉和文本信息的多模态神经网络。(Viana and Oliveira, 2017)提出了一个用于移动和云服务的文档布局分析的轻量级模型。该模型使用图像的一维信息进行推理，并与使用二维信息的模型进行比较，在实验中取得了较高的准确性。(Chen et al., 2017)介绍了一种基于卷积神经网络（CNN）的手写历史文件图像的页面分割方法。(Oliveira et al., 2018)提出了一个基于CNN的多任务逐像素预测模型。(Wick and Puppe, 2018)提出了一个用于历史文件分割的高性能全卷积神经网络（FCN）。(Grüning et al., 2019)提出了一种针对历史文献的两阶段文本行检测方法。(Soto and Yoo, 2019)将上下文信息纳入Faster R-CNN模型。该模型利用文章元素内容的局部不变性质，提高了区域检测性能。

表格检测与表格结构识别 在文档版面分析中，表格理解是一项富有挑战性的任务。有别于标题、段落等文档元素，表格的格式通常较为多变，结构也较为复杂。因此，有大量的相关工作围绕表格进行展开，其中最为主要的两个子任务分别是表格检测和表格结构识别。（1）表格检测是指确定文档中的表格的边界。（2）表格结构识别是指将表格的语义结构，包括行、列、单元格的信息按照预定义的格式抽取出来。

近年来，有许多针对表格理解这一任务提出的数据集。Marmot (Fang et al., 2012)和UNLV (Shahab et al., 2010)是较早的表格识别数据集。ICDAR会议在表格检测与识别上举办的多次竞赛提供了优质的表格数据集 (Göbel et al., 2013; Gao et al., 2019)。但这些传统表格数据集通常较小，难以发挥现代深度神经网络的优势，因此研究工作TableBank (Li et al., 2020a)利用LaTex和Office Word来自动构建了一个大规模的表格理解数据集。此后，PubTabNet (Zhong et al., 2019a)提出了一个大规模表格数据集并提供了表格结构及单元格内容辅助表格识别。TNCR (Abdallah et al., 2021)在提供表格标注的同时提供了表格类别的标注。

针对表格理解这一任务的特性，许多目标检测的方法在表格理解领域都能取得较好的效果。Faster R-CNN (Ren et al., 2016)在表格检测任务上直接应用就能达到非常好的性能。在此基础上，(Siddiqui et al., 2018)通过将可变形卷积应用在Faster R-CNN上获得了更好的性能。CascadeTabNet (Prasad et al., 2020)使用了Cascade R-CNN (Cai and Vasconcelos, 2018)模型同时完成表格检测和表格结构识别。TableSense (Dong et al., 2019) 通过增加单元格特征、添加采样算法来显著提高了表格检测能力。

除了上述两个主要的子任务，针对已解析后表格的理解也逐渐成为新的挑战。TAPAS (Herzig et al., 2020)是较早的将预训练技术引入到表格理解任务的模型。通过引入额外的位置编码层，TAPAS可以使Transformer (Vaswani et al., 2017)编码器接受结构化的表格输入。经过在大量的表格数据上进行掩码式预训练后，TAPAS在多种下游语义分析任务中显著超过了传统方法。继TAPAS后，TUTA (Wang et al., 2020a) 引入了二维坐标树来表示结构化表格的层级信息，并针对这一结构提出了基于树结构的位置表示方式和注意力机制来显示建模层次化表格。结合不同层级的预训练任务，TUTA在多个下游数据集上取得了进一步的性能提升。

5.1.2 文档信息抽取

文档信息抽取是指从大量非结构化富文本文档内容中抽取语义实体及其之间关系的技术。文档信息抽取任务对于文档类别的不同，抽取的目标实体也不尽相同。FUNSD (Guillaume Jaume, 2019)是一个文档理解数据集，其包含199张表单，每张表单中包含表单实体的键值对。CORD (Park et al., 2019)是一个票据理解数据集，并包含8个大类共54小类种实体标签。Kleister (Stanisławek et al., 2021)是一个针对长文档实体抽取任务的文档理解数据集，包含有协议和财务报表等长文本文档。DeepForm(Stray and Svetlichnaya, 2020)数据集是一个针对电视和有线电视政治广告披露表格的英文数据集。EATEN数据集是针对中文证件的信息抽取数据集，(Yu et al., 2021)在其400张子集上进一步添加了文本框标注。EPHOIE (Wang et al., 2021)数据集是一个针对中文文档数据的信息抽取数据集。XFUND (Xu et al., 2021b)是随着LayoutXLM模型提出了针对FUNSD数据集的多语言扩展版本，包含有除英文以外的七种主流语言的富文本文档。

由于富文本文档的丰富视觉信息，很多研究工作将文档信息抽取任务建模为了计算机视觉任务，通过语义分割或文本框回归等任务进行信息抽取。考虑到文档信息抽取中文本信息同样具有重要作用，通常的框架是将文档图片视为像素网格，并在该特征图上添加文本特征来获得更好的特征表示。根据添加文本特征级别的不同，这一方法的基本发展顺序呈现出了从字符级别到单词级别再到上下文级别的趋势。Chargrid(Katti et al., 2018)利用一个基于卷积的编码器-解码器网络，通过将字符进行Onehot编码来将文本信息融合到图像中。VisualWordGrid(Kerroumi et al., 2020)实现了Wordgrid(Katti et al., 2018)，通过将字符级文本信息换成单词级的word2vec特征，并融合了一定的视觉信息，提高了抽取任务的性能。BERTgrid (Denk and Reisswig, 2019)通过使用BERT获得了上下文文本表示，进一步提升了性能。ViBERTgrid (Lin et al., 2021)在BERTgrid的基础上将BERT的文本特征较早地在卷积阶段与图像特征进行融合，从而获得了较好的效果。

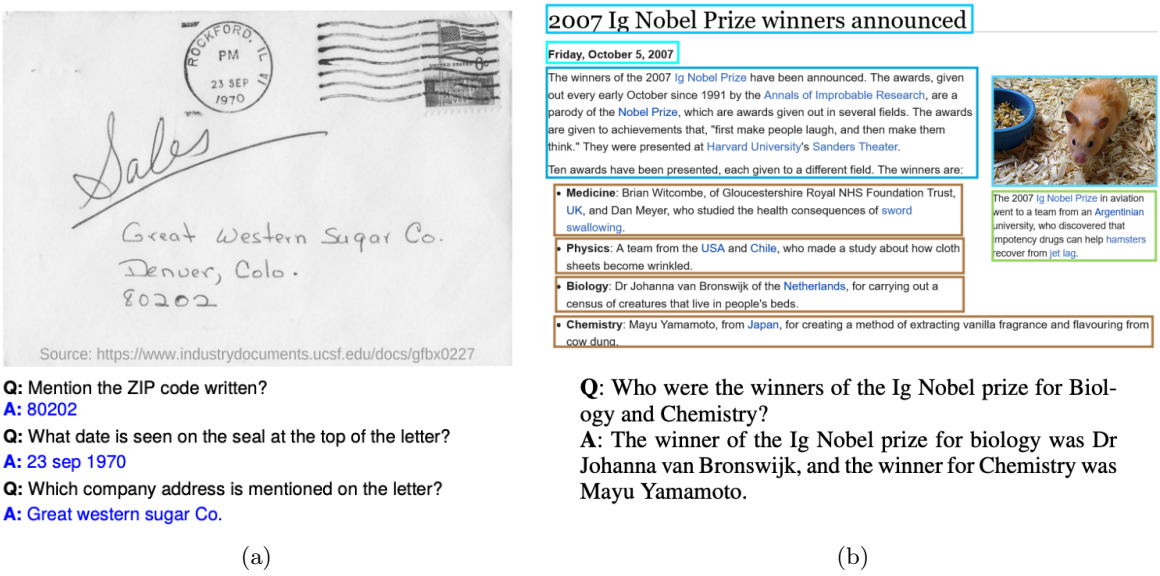
由于富文本文档中的信息仍以文本作为主体，很多研究工作将文档信息抽取任务作为特殊的自然语言理解任务。(Majumder et al., 2020)通过根据抽取目标的类别来生成目标备选，在表单任务上取得了较好的效果。TRIE (Zhang et al., 2020)联合文本检测识别与信息抽取，让两个阶段的任务互相促进从而获得更好的信息抽取效果。(Wang et al., 2020b)通过三种不同模态信息的融合来预测文本片段之间的关系，实现了对表单的层次化抽取。

非结构化的富文本文档由多个邻接的文本片段组成，那么自然可以使用图网络对非结构化富文本文档进行表示。文档中的文本片段建模为图中的节点，而文本片段之间的关系则可建模为边，这样整个文档就可以被表示为一个图网络。在章节2.2中，我们介绍了图神经网络在富文本文档中进行信息抽取的代表性工作 (Liu et al., 2019)。在此基础上，逐渐有更多的研究工作基于图神经网络展开。(Hwang et al., 2020)将文档建模为了有向图，通过依存分析的方法对文档进行信息抽取。(Riba et al., 2019)使用基于图神经网络的模型来进行发票中表格的信息抽取。(Wei et al., 2020)通过在预训练模型的输出表示上使用图卷积神经网络来建模文本布局，提高了信息抽取的性能。(Cheng et al., 2020)通过将文档表示为图结构并使用基于图的注意力机制，结合CRF在小样本学习上取得了较好的性能。PICK (Yu et al., 2021)模型通过引入一个可基于节点进行学习的图来表示文档，在发票抽取任务中取得了较好的性能。

5.1.3 文档图像分类

文档图像分类是指对文档图像进行归类标记的任务。RVL-CDIP (Harley et al., 2015)是该任务中的代表性数据集。该数据集包含16个文档图像类别共400,000张灰度图片。

由于文档图像分类仍然属于图像分类的范畴，所以针对自然图片的分类算法同样能较好的解决文档图像分类的问题。(Afzal et al., 2015)介绍了一种基于深度卷积神经网络（CNN）的文档图像分类方法用于文档图像分类。为了克服小数据集样本不足的问题，他们使用了经过Imagenet训练的Alexnet网络作为初始化，从而迁移到文档图像领域。(Afzal et al., 2017)尝试将GoogLeNet, VGG, ResNet等在自然图片领域获得成功的模型通过迁移学习的方式在文档图片上进行训练。(Tensmeyer and Martinez, 2017)通过对模型参数和数据处理的调整，使CNN模型不借助从自然图片的迁移学习就能优于此前模型的性能。(Das et al., 2018)提出了一个基于不同区域分类的深度卷积神经网络框架用于文档图像分类。该方法通过对文档的不同区域分别进行分类，最终融合多个不同区域的分类器在文档图像分类上获得了明显的性能提升。(Sarkhel and Nandi, 2019)通过引入了金字塔形的多尺度结构来抽取不同层级的特征。(Dauphinee et al.,



(a) (b)

Figure 5: 文档视觉问答任务示例

2019)通过对文档图片进行字符识别（OCR）获得文档的文本，并对图像特征和文本特征进行组合，进一步提升了分类性能。

5.1.4 文档视觉问答

文档视觉问答是一个针对文档图片的高层理解任务。具体来说，给定一张文档图片和一个针对性的问题，模型需要根据图片给出该问题的正确答案。具体的例子如图5所示。针对文档的视觉问答工作最早出现在数据集DocVQA (Mathew et al., 2021b)中，该数据集包含了超过12000个文档和对应的5000个问题。后来，出现了针对文档中图表的视觉问答工作InfographicVQA (Mathew et al., 2021a)。针对DocVQA数据集的答案较短，文档主题较单一的缺陷，有研究人员提出了VisualMRC (Tanaka et al., 2021)数据集。

不同于传统VQA任务，文档视觉问答中的文档文本对任务具有关键作用，所以现存的代表性方法都将文档图片进行字符识别（OCR）处理得到的文档文本作为重要的信息。在得到文档文本后，针对不同数据的特点，视觉问答任务被建模为不同的问题。对于DocVQA数据来说，绝大部分的问题答案都是作为文本片段存在于文档文本中的，所以主流的方法都将其建模为了机器阅读理解问题（Machine Reading Comprehension）。通过为模型提供视觉特征和文档文本，模型根据问题在给定的文档文本上进行文本片段的抽取来作为问题答案。而对于VisualMRC数据集，问题的答案通常不蕴含在文档文本片段中，需要给出较长的抽象回答。因此，在这种情况下，可行的方法是使用文本生成式的方法生成问题的答案。

5.2 支持多种下游任务的通用预训练模型

以上针对特定任务的深度学习方法针对某一项文档理解任务上能够取得较好的性能，然而这些方法主要面临两个限制：(1) 这些模型通常依赖于有限的标记数据，而忽视了挖掘大量无标注数据中的知识。对于文档理解任务尤其是其中的信息抽取任务来说，详细标注的数据是昂贵且消耗时间的。另一方面，由于富文本文档在现实生活的大量使用，存在着大量的未标注文档，而这些大量的未标注数据可以使用自监督预训练加以利用。(2) 富文本文档不仅有大量的文本信息，同时也包含丰富的版面和视觉信息。已有的针对特定任务的模型由于数据量的限制，通常只能通过预训练的CV模型或NLP模型来获取对应模态的特征，而且大部分工作只利用了单一模态的信息或者是两种特征的简单组合而不是深度交互。Transformer(Vaswani et al., 2017)在迁移学习领域的成功证明了深度上下文化（Contextualizng）对于序列建模的重要性，因此将文本和其他模态进行深度交互融合是一个较为明显的趋势。

富文本文档主要包含三种模态信息：文本、布局以及视觉信息，并且这三种模态在富文本文档中有天然的对齐特性。因此，如何对文档进行建模并且通过训练达到跨模态对齐是一

个重要的问题。LayoutLM (Xu et al., 2020)以及后续提出的LayoutLMv2 (Xu et al., 2021a)模型的提出正是针对这一方向进行的研究工作。在章节2.3中，我们详细介绍了LayoutLM这一通用文档理解预训练模型，通过将文本和布局进行联合预训练，LayoutLM在多种文档理解任务上取得了显著提升。在此基础上，又有许多后续的研究工作对这一框架进行了针对性的改进。LayoutLM在预训练过程中没有引入文档视觉信息，从而在DocVQA这类需要较强视觉感知能力的任务上效果欠佳。针对这一问题，LayoutLMv2 (Xu et al., 2021a)通过将视觉特征信息融入到预训练过程中，大大提高了模型的图像理解能力。具体来说，在结构方面，LayoutLMv2引入了空间感知自注意力机制(spatial-aware self-attention)，并将视觉特征作为输入序列的一部分。在预训练目标方面，LayoutLMv2在掩码视觉语言模型（Masked Visual-Language Model）之外又提出了文本—图像对齐（Text-Image Alignment）和文本—图像匹配（Text-Image Match）任务。通过在这两方面的改进，模型对于视觉信息的感知能力大大提高，并在包括DocVQA在内的六种下游任务中获得了显著提升。

LayoutLM模型虽然在英文数据上取得了成功，但是对于非英语世界来说文档理解任务同样重要，而LayoutXLM (Xu et al., 2021b)的提出解决了这一问题。LayoutXLM基于LayoutLMv2的模型结构，通过使用53种语言进行预训练，扩展了LayoutLM的语言支持。与此同时，相比于纯文本的跨语言模型，LayoutXLM在迁移能力上有明显的优势，这证明了不仅多语言文本之间可以进行跨语言学习，多语言富文本文档之间的还可以进行文档布局的迁移学习。

LayoutLM提出之后，许多研究工作针对这一框架进行了针对性的改进。LAMBERT (Garncarek et al., 2020) 通过使用RoBERTa作为预训练初始化获得了更好的性能。BROS (Hong et al., 2020)在引入区域掩码训练的同时在编码器阶段加入了文本空间位置信息，提高了模型对空间位置感知能力。(Li et al., 2021a)通过文本块内共享相同的位置信息并在预训练阶段引入位置信息预测的方式，也让模型具有一定的位置感知能力。LAMPRET (Wu et al., 2021)通过为模型提供更多的模态信息如字体字号、插图等，对网页文档进行建模，并结合多种层次化的预训练任务来增强模型对文本和图片的理解能力。SelfDoc (Li et al., 2021b)通过在输入阶段使用文档实体目标作为输入，结合模态适应的注意力机制，提升了模型的模态交互能力。DocFormer (Appalaraju et al., 2021)通过引入了更高清的图片输入以及图像重构的预训练任务，更加充分地利用了图像信息，从而提高了模型性能。除了语言理解之外，很多模型着眼于扩展模型的语言生成能力。一个共同的特点是都是用了Encoder-Decoder范式。TILT (Powalski et al., 2021)通过将Layout编码层引入T5模型并结合文档数据预训练，使模型能够处理文档领域的生成任务。LayoutT5和LayoutBART (Tanaka et al., 2021)在文档视觉问答任务微调阶段在T5和BART模型的基础上引入文本位置编码，来帮助模型理解并生成问题答案。

6 结语

信息处理是数字化转型的基础和前提，如今对处理能力、处理速度和处理精度也都有着越来越高的要求。以商业领域为例，电子商业文档就涵盖了采购单据、行业报告、商务邮件、销售合同、雇佣协议、商业发票、个人简历等大量繁杂的信息。机器人流程自动化（Robotic Process Automation，RPA）行业正是在这一背景下应运而生，利用人工智能技术帮助大量人工从繁杂的电子文档处理任务中解脱出来，并通过一系列配套的自动化工具提升生产力，RPA的关键核心之一就是文档智能分析技术。在过去的20年间，文档智能分析技术主要经历了三个阶段，从最初的基于启发式规则，过渡到基于统计机器学习，到近来基于深度学习的方法，极大地提升了分析性能和准确率。与此同时我们也观察到，以LayoutLM为代表的大规模自监督通用文档智能预训练模型也越来越多地受到人们的关注和使用，逐步成为构建更为复杂算法的基本单元，后续研究工作也层出不穷，促使文档智能领域加速发展。

展望未来，除了解决文档多页跨页、训练数据质量参差不齐、多任务关联性较弱以及少样本零样本学习等问题，还应该特别关注文字检测识别OCR技术与文档智能技术的结合，因为文档智能下游任务的输入通常来自于自动文字检测和识别算法，文字识别的准确性往往对于下游任务有很大的影响。此外，如何将文档智能技术与现有人类知识以及人工处理文档的技巧相结合，也是未来值得探索的一个研究课题。

参考文献

Abdelrahman Abdallah, Alexander Berendeyev, Islam Nuradin, and Daniyar Nurseitov. 2021. Tncr: Table net detection and classification dataset. *arXiv preprint arXiv:2106.15322*.

Muhammad Zeshan Afzal, Samuele Capobianco, Muhammad Imran Malik, Simone Marinai, Thomas M Breuel, Andreas Dengel, and Marcus Liwicki. 2015. Deepdocclassifier: Document classification with deep convolutional neural network. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1111–1115. IEEE.

Muhammad Zeshan Afzal, Andreas Kölsch, Sheraz Ahmed, and Marcus Liwicki. 2017. Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 883–888. IEEE.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. *arXiv preprint arXiv:2106.11539*.

Micheal Baechler and Rolf Ingold. 2011. Multi resolution layout analysis of medieval manuscripts using dynamic mlp. In *2011 International Conference on Document Analysis and Recognition*, pages 1185–1189. IEEE.

Henry S Baird, Susan E Jones, and Steven J Fortune. 1990. Image segmentation by shape-directed covers. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume 1, pages 820–825. IEEE.

Anukriti Bansal, Gaurav Harit, and Sumantra Dutta Roy. 2014. Table extraction from document images using fixed point model. In *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, pages 1–8.

Itay Bar-Yosef, Nate Hagbi, Klara Kedem, and Itshak Dinstein. 2009. Line segmentation for degraded handwritten historical documents. In *2009 10th International Conference on Document Analysis and Recognition*, pages 1161–1165. IEEE.

Philippine Barlas, Sébastien Adam, Clément Chatelain, and Thierry Paquet. 2014. A typed and handwritten text block segmentation system for heterogeneous and complex documents. In *2014 11th IAPR International Workshop on Document Analysis Systems*, pages 46–50. IEEE.

Galal M Binmakhashen and Sabri A Mahmoud. 2019. Document layout analysis: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 52(6):1–36.

Dan S Bloomberg. 1991. Multiresolution morphological approach to document image analysis. In *Proc. of the international conference on document analysis and recognition, Saint-Malo, France*.

Syed Saqib Bukhari, Faisal Shafait, and Thomas M Breuel. 2009. Script-independent handwritten textlines segmentation using active contours. In *2009 10th International Conference on Document Analysis and Recognition*, pages 446–450. IEEE.

Syed Saqib Bukhari, Mayce Ibrahim Ali Al Azawi, Faisal Shafait, and Thomas M Breuel. 2010. Document image segmentation using discriminative learning over connected components. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 183–190.

Syed Saqib Bukhari, Thomas M Breuel, Abedelkadir Asi, and Jihad El-Sana. 2012. Layout analysis for arabic historical document images using machine learning. In *2012 International Conference on Frontiers in Handwriting Recognition*, pages 639–644. IEEE.

Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162.

Jin Chen and Daniel Lopresti. 2011. Table detection in noisy off-line handwritten documents. In *2011 International Conference on Document Analysis and Recognition*, pages 399–403. IEEE.

Kai Chen, Mathias Seuret, Jean Hennebert, and Rolf Ingold. 2017. Convolutional neural networks for page segmentation of historical document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 965–970. IEEE.

Mengli Cheng, Minghui Qiu, Xing Shi, Jun Huang, and Wei Lin. 2020. One-shot text field labeling using attention and belief propagation for structure information extraction. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 340–348.

Arindam Das, Saikat Roy, Ujjwal Bhattacharya, and Swapan K Parui. 2018. Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3180–3185. IEEE.

Tyler Dauphinee, Nikunj Patel, and Mohammad Rashidi. 2019. Modular multimodal architecture for document classification. *arXiv preprint arXiv:1912.04376*.

Timo I Denk and Christian Reisswig. 2019. Bertgrid: Contextualized embedding for 2d document representation and understanding. *arXiv preprint arXiv:1909.04948*.

Harsh Desai, Pratik Kayal, and Mayank Singh. 2021. Tablex: A benchmark dataset for structure and content information extraction from scientific tables.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71.

Haoyu Dong, Shijie Liu, Shi Han, Zhouyu Fu, and Dongmei Zhang. 2019. Tablesense: Spreadsheet table detection with convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 69–76.

Ana Costa e Silva. 2009. Learning rich hidden markov models in document analysis: Table location. In *2009 10th International Conference on Document Analysis and Recognition*, pages 843–847. IEEE.

Floriana Esposito, Donato Malerba, Giovanni Semeraro, Enrico Annese, and Giovanna Scafuro. 1990. An experimental page layout recognition system for office document automatic classification: an integrated approach for inductive generalization. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume 1, pages 557–562. IEEE.

Floriana Esposito, Stefano Ferilli, Teresa MA Basile, and Nicola Di Mauro. 2008. Machine learning for digital document processing: From layout analysis to metadata extraction. In *Machine learning in document analysis and recognition*, pages 105–138. Springer.

Jing Fang, Xin Tao, Zhi Tang, Ruiheng Qiu, and Ying Liu. 2012. Dataset, ground-truth and performance metrics for table detection evaluation. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 445–449. IEEE.

James L Fisher, Stuart C Hinds, and Donald P D'Amato. 1990. A rule-based system for document image segmentation. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume 1, pages 567–572. IEEE.

Liangcai Gao, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Lang. 2019. Icdar 2019 competition on table detection and recognition (ctdar). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1510–1515.

Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. 2020. Lambert: Layout-aware (language) modeling for information extraction. *arXiv preprint arXiv:2002.08087*.

Max C. Göbel, Tamir Hassan, Ermelinda Oro, and G. Orsi. 2013. Icdar 2013 table competition. *2013 12th International Conference on Document Analysis and Recognition*, pages 1449–1453.

Tobias Grüning, Gundram Leifert, Tobias Strauß, Johannes Michael, and Roger Labahn. 2019. A two-stage method for text line detection in historical documents. *International Journal on Document Analysis and Recognition (IJDAR)*, 22(3):285–302.

Jean-Philippe Thiran Guillaume Jaume, Hazim Kemal Ekenel. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *Accepted to ICDAR-OST*.

He Guo, Xiameng Qin, Jiaming Liu, Junyu Han, Jingtuo Liu, and Errui Ding. 2019. Eaten: Entity-aware attention for single shot visual text extraction.

Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2018. Mask r-cnn.

Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349*.

Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2020. Bros: A pre-trained language model for understanding texts in document.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sep.

Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2020. Spatial dependency parsing for semi-structured document information extraction. *arXiv preprint arXiv:2005.00642*.

Thotreingam Kasar, Philippine Barlas, Sebastien Adam, Clément Chatelain, and Thierry Paquet. 2013. Learning to detect tables in scanned document images using line information. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1185–1189. IEEE.

Anoop R Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2D documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4459–4469, Brussels, Belgium, October-November. Association for Computational Linguistics.

Mohamed Kerroumi, Othmane Sayem, and Aymen Shabou. 2020. Visualwordgrid: Information extraction from scanned documents using a multimodal approach. *arXiv preprint arXiv:2010.02358*.

Koichi Kise, Akinori Sato, and Motoi Iwata. 1998. Segmentation of page images using the area voronoi diagram. *Computer Vision and Image Understanding*, 70(3):370–382.

J. Kumar, Peng Ye, and D. Doermann. 2014. Structural similarity for document image classification and retrieval. *Pattern Recognit. Lett.*, 43:119–126.

Quannan Li, Jingdong Wang, David Wipf, and Zhuowen Tu. 2013. Fixed-point model for structured labeling. In *International conference on machine learning*, pages 214–221. PMLR.

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2020a. TableBank: Table benchmark for image-based table detection and recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1918–1925, Marseille, France, May. European Language Resources Association.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020b. DocBank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 949–960, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021a. Structurallm: Structural pre-training for form understanding. *arXiv preprint arXiv:2105.11210*.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021b. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660.

Weihong Lin, Qifang Gao, Lei Sun, Zhuoyao Zhong, Kai Hu, Qin Ren, and Qiang Huo. 2021. Vibertgrid: A jointly trained multi-modal 2d document representation for key information extraction from documents. *arXiv preprint arXiv:2105.11672*.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, page 21–37.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph convolution for multimodal information extraction from visually rich documents. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 32–39, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork. 2020. Representation learning for information extraction from form-like documents. In *proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 6495–6504.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. 2021a. Infographicvqa.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021b. Docvqa: A dataset for vqa on document images.

Ajoy Mondal, Peter Lipps, and CV Jawahar. 2020. Iiit-ar-13k: a new dataset for graphical object detection in documents. In *International Workshop on Document Analysis Systems*, pages 216–230. Springer.

George Nagy and Sharad C Seth. 1984. Hierarchical representation of optically scanned documents.

Lawrence O'Gorman. 1993. The document spectrum for page layout analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 15(11):1162–1173.

Masayuki Okamoto and Makoto Takahashi. 1993. A hybrid page segmentation method. In *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93)*, pages 743–746. IEEE.

Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan. 2018. dhsegment: A generic deep-learning approach for document segmentation. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 7–12. IEEE.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: A consolidated receipt dataset for post-ocr parsing.

David Pinto, Andrew McCallum, Xing Wei, and W Bruce Croft. 2003. Table extraction using conditional random fields. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 235–242.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. *arXiv preprint arXiv:2102.09550*.

Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. 2020. Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 572–573.

Sheikh Faisal Rashid, Abdullah Akmal, Muhammad Adnan, Ali Adnan Aslam, and Andreas Dengel. 2017. Table recognition in heterogeneous documents using machine learning. In *2017 14th IAPR International conference on document analysis and recognition (ICDAR)*, volume 1, pages 777–782. IEEE.

Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Pau Riba, Anjan Dutta, Lutz Goldmann, Alicia Fornés, Oriol Ramos, and Josep Lladós. 2019. Table detection in invoice documents by graph neural networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 122–127. IEEE.

Takashi Saitoh, Michiyoshi Tachikawa, and Toshifumi Yamaai. 1993. Document image segmentation and text area ordering. In *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93)*, pages 323–329. IEEE.

Ritesh Sarkhel and Arnab Nandi. 2019. Deterministic routing between layout abstractions for multi-scale classification of visually rich documents. In *28th International Joint Conference on Artificial Intelligence (IJCAI), 2019*.

Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. 2017. Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1162–1167.

Asif Shahab, Faisal Shafait, Thomas Kieninger, and Andreas Dengel. 2010. An open approach towards the benchmarking of table structure recognition systems. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, DAS '10, page 113–120, New York, NY, USA. Association for Computing Machinery.

Zhixin Shi and Venu Govindaraju. 2004. Line separation for complex document images using fuzzy runlength. In *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.*, pages 306–312. IEEE.

Shoaib Ahmed Siddiqui, Muhammad Imran Malik, Stefan Agne, Andreas Dengel, and Sheraz Ahmed. 2018. Decnt: Deep deformable cnn for table detection. *IEEE Access*, 6:74151–74161.

Raymond W Smith. 2009. Hybrid page layout analysis via tab-stop detection. In *2009 10th International Conference on Document Analysis and Recognition*, pages 241–245. IEEE.

Carlos Soto and Shinjae Yoo. 2019. Visual detection with context for document layout analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3462–3468, Hong Kong, China, November. Association for Computational Linguistics.

Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. 2021. Kleister: Key information extraction datasets involving long documents with complex layouts.

Jonathan Stray and Stacey Svetlichnaya. 2020. Project deepform: Extract information from documents.

Don Sylwester and Sharad Seth. 1995. A trainable, single-pass algorithm for column segmentation. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pages 615–618. IEEE.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. *arXiv preprint arXiv:2101.11272*.

Chris Tensmeyer and Tony Martinez. 2017. Analysis of convolutional neural networks for document image classification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 388–393. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Matheus Palhares Viana and Dário Augusto Borges Oliveira. 2017. Fast cnn-based document layout analysis. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1173–1180.

Yalin Wang, Robert Haralick, and Ihsin T Phillips. 2000. Improvement of zone content classification by using background analysis. In *Fourth IAPR International Workshop on Document Analysis Systems.(DAS2000)*. Citeseer.

Yalin Wang, Ihsin T Phillips, and Robert M Haralick. 2002. Table detection via probability optimization. In *International Workshop on Document Analysis Systems*, pages 272–282. Springer.

Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Shi Han, and Dongmei Zhang. 2020a. Structure-aware pre-training for table understanding with tree-based transformers. *arXiv preprint arXiv:2010.12537*.

Zilong Wang, Mingjie Zhan, Xuebo Liu, and Ding Liang. 2020b. Docstruct: A multimodal method to extract hierarchy structure in document for general form understanding. *arXiv preprint arXiv:2010.11685*.

Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. 2021. Towards robust visual information extraction in real world: New dataset and novel solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2738–2745.

Yalin Wangt, Ihsin T Phillipst, and Robert Haralick. 2001. Automatic table ground truth generation and a background-analysis-based table structure extraction method. In *Proceedings of Sixth International Conference on Document Analysis and Recognition*, pages 528–532. IEEE.

Hao Wei, Micheal Baechler, Fouad Slimane, and Rolf Ingold. 2013. Evaluation of svm, mlp and gmm classifiers for layout analysis of historical documents. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1220–1224. IEEE.

Mengxi Wei, Yifan He, and Qiong Zhang. 2020. Robust layout-aware ie for visually rich documents with pre-trained language models. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2367–2376.

Christoph Wick and Frank Puppe. 2018. Fully convolutional neural networks for page segmentation of historical document images. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 287–292. IEEE.

Kwan Y. Wong, Richard G. Casey, and Friedrich M. Wahl. 1982. Document analysis system. *IBM journal of research and development*, 26(6):647–656.

Chung-Chih Wu, Chien-Hsing Chou, and Fu Chang. 2008. A machine-learning approach for analyzing document layout structures with two reading orders. *Pattern recognition*, 41(10):3200–3213.

Te-Lin Wu, Cheng Li, Mingyang Zhang, Tao Chen, Spurthi Amba Hombaiah, and Michael Bendersky. 2021. Lampret: Layout-aware multimodal pretraining for document understanding. *arXiv preprint arXiv:2104.08405*.

Yi Xiao and Hong Yan. 2003. Text region extraction in a document image based on the delaunay tessellation. *Pattern Recognition*, 36(3):799–809.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1192–1200, New York, NY, USA. Association for Computing Machinery.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021a. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, Online, August. Association for Computational Linguistics.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021b. LayoutXLM: Multimodal pre-training for multilingual visually-rich document understanding.

Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C. Lee Giles. 2017a. Learning to extract semantic structure from documents using multimodal fully convolutional neural network.

Xiaowei Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C. Lee Giles. 2017b. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4342–4351.

Antonio Jimeno Yepes, Xu Zhong, and Douglas Burdick. 2021. Icdar 2021 competition on scientific literature parsing.

Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. 2021. Pick: Processing key information extraction from documents using improved graph learning-convolutional networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4363–4370. IEEE.

Peng Zhang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Jing Lu, Liang Qiao, Yi Niu, and Fei Wu. 2020. Trie: End-to-end text reading and information extraction for document understanding. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1413–1422.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2019a. Image-based table recognition: data, model, and evaluation. *arXiv preprint arXiv:1911.10683*.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019b. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, Sep.