DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis

Birgit Pfitzmann
IBM Research
Rueschlikon, Switzerland
bpf@zurich.ibm.com

Christoph Auer
IBM Research
Rueschlikon, Switzerland
cau@zurich.ibm.com

Michele Dolfi
IBM Research
Rueschlikon, Switzerland
dol@zurich.ibm.com

Ahmed S. Nassar
IBM Research
Rueschlikon, Switzerland
ahn@zurich.ibm.com

Peter Staar
IBM Research
Rueschlikon, Switzerland
taa@zurich.ibm.com

ABSTRACT

Accurate document layout analysis is a key requirement for high-quality PDF document conversion. With the recent availability of public, large ground-truth datasets such as PubLayNet and DocBank, deep-learning models have proven to be very effective at layout detection and segmentation. While these datasets are of adequate size to train such models, they severely lack in layout variability since they are sourced from scientific article repositories such as PubMed and arXiv only. Consequently, the accuracy of the layout segmentation drops significantly when these models are applied on more challenging and diverse layouts. In this paper, we present *DocLayNet*, a new, publicly available, document-layout annotation dataset in COCO format. It contains 80863 manually annotated pages from diverse data sources to represent a wide variability in layouts. For each PDF page, the layout annotations provide labelled bounding-boxes with a choice of 11 distinct classes. DocLayNet also provides a subset of double- and triple-annotated pages to determine the inter-annotator agreement. In multiple experiments, we provide baseline accuracy scores (in mAP) for a set of popular object detection models. We also demonstrate that these models fall approximately 10% behind the inter-annotator agreement. Furthermore, we provide evidence that DocLayNet is of sufficient size. Lastly, we compare models trained on PubLayNet, DocBank and DocLayNet, showing that layout predictions of the DocLayNet-trained models are more robust and thus the preferred choice for general-purpose document-layout analysis.

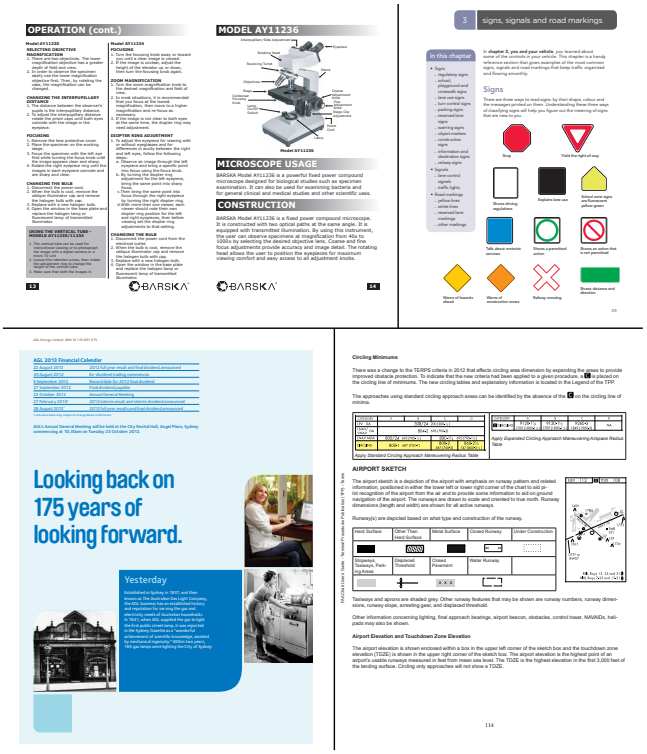

Figure 1: Four examples of complex page layouts across different document categories

CCS CONCEPTS

• **Information systems** → **Document structure**; • **Applied computing** → **Document analysis**; • **Computing methodologies** → **Machine learning**; **Computer vision**; *Object detection*;

KEYWORDS

PDF document conversion, layout segmentation, object-detection, data set, Machine Learning

ACM Reference Format:
Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter Staar. 2022. DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539043

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3539043

DocLayNet：一个用于文档布局分析的大型人工标注数据集

毕尔吉特·皮茨曼
IBM 研究 瑞士吕施利贡 bpf@zurich.ibm.com

克里斯托夫·奥尔
IBM研究 瑞士吕施利孔 cau@zurich.ibm.com

米歇尔·多尔菲 IBM 研究 苏黎世理工，瑞士 dol@zurich.ibm.com

纳瑟尔·阿赫迈德 S. IBM 研究院 瑞士吕施利孔 ahn@zurich.ibm.com

Peter Staar IBM 研究 瑞士吕施利贡 taa@zurich.ibm.com

摘要

精确的文档布局分析是高质量PDF文档转换的关键要求。随着PubLayNet和DocBank等大型公共真实数据集的出现，深度学习模型在布局检测和分割方面已被证明非常有效。虽然这些数据集的规模足够训练此类模型，但由于仅来自PubMed和arXiv等科学文章库，因此在布局多样性方面严重不足。因此，当这些模型应用于更具挑战性和多样化的布局时，布局分割的准确性会显著下降。在本文中，我们介绍了DocLayNet，这是一个新公开可用的COCO格式的文档布局标注数据集。它包含从多样化数据源中手动标注的80863页，以代表布局的广泛多样性。对于每个PDF页面，布局标注提供了标记的边界框，并有11个不同类别的选择。DocLayNet还提供了一部分双重和三重标注的页面以确定标注者之间的一致性。在多个实验中，我们提供了一组流行目标检测模型的基准准确度分数（以mAP表示）。我们还展示了这些模型与标注者之间的一致性约相差10%。此外，我们提供了DocLayNet规模足够大的证据。最后，我们比较了在PubLayNet、DocBank和DocLayNet上训练的模型，结果显示DocLayNet训练的模型的布局预测更加稳健，因此是通用文档布局分析的首选。

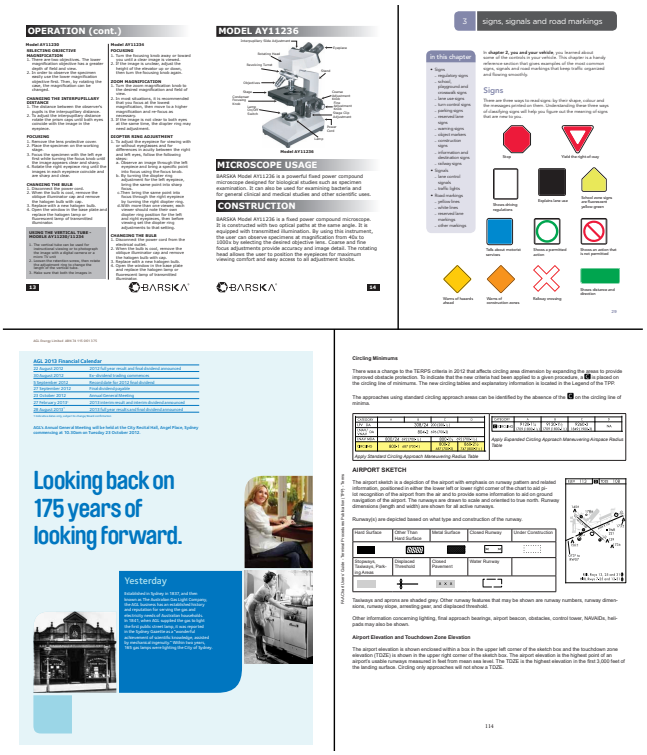

图 1：不同文档类别中复杂页面布局的四个示例

CCS 概念

• 信息系统 → 文档结构；• 应用计算 → 文档分析；• 计算方法论 → 机器学习；计算机视觉；物体检测；

关键词

PDF文档转换，布局分割，目标检测，数据集，机器学习

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3539043

ACM参考格式：
Birgit Pfitzmann、Christoph Auer、Michele Dolfi、Ahmed S. Nassar 和 Peter Staar。2022年。DocLayNet：用于文档布局分析的大型人工标注数据集。发表于第28届ACM SIGKDD知识发现与数据挖掘会议（KDD '22），2022年8月14-18日，美国华盛顿特区。ACM, New York, NY, USA, 共9页。https://doi.org/10.1145/3534678.3539043

1 INTRODUCTION

Despite the substantial improvements achieved with machine-learning (ML) approaches and deep neural networks in recent years, document conversion remains a challenging problem, as demonstrated by the numerous public competitions held on this topic [1–4]. The challenge originates from the huge variability in PDF documents regarding layout, language and formats (scanned, programmatic or a combination of both). Engineering a single ML model that can be applied on all types of documents and provides high-quality layout segmentation remains to this day extremely challenging [5]. To highlight the variability in document layouts, we show a few example documents from the DocLayNet dataset in Figure 1.

A key problem in the process of document conversion is to understand the structure of a single document page, i.e. which segments of text should be grouped together in a unit. To train models for this task, there are currently two large datasets available to the community, PubLayNet [6] and DocBank [7]. They were introduced in 2019 and 2020 respectively and significantly accelerated the implementation of layout detection and segmentation models due to their sizes of 300K and 500K ground-truth pages. These sizes were achieved by leveraging an automation approach. The benefit of automated ground-truth generation is obvious: one can generate large ground-truth datasets at virtually no cost. However, the automation introduces a constraint on the variability in the dataset, because corresponding structured source data must be available. PubLayNet and DocBank were both generated from scientific document repositories (PubMed and arXiv), which provide XML or LaTeX sources. Those scientific documents present a limited variability in their layouts, because they are typeset in uniform templates provided by the publishers. Obviously, documents such as technical manuals, annual company reports, legal text, government tenders, etc. have very different and partially unique layouts. As a consequence, the layout predictions obtained from models trained on PubLayNet or DocBank is very reasonable when applied on scientific documents. However, for more *artistic* or *free-style* layouts, we see sub-par prediction quality from these models, which we demonstrate in Section 5.

In this paper, we present the DocLayNet dataset. It provides page-by-page layout annotation ground-truth using bounding-boxes for 11 distinct class labels on 80863 unique document pages, of which a fraction carry double- or triple-annotations. DocLayNet is similar in spirit to PubLayNet and DocBank and will likewise be made available to the public[1] in order to stimulate the document-layout analysis community. It distinguishes itself in the following aspects:

(1) *Human Annotation*: In contrast to PubLayNet and DocBank, we relied on human annotation instead of automation approaches to generate the data set.
(2) *Large Layout Variability*: We include diverse and complex layouts from a large variety of public sources.
(3) *Detailed Label Set*: We define 11 class labels to distinguish layout features in high detail. PubLayNet provides 5 labels; DocBank provides 13, although not a superset of ours.
(4) *Redundant Annotations*: A fraction of the pages in the DocLayNet data set carry more than one human annotation. This enables experimentation with annotation uncertainty and quality control analysis.
(5) *Pre-defined Train-, Test- & Validation-set*: Like DocBank, we provide fixed train-, test- & validation-sets to ensure proportional representation of the class-labels. Further, we prevent leakage of unique layouts across sets, which has a large effect on model accuracy scores.

All aspects outlined above are detailed in Section 3. In Section 4, we will elaborate on how we designed and executed this large-scale human annotation campaign. We will also share key insights and lessons learned that might prove helpful for other parties planning to set up annotation campaigns.

In Section 5, we will present baseline accuracy numbers for a variety of object detection methods (Faster R-CNN, Mask R-CNN and YOLOv5) trained on DocLayNet. We further show how the model performance is impacted by varying the DocLayNet dataset size, reducing the label set and modifying the train/test-split. Last but not least, we compare the performance of models trained on PubLayNet, DocBank and DocLayNet and demonstrate that a model trained on DocLayNet provides overall more robust layout recovery.

2 RELATED WORK

While early approaches in document-layout analysis used rule-based algorithms and heuristics [8], the problem is lately addressed with deep learning methods. The most common approach is to leverage object detection models [9–15]. In the last decade, the accuracy and speed of these models has increased dramatically. Furthermore, most state-of-the-art object detection methods can be trained and applied with very little work, thanks to a standardisation effort of the ground-truth data format [16] and common deep-learning frameworks [17]. Reference data sets such as PubLayNet [6] and DocBank provide their data in the commonly accepted COCO format [16].

Lately, new types of ML models for document-layout analysis have emerged in the community [18–21]. These models do not approach the problem of layout analysis purely based on an image representation of the page, as computer vision methods do. Instead, they combine the text tokens and image representation of a page in order to obtain a segmentation. While the reported accuracies appear to be promising, a broadly accepted data format which links geometric and textual features has yet to establish.

3 THE DOCLAYNET DATASET

DocLayNet contains 80863 PDF pages. Among these, 7059 carry two instances of human annotations, and 1591 carry three. This amounts to 91104 total annotation instances. The annotations provide layout information in the shape of labeled, rectangular bounding-boxes. We define 11 distinct labels for layout features, namely *Caption*, *Footnote*, *Formula*, *List-item*, *Page-footer*, *Page-header*, *Picture*, *Section-header*, *Table*, *Text*, and *Title*. Our reasoning for picking this particular label set is detailed in Section 4.

In addition to open intellectual property constraints for the source documents, we required that the documents in DocLayNet adhere to a few conditions. Firstly, we kept scanned documents

[1] https://developer.ibm.com/exchanges/data/all/doclaynet

1 介绍

尽管近年来机器学习（ML）方法和深度神经网络取得了实质性改进，文档转换仍然是一个具有挑战性的问题，众多关于这一主题的公开竞赛证明了这一点[1-4]。挑战源于PDF文档在布局、语言和格式（扫描的、程序生成的或两者的组合）方面的巨大差异。至今为止，设计一个可以应用于所有类型文档并提供高质量布局分割的单一ML模型仍然极其困难[5]。为了突出文档布局的多样性，我们在图1中展示了来自DocLayNet数据集的一些示例文档。

在文档转换过程中，一个关键问题是理解单个文档页面的结构，即哪些文本段落应当在一个单元中进行分组。为了训练该任务的模型，目前社区中有两个大型数据集可用，分别是 PubLayNet [6] 和 DocBank [7]。它们分别于2019年和2020年推出，由于其大小分别为30万和50万页的真实数据，因此显著加速了布局检测和分割模型的实施。这些规模是通过利用自动化方法实现的。自动生成真实数据的好处显而易见：几乎可以零成本生成大型真实数据集。然而，自动化给数据集的多样性引入了一定的限制，因为必须有相应的结构化源数据可用。PubLayNet 和 DocBank 都是从科学文档库（PubMed 和 arXiv）生成的，这些库提供 XML 或 LATEX 源。这些科学文档呈现出有限的布局多样性，因为它们是由出版商提供的统一模板排版的。显然，技术手册、公司年度报告、法律文本、政府招标文件等文档在布局上非常不同，并且部分具有独特性。因此，从 PubLayNet 或 DocBank 训练的模型获得的布局预测在应用于科学文档时非常合理。然而，对于更具艺术性或自由风格的布局，这些模型的预测质量较差，我们在第5节中对此进行演示。

在本文中，我们介绍了DocLayNet数据集。它为80863个独特的文档页面提供逐页布局注释真值，使用边界框标注11个不同类别标签，其中一部分页面具有双重或三重注释。DocLayNet在精神上类似于PubLayNet和DocBank，并将同样公开提供给公众，以激励文档布局分析社区。它在以下方面具有独特之处：

(1) 人工标注：与 PubLayNet 和 DocBank 相比，我们依靠人工标注而不是自动化方法来生成数据集。 (2) 大布局多样性：我们包含了来自大量公共来源的多样且复杂的布局。 (3) 详细的标签集：我们定义了11个类别标签，以高度详细地区分布局特征。PubLayNet 提供了5个标签；DocBank 提供了13个标签，尽管并不是我们的超集。 (4) 冗余标注：在 DocLayNet 数据集中，一部分页面有一个以上的人工标注。这使得可以对注释的不确定性和质量控制分析进行实验。(5) 预定义的训练集、测试集和验证集：与 DocBank 类似，我们提供固定的训练集、测试集和验证集，以确保类别标签的比例代表性。此外，我们防止在各个集合之间出现独特布局的泄漏，这对模型准确性分数有很大影响。

上述所有方面在第3节中进行了详细说明。在第4节中，我们将详细阐述我们如何设计和执行这次大规模的人类标注活动。我们还将分享关键的见解和经验教训，这些可能对计划开展标注活动的其他方有所帮助。

在第 5 节中，我们将展示在 DocLayNet 上训练的多种目标检测方法（Faster R-CNN、Mask R-CNN 和 YOLOv5）的基线准确性数据。我们进一步展示了模型性能如何受到 DocLayNet 数据集大小变化、减少标签集和修改训练/测试拆分的影响。最后但并非最不重要的，我们比较了在 PubLayNet、DocBank 和 DocLayNet 上训练的模型的性能，并证明在 DocLayNet 上训练的模型可以提供整体更稳健的布局恢复。

2 相关工作

虽然早期的文档布局分析方法使用基于规则的算法和启发式方法[8]，但最近该问题通过深度学习方法得到了解决。最常见的方法是利用对象检测模型[9 – 15]。在过去的十年中，这些模型的准确性和速度都显著提高。此外，由于对真实数据格式的标准化努力[16]及通用深度学习框架[17]，大多数最先进的目标检测方法可以通过非常少的工作进行训练和应用。引用数据集如PubLayNet[6]和DocBank以普遍接受的COCO格式提供他们的数据[16]。

最近，社区中出现了用于文档布局分析的新型机器学习模型[18 – 21]。这些模型并不像计算机视觉方法那样，仅仅基于页面的图像表示来处理布局分析问题。相反，它们结合了页面的文本标记和图像表示，以获得分割。尽管所报告的准确率似乎很有前景，但尚未建立一个广泛接受的链接几何和文本特征的数据格式。

3 DOCLAYNET 数据集

DocLayNet 包含 80863 个 PDF 页面。其中，7059 个页面有两组人工标注，1591 个页面有三组。这总计 91104 个标注实例。标注以标记的矩形边界框的形式提供布局信息。我们定义了 11 个不同的布局特征标签，分别是 Caption（标题）、Footnote（脚注）、Formula（公式）、List-item（列表项）、Page-footer（页面底部）、Page-header（页面顶部）、Picture（图片）、Section-header（节标题）、Table（表格）、Text（文本）和 Title（标题）。我们选择这组特定标签的原因在第 4 节详细说明。

除了源文档开放获取的限制外，我们还要求 DocLayNet 中的文档符合一些条件。首先，我们保留扫描文档

[1]https://developer.ibm.com/exchanges/data/all/doclaynet

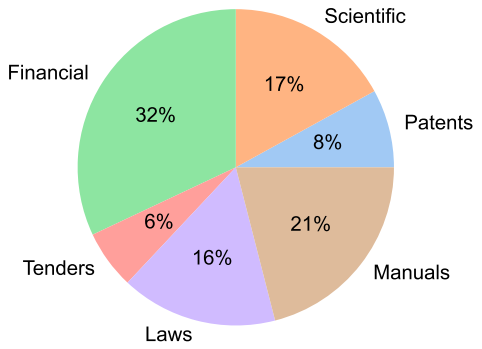

Figure 2: Distribution of DocLayNet pages across document categories.

to a minimum, since they introduce difficulties in annotation (see Section 4). As a second condition, we focussed on medium to large documents (> 10 pages) with technical content, dense in complex tables, figures, plots and captions. Such documents carry a lot of information value, but are often hard to analyse with high accuracy due to their challenging layouts. Counterexamples of documents not included in the dataset are receipts, invoices, hand-written documents or photographs showing "text in the wild".

The pages in DocLayNet can be grouped into six distinct categories, namely *Financial Reports*, *Manuals*, *Scientific Articles*, *Laws & Regulations*, *Patents* and *Government Tenders*. Each document category was sourced from various repositories. For example, Financial Reports contain both *free-style* format annual reports[2] which expose company-specific, artistic layouts as well as the more formal SEC filings. The two largest categories (*Financial Reports* and *Manuals*) contain a large amount of free-style layouts in order to obtain maximum variability. In the other four categories, we boosted the variability by mixing documents from independent providers, such as different government websites or publishers. In Figure 2, we show the document categories contained in DocLayNet with their respective sizes.

We did not control the document selection with regard to language. The vast majority of documents contained in DocLayNet (close to 95%) are published in English language. However, DocLayNet also contains a number of documents in other languages such as German (2.5%), French (1.0%) and Japanese (1.0%). While the document language has negligible impact on the performance of computer vision methods such as object detection and segmentation models, it might prove challenging for layout analysis methods which exploit textual features.

To ensure that future benchmarks in the document-layout analysis community can be easily compared, we have split up DocLayNet into pre-defined train-, test- and validation-sets. In this way, we can avoid spurious variations in the evaluation scores due to random splitting in train-, test- and validation-sets. We also ensured that less frequent labels are represented in train and test sets in equal proportions.

[2] e.g. AAPL from https://www.annualreports.com/

Table 1 shows the overall frequency and distribution of the labels among the different sets. Importantly, we ensure that subsets are only split on full-document boundaries. This avoids that pages of the same document are spread over train, test and validation set, which can give an undesired evaluation advantage to models and lead to overestimation of their prediction accuracy. We will show the impact of this decision in Section 5.

In order to accommodate the different types of models currently in use by the community, we provide DocLayNet in an *augmented* COCO format [16]. This entails the standard COCO ground-truth file (in JSON format) with the associated page images (in PNG format, 1025×1025 pixels). Furthermore, custom fields have been added to each COCO record to specify document category, original document filename and page number. In addition, we also provide the original PDF pages, as well as sidecar files containing parsed PDF text and text-cell coordinates (in JSON). All additional files are linked to the primary page images by their matching filenames.

Despite being cost-intense and far less scalable than automation, human annotation has several benefits over automated ground-truth generation. The first and most obvious reason to leverage human annotations is the freedom to annotate any type of document without requiring a programmatic source. For most PDF documents, the original source document is not available. The latter is not a hard constraint with human annotation, but it is for automated methods. A second reason to use human annotations is that the latter usually provide a more natural interpretation of the page layout. The human-interpreted layout can significantly deviate from the programmatic layout used in typesetting. For example, "invisible" tables might be used solely for aligning text paragraphs on columns. Such typesetting tricks might be interpreted by automated methods incorrectly as an actual table, while the human annotation will interpret it correctly as *Text* or other styles. The same applies to multi-line text elements, when authors decided to space them as "invisible" list elements without bullet symbols. A third reason to gather ground-truth through human annotation is to estimate a "natural" upper bound on the segmentation accuracy. As we will show in Section 4, certain documents featuring complex layouts can have different but equally acceptable layout interpretations. This natural upper bound for segmentation accuracy can be found by annotating the same pages multiple times by different people and evaluating the inter-annotator agreement. Such a baseline consistency evaluation is very useful to define expectations for a good target accuracy in trained deep neural network models and avoid overfitting (see Table 1). On the flip side, achieving high annotation consistency proved to be a key challenge in human annotation, as we outline in Section 4.

4 ANNOTATION CAMPAIGN

The annotation campaign was carried out in four phases. In phase one, we identified and prepared the data sources for annotation. In phase two, we determined the class labels and how annotations should be done on the documents in order to obtain maximum consistency. The latter was guided by a detailed requirement analysis and exhaustive experiments. In phase three, we trained the annotation staff and performed exams for quality assurance. In phase four,

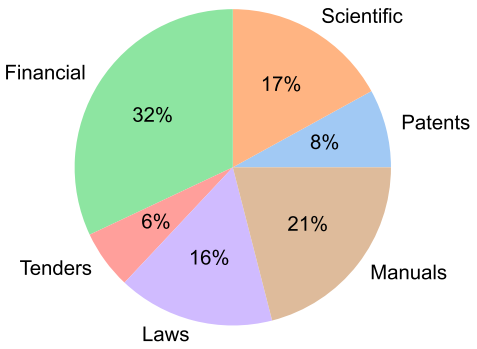

图 2：DocLayNet 页面在文档类别中的分布。

将其降至最低，因为它们在注释中引入了困难（见第4节）。作为第二个条件，我们专注于内容技术性强的中大型文档（> 10页），其中包含复杂的表格、图形、图标和说明。此类文档承载了大量信息价值，但由于其排版复杂，往往难以精确分析。未包含在数据集中的文档反例包括收据、发票、手写文档或显示“ 野外文本 ”的照片。

在DocLayNet中，页面可以分为六个不同的类别，即财务报告、手册、科学文章、法律法规、专利和政府招标。每个文档类别来源于不同的存储库。例如，财务报告包含既有公司特定的、艺术布局的自由格式年度报告2，也有更正式的SEC文件。最大的两个类别（财务报告和手册）包含大量自由格式布局，以获得最大的多样性。在其他四个类别中，我们通过混合来自不同提供者的文档，如不同的政府网站或出版商，提升了多样性。在图2中，我们展示了DocLayNet中包含的文档类别及其相应的大小。

我们没有根据语言控制文档选择。DocLayNet 中的绝大多数文档（接近 95%）以英语发布。然而，DocLayNet 也包含一些其他语言的文档，如德语（2.5%）、法语（1.0%）和日语（1.0%）。虽然文档语言对计算机视觉方法的性能影响可以忽略不计，如对象检测和分割模型，但对于利用文本特征的布局分析方法来说，可能会是一个挑战。

为了确保文档布局分析社区中的未来基准可以轻松比较，我们将 DocLayNet 分割为预定义的训练集、测试集和验证集。通过这种方式，我们可以避免由于在训练集、测试集和验证集中的随机分割而导致的评估分数的虚假变化。我们还确保在训练集和测试集中，较少出现的标签以相同比例表示。

[2]e.g. AAPL from https://www.annualreports.com/

表 1 显示了不同集合中标签的总体频率和分布。重要的是，我们确保子集仅在完整文档边界上进行拆分。这样可以避免同一文档的页面分散在训练集、测试集和验证集中，从而使模型获得不希望的评估优势，并导致其预测准确性的过高估计。我们将在第 5 节中展示这一决定的影响。

为了适应社区目前使用的不同类型的模型，我们以扩展的COCO格式[16]提供DocLayNet。这包括标准的COCO真实值文件（JSON格式）和关联的页面图像（PNG格式，1025×1025像素）。此外，还为每个COCO记录添加了自定义字段，以指定文档类别、原始文档文件名和页码。此外，我们还提供原始PDF页面，以及包含解析的PDF文本和文本单元格坐标的旁车文件（JSON格式）。所有附加文件都通过匹配的文件名链接到主要页面图像。

尽管相对于自动化，人类标注成本高且扩展性差得多，但在人类标注相较于自动生成的真实数据方面有几个优势。利用人类标注的第一个也是最显而易见的原因是，可以自由标注任何类型的文档而不需要程序化来源。对于大多数PDF文档，原始的源文档通常不可用。对于人类标注来说，这并不是一个严格的约束，但对于自动化方法却是如此。使用人类标注的第二个原因是后者通常能提供更自然的页面布局解释。人类解释的布局可能会显著偏离排版中使用的程序化布局。例如，“ 隐形 ”表格可能仅用于对齐栏中文本段落。这种排版技巧可能会被自动化方法错误解释为实际的表格，而人类标注会正确地将其解释为文本或其他样式。同样的情况适用于多行文本元素，当作者决定将它们间隔为没有符号的“ 隐形 ”列表元素时。通过人类标注收集真实数据的第三个原因是估计分段准确性的“ 自然 ”上限。正如我们将在第4节中展示的，某些具有复杂布局的文档可能具有不同但同样可接受的布局解释。通过不同的人多次标注相同页面并评估标注者之间的一致性，可以找到分段准确性的这种自然上限。这种基线一致性评估对于定义训练深度神经网络模型良好目标准确性的期望值并避免过拟合非常有用（见表1）。另一方面，实现高标注一致性被证明是人类标注的一个关键挑战，正如我们在第4节中概述的那样。

4 标注活动

标注活动分为四个阶段进行。在第一阶段，我们识别并准备了用于标注的数据源。在第二阶段，我们确定了类别标签以及如何在文档上进行标注以获得最大的一致性。后者是由详细的需求分析和全面的实验指导的。在第三阶段，我们对标注人员进行了培训，并通过考试进行质量保证。在第四阶段，

Table 1: DocLayNet dataset overview. Along with the frequency of each class label, we present the relative occurrence (as % of row "Total") in the train, test and validation sets. The inter-annotator agreement is computed as the mAP@0.5-0.95 metric between pairwise annotations from the triple-annotated pages, from which we obtain accuracy ranges.

		% of Total			triple inter-annotator mAP @ 0.5-0.95 (%)						
class label	Count	Train	Test	Val	All	Fin	Man	Sci	Law	Pat	Ten
Caption	22524	2.04	1.77	2.32	84-89	40-61	86-92	94-99	95-99	69-78	n/a
Footnote	6318	0.60	0.31	0.58	83-91	n/a	100	62-88	85-94	n/a	82-97
Formula	25027	2.25	1.90	2.96	83-85	n/a	n/a	84-87	86-96	n/a	n/a
List-item	185660	17.19	13.34	15.82	87-88	74-83	90-92	97-97	81-85	75-88	93-95
Page-footer	70878	6.51	5.58	6.00	93-94	88-90	95-96	100	92-97	100	96-98
Page-header	58022	5.10	6.70	5.06	85-89	66-76	90-94	98-100	91-92	97-99	81-86
Picture	45976	4.21	2.78	5.31	69-71	56-59	82-86	69-82	80-95	66-71	59-76
Section-header	142884	12.60	15.77	12.85	83-84	76-81	90-92	94-95	87-94	69-73	78-86
Table	34733	3.20	2.27	3.60	77-81	75-80	83-86	98-99	58-80	79-84	70-85
Text	510377	45.82	49.28	45.00	84-86	81-86	88-93	89-93	87-92	71-79	87-95
Title	5071	0.47	0.30	0.50	60-72	24-63	50-63	94-100	82-96	68-79	24-56
Total	1107470	941123	99816	66531	82-83	71-74	79-81	89-94	86-91	71-76	68-85

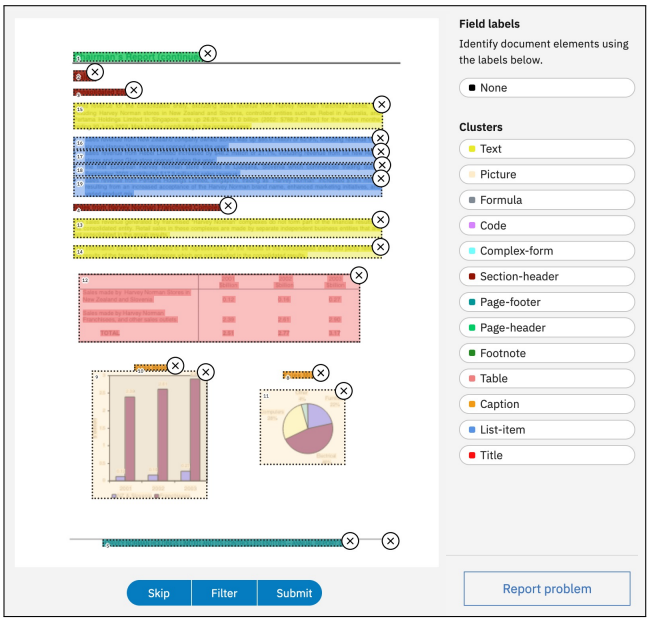

Figure 3: Corpus Conversion Service annotation user interface. The PDF page is shown in the background, with overlaid text-cells (in darker shades). The annotation boxes can be drawn by dragging a rectangle over each segment with the respective label from the palette on the right.

we distributed the annotation workload and performed continuous quality controls. Phase one and two required a small team of experts only. For phases three and four, a group of 40 dedicated annotators were assembled and supervised.

Phase 1: Data selection and preparation. Our inclusion criteria for documents were described in Section 3. A large effort went into ensuring that all documents are free to use. The data sources include publication repositories such as arXiv[3], government offices, company websites as well as data directory services for financial reports and patents. Scanned documents were excluded wherever possible because they can be rotated or skewed. This would not allow us to perform annotation with rectangular bounding-boxes and therefore complicate the annotation process.

Preparation work included uploading and parsing the sourced PDF documents in the Corpus Conversion Service (CCS) [22], a cloud-native platform which provides a visual annotation interface and allows for dataset inspection and analysis. The annotation interface of CCS is shown in Figure 3. The desired balance of pages between the different document categories was achieved by selective subsampling of pages with certain desired properties. For example, we made sure to include the title page of each document and bias the remaining page selection to those with figures or tables. The latter was achieved by leveraging pre-trained object detection models from PubLayNet, which helped us estimate how many figures and tables a given page contains.

Phase 2: Label selection and guideline. We reviewed the collected documents and identified the most common structural features they exhibit. This was achieved by identifying recurrent layout elements and lead us to the definition of 11 distinct class labels. These 11 class labels are *Caption*, *Footnote*, *Formula*, *List-item*, *Page-footer*, *Page-header*, *Picture*, *Section-header*, *Table*, *Text*, and *Title*. Critical factors that were considered for the choice of these class labels were (1) the overall occurrence of the label, (2) the specificity of the label, (3) recognisability on a single page (i.e. no need for context from previous or next page) and (4) overall coverage of the page. Specificity ensures that the choice of label is not ambiguous, while coverage ensures that all meaningful items on a page can be annotated. We refrained from class labels that are very specific to a document category, such as *Abstract* in the *Scientific Articles* category. We also avoided class labels that are tightly linked to the semantics of the text. Labels such as *Author* and *Affiliation*, as seen in DocBank, are often only distinguishable by discriminating on

[3] https://arxiv.org/

表 1：DocLayNet 数据集概况。除了每个类别标签的频率外，我们还展示了其在训练、测试和验证集中的相对出现率（占“总计”行的百分比）。标注者之间的一致性计算为三重标注页面的成对标注之间的 mAP@0.5-0.95 指标，从中我们得出准确性范围。

		% of Total			triple inter-annotator mAP @ 0.5-0.95 (%)						
class label	Count	Train	Test	Val	All	Fin	Man	Sci	Law	Pat	Ten
Caption	22524	2.04	1.77	2.32	84-89	40-61	86-92	94-99	95-99	69-78	n/a
Footnote	6318	0.60	0.31	0.58	83-91	n/a	100	62-88	85-94	n/a	82-97
Formula	25027	2.25	1.90	2.96	83-85	n/a	n/a	84-87	86-96	n/a	n/a
List-item	185660	17.19	13.34	15.82	87-88	74-83	90-92	97-97	81-85	75-88	93-95
Page-footer	70878	6.51	5.58	6.00	93-94	88-90	95-96	100	92-97	100	96-98
Page-header	58022	5.10	6.70	5.06	85-89	66-76	90-94	98-100	91-92	97-99	81-86
Picture	45976	4.21	2.78	5.31	69-71	56-59	82-86	69-82	80-95	66-71	59-76
Section-header	142884	12.60	15.77	12.85	83-84	76-81	90-92	94-95	87-94	69-73	78-86
Table	34733	3.20	2.27	3.60	77-81	75-80	83-86	98-99	58-80	79-84	70-85
Text	510377	45.82	49.28	45.00	84-86	81-86	88-93	89-93	87-92	71-79	87-95
Title	5071	0.47	0.30	0.50	60-72	24-63	50-63	94-100	82-96	68-79	24-56
Total	1107470	941123	99816	66531	82-83	71-74	79-81	89-94	86-91	71-76	68-85

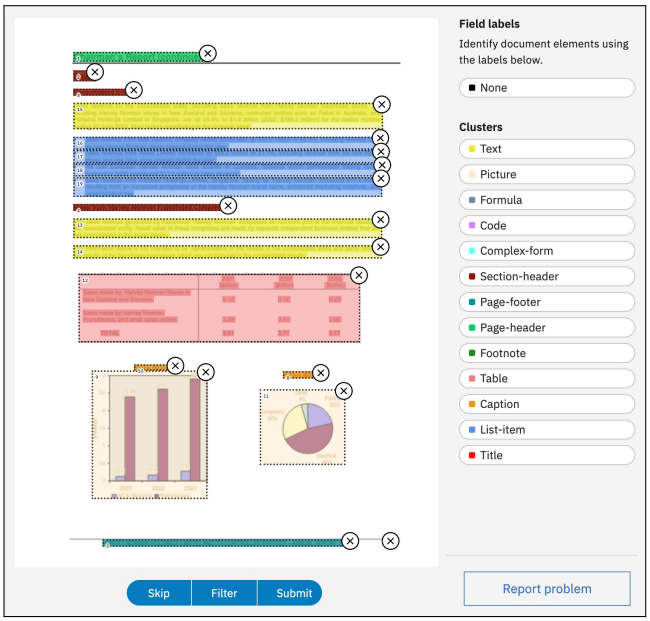

图3：语料库转换服务批注用户界面。PDF页面显示在背景中，叠加着文本单元（较深的阴影）。可以通过从右侧调色板中拖动一个矩形框选取各个段落并加上相应标签来绘制批注框。

我们分配了标注工作量并进行持续的质量控制。第一阶段和第二阶段只需要一个小型的专家团队。对于第三阶段和第四阶段，组建了一个由40名专职标注员组成的团队并进行了监督阶段 1：数据选择和准备。我们在第 3 节中描述了文档的纳入标准。我们花费了大量精力确保所有文档均可免费使用。数据源包括出版物存储库，例如 arXiv3、政府部门、公司网站，以及用于财务报告和专利的数据目录服务。尽可能排除扫描的文档，因为它们可能会旋转或倾斜。这将不允许我们使用矩形边界框进行标注，因此会使标注过程复杂化。

准备工作包括在语料转换服务（CCS）[22]中上传和解析来源的PDF文件，该服务是一个云原生平台，提供可视化标注界面，并允许进行数据集检查和分析。CCS的标注界面如图3所示。通过选择性地子抽样具有某些期望属性的页面，实现了不同文档类别之间页面的期望平衡。例如，我们确保包括每个文档的标题页，并将剩余页面选择偏向于包含图形或表格的页面。后者是通过利用来自PubLayNet的预训练目标检测模型来实现的，这帮助我们估算每个页面包含多少个图形和表格。

第2阶段：标签选择和指南。我们审查了所收集的文档，并识别出它们展示的最常见的结构特征。通过识别重复出现的布局元素，我们定义了11个不同的类别标签。这11个类别标签是：标题、脚注、公式、列表项、页脚、页眉、图片、章节标题、表格、文本和标题。选择这些类别标签时考虑的关键因素是：(1) 标签的整体出现频率；(2) 标签的具体性；(3) 在单页上的可识别性（即不需要前后页的上下文）；(4) 页面整体覆盖率。具体性确保了标签的选择不具歧义性，而覆盖率确保了页面上所有有意义的项目均可被标注。我们避免使用对文档类别非常专有的类别标签，例如，科学文章类别中的摘要。我们还避免使用与文本语义紧密相关的类别标签。在DocBank中，作者和单位等标签通常只能通过区分来识别。

[3]https://arxiv.org/

the textual content of an element, which goes beyond visual layout recognition, in particular outside the *Scientific Articles* category.

At first sight, the task of visual document-layout interpretation appears intuitive enough to obtain plausible annotations in most cases. However, during early trial-runs in the core team, we observed many cases in which annotators use different annotation styles, especially for documents with challenging layouts. For example, if a figure is presented with subfigures, one annotator might draw a single figure bounding-box, while another might annotate each subfigure separately. The same applies for lists, where one might annotate all list items in one block or each list item separately. In essence, we observed that challenging layouts would be annotated in different but plausible ways. To illustrate this, we show in Figure 4 multiple examples of plausible but inconsistent annotations on the same pages.

Obviously, this inconsistency in annotations is not desirable for datasets which are intended to be used for model training. To minimise these inconsistencies, we created a detailed annotation guideline. While perfect consistency across 40 annotation staff members is clearly not possible to achieve, we saw a huge improvement in annotation consistency after the introduction of our annotation guideline. A few selected, non-trivial highlights of the guideline are:

(1) Every list-item is an individual object instance with class label *List-item*. This definition is different from PubLayNet and DocBank, where all list-items are grouped together into one *List* object.
(2) A *List-item* is a paragraph with hanging indentation. Single-line elements can qualify as *List-item* if the neighbour elements expose hanging indentation. Bullet or enumeration symbols are not a requirement.
(3) For every *Caption*, there must be exactly one corresponding *Picture* or *Table*.
(4) Connected sub-pictures are grouped together in one *Picture* object.
(5) Formula numbers are included in a *Formula* object.
(6) Emphasised text (e.g. in italic or bold) at the beginning of a paragraph is not considered a *Section-header*, unless it appears exclusively on its own line.

The complete annotation guideline is over 100 pages long and a detailed description is obviously out of scope for this paper. Nevertheless, it will be made publicly available alongside with DocLayNet for future reference.

Phase 3: Training. After a first trial with a small group of people, we realised that providing the annotation guideline and a set of random practice pages did not yield the desired quality level for layout annotation. Therefore we prepared a subset of pages with two different complexity levels, each with a practice and an exam part. 974 pages were reference-annotated by one proficient core team member. Annotation staff were then given the task to annotate the same subsets (blinded from the reference). By comparing the annotations of each staff member with the reference annotations, we could quantify how closely their annotations matched the reference. Only after passing two exam levels with high annotation quality, staff were admitted into the production phase. Practice iterations

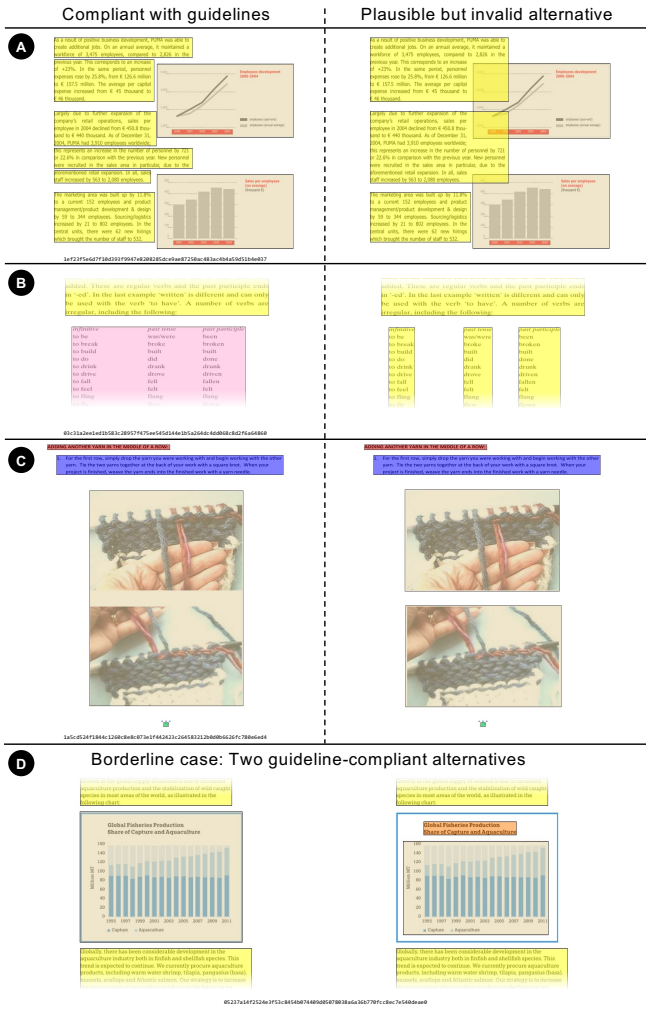

Figure 4: Examples of plausible annotation alternatives for the same page. Criteria in our annotation guideline can resolve cases A to C, while the case D remains ambiguous.

were carried out over a timeframe of 12 weeks, after which 8 of the 40 initially allocated annotators did not pass the bar.

Phase 4: Production annotation. The previously selected 80K pages were annotated with the defined 11 class labels by 32 annotators. This production phase took around three months to complete. All annotations were created online through CCS, which visualises the programmatic PDF text-cells as an overlay on the page. The page annotation are obtained by drawing rectangular bounding-boxes, as shown in Figure 3. With regard to the annotation practices, we implemented a few constraints and capabilities on the tooling level. First, we only allow non-overlapping, vertically oriented, rectangular boxes. For the large majority of documents, this constraint was sufficient and it speeds up the annotation considerably in comparison with arbitrary segmentation shapes. Second, annotator staff were not able to see each other's annotations. This was enforced by design to avoid any bias in the annotation, which could skew the numbers of the inter-annotator agreement (see Table 1). We wanted

一个元素的文本内容，超越视觉布局识别，特别是在科学文章类别之外。

乍一看，视觉文档布局解释的任务似乎直观，足以在大多数情况下获得合理的标注。然而，在核心团队的早期试运行中，我们观察到在许多情况下，标注者使用不同的标注风格，特别是对于布局较复杂的文档。例如，如果一个图形带有子图，有的标注者可能会画一个单一的图形边界框，而另一些则可能会分别标注每个子图。对于列表也是如此，有的人可能会将所有列表项标注在一个块中，而另一些则会分别标注每个列表项。实质上，我们观察到挑战性的布局会被以不同但合理的方式标注。为此，我们在图4中展示了同一页面上多个合理但不一致的标注示例。

显然，对于用于模型训练的数据集来说，这种注释不一致性是不可取的。为了尽量减少这些不一致性，我们创建了一个详细的注释指南。虽然在40名注释员工中达成完美的一致性显然是不可能的，但在引入我们的注释指南后，我们看到了注释一致性的大幅改善。指南中一些精选的、非平凡的亮点是：

(1) 每个列表项都是具有类标签列表项的单独对象实例。与 PubLayNet 和 DocBank 的定义不同，在这些定义中，所有列表项都被组合成一个列表对象。(2) 列表项是具有悬挂缩进的段落。如果相邻元素暴露出悬挂缩进，则单行元素可以被视为列表项。项目符号或枚举符号不是必需的。(3) 对于每个标题，必须有且只有一个对应的图片或表格。(4) 连接的子图片被组合到一个图片对象中。(5) 公式对象中包含了公式编号。(6) 段落开头的强调文本（例如，斜体或粗体）不被视为节标题，除非它单独出现在自己的一行中。

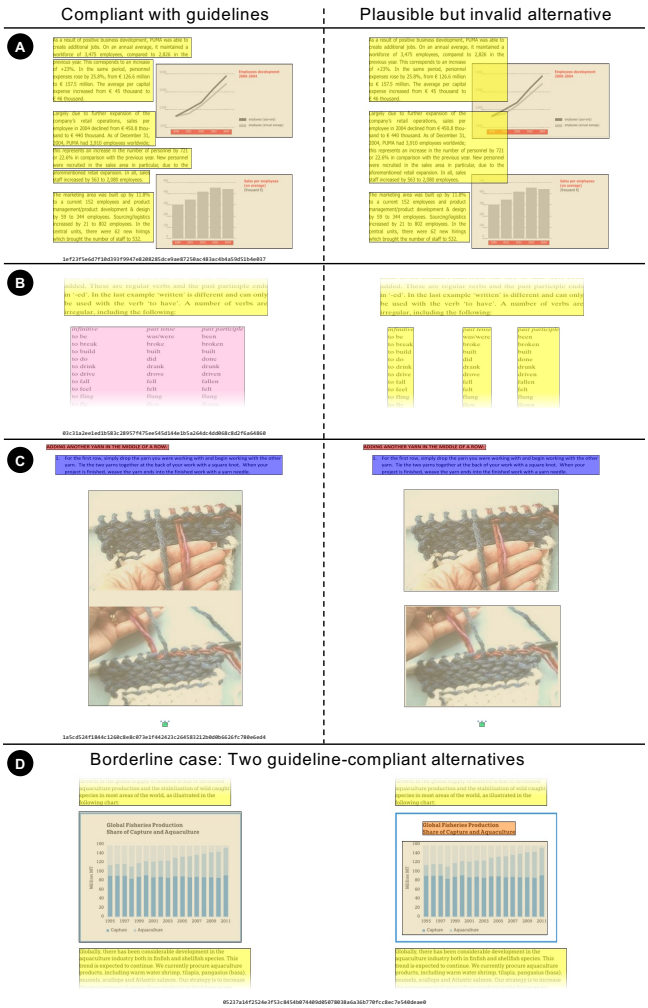

图 4：同一页面的合理注释替代方案示例。我们的注释指南中的标准可以解决案例 A 到 C，而案例 D 仍然存在歧义。

完整的注释指南长度超过100页，显然详细描述超出了本文的范围。不过，它将与DocLayNet一起公开供将来参考。

阶段3：训练。在对一小组人员进行首次尝试后，我们意识到提供注释指南和一组随机练习页面并没有达到所需的版面注释质量水平。因此，我们准备了一部分具有两种不同复杂性水平的页面，每种复杂性水平均包含一个练习部分和一个考试部分。由一名熟练的核心团队成员对974页进行了参考注释。随后，注释人员被分配任务，对相同的页面子集进行注释（参考内容对他们是保密的）。通过将每个员工的注释与参考注释进行比较，我们可以量化他们的注释与参考匹配的紧密程度。只有通过两个考试水平并达到高质量的注释后，员工才能被允许进入生产阶段。实践迭代

在为期12周的时间内进行了这些工作，之后40名最初分配的标注者中有8名未能通过标准。

第4阶段：生产标注。之前选择的80K页由32名标注者使用定义的11个类别标签进行标注。这个生产阶段花费了大约三个月的时间完成。所有标注都是通过CCS在线创建的，CCS将程序化的PDF文本框显示为页面上的覆盖层。页面标注是通过绘制矩形边界框获得的，如图3所示。关于标注实践，我们在工具层面实现了一些限制和功能。首先，我们只允许非重叠、垂直方向的矩形框。对于绝大多数文档来说，这个限制是足够的，并且相比于随意的分段形状，这大大加快了标注速度。其次，标注人员无法看到彼此的标注。这是通过设计强制执行的，以避免标注中的任何偏见，这可能会扭曲标注人员之间协议的数字（见表1）。我们想要

Table 2: Prediction performance (mAP@0.5-0.95) of object detection networks on DocLayNet test set. The MRCNN (Mask R-CNN) and FRCNN (Faster R-CNN) models with ResNet-50 or ResNet-101 backbone were trained based on the network architectures from the *detectron2* model zoo (Mask R-CNN R50, R101-FPN 3x, Faster R-CNN R101-FPN 3x), with default configurations. The YOLO implementation utilized was YOLOv5x6 [13]. All models were initialised using pre-trained weights from the COCO 2017 dataset.

	human	MRCNN		FRCNN	YOLO
		R50	R101	R101	v5x6
Caption	84-89	68.4	71.5	70.1	77.7
Footnote	83-91	70.9	71.8	73.7	77.2
Formula	83-85	60.1	63.4	63.5	66.2
List-item	87-88	81.2	80.8	81.0	86.2
Page-footer	93-94	61.6	59.3	58.9	61.1
Page-header	85-89	71.9	70.0	72.0	67.9
Picture	69-71	71.7	72.7	72.0	77.1
Section-header	83-84	67.6	69.3	68.4	74.6
Table	77-81	82.2	82.9	82.2	86.3
Text	84-86	84.6	85.8	85.4	88.1
Title	60-72	76.7	80.4	79.9	82.7
All	82-83	72.4	73.5	73.4	76.8

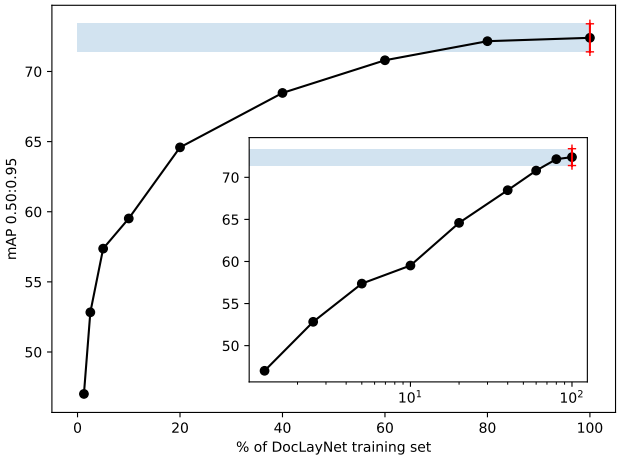

Figure 5: Prediction performance (mAP@0.5-0.95) of a Mask R-CNN network with ResNet50 backbone trained on increasing fractions of the DocLayNet dataset. The learning curve flattens around the 80% mark, indicating that increasing the size of the DocLayNet dataset with similar data will not yield significantly better predictions.

to avoid this at any cost in order to have clear, unbiased baseline numbers for human document-layout annotation. Third, we introduced the feature of *snapping* boxes around text segments to obtain a pixel-accurate annotation and again reduce time and effort. The CCS annotation tool automatically shrinks every user-drawn box to the minimum bounding-box around the enclosed text-cells for all purely text-based segments, which excludes only *Table* and *Picture*. For the latter, we instructed annotation staff to minimise inclusion of surrounding whitespace while including all graphical lines. A downside of snapping boxes to enclosed text cells is that some wrongly parsed PDF pages cannot be annotated correctly and need to be skipped. Fourth, we established a way to flag pages as *rejected* for cases where no valid annotation according to the label guidelines could be achieved. Example cases for this would be PDF pages that render incorrectly or contain layouts that are impossible to capture with non-overlapping rectangles. Such rejected pages are not contained in the final dataset. With all these measures in place, experienced annotation staff managed to annotate a single page in a typical timeframe of 20s to 60s, depending on its complexity.

5 EXPERIMENTS

The primary goal of DocLayNet is to obtain high-quality ML models capable of accurate document-layout analysis on a wide variety of challenging layouts. As discussed in Section 2, object detection models are currently the easiest to use, due to the standardisation of ground-truth data in COCO format [16] and the availability of general frameworks such as *detectron2* [17]. Furthermore, baseline numbers in PubLayNet and DocBank were obtained using standard object detection models such as Mask R-CNN and Faster R-CNN. As such, we will relate to these object detection methods in this paper and leave the detailed evaluation of more recent methods mentioned in Section 2 for future work.

In this section, we will present several aspects related to the performance of object detection models on DocLayNet. Similarly as in PubLayNet, we will evaluate the quality of their predictions using mean average precision (mAP) with 10 overlaps that range from 0.5 to 0.95 in steps of 0.05 (mAP@0.5-0.95). These scores are computed by leveraging the evaluation code provided by the COCO API [16].

Baselines for Object Detection

In Table 2, we present baseline experiments (given in mAP) on Mask R-CNN [12], Faster R-CNN [11], and YOLOv5 [13]. Both training and evaluation were performed on RGB images with dimensions of 1025×1025 pixels. For training, we only used one annotation in case of redundantly annotated pages. As one can observe, the variation in mAP between the models is rather low, but overall between 6 and 10% lower than the mAP computed from the pairwise human annotations on triple-annotated pages. This gives a good indication that the DocLayNet dataset poses a worthwhile challenge for the research community to close the gap between human recognition and ML approaches. It is interesting to see that Mask R-CNN and Faster R-CNN produce very comparable mAP scores, indicating that pixel-based image segmentation derived from bounding-boxes does not help to obtain better predictions. On the other hand, the more recent Yolov5x model does very well and even out-performs humans on selected labels such as *Text*, *Table* and *Picture*. This is not entirely surprising, as *Text*, *Table* and *Picture* are abundant and the most visually distinctive in a document.

表 2：DocLayNet 测试集上目标检测网络的预测性能（mAP@0.5-0.95）。使用 ResNet-50 或 ResNet-101 主干的 MRCNN（Mask R-CNN）和 FRCNN（Faster R-CNN）模型基于 detectron 2 模型库中的网络架构进行了训练（Mask R-CNN R50，R101-FPN 3x，Faster R-CNN R101-FPN 3x），使用默认配置。YOLO 的实现使用的是 YOLOv5x6 [13]。所有模型均使用 COCO 2017 数据集的预训练权重进行初始化。

	human	MRCNN		FRCNN	YOLO
		R50	R101	R101	v5x6
Caption	84-89	68.4	71.5	70.1	77.7
Footnote	83-91	70.9	71.8	73.7	77.2
Formula	83-85	60.1	63.4	63.5	66.2
List-item	87-88	81.2	80.8	81.0	86.2
Page-footer	93-94	61.6	59.3	58.9	61.1
Page-header	85-89	71.9	70.0	72.0	67.9
Picture	69-71	71.7	72.7	72.0	77.1
Section-header	83-84	67.6	69.3	68.4	74.6
Table	77-81	82.2	82.9	82.2	86.3
Text	84-86	84.6	85.8	85.4	88.1
Title	60-72	76.7	80.4	79.9	82.7
All	82-83	72.4	73.5	73.4	76.8

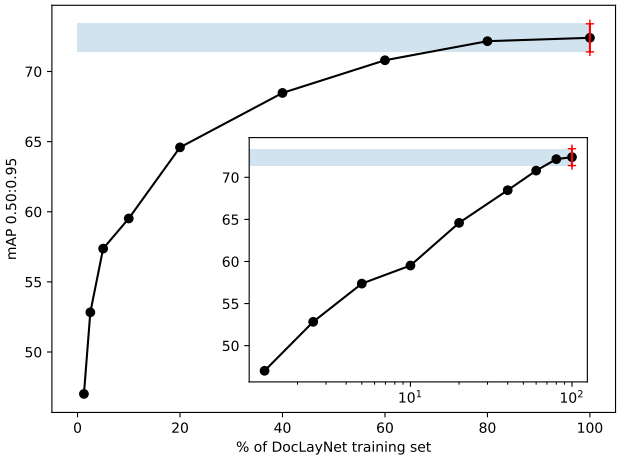

图5：使用ResNet50主干网络的Mask R-CNN网络在不断增加的DocLayNet数据集比例上的预测性能（mAP@0.5-0.95）。学习曲线在约80%处趋于平坦，这表明通过添加相似数据来增加DocLayNet数据集的大小不会显著提高预测效果。

为了以任何代价避免这种情况，以便为人工文档布局注释提供明确且无偏见的基准数据。第三，我们引入了将文本段落周围的框吸附的功能，以获得像素精确的注释，并再次减少时间和精力。CCS注释工具会自动缩小每个用户绘制的框至所有纯文本段落内封闭文本单元的最小边界框，只排除表格和图片。对于后者，我们要求注释人员在包括所有图形线条的同时尽量减少包含周围的空白。将框吸附到封闭的文本单元上的一个缺点是，一些错误解析的PDF页面不能正确注释，需要跳过。第四，我们建立了一种标记页面为拒绝的方式，以应对根据标签指南无法进行有效注释的情况。这种案例的例子包括显示错误的PDF页面或包含无法用非重叠矩形捕获的布局。这样的被拒绝页面不包含在最终的数据集中。有了所有这些措施，经验丰富的注释人员通常可以在20秒到60秒的时间内完成单页的注释，这取决于其复杂性。

5 实验

DocLayNet的主要目标是获得能够在各种复杂布局上进行准确文档布局分析的高质量机器学习模型。如第2节所讨论，目前，目标检测模型由于COCO格式[16]的标准化地面真实数据和像detectron2[17]这样的通用框架的可用性，是最容易使用的。此外，PubLayNet和DocBank中的基准数据是使用标准目标检测模型（例如Mask R-CNN和Faster R-CNN）获得的。因此，我们将在本文中涉及这些目标检测方法。

本文，并将第 2 节中提到的较新方法的详细评估留作未来工作。

在本节中，我们将展示与DocLayNet上目标检测模型性能相关的几个方面。与在PubLayNet上的做法类似，我们将使用平均精度均值（mAP）评估其预测质量，使用10个重叠率，范围从0.5到0.95，每隔0.05递增（mAP@0.5-0.95）。这些得分是通过利用COCO API [16]提供的评估代码计算的。

目标检测基线

在表2中，我们展示了Mask R-CNN [12]、Faster R-CNN [11]和 YOLOv5 [13] 的基线实验（以 mAP 给出）。训练和评估均在尺寸为 1025×1025 像素的 RGB 图像上进行。在训练中，对于冗余标注的页面，我们只使用了一个标注。可以观察到，不同模型之间的 mAP 差异相对较小，但总体上比从三重标注页面的成对人工标注计算的 mAP 低 6% 到 10%。这很好地表明 DocLayNet 数据集为研究界提供了一个有价值的挑战，旨在缩小人类识别与机器学习方法之间的差距。有趣的是，Mask R-CNN 和 Faster R-CNN 产生了非常相近的 mAP 分数，这表明从边界框衍生的基于像素的图像分割并没有帮助获得更好的预测。另一方面，较新的 Yolov5x 模型表现非常好，甚至在某些标签（例如文本、表格和图片）上超越了人类。这并不完全令人意外，因为在文档中文本、表格和图片是丰富且视觉上最具特色的。

Table 3: Performance of a Mask R-CNN R50 network in mAP@0.5-0.95 scores trained on DocLayNet with different class label sets. The reduced label sets were obtained by either down-mapping or dropping labels.

Class-count	11	6	5	4
Caption	68	Text	Text	Text
Footnote	71	Text	Text	Text
Formula	60	Text	Text	Text
List-item	81	Text	82	Text
Page-footer	62	62	-	-
Page-header	72	68	-	-
Picture	72	72	72	72
Section-header	68	67	69	68
Table	82	83	82	82
Text	85	84	84	84
Title	77	Sec.-h.	Sec.-h.	Sec.-h.
Overall	72	73	78	77

Table 4: Performance of a Mask R-CNN R50 network with document-wise and page-wise split for different label sets. Naive page-wise split will result in ~10% point improvement.

Class-count	11		5	
Split	Doc	Page	Doc	Page
Caption	68	83		
Footnote	71	84		
Formula	60	66		
List-item	81	88	82	88
Page-footer	62	89		
Page-header	72	90		
Picture	72	82	72	82
Section-header	68	83	69	83
Table	82	89	82	90
Text	85	91	84	90
Title	77	81		
All	72	84	78	87

Learning Curve

One of the fundamental questions related to any dataset is if it is "large enough". To answer this question for DocLayNet, we performed a data ablation study in which we evaluated a Mask R-CNN model trained on increasing fractions of the DocLayNet dataset. As can be seen in Figure 5, the mAP score rises sharply in the beginning and eventually levels out. To estimate the error-bar on the metrics, we ran the training five times on the entire data-set. This resulted in a 1% error-bar, depicted by the shaded area in Figure 5. In the inset of Figure 5, we show the exact same data-points, but with a logarithmic scale on the x-axis. As is expected, the mAP score increases linearly as a function of the data-size in the inset. The curve ultimately flattens out between the 80% and 100% mark, with the 80% mark falling within the error-bars of the 100% mark. This provides a good indication that the model would not improve significantly by yet increasing the data size. Rather, it would probably benefit more from improved data consistency (as discussed in Section 3), data augmentation methods [23], or the addition of more document categories and styles.

Impact of Class Labels

The choice and number of labels can have a significant effect on the overall model performance. Since PubLayNet, DocBank and DocLayNet all have different label sets, it is of particular interest to understand and quantify this influence of the label set on the model performance. We investigate this by either down-mapping labels into more common ones (e.g. *Caption*→*Text*) or excluding them from the annotations entirely. Furthermore, it must be stressed that all mappings and exclusions were performed on the data before model training. In Table 3, we present the mAP scores for a Mask R-CNN R50 network on different label sets. Where a label is down-mapped, we show its corresponding label, otherwise it was excluded. We present three different label sets, with 6, 5 and 4 different labels respectively. The set of 5 labels contains the same labels as PubLayNet. However, due to the different definition of lists in PubLayNet (grouped list-items) versus DocLayNet (separate list-items), the label set of size 4 is the closest to PubLayNet, in the assumption that the *List* is down-mapped to *Text* in PubLayNet. The results in Table 3 show that the prediction accuracy on the remaining class labels does not change significantly when other classes are merged into them. The overall macro-average improves by around 5%, in particular when *Page-footer* and *Page-header* are excluded.

Impact of Document Split in Train and Test Set

Many documents in DocLayNet have a unique styling. In order to avoid overfitting on a particular style, we have split the train-, test- and validation-sets of DocLayNet on document boundaries, i.e. every document contributes pages to only one set. To the best of our knowledge, this was not considered in PubLayNet or DocBank. To quantify how this affects model performance, we trained and evaluated a Mask R-CNN R50 model on a modified dataset version. Here, the train-, test- and validation-sets were obtained by a randomised draw over the individual pages. As can be seen in Table 4, the difference in model performance is surprisingly large: page-wise splitting gains $\tilde{1}0\%$ in mAP over the document-wise splitting. Thus, random page-wise splitting of DocLayNet can easily lead to accidental overestimation of model performance and should be avoided.

Dataset Comparison

Throughout this paper, we claim that DocLayNet's wider variety of document layouts leads to more robust layout detection models. In Table 5, we provide evidence for that. We trained models on each of the available datasets (PubLayNet, DocBank and DocLayNet) and evaluated them on the test sets of the other datasets. Due to the different label sets and annotation styles, a direct comparison is not possible. Hence, we focussed on the common labels among the datasets. Between PubLayNet and DocLayNet, these are *Picture*,

表 3：在使用不同类别标签集的 DocLayNet 上训练的 Mask R-CNN R50 网络在 mAP@0.5-0.95 分数中的性能表现。缩减的标签集是通过向下映射或丢弃标签获得的。

Class-count	11	6	5	4
Caption	68	Text	Text	Text
Footnote	71	Text	Text	Text
Formula	60	Text	Text	Text
List-item	81	Text	82	Text
Page-footer	62	62	-	-
Page-header	72	68	-	-
Picture	72	72	72	72
Section-header	68	67	69	68
Table	82	83	82	82
Text	85	84	84	84
Title	77	Sec.-h.	Sec.-h.	Sec.-h.
Overall	72	73	78	77

表 4：Mask R-CNN R50 网络在不同标签集的文档级和页面级拆分性能。简单的页面级拆分将导致 ~ 10% 点的改进。

Class-count	11		5	
Split	Doc	Page	Doc	Page
Caption	68	83		
Footnote	71	84		
Formula	60	66		
List-item	81	88	82	88
Page-footer	62	89		
Page-header	72	90		
Picture	72	82	72	82
Section-header	68	83	69	83
Table	82	89	82	90
Text	85	91	84	90
Title	77	81		
All	72	84	78	87

学习曲线

与任何数据集相关的基本问题之一是它是否“足够大”。为了回答这个与DocLayNet有关的问题，我们进行了数据消融研究，在其中我们评估了一个在不断增加的DocLayNet数据集比例上训练的Mask R-CNN模型。如图5所示，mAP评分在开始时急剧上升，最终趋于平稳。为了估计指标上的误差范围，我们在整个数据集上进行了五次训练。这导致了1%的误差范围，图5中的阴影区域展示了这一点。在图5的插图中，我们展示了完全相同的数据点，但x轴采用对数刻度。如所预期的那样，在插图中，mAP评分作为数据尺寸的函数呈线性增长。曲线最终在80%和100%标记之间趋于平稳，且80%标记落在100%标记的误差范围内。这很好地表明，通过进一步增加数据大小，模型不会显著改进。相反，它可能更受益于改进的数据一致性（如第3节中讨论的），数据增强方法[23]，或增加更多文档类别和风格。

类标签的影响

标签的选择和数量会对整体模型性能产生显著影响。由于PubLayNet、DocBank和DocLayNet的标签集各不相同，了解和量化标签集对模型性能的影响特别重要。我们通过将标签映射为更常见的标签（例如Caption$v*$Text）或从标注中完全排除来调查这一影响。此外，必须强调的是，所有映射和排除都在模型训练前对数据进行。在表3中，我们展示了Mask R-CNN R50网络在不同标签集上的mAP分数。对于下映射的标签，我们显示其对应的标签，否则，标签将被排除。我们展示三种不同的标签集，分别包含6、5和4个不同的标签。5个标签的集合包含与PubLayNet相同的标签。然而，由于定义不同

在PubLayNet（分组列表项）与DocLayNet（单独列表项）中的列表比较，假设列表在PubLayNet中被映射到文本时，尺寸为4的标签集最接近PubLayNet。表3中的结果表明，当其他类合并到剩余类标签中时，其预测准确率并没有显著变化。总体宏平均值提高了大约5%，尤其是当不包括页脚和页眉时。

文档拆分对训练集和测试集的影响

DocLayNet中的许多文档具有独特的风格。为了避免在某种特定风格上过拟合，我们在文档边界上拆分了DocLayNet的训练集、测试集和验证集，即每个文档仅为一个集合贡献页面。据我们所知，这一点在PubLayNet或DocBank中没有被考虑。为了量化这对模型性能的影响，我们在修改后的数据集版本上训练和评估了一个Mask R-CNN R50模型。在这里，训练集、测试集和验证集是通过对单个页面的随机抽取获得的。如表4所示，模型性能的差异惊人地大：按页面拆分比按文档拆分在mAP上多获得了10%的提升。因此，DocLayNet的随机页面拆分很容易导致模型性能被意外高估，应该避免。

数据集比较

在整篇论文中，我们声称DocLayNet更广泛的文档布局种类导致了更强大的布局检测模型。在表5中，我们为此提供了证据。我们在每个可用的数据集（PubLayNet、DocBank和DocLayNet）上训练了模型，并在其他数据集的测试集上对它们进行了评估。由于标签集和注释风格的不同，无法进行直接比较。因此，我们专注于数据集之间的通用标签。在PubLayNet和DocLayNet之间，这些是图片，

Table 5: Prediction Performance (mAP@0.5-0.95) of a Mask R-CNN R50 network across the PubLayNet, DocBank & DocLayNet data-sets. By evaluating on common label classes of each dataset, we observe that the DocLayNet-trained model has much less pronounced variations in performance across all datasets.

Training on	labels	Testing on PLN	DB	DLN
PubLayNet (PLN)	Figure	96	43	23
	Sec-header	87	-	32
	Table	95	24	49
	Text	96	-	42
	total	93	34	30
DocBank (DB)	Figure	77	71	31
	Table	19	65	22
	total	48	68	27
DocLayNet (DLN)	Figure	67	51	72
	Sec-header	53	-	68
	Table	87	43	82
	Text	77	-	84
	total	59	47	78

Section-header, *Table* and *Text*. Before training, we either mapped or excluded DocLayNet's other labels as specified in table 3, and also PubLayNet's *List* to *Text*. Note that the different clustering of lists (by list-element vs. whole list objects) naturally decreases the mAP score for *Text*.

For comparison of DocBank with DocLayNet, we trained only on *Picture* and *Table* clusters of each dataset. We had to exclude *Text* because successive paragraphs are often grouped together into a single object in DocBank. This paragraph grouping is incompatible with the individual paragraphs of DocLayNet. As can be seen in Table 5, DocLayNet trained models yield better performance compared to the previous datasets. It is noteworthy that the models trained on PubLayNet and DocBank perform very well on their own test set, but have a much lower performance on the foreign datasets. While this also applies to DocLayNet, the difference is far less pronounced. Thus we conclude that DocLayNet trained models are overall more robust and will produce better results for challenging, unseen layouts.

Example Predictions

To conclude this section, we illustrate the quality of layout predictions one can expect from DocLayNet-trained models by providing a selection of examples without any further post-processing applied. Figure 6 shows selected layout predictions on pages from the test-set of DocLayNet. Results look decent in general across document categories, however one can also observe mistakes such as overlapping clusters of different classes, or entirely missing boxes due to low confidence.

6 CONCLUSION

In this paper, we presented the DocLayNet dataset. It provides the document conversion and layout analysis research community a new and challenging dataset to improve and fine-tune novel ML methods on. In contrast to many other datasets, DocLayNet was created by human annotation in order to obtain reliable layout ground-truth on a wide variety of publication- and typesetting-styles. Including a large proportion of documents outside the scientific publishing domain adds significant value in this respect.

From the dataset, we have derived on the one hand reference metrics for human performance on document-layout annotation (through double and triple annotations) and on the other hand evaluated the baseline performance of commonly used object detection methods. We also illustrated the impact of various dataset-related aspects on model performance through data-ablation experiments, both from a size and class-label perspective. Last but not least, we compared the accuracy of models trained on other public datasets and showed that DocLayNet trained models are more robust.

To date, there is still a significant gap between human and ML accuracy on the layout interpretation task, and we hope that this work will inspire the research community to close that gap.

REFERENCES

[1] Max Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. Icdar 2013 table competition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1449–1453, 2013.

[2] Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. Icdar2017 competition on recognition of documents with complex layouts - rdcl2017. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1404–1410, 2017.

[3] Hervé Déjean, Jean-Luc Meunier, Liangcai Gao, Yilun Huang, Yu Fang, Florian Kleber, and Eva-Maria Lang. ICDAR 2019 Competition on Table Detection and Recognition (cTDaR), April 2019. http://sac.founderit.com/.

[4] Antonio Jimeno Yepes, Peter Zhong, and Douglas Burdick. Competition on scientific literature parsing. In *Proceedings of the International Conference on Document Analysis and Recognition*, ICDAR, pages 605–617. LNCS 12824, Springer-Verlag, sep 2021.

[5] Logan Markewich, Hao Zhang, Yubin Xing, Navid Lambert-Shirzad, Jiang Zhexin, Roy Lee, Zhi Li, and Seok-Bum Ko. Segmentation for document layout analysis: not dead yet. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–11, 01 2022.

[6] Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. Publaynet: Largest dataset ever for document layout analysis. In *Proceedings of the International Conference on Document Analysis and Recognition*, ICDAR, pages 1015–1022, sep 2019.

[7] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, COLING, pages 949–960. International Committee on Computational Linguistics, dec 2020.

[8] Riaz Ahmad, Muhammad Tanvir Afzal, and M. Qadir. Information extraction from pdf sources based on rule-based system using integrated formats. In *SemWebEval@ESWC*, 2016.

[9] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, pages 580–587. IEEE Computer Society, jun 2014.

[10] Ross B. Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision*, ICCV, pages 1440–1448. IEEE Computer Society, dec 2015.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, ICCV, pages 2980–2988. IEEE Computer Society, Oct 2017.

[13] Glenn Jocher, Alex Stoken, Ayush Chaurasia, Jirka Borovec, NanoCode012, TaoXie, Yonghye Kwon, Kalen Michael, Liu Changyu, Jiacong Fang, Abhiram V, Laughing, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Jebastin Nadar, imyhxy, Lorenzo Mammana, Alex Wang, Cristi Fati, Diego Montes, Jan Hajek, Laurentiu

表5：Mask R-CNN R50网络在PubLayNet、DocBank和DocLayNet数据集上的预测性能（mAP@0.5-0.95）。通过评估每个数据集的常见标签类别，我们观察到DocLayNet训练的模型在所有数据集中的性能变化较小。

Training on	labels	Testing on PLN	DB	DLN
PubLayNet (PLN)	Figure	96	43	23
	Sec-header	87	-	32
	Table	95	24	49
	Text	96	-	42
	total	93	34	30
DocBank (DB)	Figure	77	71	31
	Table	19	65	22
	total	48	68	27
DocLayNet (DLN)	Figure	67	51	72
	Sec-header	53	-	68
	Table	87	43	82
	Text	77	-	84
	total	59	47	78

部分标题、表格和文本　　在训练之前，我们根据表3中指定的内容，将DocLayNet的其他标签进行映射或排除，并且将PubLayNet的列表转为文本。请注意，列表的不同聚类（按列表元素与整个列表对象）自然会降低文本的mAP得分。

为了比较 DocBank 和 DocLayNet，我们仅在每个数据集的图片和表格聚合上进行了训练。我们不得不排除文本，因为在 DocBank 中，连续的段落通常被组合为一个对象。这种段落分组与 DocLayNet 中的单独段落不兼容。如表 5 所示，DocLayNet 训练的模型与以前的数据集相比表现更好。值得注意的是，尽管在其自身的测试集上，PubLayNet 和 DocBank 训练的模型表现得很好，但在外来数据集上的表现要低得多。虽然这也适用于 DocLayNet，但差异远不那么明显。因此，我们得出结论，DocLayNet 训练的模型总体上更加稳健，并且将在应对具有挑战性的未知布局时产生更好的结果。

示例预测

总结本节内容，我们通过提供一些未经过进一步后处理的示例，来展示从DocLayNet训练的模型所期望的布局预测质量。图6显示了DocLayNet测试集页面上的所选布局预测。总体来看，结果在各类文档中表现良好，但也可以观察到一些错误，例如不同类别的簇重叠，或由于置信度低而导致的完全遗漏的框。

6 结论

在本文中，我们展示了 DocLayNet 数据集。它为文档转换和布局分析研究社区提供了一个新的具有挑战性的数据集，以改进和微调新颖的机器学习方法。与许多其他数据集相比，DocLayNet 是通过人工标注创建的，以便在各种出版和排版风格上获得可靠的布局真实值。包括在科学出版领域以外的大量文档在这方面增加了重要的价值。

从数据集中，我们一方面推导出了文档布局标注的人工性能参考指标（通过双重和三重标注），另一方面评估了常用目标检测方法的基线性能。我们还通过数据消融实验展示了各种与数据集相关的方面对模型性能的影响，包括从规模和类别标签的角度。最后但同样重要的是，我们比较了在其他公共数据集上训练的模型的准确性，并表明DocLayNet训练的模型更具鲁棒性。

迄今为止，人类和机器学习在布局解释任务上的准确性仍然存在显著差距，我们希望这项工作能够激励研究界缩小这一差距。

参考文献

[1] Max Göbel, Tamir Hassan, Ermelinda Oro 和 Giorgio Orsi。ICDAR 2013 表格竞赛。在 2013 年第 12 届国际文档分析与识别会议中，第 1449-1453 页，2013 年。[2] Christian Clausner, Apostolos Antonacopoulos 和 Stefan Pletschacher。ICDAR2017 复杂布局文档识别竞赛 - RDCL2017。在 2017 年第 14 届 IAPR 国际文档分析与识别会议 (ICDAR) 中，第一卷，第 1404-1410 页，2017 年。[3] Hervé Déjean, Jean-Luc Meunier, Liangcai Gao, Yilun Huang, Yu Fang, Florian Kleber 和 Eva-Maria Lang。ICDAR 2019 表格检测和识别竞赛 (cTDaR)，2019 年 4 月。http://sac.founderit.com/。[4] Antonio Jimeno Yepes, Peter Zhong 和 Douglas Burdick。科学文献解析竞赛。在国际文档分析与识别会议记录中，ICDAR，第 605-617 页。LNCS 12824, Springer-Verlag, 2021 年 9 月。[5] Logan Markewich, Hao Zhang, Yubin Xing, Navid Lambert-Shirzad, Jiang Zhexin, Roy Lee, Zhi Li 和 Seok-Bum Ko。文档布局分析的分割：尚未过时。国际文档分析与识别期刊 (IJDAR)，第 1-11 页，2022 年 1 月。[6] Xu Zhong, Jianbin Tang 和 Antonio Jimeno-Yepes。Publaynet：迄今为止最大的文档布局分析数据集。在国际文档分析与识别会议记录中，ICDAR，第 1015-1022 页，2019 年 9 月。[7] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li 和 Ming Zhou。Docbank：文档布局分析的基准数据集。在第 28 届计算语言学国际会议（COLING）会议记录中，第 949-960 页。国际计算语言学委员会，2020 年 12 月。[8] Riaz Ahmad, Muhammad Tanvir Afzal 和 M. Qadir。基于集成格式的规则系统从 PDF 源中提取信息。在 SemWebEval@ESWC, 2016。[9] Ross B. Girshick, Jeff Donahue, Trevor Darrell 和 Jitendra Malik。用于准确的对象检测和语义分割的丰富特征层次。在 IEEE 计算机视觉与模式识别会议 (CVPR) 中，第 580-587 页。IEEE 计算机协会, 2014 年 6 月。

[10] Ross B. Girshick. Fast R-CNN. 在2015 IEEE国际计算机视觉会议，ICCV 2015，智利圣地亚哥，第1440-1448页。IEEE计算机学会，2015年12月。

[11] Shaoqing Ren, Kaiming He, Ross Girshick, 和 Jian Sun。Faster r-cnn：利用区域提议网络实现实时目标检测。《IEEE 模式分析与机器智能汇刊》，39(6):1137 – 1149, 2017。

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross B. Girshick。Mask R-CNN。在IEEE国际计算机视觉会议，ICCV，页面2980-2988。IEEE计算机学会，2017年10月。

[13] Markus Jocher, Alex Stoken, Ayush Chaurasia, Jirka Borovec, Yonghye Kwon, Kalen Michael, Liu Changyu, Jiacong Fang, Abhiram V, Laughing, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Jebastin Nadar, imyhxy, Lorenzo Mammana, Alex Wang, Cristi Fati, Diego Montes, Jan Hajek, Laurentiu

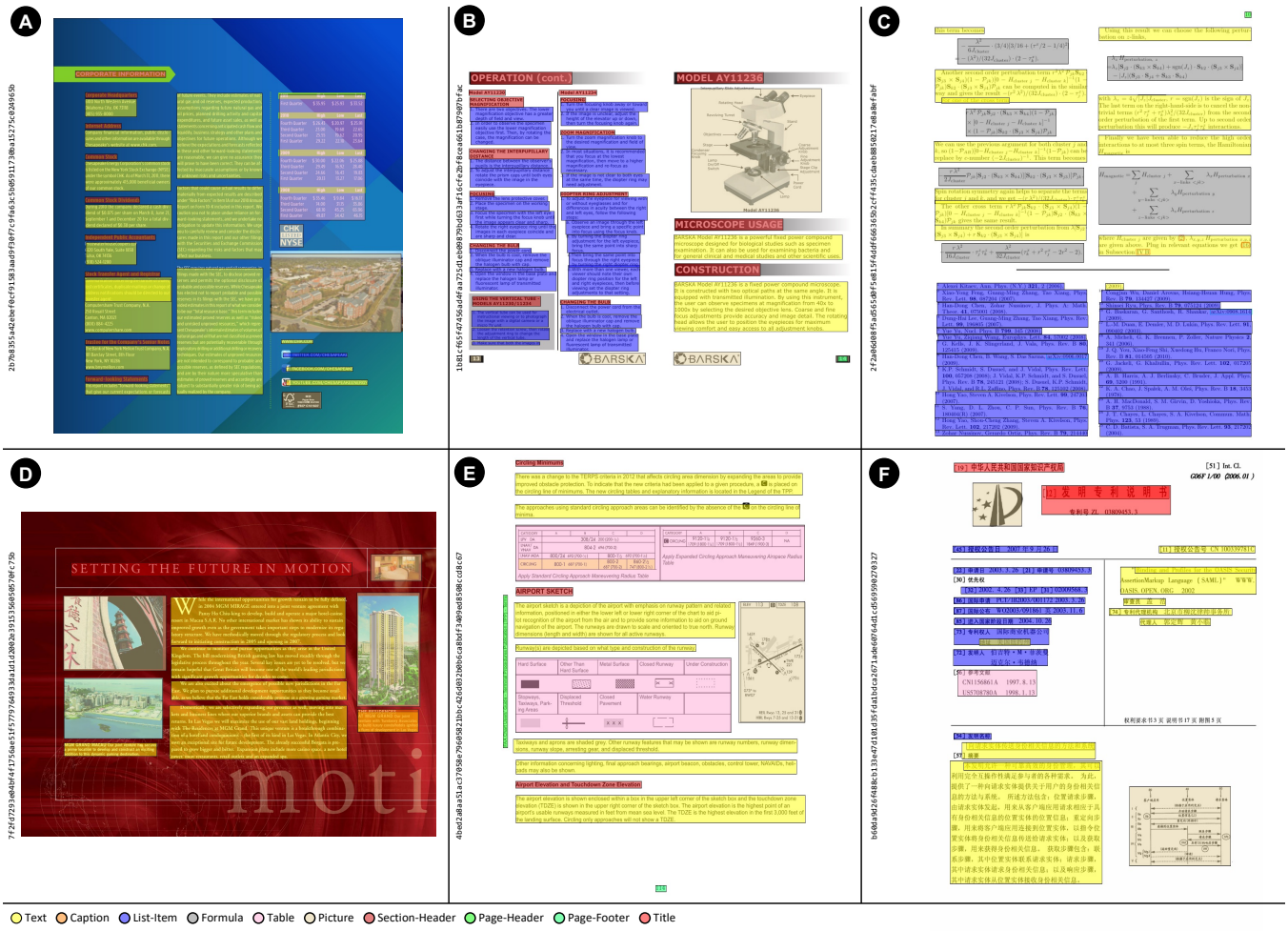

Figure 6: Example layout predictions on selected pages from the DocLayNet test-set. (A, D) exhibit favourable results on coloured backgrounds. (B, C) show accurate list-item and paragraph differentiation despite densely-spaced lines. (E) demonstrates good table and figure distinction. (F) shows predictions on a Chinese patent with multiple overlaps, label confusion and missing boxes.

Diaconu, Mai Thanh Minh, Marc, albinxavi, fatih, oleg, and wanghao yang. ultralytics/yolov5: v6.0 - yolov5n nano models, roboflow integration, tensorflow export, opencv dnn support, October 2021.

[14] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020.

[15] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. *CoRR*, abs/1911.09070, 2019.

[16] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context, 2014.

[17] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2, 2019.

[18] Nikolaos Livathinos, Cesar Berrospi, Maksym Lysak, Viktor Kuropiatnyk, Ahmed Nassar, Andre Carvalho, Michele Dolfi, Christoph Auer, Kasper Dinkla, and Peter W. J. Staar. Robust pdf document conversion using recurrent neural networks. In *Proceedings of the 35th Conference on Artificial Intelligence*, AAAI, pages 15137–15145, feb 2021.

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

[20] Shoubin Li, Xuyan Ma, Shuaiqun Pan, Jun Hu, Lin Shi, and Qing Wang. Vtlayout: Fusion of visual and text features for document layout analysis, 2021.

[21] Peng Zhang, Can Li, Liang Qiao, Zhanzhan Cheng, Shiliang Pu, Yi Niu, and Fei Wu. Vsr: A unified framework for document layout analysis combining vision, semantics and relations, 2021.

[22] Peter W J Staar, Michele Dolfi, Christoph Auer, and Costas Bekas. Corpus conversion service: A machine learning platform to ingest documents at scale. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, pages 774–782. ACM, 2018.

[23] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

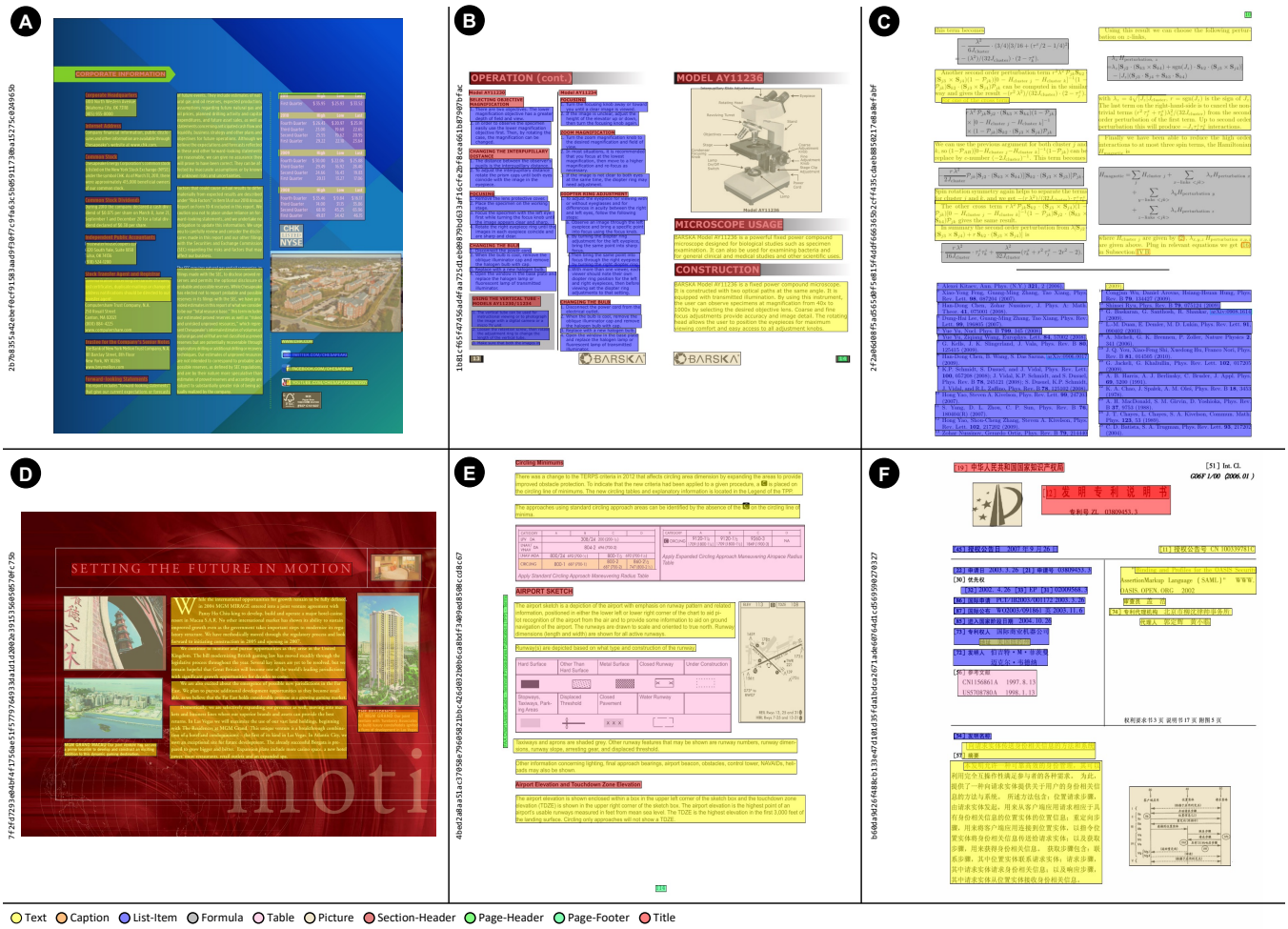

图 6：DocLayNet 测试集选定页面的示例布局预测。(A, D) 在彩色背景上展示了良好的结果。(B, C) 展示了尽管行距密集但列表项和段落的准确区分。(E) 显示了良好的表格和图形区分。(F) 显示了在中文专利上的预测，存在多个重叠、标签混淆和缺失框。

Diaconu，Mai Thanh Minh，Marc，albinxavi，fatih，oleg，以及wanghao yang。ul- tralytics/yolov5: v6.0 - yolov5n nano模型，roboflow集成，tensorflow导出，opencv dnn支持，2021年10月。[14] Nicolas Carion，Francisco Massa，Gabriel Synnaeve，Nicolas Usunier，Alexander Kirillov，以及Sergey Zagoruyko。基于转换器的端到端目标检测。CoRR，abs/2005.12872，2020。[15] Mingxing Tan，Ruoming Pang，和Quoc V. Le。Efficientdet：可扩展且高效的目标检测。CoRR，abs/1911.09070，2019。[16] Tsung-Yi Lin，Michael Maire，Serge J. Belongie，Lubomir D. Bourdev，Ross B. Gir- shick，James Hays，Pietro Perona，Deva Ramanan，Piotr Dollár，以及C. Lawrence Zitnick。Microsoft COCO：情境中的常见物体，2014。[17] Yuxin Wu，Alexander Kirillov，Francisco Massa，Wan-Yen Lo，以及Ross Girshick。Detectron2，2019。[18] Nikolaos Livathinos，Cesar Berrospi，Maksym Lysak，Viktor Kuropiatnyk，Ahmed Nassar，Andre Carvalho，Michele Dolfi，Christoph Auer，Kasper Dinkla，以及Peter W. J. Staar。使用递归神经网络进行鲁棒的PDF文档转换。于第35届人工智能会议论文集，AAAI，第15137 – 15145页，2021年2月。[19] Yiheng Xu，Minghao Li，Lei Cui，Shaohan Huang，Furu Wei，以及Ming Zhou。Layoutlm：文档图像理解的文本与布局的预训练。于第26届ACM SIGKDD国际知识发现与数据挖掘会议论文集，KDD，第1192 – 1200页，美国纽约，2020。计算机机器协会。

[20] 李寿斌, 马绪彦, 潘帅群, 胡军, 史林, 王青. Vtlayout: 视觉和文本特征融合用于文档布局分析, 2021. [21] 张鹏, 李灿, 乔良, 程展展, 蒲土良, 牛一, 吴飞. Vsr: 结合视觉, 语义和关系的统一文档布局分析框架, 2021. [22] Peter W J Staar, Michele Dolfi, Christoph Auer, 和 Costas Bekas. Corpus conversion service: 大规模文档摄取的机器学习平台. 在第24届ACM SIGKDD知识发现与数据挖掘国际会议,KDD, 第774-782页. ACM, 2018. [23] Connor Shorten 和 Taghi M. Khoshgoftaar. 深度学习图像数据增强的调查. 大数据杂志, 6(1):60, 2019.