arXiv:2301.10781v1 [cs.DL] 25 Jan 2023

科学文章图表与标题提取的文档布局分析泛化能力

Jill Naiman[1], [2*]

[1*]信息科学学院，伊利诺伊州大学，厄巴纳-尚佩恩，Daniel ST，尚佩恩，61820，伊利诺伊州，美国。[2]伊利诺伊州大学国家超级计算应用中心，厄巴纳-尚佩恩，W Clark ST，厄巴纳，61801，伊利诺伊州，美国。

通讯作者。电子邮箱：jnaiman@illinois.edu；

摘要

泛化能力不足——即在一个数据集上训练的模型无法对另一个数据集给出准确结果——是文档布局分析领域中一个已知的问题。因此，当模型用于定位科学文献中的重要页面对象（如图、表格、标题和数学公式）时，往往无法成功应用于新的领域。尽管已有多种解决方案被提出，包括更新、改进的深度学习模型、更大规模的人工标注数据集，以及生成大规模合成数据集，但迄今为止，仍没有一种“灵丹妙药”能将针对特定领域或历史时期训练的模型迁移到新的领域。在此，我们介绍正在进行的工作：将我们的文档布局分析模型从历史天体物理学文献迁移至 HathiTrust 美国联邦文献馆藏中更大规模的科学文档语料。我们以此为例，突出文档版面分析社区在泛化能力方面面临的一些问题，并探讨应对这些问题的若干挑战与可能的解决方案。本项目的所有代码均可在 The Reading Time Machine 的 GitHub 仓库中获取：https://github.com/ReadingTimeMachine/htrcshort conf。

关键词： 学术文档处理、文档布局分析、天文学。

1 引言

科学领域的大量信息通过学术文献传播，而这些信息是通过纯文本以外的页面组成部分传递的——也就是图、表格和数学公式。因此，这类文档对象对科学家以及研究科学共同体和科学过程的人士尤为重要 [1]。

虽然较新的学术文章是“数字原生”的，也就是说，文献以便于提取页面对象的格式存储（例如 XML），但文章并非都如此。这些文献发表于互联网 [2]广泛普及并具备访问能力之前。页面对象提取的便利性有时也适用于基于矢量的 PDF，即“基于规则”的 PDF，其文件类型包含用于呈现文章页面的指令。如果已知矢量 PDF 的格式，则可通过搜索关键词的启发式方法 [3, 4]提取文本和图像，并将表格/图及其图注作为配对项提取 [4, 5]。然而，启发式方法

从矢量 PDF 文档中提取和索引图往往并非易事，可能导致页面对象错误或缺失 [6]。

近年来，人们通常采用深度学习方法，通过多种方法从原生数字化和扫描文档中提取页面对象 [6–8]。对于原生数字化文档，这些深度学习方法会与基于启发式方法得出的结果结合，在后处理步骤中使用 [9]。然而，对于历史扫描文档，这些方法面临许多挑战 [10–12]。

鉴于 HathiTrust 为研究人员提供了大量此类数据产品，值得探讨在尝试提取和分类这些页面对象时固有的一些问题。下文将聚焦于文档版面分析社区中的一个主要问题，即启发式模型和深度学习模型的泛化能力难题——当模型针对某种特定类型的页面（即电子学位论文）进行训练时，往往无法在另一种类型的页面（即学术文章）上取得良好效果。

我们工作的自然延伸，是在其他科学领域或期刊中提取图及其图注。HathiTrust 中可能包含数百万个此类文章页面，因此，将我们的模型应用于这些文章，或许会是一项富有成效的工作。为了展示其中一些问题，我们将模型应用于美国联邦文献[3] 馆藏中随机选取的文章页面。为了在比较中尽可能有利于模型取得成功，我们使用“天文学”这一搜索字段缩小该馆藏范围，并筛选出仅包含英文会议、期刊和手稿文献且可“全文浏览”的结果，总计 56,282 条记录。作为示例，我们选取 ≈350 筛选后馆藏中六篇文章的随机页面，并使用 [12, 14] Make-Sense.ai[4][15]中的图和图注类别定义对其进行标注。将我们的模型应用于这些数据后，在不同交并比阈值下得到的精确率、召回率和 F1 分数，以藏青色实线和虚线显示于图 1中。可以看出，这些线明显低于我们在原始数据集中得到的 F1≈90%。

2 泛化能力问题

泛化能力不足是文档布局分析领域中一个众所周知且普遍存在的问题 [，例如 6, 13]。不仅出版物类型的变化，单是出版年份的变化，也可能大幅降低页面对象提取方法的准确率，尤其是对于未针对这类文档进行专门训练的模型 [10, 12, 14]。

我们先前的工作旨在从天体物理数据系统（ADS）[1] 收藏的部分“数字化前”天体物理学文献中提取图及其图注，利用了文章页面的灰度和光学字符识别（OCR）特征 [12]。我们的模型在数据集上取得了很高的准确度——对于交并比（IOU）为 0.9 的指标，我们得到的 F1 分数为 $\geq 90\%$[2][12, 14]。

一种可能的解释很简单：我们的 ≈6000 标注页面数据集虽然足够大，可以训练出像我们这样基于 YOLO 的模型， [18, 19]但仍然太小，无法创建出具有良好泛化能力的模型。为了验证这一可能性，我们还应用了一个[5] Facebook 的 detectron2 [16]版本，该版本是在 PubLayNet[6] 数据集[17]上训练的。由于 PubLayNet 数据集包含超过 100 万篇文章，因此如果确实是训练集的规模影响了模型的泛化能力，那么这应该能提供足够多的训练样本，使 detectron2 具有良好的泛化能力。然而，正如图 1 中的橙色实线和虚线所示，事实并非如此。虽然 detectron2 模型在图方面的 F1 分数高于我们的模型，但其在图注方面表现较差（不过，正如 [12] 所讨论的，这可能是由于“图注”类与“文本”类之间存在不一致，后者是该版本中唯一经过训练的类别

[1] https://ui.adsabs.harvard.edu/ [2] F1 分数是精确率（prec）和召回率（rec）的组合指标，其中 $F1 = 2 \times prec \times rec/(prec + rec)$，其中 $prec = TP/(TP + FP)$，$rec = TP/(TP + FN)$，分别表示真正例（TP）、假正例（FP）和假负例（FN）的组合指标。

[3] https://babel.hathitrust.org/cgi/mb?a=listis&c=2062901859 [4] https://www.makesense.ai/ [5] https://github.com/JPLeoRX/detectron2-publaynet [6] https://github.com/ibm-aur-nlp/PubLayNet

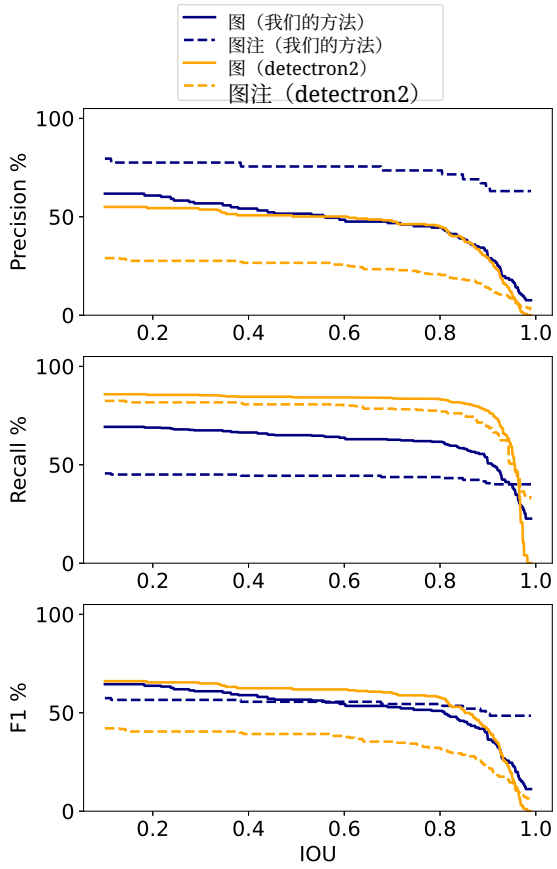

图 1：精确率（上）、召回率（中）和 F1 分数（下）在不同交并比阈值下的表现：我们的模型 [12, 14]（蓝线）以及 detectron2 [16]在 PubLayNet 文档数据集上训练的模型 [17]（橙线）。这两个模型都无法很好地泛化到 HathiTrust 文档集。如底部子图中的 F1 分数所示，在较低交并比阈值下，detectron2 对图（实线）的表现更好；在较高交并比阈值下，我们的模型对图注（虚线）以及图和图注的表现更好。指标根据从六篇 HathiTrust 文章中随机选取的 363 个页面计算得出（详情见正文）。

（即在该版本的 detectron2 中训练的类别）。此外，与我们的模型相比，detectron2 模型在较高交并比水平下的指标下降幅度略大。正如 [14] – 中所阐述的，在使用像我们这种基于 YOLO 的模型，以及 detectron2 中使用的 Faster/Mask-RCNN 模型等目标检测方法时，保持较高的准确率在较高交并比阈值下的表现对于从页面中“提取”页面对象并保留其可读性和整体实用性至关重要。因此，两个模型在高交并比下准确率下降令人担忧。

两个模型准确率显著下降的一种可能原因是，它们训练所用的数据与现在要应用于其中的新 HathiTrust 数据过于“不同”。图 2 定性展示了这一点，其中显示了我们的模型和 detectron2 在 HathiTrust 数据集中的一个页面（左侧两个子图）以及我们最初的 ADS 天体物理数据集中的一个页面（右侧两个子图）上的表现。我们的模型和 detectron2 都漏掉了不熟悉的版面元素，例如 HathiTrust 页面上重叠的图以及“浮动”的图注。

我们的模型在自己的数据集上表现最好，这也在情理之中——它是专门针对天体物理学文献训练的，而 detectron2 模型则是基于 PubMed 的医学科学文献训练的 [17]。

图 2展示了不同类型文档可能存在的准确率差异——在本例中，是“曼哈顿风格”（学术）文档与“非曼哈顿风格”（非学术）文档之间的准确率差异。此外，detectron2 未能在天体物理学数据集（图 2最右侧的图）中准确找到边界框，这凸显了不同模型对文档页面布局变化的敏感程度。

3 答案永远是更多数据

考虑到字体大小和页面对象间距看似细微的差异，就可能导致机器学习模型的准确率大幅下降，一个自然而然的解决方案或许是扩大用于训练的数据集的规模和范围。理论上虽简单，实践中却可能遇到若干值得关注的问题。

3.1 不一致的页面对象定义

在启动大规模数据收集计划之前，必须明确他们具体要在文档页面上寻找什么

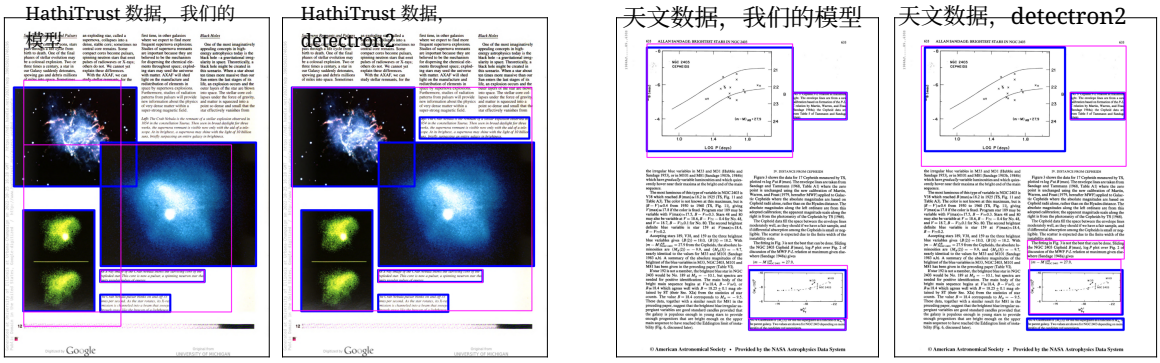

图 2：真实标注框（蓝色）以及检测到的框（洋红色），分别来自本文介绍的模型 [12]和 detectron2 模型 [16]（在 PubLayNet 数据集上训练）[17]。这两个模型在 HathiTrust 页面示例上表现都不佳（左侧两个子图），但应用于 ADS 天体物理数据集页面中的页面对象时，准确度更高（右侧两个子图）。

。正如 [14]，中所述，不同的数据集通常采用不同的页面对象定义进行标注，例如图和图注。举例来说，有时图注会包含在图框定义中；有时图中的每个子图都会被标注为一幅图；而在其他情况下，图中的所有子图都被视为位于同一个图框内（见 [14]的图 2）。

不同的标注员处理同一数据集时，可能会对类别定义产生分歧，导致同一数据集中的数据不一致 [13, 20]。尽管人们一直在推动采用标准化方法来定义页面对象类别 [，例如 21]，，但参与文档布局分析的不同群体至今仍未全面采用这些方法。

此外，这些不同类别的定义可能取决于使用场景。在 [12, 14]，所介绍的工作中，我们根据最终目标——在 Astronomy Explorer (AIE) 上托管历史图像及其图注[7]，并像处理新发表的文章那样自动完成这一过程——来定义图和图注类别。因此，即使能够制定一致的标注“规范”（如 [12, 14]中所做的那样），这些定义也可能无法满足所有文档布局分析应用的需求。

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

3.2 人工标注成本高昂

假设某个群体或领域对页面对象类别的定义达成一致，那么生成数据的一种方式是聘请一大批标注人员，对文章页面进行人工分类。DocLayNet 数据集 [13]便是此类工作的一个例子。这项大规模标注工作展示了：为了生成大量多样化的文档供机器学习模型使用，需要满足许多后勤和资源方面的要求。

DocLayNet 数据集包含 80,000 多个经过人工标注的页面。该数据集由 ~40名受监督的标注员和一小组专家耗时 6 个多月完成，而其中只有“很小一部分”页面由多名标注员标注 [13]。此外，尽管标注过程包括一份 >100 页面标注指南和为期 12 周的培训期，但在这“很小一部分”经过双重和三重标注的页面中，针对其 11 种页面对象类别的编码员间一致性仅为 ≈80-85%[13]。

DocLayNet 数据集可谓是文档布局分析领域最先进的人工标注数据集之一。生成该数据集需要投入大量的后勤资源和时间，而各类别的标注者间一致性仅为 80% 至 85%，这充分说明这些手工标注任务难以扩展。与我们的工作相关的是，该数据集

不包含扫描页面——DocLayNet 的作者指出，这些页面通常存在扭曲或颜色不均匀的问题，因而会给标注过程带来更多潜在的不确定性和错误 [13]。

3.3 最少历史合成数据

鉴于制作 DocLayNet 等数据集所需的时间，另一种选择是生成可基于文章源文件构建的“合成数据”。

大型“基准”数据集是通过挖掘 XML [（例如 PubLayNet 17] ）和 LaTeX 源文件创建的，且可能采用或不采用弱监督 [（例如 TableBank 和 DocBank 22, 23]）。然而，此类数据集中的大多数文章都是近期发表的——例如，PubLayNet 数据集中只有≈3%的文章发表于 1997 年之前。

鉴于历史合成数据集匮乏，人们曾多次尝试通过在页面上加入人工扭曲、旋转、模拟灰尘和随机噪声等效果，人为地让这些较新的文档“变旧”。虽然许多研究关注的是文章老化对下游任务的影响，例如挖掘历史事件相关 OCR 文本 [24] 或进行命名实体识别 [25], ，但近期也有一些研究关注老化过程对页面对象定位 [10, 11] 的影响，以及为历史文档生成新的训练集 [26]。

合成历史文档数据无疑是扩大训练数据集规模的一条有前景的途径；然而，许多数据集仍然相对较小，而且是针对特定类型的历史文档构建的 [，例如 27] 中主要以书籍为主的数据集。

4.....除非答案在于更好的模型?

鉴于训练数据收集与生成方面存在问题，另一条值得探索的途径是开发更好的模型，以便利用有限的可用训练数据。

一种可能的方法是使用在大型文档布局分析或目标检测数据集上训练的模型，并通过迁移学习对特定历史文档集进行“微调”。有若干文档布局分析研究采用了这一策略[28–30]。不过，对于历史文档，有证据表明，迁移学习对页面对象定位准确率的提升可能微乎其微 [11]。此外，这要求每页的格式都相同，因此更难将新特征纳入训练过程 [12]。

除了迁移学习之外，还有一些模型不再将目标检测直接迁移到文档布局分析任务，而是将文本数据纳入训练特征 [12, 14, 20]。这些模型通常是“多模态”的，因为它们融合了图像分类与分割领域的机器学习方法，以及自然语言处理或类似技术对文本的处理方法 [31]。

此外，人们也在开发专门应对训练数据匮乏问题的模型。这些“单样本”或“少样本”模型有望大幅减少为特定领域创建专用文档数据集所需投入的时间和资源 [32]。

能够利用迁移学习、文本类文档组件且只需少量训练实例的模型组合，有望降低分析特定领域历史文献的“成本”。然而，目前大多数最先进的模型难以达到其他目标检测应用所具备的高准确率水平 [6, 13]。

5 开个玩笑！当然，真正的 答案是所有事情！

鉴于目前并不存在能在不同领域中完美定位页面对象的“灵丹妙药”，提高文档版面分析准确率所需的解决方案，很可能是将这里讨论的多种方法结合成一个流程。

在我们自己的工作中，我们正尝试将针对天体物理学文献调优的模型迁移到 HathiTrust 中更广泛的历史科学语料库，并从三个主要方面着手解决泛化问题：文档特定的

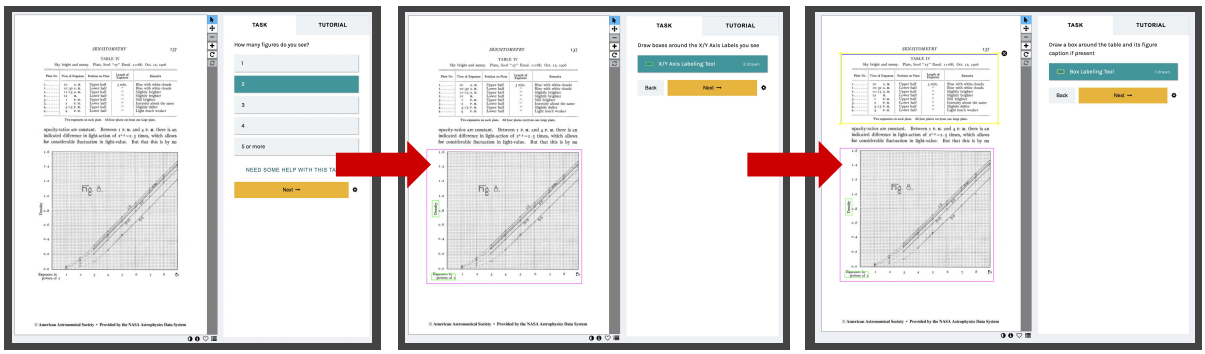

图 3：公民科学平台 Zooniverse 上一个页面对象检测界面的原型示例，展示了该界面的多个步骤：引导标注员先识别页面上的图的数量（左），再为图及其坐标轴标签绘制方框（中），并识别表格等其他有趣的页面对象（右）。我们设计该界面的目标是，高效引导公民科学合作伙伴完成科学文章的标注，从而获得较高的标注员间一致性，并确保该界面能够大规模实施。

机器学习模型、用于扩大标注过程规模的公民科学，以及生成经过适当老化过程处理的大型合成数据集。

我们当前的模型使用经过修改的目标检测模型来检测图及其图注，该模型同时利用光学字符识别、语言和灰度特征 [12, 14]。鉴于其架构（基于 YOLO），它的假阳性率较低；不过，我们计划将其与 detectron2 （基于 Mask/Faster-RCNN）等模型集成。这些模型具有不同的错误特征，因此有望提高模型准确率。此外，我们计划研究迁移学习方案，例如少样本模型 [例如32] ，以减少将模型微调至新领域所需的训练实例数量。

为创建一个经过筛选的金标准数据集，我们目前正在设计一个标注界面，该界面将部署在 Zooniverse[8] 平台上——其原型如图 3所示。Zooniverse 平台汇聚了超过一百万名公民科学家，他们为专业科学家的工作提供帮助。该平台提供了多种选项，能够为与公民科学家成功合作、创建准确标注所需的支持体系提供便利。借助教程、讨论论坛以及通过金标准数据集帮助公民科学家“测试”其知识水平等工具 [33],

[8] https://www.zooniverse.org/

培训标注员的大部分工作都可以大规模开展。由更多人参与标注后，编码员间一致性以及机器学习模型的容错能力等指标便能得到更全面的量化 [34–36]。

最后，我们目前正在更深入地研究科学文章文档的老化过程。尽管这项工作仍处于初步阶段且正在进行中，但我们计划在完成页面老化过程的量化后，利用合成数据集来训练模型。

过去的工作通常集中于提高文档布局分析准确性的某一种主要方法（模型、精选数据集和合成数据），而社区近期的工作则表明，结合这三种方法至关重要。因此，我们目前的工作不仅旨在发表我们的研究成果，也希望分享我们用来指导思路的过程，并期望这些内容能对文档版面分析社区的其他成员有所帮助。

这项工作得到了菲德勒奖学金和美国国家航空航天局天体物理学数据分析计划拨款（20-ADAP20-0225）的支持。

参考文献

[1] Maltese，A.V.，Harsh，J.A.，Svetina，D.：数据可视化素养：探究新手—专家连续体中的数据解读。大学科学期刊

Teaching 45(1),84–90（2015）。出版方：美国国家科学教师协会。访问日期：2022年5月18日

[2] Pepe, A., Goodman, A., Muench, A：ADS全天巡天。载于：Ballester, P., Egret, D., Lorente, N.P.F.（编）第二十一届天文数据分析软件与系统会议。太平洋天文学会会议系列，第461卷，第275页（2012年）[3] 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[4] 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）[5] GROBID。GitHub（2008年至2021年）[6]Bhatt, J., Hashmi, K.A., Afzal, M.Z., Stricker, D：基于深度神经网络的页面图形对象检测综述。《应用科学》11（12）（2021年）。https://doi.org/10.3390/app11125344[7] S aha, R., Mondal, A., Jawahar, C.V：文档图像中的图形对象检测。载于：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[8] Yang, X., Yumer, E., Asente, P., Kraley, M., Kifer, D., Giles, C.L：使用多模态全卷积神经网络学习从文档中提取语义结构。载于：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[9] Siegel, N., Lourie, N., Power, R., Ammar, W：利用远程监督神经网络提取科学图表。arXiv电子预印本，1804–02445（2018年）https://arxiv.org/abs/1804.02445 [cs.DL]

[10]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表的基准数据集。arXiv 预印本，2106–15320（2021年）https://arxiv.org/abs/2106.15320 [cs.CV][11] Kahu, S.Y.：从扫描版电子学位论文中提取图表。硕士论文，弗吉尼亚理工大学（2020）[12] Naiman, J.P., Williams, P.K.G., Goodman, A.：混合栅格与矢量 PDF 的图表及图注提取：利用光学字符识别特征对天文学文献进行数字化。在：Silvello, G., Corcho, O., Manghi, P., Di Nunzio, G.M., Golub, K., Ferro, N., Poggi, A.（编），《数字图书馆理论与实践的联系》，第52–67 页。施普林格，Cham （2022）[13] Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A.S., Staar, P.W.J.：DocLayNet：用于文档版面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 [cs.CV][14] Naiman, J.P., G., W.P.K., Goodman, A.：《The Reading Time Machine》图表定位数据集。未发表论文（2022）[15] Skalski, P.：Make Sense。https://github.com/SkalskiP/make-sense/ （2019）[16]Wu, Y., Kirillov, A., Massa, F., Lo, W.- Y., Girshick, R.：Detectron2。https://github.com/facebookresearch/detectron2 （2019）[17] 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][18] Bochkovskiy, A., Wang, C.-Y., Liao, H.- Y.M.：YOLOv4：目标检测的最佳速度与准确率（2020）

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

[20]Younas, J., Rizvi, S.T.R., Malik, M.I., Shafait, F., Lukowicz, P., Ahmed, S.: Ffd：从文档图像中检测图形和公式。载于：2019年数字图像计算：技术与应用（DICTA），第 1–7 页（2019年）。https://doi.org/10.1109/DICTA47822.2019.8945972

[21]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论文集，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/访问日期：2022-09- 20[22] Li, M., Cui, L., Huang,S., Wei, F., Zhou, M., Li, Z.: TableBank：用于表格检测与识别的基准数据集。arXiv 预印本，1903–01949（2019年）https://arxiv.org/abs/1903.01949 [cs.CV[23]Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.: DocBank：用于文档版面分析的基准数据集。arXiv 预印本，2006–01038（2020年）https://arxiv.org/abs/2006.01038 [cs.CL][24] Boros, E., Nguyen, N.K., Lejeune, G., Doucet, A.: 评估 OCR 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。访问日期：2022-09-28[25] Hamdi,A.,Linhares Pontes, E., Sid`ere, N., Coustaty, M., Doucet, A.: 深入分析 OCR 错误对命名实体识别与链接的影响。《自然语言工程》（2022年）。https://doi.org/10.1017/S1351324922000110。出版方：剑桥大学出版社（CUP）。

访问日期：2022-09-28

[26]Bartz, C., Raetz, H., Otholt, J., Meinel, C., Yang, H.: 风格合成：使用合成数据对历史文档进行语义分割。arXiv 预印本，2107– 06777（2021年）https://arxiv.org/abs/2107.06777 [cs.CV]

[27] Monnier, T., Aubry, M.: docExtractor：即用型历史文档元素提取工具。载于：ICFHR（2020）

[28]Boukhers, Z., Beili, N., Hartmann, T., Goswami,P.,Zafar,M.A.:MexPub：从德国出版物中提取元数据的深度迁移学习。载于：2021年 ACM/IEEE 数字图书馆联合会议 (JCDL)，第 250–253 页（2021年）。https://doi.org/10.1109/JCDL52503.2021.00076

[29]Schreiber, S., Agne, S., Wolf, I., Dengel, A., Ahmed, S.: DeepDeSRT：用于文档图像中表格检测和结构识别的深度学习。载于：2017年 IAPR 第 14 届国际文档分析与识别会议 (ICDAR)，第 1162–1167 页。IEEE，京都（2017年）。https://doi.org/10.1109/ICDAR.2017.192。http://ieeexplore.ieee.org/document/8270123/访问日期：2022-10-30

[30]Dong, H., Cheng, Z., He, X., Zhou, M., Zhou, A., Zhou, F., Liu, A., Han, S., Zhang, D.: 表格预训练：模型架构、预训练目标及下游任务综述。arXiv 预印本，2201–09745 (2022)https://arxiv.org/abs/2201.09745 [cs.CL]

[31]Boukhers, Z., Bouabdallah, A.：用于从科学PDF文档中提取元数据的视觉与自然语言：一种多模态方法。载于：第22届ACM/IEEE数字图书馆联合会议论文集，第1–5页。ACM，德国科隆（2022）。https://doi.org/10.1145/3529372.3533295。https://dl.acm.org/doi/10.1145/3529372.3533295访问日期：2022-10-27[32]Singh, P., Varadarajan, S., Singh, A.N.,

Srivastava, M.M.：基于少样本目标检测的多领域文档布局理解。载于：Campilho, A., Karray, F., Wang, Z.（编）《图像分析与识别》。计算机科学讲义，第89–99页。施普林格，Cham（2020）。
https://doi.org/10.1007/978-3-030-50516-5 8

[33]Eisner, N.L., Barrag´an, O., Lintott, C., Aigrain, S., Nicholson, B., Boyajian, T.S., Howell, S., Johnston, C., Lakeland, B., Miller, G., McMaster, A., Parviainen, H., Safron, E.J., Schwamb, M.E., Trouille, L., Vaughan, S., Zicher, N., Allen, C., Allen, S., Bouslog, M., Johnson, C., Simon, M.N., Wolfenbarger, Z., Baeten, E.M.L., Bundy, D.M., Hoffman, T：行星猎人TESS II：TESS 首两年的发现。《皇家天文学会月报》**501**(4), 4669–4690 (2021年)
https://arxiv.org/abs/2011.13944 [天体物理学.EP]。
https://doi.org/10.1093/mnras/staa3739

[34]Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., Nichol, R.C., Sza- lay, A., Andreescu, D., Murray, P., Vandenberg, J.：《星系动物园》：通过目视检查斯隆数字巡天中的星系得出的形态。《皇家天文学会月报》**389**(3), 1179–1189（2008）
https://arxiv.org/abs/0804.4483 [astro-ph]。
https://doi.org/10.1111/j.1365-2966.2008.13689.x

[35] Schwamb, M.E., Lintott,C.J., Fischer, D.A., Giguere, M.J., Lynn, S., Smith, A.M., Brewer, J.M., Parrish, M., Schawinski, K., Simpson, R.J.：《行星猎人：评估开普勒短周期行星清单》。《天体物理学杂志》**754**(2), 129（2012年）
https://arxiv.org/abs/1205.6769 [astro-ph.EP]。
https://doi.org/10.1088/0004-637X/754/2/129

[36] Johnson, L.C., Seth, A.C., Dalcanton, J.J., Wallace, M.L., Simpson, R.J., Lintott, C.J., Kapadia, A., Skillman, E.D., Caldwell, N., Fouesneau, M., Weisz, D.R.,Williams, B.F., Beerman, L.C., Gouliermis, D.A., Saraje- dini, A：《PHAT恒星团巡天》II：仙女座项目星团目录。《天体物理学杂志》**802**(2), 127（2015）https://arxiv.org/abs/1501.04966 [astro-ph.GA]。
https://doi.org/10.1088/0004-637X/802/2/127