中图法分类号:TP18; TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-18

论文引用格式: Li Xiao-Hui, Zhou Yu, Peng Liang-Rui, Chen Shan-Xiong, Lian Zhou-Hui, Gao Liang-Cai, Yin Xu-Cheng, Liu Cheng-Lin. The Technological Evolution from Text Recognition to Intelligent Document Parsing[J/OL]. Journal of Image and Graphics, XXXX:1-18. DOI: 10.11834/jig.260148. (李晓辉, 周宇, 彭良瑞, 陈善雄, 连宙辉, 高良才, 殷绪成, 刘成林. 从文本识别到智能文档解析的技术演进[J/OL]. 中国图象图形学报,XXXX:1-18. DOI: 10.11834/jig.260148.)[DOI:10.11834/jig.260148]

从文本识别到智能文档解析的技术演进

“文档图像微沙龙”系列活动前沿综述

李晓辉[1],周宇[2],彭良瑞[3],陈善雄[4],连宙辉[5],高良才[5],殷绪成[6],刘成林[1,7]

1. 中国科学院自动化研究所　多模态人工智能系统全国重点实验室,北京　100190; 2. 南开大学　计算机学院　&　密码与网络空间安全学院　VCIP　&　TMCC　&　DISSec,天津　300350; 3. 清华大学　电子工程系,北京　100084; 4. 西南大学　计算机与信息科学学院,重庆　400715; 5. 北京大学　王选计算机研究所,　北京　100871; 6. 北京科技大学　计算机与通信工程学院,北京　100083; 7. 中国科学院大学　人工智能学院,北京　100049

摘　要: 文档图像分析与识别(Document Image Analysis and Recognition, DIAR)作为连接物理世界与数字信息的关键桥梁,其技术体系正经历从传统任务驱动向大模型时代智能理解的深刻变革。本文基于中国图象图形学学会文档图像分析与识别专业委员会主办的“文档图像微沙龙”系列学术活动,系统梳理并凝练了近年来中国青年学者在该领域的代表性成果。文章以技术演进为脉络,首先回顾了文字检测、识别、公式与表格等核心基础任务的创新突破,重点阐述了开放集识别、自监督学习等前沿范式;进而探讨了从独立任务到端到端联合优化的系统性进展;最后,聚焦于大模型时代下智能文档解析的新范式,深入剖析了专用光学字符识别(Optical Character Recognition, OCR)大模型、多模态文档解析框架以及评估体系构建等关键方向。本文旨在勾勒DIAR领域从精细化单点技术到智能化系统集成、再到认知级语义理解的发展全景,为构建高鲁棒性、可解释且高效的通用文档智能基座提供理论参考与实践指引。本文提及的算法、数据集和评估指标已汇总至https://github.com/xhli-git/Micro-Salon-Survey.

关键词: 文档图像分析;光学字符识别;大视觉语言模型;端到端学习;自监督学习;智能文档解析;评估基准

The Technological Evolution from Text Recognition to Intelligent Document Parsing

A Survey of Advances from the "Document Image Micro-Salon" Seminar Series

Li Xiao-Hui[1], Zhou Yu[2], Peng Liang-Rui[3], Chen Shan-Xiong[4], Lian Zhou-Hui[5], Gao Liang-Cai[5], Yin Xu-Cheng[6], Liu Cheng-Lin[1,7]

1. National Key Laboratory of Multi-Modal Artificial Intelligence System, Institute of Automation of Chinese Academy of Sciences, Beijing 100190, China; 2. VCIP & TMCC & DISSec, College of Computer Science & College of Cryptology and Cyber Science, Nankai University, Tianjin 300350, China; 3. Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China; 4. College of Computer

收稿日期:2026-03-24;**修回日期**:2026-05-07

* **通信作者**:刘成林,男,中国科学院自动化研究所研究员,主要研究方向为模式识别、机器学习、文档分析与识别。E-mail: liucl@nlpr.ia.ac.cn

基金项目:国家自然科学基金联合基金重点项目(项目编号:U23B2029);国家自然科学基金面上项目(项目编号:62376266)

Supported by: National Natural Science Foundation of China (Grant NO U23B2029); National Natural Science Foundation of China (Grant NO 62376266)

and Information Science, Southwest University, Chongqing 400715, China; 5. Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China; 6. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China; 7. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Document Image Analysis and Recognition (DIAR) stands as a pivotal technological bridge, transforming static physical documents into dynamic, structured digital information. This field is currently undergoing a profound paradigm shift, evolving from a collection of isolated, task-specific pipelines towards an era of holistic, intelligent understanding powered by large-scale models. This comprehensive survey, synthesized from over twenty presentations at the "Document Image Micro-Salon" seminar series organized by the Technical Committee on Document Image Analysis and Recognition of the China Society of Image and Graphics, offers a systematic review of the groundbreaking contributions made by young Chinese scholars in recent years. We trace the trajectory of this evolution across three interconnected layers: the continuous refinement of foundational tasks, the architectural unification of these tasks into end-to-end systems, and the emergence of new intelligent parsing paradigms driven by large vision-language models (LVLMs). At the foundational layer, significant innovations have redefined core challenges. In text recognition, research has moved decisively beyond closed-set, fully supervised settings. Pioneering work has formalized and addressed the Open-Set Text Recognition (OSTR) problem, enabling models to recognize novel characters—crucial for applications like ancient manuscript digitization or minority language processing—without costly retraining, through frameworks that learn to map character labels to visual prototypes. Concurrently, the field has embraced self-supervised and semi-supervised learning to overcome the bottleneck of expensive annotations. Novel frameworks that synergistically combine contrastive learning with masked image modeling have demonstrated remarkable gains in general representation power, while unified architectures for joint supervised and self-supervised learning have effectively bridged the domain gap between synthetic and real-world data. Furthermore, there has been a concerted push towards more efficient and universal models. Architectures like SVTR have shown that a single, pure vision model can outperform traditional hybrid designs, and new decoding strategies have reconciled the speed-accuracy trade-off in sequence prediction. The frontier has even expanded to instruction-guided recognition, where models are trained to understand text by predicting rich character attributes, showcasing a deeper level of semantic interaction. The deep structural complexities of mathematical expressions and tables have also seen dedicated advances. For handwritten mathematical expression recognition (HMER), graph-to-graph learning paradigms have enabled explicit modeling of hierarchical symbol relationships, while structured string decoders have offered a balanced approach between sequential and tree-based methods. More recently, leveraging the power of pre-training, specialized LVLMs for HMER have been developed, featuring hierarchical adapters that separately handle primitive character recognition and structural relationship inference, achieving state-of-the-art results that surpass both general-purpose and previous specialized models. In table structure recognition (TSR), solutions have been tailored to the wild variations found in real-world documents. Innovations include cycle-pairing modules for robust cell detection in distorted tables and visual-aligned sequential modeling that ensures precise physical bounding box prediction by enriching logical representations with fine-grained local visual features. The ultimate goal is a unified framework capable of end-to-end, multi-turn conversational parsing of complex tabular data. Building upon these refined components, the second layer of progress focuses on system-level integration. The traditional, error-prone pipeline of "detect-then-recognize" is being replaced by deeply fused, end-to-end architectures. Models like Text Perceptron and MANGO have pioneered methods to jointly optimize detection and recognition, using techniques such as fiducial point regression and mask attention to directly read character sequences from feature maps, thereby eliminating intermediate cropping steps and enabling seamless handling of arbitrary text shapes. This trend culminates in the vision of General OCR (OCR-2.0), embodied by unified transformer models that can process all forms of human-readable optical signals—text, formulas, tables, and charts—in a single, elegant framework. Yet, the field pragmatically acknowledges that meticulously engineered, high-performance pipelines remain highly valuable, as evidenced by open-source parsers like MinerU that integrate multiple SOTA modules with sophisticated post-processing rules to achieve production-grade quality. Finally, the advent of the large model era has catalyzed a third wave of innovation, giving rise to new intelligent document parsing paradigms. Recognizing the limitations of generic LVLMs in professional OCR tasks (e.g., hallucina-

tion，inefficiency），the community has developed a new generation of specialized，lightweight OCR-LVMs. These models，such as PaddleOCR-VL and HunyuanOCR，strategically balance the benefits of end-to-end learning with the reliability of modular design，often through a two-stage “analyze-then-parse” workflow that first predicts a layout and then performs parallel，element-wise content recognition. This approach preserves structural integrity while maintaining high efficiency. To support this rapid development，a new ecosystem of comprehensive evaluation benchmarks has emerged. These benchmarks，including OCRBench，OmniDocBench，and TextHalu-Bench，move far beyond simple accuracy metrics. They provide systematic，fine-grained assessments of a model’s capabilities across diverse languages，layouts，and degradation conditions，while also quantifying critical factors like its susceptibility to semantic hallucination and the cascading impact of its errors on downstream applications like RAG systems. In summary，this survey delineates a clear and vibrant path of DIAR’s evolution—from the meticulous engineering of individual components，through their synergistic integration into robust systems，to the cognitive-level semantic understanding promised by the new generation of intelligent document parsers. It provides a crucial reference for the ongoing effort to build a universal document intelligence foundation that is not only highly accurate but also robust，interpretable，efficient，and truly capable of serving the complex demands of the real world. The algorithms，datasets，and evaluation metrics mentioned in this paper have been compiled athttps://github. com/xhli-git/Micro-Salon-Survey.

Key words: Document Image Analysis; Optical Character Recognition (OCR); Large Vision-Language Models; End-to-End Learning; Self-Supervised Learning; Intelligent Document Parsing; Evaluation Benchmarks

论文引用格式:DOI:10. 11834/jig. 260148

0 引　言

近年来,人工智能、计算机视觉与自然语言处理的深度融合,推动文档图像分析与识别(Document Image Analysis and Recognition，DIAR)技术迈入新阶段。作为连接物理文档与数字信息的关键桥梁,文档图像分析与识别在教育、金融、政务、出版及智能办公等场景中展现出日益重要的应用价值。尤其在大模型时代背景下,传统光学字符识别(OCR)正加速向“OCR-2. 0”范式演进——其核心不再局限于文本转录,而是转向对文档中多模态结构化内容的联合理解,涵盖文字、数学公式、表格、图表以及整体版面语义的端到端解析。

为促进学术界与产业界的深度交流,中国图象图形学学会文档图像分析与识别专业委员会于2021年9月开始发起“文档图像微沙龙”系列学术活动,每月邀请一线青年学者分享顶会顶刊研究成果、技术挑战与行业实践。截至2026年2月5日,该活动已成功举办48期。本综述文章基于该系列沙龙中二十余场主题报告,围绕文字识别、公式识别、表格理解、版面分析与端到端文档解析等方向,系统梳理当前研究的核心问题、关键技术路径与未来发展趋势,见图1。本文旨在凝练中国青年学者在文档图像分析与识别领域的代表性成果,形成具有学术深度与实践指导意义的阶段性总结,为后续研究提供参考,并推动该领域在理论创新与产业落地上的协同发展。

1 基础任务的精进与革新

传统DIAR研究常被解构为一系列独立但又紧密关联的子任务,如文字检测、文字识别、公式识别、表格识别等。当前,文字检测和文字识别一般以文本行为处理对象,又称为文本检测(Text Detection)和文本识别(Text Recognition)。近年来,中国青年学者在这些基础任务上不断深耕,不仅在性能上屡破纪录,更在问题定义、学习范式和模型设计上带来了深刻的革新,详见表1。

1.1 文字识别:从闭集到开放,从全监督到自监督

文字识别是DIAR的基石。早期研究主要聚焦于提升在特定数据集上的闭集(Closed-Set)识别精度。然而,现实世界的应用场景远比实验室复杂,新字符、新字体、新语言层出不穷,对模型的泛化能力提出了严峻挑战。

1.1.1 开放集文字识别的探索

针对传统模型在面对训练集未见字符时完全失效的问题,北京科技大学刘畅等人首次形式化了开放集文字识别(Open-Set Text Recognition，OSTR)任

务，并提出了标签到原型学习（Label-to-Prototype Learning）框架（Liu 等，2022）。其核心动机在于，现实应用中（如古籍数字化、小语种处理）不断涌现新字符，而重新标注和训练模型成本高昂。该框架包含三个核心组件：1）标签到原型模块，通过一个可泛化的 ProtoCNN 网络，将字符标签（以 Noto 字体字形为输入）映射为其视觉原型（类中心）；2）开放集预测器，通过计算输入特征与动态生成原型的余弦相似度进行分类，并引入一个可学习的拒识阈值来处理完全无关的“集外”样本；3）拓扑保持变换网络（Topology-Preserving Transformation Network，TPTNet），一种轻量级特征级网格变换模块，用于校正不规则文本。此方法使得模型无需重训练即可适应新字符，并具备新类发现的能力。

在此基础上，他们进一步洞察到，现有模型的字符特征常与语言上下文（如 n-gram 统计）混合，在面对新字符时，模型会因上下文不匹配而将其“纠正”为错误的已知字符，这构成了一个严重的偏置源。为此，他们提出了字符-上下文解耦（Character-Context Decoupling，CCD）框架（Liu 等，2023）。该方法将上下文信息分解为时序信息（字符顺序与词长）和语言信息（n-gram 等），并分别通过两个模块进行隔离：1）分离式时序注意力（Decoupled Temporal Attention，DTA）模块，通过切断梯度回传，仅用时序信息预测词长并采样字符特征，避免其污染视觉特征；2）解耦上下文锚点（Decoupled Context Anchor，DCA）机制，在训练时显式建模并分离语言信息的影响，在开放集评估时则忽略此不可靠的先验，仅依赖纯净的视觉特征进行识别。实验表明，CCD 在识别新字符（如日文假名）上性能显著优于基线，有效缓解了上下文偏置问题。

1.1.2 自监督与半监督学习的突破

在数据利用方面，高昂的标注成本一直是制约模型性能的瓶颈。华南理工大学杨明锟等人受人类通过“读”（判别）与“写”（生成）双重行为学习文字的启发，开创性地将对比学习（Contrastive Learning，CL）与掩码图像建模（Masked Image Modeling，MIM）相融合，提出了 DiG（Discriminative and Generative）自监督预训练框架（Yang 等，2022）。该框架首次在文字识别领域引入生成式学习目标：

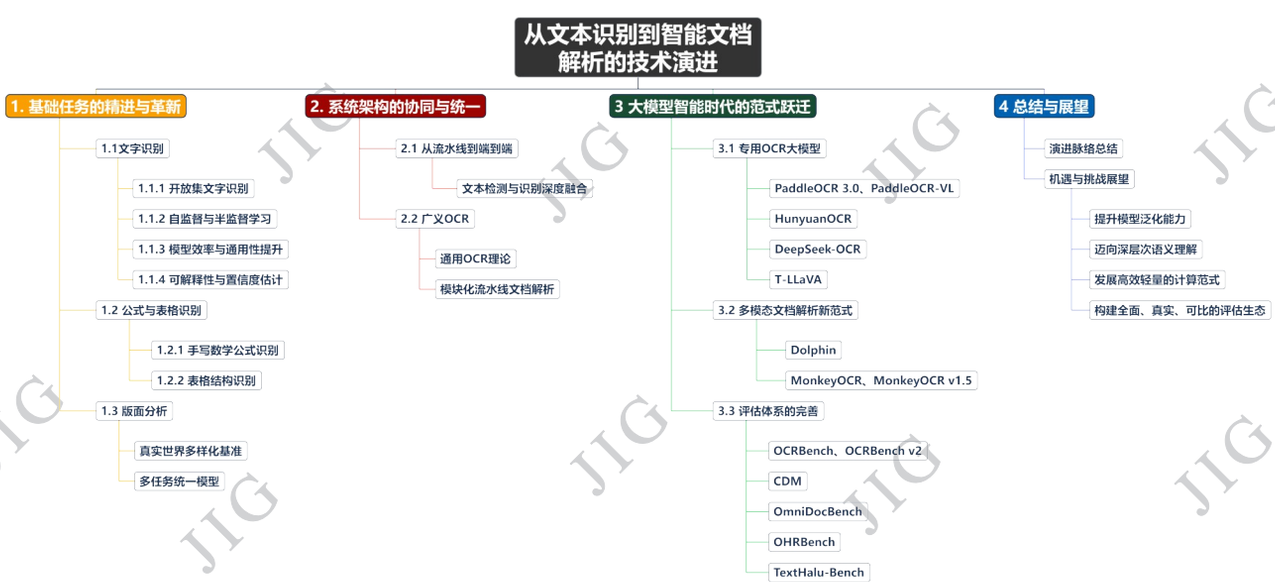

图 1　本文组织架构图

Fig. 1　Organizational chart of this article

输入图像生成掩码视图和增强视图，前者用于 MIM 任务以重建被掩码的像素（模拟“写”），后者与掩码视图共同用于 CL 任务以学习不变表征（模拟“读”）。两者共享一个 Vision Transformer（ViT）编码器，并辅以动量分支稳定训练。DiG 在 15.77M 无标签真实图像上预训练后，在 11 个文本识别基准上平均超越 SOTA 方法 5.3%，并在不规则文本上提升尤为显著（10.2%-20.2%）。此外，DiG 预训练模型能有效迁移到文本分割和超分辨率等下游任务，展现了强大的通用表征能力。

上海交通大学官同坤等人则针对自监督学习中存在的表征单元粗糙(如图像块/序列块导致字符特征混合)、数据增强不灵活(大几何变换破坏序列对应关系)及两阶段训练引发的“真实到合成”域漂移等问题，提出了CCDPlus框架(Guan等，2025)。其核心是构建一个联合监督与自监督学习(Joint Supervised and Self-supervised Learning，JSSL)的统一框架。在监督分支(使用合成数据)引入弱监督字符分割(Weakly Supervised Character Segmentation，WCS)模块，通过聚类生成伪标签学习文本前景，并利用识别器注意力权重指导字符水平位置预测，从而在线生成精确的字符结构。在自监督分支(使用无标签真实数据)，利用已训练的WCS模块在线提取字符结构作为细粒度表征单元，并通过字符到字符的蒸馏，在不同增强视图间对齐同一字符的特征。该统一框架有效弥合了合成与真实数据间的域间隙，在更具挑战性的Union14M-L数据集上提升达6.1%。

华中科技大学罗东亮等人则将目光投向了更具挑战性的半监督端到端文字识别(Semi-Supervised End-to-end Text Spotting，SSTS)。他们提出的SemiETS框架(Luo等，2025)，系统性地解决了SSTS中伪标签质量差的两大核心挑战：任务间不一致(检测与识别伪标签难以同时准确)和师生模型间不一致(因区域错位导致识别监督模糊)。该框架基于师生架构，包含两个关键模块：1)渐进式样本分配(Progressive Sample Assignment，PSA)，通过联合检测-识别约束的代价函数进行匹配，并分层筛选出可靠的“仅检测”(Det-only)和“端到端”(E2E)伪标签，以抑制噪声；2)互挖掘策略(Mutual Mining Strategy，MMS)，利用双向一致性来增强监督信号：一方面通过空间感知一致性集成(Spatial Consistency Integration，SCI)，利用多边形DIoU衡量师生预测区域的对齐度，动态调整识别损失权重以缓解模糊监督；另一方面通过内容感知区域校准(Content-aware Region Calibration，CRC)，利用Levenshtein距离衡量师生识别结果的差异，动态放大回归损失以校准不精确的检测框。实验表明，SemiETS在低标注比例下(如0.5%)性能远超现有SSL方法，甚至超越全监督强基线，展现出优秀的域适应能力。

1.1.3　模型效率与通用性的提升

在模型效率与通用性方面，复旦大学杜永坤等人的工作尤为突出。他们提出的SVTR(Single Visual model for Text Recognition)(Du等，2022)摒弃了传统的“视觉特征提取+序列建模”混合架构，旨在仅用单一视觉模型完成端到端识别。其动机源于观察到：若能直接从图像中提取兼具局部笔画细节与全局字符间依赖的判别性特征，则可完全摒弃RNN或Transformer解码器。方法上，SVTR首先将输入图像通过渐进式重叠卷积嵌入分解为“字符组件”(character components)；随后采用三阶段、高度递减的骨干网络，在各阶段交替堆叠局部混合块(Local Mixing，通过滑动窗口自注意力捕获字符内笔画模式)和全局混合块(Global Mixing，通过全图自注意力建模字符间长程依赖)，形成多粒度特征感知；最后通过简单的线性分类器并行预测字符。大型模型SVTR-L在英文基准上达到SOTA水平，并在中文数据集上以显著优势(最高+9.6%)超越现有方法，同时推理更快。

在此基础上，他们进一步设计了上下文感知并行解码器(Context Perception Parallel Decoder，CPPD)(Du等，2025a)，以解决自回归(AR)解码器精度高但速度慢、并行解码(PD)速度快但精度低的矛盾。CPPD通过两个专用模块在单次前向传播中重建AR解码器的上下文引导能力：1)字符计数(Character Counting，CC)，预测词表中每个字符的出现次数；2)字符排序(Character Ordering，CO)，推断内容无关的字符占位符及其阅读顺序。二者均采用交叉注意力机制，并由专用侧损失引导学习。最终，增强后的特征用于并行字符预测。基于SVTR的CPPD模型在英/中文基准上达到SOTA精度，且推理速度比AR基线快约8倍。

更进一步，他们提出了指令引导的文本识别(Instruction-Guided Text Recognition，IGTR)新范式(Du等，2025b)。IGTR将STR重构为一个指令学习问题，通过预测字符属性(如频率、位置、状态等)来深化对文本图像的理解。方法上，首先构建丰富的条件，问题，答案　三元组指令集，用于描述字符属性及模拟不同识别流程。模型架构包含一个轻量级指令编码器、一个跨模态特征融合模块和一个多任务答案头。实验表明，IGTR在英/中文基准上均显著超越SOTA模型，且模型小(24.1M)、推理快(4-10ms)。通过调整指令采样策略，IGTR能优雅地解决稀有字符和形近字符的识别难题。

华中科技大学屈亚东等人则关注半监督场景下的性能上限问题，提出了ViSu方法（Qu等，2024）。该方法通过将在线生成策略引入Mean-Teacher架构，使模型能够通过简单的合成数据训练，具备泛化到复杂字符（如严重形变的字符和艺术字）的能力。在归纳阶段，设计了一种基于对比学习的损失函数，从理论上纠正了将部分正样本误视为负样本的公式偏差，提高了类内分布的紧致性，增强了模型对字符视觉形态的鲁棒性。

1.1.4　文字识别可解释性和置信度提升

尽管文字识别技术近年来进展迅速，但序列到序列（Sequence-to-Sequence）的主流方法难以提供精确的字符定位与分割，导致识别结果缺乏可解释性，且难以对错误字符进行有效拒识；而传统显式分割方法又严重依赖昂贵的字符级标注数据。针对这一矛盾，于明明等人提出了一种统一字符分割与识别的手写中文文本行识别框架（Yu等，2024）。该方法创新性地将识别建模为基于滑动窗口的候选字符分类与边界框回归任务，并设计了联合训练范式：利用字符级标注的合成数据进行显式对齐监督，同时利用仅行级标注的真实数据通过CTC映射生成隐式软标签进行弱监督学习。该框架无需真实数据的字符级标注即可实现高精度分割与识别，在ICDAR-2013数据集上取得了98.13%的正确率（CR）和95.10%的分割F1分数，且在仅需0.1%字符级标注的半监督设置下仍保持优异性能，显著提升了模型的可解释性与数据利用效率。

在获得精确字符定位的基础上，针对深度神经网络普遍存在的“过度自信”问题（即错误识别样本的置信度虚高），刘扬扬等人进一步提出了一种融合多源上下文的字符级置信度估计方法（Liu等，2024b），旨在优化高可靠性场景下的错误拒识性能。该方法基于贝叶斯概率公式，构建了一个统一的置信度校准框架，有机融合了重训练的字符形状分类器得分、语言模型（BERT-like）提供的语义上下文得分以及改进的几何上下文得分。其核心贡献在于提出了双边二元几何特征（BiBG），通过同时考量当前字符与前驱、后继字符的相对位置关系，克服了传统几何特征仅关注单向关系的局限。实验表明，该方法在ICDAR2013数据集上显著优于最大软最大概率（MSP）基线及其他校准策略，将精度达99%时的拒绝率（RP99）从基线的25.27%大幅降至13.66%，即使在识别器已包含语言模型的情况下，仍能进一步将RP99从11.47%优化至6.82%。这两项工作分别从“结构可解释性”与“置信度可靠性”两个维度，为构建高可信、可交互的文字识别系统提供了关键技术支持。

1.2　公式与表格识别：面向结构化内容的深度解析

数学公式和表格是文档中承载高密度、高价值信息的核心元素，其二维乃至多维的结构特性给传统一维序列模型带来了巨大挑战。

表1　OCR基础任务相关工作列表

Table 1　List of related works on fundamental OCR tasks

任务	工作	核心方法	实验结果	主要贡献	开源链接
文字识别	OSTR (Liu等，2022)	提出标签到原型学习框架，含ProtoCNN、开放集预测器和TPTNet	模型无需重训练即可识别新字符，并能发现新字符	首次形式化开放集文字识别任务，解决新字符泛化问题	https://github.com/lancercat/OSOCR
	CCD (Liu等，2023)	通过DTA模块和DCA机制分离时序与语言上下文信息	在日文假名等新字符识别上显著优于基线，缓解上下文偏置	揭示并解决了上下文导致新字符被错误纠正的偏置问题	https://github.com/lancercat/VSDF
	DiG (Yang等，2022)	融合对比学习（CL）与掩码图像建模（MIM），共享ViT编码器	在11个基准平均超越SOTA 5.3%，不规则文本提升10.2%–20.2%	首次在文字识别中引入生成式自监督目标，提升通用表征能力	https://github.com/ayumiymk/DiG
	CCDPlus (Guan等，2025)	构建联合监督与自监督学习（JSSL）框架，引入WCS模块和字符级蒸馏	在Union14M-L上提升达6.1%	统一框架弥合合成与真实数据域间隙，解决表征粗糙等问题	https://github.com/Tongkun-Guan/CCD

表1续表

任务	工作	核心方法	实验结果	主要贡献	开源链接
	SemiETS (Luo等,2025)	基于师生架构,提出PSA筛选可靠伪标签和MMS增强监督信号	在0.5%标注比例下超越全监督基线,展现强域适应能力	系统解决半监督端到端文字识别中伪标签质量差的核心挑战	https://github.com/DrLuo/SemiETS
	SVTR (Du等,2022)	单一视觉模型,交替局部/全局混合块提取多粒度特征	英文达SOTA,中文最高+9.6%,推理更快	摒弃传统混合架构,证明单一视觉模型可高效完成识别	https://github.com/PaddlePaddle/PaddleOCR
	CPPD (Du等,2025a)	并行解码器中引入字符计数(CC)和字符排序(CO)模块	达SOTA精度,推理速度比自回归快约8倍	解决并行解码精度低问题,在单次前向重建上下文引导能力	https://github.com/PaddlePaddle/PaddleOCR/blob/dygraph/doc/doc_en/algorithm_rec_cppd_en.md
	IGTR (Du等,2025b)	将识别重构为指令学习,预测字符属性,使用三元组指令集	超越SOTA,模型小(24.1M)、快(4-10ms),擅长稀有/形近字符	提出指令引导新范式,通过属性预测深化文本理解	https://github.com/Topdu/OpenOCR
	ViSu (Qu等,2024)	在Mean-Teacher中引入在线生成策略和对比损失	能泛化到严重形变字符和艺术字	提升半监督场景下对复杂字符的鲁棒性和性能上限	https://github.com/qqqyd/ViSu
	统一字符分割与识别 (Yu等,2024)	建模为滑动窗口候选分类与回归,联合显式/弱监督训练	ICDAR-2013上CR 98.13%,F1 95.10%,0.1%标注下效果仍优	无需真实字符级标注实现高精度分割与识别,提升可解释性	--
	多源上下文字符级置信度估计 (Liu等,2024b)	融合形状、语义、几何(BiBG)上下文,基于贝叶斯校准	RP99从25.27%降至13.66%,有语言模型时从11.47%降至6.82%	提出高可靠性拒识方案,显著提升置信度校准效果	--
公式识别	G2G (Wu等,2021)	图到图学习,GNN-GNN架构,子图注意力机制	CROHME 2014/2016 ExpRate达54.46%/52.05%	首次显式建模公式层次结构并实现符号分割	--
	SSD (Wu等,2022)	结构化字符串解码器,用占位符展平LaTeX,带记忆队列注意力	CROHME 2014 ExpRate达53.1%	平衡字符串与树解码器优缺点,显式建模层次结构	--
	VLPG (Guo等,2025a)	视觉-语言预训练,定位与语言建模pretext任务,两步微调	CROHME各年ExpRate提升至60%+,OffRaSHME达67.75%	利用未配对数据缓解标注稀缺,提升泛化能力	https://github.com/guohy17/VLPG
	HiE-VL (Guo等,2025b)	分层大模型,SAM视觉编码器,"原始/结构"双适配器	CROHME 2014 ExpRate达73.3%,远超GPT-4V和SOTA专用模型	首个专为HMER设计的LVLM,证明其在复杂公式识别领先潜力	https://github.com/guohy17/HiE-VL
表格识别	WTW (Long等,2021)	CenterNet加循环配对模块,检测中心点与顶点,配对损失	WTW上TEDS指标绝对优势24.6%,ICDAR2019达SOTA	解决野外复杂表格单元格检测与分组难题	https://github.com/wangwen-whu/WTW-Dataset
	VAST (Huang等,2023)	坐标序列解码器(自回归预测边界框)+视觉对齐损失	六个基准上逻辑与物理结构指标均达SOTA	解决生成式TSR物理结构预测不准问题,丰富局部视觉信息	--

表1续表

任务	工作	核心方法	实验结果	主要贡献	开源链接
	TabPedia (Zhao等,2024a)	基于LVLM,概念协同机制,中介令牌融合多任务/多源嵌入	多基准达SOTA或极具竞争力,支持端到端多轮对话解析	统一VTU多任务,构建复杂理解新基准ComTQA	https://github.com/zhaowc-ustc/TabPedia
版面分析	M6Doc (Cheng等,2023)	发布"六多"特性真实世界文档数据集	TransDLANet在M6Doc上达64.5% mAP	填补中文及真实文档数据空白,推动细粒度布局分析	https://github.com/HCIILAB/M6Doc
	DocLayout-YOLO (Zhao等,2024b)	布局合成(二维装箱)生成DocSynth-300K,GL-CRM多尺度模块	DocStructBench上78.8% mAP,85.5 FPS	解决预训练数据同质化,平衡效率与精度	https://github.com/opendatalab/DocLayout-YOLO
	DocSAM (Li等,2025a)	统一框架,Sentence-BERT语义查询+实例查询,开集分类	在M6Doc等复杂基准上精度具竞争力,支持48数据集联合训练	打破任务壁垒,证明单一模型处理多格式/语言/标注的强大泛化能力	https://github.com/xhli-git/DocSAM
端到端文本检测与识别	Text Perceptron (Qiao等,2020)	分割检测+STM/TPS形变校正,端到端微调控制点	在Total-Text和SCUT-CTW1500显著领先	融合检测与识别,实现不规则文本端到端优化	--
	MANGO (Qiao等,2021)	单阶段,PMA模块直接从特征图读取字符,无RoI裁剪	多基准达SOTA或竞争性能,推理更快	支持任意形状文本识别,仅需粗定位即可端到端训练	--

1.2.1 手写数学公式识别

在手写数学公式识别(Handwritten Mathematical Expression Recognition, HMER)领域,中国科学院自动化研究所吴金文等人提出了一种图到图(Graph-to-Graph, G2G)的学习范式(Wu等,2021)。该方法将问题建模为图到图的学习,以克服现有方法无法充分挖掘公式层次化结构信息以及无法显式进行符号分割的缺陷。具体而言,首先将输入的在线笔画序列构建为源图,节点为笔画,边由视线算法(Line-of-Sight, LOS)和时序关系定义;将输出的符号布局树(Symbol Layout Tree, SLT)构建为目标图,显式编码父-子、兄弟等层次关系。模型采用GNN-GNN架构:编码器GNN学习源图的结构化表示,解码器GNN则基于目标图的层次约束生成节点。关键创新在于子图注意力机制,它通过正则化注意力分布,引导解码器聚焦于与当前目标节点对应的源图子图,从而在端到端训练中实现了显式的符号分割。实验在CROHME 2014/2016数据集上进行,G2G模型将表达式识别率(ExpRate)分别提升至54.46%和52.05%,显著超越SOTA。

科大讯飞吴浩等人则另辟蹊径,提出了结构化字符串解码器(Structural String Decoder, SSD)(Wu等,2022)。该方法巧妙地平衡了字符串解码器(泛化能力差)和树解码器(语言模型弱)的优缺点。首先,将LaTeX标记通过广度优先遍历转换为结构化字符串,用占位符p和结束符e展平嵌套结构,形成混合符号与关系的序列表示。SSD采用GRU进行语言建模,并设计了带记忆队列的条件注意力机制:当遇到p时,将其隐藏状态和注意力权重入队作为子表达式的解码条件;解码子表达式时,利用队列中的条件信息引导注意力,实现层次结构的显式建模。此外,引入占位符混洗策略(shuffle-p)增强训练稳定性。在CROHME数据集上的实验表明,SSD在2014/2016/2019测试集上均显著超越SOTA,例如2014年ExpRate达53.1%。

在此基础上,郭宏宇等人针对HMER中成对标注数据稀缺的难题,提出了基于图解析框架的视觉-语言预训练范式VLPG(Guo等,2025a)。该方法利用未配对的符号图像与LaTeX语料,设计了定位pretext任务(构建伪图像预训练符号检测器)与语言建模任务(基于文本符号布局树预训练结构理解模块),并通过两步微调实现视觉图与文本图的对齐。实验显示,VLPG在CROHME 2014/2016/2019/2023上的表达式识别率(ExpRate)分别从基线的53.57%/54.40%/56.99%/56.47%提升至60.41%/60.51%/62.34%/62.20%,显著优于GETD等SOTA

方法，且在OffRaSHME上达到67.75%，有效缓解了数据依赖并提升了模型泛化能力。

进一步地，为解决通用大视觉语言模型（LVLM）在公式识别中局部感知弱且忽略层次结构的问题，郭宏宇等人提出了首个专为HMER设计的分层大模型HiE-VL（Guo等，2025b）。该模型采用基于SAM的高分辨率视觉编码器保留细粒度信息，并创新设计分层适配器：通过“原始适配器”并行识别字符基元，“结构适配器”推断符号间关系，配合两阶段感知增强预训练与指令微调策略。实验表明，HiE-VL在CROHME 2014上ExpRate达73.3%，不仅远超GPT-4V（34.0%）等通用模型，更超越当前SOTA专用模型CAN-ABM（65.89%），在HME100K上也取得64.2%的优异成绩，首次证明了LVLM在复杂公式识别领域的领先潜力。

1.2.2　表格结构识别

表格结构识别（Table Structure Recognition，TSR）正从规整的电子表格向复杂的自然场景跨越。

阿里巴巴达摩院龙如蛟等人针对野外复杂表格（存在严重形变、弯曲或遮挡），发布了大规模数据集WTW（Wild Table in the Wild）（Long等，2021），并提出了Cycle-CenterNet方法。其核心动机是解决野外复杂条件下单元格的精确检测与分组难题。方法上，在CenterNet基础上引入循环配对模块（Cycle-Pairing Module），同时检测单元格的中心点和顶点，并利用相邻单元格共享顶点的几何特性进行分组。为此，设计了新颖的配对损失（Pairing Loss），通过动态加权机制，重点优化那些中心点与顶点预测未能相互指向的困难样本对。WTW数据集包含14,581张涵盖倾斜、弯曲、遮挡等7类挑战的图像。实验表明，Cycle-CenterNet在WTW上以24.6%的绝对优势（TEDS指标）超越基线，并在ICDAR2019上达到SOTA。

华为黄永帅等人则聚焦于端到端生成式TSR方法中物理结构（单元格包围框）预测不精确的共性问题，提出了VAST（Visual-Aligned Sequential coordinate modeling for Table structure recognition）框架（Huang等，2023）。其动机源于现有方法（如TableFormer）的坐标回归解码器依赖于逻辑结构（HTML）解码器的全局表征，而该表征缺乏精确定位所需的局部视觉细节。为此，VAST引入两大核心创新：1）坐标序列解码器（Coordinate Sequence Decoder），将每个非空单元格的边界框（左、上、右、下）建模为一个离散的语言序列，利用自回归方式逐坐标预测，以显式捕获坐标间的内部依赖关系；2）视觉对齐损失（Visual-Alignment Loss），在训练阶段，通过对比学习强制逻辑解码器输出的非空单元格表征与其对应的CNN视觉特征（经RoIAlign提取）在嵌入空间中对齐，从而丰富其局部视觉信息。整个框架采用CNN编码器与两个级联的Transformer解码器构成。实验在PubTabNet、FinTabNet、ICDAR2013等六个基准上验证了有效性，VAST在逻辑结构（S-TEDS）和物理结构（AP50，CAR F1）指标上均达到SOTA。

中国科学技术大学赵伟超等人则站在更高维度，提出了TabPedia统一框架（Zhao等，2024a）。该框架基于大视觉语言模型（Large Vision-Language Model，LVLM），旨在解决现有VTU（Visual Table Understanding）方法多为任务特定、流程割裂的问题。其核心创新在于概念协同机制（Concept Synergy）：将不同VTU任务（表格检测、结构识别、区域查询、问答）和多源视觉嵌入（来自低分辨率ViT-L和高分辨率Swin-B双编码器）抽象为“概念”，并通过引入中介令牌（Meditative Tokens）到LLM（Vicuna-7B）中，使其能自适应地融合所需信息。为评估真实场景下的复杂理解能力，作者构建了新基准ComTQA，包含约9,000个涉及多答案、计算与推理的QA对。实验表明，TabPedia在多个公共基准上均达到SOTA或极具竞争力的性能，并能通过多轮对话实现端到端的全图表格解析。

1.3　版面分析：面向真实世界的多样化基准与方法

文档版面分析（Document Layout Analysis，DLA）是理解文档语义的前提。然而，长期以来，公共数据集多局限于PDF格式的英文科学论文，缺乏对真实世界多样性的覆盖。

华南理工大学郑晓怡等人直面这一挑战，发布了M^{6}Doc数据集（Cheng等，2023）。该数据集以其“六多”特性——多格式（Multi-Format：扫描、拍照、PDF）、多类型（Multi-Type：科学文章、教科书、试卷、杂志、报纸、笔记、书）、多布局（Multi-Layout：矩形、曼哈顿、非曼哈顿、多列曼哈顿）、多语言（Multi-Language：中、英）、多注释类别（Multi-Annotation Category：74类，237,116个实例）和现代化（Modern）——构建了一个极具挑战性的新基准，典型样本示例见图2（插图来自Cheng等，2023）。M^{6}Doc的发布

不仅填补了中文和真实文档数据的空白，更推动了细粒度逻辑布局分析的研究。为验证其有效性，该团队同时提出了基于Transformer的TransDLANet模型，采用自适应元素匹配机制和动态交互解码器，在M^{6}Doc上取得了64.5% mAP的SOTA性能。

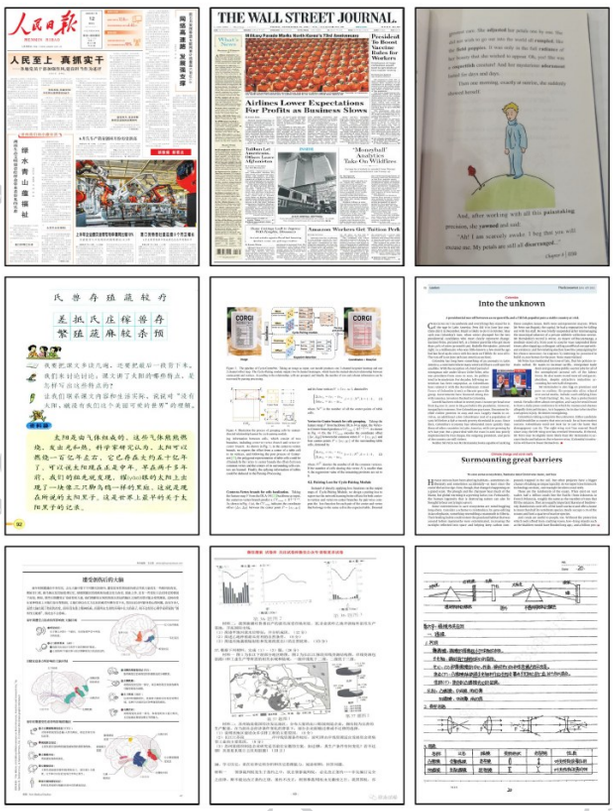

图2　M^{6}Doc数据集典型样本示例

Fig. 2　Typical examples from the M^{6}Doc dataset

上海人工智能实验室王斌等人则从效率与精度平衡的角度出发，提出了DocLayout-YOLO(Zhao等，2024b)。为解决现有预训练数据同质化问题，他们创新性地将布局合成建模为二维装箱问题，利用Mesh-candidate BestFit算法，自动生成了大规模、多样化的DocSynth-300K合成数据集。在模型层面，为应对文档元素尺度变化大的挑战，设计了全局到局部可控感受野模块(Global-to-Local Controllable Receptive Module，GL-CRM)，通过分层架构(全局页级、块级、局部语-义级)自适应地捕获多尺度特征。此外，他们构建了更具挑战性的DocStructBench基准以全面评估性能。实验表明，DocLayout-YOLO在DocStructBench上以85.5 FPS的速度达到78.8% mAP，显著超越现有SOTA方法。

中国科学院自动化研究所李晓辉等人进一步打破了任务壁垒，提出了统一文档图像分割框架DocSAM(Li等，2025a)。该方法摒弃了针对特定任务的专用模型设计，创新性地将版面分析、多粒度文本分割及表格结构识别统一建模为实例与语义分割的组合任务。DocSAM利用Sentence-BERT将类别名称映射为语义查询，与可学习的实例查询在混合查询解码器中进行深度交互，实现了基于原型匹配的开集分类，从而支持在包含PubLayNet、M6Doc等48个异构数据集上的联合训练。实验显示，DocSAM不仅在M6Doc等复杂基准上取得了具有竞争力的检测精度(如微调后mAP显著提升)，更证明了单一模型在处理多格式、多语言及多标注体系文档时的强大泛化能力与资源效率，为构建通用文档智能基座模型提供了新的范式。

2　系统架构的协同与统一

随着基础任务性能的不断提升，研究者们开始关注如何打破各子任务间的壁垒，通过协同优化来构建更鲁棒、更高效的端到端系统。

2.1　从流水线到端到端：检测与识别的深度融合

传统的文字识别系统通常采用“先检测后识别”的流水线架构，这种设计容易导致误差累积。海康威视乔梁等人提出的Text Perceptron框架(Qiao等，2020)，通过形状变换模块(Shape Transformation Module，STM)显式学习文本边界上的控制点(fiducial points)，并利用薄板样条(Thin Plate Spline，TPS)将不规则特征区域直接变换为规则形态，从而无缝桥接检测与识别。检测器采用高效的分割范式，将文本区域细分为中心、头、尾及上下边界四类，并辅以角点和边界偏移回归，以精确捕获文本几何形状和潜在阅读顺序。STM生成的控制点位置可通过识别损失进行端到端微调，实现全局优化。该方法在Total-Text和SCUT-CTW1500两个不规则文本基准上显著超越先前方法。

在此基础上，他们进一步提出了MANGO(Mask Attention Guided One-stage Scene Text Spotter)(Qiao等，2021)，一种新颖的单阶段文本识别器。其核心动机是摒弃显式的RoI裁剪，直接从粗略定位的文本区域中读取字符序列。方法上，MANGO引入位置感知掩码注意力(Position-aware Mask Attention，PMA)模块，包含实例级(IMA)和字符级(CMA)两个

子模块。IMA利用动态卷积将不同文本实例的特征分配到特征图的不同通道；CMA则在IMA基础上进一步生成每个字符的注意力掩码。通过这种方式，模型能直接从原特征图中提取实例/字符特征，并送入轻量级序列解码器进行批量识别。该框架可仅用粗略的位置信息（如矩形框或中心点）和文本标注进行端到端训练，并天然适应任意形状文本。实验表明，MANGO在多个规则与不规则文本基准上均取得竞争性甚至SOTA的性能，且推理速度优于多数两阶段方法。

2.2 广义OCR：迈向统一的文档理解模型

面对日益复杂的文档内容（文本、公式、表格、图表等），模块化的流水线方案愈发显得笨重且难以维护。中国科学院大学魏浩然等人提出了通用OCR理论（General OCR Theory，OCR-2.0）（Wei等，2024），其实现模型GOT（General Optical character recognition Transformer）是一个统一、优雅、端到端的580M参数模型，能够处理所有人工光学信号。GOT采用简洁的编码器-解码器架构：编码器基于ViT-Det，具备高压缩率（1024x1024图像压缩至256个token）；解码器采用0.5B参数的Qwen语言模型，支持8K长上下文。为高效训练，采用三阶段策略：1）解耦预训练编码器；2）联合训练编码器-解码器以注入多任务知识；3）仅微调解码器以定制新功能（如区域交互式OCR、动态分辨率、多页OCR）。为此，构建了覆盖六大类任务的合成数据引擎。实验表明，GOT在中英文文档、场景文本、公式、表格、图表等多项任务上均达到或超越SOTA，且可部署于消费级GPU。

上海人工智能实验室的MinerU（Wang等，2024）则代表了另一条务实的路径。作为一个高精度、一体化的开源文档解析器，MinerU集成了五个SOTA子模型（布局检测、公式检测、表格识别、公式识别、OCR），其中布局和公式检测模型在自建的21K/2.9K页多样化数据集上微调，显著优于通用模型。后处理阶段通过规则解决边界框重叠、跨栏/跨页段落合并及页眉页脚过滤，确保阅读顺序正确。实验表明，MinerU在学术论文、教材、试卷等11类文档上均取得高质量提取效果，其公式识别模型（UniMERNet）性能媲美Mathpix。MinerU的成功证明，在大模型时代，精心设计的流水线方案依然具有强大的生命力和实用价值。

3 大模型智能时代的范式跃迁

大模型的崛起为DIAR领域带来了前所未有的机遇与挑战。研究者们开始探索如何将大视觉语言模型（LVLM）的能力与DIAR的专业需求相结合，催生了一系列新的研究范式，详见表2。

3.1 专用OCR大模型：精度与效率的再平衡

尽管通用LVLM展现出强大的零样本能力，但其在专业OCR任务上的精度、效率和幻觉问题仍不容忽视。为此，多家机构纷纷推出了轻量级、高性能的专用OCR大模型。

百度推出了PaddleOCR 3.0（Cui等，2025a），一个面向大模型时代的开源OCR与文档解析工具包。其动机源于大语言模型（LLM）和检索增强生成（RAG）对高质量、结构化文档数据的迫切需求，以及现有OCR方案在多语言、复杂布局和语义理解上的不足。PaddleOCR 3.0包含三大核心方案：1）PP-OCRv5：轻量级（<100M参数）高精度文本识别模型，支持中、英、日等多语言统一识别，并显著提升了手写体与古籍文本的鲁棒性；2）PP-StructureV3：端到端文档解析系统，集成布局分析、表格/公式/图表识别等模块，可将文档转换为结构化JSON/Markdown；3）PP-ChatOCRv4：结合轻量OCR与LLM的关键信息抽取系统，支持多轮问答式交互。实验表明，PP-OCRv5在多场景OCR任务上超越了Qwen2.5-VL-72B等十亿级VLM；PP-StructureV3在OmniDocBench上达到SOTA；PP-ChatOCRv4在自建多场景QA基准上Recall@1达85.55%。

在此基础上，百度又推出了PaddleOCR-VL（Cui等，2025b），一个面向多语言文档解析的高效视觉语言模型（VLM）方案。其动机在于解决现有端到端VLM计算开销大、易产生幻觉，而传统流水线方法集成复杂、错误累积的问题。方法采用两阶段解耦架构：第一阶段由轻量级模型PP-DocLayoutV2执行布局分析与阅读顺序预测；第二阶段将裁剪出的元素区域送入核心模型PaddleOCR-VL-0.9B进行并行识别。该0.9B模型融合了NaViT风格的动态高分辨率视觉编码器与ERNIE-4.5-0.3B语言模型，并通过精心设计的两阶段训练（大规模对齐+指令微调）及包含3000万样本的高质量数据集进行优化，支持109种语言。实验在OmniDocBench v1.0/v1.5和

olmOCR-Bench 等多个基准上验证，PaddleOCR-VL 在文本、表格、公式、图表及阅读顺序等任务上均达到SOTA，性能超越 GPT-4o、Qwen2. 5-VL-72B 等大模型及专用方案，同时推理速度更快、资源消耗更低，典型文档解析结果见图 3（插图来自 Cui 等，2025b）。

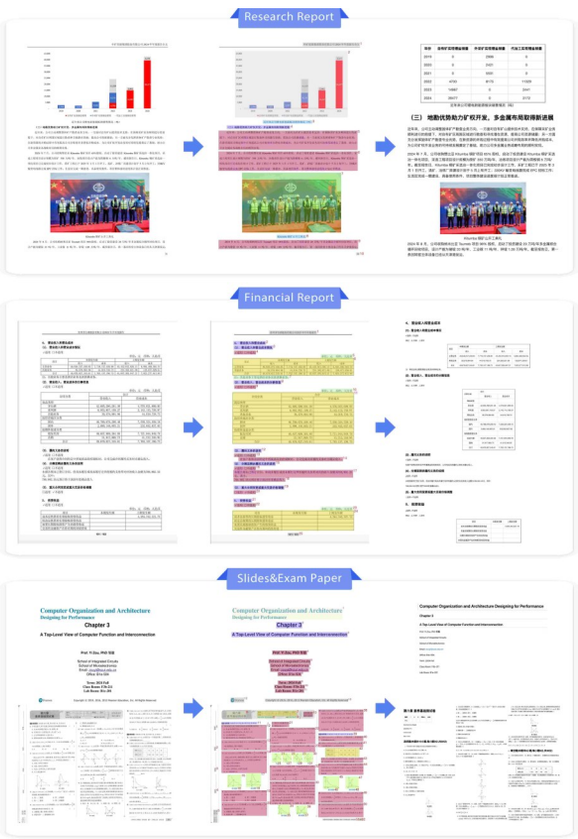

图 3　PaddleOCR-VL 文档解析结果示例

Fig. 3　Document parsing results using PaddleOCR-VL

腾讯混元的 HunyuanOCR（Team 等，2025）是一个面向商业应用的开源轻量级（1B 参数）视觉语言模型（VLM），旨在统一解决文本检测、解析、信息抽取、图文问答与翻译等多样化 OCR 任务。其动机源于传统流水线方案存在错误累积与维护成本高，而通用大 VLM 则效率低下且在专业 OCR 任务上表现不佳。方法上，HunyuanOCR 采用端到端架构，由原生分辨率 ViT 视觉编码器、自适应 MLP 连接器和轻量 LLM 组成，并通过四阶段预训练（对齐、多模态、长上下文、指令微调）与基于 GRPO 算法的强化学习（RL）进行优化，其中 RL 针对不同任务（如检测、翻译）设计了可验证或 LLM-as-a-Judge 的奖励机制。实验表明，该模型在 OmniDocBench、OCRBench 等多个基准上超越了更大规模的开源模型（如 Qwen3-VL-4B）及商业 API，在 ICDAR 2025 DIMT 挑战赛（小模型赛道）中夺冠。

DeepSeek 提出的 DeepSeek-OCR（Wei 等，2025a）更是将视觉模态视为一种高效的压缩介质，旨在探索利用视觉模态作为高效压缩介质来解决大语言模型（LLM）处理长文本时的计算瓶颈。其核心动机是“一图胜千言”，即单张文档图像可承载丰富文本信息，但仅需极少的视觉 token。方法上，DeepSeek-OCR 由两部分构成：1）DeepEncoder，一个新颖的视觉编码器，通过串联窗口注意力（SAM）与全局注意力（CLIP）组件，并在其间引入 16 倍卷积压缩模块，在高分辨率输入下维持低激活内存与极少视觉 token（如 1024x1024 图像仅输出 256 个 token）；2）DeepSeek3B-MoE 解码器，用于从压缩的视觉 token 中重建文本。模型支持多分辨率模式（Tiny 至 Gundam），以适应不同压缩比需求。实验表明，在 Fox 基准上，9-10 倍压缩比下 OCR 精度达 97%，20 倍时仍保持约 60%；在 OmniDocBench 上，仅用 100 个视觉 token 即超越 GOT-OCR2. 0（256 token），用<800 token 超越 MinerU2. 0（～7000 token）。

针对 LVLM 处理文档时的“分辨率 - 感知悖论”（下采样丢失细节 vs. 固定切片破坏语义），魏梦泽等人提出 T-LLaVA（Wei 等，2025b）。该模型旨在兼顾文本显著性与完整性，核心创新包括：1）文本密度激活器，动态分析字符分辨率与密度以智能决策切片策略；2）比率 - 分辨率自适应切片（RRS），综合图像宽高比与分辨率，通过评分函数优选切片方案以匹配 ViT 输入并保留结构；3）特征融合机制，利用分隔符有序整合全局与局部特征以维持空间拓扑。实验显示，T-LLaVA 在场景文本、公式及整页文档等基准上表现优异，其<1B 轻量模型性能超越多个>7B 大型通用模型。

3. 2　多模态文档解析新范式：分析–再解析与三元组

为了在文档图像解析中兼顾精度、效率与布局结构的完整性，研究者们提出了多种创新性的解析范式。

针对文档图像解析中现有方法的两大瓶颈——集成式方案（多专家模型）存在协调开销与效率低下，而端到端自回归生成式方案则面临长文档布局结构退化与效率瓶颈，字节跳动冯浩等人提出 Dol-

phin模型(Feng等,2025)。其核心思想是“分析-再解析”(analyze-then-parse)范式。方法上,Dolphin采用两阶段策略:第一阶段进行页面级布局分析,以阅读顺序生成包含元素类型与边界框的异构锚点序列;第二阶段将这些锚点作为提示(prompt),并行地对每个裁剪出的元素区域进行内容解析(如文本转Markdown、表格转HTML、公式转LaTeX)。为支持训练,作者构建了包含超3000万样本的大规模多粒度数据集。实验表明,Dolphin在自建的复杂文档基准Dolphin-Page及多个公开元素级基准(如PubTabNet, Formula-SPE/SCE/CPE)上均达到SOTA性能,且凭借轻量架构(322M参数)与并行机制,在运行效率(FPS)上显著优于GPT-4o、Claude等大模型及集成方案。

此外,华中科技大学李长等人也提出了MonkeyOCR及其配套数据集MonkeyDoc(Li等,2025b)。其核心是结构-识别-关系(SRR)三元范式:首先进行块级结构检测以精准分割语义区域;其次利用统一的3B大视觉语言模型(LMM)对各区域并行进行内容识别,避免错误传播;最后通过块级阅读顺序预测重建逻辑关系。为支撑训练,作者构建了迄今最全面的文档解析数据集MonkeyDoc,包含390万中英双语实例,覆盖十余种文档类型,并通过多阶段流程(整合、合成、专家标注)确保质量。实验表明,MonkeyOCR在OmniDocBench上全面超越SOTA,尤其在公式(+15.0% CDM)和表格(+8.6% TEDS)识别上提升显著;其3B模型甚至优于72B的Qwen2.5-VL和Gemini 2.5 Pro,并可在单张3090 GPU上高效部署(多页处理速度达0.84页/秒)。优点在于精度与效率兼得、泛化能力强;局限在于性能高度依赖初始结构检测的准确性。

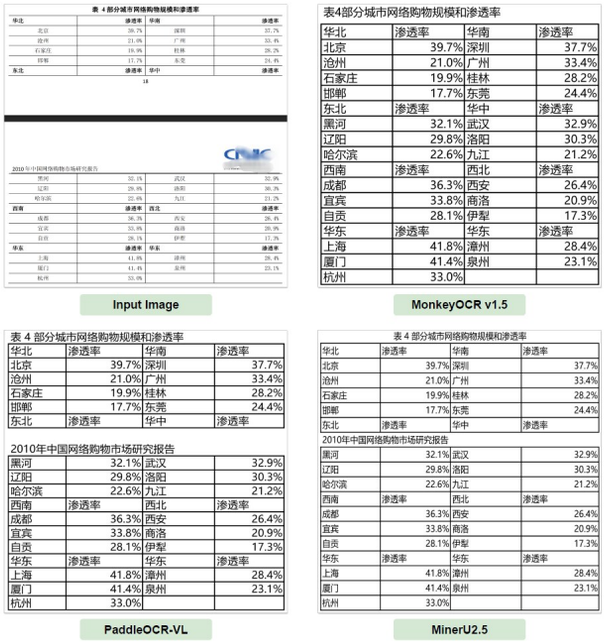

图4　跨页表格识别效果对比

Fig. 4　Comparison of cross-page table recognition

进一步地,针对复杂文档(含多级表格、嵌入图像、跨页结构)解析的挑战,华中科技大学又提出MonkeyOCR v1.5(Zhang等,2025b)。其动机是解决现有模型在复杂场景下鲁棒性不足的问题。方法上,采用两阶段统一框架:第一阶段用大视觉语言模型(VLM)联合预测文档布局与阅读顺序,确保结构与序列一致性;第二阶段对检测区域进行并行内容识别。为提升复杂表格处理能力,引入基于视

表2　文档解析相关工作列表

Table 2　List of related work on document parsing

任务	工作	核心方法	实验结果	主要贡献	开源链接
文档解析	GOT(Wei等,2024)	580M统一编码器-解码器模型,三阶段训练,合成数据引擎	多项任务达或超SOTA,可部署于消费级GPU	提出通用OCR理论,实现所有人工光学信号的统一处理	https://github.com/Ucas-HaoranWei/GOT-OCR2.0
	MinerU(Wang等,2024)	集成五个SOTA子模型,规则后处理解决布局问题	11类文档高质量提取,公式识别媲美Mathpix	证明精心设计的流水线在大模型时代仍有强大实用价值	https://github.com/opendatalab/MinerU
	PaddleOCR 3.0(Cui等,2025a)	包含PP-OCRv5、PP-StructureV3、PP-ChatOCRv4三大方案	PP-OCRv5超越十亿级VLM,PP-StructureV3在OmniDocBench达SOTA	面向大模型时代需求,提供开源、高精度、多语言OCR工具包	https://github.com/PaddlePaddle/PaddleOCR

表2续表

任务	工作	核心方法	实验结果	主要贡献	开源链接
	PaddleOCR-VL (Cui等, 2025b)	两阶段解耦架构(布局分析+并行识别),0.9B高效VLM	多基准上超越GPT-4o等大模型,速度更快、资源消耗更低	平衡端到端VLM与流水线优缺点,实现高效多语言文档解析	https://github.com/PaddlePaddle/PaddleOCR
	HunyuanOCR (Team等,2025)	1B端到端VLM,四阶段预训练+GRPO强化学习	多基准超越更大开源模型及商业API,小模型赛道夺冠	统一解决多样化OCR任务,兼顾效率与专业精度	https://github.com/Tencent-Hunyuan/HunyuanOCR
	DeepSeek-OCR (Wei等, 2025a)	DeepEncoder视觉压缩(1024x1024→256 token)+MoE解码器	100 token超越GOT-OCR2.0(256 token),<800 token超越MinerU2.0	将视觉作为高效压缩介质,缓解LLM长文本计算瓶颈	http://github.com/deepseek-ai/DeepSeek-OCR
	T-LLaVA (Wei等,2025b)	文本密度激活器+比率-分辨率自适应切片(RRS)+特征融合	<1B模型性能超越多个>7B通用模型	解决"分辨率-感知悖论",兼顾文本细节与结构完整性	--
	Dolphin (Feng等,2025)	"分析-再解析"范式:布局分析生成锚点,再并行解析内容	元素级与页面级基准达SOTA,运行效率显著优于大模型	兼顾精度、效率与布局结构完整性,避免错误传播与退化	https://github.com/ByteDance/Dolphin
	MonkeyOCR (Li等,2025b)	SRR三元范式(结构-识别-关系),3B LMM并行识别	OmniDocBench全面超越SOTA,优于72B模型,单卡高效部署	精度与效率兼得,构建最全面双语文档数据集	https://github.com/Yuliang-Liu/MonkeyOCR
	MonkeyOCR v1.5 (Zhang等, 2025b)	两阶段框架(联合布局/顺序预测+并行识别),IDTP/TGTM/RL优化	OmniDocBench v1.5总体分92.9%,复杂子集领先8.2%	显著提升复杂文档(多级表格、嵌入图像、跨页)鲁棒性	https://github.com/Yuliang-Liu/MonkeyOCR
评测基准	OCRBench (Liu等,2024a)	覆盖5大任务、29数据集、1000问答对的系统性评估	揭示LMM在手写/中文/HMER上平均差距超50%	首个系统性LMM文本能力评估框架,揭示其依赖语义先验的短板	https://github.com/Yuliang-Liu/MultimodalOCR
	OCRBench v2 (Fu等,2024)	扩展至31场景、23子任务、1万问答对,含私有测试集	前沿模型总分普遍<50,暴露五大核心局限	通过高熵指令与私有集精准诊断LMM深层缺陷	https://99franklin.github.io/ocrbench_v2
	CDM (Wang等,2025)	将LaTeX渲染为彩色图像,匈牙利匹配+RANSAC剔除无效匹配	评分与人类偏好高度一致(96%)	首创图像域公式评估,消除LaTeX风格差异带来的偏差	https://github.com/opendatalab/UniMERNet/tree/main/cdm
	OmniDocBench (Ouyang等, 2025)	覆盖9类文档、19布局、15属性,多层次评估体系	揭示流水线与VLM各自优势场景	构建全面真实文档基准,支持细粒度能力分析	https://github.com/opendatalab/OmniDocBench
	OHRBench (Zhang等, 2025a)	定义语义/格式两类OCR噪声,评估其对RAG的级联损害	最优OCR方案仍使RAG性能比真实数据低14%+	首次量化OCR噪声对下游RAG系统的负面影响	https://github.com/opendatalab/OHR-Bench
	TextHalu-Bench (Shu等, 2025)	构建幻觉基准,提出ZoomText+GLC无需训练的消除框架	有效抑制LVLM在OCR任务中的语义幻觉	首个专门针对OCR幻觉的基准与解决方案	https://github.com/shuyansy/MLLM-Semantic-Hallucination

觉一致性的强化学习(GRPO),通过渲染-比对机制自监督优化表格HTML结构,无需额外标注。此外,设计图像解耦表格解析(IDTP)模块,通过掩码-占位-重插入流程,有效处理含嵌入图像的表格;并提出类型引导表格合并(TGTM)策略,结合规则匹配与BERT语义判别,可靠重建跨页/跨栏表格。实验在OmniDocBench v1.5上验证,MonkeyOCR v1.5以92.9%的总体分超越PPOCR-VL和MinerU2.5,在OCRFlux-Complex子集上领先8.2%,并在报纸等复杂版式上表现最优。典型方法跨页表

格识别效果对比示例见图4(插图来自Zhang等，2025b)。

3.3 评估体系的完善:构建更公平、更全面的基准

随着技术的快速发展,构建科学、全面的评估体系变得至关重要。

针对大型多模态模型(LMM)在文本视觉任务中能力边界不明的问题,华中科技大学刘禹良等人联合微软研究院等机构提出了首个系统性评估框架OCRBench(Liu等,2024a)。该基准涵盖29个数据集及1000个人工校正问答对,重点评测文本识别、场景与文档VQA、关键信息提取(KIE)及手写数学表达式识别(HMER)五大任务。研究发现,尽管LMM在常规及艺术文本上表现优异,甚至超越部分监督式SOTA,但在手写体、中文、无语义文本及HMER任务上存在显著短板,平均性能差距超50%。分析表明,这主要归因于LMM过度依赖语义先验而忽视字符形状感知,且受限于输入分辨率导致细粒度特征丢失。作为开创性工作,OCRBench揭示了模型内在机制,但在多语言覆盖及检测任务上仍有局限。

为突破现有基准在复杂任务覆盖不足及性能饱和的瓶颈,同一团队联合华南理工大学、阿德莱德大学及字节跳动等单位推出了升级版OCRBench v2(Fu等,2024)。该大规模双语基准扩展至31种场景和23个子任务,新增指代、检测、元素解析、数学计算及逻辑推理等8项核心能力评估,包含1万个人工验证问答对及1500张私有测试图像。评测显示,包括GPT-4o在内的前沿模型总分普遍低于50分,暴露出低频文本识别困难、细粒度空间感知弱(定位IoU仅12.9%)、抗旋转布局能力差(性能下降超55%)、结构化解析欠缺及逻辑推理短板五大局限。OCRBench v2通过高熵指令与私有测试集有效规避数据污染,精准诊断了LMM在真实应用中的深层缺陷,虽暂未涉及长文档多页关联,但为后续模型优化提供了关键依据。

上海人工智能实验室王斌等人在这方面也做出了系统性贡献。他们提出了CDM(Character Detection Matching)(Wang等,2025),将公式识别评估从易受LaTeX风格影响的文本域转移到更符合人类直观的图像域。方法上,CDM首先将预测和真实LaTeX渲染为图像,并通过为每个token分配唯一颜色来精确定位字符级边界框;随后采用匈牙利算法进行元素级匹配,匹配成本综合考虑token一致性、位置接近度和顺序相似性;最后通过RANSAC算法剔除因结构错误(如上下标颠倒)导致的无效匹配,并以F1分数作为最终得分。实验表明,CDM与人类偏好高度一致(96%),能有效消除因表达风格差异带来的评分偏差。

他们构建的OmniDocBench(Ouyang等,2025)覆盖9类真实文档(如教材、报纸、手写笔记),并提供19种布局类别、15种属性标签(如语言、背景、旋转)及阅读顺序等丰富注释。评估方法上,设计了灵活的多层次体系:端到端评估、任务级评估(布局、OCR、表格、公式识别)和属性级评估。实验系统评测了主流方法,揭示了流水线工具(如MinerU)在常规文档上精度高,而通用VLM(如Qwen2-VL)在非常规格式(如笔记)和退化条件下(如模糊扫描)更具鲁棒性。

此外,他们还发布了OHRBench(Zhang等,2025a),首次量化了OCR环节的噪声对下游检索增强生成(Retrieval-Augmented Generation, RAG)系统性能的级联损害。OHRBench包含来自7个真实领域的8,561页文档图像及8,498个QA对。作者系统性地定义了两类OCR噪声:语义噪声(预测错误)和格式噪声(表示不一致)。实验全面评估了主流OCR方案,发现即使最优方案在RAG整体性能上仍比真实数据低14%以上。深入分析表明,语义噪声对所有RAG组件均有显著负面影响,而格式噪声的影响则因检索器和LLM而异。

针对LVLM普遍存在的语义幻觉问题,意大利特伦托大学舒言等人构建了首个专门的TextHalu-Bench基准(Shu等,2025),并提出了一种无需训练的幻觉消除框架。该框架通过ZoomText定位文本区域,并利用Grounded Layer Correction (GLC)自适应融合最聚焦的中间层表示,有效抑制了幻觉的产生,为提升LVLM在OCR任务中的可靠性提供了有效工具。

4 总结与展望

本文以“文档图像微沙龙”系列学术活动为观察窗口,系统梳理了近年来中国青年学者在文档图像分析与识别(Document Image Analysis and Recognition, DIAR)领域取得的前沿进展。从文字检测与

识别、数学公式解析、表格结构理解等基础任务的持续精进，到面向多模态、多任务的端到端统一架构设计，再到大模型驱动下智能文档解析范式的深刻跃迁，一条清晰而富有活力的技术演进脉络已然形成。

展望未来，DIAR领域仍面临诸多关键挑战，同时也孕育着广阔的发展机遇：

第一，**提升模型泛化能力仍是核心目标**。当前方法在标准印刷体文本上已取得显著成效，但在面对古籍文献、少数民族文字、手写体、艺术字形、低质量扫描件等复杂或极端场景时，性能仍显不足。如何通过开放集学习（Open-Set Learning）、自监督/弱监督预训练、跨域迁移、持续学习（包括类别增量学习和语种增量学习）等机制，构建更具鲁棒性与适应性的通用文档理解模型，是亟待深入探索的方向。

第二，**迈向深层次语义理解是必然趋势**。未来的文档智能不应止步于像素到文本的映射，而需实现从“感知内容”到“理解语义”乃至“推理知识”的跨越。这要求模型不仅能够准确提取结构化信息，还需具备上下文感知、逻辑关联、实体消歧乃至常识推理能力，从而支持从海量非结构化文档中自动挖掘高价值知识。同时，为了提升文档语义推理的可靠性（减少幻觉）和结果可追溯性，需要文档版面分析和识别的中间结果具有可解释性。

第三，**发展高效轻量的计算范式势在必行**。尽管大模型在性能上展现出强大潜力，但其高昂的计算成本与资源消耗严重制约了在移动端、嵌入式设备及实时应用场景中的部署。因此，亟需在模型压缩、知识蒸馏、神经架构搜索（NAS）、稀疏激活等方向取得突破，在保障精度的同时显著降低推理延迟与能耗，推动DIAR技术真正走向产业落地。

第四，**构建全面、真实、可比的评估生态至关重要**。当前研究仍受限于评测基准的单一性与理想化假设。未来应大力推动如OmniDocBench、OHRBench、M6Doc、WTW、TextHalu-Bench等新型基准数据集的建设与应用——这些基准强调多语言、多版式、多模态融合，并关注幻觉检测、逻辑一致性、用户意图对齐等高阶能力，能够更真实地反映实际应用需求，引导学术研究从“刷榜”转向解决真实世界问题。

我们坚信，在广大青年学者的持续创新与协同攻关下，文档图像分析与识别技术将持续突破理论边界、拓展应用场景，为构建更加智能、高效、可信的数字基础设施和知识服务体系提供坚实支撑，进而赋能教育、金融、政务、文化遗产保护等多个关键领域，助力国家数字化战略纵深发展。

致谢：本综述主要基于中国图象图形学学会文档图像分析与识别专业委员会主办的“文档图像微沙龙”系列学术报告整理而成，谨向所有报告嘉宾、主持人及组织专家致以诚挚感谢！

参考文献（References）

Cheng H Y, Zhang P R, Wu S H, Zhang J X, Zhu Q Y, Xie Z C, et al. 2023. M^{6}doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE: 15138 - 15147.

Cui C, Sun T, Lin M H, Gao T Q, Zhang Y B, Liu J X, et al. 2025a. PaddleOCR 3.0 technical report[EB/OL].[2026-02-06]. https://arxiv.org/pdf/2507.05595.pdf.

Cui C, Sun T, Liang S Y, Gao T Q, Zhang Z L, Liu J X, et al. 2025b. PaddleOCR-VL: Boosting multilingual document parsing via a 0.9B ultra-compact vision-language model[EB/OL]. [2026-02-06]. https://arxiv.org/pdf/2510.14528.pdf.

Du Y K, Chen Z N, Jia C Y, Yin X T, Zheng T L, Li C X, et al. 2022. SVTR: Scene text recognition with a single visual model[EB/OL]. [2026-02-06]. https://arxiv.org/pdf/2205.00159.pdf.

Du Y K, Chen Z N, Jia C Y, Yin X T, Li C X, Du Y N, et al. 2025a. Context perception parallel decoder for scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Du Y K, Chen Z N, Su Y C, Jia C Y and Jiang Y G. 2025b. Instruction-guided scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Fang S C, Xie H T, Wang Y X, Mao Z D and Zhang Y D. 2021. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE: 7098 - 7107.

Feng H, Wei S, Fei X, Shi W, Han Y D, Liao L, et al. 2025. Dolphin: Document image parsing via heterogeneous anchor prompting[EB/OL].[2026-02-06]. https://arxiv.org/pdf/2505.14059.pdf.

Fu L, Kuang Z B, Song J J, Huang M X, Yang B and Li Y Z, et al. 2024. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning[EB/OL].[2025-01-02]. https://arxiv.org/pdf/2501.00321.pdf

Guan T K, Shen W and Yang X K. 2025. CCDPlus: Towards accurate character to character distillation for text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Guo H Y, Wang C, Yin F, Li X H and Liu C L. 2025a. Vision-language pre-training for graph-based handwritten mathematical expression recognition. Pattern Recognition, 162: 111346 [DOI: 10.1016/j.patcog.2025.111346]

Guo H Y, Yin F, Xu J and Liu C L. 2025b. HiE-VL: A large vision-language model with hierarchical adapter for handwritten mathematical expression recognition//Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad: IEEE: 1-5 [DOI: 10.1109/ICASSP49660.2025.10889999]

Hu J S, Wu H, Chen M J, Liu C Y, Wu J J, Yin S, et al. 2023. Handwritten chemical structure image to structure-specific markup using random conditional guided decoder//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM: 8114 - 8124.

Huang Y S, Lu N, Chen D P, Li Y B, Xie Z C, Zhu S G, et al. 2023. Improving table structure recognition with visual-alignment sequential coordinate modeling//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE: 11134 - 11143.

Li X H, Yin F and Liu C L. 2025a. DocSAM: Unified document image segmentation via query decomposition and heterogeneous mixed learning//Proceedings of the Computer Vision and Pattern Recognition Conference: 15021-15032

Li Z, Liu Y L, Liu Q, Ma Z Y, Zhang Z Y, Zhang S, et al. 2025b. MonkeyOCR: Document Parsing with a Structure-Recognition-Relation Triplet Paradigm[EB/OL]. [2026-02-06]. https://arxiv.org/pdf/2506.05218.pdf.

Long R J, Wang W, Xue N, Gao F Y, Yang Z B, Wang Y P, et al. 2021. Parsing table structures in the wild//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE: 944 - 952.

Liu C, Yang C, Qin H B, Zhu X B, Liu C L and Yin X C. 2023. Towards open-set text recognition via label-to-prototype learning. Pattern Recognition, 134: 109109.

Liu C, Yang C and Yin X C. 2022. Open-set text recognition via character-context decoupling//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE: 4523 - 4532.

Liu Y L, Li Z, Huang M X, Yang B, Yu W W and Li C Y, et al. 2024a. OCRBench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences, 67 (12): 220102 [DOI: 10.1007/s11432-024-4156-8]

Liu Y Y, Chen Y, Yin F and Liu C L. 2024b. Context-aware confidence estimation for rejection in handwritten chinese text recognition//Proceedings of the International Conference on Document Analysis and Recognition. Cham: Springer: 134-151 [DOI: 10.1007/978-3-031-70684-4_9]

Luo C J, Lin Q X, Liu Y L, Jin L W and Shen C H. 2021. Separating content from style using adversarial learning for recognizing text in the wild. International Journal of Computer Vision, 129(4): 960 - 976.

Luo C J, Zhu Y Z, Jin L W and Wang Y P. 2020. Learn to augment: Joint data augmentation and network optimization for text recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE: 13746 - 13755.

Luo D L, Zhu H S, Zhang Z Y, Liang D K, Xie X D, Liu Y L, et al. 2025. SemiETS: Integrating Spatial and Content Consistencies for Semi-Supervised End-to-end Text Spotting//Proceedings of the Computer Vision and Pattern Recognition Conference. Piscataway: IEEE: 9329 - 9338.

Ouyang L K, Qu Y, Zhou H B, Zhu J W, Zhang R, Lin Q S, et al. 2025. OmniDocBench: Benchmarking diverse PDF document parsing with comprehensive annotations//Proceedings of the Computer Vision and Pattern Recognition Conference. Piscataway: IEEE: 24838 - 24848.

Qiao L, Tang S L, Cheng Z Z, Xu Y L, Niu Y, Pu S L, et al. 2020. Text Perceptron: Towards end-to-end arbitrary-shaped text spotting//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 34(07): 11899 - 11907.

Qiao L, Chen Y, Cheng Z Z, Xu Y L, Niu Y, Pu S L, et al. 2021. MANGO: A mask attention guided one-stage scene text spotter//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 35(3): 2467 - 2476.

Qin X G, Zhou Y, Guo Y H, Wu D Y, Tian Z H, Jiang N, et al. 2021. Mask is all you need: Rethinking Mask R-CNN for dense and arbitrary-shaped scene text detection//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM: 414 - 423.

Qu Y D, Wang Y X, Zhou B B, Wang Z X, Xie H T and Zhang Y D. 2024. Boosting semi-supervised scene text recognition via viewing and summarizing. Advances in Neural Information Processing Systems, 37: 105503 - 105527.

Shu Y, Lin H G, Liu Y X, Zhang Y, Zeng G Y, Li Y, et al. 2025. When Semantics Mislead Vision: Mitigating Large Multimodal Models Hallucinations in Scene Text Spotting and Understanding [EB/OL].[2026-02-06]. https://arxiv.org/pdf/2506.05551.pdf.

Team H V, Lyu P Y, Wan X Y, Li G L, Peng S P, Wang W N, et al. 2025. HunyuanOCR technical report[EB/OL]. [2026-02-06]. https://arxiv.org/pdf/2511.19575.pdf.

Wang B, Xu C, Zhao X M, Ouyang L K, Wu F, Zhao Z Y, et al. 2024. MinerU: An open-source solution for precise document content extraction[EB/OL].[2026-02-06]. https://arxiv.org/pdf/2409.18839.pdf.

Wang B, Wu F, Ouyang L K, Gu Z C, Zhang R, Xia R Q, et al. 2025. Image Over Text: Transforming Formula Recognition Evaluation with Character Detection Matching//Proceedings of the Computer Vision and Pattern Recognition Conference. Piscataway: IEEE: 19681 - 19690.

Wei H R, Liu C L, Chen J Y, Wang J, Kong L Y, Xu Y M, et al. 2024. General OCR theory: Towards OCR-2.0 via a unified end-to-end model[EB/OL]. [2026-02-06]. https://arxiv.org/pdf/2409.01704.pdf.

Wei H R, Sun Y F and Li Y K. 2025a. DeepSeek-OCR: Contexts optical compression[EB/OL]. [2026-02-06]. https://arxiv.org/pdf/2510.18234.pdf.

Wei M Z, Yang C, Liang M, Zhou F, Zhu X B and Yin X C. 2025b. T-LLaVA: An Effective Saliency-Aware Slicing Strategy for Text Recognition//International Conference on Document Analysis and Recognition. Cham: Springer Nature Switzerland: 351 - 369.

Wu J J, Hu J S, Chen M J, Dai L R, Niu X J and Wang N. 2022. Structural string decoder for handwritten mathematical expression recognition//2022 26th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE: 3246 - 3251.

Wu J W, Yin F, Zhang Y M, Zhang X Y and Liu C L. 2021. Graph-to-graph: towards accurate and interpretable online handwritten mathematical expression recognition//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 35(4): 2925 - 2933.

Yang M K, Liao M H, Lu P, Wang J, Zhu S G, Luo H L, et al. 2022. Reading and writing: Discriminative and generative modeling for self-supervised text recognition//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM: 4214 - 4223.

Yu M M, Zhang H, Yin F and Liu C L. 2024. An approach for handwritten Chinese text recognition unifying character segmentation and recognition. Pattern Recognition, 151: 110373 [DOI: 10.1016/j.patcog.2024.110373]

Zhang J Y, Zhang Q T, Wang B, Ouyang L K, Wen Z C, Li Y, et al. 2025a. OCR hinders RAG: Evaluating the cascading impact of OCR on retrieval-augmented generation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE: 17443 - 17453.

Zhang J R, Liu Y L, Wu Z J, Pang G S, Ye Z L, Zhong Y P, et al. 2025b. MonkeyOCR v1.5 Technical Report: Unlocking Robust Document Parsing for Complex Patterns[EB/OL]. [2026-02-06]. https://arxiv.org/pdf/2511.10390.pdf.

Zhao W C, Feng H, Liu Q, Tang J Q, Wu B H, Liao L, et al. 2024a. TabPedia: Towards comprehensive visual table understanding with concept synergy. Advances in Neural Information Processing Systems, 37: 7185 - 7212.

Zhao Z Y, Kang H R, Wang B and He C H. 2024b. DocLayout-YOLO: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception[EB/OL]. [2026-02-06]. https://arxiv.org/pdf/2410.12628.pdf.

作者简介

周宇，男，南开大学计算机学院 & 密码与网络空间安全学院教授，主要研究方向为计算机视觉、多模态人工智能、具身智能、大模型等。E-mail：yzhou@nankai.edu.cn

彭良瑞，女，清华大学电子工程系副研究员，主要研究方向为机器学习与计算机视觉、智能图文信息处理。E-mail：penglr@tsinghua.edu.cn

陈善雄，男，西南大学计算机与信息科学学院教授，主要研究方向为图像文档处理，古籍数字化保护。E-mail：csxpml@163.com

连宙辉，男，北京大学王选计算机研究所长聘副教授，主要研究方向为计算机图形学、文字图形图像生成、三维视觉。E-mail：lianzhouhui@pku.edu.cn

高良才，男，北京大学王选计算机研究所副教授，主要研究方向为模式识别。E-mail：glc@pku.edu.cn

殷绪成，男，北京科技大学计算机与通信工程学院教授，主要研究方向为模式识别与计算机视觉、文字识别、工业智能技术及应用。 E-mail：xuchengyin@ustb.edu.cn