arXiv:2302.11583v1 [cs.DL] 22 Feb 2023

高度局部化图及图注的历史天体物理学文献数字化

Jill P. Naiman[1, 2*]、Peter K. G. Williams[3] 和 Alyssa Goodman[3]

[1*]信息科学学院，伊利诺伊大学厄巴纳-香槟分校，Daniel ST，尚佩恩，61820，伊利诺伊州，美国。[2]国家超级计算应用中心，伊利诺伊大学厄巴纳-香槟分校，W Clark ST，厄巴纳，61801，伊利诺伊州，美国。[3]哈佛-史密森天体物理中心，GardenST，剑桥，02138，马萨诸塞州，美国。

*通讯作者。电子邮箱：jnaiman@illinois.edu；贡献作者：pwilliams@cfa.harvard.edu；agoodman@cfa.harvard.edu；

摘要

20世纪90年代末“数字化时代”之前发表的科学文章中，图往往“困”在扫描页面里。尽管提取图及其图注方面已有进展，但目前仍没有稳健的方法来完成这一流程。我们提出了一种基于YOLO的方法，可用于经过光学字符识别（OCR）处理的扫描页面，并同时利用灰度信息和OCR特征。我们致力于将目标检测领域的交并比（IOU）指标引入文档版面分析，并将“高精度定位”量化为IOU达到0.9。将该方法应用于NASA天体物理学数据系统（ADS）收录的天体物理学文献后，我们发现，在IOU阈值为0.9时，图和图注的F1分数分别为90.9%和92.2%，显著优于其他最先进的方法。

关键词：学术文档处理、文档版面分析、天文学。

1 引言

随着数据集规模不断扩大，科学出版物的发布速度也持续加快，科学家需要使用自动化方法来解析这些不断增长的数据产品，其中也包括学术文献本身。易于访问且经过妥善整理的数据产品不仅是开放科学的重要组成部分 [1, 2], ，也受到大多数主要科学期刊的强烈鼓励，并被纳入投稿流程 [3]。然而，以图、表格和公式形式存在于学术文献中的数据产品，尤其是 ∼1997之前发表的“数字化前”时代的文献，无法用于整理，除非开发出能够提取这些重要信息的方法。

文章中不同版面元素的提取是科学数据整理的重要组成部分，过去几年中，表格、图及其图注等元素的提取准确率显著提高 [4–7]。文档版面分析中的一个重要研究领域是对 PDF 文件进行“挖掘”，因为较新的 PDF 文件通常采用“矢量”格式——文档由一组指令渲染而成，而不是

与栅格格式中的逐像素处理不同，理论上可以解析这组指令，以确定图、图注和表格的位置 [8–10]。

然而，这种解析并非易事，人们已经开发出许多方法来完成这一任务。如果 PDF 的矢量格式结构良好，那么就可以通过解析这种已知的 PDF 格式来提取文本和图像。现有一些软件包可以从此类 PDF 文件中输出文本和/或图像 [11]。提取此类信息后，还可以使用多种方法，将原始图和文本整理成图—图注对、表格及其他布局组件（例如章节标题、数学公式）。从历史上看，其中一些最常用的是启发式方法：根据关键词（如“Fig.”或“Figure”）将文本块分类为图注或表格图注 [12, 13]。

近年来，深度学习方法在矢量和栅格文档中广受欢迎 [14–16]，，其中包括使用语义分割方法的方案 [17] 以及目标检测方法 [18]。当有矢量 PDF 可用时，这些深度学习方法通常会与启发式方法结合，在文档版面分析过程中提取文本 [15]。虽然这些方法对于从近期学术文献中提取数据产品至关重要，但数字化前文献通常会以数字平台中的扫描文档形式收录，其中较早的文章扫描分辨率各不相同；而且，使用较新文章训练集开发的深度学习方法在处理这些数字化前文献时往往表现不佳 [19]。此外，与“原生数字化”的科学文献相比，历史文档的布局、字体和文章样式通常有所不同[19]。在这些情况下，必须使用光学字符识别（OCR）进行文本提取，并从原始 OCR 结果中提取图和表格。应用于由 OCR 生成文本的栅格 PDF 时，使用较新文档或基于矢量 PDF 训练的深度学习文档版面分析方法通常不够稳健 [17, 19]。尽管针对 OCR 页面改进这些方法的工作已经取得进展，尤其是在电子学位论文（ETD）方面 [19]，，但从这些较早的栅格文档中提取版面元素仍有大量工作可做。

大型的基于栅格图像的“基准数据集”确实存在，但其中往往大多数都是由较新的文章构成。例如，广泛使用的 PubLayNet 数据集 [20] 中，1997年之前的文章仅约占 2.6%；而专注于历史科学文章的基准数据集则较难获取[19]。此外，不同数据集对版面元素（图、表格及其图注）的定义可能有所不同 [4, 20]，，而大型手工标注基准数据集中的版面元素定义也可能存在不一致之处 [21]。

下文将介绍一种新方法，用于从美国国家航空航天局天体物理数据系统（ADS）[1]中数字化前的科学文献馆藏数据集中提取图和图注。该数据集同时包含基于矢量和栅格的 PDF 文件。我们将深度学习目标检测方法与启发式技术相结合，应用于文章页面扫描件以及通过 Tesseract OCR 引擎 [22]处理扫描件生成的文本特征，并在后处理步骤中，将从矢量 PDF 文件中提取图注的结果与 pdffigures2[13] 得到的结果合并。

虽然我们的模型着眼于天文学文献的数字化——天文学是最早的“大数据”科学之一 [23, 24]——但由于我们的方法高度依赖通过光学字符识别生成的特征，因此也可扩展应用于其他已完成光学字符识别的大型科学文献馆藏 [，例如 HathiTrust 美国联邦文献馆藏[2], 25]。此外，我们的处理流程设计很大程度上受到数据（天文学文献）和模型预期用户（科学家和数字图书馆员）的影响。因此，我们采用开源软件（例如 Tesseract）以及两个群体都使用的编程语言（例如 Python）。本文的结构考虑到了这种广泛的潜在用途： 第2节详细介绍我们的数据集，并概述处理流程的设计考量；第4节讨论模型架构和准确性，尤其是第4.1节讨论我们的方法对其他“基准”数据集的可推广性。这一点与其他领域以及更广泛的文档布局

[1]https://ui.adsabs.harvard.edu/[2] https://babel.hathitrust.org/cgi/mb?a=listis&c=2062901859

分析领域相关的内容，是我们对提取指标之间关系的讨论。在第4.1节中，我们探讨如何使用流行的目标检测指标交并比（IOU），量化从局部页面对象中提取的信息。我们在第5节概述了这项工作的未来计划。所有代码均可在 GitHub 上获取[3]。

2 设计考量与数据预处理

2.1 数据

本研究所使用的数据集是NASA天体物理学数据系统（ADS）所归档的英文文献子集，涵盖“数字化时代”之前的文章——出版年份 $\lesssim$ 1997——如图1顶部面板所示。这些文章涵盖了《天文学杂志》（AJ）、《天体物理学杂志》（ApJ）以及《天体物理学杂志增刊》（ApJS）自1852年至1997年的出版物（图1中部面板）。此外，该数据集也是天体物理数据系统全天巡天（ADSASS）所收录文献的子集。ADSASS旨在根据文本中所研究的对象，将每篇文章与其在天空中的位置相关联 [26, 27]。本研究选取了被认为包含天空图像的文章，依据是对页面颜色分布进行启发式判定 [26]。

图1的下方图版展示了我们较大的数据库子集中，标注为图、图注、表格和数学公式的部分的分布。这些已标注页面是从通过 [26]的启发式方法找到的所有文章中随机选取的。在这项工作中，我们将重点关注图和图注的提取；不过，随本文提供的可下载数据中也包含表格和数学公式的标注，因为文档版面分析领域的研究者通常也会关注这些元素 [20, 21, 28]。我们在第 2.4 节中更详细地讨论了标注流程。

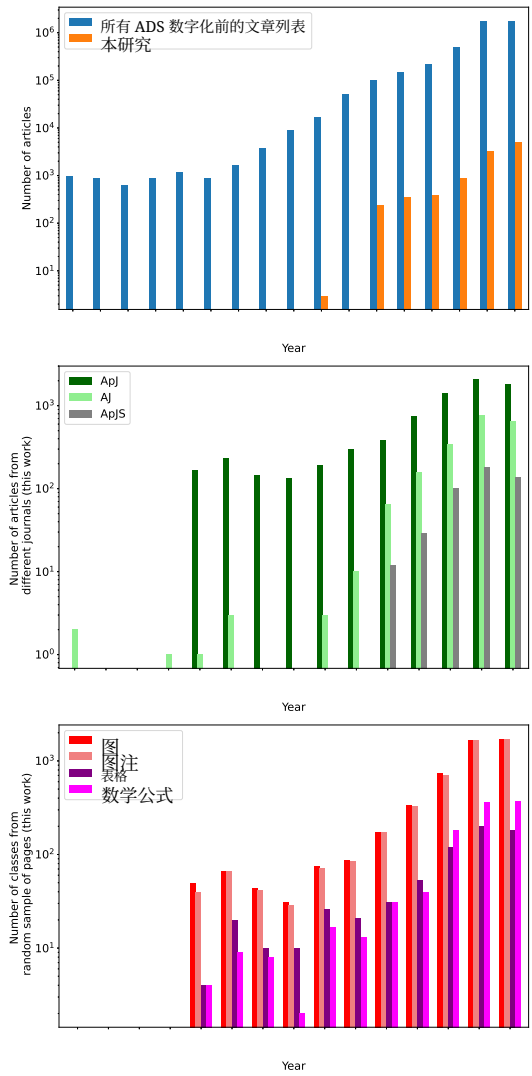

图 1：上方面板显示了所有数字化前的ADS条目以及构成我们已标注数据库的子集的年代分布。中间面板显示了整个数据库按时间和出版商划分的分布，共计 ~10k篇文章（这三家出版商的数字化前ADS条目总数为 ~56k篇）。下方面板显示了完整数据库文章中抽取的主要已标注子集页面里，每年图、图注、数学公式和表格的分布。共计5515页的已标注子集中包含：5010个图、4925个图注[4]，1040 数学公式、677个表格。另有553页不含其他页面对象。

[3]所有项目详情和完整数据：https://github.com/ReadingTimeMachine

[4]有若干整页图没有图注，因此图的数量略多于图注。

本文涉及的文章包括采用矢量 PDF 格式（因此可能可由“PDF 挖掘工具”解析）的文章，以及采用栅格 PDF 格式的文章；其中大多数文章采用栅格 PDF 格式，而采用矢量 PDF 格式的文章所占比例随着文章年代变早而降低。

在此，我们将“PDF挖掘”定义为：对用于构建矢量 PDF 文档的一组“指令”进行逆向工程，以找到文本、表格和图等页面对象的位置和内容 [11, 13, 29–33]。除了这些纯启发式方法外，较新的方法通常还会结合启发式方法，在栅格化页面图像上运用深度学习来定位和提取页面对象；与单独使用纯启发式方法相比，这种方法通常能获得更高的精确度 [例如 15]。然而，由于这些较新方法依赖启发式方法进行文本提取，在无法访问矢量 PDF 中可解析的指令时，它们还不足以准确提取页面对象，因此已被证明无法准确处理历史文档 [19, 34]。为简化分析，以下内容仅讨论基于纯启发式方法的 PDF 挖掘软件，因为这些工具的可解析性能够估算我们的文章中有多少是以矢量 PDF 格式的一组指令编码的。

要确定一篇文章能否通过 PDF 挖掘工具解析，需要仔细检查每个 PDF 页面，使用挖掘软件解析输出，并量化正确解析的单词、图和表格数量。例如，对几个测试页面使用 PDF 图像提取工具 pdfimages[5] ，通常会得到损坏的输出图像文件。然而，要将这一流程扩展至完整数据集中的所有文章，需要耗费大量时间和计算资源。因此，我们必须通过其他方式估算 PDF文件的可解析性。

作为对可解析性的初步估计，我们将 PDF 挖掘软件 GROBID [11] 和 pdffig- ures2 [13] 应用于与我们手动标注页面相关联的文章，并使用两个指标计算其“可解析性”。我们检查解析后的文章输出，其中图号和表号均从 1 开始，并以 1 为增量均匀递增至最大图号和表号。如果这些指标不成立，很可能是因为缺少某个图或表，或者某个图或表被重复计数。对于这些估计，我们选择 GROBID 和 pdffigures2，是因为它们是广泛使用的 PDF 挖掘软件 [13, 35] ，并且经常用于从科学文献 [36–38]中专门提取文本和页面对象。

在本分析中，我们假设编号系统采用整数（例如 Figure 5 或 Fig. 5）或罗马数字（例如 Table IV 或 Tab. IV）。在 pdffigures2 中被标记为表格或图、但不遵循这些编号惯例的对象，占所有对象的 0.7%。在 GROBID 中，这些非标准编号系统占所有对象的 1.3%。我们还假设，同一页面对象不会混用这些编号系统，但同一篇文章中可能同时使用这两种系统，因为表格通常用罗马数字编号，而图则用整数编号。因此，若整数或罗马数字系统能够生成从 1 开始且单调递增的图号和表号，我们就将该 PDF 视为可解析。

这些估算结果如图2所示。GROBID 和 pdffigures2在 ≈1945-1990 年间对文章的解析效果最好，如图2上图所示，其中GROBID 在这段时间内表现更佳。总体而言，以表格衡量的可解析性更高，这是意料之中的，因为图通常可能用“Fig”或“Figure”以外的词语标注（例如该数据集中的“图版”），而表格极少使用“Table”或“Tab”以外的词语标注。

GROBID在图表解析方面的能力于≈1980-1990 年间达到顶峰（图2中图），表格解析能力也于此时达到顶峰（图2下图），这很可能是由于这一时期《天文学杂志》（AJ）的文章数量增加所致（见图1中图）；这些文章的可解析性似乎与另外两种期刊格式相当，或略高一些。

对所有文章进行汇总后，使用整数或罗马数字时，pdffigures2 对图（表格）的可解析性为 ≈0.7%（1.1%）。对于 GROBID，图（表格）的这一百分比升至 ≈9.0%（33.7%）。我们的估算未考虑任何错误发现的图或表格。将假阳性纳入我们的可解析性指标，可能会

[5] https://www.xpdfreader.com/pdfimages-man.html

降低此处报告的这些工具的准确率。除了未考虑假阳性的影响外，我们的准确率还可能进一步偏低，因为我们的估算未对提取文本的正确性进行任何检查。

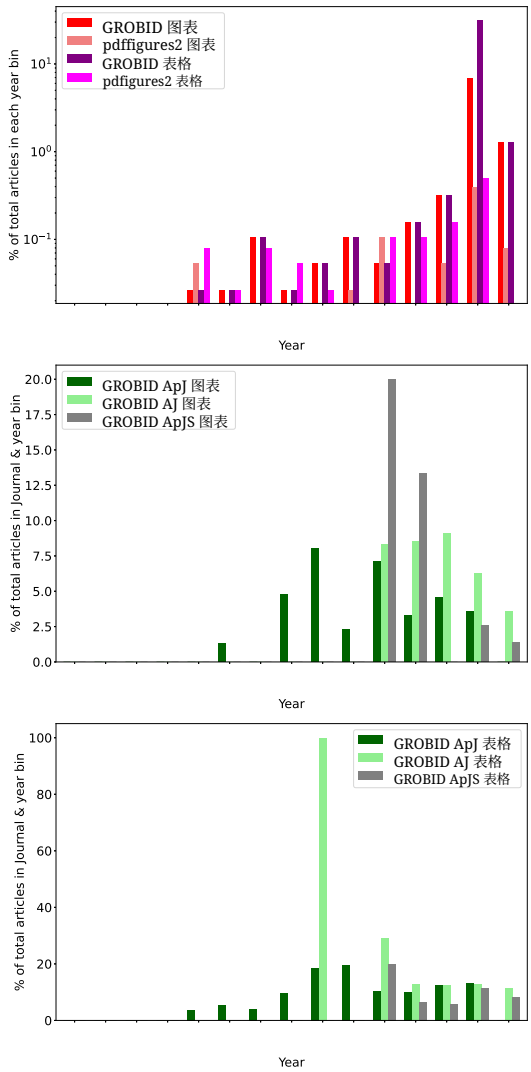

图 2：PDF“可解析性”通过 PDF 挖掘工具 GROBID 和 pdffigures2 进行估算，以整数或罗马数字形式表示，为每个 10 年区间内文章总数的百分比。如果按顺序排列的图（表格）列表从图（表格）编号“1”开始，且编号逐一递增，则认为该文章可解析。我们估算数据集中所有文章的可解析性为 ≈1-33%（详情请参阅正文）。

最后，我们在此指出，除本文测试的三种工具外，还有许多其他 PDF 解析器（pdfimages、GROBID、pdffigures2，我们重点关注后两者），或许能进一步提升可解析性 [29–33]。

2.2 流程开发

我们构建这一数据集的最终目标，是将图-图题对托管在天文学图像浏览器（AIE）数据库[6]中。目前，AIE 会自动收录一部分原生数字化的 ADS 馆藏内容——即美国天文学会期刊（AAS）中的文章——及其图-图题对。因此，此处介绍的处理流程从对页面进行初步光学字符识别开始，最终通过识别图周围的区域以及图注区域中的 OCR 识别文字，提取图及其图注，并将其托管在 AIE 等平台上。

由于本研究的受众可能由科学家和数字图书馆员共同组成，我们将工作重点放在开发基于 Python 的处理和训练处理流程上，因为这种语言在这两类人群中具有很高的通用性。图3 展示了我们完整处理流程的概要——从数据生成、标注，到训练深度学习模型，最后对模型结果进行后处理。

2.3 光学字符识别与图像处理

我们的深度学习模型使用光学字符识别特征进行训练（详见第3节了解模型详情），这些特征是通过使用Tesseract的光学字符识别引擎处理每个页面生成的 [22]。我们使用Tesseract参数来查找页面上的所有英文（lang=eng）文本，不假设行或段落的任何格式（psm=12）、页面旋转方向（OSD，psm=12），并使用长短期记忆网络光学字符识别引擎处理页面（oem=1）。这些参数使Tesseract能够查找页面上的所有文本，而不受文本位于段落中还是图像上的影响。Tesseract 不会明确定位或标记公式或图像。这些参数使我们能够准确地找到页面上的大多数文本边界框；然而，生成的光学字符识别结果

[6] http://www.astroexplorer.org/

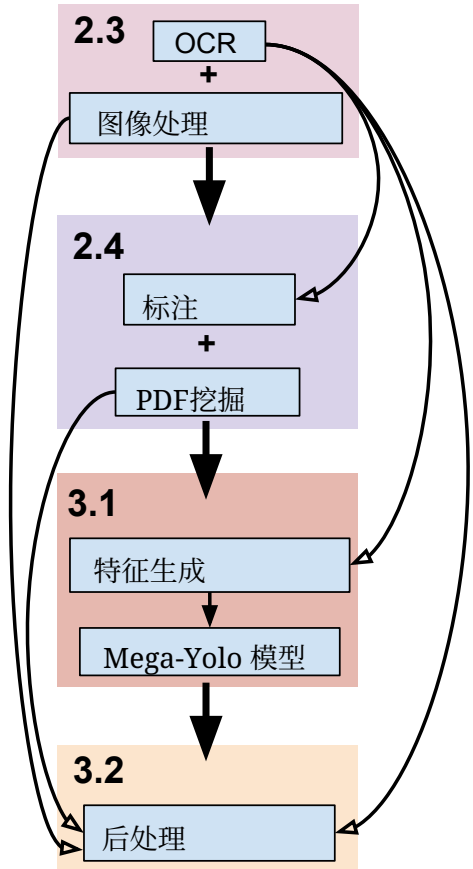

图 3：我们的整体处理流程分为四个主要步骤。光学字符识别和图像处理将在第2.3节中讨论。类别定义和“代码本”，以及标注和 PDF 挖掘将在第 2.4 节中讨论。特征选择和深度学习模型的描述见第3.1节，后处理技术将在第 3.2 节中讨论。（更详细的步骤分解请参见附录6 和图10 。）

文本可能含有噪声，且不包含非英语字符。在文档版面分析领域，清理含噪、混合语言的光学字符识别文本是一项持续深入的研究 [39–42] ，超出了本文的范围，但将作为未来的研究方向 [43, 44]。

每个随机选取的 PDF 页面都会被处理为 TIFF格式（暂存以供光学字符识别）和 JPEG格式（稍后在处理流程中用于提取灰度特征）。原始 PDF文章以高分辨率 TIFF扫描件的形式存储于 ADS 中 [45]。

由于 PDF文章保留了原始扫描件的分辨率，且 ADS 更充分支持批量下载，因此本研究采用这些文章格式。我们从文章 PDF 中以高分辨率（每英寸点数=600）提取 TIFF 和 JPEG，并将生成的图像长宽均缩小一半，以避免像素化，且不实施抗锯齿。证据表明，TIFF 图像用于光学字符识别时比其他图像格式产生的错误更少 [46]。临时 TIFF 图像通过 Python封装器[7] 传入 Tesseract，并利用 Tesseract 针对整页光学字符识别的优化功能。输出以 hOCR格式存储。在此步骤中，使用 OpenCV [47] 及其关联的 Python封装器[8] 进行基本图像预处理（阈值处理和高斯滤波），以处理页面上的各种瑕疵。

除了深度学习模型，我们还使用图像处理技术，通过启发式方法查找潜在的图框。利用光学字符识别所得单词的位置遮蔽文本，再通过基于 OpenCV 构建的基础形状查找器处理修改后的页面。该查找器经过调校，可识别矩形（具有四个角，且边由大致平行的线段构成）。我们将这个“矩形查找器”应用于页面的多个滤波版本（方向梯度直方图、膨胀、颜色反转以及不同级别的阈值处理）。随后，根据方形角点的位置对矩形列表进行 K均值聚类，并剔除尺寸较小的伪影矩形，同时根据纵横比排除可能是色条而非图的矩形。

对页面进行光学字符识别并使用 OpenCV 查找形状，每页大约需要 20–25 秒（在配备 64 Gb 随机存取存储器的 Apple M1 Max 六核上测试）。

2.4 标注和类别定义

在深入探讨我们的深度学习模型的细节之前，我们先介绍标注流程的几个方面，以便清楚地

[7] https://pypi.org/project/pytesseract/[8] http://pypi.org/project/opencv-python/

理解我们的模型试图在每个文章页面上定位什么。

我们首先定义图和图注的类别，因为文献中经常存在分歧，同一数据集的标注者之间偶尔也会有分歧 [4, 21]。如图 4所示，我们将图定义为单个页面上一个或多个分面的集合；在科学文章中，这些分面会被称为一个图（即“图 3”）。这不同于其他研究，后者通常将每个“子图”视为单独的图 [21]。此外，图的定义包括所有坐标轴标签和标题。当图跨越多个页面时（通常会通过“图 1b”或“图 5，续”之类的图注加以标示），每个图所在的页面都会被分类为单独的图。如果图注在水平方向上超出其对应的图，则将该图水平延伸至图注的边缘（见图 6中的品红色线条）。这些定义保持了与其他定义的一致性（例如 [20]），同时通过不重叠的框定义图和图注区域。最后，除非图注位置异常，图边界框不包括图注，这与其他定义相反 [，例如 4, 19, 20],由于我们的目标之一是提取图-图题对，因此必须在文章页面上区分这两种不同的对象。

图注定义为与这些图对象相关联的说明文字。虽然许多图没有图注，但除少数情况外，大多数图注都与其对应的图位于同一文章页面上。当可能存在多个图注时（例如，较长的图注旁还有类似“图 4b”的子图图注），如果较长的图注与该图位于同一页面上，我们就将其选作与该图关联的图注。如果某个页面上的图只有子图图注，我们就将该子图图注定义为该图的图注。

使用 MakeSense.ai [48] 结合 JPEG格式的图像对文档布局对象进行人工分类。完成人工标注后，会使用 OCR 结果进行检查。具体而言，会修改图注的边界框，使其涵盖 OCR 词框，因为这些词框通常会发生偏移，或大于原始灰度图像中呈现的文本文章页面。当 OCR 结果较差，未能捕获图注的全部内容时，视觉框与处理后框之间就会产生偏移。例如，如果页面倾斜严重、扫描噪声过大或文本颜色过浅，Tesseract 可能会漏掉图注中的许多词语，从而导致其边界框明显小于标注人员通过视觉检测得到的边界框。即使在理想条件下，单个字母的部分笔画也可能被截断，因此未被纳入 OCR 边界框[9] 不过，由于我们最终将提取 OCR 文本（以便在未来工作中托管于 AIE），因此会忽略这些边缘情况——在我们的标注数据集中仅报告过一个实例。通过在此处以及未来的标注活动中纳入这些带噪实例，我们的训练数据将包含完整文章语料库中 OCR 页面预计会出现的噪声。这是标注流程的关键部分，可确保我们定位感兴趣的文本信息，而不是只在灰度图像上定位一个边界框，后者通常会与 OCR 结果发生偏移。

最后，在此阶段，我们将 PDF 页面传入 PDF 挖掘软件 pdf- figures2 [13] ，以提取矢量 PDF 中的所有图和图注框。由于我们不一定事先知道 $a\ priori$ 哪些页面将以矢量 PDF 格式存储，因此我们对所有页面运行 pdffigures2。找到的图和图注将被存储起来，以便在后处理步骤中与模型的结果结合（参见第 3.2 节中的步骤 3）。

修改后的人工标注和 PDF 挖掘的结果均以 YOLO 标注风格的 XML 文件 [50, 51] 存储。

3 深度学习模型设计与训练

下文将讨论我们开发的一种深度学习模型。该模型高度依赖从栅格 PDF 的 OCR 结果中提取的特征，以利用数据集中此类 PDF 文件占绝大多数这一特点。

[9]这是 OCR 引擎的常见问题，也经常有人在线提问，尤其是针对 Tesseract 的情况，例如：https://stackoverflow.com/questions/57033120/bounding-boxes-around-characters-for-tesseract-4 0 0 1 ---beta-例如。[49]

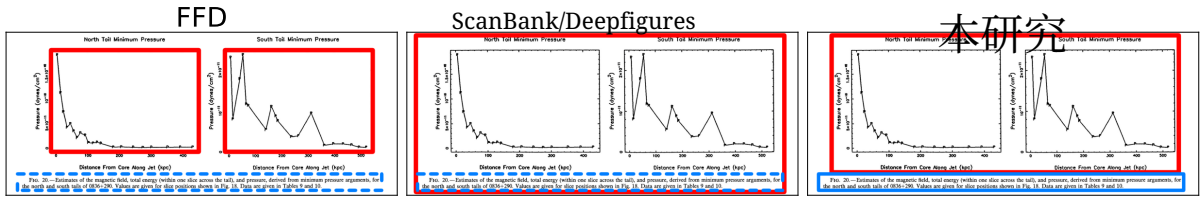

图 4：说明“编码本”如何定义图和图注（右），并与其他研究（左、中）进行比较。例如，有些研究 [例如 21]（左侧红框）通常会将子图拆分为单独的图，而另一些研究 [15, 19, 34]（中间红框）则将图注和图合并为一个“图”框。其他方法中的说明文字框（以蓝色显示）用虚线方框标出，因为实际应用往往只关注图的检测 [例如 34]。请注意，由于各数据集之间没有重叠，其他研究中的框仅用于示意。有关我们与其中若干数据集的类别定义和方法的比较，请参见 第4.2节。

3.1 模型设计与特征选择

现代典型方法依靠深度学习技术来检测页面上的版面元素[14]，，通常结合启发式方法 [15]。这些方法涵盖了使用 YOLO 等模型进行目标检测 [15, 19, 52]，以及近来使用 Faster R-CNN [21, 53–56]和 Mask-CNN [57–59]的方法。此外，还有一些研究提出了利用语义分割 [17] 和全卷积网络 [60, 61]，进行逐像素分割的方法，包括“全卷积实例分割” [62–64]。

这些模型通常会使用从文章页面提取的各种特征作为输入，与未经处理的页面并用，或取而代之。一些较受欢迎的近期方法会运用图像处理和计算机视觉技术 [21]，包括连通分量分析 [65–68]和距离变换 [56]，或它们的某些组合 [69]。

虽然许多方法旨在任何光学字符识别过程开始之前检测页面对象 [例如 6、70–72]，但我们在此实现的方法可在光学字符识别之后应用，旨在支持未来将我们的工作扩展到大型数字档案；这些档案由此前经过光学字符识别处理的文章构成，例如 HathiTrust 中的文章，以及互联网档案馆中的历史文档 [73–75]。下文中，我们采用 TensorFlow 实现的 YOLO-v5[10] [76, 77]，并将工作重点放在特征探索上，利用一组从光学字符识别输出本身提取的特征，目标是选出数量最少的“最佳”特征，并据此构建我们的模型。

除了原始灰度页面，还可以使用多种主要从 Tesseract 的 hOCR 输出中提取的特征。为尽量减少存储空间，每个特征都会缩放为无符号 8 位整数，作为一个“颜色”通道，组成 512x512（像素）的多层图像，并输入能够处理三个以上颜色通道的“mega-YOLO”模型。从 Tesseract 以 hOCR 格式输出的特征包括：

- 字体大小（*fs*）：每个词边界框的字体大小通过减去页面字体大小中位数，再除以标准差进行归一化。字体超出五个标准差的边界框将被忽略。

- 内容区域（c_b）：自动页面分割得到的“内容区域”包含大型文本块，有时也会将图纳入独立的“内容区域”，但并不总是如此。

- 段落（p_b）：自动分割得到的词组，这些词组很可能是段落。通常与“内容区域”重叠。

- 上伸部（*asc*）：根据排版定义，指词中高于大写字母顶部的字母部分所占的量（例如字母“h”的顶部）。上伸部通过减去每页的中位值进行归一化。

- 下伸部 *(dec)*：一种排版元素，指单词中位于字母“底部”以下的部分（例如字母“g”的底部弯钩）。

[10] https://github.com/jahongir7174/YOLOv5-tf
https://github.com/jmpap/YOLOV2-Tensorflow-2.0

下伸部通过减去每个页面的中位数值进行归一化。

- 词语置信度 *(wc)*：每个词的置信度百分比
- 词语旋转角度 *(t_{ang})*：词语以 0°、 180°和 270°为步长进行旋转。

从页面扫描图像和光学字符识别生成的文本中提取的其他特征包括：

- 灰度 *(gs)*：使用页面的亮度将图像转换为灰度。大多数图像本身已经是灰度图，少数彩色图像则会转换为灰度图。
- 词语中字母所占比例 *(%l)*：词语中作为字母的字符所占的百分比（缩放至 125–255，以便为扫描页面中不包含任何词语的空格保留“真正的零值”）。
- 词语中数字所占比例 *(%n)*：词语中作为数字的字符所占的百分比（缩放至 125–255）。
- 标点符号 *(p)*：标点符号标记为 250，非标点字符标记为 125（为空的非词语空格保留 0）。
- *spaCy* 词性 *(SP)*：spaCy 的 [78] 19 英语中的“词性”（名词、动词等）条目
- *spaCy*细粒度词性标签 *(ST)*：包含 57 个值的更详细词性标签
- *spaCy* 依存句法分析 *(SD)*：用于指定不同词语之间关系的 51 个“句法依存关系”标签

图 5 展示了单个页面中这些特征的部分示例（灰度（gs）、字体大小（fs和 spaCy 依存句法分析（SD）），以及它们在标注数据集所有页面中的分布。为便于说明，我们保留了图 5 左上分面中的灰度图像，未对其反转；不过，作为特征使用时，灰度图像会进行反转，这在文档布局应用中很常见 [例如 58]。图 5 顶部的分面展示了由不同特征层突出显示的不同区域。灰度（左上分面）让读者能够直观地定位图，而图中的 OCR 词语也以较大的字体大小标示出图所在的区域（顶部中间分面中的红色和黄色块）。尽管这些检测结果往往有误，但这种大字体大小是 Tesseract 在图中识别出的词语的典型特征。此外，紧邻图下方的图注以比页面其余文本（大多为绿色）更小的字体大小（偏蓝色）突出显示。字体大小上这些相当明显的区域差异，并未体现在 spaCy 依存句法分析参数所呈现的模式中（右上分面）。在这里，OCR边界框的颜色在图、图注或页面其他文本所在区域之间，没有呈现任何明显的规律。

这些模式也体现在数据集中所有扫描页面上这三种特征的分布中，如图 5下方分面中的直方图所示。灰度和字体大小这两个示例特征显示，图、图注和页面“其余部分”这三类的分布存在差异：图内的灰度分布更均匀（图 5左下方的直方图），而与图的字体大小分布相比，图注的字体大小分布在较高值处出现峰值（图 5中下方的直方图）。其他特征的趋势较难判断，图 5 右下方的直方图便展示了这一点：对于 spaCy 依存句法分析这一特征，图、图注和页面其余部分之间的区别不太明显。

除了可能有助于区分页面区域的类型外，这些特征在形态上也类似于连通分量等计算机视觉特征 [65–68],因此，它们可自然地扩展此类研究（例如参见 [21]中的图 2，并与我们的图 5进行比较）。

在第4节中，我们讨论效果最佳的模型，该模型采用（灰度、上伸部、下伸部、词语置信度、词语中数字所占比例、词语中字母所占比例、标点符号、词语旋转角度和 spaCy 词性）作为输入特征集。

3.2 后处理流程

选定特征并训练模型后，我们将最终检测框与OCR词框和段落框，以及通过启发式方法找到的任何图注和图合并，并在亚像素级别对其进行修改（交并比（IOU）计算的结果会四舍五入到最近的像素，以匹配真实标注框的精度）。

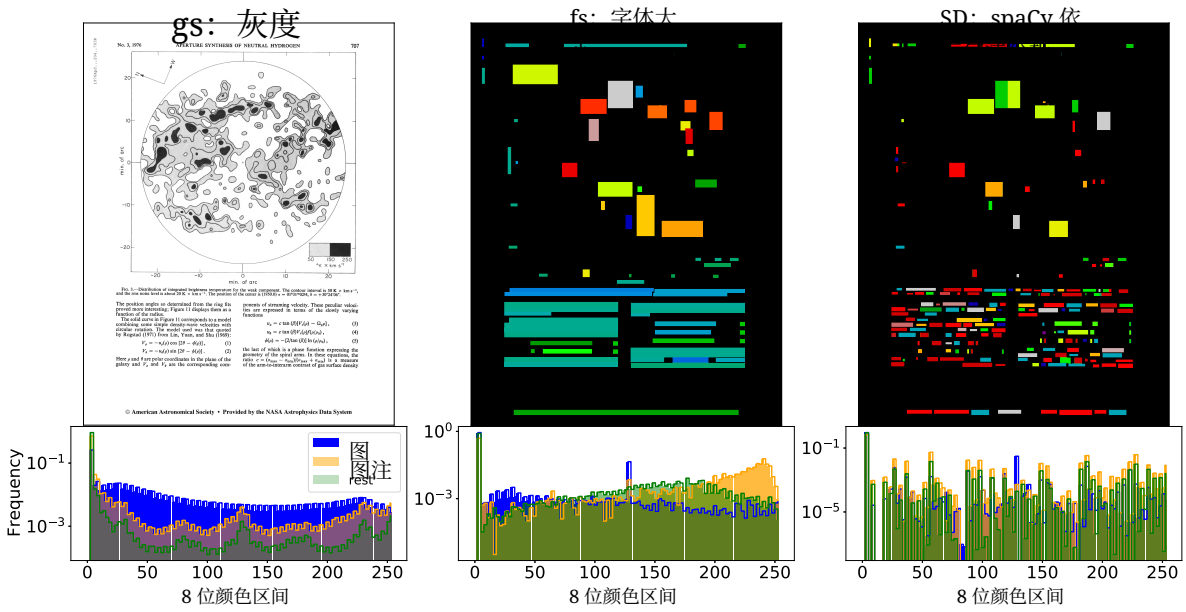

图5：单个页面（上排）中选定特征的示例，以及这些特征在整个标注数据集中的分布（下排）。所有特征均已重新缩放至 8 位颜色分箱（见第3节1）。此处，“其余部分”指页面中不包含图或图注的其余区域。部分特征的分布差异明显（例如灰度和字体大小），而其他特征则没有明显差异（例如 spaCy 依存句法分析）。注意：为便于绘图，我们保留了未反转的灰度图像，但在模型中将其用作特征时，图像会进行反转。

计算以匹配真实标注框的精确率）。

后处理是文档布局分析中的常见做法 [69, 79], ，但不同实现之间往往有所不同，有时也不会纳入最终处理流程 [80]。图 6展示了在处理流程的每个后处理步骤中，检测框和 F1分数的变化情况；在此过程中，我们将真实框与模型在各个后处理步骤中检测到的框进行比较：

- 步骤 1：“原始”检测框是通过非极大值抑制筛选得到的，该方法将多个相互重叠的候选框（即被标记为很可能包含特定类别页面对象的区域）合并为一个边界框 [52, 81–85]。

- 步骤 2：如果两个检测框的交并比大于 ≥ 0.25，则移除得分较低的框，从而减少假阳性（FP）

- 步骤 3：将 pdffigures2 找到的图注框替换为深度学习模型在两者重叠时识别出的图注框，从而提高图注真正例（TP）率，并在较高的交并比（IOU）阈值下减少假正例（FP）和假阴性（FN）。

- 步骤 4：首先对填充后的 OCR 文本框图像应用高斯模糊滤波器，以启发式方式检测图注框。对于图像中与文本框重叠、且文本框通过模糊搜索匹配到“Fig.”、“Figure”和“Plate”等词语的轮廓，将其标记为启发式检测结果。如果一个启发式检测得到的图注框与 mega-YOLO 检测到的框重叠，我们会取启发式框的顶部（通常更准确），并分别取两个框左侧坐标的最小值以及右侧和底部坐标的最大值。这样一来，真正例（TP）总体增加，而假负例（FN）和假阳性（FP）均减少。

- 步骤 5：根据检测框与 OCR 词框和段落框的重叠区域扩展检测到的图注，并允许在水平方向进行多次“增长”迭代。只有当段落框和词框的中心与

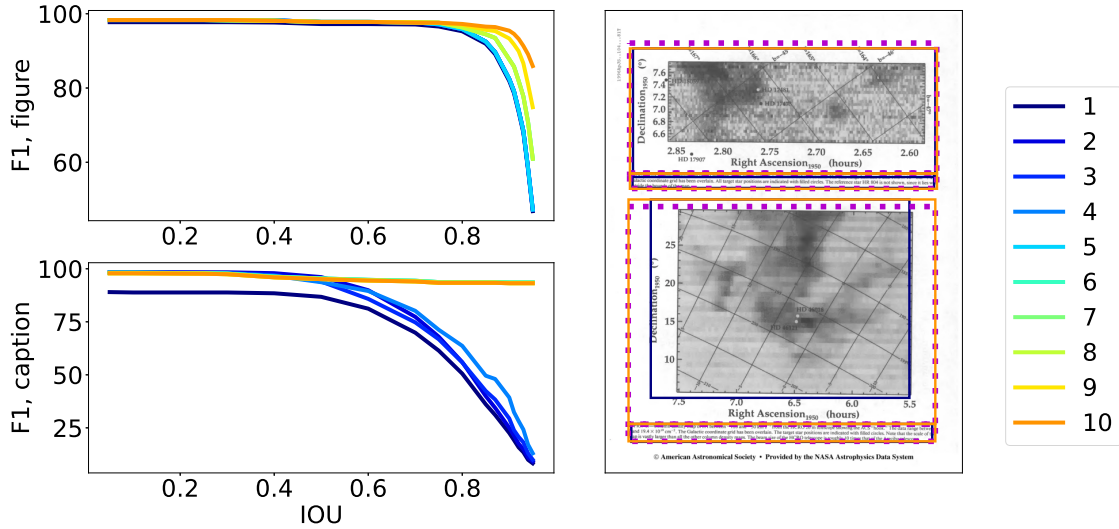

图 6：后处理步骤对模型 12（m12，见 表 1 和 表 2）F1值（左图）的影响。后处理会在较大的交并比值下引起指标变化——对于图和图注，交并比≳0.8 和交并比≳0.6。右图展示了单个页面中初始检测框（步骤 1，深蓝色）和最终检测框（步骤 10，橙色）与真实框（粗品红色虚线）的对比，从而描绘了这些变化。

检测框重叠时，才会根据段落框和词框扩展检测框。

• 步骤 6：如果检测到的图框与通过图像处理获得的矩形重叠（如 第 2.3 节所述），则扩展检测框以包含该图像处理矩形。这会提高图在较大交并比阈值下的真正率。

•步骤 7：丢弃所有面积大于页面面积 75% 的已找到图注，从而使图注的假正例略有减少。

• 步骤 8：通过最小化图注标题中心与特定图底部之间的距离，将图注与图配对。利用页面的旋转信息以及 OCR 词语的整体旋转情况来确定图的“底部”。页面上任何没有关联图的图注都会被删除，从而使假正例减少。

• 步骤 9：将找到的图框向下延伸至其关联图注的顶部，从而提高高交并比阈值下图和图注的真正例数量。

• 步骤10：如果图注的横向延伸范围超出了其对应的图，则将该图横向扩展至图注的边缘

。这会提高高交并比阈值下图的真正率。

步骤 9 和 10 类似于 第 2.4 节中所述的针对标注框的步骤。图 6 所示指标受到的影响适中，主要影响高交并比阈值（交并比≳0.9）下图的结果。

3.3 特征选择消融实验

为了确定能够在最大限度减少特征内存占用的同时生成最准确模型的特征集合，我们进行了一系列消融实验，汇总于表 1 和表 2中。在所有特征选择实验中，我们将 75% 的数据用于训练集，15% 用于验证，10% 用于测试数据集。表 2中的结果展示了该特征选择测试数据集上的表现。

由于测试十四种不同特征的所有组合在计算上代价过高，我们首先采用逐次纳入一组或两组特征的策略，直到模型包含全部十四种特征，如表 1粗横线以上所示。

与其他研究类似 [，例如 17]，，这些参数组合力求遵循一种“直观的”

模型	描述
m1	gs
m2	gs +fs
m3	gs +fs +asc +dec
m4	gs +fs +asc +dec + wc
m5	gs + fs + asc + dec + wc + %n + %l +p
m6	gs + fs + asc + dec + wc + %n + %l +p + t^{ang}
m7	gs + fs + asc + dec + wc + %n + %l +p + t_{ang} + SP
m8	gs + fs + asc + dec + wc + %n + %l +p + t_{ang} + SP + ST + SD
m9	gs + fs + asc + dec + wc + %n + %l +p + t_{ang} + SP + ST + SD + p_b
m10	gs + fs + asc + dec + wc + %n + %l +p + t_{ang} + SP + ST + SD + p_b + c_b
m11	gs +fs +wc + %n + %l +p + t_{ang} + SP
m12	**gs + asc + dec + wc + %n +%l + p** + t_{ang} + **SP**
m13	gs + asc + dec + wc + %n + p + t_{ang} + SP
m14	gs + asc + dec + wc + %n + %l + t_{ang} + SP
m15	gs + asc + dec + wc + %n + %l + p +t^{ang}

表 1：使用第3节中讨论的特征进行的消融实验。所有模型都包含后处理（第 3.2 节）。根据表 2中的指标和第 3.3 节中的讨论确定的“最佳”模型是以粗体突出显示的模型 12（m12）。

复杂度递增的模型构建方式。例如，我们的第一个模型 (m1) 仅使用灰度 (gs) 图像作为特征，简单地模拟了对未经处理的页面应用 YOLO。随后，我们添加“主要”光学字符识别特征，在此将其定义为直接来自光学字符识别引擎、未经额外处理的特征。由于图注和图中坐标轴标签等对象的字体大小 (fs) 通常与文章正文在视觉上有所不同（见 图 5），我们在 m2 中加入这一特征；接着加入上伸部 (asc) 和下伸部 (dec)，因为不同文本对象中常见的字体差异也会导致这些特征发生变化 (m3)。对于图中可能出现的虚假光学字符识别检测结果，其词语置信度 (wc) 可能较低；我们在下一个模型 (m4) 中加入了这一特征。

此处将次级光学字符识别特征定义为从光学字符识别引擎生成的词语中提取的特征。数字、字母和标点符号（分别以 %n、%l 和 p 表示）在 OCR 词语中的比例，通常会因图注或正文中的词语与坐标轴标签上或图内找到的 OCR 词语而有所不同；词语的旋转角度（m5 和 m6）也是如此。与原始 OCR 数据相距更远的是根据识别出的 OCR 词语提取的语言特征（m7 至 m9 中的 spaCy SP、ST、SD），最后还有将词语分组为段落和大词块的特征（m10 中的 p_b 和 c_b）。

在这十个模型中，我们选择准确度最高的模型；这里将其定义为：对图及其图注均具有较高的 F1分数，同时保持较低的假阳性分数（假正例）。表 2中的前十个模型里，模型7是“最佳”模型。随后，我们从该模型中按下方所示的组合减去一个或两个特征；这些组合位于表 1中粗横线下方。采用相同的选择标准后，我们选出模型 12 作为整体“最佳”模型。该模型包含以下特征：灰度、上伸部、下伸部、词语置信度、词语中数字所占比例、词语中字母所占比例、标点符号、词语旋转以及 spaCy 词性，如表 2所示。该模型不仅结合了灰度页面，还结合了许多主要的光学字符识别特征；直观而言，这些特征在正文、图和图注区域之间会有所不同。添加 spaCy 词性标注（SP）这一辅助特征，表明不同文本区域之间的词语使用可能存在差异，这与先前的研究结果一致 [17, 86, 87]。

采用的优化器为 Adam，批量大小为 $\beta_1 = 0.937$ $\beta_2 = 0.999$。学习率通过余弦调度器进行调度，该调度器取决于初始学习率、训练轮次和批量大小。实际上，将其应用于我们的模型时，学习率会在

初始训练轮次内线性增加至原来的 ∼1.6 倍（之后保持不变）。我们从一组较小的学习率（0.008、0.004、0.0004、0.0002）中选定了最优初始学习率 0.004。所有实验均训练 150 轮，并在此期间收敛（通过验证损失进行跟踪）。未应用数据增强。训练在 Tesla V100-SXM2 GPU 上进行，每轮平均耗时 ∼6.5 分钟。

4 结果

为量化我们的“最佳”模型（模型 12）在未见数据上的结果，我们另外标注了 ≈600页，作为“最终测试数据集”，其中包含≈500 图及图注的真实标注（分别为 490 和 487 个）。包括后处理在内，在配备 64 Gb 随机存取存储器的 Apple M1 Max 单核上，每页评估平均耗时 1.8 秒。该数据集中图和图注在 10 年时间区间内的分布显示在图 7的上方面板中。

我们的模型在该最终测试数据集上的性能指标（采用多个交并比阈表3值）见表3。对于图而言，在所有交并比阈值下，真正例（TP）和假负例（FN）相对稳定，而在图注的交并比=0.8 阈值下，假阳性略有增加，真正例则有所下降。我们通过五折交叉验证，基于复合指标估算误差；为清晰起见，仅显示真正例、假阳性和假负例指标的平均值。

图注和图的 F1 分数随时间（以 10 年为增量）的分布显示在下方分面中，见图 7。与图2所示图中 PDF 可解析性的分布相比，我们的方法在更早的发表时间段下降得较少。

下文将通过与其他方法和数据集进行比较，特别是在高度定位的情况下，对这些结果进行语境化分析。

4.1 高度定位的页面对象——对图形提取的意义

在将我们的模型与其他模型进行比较之前，我们首先定义文档版面分析语境中“高度定位”的含义。

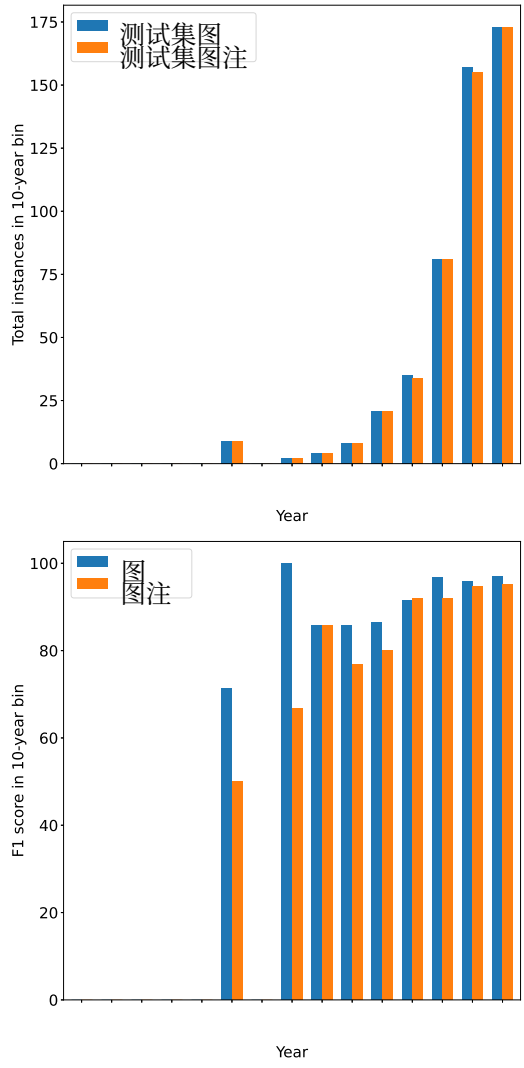

图7：最终测试数据集中图和图注的分布（上方分面），以及每个时间区间内 F1 分数的分布（下方分面），交并比为=0.8 阈值。时间区间的选择与图1下方的两个分面和图2中的所有分面保持一致。不过，没有来自 1903 年之前文章的页面。

以往研究指出，将目标检测（和分割）指标直接用于文档版面分析存在问题[89]，，并且在某些情况下，研究者为文档版面分析及其相关流程专门开发了新指标 [90]。具体到我们的

	TP		FP		FN		Prec		Rec		F1	
	fig	cap	fig	cap	fig	cap	fig	cap	fig	cap	fig	cap
m1	90.3	88.3	11.3	10.0	1.8	4.3	88.9	89.8	98.0	95.4	93.3	92.5
m2	89.5	86.9	10.9	8.6	2.6	6.4	89.2	91.0	97.2	93.2	93.0	92.1
m3	85.9	86.9	17.9	12.5	1.8	5.1	82.8	87.4	97.9	94.4	89.7	90.8
m4	90.5	88.7	10.1	8.8	2.0	3.7	90.0	91.0	97.8	96.0	93.7	93.4
m5	84.5	87.3	15.7	7.2	2.2	7.0	84.3	92.4	97.4	92.6	90.4	92.5
m6	89.5	89.1	11.9	9.0	2.0	4.3	88.3	90.8	97.8	95.4	92.8	93.0
m7	92.8	88.1	8.0	9.0	1.4	4.1	92.0	90.7	98.5	95.6	95.1	93.1
m8	90.5	90.0	9.1	7.2	2.0	3.9	90.9	92.6	97.8	95.9	94.2	94.2
m9	84.3	84.2	13.5	7.4	4.0	9.4	86.2	91.9	95.4	89.9	90.6	90.9
m10	88.7	87.5	11.5	9.6	1.8	4.3	88.6	90.1	98.0	95.3	93.0	92.6
m11	90.5	92.4	10.5	6.8	0.8	1.8	89.6	93.2	99.1	98.0	94.1	95.6
m12	**92.2**	**89.1**	**6.4**	**6.6**	**2.4**	**4.9**	**93.5**	**93.1**	**97.4**	**94.8**	**95.4**	**94.0**
m13	92.8	88.7	7.8	8.4	2.0	4.3	92.2	91.4	97.9	95.4	95.0	93.3
m14	87.3	88.7	15.9	8.4	1.2	6.4	84.6	91.4	98.6	93.3	91.1	92.3
m15	89.9	89.5	8.7	5.9	2.4	5.1	91.2	93.8	97.4	94.6	94.2	94.2

表 2：模型指标，模型详见表 1。特征选择测试数据集中使用了 497 个图（图）和 488 条图注（图注）。图和图注的交并比均为 0.9。真正例、假正例和假负例分别以总实例数的百分比表示。

	交并比=0.1		交并比=0.6		交并比=0.8	
	图	图注	图	图注	图	图注
TP	96.5	92.6	94.3	88.9	93.3	87.5
FP	2.9	2.7	5.1	6.4	6.1	7.8
FN	3.5	6.0	3.5	6.0	3.5	6.0
Prec	97.1±1.5	97.1±1.1	94.7±2.3	93.2±2.3	93.8±3.0	91.8±1.6
Rec	96.5±1.6	94.0±2.7	96.4±1.7	93.8±2.8	96.4±1.0	93.7±1.5
F1	96.7±0.8	95.5±1.2	95.5±1.7	93.4±1.0	95.1±1.6	92.7±1.4

表3：我们的主模型在最终测试数据集上针对不同交并比阈值的指标。测试数据集中有490个图和487个图注。使用COCO[0.5:0.95:0.05] 方案 [88]时，图和图注的平均精度分别为91.5%和87.9%。每个类别中的真正例、假正例和假负例以该类别总实例数的百分比表示。误差根据五折交叉验证的标准差计算。

基于YOLO的应用中，目标检测所用的交并比指标会对所有相交区域进行同等加权，而眼动追踪研究表明，解读图表的关键元素包括图表的x轴和y轴及其标签 [，例如 91]，它们通常位于图的边缘。因此，将目标检测指标（在本例中为交并比测量）用于文档版面分析时，应量化该指标在多大程度上能够反映边界框边缘处这些重要图形元素被遗漏的次数。

虽然我们当前的标注方法不会分别跟踪x轴和y轴标签，但我们通过特定真实标注图形的“超出区域”和“损失面积”来估计这种影响与检测框的交并比。这些区域以左侧分面中某一图的示例表示，见图 8。位于真实标注框（洋红色框）外部且位于检测框（橙色框）内部的像素，其总和用于计算“超出区域”（图中左上图板的绿色阴影区域，见图 8）；而真实标注框内部但检测框外部的所有像素之和则为“损失面积”（图中左下图板的绿色阴影区域，见图 8）。当超出区域占真实标注框面积的 ∼10% 时，部分图注会被纳入图框（绿色阴影区域，左上图板）。尽管这些纳入的信息并非提取图的理想附加内容，但如果将其包含在 AIE 等托管网站或其他

图形分析应用（例如用于研究颜色分布及科学图其他元素的数据集 [92, 93]）中，则不太可能使查看者产生困惑。相比之下，我们通过左下图板中的绿色阴影区域展示了损失面积为 ∼5% 时的显著影响。当真实标注框与检测框之间的偏移量更大时，纵轴标签很可能不会被纳入提取的图中。先前的研究表明，坐标轴标签对于理解图至关重要，因此，排除这些标签将使提取的图对读者而言无法使用 [例如91]。

这些示例为我们选择信息“超出”和图中信息“丢失”的可接受界限提供了依据。鉴于我们的应用场景是在 AIE 平台上托管，我们关注的是尽量减少信息损失，而信息超出则是次要考虑。因此，我们将多余区域的可接受界限设为 10%，损失区域的可接受界限设为 5%。

我们在图 8右侧分面中的联合图里，展示了所有真正例检测框的超出区域和损失区域分布，并将其作为这些配对计算所得交并比的函数。这些分布不包括那些没有对应真实框（或检测框）的检测框（或真实框）。如超出区域（右上）联合图与损失区域（右下）联合图的对比所示，两种分布随交并比变化的形状相似，但损失区域总体上小于超出区域。这并不出乎意料，因为第 3.2 节所述的许多后处理步骤都涉及扩大图周围的框。虽然图 8仅统计图的真实框—检测框配对，但图注的超出区域和损失区域分布也相似，故为简洁起见，此处不再展示。

在图 8右侧面板的水平虚线上，我们展示了所提出的可接受超额面积（10%）和面积损失（5%）阈值估计值。如图 8中联合图顶部和侧边的边际直方图线所示，大多数超额/损失面积分布都落在这些阈值内。

确定这些截断值后，我们计算其中包含这些截断值范围内 90% 数据的最小交并比，以避免纳入交并比极低、但超出区域或缺失区域值仍很小的离群值。对于超出区域和缺失区域，计算结果均为交并比 ~ 0.9 ，如图中竖线所示两个分面以及顶部边际交并比图图 8。我们强调，在此阶段，这些数字仅为估算值；将目标检测指标转换为文档版面分析指标时，恰当的截断值究竟应如何确定，仍有待未来研究。

下文中，我们将 IOU $= 0.9$定义为衡量“高定位精度”的交并比指标，并使用这一阈值来量化与其他模型相比，我们模型的稳健性。

4.2 页面局部对象的基准测试（交并比=0.9）

由于我们方法的最终目标是从扫描页面中提取图及其图注，因此我们评估模型在高定位精度下于该数据集上的表现（采用我们定义的交并比阈值 0.9）。如表4的最后一行和最后一列所示，当交并比为 0.9 时，图检测和图注检测的 F1 分数分别为 90.9%（92.2%）。

为了便于与其他文档版面分析方法进行比较，我们从两个方面评估本方法：一是量化其他模型在我们数据集上的表现，二是比较我们的模型在其他文档版面分析数据集上的表现。

用于比较的数据选自 ScanBank 数据集中的电子学位论文（ETD）以及 PubLayNet 数据集中的文章页面。这两类数据分别代表科学文献的两种常见形式（学位论文和经同行评审的出版物），多年来一直是科学领域数字化工作的常见目标 [13, 19, 34, 58, 71, 94–96]。

电子学位论文代表了科学家研究生阶段（硕士和/或博士）工作的成果，其研究是信息科学领域中一个活跃且不断发展的方向 [97]。ScanBank 数据集中的电子学位论文（与 ScanBank 目标检测模型 [34] 一同发布）收集自 MITDSpacerepository[11]，，但仅限于 1990 年之前发表的论文，以确保它们采用栅格格式，而非矢量格式 [19]。ScanBank 数据集包含来自 70

[11]MIT DSpace 是一个存储库，其中收录了“经同行评审的文章、技术报告、工作论文、学位论文等” https://dspace.mit.edu/

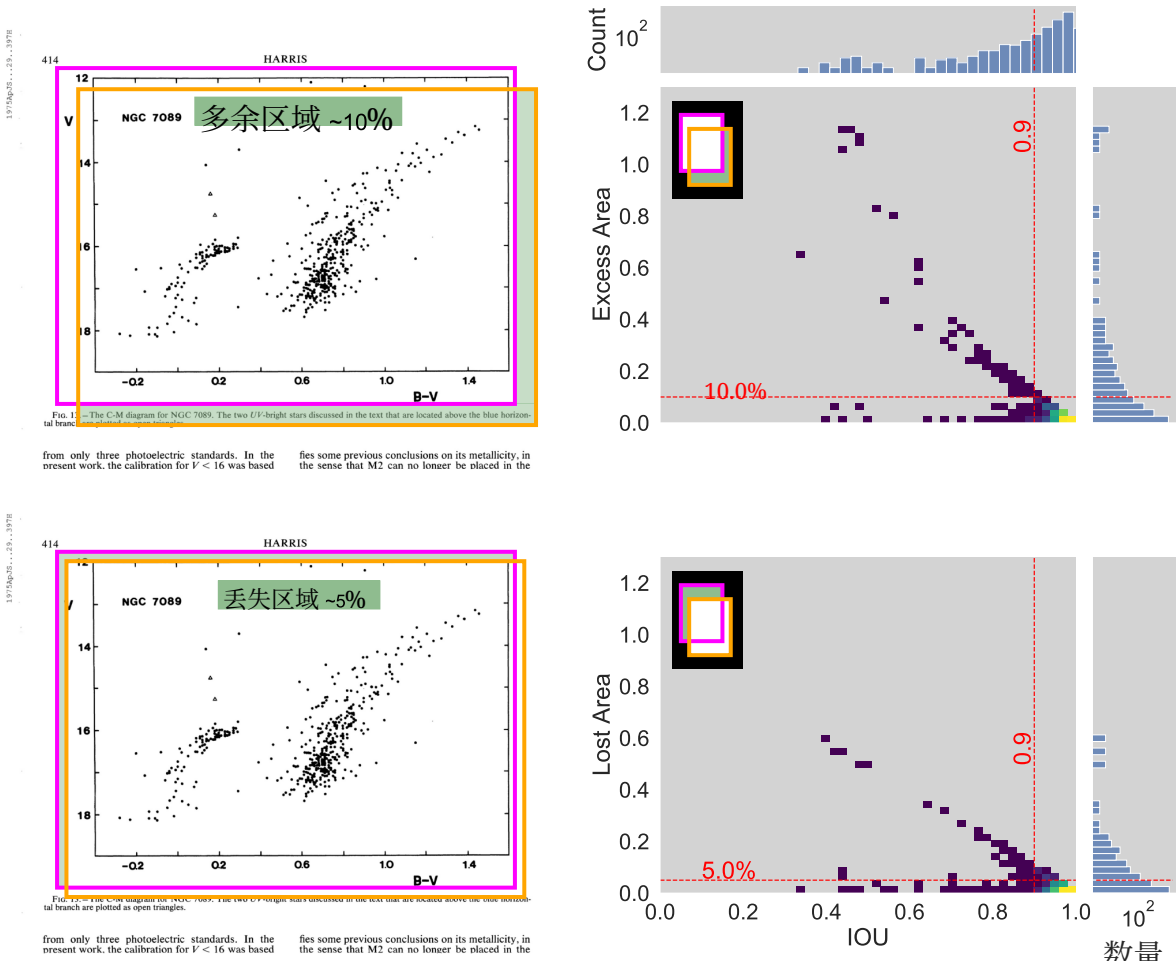

图 8：图中展示了“多余区域”（左上图）和“丢失区域”（左下图）的示例，以及训练集和验证数据集中图的真实框与检测框配对在所有交并比下，超出区域（右上图）和丢失区域（右下图）相对于真实标注框的面积百分比分布。联合图左上部分图示中的绿色阴影区域也说明了这些区域相对于真实框（洋红色）和检测框（橙色）的定义。“高度精确定位”的页面对象所选用的交并比为 0.9（更多详情见正文）。

1900 至 1990 年间的电子学位论文涵盖了 36 个不同领域，其中包括科学、技术、工程和数学领域（如土木/机械工程、物理学、化学）以及其他领域（如人文学科、政治学、现代语言与语言学）。左侧面板展示了该数据集中一篇电子学位论文中包含图及其图注的页面示例，见图 9。与我们的数据相比（右侧面板展示了一个页面示例，见图 9），电子学位论文中的文本间距更宽，字体、行间距和页面间距也更大。尽管电子学位论文可以有许多不同的形式，不同领域的科学出版物通常具有较大的行距和页面间距。

与我们的数据集和 ScanBank 数据集在视觉上的差异相比，PubLayNet 数据集中的一个示例页面 [20] 显示在图 9的中间分面中。PubLayNet 数据集的规模远大于我们的数据集和 ScanBank数据集，包含 ~360k 个页面，这些页面是从 PubMed 上托管的一百多万篇“后数字时代”文章中提取的[12]。除了差异

[12]https://pubmed.ncbi.nlm.nih.gov/

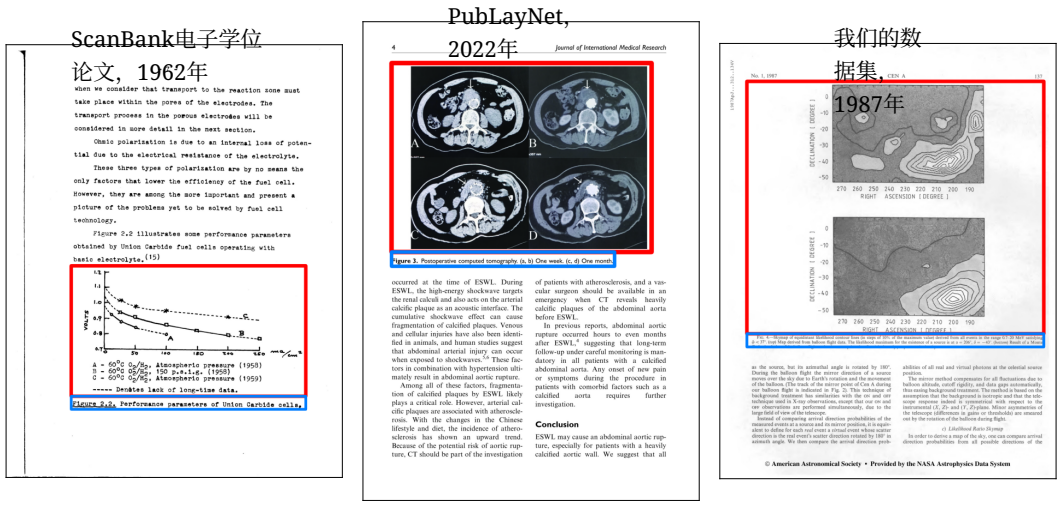

图 9：历史电子学位论文ScanBank 合集中的示例页面 [19] （左侧分面）、PubLayNet 数据集 [20] （中间分面）和我们的数据集（右侧分面）。这些分面展示了这三个数据集之间的差异（字体、行间距、页面尺寸）和相似之处（图表与图注配对及其位置）。这些页面已按照我们对图（红色框）和图注（蓝色框）的定义进行了标注。

在页面尺寸上，PubLayNet 数据集在视觉上与我们的数据集相似，如图 9 中间和右侧分面的对比所示——两个分面中的字体大小、页面布局和行间距都相似。

ScanBank 和 PubLayNet 数据集都包含页面布局标注。ScanBank 数据集包含图、表格及其图注类别，而 PubLayNet 数据集包含文本、标题、列表、表格和图类别。这两个数据集的文章数量差异凸显了训练模型以数字化数字化前文献时持续存在的问题——历史文档布局分析数据集需要人工标注 [ScanBank 19],而较新的数字化后文章则可以从其在线存储格式中挖掘出来 [例如 XML、PubLayNet 20]。

这些数据集在不同方面与我们的数据集有所重叠。ScanBank 的年代分布与我们的数据集相似，但其涵盖的领域和出版渠道更加多样。PubLayNet 数据集中的页面在版式上与我们的数据集更为相似，也和我们的数据集一样，属于科学、技术、工程和数学领域；不过，其中的文章发表时间比我们数据集中的文章更近。

鉴于这些差异与相似之处，接下来我们不仅会比较我们的模型在我们数据集上的表现，还会比较在这些数据集上训练的模型在我们数据集上的表现。

表4 展示了其他深度学习模型在我们的最终测试数据集上的表现。在此，我们使用 Scan-Bank [19, 34]（基于 DeepFigures [15]，该模型是在 ScanBank 电子学位论文上训练的）以及一个版本的detectron2 [80]，该版本是在 PubLayNet 数据集上训练的[20]。

ScanBank和 detectron2 用于比较，因为它们应用于栅格格式文章（而非像pdffigures2 [13] 这样的基于矢量的方法；正如第 2.1 节所讨论的，这会导致它们在我们的数据集上准确率较低。ScanBank 和 ScanBank和 detectron2 对图和图注的定义与我们的并不完全相同，因此，为便于比较，我们做了一些近似处理并作出了一些假设。

如 [34], 中所述，ScanBank 将图定义为包含图注，而我们的图定义不包括图注。为了与我们的结果进行比较，最初在图注与图相关联时，我们通过将真正的图框重新定义为图框与图注框的组合来进行指标计算，

然而，我们发现，如果改用我们对图和图注的定义，使用ScanBank 进行检测所得的指标会达到最优，因此，在与该模型进行所有比较时，我们采用自己对图和图注的定义。由于 detectron2 并不会专门寻找图注框，而是定位通用的“文本”框，因此我们将detectron2检测到的图注定义为中心点最接近已找到图中心的那些框。为了单独测试我们的后处理方法所产生的影响，我们将部分后处理步骤应用于由 ScanBank 和 detectron2 生成的结果，并在表4中将其与采用和未采用后处理的我们的方法进行比较。对这些其他模型的结果应用后处理时，我们仅执行至第 3.2 节 所述的“步骤 5”为止，因为我们发现，这样可以优化 ScanBank 和 detectron2 在表4中报告的指标。

如表4所示，ScanBank 在我们的最终测试数据集上，无论是否进行后处理，表现都不理想。具体而言，ScanBank无法可靠地检测图注，因为假阴性率（FN）较高。此外，如假正率（FP）较高所示，有很大一部分图和图注被错误检测，或定位不准确。这在一定程度上是意料之中的，因为 ETD 格式在视觉上与我们数据集中的文章截然不同，包括字体和图注位置均有差异。应用后处理后，图注指标显著提升，真正例（TP）率增加了 ≈30%，FP 率降低了 ≈20%。

对于未进行后处理的 detectron2 模型，其真正率略高于 ScanBank，但假正率与 ScanBank 相近。其假阴性率比我们的模型（无论是否进行后处理）低几个百分点，这可能部分归因于基于 YOLO 的模型（我们的数据集）与基于 Mask-RCNN 的目标检测模型（detectron2）已知的错误特征差异 [76]。后处理显著提高了图注的真正率，增幅为 ≈35%，并使假正率降低了 ≈40%。应用后处理后，图的真正率也适度提升了 ≈10%。

后处理（使用全部步骤）对模型结果的影响最大——图和图注的真正率分别提高了 ≈25% 和 ≈60%。这并不令人意外，因为我们的后处理方法是基于扫描页面训练数据开发的。此外，我们采用的是基于YOLO的模型，用于检测边界框，而非分割方法；后者在较高的交并比阈值下往往能取得更高的真正率。后处理流程通过调整框的大小，使其更贴合图注文字和图的精确位置，从而“模拟”分割并提高交并比重叠度。

综合来看，表4 的结果表明，在高交并比条件下，其他模型对我们的数据集至多只有中等程度的泛化能力，而且只有应用我们的后处理流程后才能达到这一水平。由于我们的后处理步骤不仅需要灰度扫描页面，还需要其光学字符识别输出，因此这些额外开销（包括执行光学字符识别和后处理步骤）大幅抵消了其他方法在页面处理速度方面的优势。

这种缺乏可推广性的现象是文档版面分析领域一个众所周知的问题，[例如14] 我们的模型也不例外。表5 量化了我们的模型在一组来自ScanBank “金标准”数据集 [19, 34] 的电子学位论文（ETD），以及从 PubLayNet 中选取的非商业文章页面上的表现（选择非商业页面是为了获取高分辨率扫描件并执行我们的方法所需的光学字符识别） [20]。此处，我们展示了模型在这组人工标注的电子学位论文和 PubLayNet 页面上的表现，采用的是我们对图和图注的定义，如图 9左侧（ScanBank）和中间分面（PubLayNet）中的红色（图）框与蓝色（图注）框所示。此外，为便于比较，我们在表5 中，括号内列出了 detectron2（ScanBank）在来自 PubLayNet（电子学位论文）数据集的同一组页面上的结果，并采用了与表5中相同的图注框近似方法表4。

ScanBank 和 detectron2 在各自训练数据集中的图上表现都优于我们的模型，其真正例率分别比我们的检测结果高出 ≈10% 和 ≈30%。假阳性率通常相近（ScanBank）或更高（detectron2 高出 ≈20%），假阴性率也更高。

	ScanBank 无 PP		ScanBank w/PP		detectron2* 无 PP		detectron2* w/PP		Ours 无 PP		Ours w/PP	
	fig	cap	fig	cap	fig	cap†	fig	cap†	fig	cap	fig	cap
TP	69.9	29.0	69.3	52.8	72.0	46.4	81.0	80.9	58.2	23.2	85.7	86.7
FP	71.4	28.8	43.6	8.7	41.8	68.2	27.1	22.4	45.3	82.3	13.7	8.6
FN	1.7	42.8	2.5	40.7	0.6	1.6	1.2	4.9	3.1	5.1	3.5	6.0
Prec	49.5	50.2	61.4	85.9	63.3	40.5	74.9	78.3	56.2	22.0	86.2	90.9
Rec	97.6	40.4	96.5	56.5	99.2	96.6	98.5	94.3	95.0	81.9	96.1	93.6
F1	65.7	44.8	75.0	68.1	77.2	57.1	85.1	85.6	70.6	34.7	90.9	92.2

* 测试版本的detectron2是在 PubLayNet 数据集上训练的 [80]。

† 此处，图注是被分类为“文本”、且最靠近图中心的框。

表4：ScanBank [19, 34]和 detectron2 [80] 在我们的最终测试数据集上的性能指标。交并比为 0.9。TP、FP、FN 占真实实例总数的百分比。为便于比较，列出了采用后处理（“w/PP”）和未采用后处理（“No PP”）的模型。ScanBank 或 detectron2 均未使用我们的数据集进行重新训练或迁移学习。所有指标的五折交叉验证误差为 ~1-2%。

	PubLayNet（非商业）		电子学位论文（ScanBank）	
	图 207	图注† 201	图 197	图注 140
TP	55.1(83.1)	50.2(55.2)	20.8(32.5)	9.3(0.7)
FP	45.4(25.6)	23.4(56.7)	89.1(84.8)	12.4(10.7)
FN	12.1(1.9)	29.4(1.5)	32.1(19.3)	83.4(90.0)
Prec	54.8(76.4)	68.2(49.3)	18.9(27.7)	42.9(6.2)
Rec	82.0(97.7)	63.1(97.4)	39.3(62.7)	10.1(0.8)
F1	65.7(85.8)	65.6(65.5)	25.5(38.4)	16.3(1.4)

† 对于 detectron2，我们假设 假设一个被分类为“文本”的框，它距离检测到的图中心最近，那么它就是该图对应的图注 n.

表5：我们的模型在 Pub- LayNet 数据集的非商业页面以及 ScanBank“金标准”数据集中的电子学位论文页面上的性能指标 [20] （交并比为=0.9）。括号中列出了在这些数据集上训练的模型的指标，以供比较：在 PubLayNet 数据集上训练的 detectron2，以及在电子学位论文数据集上训练的 ScanBank。比较结果应视为初步估计，详情请参阅正文。

使用我们的模型，图注的结果往往相当（ detectron2）或更好（ScanBank）；不过，我们再次提醒，所有数据集对图注和图的定义并不一致，因此这些比较只能视为估算。

尽管这些结果表明，我们的模型或许能更好地推广到图注识别，但还需要在更多数据集上进行测试，才能进一步验证这一结论。然而，作者提醒，任何现有的文档布局分析模型通常都难以实现良好的可推广性 [6, 98]。这种可推广性不足似乎与训练数据集规模无关 [98]。更复杂的是，不同领域对图、图注及其他页面对象的定义可能有所不同，甚至同一领域内不同文档的标注人员也可能存在差异 [21, 99], ，这使得对多样化文档集进行高准确率的“蛮力”人工分类成为一个并非易事的问题 [99]。

5 讨论与未来工作

本文重点研究天文学科学文献中图和图注的定位。我们展示了一个基于 YOLO 的深度学习模型的结果，该模型基于从扫描页面中提取的特征进行训练，hOCR 输出以及Tesseract 光学字符识别引擎 [22],和 spaCy 的文本处理输出 [78]。

由于我们的数据集同时包含矢量 PDF 和栅格 PDF 格式，我们使用 PDF 挖掘软件 pdffigures2 [13]和 GROBID [11]对文章进行处理，以测试其“PDF 可解析性”。我们发现，最多只有 $\approx 33\%$ 的文章可解析，因此采用基于目标检测的方法来查找图和图注。

我们花费了大量精力，精确定义“图”和“图注”类别，以避免与其他数据集 [（例如 21]）中可能存在的分类差异，并与我们的目标保持一致：不仅要定位这些对象，还要从扫描页面中提取它们，使其能够独立于原始文章单独托管。

通过消融实验，我们发现，结合页面属性和 hOCR 属性（灰度、上伸部、下伸部、词语置信度、词语中数字所占比例、词语中字母所占比例、标点符号、词语旋转角度和 spaCy 词性）能使我们的模型在检测图及其图注方面达到最佳性能。将我们的模型与其他常用于文档版面分析的深度学习模型（ScanBank [19, 34]和 detectron2 [80]）进行比较后，我们发现，在我们的数据集上，我们的模型表现更好，尤其是在较高的交并比阈值（IOU=0.9）下，图注检测效果尤为突出。

这一 IOU 截止值的设定依据是对真实框与检测框配对的“损失面积”和“超出区域”进行分析，因此，我们将 IOU=0.9 截止值定义为在将基于 YOLO的目标检测模型应用于文档布局分析问题时，“高度局部化”页面对象的判定标准。为符合我们的提取目标，我们的模型假阳性率相对较低，从而最大限度地减少错误页面对象的提取。

与其他深度学习模型在高 IOU 条件下对我们数据集的泛化能力较弱相似，我们的模型在检测 ScanBank 的电子学位论文集 [19]和 PubLayNet 的非商业用途扫描页面 [20] 子集中的图时，泛化能力不佳；这些数据用于训练我们在比较中采用的 detectron2 版本。相比之下，我们的模型在图注检测方面的泛化能力显著更强，真正率相当或更高，假阳性率更低。此外，我们还表明，我们的后处理流程提高了所有模型的性能，尤其是在图注检测方面。不过，我们提醒读者，鉴于我们的模型与其他模型对图形和图注的定义不同，这些比较结果仅为估算值。

总的来说，我们的工作首先仔细定义了页面对象的类别，然后定义“高定位”并测试了我们模型以及 ScanBank 和 detectron2 模型的泛化能力，这些工作凸显了文档布局分析领域需要认真思考如何将目标检测和分割方法应用于页面对象的提取。在量化提取的准确度之前，首先明确我们 intended 提取哪些信息至关重要。这与其他研究形成对比，后者通常使用平均精度均值或交并比=0.8 作为比较指标，却没有量化这如何转化为从所提取页面对象 [17, 21, 52, 58, 59, 61]中丢失的信息。

我们的工作依赖于相对较小的一组扫描页面（~6000）。尽管本文中图和图注的结果超过了训练基于 YOLO 的模型所需的每类~2000 实例数量估计值 [50, 51]，但我们的数据包含许多复杂布局的边缘情况，我们预计增加数据量将改善这些页面的结果。此外，我们的定义不会将分布在多个页面的分面中的图关联起来（例如，位于不同页面上的图 1a 和图 1b）。关联此类页面需要提取图注，并使用后校正方法对其 OCR 结果进行去噪 [，例如 42] ，这是当前正在开展的工作主题 [44]。

由于我们的模型依赖三个以上的特征通道，因此在预训练的基于YOLO的模型上进行迁移学习并不那么直接，但仍可能是未来工作中利用我们小型数据集的一种方式。

此外，我们当前的方法并未测试常用图像处理特征（例如连通分量 [21]）或对基于YOLO的方法而言“非标准”的损失函数/处理技术 [100]在我们的数据集上的效果。我们还使用用于语言特征生成的“标准”spaCy 软件包，而非科学领域专用的 spaCy 版本（例如 SciSpaCy[13]）。未来纳入这些特征进行测试，或许能提升模型性能。

尽管我们的所有模型都能在 150 个训练轮次内收敛，但这并未包含任何数据增强。由于我们的模型不仅使用灰度信息，还使用 hOCR 属性，因此，典型的数据增强操作（例如翻转、改变饱和度）并不适用于所有特征层。不过，恰当应用的数据增强（例如灰度层的对比度调整）很可能会使模型精度高于此处报告的指标。若计算限制允许，未来开展一项涵盖范围更广的超参数调优研究，测试本文所测试的多个学习率之外的取值，并进一步进行特征选择分析，也将有助于改进我们的工作。

最后，鉴于本文提出的基于YOLO的方法与其他 Mask-RCNN/Faster-RCNN文档版面分析模型（例如 detectron2）之间的错误特征存在差异，结合这两种方法的集成模型很可能会进一步提升模型性能。

尽管本研究使用了天文学专用数据集，但结果表明，与其他模型相比，它具有更强的可推广性潜力 [例如 19, 20, 80]，然而，其准确率仍远未达到在其他领域大规模部署、用于处理历史文献所需的水平。正如以往研究所指出的 [98]，“可推广性问题”很可能需要通过审慎定义页面对象类别、开展大规模标注活动，并结合本研究开发的模型等方法，对“高精度定位”进行量化，才能得到解决。

本研究得到 Fiddler Fellowship 和美国国家航空航天局天体物理学数据分析计划资助项目（20-ADAP20-0225）的支持。

[13]https://allenai.github.io/scispacy/。

参考文献

[1]Sandy, H.M., Mitchell, E., Corrado, E.M., Budd, J., West, J.D., Bossaller, J., VanScoy, A.：为开放研究提供依据：对可重复性的影响以及透明度。《信息科学与技术协会学报》54(1), 583–586（2017年）
https://arxiv.org/abs/https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.2017.14505401079。https://doi.org/10.1002/pra2.2017.14505401079

[2]Sohmen, L., Charbonnier, J., Bl¨umel, I., Wartena, C., Heller, L.：科学开放获取出版物中的图。在：数字图书馆理论与实践国际会议，第220–226页（2018年）。施普林格

[3] Mayernik, M.S., Hart, D.L.,Maull, K.E., Weber, N.M.：评估和追踪科研基础设施的成果与影响。《信息科学与技术协会学报》68(6), 1341–1359（2017年）
https://arxiv.org/abs/https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23721。https://doi.org/10.1002/asi.23721

[4] Gao, L., Yi,X., Jiang, Z., Hao, L.,Tang, Z.：ICDAR2017 页面对象检测竞赛。载于：2017年第14届 IAPR 国际文档分析与识别会议（ICDAR），第 01 卷，第 1417–1422 页（2017年）。
https://doi.org/10.1109/ICDAR.2017.231

[5]Zhong, X., Tang, J., Jimeno Yepes, A.：PubLayNet：迄今最大的文档版面分析数据集。arXiv 预印本，1908–07836（2019年）
https://arxiv.org/abs/1908.07836 [cs.CL]

[6]Bhatt, J., Hashmi, K.A., Afzal, M.Z., Stricker, D.：基于深度神经网络的图形页面对象检测综述。《应用科学》11(12)（2021年）。https://doi.org/10.3390/app11125344

[7]Lehenmeier, C., Burghardt, M., Mischka,

B.：版面检测与表格识别——历史文档和手写表格数据数字化的近期挑战。载于：数字图书馆理论与实践国际会议，第229–242页（2020年）。施普林格[8]Klampfl, S., Kern, R.：一种从数字科学文章中提取正文和目录的无监督机器学习方法。载于：数字图书馆理论与实践国际会 议，第144–155页（2013年）。施普林格[9]Bai, K., Mitra, P., Giles, C.L., Liu, Y.：从数字文档中自动提取表格元数据。载于：第六届 ACM/IEEE-CS 数字图书馆联合会议论文集（JCDL'06），第339–340页（2006年）。IEEE[10]Choudhury, S.R., Tuarob, S., Mitra, P., Rokach, L., Kirk, A., Szep, S., Pellegrino, D., Jones, S., Giles, C.L.：化学数字图书馆的图搜索引擎架构。载于：第十三届 ACM/IEEE-CS 数字图书馆联合会议论文集，第369–370页（2013年）[11] Lopez, P.：GROBID。GitHub（2008—2021年）。https://github.com/kermitt2/grobid访问日期：2023-02-03[12] Choudhury, S.R., Mitra, P., Kirk, A., Szep, S., Pellegrino, D., Jones, S., Giles, C.L.：从数字文档中提取图元数据。载于：2013年第12届国际文档分析与识别会议，第135–139页（2013年）。https://doi.org/10.1109/ICDAR.2013.34[13] Clark, C., Divvala, S.：Pdffigures 2.0：从研究论文中挖掘图。载于：2016年 IEEE/ACM 数字图书馆联合会议（JCDL），第143–152页（2016年）[14]Bhatt, J., Hashmi, K.A., Afzal, M.Z., Stricker, D.：基于深度神经网络的图形页面对象检测综述。《应用科学》**11**(12)（2021年）。https://doi.org/10.3390/app11125344

[15] Siegel, N., Lourie, N., Power, R., Ammar, W.：利用远程监督神经网络提取科学图。arXiv 电子预印本，1804–02445（2018年）https://arxiv.org/abs/1804.02445 [cs.DL]

[16]Sinha, S., Hashmi, K.A., Pagani, A., Liwicki, M., Stricker, D., Afzal, M.Z.：重新思考用于扫描文档图像中图形对象检测的可学习提议。《应用科学》**12**(20)，10578（2022年）。https://doi.org/10.3390/app122010578。编号：20 出版商：Multidisciplinary Digital Publishing Institute

[17]Yang, X., Yumer, E., Asente, P., Kraley, M., Kifer, D., Giles, C.L.：学习使用多模态全卷积神经网络从文档中提取语义结构。在：2017年 IEEE 计算机视觉与模式识别会议（CVPR），第4342–4351 页（2017年）。https://doi.org/10.1109/CVPR.2017.462[18] Saha, R., Mondal, A., Jawahar, C.V.：文档图像中的图形对象检测。在：2019年国际文档分析与识别会议（ICDAR），第 51–58 页（2019年）。https://doi.org/10.1109/ICDAR.2019.00018[19] Yashwant Kahu, S., Ingram, W.A., Fox, E.A., Wu, J.：ScanBank：用于从扫描版电子学位论文中提取图形的基准数据集。arXiv 电子预印本，2106–15320（2021年）https://arxiv.org/abs/2106.15320 [计算机视觉[20] Zhong, X., Tang, J., Jimeno Yepes, A.：PubLayNet：迄今最大的文档布局分析数据集。arXiv 电子预印本，1908–07836（2019年）https://arxiv.org/abs/1908.07836 [cs.CL][21] Younas, J., Rizvi, S.T.R., Malik, M.I., Shafait, F., Lukowicz, P., Ahmed, S.：FFD：从文档图像中进行图形与公式检测。在：2019年数字图像计算：技术与应用会议（DICTA），第 1–7 页（2019年）。
h
t
t
ps://
doi.org/10.1109/DICTA47822.2019.8945972

[22] Smith, R.：Tesseract 光学字符识别引擎概述。载于：第九届文档分析与识别国际会议论文集——第 02 卷。ICDAR '07，第 629–633 页。IEEE计算机学会，美国（2007年）

[23]Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., Robinson, G.E.：大数据：天文学还是基因组学？《公共科学图书馆·生物学》**13(7)**，1–11（2015年）。https://doi.org/10.1371/journal.pbio.1002195

[24]Smith, L., Arcand, K., Smith, R., Bookbinder, J., Smith, J.：捕捉爆发恒星的多重面貌：传播复杂且不断演变的天文学数据。JCOM《科学传播期刊》**16**，16050202（2017年）。https://doi.org/10.22323/2.16050202[25] Bagga, S., Piper, A.：HATHI 1M：介绍 Hathi Trust 收录的百万页英文历史散文数据集。哈佛 Dataverse（2021年）。https://doi.org/10.7910/DVN/HAKKUA。https://doi.org/10.7910/DVN/HAKKUA[26] Pepe, A., Goodman, A., Muench, A.：《ADS全天巡天》。载于：Ballester, P., Egret, D., Lorente, N.P.F.（编）《天文学数据分析软件与系统 XXI》。太平洋天文学会会议系列，第 461 卷，第 275 页（2012年）[27] Accomazzi, A., Kurtz, M.J., Henneken, E. A., Grant, C.S., Thompson, D., Chyla, R., Holachek, A., Sudilovsky, V., Murray, S.S.：ADS 中功能与整理支持的改进。载于：美国天文学会第 225 届会议摘要。美国天文学会会议摘要，第 225 卷，第 336–55 页（2015年）[28] Iwatsuki, K., Sagara, T., Hara, T., Aizawa, A.：检测科学文档中的行内数学表达式。载于：2017年 ACM 文档工程研讨会论文集。DocEng '17，第 141–144 页。计算机协会，美国纽约州纽约（2017年）。https://doi.org/10.1145/3103010.3121041。https://doi.org/10.1145/3103010.3121041

[29]弗赖堡：算法，U.，团队，D.S.：pdfact。GitHub（2016–2023年）。https://github.com/ad-freiburg/pdfact访问日期：2023-02-03[30]K, P.：pdfedit。GitHub（2006–2012年）。https://github.com/nullishzero/PDFEdit访问日期：2023-02-03[31] Shinyama, Y.：pdfminer.six。GitHub（2018–2022年）。https://github.com/pdfminer/pdfminer.six访问日期：2023-02-03[32]开发者，T.P.：pdftocairo。Poppler 开发者（2005–2011年）。https://manpages.ubuntu.com/manpages/trusty/man1/pdftocairo.1.html访问日期：2023-02-03[33]Fenniak, M., Stamy, M., pubpub-zz, Thoma, M., Peveler, M., exiledkingcc, PyPDF2 贡献者：PyPDF2 库（2022年）。https://pypi.org/project/PyPDF2/[34]Kahu, S.Y.：从扫描版电子学位论文中提取图形。硕士学位论文，弗吉尼亚理工大学（2020年）。https://vtechworks.lib.vt.edu/handle/10919/100113[35] Lopez, P.：Grobid：结合自动书目数据识别与学术出版物术语提取。载于：数字图书馆研究与先进技术：第13届欧洲会议，ECDL 2009，希腊科孚岛，2009年9月27日至10月2日。论文集13，第473–474页（2009年）。施普林格[36] Romary, L., Lopez, P.：Grobid：科学出版物信息抽取。ERCIM 新闻**100**（2015年）[37]Li, P., Jiang, X., Shatkay, H.：从科学出版物中提取图形和图注。载于：第27届 ACM 国际信息与知识管理会议论文集，第

1595–1598（2018年）

[38] Yu, C.-N., Levy, C.C., Saniee, I.：历史技术文献中的图像提取卷积神经网络。载于：2017年第14届IAPR国际文档分析与识别会议（ICDAR），第1卷，第789–795页（2017年）。IEEE

[39] Subramani, N., Matton, A., Greaves, M., Lam, A.：用于光学字符识别和文档理解的深度学习方法综述。arXiv. arXiv:2011.13534 [cs]（2021年）。http://arxiv.org/abs/2011.13534

[40] Etter, D., Rawls, S., Carpenter, C., Sell, G.：光学字符识别的合成配方。载于：2019年国际文档分析与识别会议（ICDAR），第864–869页。IEEE，悉尼，澳大利亚（2019年）。https://doi.org/10.1109/ICDAR.2019.00143。https://ieeexplore.ieee.org/document/8978033/

[41]Boros, E., Nguyen, N.K., Lejeune, G., Doucet, A.：评估光学字符识别噪声对数字化文档中多语言事件检测的影响。《国际数字图书馆期刊》**23**（3），241–266（2022年）。https://doi.org/10.1007/s00799-022-00325-2

[42] Ramirez-Orta, J.,Xamena, E.,Maguitman, A., Milios, E., Soto, A.J.：利用大型字符序列到序列模型集成进行光学字符识别后文档校正。技术报告 arXiv:2109.06264，arXiv（2022年1月）。https://doi.org/10.48550/arXiv.2109.06264. arXiv:2109.06264[cs] 类型：文章。http://arxiv.org/abs/2109.06264

[43] Zhu, D., Naiman, J.P., G., W.P.K., Goodman, A.：《使用Tesseract引擎进行OCR：基于合成数据的参数研究》。未发表论文（2023年）

[44]Cosillo, M., Naiman, J.P., G., W.P.K., Goodman, A.：《历史科学文本的OCR后校正》。未发表论文（2023年）

[45] Grant, C.S.、Accomazzi, A.、Eichhorn, G.、Kurtz, M.J.、Murray, S.S.：美国国家航空航天局天体物理学数据系统：数据馆藏。《天文学与天体物理学增刊系列》**143**（1），111–135（2000年）。https://doi.org/10.1051/aas:2000173

[46]Hasan Choudhury, M., Jayanetti, H.R., Wu, J., Ingram, W.A., Fox, E.A.：《结合扫描版电子学位论文视觉特征的自动元数据提取》。arXiv电子预印本，2107–00516（2021年）https://arxiv.org/abs/2107.00516 [cs.DL]

[47]Bradski, G.：《OpenCV库》。《Dr. Dobb's软件工具杂志》（2000年）

[48]Skalski, P.：Make Sense。https://github.com/SkalskiP/make-sense/ (2019年)

[49] W., A.: OCR偏移图像。StackOverflow（2019年）。https://i.stack.imgur.com/qDUFT.png 访问日期：2023年02月06日

[50]Bochkovskiy, A., Wang, C.-Y., Liao, H.- Y.M.: YOLOv4：目标检测的最佳速度与准确率（2020年）

[51]Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M.: Scaled-YOLOv4：跨阶段部分网络的扩展。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第13029–13038页（2021年）

[52]Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once：统一的实时目标检测。arXiv电子预印本，1506–02640（2015年）https://arxiv.org/abs/1506.02640 [计算机视觉]

[53] Ren,S., He,K.,Girshick, R., Sun, J.：Faster R-CNN：借助区域建议网络实现实时目标检测。arXiv 电子预印本，1506–01497（2015年）https://arxiv.org/abs/1506.01497 [计算机视觉]

[54]Schreiber, S., Agne, S., Wolf, I., Dengel, A., Ahmed, S.: Deepdesrt：用于检测文档图像中的表格并识别其结构的深度学习方法。载于：201714th

IAPR 文档分析与识别国际会议（ICDAR），第 01 卷，第 1162–1167 页（2017年）。https://doi. org/10.1109/ICDAR.2017.192[55] Vo, N.D., Nguyen, K., Nguyen, T.V., Nguyen, K.: 用于页面对象检测的深度目标检测器集成方法。载于：第十二届普适信息管理与通信国际会议论文集。IMCOM '18。计算机协会，纽约，NY，美国（2018年）。https://doi.org/10.1145/ 3164541.3164644. https://doi.org/10.1145/ 3164541.3164644 [56] Gilani, A., Qasim, S.R., Malik, I., Shafait, F.: 使用深度学习进行表格检测。载于：2017年IAPR第14届文档分析与识别国际会议（ICDAR），第 01 卷，第 771–776 页（2017年）。https://doi.org/10.1109/ICDAR.2017.131[57] He, K., Gkioxari, G., Doll´ar, P., Girshick, R.: Mask R-CNN。arXiv 电子预印本，1703–06870（2017年）https://arxiv.org/abs/1703.06870 [计算机视觉] [58] Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.: DocBank：文档版面分析基准数据集。arXiv 电子预印本，2006–01038（2020年）https:// arxiv.org/abs/2006.01038 [cs.CL][59] Agarwal, M., Mondal, A., Jawahar, C.V.: CDeC-Net：用于文档图像中表格检测的复合可变形级联网络。arXiv 电子预印本，2008–10831（2020年）https://arxiv.org/abs/2008.10831 [计算机视觉] [60] Kr¨ahenb¨uhl, P., Koltun, V.: 具有高斯边缘势的全连接条件随机场中的高效推理。arXiv 电子预印本，1210–5644（2012年）https://arxiv.org/abs/1210.5644 [计算机视觉] [61] Kavasidis, I., Palazzo, S., Spampinato, C., Pino, C., Giordano, D., Giuffrida, D., Messina, P.: 用于数字化文档中表格和图表检测的基于显著性的卷积神经网络。arXiv 电子预印本，1804–06236（2018年）https://arxiv.org/abs/1804.06236 [计算机视觉]

[62] Dai, J., He, K., Li, Y., Ren, S., Sun, J.：实例敏感的全卷积网络。arXiv 电子预印本，1603–08678（2016）https://arxiv.org/abs/1603.08678 [计算机视觉][63]Dai, J., Li, Y., He, K., Sun, J.：R-FCN：通过基于区域的全卷积网络进行目标检测。arXiv 电子预印本，1605–06409（2016）https://arxiv.org/abs/1605.06409 [计算机视觉][64] Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y.：全卷积实例感知语义分割。arXiv 电子预印本，1611– 07709（2016）https://arxiv.org/abs/1611.07709 [计算机视觉][65] Ha, J., Haralick, R.M., Phillips, I.T.：使用连通分量边界框的递归 x-y 切分。载于：第三届国际文档分析与识别会议论文集，第 2 卷，第 952–9552 页（1995）。https://doi.org/10.1109/ICDAR.1995.602059[66] Chowdhury, S.P., Mandal, S., Das, A.K., Chanda, B.：从文档图像中自动分割数学区域。载于：第七届国际文档分析与识别会议，2003 年。会议论文集，第 755–759 页（2003）。https://doi.org/10.1109/ICDAR.2003.1227763[67] Cronje, J.：从专利图纸图像中进行图形检测和部件标签提取。南非模式识别协会第 23 届年度研讨会（2012 年）https://arxiv.org/abs/http://researchspace.csir.co.za/dspace/handle/10204/6464 [68] Bukhari, S.S., Al Azawi, M.I.A.,Shafait, F., Breuel, T.M.：使用连通分量上的判别式学习进行文档图像分割。载于：第九届 IAPR 文档分析系统国际研讨会论文集。DAS '10，第 183–190 页。计算机协会，纽约，纽约州，美国（2010 年）。https://doi.org/10.1145/1815330.1815354。https://doi.

org/10.1145/1815330.1815354

[69]Younas, J., Siddiqui, S.A., Munir, M., Malik, M.I., Shafait, F., Lukowicz, P., Ah med, S.: Fi-fo 检测器：使用可变形网络进行图和公式检测。《应用科学》**10**（18）（2020年）。https://doi.org/10.3390/app10186460

[70]Arlazarov, V.V., Andreeva, E.I., Bula- tov, K.B., Nikolaev, D.P., Petrova, O.O., Savelev, B.I., Slavin, O.A.: 《文档图像分析与识别：综述》。《计算机光学》**46**(4), 567– 589（2022年）。https://doi.org/10.18287/2412-6179-CO-1020[71] Binmakhashen, G.M., Mahmoud, S.A.: 《文档版面分析：全面综述》。《ACM计算调查》**52**(6), 109–110936（2019年）。https://doi.org/10.1145/3355610[72]Kosaraju, S.: 《文档版面分析与识别系统》。计算机科学理学硕士论文（2019年）
https://arxiv.org/abs/https://digitalcommons.kennesaw.edu/cs etd/28/[73]Christenson, H.: HathiTrust。《图书馆资源与技术服务》**55**(2), 93–102（2011年）[74] Ribaric, T.: 为DSpace存储库平台自动准备来自互联网档案馆的学位论文材料。Code4LibJour- nal (8)（2009年）[75]Wajer, M.: 使用Tesseract和hOCR在互联网档案馆进行光学字符识别。互联网档案馆（2020年–2022年）。https://archive.org/developers/ocr.html访问日期：2023年02月06日[76]Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once：统一的实时目标检测。arXiv预印本，1506–02640（2015年）https://arxiv.org/abs/1506.02640 [计算机视觉[77]Jocher, G., Stoken, A., Borovec, J.,NanoCode012, ChristopherSTAN, Changyu, L., Laughing, tkianai, Hogan, A., lorenzomammana, yxNONG, AlexWang1900, Diaconu, L., Marc, wang- haoyang0106, ml5ah, Doug, Ingham, F., Frederik, Guilhen, Hatovix, Poznan- ski, J., Fang, J., Yu, L., changyu98, Wang,M., Gupta,N., Akhtar, O., PetrD- voracek, Rai, P.: ultralytics/yolov5：v3.1 - 错误修复与性能改进。Zenodo（2020年）。https://doi.org/10.5281/zenodo.4154370。https://doi.org/10.5281/zenodo.4154370

[78]Honnibal, M., Montani, I., Van Lan- deghem, S., Boyd, A.: spaCy：Python 中的工业级自然语言处理（2020年）。https://doi.org/10.5281/zenodo.1212303

[79] Yi, X., Gao, L., Liao, Y., Zhang, X., Liu, R., Jiang, Z.: 基于卷积神经网络的文档图像页面对象检测。载于：2017年第14届 IAPR 国际文档分析与识别会议（ICDAR），第01卷，第230–235页（2017年）。https://doi.org/10.1109/ICDAR.2017.46

[80] Wu, Y.、Kirillov, A.、Massa, F.、Lo, W.-Y.、Girshick, R.: 《Detectron2》。https://github.com/facebookresearch/detectron2 （2019年）

[81]Girshick, R.: 《快速R-CNN》。arXiv 电子预印本，1504–08083（2015年）https://arxiv.org/abs/1504.08083 [计算机视觉]。https://doi.org/10.48550/arXiv.1504.08083

[82]Girshick, R., Donahue, J., Darrell, T., Malik, J.: 用于精准目标检测和语义分割的丰富特征层级。arXiv 预印本，1311–2524（2013年）https://arxiv.org/abs/1311.2524 [计算机视觉]。https://doi.org/10.48550/arXiv.1311.2524[83] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN：基于区域建议网络的实时目标检测。arXiv 预印本，1506–01497（2015年）https://arxiv.org/abs/1506.01497 [计算机视觉]。https://doi.org/10.48550/arXiv.1506.01497[84] Felzenszwalb, P.F., Girshick, R.B.,

McAllester, D., Ramanan, D.：通过判别式训练的基于部件的模型进行目标检测。IEEE 模式分析与机器智能汇刊**32(9)**，1627–1645（2010年）。https://doi.org/10.1109/TPAMI.2009.167

[85] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD：单次多框检测器。arXiv 预印本，1512–02325（2015年）https://arxiv.org/abs/1512.02325 [计算机视觉]。https://doi.org/10.48550/arXiv.1512.02325[86] Yang, H., Hsu, W.：用于文档布局理解的基于Transformer的方法。载于：2022年IEEE国际图像处理会议（ICIP），第4043–4047页（2022年）。https://doi.org/10.1109/ICIP46576.2022.9897491 [87]Singhal, T., Liu, J., Blessing, L.T.M., Lim,K.H.：使用领域专用词嵌入和主题建模分析科学出版物。载于：2021年IEEE大数据国际会议（Big Data），第4965–4973页（2021年）。https://doi.org/10.1109/BigData52589.2021.9671598[88] Lin, T.-Y.,Maire, M., Belongie, S., Bour-dev,L., Girshick, R., Hays, J., Perona,P., Ramanan, D., Zitnick,C.L.,Doll´ar, P.：Microsoft COCO：上下文中的常见物体。arXiv 预印本，1405–0312（2014年）https://arxiv.org/abs/1405.0312 [计算机视觉[89] Wick, C., Puppe, F.：用于历史文档图像页面分割的全卷积神经网络。载于：2018年第13届IAPR文档分析系统国际研讨会（DAS），第287–292页（2018年）。https://doi.org/10.1109/DAS.2018.39[90] Pletschacher, S., Antonacopoulos, A.：PAGE（页面分析与真实标注元素）格式框架。载于：2010年第20届国际模式识别会议，第257–260页。IEEE，土耳其伊斯坦布尔（2010年）。https://doi.org/10.1109/ICPR.2010.72。http://ieeexplore.ieee.org/document/5597587/

[91]Bylinskii, Z., Borkin, M.A., Kim, N.W., Pfister, H., Oliva, A.：用于大规模评估和比较信息可视化的眼动注视指标。载于：Burch, M.,Chuang, L.,Fisher, B., Schmidt, A., Weiskopf, D.（编），《眼动追踪与可视化》，第235–255页。施普林格，Cham（2017年）

[92] Giannakopoulos, T.,Foufoulas, I.,Stamato- giannakis, E., Dimitropoulos, H., Manola, N., Ioannidis, Y.：基于视觉的科学文献图分类。载于：第24届万维网国际会议论文集，第1059–1060页（2015年）

[93]Davila, K., Setlur, S., Doermann, D., Kota, B.U., Govindaraju, V.：图表挖掘：自动图表分析方法综述。《IEEE模式分析与机器智能汇刊》**43**(11), 3799–3819 (2021年)。https://doi.org/10.1109/TPAMI.2020.2992028[94] Nagy, G., Seth, S., Viswanathan, M.：面向技术期刊的文档图像分析系统原型。Computer **25**(7), 10–22 (1992)。https://doi.org/10.1109/2.144436。会议名称：Computer[95] Jiang, Z., Deng, H., Wu, Z., Guo, J., Sun, S., Mijovic, V., Yang, Z., Lou, J.-G., Zhang, D.：UniLayout：驾驭用于图形布局生成的统一序列到序列转换器。arXiv电子印本，2208–08037（2022年）https://arxiv.org/abs/2208.08037 [计算机视觉]。https://doi.org/10.48550/arXiv.2208.08037[96] Silajev, I., Victor, N., Mortimer, P.：使用LayoutLMv3进行语义表格检测。arXiv电子印本，2211–15504（2022年）https://arxiv.org/abs/2211.15504 [计算机视觉]。https://doi.org/10.48550/arXiv.2211.15504[97] Gupta, B.M., Dhawan, S.M.：“图书馆电子资源”全球研究综述：1994—2017年的科学计量学评估。《COLLNET科学计量学与信息管理期刊》**13**(1), 167–181 (2019年)

https://arxiv.org/abs/https://doi.org/10.1080/09737766.2019.1583836。 https://doi.org/10.1080/09737766.2019.1583836

[98]Naiman, J.P.：《科学文章图与图注提取中的文档布局分析可推广性》。arXiv 预印本，2301–10781（2023年）https://arxiv.org/abs/2301.10781 [计算机视觉]。https://doi.org/10.48550/arXiv.2301.10781

[99] Pfitzmann, B.、Auer, C.、Dolfi, M.、Nassar, A.S.、Staar, P.W.J.：《DocLayNet：用于文档布局分析的大型人工标注数据集》。arXiv 预印本，2206–01062（2022年）https://arxiv.org/abs/2206.01062 [计算机视觉]

[100] Neubeck，A.，Van Gool，L.：高效的非极大值抑制。载于：第18届国际模式识别会议（ICPR'06），第3卷，第850–855页（2006）。https://doi.org/10.1109/ICPR.2006.479

6 附录

展开后的完整架构图由图 3 展示，见图10。首先使用Tesseract 光学字符识别引擎处理扫描页面，并通过图像处理查找页面上的方形（左侧粉色面板，第 2.3 节）。然后进行 PDF挖掘以查找图或图注，但实际上只有一小部分扫描件包含可由 PDF 解析的页面对象，而且这些对象通常只有图注（左侧紫色面板，第 2.4 节）。接着，从灰度图像和 OCR 数据中提取特征，再将其输入 Mega-Yolo 模型，生成原始检测框（左侧红色面板，第 3.1 节）。最后，利用 PDF 挖掘检测到的图注、图像处理检测到的图框以及 OCR 词边界框，对原始检测框进行清理，得到输出检测框（左下方橙色面板，第 3.2 节）。

为了生成真实框，对同一页面进行人工标注（右侧紫色面板）。随后利用 OCR 词边界框扩充人工标注框，生成准确的图和图注框（右下方橙色面板）。

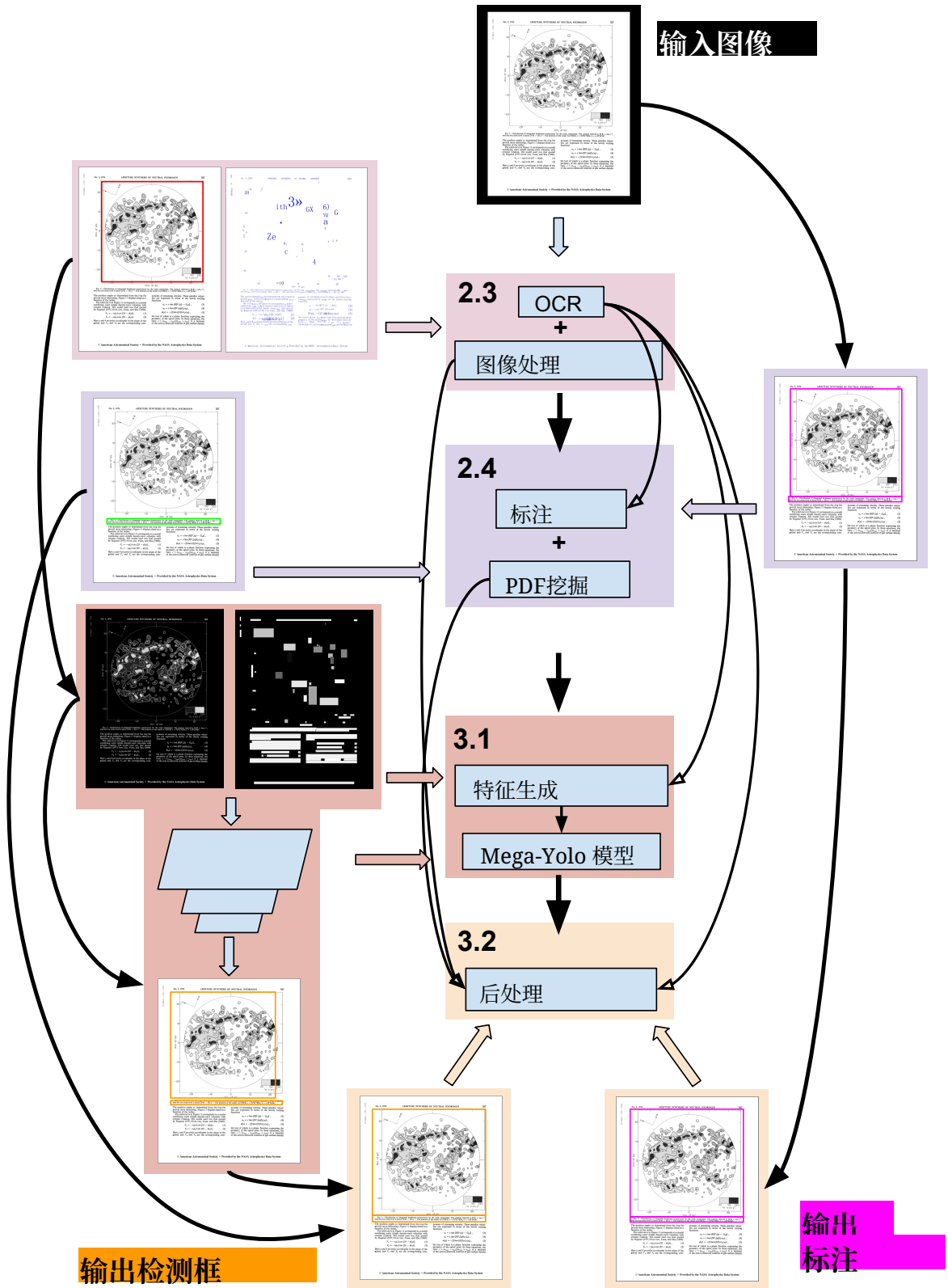

图10：详细架构图展示了扫描页面如何通过我们的完整模型流水线进行处理（左侧分面），包括光学字符识别和图像处理、PDF挖掘、特征生成，以及随后将特征输入 Mega-Yolo 模型，模型输出最终生成检测框。右侧展示了在扫描页面上执行的标注流程，用于生成真实值模型，以便与我们模型生成的检测框进行比较。