M^{6}Doc: A Large-Scale Multi-Format, Multi-Type, Multi-Layout, Multi-Language, Multi-Annotation Category Dataset for Modern Document Layout Analysis

Hiuyi Cheng[1], Peirong Zhang[1], Sihang Wu[2], Jiaxin Zhang[1],
Qiyuan Zhu[2], Zecheng Xie[2], Jing Li[2], Kai Ding[3], and Lianwen Jin[1]*
[1]South China University of Technology
[2]Huawei Cloud Computing Technologies Co., Ltd.
[3]IntSig Information Co., Ltd.
{eechenghiuyi, eeprzhang, msjxzhang}@mail.scut.edu.cn, eelwjin@scut.edu.cn, danny_ding@intsig.net,
{wusihang2, zhuqiyuan2, xiezecheng1, lijing260}@huawei.com

Abstract

Document layout analysis is a crucial prerequisite for document understanding, including document retrieval and conversion. Most public datasets currently contain only PDF documents and lack realistic documents. Models trained on these datasets may not generalize well to real-world scenarios. Therefore, this paper introduces a large and diverse document layout analysis dataset called M^{6}Doc. The M^6 designation represents six properties: (1) Multi-Format (including scanned, photographed, and PDF documents); (2) Multi-Type (such as scientific articles, textbooks, books, test papers, magazines, newspapers, and notes); (3) Multi-Layout (rectangular, Manhattan, non-Manhattan, and multi-column Manhattan); (4) Multi-Language (Chinese and English); (5) Multi-Annotation Category (74 types of annotation labels with 237,116 annotation instances in 9,080 manually annotated pages); and (6) Modern documents. Additionally, we propose a transformer-based document layout analysis method called TransDLANet, which leverages an adaptive element matching mechanism that enables query embedding to better match ground truth to improve recall, and constructs a segmentation branch for more precise document image instance segmentation. We conduct a comprehensive evaluation of M^{6}Doc with various layout analysis methods and demonstrate its effectiveness. TransDLANet achieves state-of-the-art performance on M^{6}Doc with 64.5% mAP. The M^{6}Doc dataset will be available at https://github.com/HCIILAB/M6Doc.

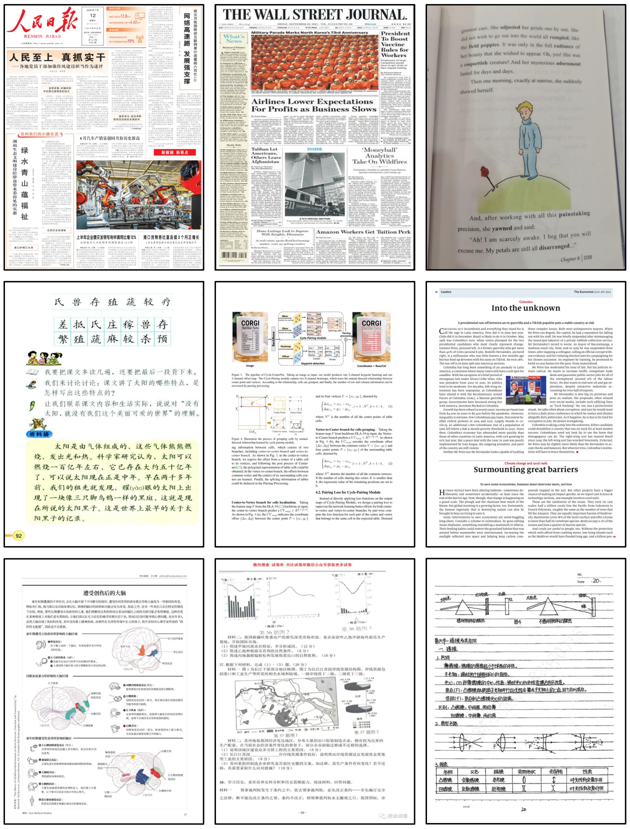

人民日报
THE WALL STREET JOURNAL.
Airlines Lower Expectations For Profits as Business Slows
Into the unknown
Surmounting great barriers

Figure 1. Examples of complex page layouts across different document formats, types, layouts, languages.

1. Introduction

Document layout analysis (DLA) is a fundamental preprocessing task for modern document understanding and

*Corresponding Author.

digitization, which has recently received increasing attention [25]. DLA can be classified into physical layout analysis and logical layout analysis [15]. Physical layout analysis considers the visual presentation of the document and distinguishes regions with different elements such as text, image, and table. Logical layout analysis distinguishes the semantic structures of documents according to the meaning and assigns them to different categories, such as chapter heading, section heading, paragraph, and figure note.

Currently, deep learning methods have dominated DLA, which require a plethora of training data. Some datasets have been proposed in the community to promote the development of DLA, as shown in Table 1. However, these datasets have several limitations. (1) Small size. Early DLA datasets, such as PRImA [1] and DSSE200 [41], were small-scale and contained only hundreds of images. (2) Limited document format. The formats of current public large-scale datasets such as PubLayNet [44], DocBank [17], and DocLayNet [29], are all PDF documents. It presents a huge challenge to evaluate the effectiveness of different methods in realistic scenarios. (3) Limited document diversity. Most datasets include only scientific articles, which are typeset using uniform templates and severely lack variability. Although DocLayNet [41] considers documents of seven types, they are not commonly used. The lack of style diversity would prejudice the development of multi-domain general layout analysis. (4) Limited document languages. Most datasets' language is English. Since the text features of documents in different languages are fundamentally different, DLA methods may encounter domain shift problems in different languages, which remain unexplored. (5) Few annotation categories. The annotation categories of current datasets are not sufficiently fine-grained, preventing more granular layout information extraction.

To promote the development of fine-grained logical DLA in realistic scenarios, we have built the Multi-Format, Multi-Type, Multi-Layout, Multi-Language, and Multi-Annotation Categories Modern document (M^6Doc) dataset. M^6Doc possesses several advantages. Firstly, M^6Doc considers three document formats (scanned, photographed, and PDF) and seven representative document types (scientific articles, magazines, newspapers, etc.). Since scanned/photographed documents are commonly seen and widely used, the proposed M^6Doc dataset presents great diversity and closely mirrors real-world scenarios. Secondly, M^6Doc contains 74 document annotation categories, which are the most abundant and fine-grained up to date. Thirdly, M^6Doc is the most detailed manually annotated DLA dataset, as it contains 237,116 annotation instances in 9,080 pages. Finally, M^6Doc includes four layouts (rectangular, Manhattan, non-Manhattan, and multi-column Manhattan) and two languages (Chinese and English), covering more comprehensive layout scenarios. Several examples of the M^6Doc dataset are shown in Figure 1.

In addition, we propose a transformer-based model, TransDLANet, to perform layout extraction in an instance segmentation manner effectively. It adopts a standard Transformer encoder without positional encoding as a feature fusion method and uses an adaptive element matching mechanism to enable the query vector to better focus on the unique features of layout elements. This helps understand the spatial and global interdependencies of distinct layout elements and also reduces duplicate attention on the same instance. Subsequently, a dynamic decoder is exploited to perform the fusion of RoI features and image features. Finally, it uses three parameter-shared multi-layer perception (MLP) branches to decode the fused interaction features for multi-task learning.

The contributions of this paper are summarized as follows:

- M^6Doc is the first layout analysis dataset that contains both real-world (photographed and scanned) files and born-digital files. Additionally, it is the first dataset that includes Chinese examples. It has several representative document types and layouts, facilitating the development of generic layout analysis methods.
- M^6Doc is the most fine-grained logical layout analysis categories. It can serve as a benchmark for several related tasks, such as logical layout analysis, formula recognition, and table analysis.
- We propose the TransDLANet, a Transformer-based method for document layout analysis. It includes a Transformer-like encoder to better capture the correlation between queries, a dynamic interaction decoder, and three multi-ayer perceptron branches with shared parameters to decode the fused interaction features for multi-task learning.

2. Related Works

2.1. Modern Layout Analysis Dataset

A variety of modern layout analysis datasets have been created in recent years. In 2009, Antonacopoulos et al. [1] presented the PRImA dataset, which was the first commonly used real-world dataset with 305 images of magazines and scientific articles. In 2019, Zhong et al. [44] published the PubLayNet dataset, which contains over 360,000 page samples annotated with typical document layout elements such as text, heading, list, graphic, and table. Annotations were automatically generated by matching PDFs and XML formats of articles from the PubMed Central Open Access subset. In 2020, researchers at Microsoft Research Asia built the DocBank dataset [17], which contains 500,000 document pages and fine-grained token-level annotations for document layout analysis. It was developed

Table 1. Modern Document Layout Analysis Datasets. **A.M.** denotes the annotating means.

Dataset	**#Image**	**#Class**	**#Instance**	**A.M.**	**Format**	**Document Type**	**Language**
DSSE200 [41]	200	6	-	Automatic	PDF	Magazines, Academic papers.	English
DAD [23]	5,980	5	90,923	Automatic	PDF	Articles	English
PubMed [16]	12,871	5	257,830	Automatic	PDF	Articles	English
Chn [16]	8,005	5	203,456	Automatic	PDF	Chinese Wikipedia pages	Chinese
PubLayNet [44]	360K	5	3,311,660	Automatic	PDF	Articles	English
DocBank [17]	500K	13	-	Automatic	PDF	Articles	English
DocLayNet [29]	80,863	11	1,107,470	Manual	PDF	Financial Reports, Manuals, Scientific Articles, Laws & Regulations, Patents, Government Tenders.	English, German, French, Japanese
PRImA [1]	305	10	-	Automatic	Scanned	Magazine, Technical article, Forms, Bank statements, Advertisements	English
BCE-Arabic-v1 [33]	1,833	3	-	Automatic	Scanned	Arabic books	Arabic
BCE-Arabic-v2 [7]	9,000	21	-	Automatic	Scanned	Arabic books	Arabic
M^{6}Doc (**Ours**)	9,080	74	237,116	Manual	PDF, Scanned, Photographed	Scientific articles, Textbooks, Books, Test papers, Magazines, Newspapers, Notes	English, Chinese

based on a large number of PDF files of papers compiled by the LaTeX tool. Unlike the conventional manual annotating process, they approach obtaining high-quality annotations using a weakly supervised approach in a simple and efficient manner. In 2022, IBM researchers presented the DocLayNet dataset [29], which contains 80,863 manually annotated pages. It contains six document types (technical manuals, annual company reports, legal text, and government tenders), 11 categories of annotations, and four languages (English documents close to 95%). A few pages in the DocLayNet dataset have multiple manual annotations, which allows for experiments in annotation uncertainty and quality control analysis.

However, the predominant document format for large datasets is PDF, not scanned and photographed images as in real-world scenarios. Only a few public datasets include real-world data. The variety of layouts in current public datasets is still very limited and is not conducive to the development of logical layout analysis. Currently, 95% of the publicly available datasets are English documents, which are largely unsuitable for Asian language documents. To this end, we propose the M^6Doc dataset to facilitate the development of layout analysis.

2.2. Deep Learning for Layout Analysis

Earlier layout analysis methods [13, 24, 26, 28, 39] used rule-based and heuristic algorithms, so they were limited to applications on certain simple types of documents, and the generalization performance of such methods was poor. However, with the development of deep learning, DLA methods based on deep learning have been developed to tackle challenging tasks. Mainstream approaches include object detection-based models [2, 16, 30], segmentation-based models [4, 15, 40], and multi-modal methods [27, 41, 43]. For example, Li et al. [16] considered DLA as an object detection task and added a domain adaptation module to study cross-domain document object detection tasks. Lee et al. [15] used segmentation methods to solve DLA problems and introduced trainable multiplication layer techniques for improving the accuracy of object boundary detection to improve the performance of pixel-level segmentation networks. Zhang et al. [43] proposed a unified framework for multi-modal layout analysis by introducing semantic information in a new semantic branch of Mask R-CNN [9] and a module for modeling element relationships. Behind their success, large datasets are required for training and evaluating the models.

However, the lack of a multi-format, multi-type, multi-language, and multi-label categorized logical layout analysis dataset makes it difficult for current methods to obtain good results in real-world and other language scenarios. Moreover, a data format that links visual and textual features has not yet been established for multi-modal tasks.

3. M^6Doc Dataset

The M^6Doc dataset contains a total of 9,080 modern document images, which are categorized into seven subsets, *i.e.*, scientific article (11%), textbook (23%), test paper (22%), magazine (22%), newspaper (11%), note (5.5%), and book (5.5%) according to their content and layouts. It contains three formats: PDF (64%), photographed documents (5%), and scanned documents (31%). The dataset includes a total of 237,116 annotated instances.

The M^6Doc datasets were collected from various sources, including arXiv[1], the official website of the Chinese People's Daily[2], and VKontakte[3]. The source and composition of different subsets are shown below.

- The scientific article subset includes articles obtained by searching with the keywords "Optical Character Recognition" and "Document Layout Analysis" on arXiv. PDF files were then downloaded and converted to images.

[1] https://arxiv.org/

[2] http://paper.people.com.cn/

[3] https://vk.com/

Table 2. $M^6 Doc$ dataset overview.

Category	Training		Validate		Test		Category	Training		Validate		Test	
	Number	%	Number	%	Number	%		Number	%	Number	%	Number	%
background	0	0.000	0	0.000	0	0.000	institute	60	0.042	9	0.039	28	0.040
QR code	59	0.041	15	0.065	23	0.032	jump line	381	0.266	63	0.271	180	0.254
advertisement	257	0.180	45	0.194	145	0.205	kicker	516	0.361	91	0.392	257	0.363
algorithm	12	0.008	3	0.013	12	0.017	lead	664	0.464	109	0.470	285	0.402
answer	165	0.115	30	0.129	77	0.109	marginal note	238	0.166	37	0.159	101	0.143
author	2,424	1.695	403	1.736	1,188	1.676	matching	7	0.005	1	0.004	8	0.011
barcode	10	0.007	1	0.004	3	0.004	mugshot	73	0.051	11	0.047	46	0.065
bill	3	0.002	2	0.009	3	0.004	option	3,198	2.236	515	2.219	1,577	2.225
blank	189	0.132	58	0.250	90	0.127	ordered list	1,012	0.707	172	0.741	510	0.720
bracket	863	0.603	164	0.707	273	0.385	other question number	42	0.029	3	0.013	31	0.044
breakout	411	0.287	72	0.310	188	0.265	page number	4,782	3.343	803	3.460	2,383	3.363
byline	1,276	0.892	185	0.797	660	0.931	paragraph	65,642	45.891	10,575	45.562	33,069	46.664
caption	3,508	2.452	605	2.607	1,766	2.492	part	524	0.366	89	0.383	283	0.399
catalogue	39	0.027	10	0.043	19	0.027	play	10	0.007	3	0.013	2	0.003
chapter title	245	0.171	33	0.142	124	0.175	poem	98	0.069	18	0.078	33	0.047
code	62	0.043	7	0.030	31	0.044	reference	149	0.104	23	0.099	62	0.087
correction	9	0.006	1	0.004	6	0.008	sealing line	3	0.002	2	0.009	5	0.007
credit	1,523	1.065	255	1.099	728	1.027	second-level question number	2,773	1.939	377	1.624	1,330	1.877
dateline	901	0.630	140	0.603	482	0.680	second-level title	273	0.191	48	0.207	140	0.198
drop cap	414	0.289	71	0.306	234	0.330	section	2,508	1.753	408	1.758	1,228	1.733
editor's note	39	0.027	4	0.017	9	0.013	section title	897	0.627	171	0.737	442	0.624
endnote	35	0.024	4	0.017	19	0.027	sidebar	54	0.038	10	0.043	27	0.038
examinee information	8	0.006	2	0.009	6	0.008	sub section title	567	0.396	107	0.461	269	0.380
fifth-level title	13	0.009	2	0.009	20	0.028	subhead	1,998	1.397	394	1.698	1,069	1.508
figure	7,614	5.323	1,242	5.351	3,762	5.309	subsub section title	101	0.071	21	0.090	71	0.100
first-level question number	5,669	3.963	930	4.007	2,740	3.866	supplementary note	986	0.689	158	0.681	487	0.687
first-level title	586	0.410	81	0.349	292	0.412	table	821	0.574	146	0.629	409	0.577
flag	30	0.021	5	0.022	12	0.017	table caption	287	0.201	41	0.177	143	0.202
folio	1,442	1.008	213	0.918	685	0.967	table note	8	0.006	2	0.009	5	0.007
footer	1,984	1.387	310	1.336	987	1.393	teasers	32	0.022	7	0.030	7	0.010
footnote	295	0.206	49	0.211	139	0.196	third-level question number	240	0.168	36	0.155	102	0.144
formula	1,3090	9.151	2,058	8.867	6,191	8.736	third-level title	146	0.102	44	0.190	94	0.133
fourth-level section title	15	0.010	3	0.013	19	0.027	title	201	0.141	35	0.151	100	0.141
fourth-level title	70	0.049	13	0.056	66	0.093	translator	73	0.051	11	0.047	38	0.054
header	1,877	1.312	297	1.280	969	1.367	underscore	3,687	2.578	590	2.542	1,717	2.423
headline	4,115	2.877	643	2.770	1,981	2.795	unordered list	497	0.347	84	0.362	271	0.382
index	214	0.150	36	0.155	100	0.141	weather forecast	10	0.007	3	0.013	3	0.004
inside	16	0.011	1	0.004	5	0.007	**Total**	**143,040**	**100**	**23,210**	**100**	**70,866**	**100**

- The textbook subset contains 2,080 scanned document images from textbooks for three grades (elementary, middle, and high school) and nine subjects (Chinese, Math, English, Physics, Chemistry, Biology, History, Geography, and Politics).
- The test paper subset consists of 2,000 examination papers covering the same nine subjects as the textbook subset.
- The magazine subset includes 1,000 Chinese and English magazines in PDF format, respectively. The Chinese magazines were sourced from five publishers: Global Science, The Mystery, Youth Digest, China National Geographic, and The Reader. The English magazines were sourced from five American publishers: The New Yorker, New Scientist, Scientific American, The Economist, and Time USA.
- The newspaper subset contains 500 PDF document images from the Chinese People's Daily and the Wall Street Journal.
- The note subset consists of students' handwritten notes in nine subjects, including 500 scanned pages.
- The book subset contains 500 photographed images, which were acquired from 50 books with 10 pages each. Each book has a distinct layout, resulting in considerable diversity in this subset.

For a fair evaluation, we divided the dataset into training, validation, and test sets in a ratio of 6:1:3. We also ensured that the different labels were in equal proportions in the three sets. Table 2 summarizes the overall frequency and distribution of labels in different sets.

4. Data Annotation

Label definition. To ensure that the definition of document layout elements is reasonable and traceable, we reviewed relevant information, such as layout knowledge and layout design. We also used knowledge from the book "Page Design: New Layout & Editorial Design(2019)" [34] and referred to YouTube video explanations regarding magazine[1] and newspaper[2] layouts. In most cases, we followed the Wikipedia definition. Consequently, we defined 74 detailed document annotation labels. The key factors in selecting these annotation labels include (1) the commonality of annotation labels between different document types, (2) the specificity of labels between different document types, (3) the frequency of labels, and (4) the recognition of independent pages. We first unified the labels between different documents to the maximum extent and then defined the labels for certain document types for differentials. Commonality and specificity ensure that the defined labels can adapt

[1] https://www.youtube.com/watch?v=7sSJtScnsjE
[2] https://www.youtube.com/watch?v=LcsOuGcaqZs

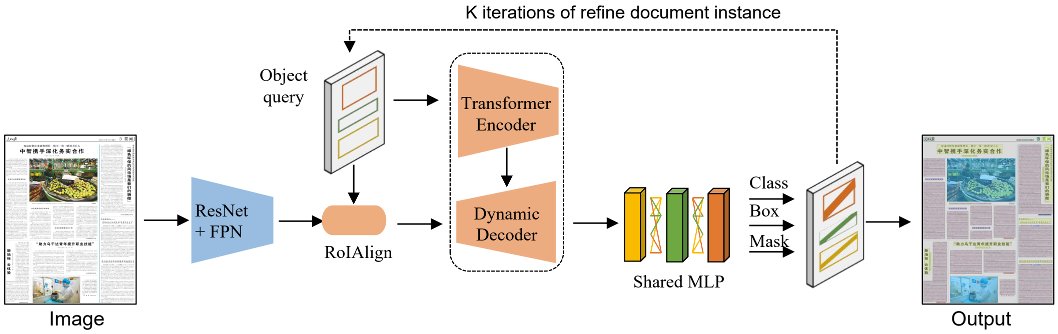


Figure 2. The pipeline of TransDLANet contains four main components: 1) a CNN-based backbone; 2) a transformer encoder; 3) a dynamic decoder that decodes the instance-level features; and 4) three shared multi-layer perceptron(MLP) branches that obtain the classification confidence, bounding boxes, and segmentation mask of the document instance region.

to multiple document types, which implies that a more detailed logical layout analysis for a certain type of document can be performed. It differs from how labels are defined in DocBank, PubLayNet, and DocLayNet, which all ignore defining specific labels for different document types.

Annotation guideline. We provide a detailed annotation guideline (over 170 pages) and some typical annotation examples. 47 annotators performed the annotation task strictly according to the guidelines.

Several key points of the guideline that are different from DocBank, PubLayNet, and DocLayNet are summarized as follows:

- We distinguish table caption and figure caption into two categories.
- We distinguish the ordered list and unordered list into two categories.
- All list-items are grouped together into one list object. This definition differs from DocLayNet, which considers single-line elements as list-item if the list-item are paragraphs with hanging indentation.
- Bold emphasized text at the beginning of a paragraph is not considered a heading unless it appears on a separate line or with heading formatting, such as 1.1.1.
- The headings at different levels are defined in detail.
- The formulas inside the paragraphs are marked.

The annotation results showed that different annotators interpreted ambiguous scenarios differently, such as (1) in the absence of obvious borders, it is sometimes difficult to determine whether a region is a table or a paragraph; (2) whether images with sub-images should be annotated separately or holistically; and (3) in the absence of obvious markers or separators, it is sometimes difficult to determine whether a paragraph is a list item or a body. It was difficult to unify the consistency of the results of the 47 annotators. Therefore, we provided consistent annotation requirements for ambiguous scenarios. To further ensure the consistency of the annotation results, all data were finally checked by the author. The annotation files followed the MS COCO annotation format [20] for object detection. Detailed annotation guidelines and the M^6Doc dataset will be available for reference. A more detailed labeling process is provided in the Supplementary Material.

5. TransDLANet

Our method closely follows the framework of ISTR [10], but differs at its core by leveraging an adaptive element matching mechanism that enables query embedding to better match ground truth and improve recall. We use the transformer encoder as a characteristic fusion method without position encoding and construct a segmentation branch for more precise document image instance segmentation. Additionally, we use three multi-layer perception(MLP) branches with shared parameter for multi-task learning.

TransDLANet architecture. The overall architecture is depicted in Figure 2. We use a CNN-based backbone to extract document image features, and RoIAlign to extract the image features for the pre-defined query vectors. The Transformer encoder performs self-attentive feature learning on query embedding vectors and uses an adaptive element matching mechanism to enhance further the association between document instances encoded by the query vectors. The dynamic interaction-based decoding module (Dynamic Decoder) fuses the query vector with the features of the bounding box image region obtained by the query vector through the RoIAlign. Three shared parameter MLP branches are used for decoding the classification confidence, the bounding boxes' coordinate position, and the segmentation mask of the document instance region. Finally, we repeat this process for K iterations to refine the final document instance.

6. Experiment

6.1. Datasets

Our experiments are conducted on a number of commonly known document layout analysis benchmarks, including DocBank [17], PubLayNet [44], and DocLayNet [29].

6.2. Implementation Details

We adopted ResNet-101 pretrained on ImageNet [6] as our model's backbone. We used the AdamW optimizer [21] to train the model, setting the base learning rate to 2×10^{-5}. The default training epoch was set to 500, and the learning rates descended to 2×10^{-6} and 2×10^{-7} at 50% and 75% of the training epochs, respectively. During training, we used random crop augmentations and scaled the input images such that the shortest side was at least 704-896 pixels and the longest side was at most 1333 pixels to ensure optimal performance.

6.3. Significance of M^6Doc

Due to the inconsistency in labeling across different datasets, it is not feasible to directly compare the mAP scores. Consequently, we have used visualization results to perform our analysis. The Supplementary Material includes the results of qualitative experiments in which we mapped the labels of M^6Doc to labels of other datasets.

Significance of format diversity. Models trained on existing benchmark datasets such as DocBank, DocLayNet, and PubLayNet are not effective in processing some novel scenarios proposed in M^6Doc, such as scanned and photographed images. The specific analysis is as follows: the first row of the first three columns in Figure 3 (a) demonstrates that models trained on DocBank, DocLayNet, and PubLayNet are not effective in identifying document instances in scanned handwritten notes, likely due to the differences between handwritten and printed documents. Rows 3 and 4 of the first two columns reveal that the models trained on DocBank and PubLayNet are unable to process the scanned textbook and photographed book datasets, likely due to the complex backgrounds and tilting and brightness variation phenomena in these images. However, models trained on M^6Doc well handle scanned and photographed images, as shown in columns 4-6 of Figure 3 (a). These results suggest that providing a training set containing scanned and photographed images is crucial for developing models that can handle diverse document formats.

Significance of type diversity. The importance of type diversity is demonstrated in Figure 3 (a), where models trained on DocBank and PubLayNet fail to understand layouts for the new document types (note, textbook, book, newspaper) introduced in M^6Doc. This is due to the fact that DocBank and PubLayNet are limited to only one document type with restricted layouts. In contrast, M^6Doc provides a diverse set of document types and complex layouts, which enables trained models to generalize well on DocBank and PubLayNet. Additionally, as seen in Figure 3 (c), DocLayNet and M^6Doc have different data sources, resulting in significantly different layouts. As a result, models trained on M^6Doc or DocLayNet do not perform well on each other. Hence, the need for diverse document types in a dataset is crucial for addressing generic layout analysis.

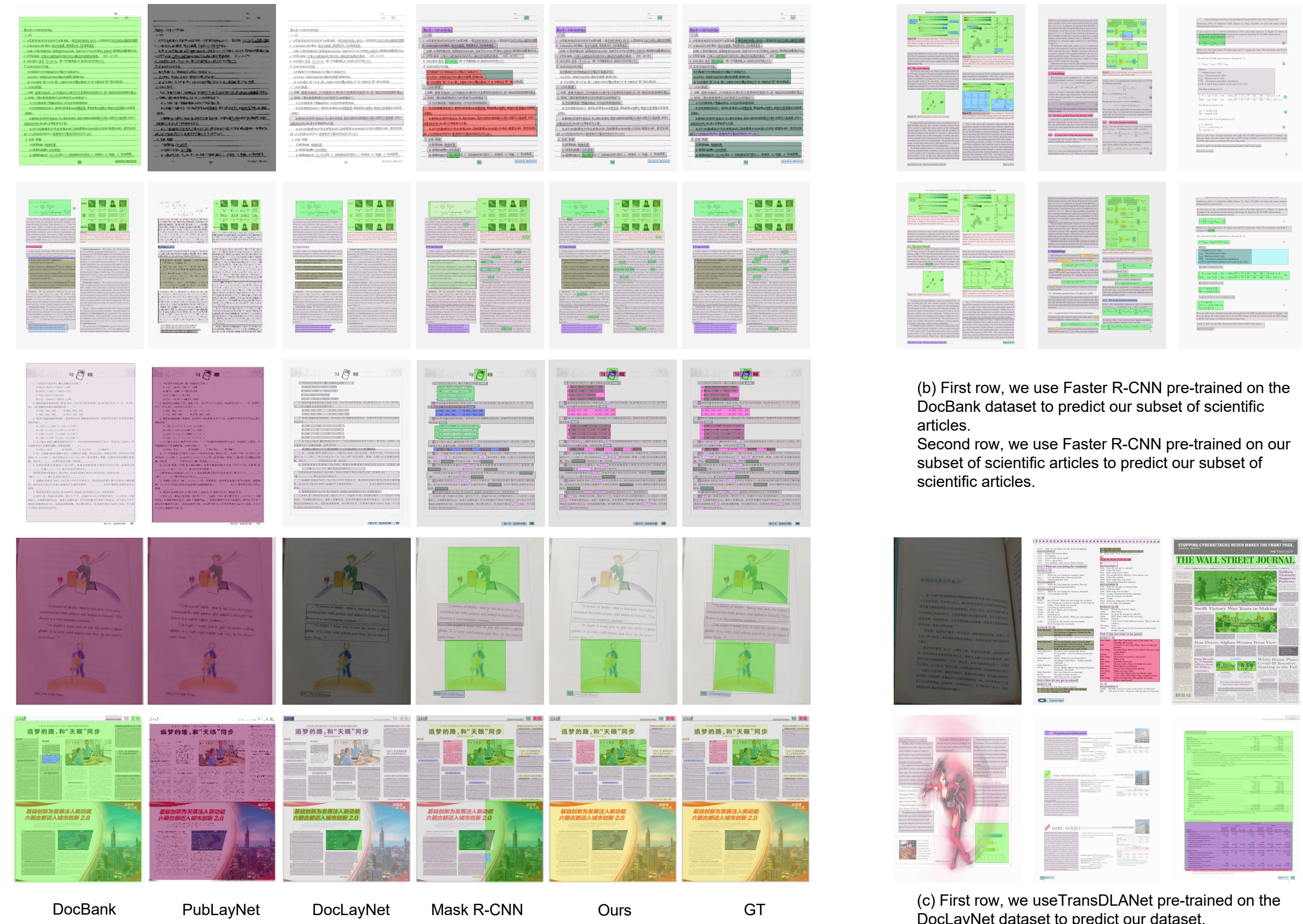
Significance of detailed labels. The importance of detailed labels is demonstrated by comparing the performance of models trained on DocBank and the scientific article subset of M^6Doc, which have the same layout distribution. In our experiments, we used the Faster-RCNN model to predict the test set of the scientific article subset. Figure 3 (b) shows that the model trained on DocBank tends to detect large paragraphs of text while ignoring formulas, likely due to the large region of paragraph annotation used in DocBank. However, our model trained on the scientific article subset is able to avoid this issue and achieve more accurate segmentation results on the same test set by using more detailed labels. It's worth noting that the scientific article subset only contains 600 images, yet adding more detailed labels improved the model's performance. This suggests that having more labels with fewer data may be more beneficial than having fewer labels with more data.

6.4. Comparisons with object detection and instance segmentation methods

In this section, we present the results of a thorough evaluation of M^6Doc using different layout analysis techniques, which could serve as a benchmark for performance comparison. Further experiments on the performance of TransDLANet on nine sub-datasets of M^6Doc are included in the Supplementary Material.

We used RetianNet [19], YOLOv3 [31], GFL [18], FCOS [35], FoveaBox [14], Faster R-CNN [32], Cascade R-CNN [3], Mask R-CNN [9], Cascade Mask R-CNN [3], Deformable DETR [45], and ISTR [11] as object detection baselines, while used HTC [5], SCNet [36], QueryInst [8], SOLO [37], and SOLOv2 [38] as instance segmentation baselines to evaluate the M^6Doc dataset. As the M^6Doc dataset consists of different document types and layouts, and the instances have varying scales, it is challenging for anchor-based regression detection models to set up anchors that can fit all document instances. Therefore, the anchor ratios were adjusted to [0.0625, 0.125, 0.25, 0.5, 1.0, 2.0, 4.0, 8.0, 16.0] instead of the original three anchor ratios [0.5, 1.0, 2.0] for anchor-based models. For pure bounding box methods, the segmentation metrics were calculated using the detected bounding box as the segmentation mask. For pure instance segmentation methods, the minimum bounding rectangle was used to calculate the metrics for the bounding box. The results of the experiments are presented in the following sections.

As shown in Table 3, Mask R-CNN produced lower performance than Faster R-CNN. The same conclusion was reached for the DocLayNet dataset, as shown in Table 5. It indicates that pixel-based image segmentation degrades performance when the dataset contains more complex document layouts. On the other hand, the recall rates of anchor-



(a) The first three columns on the left show the results obtained by our proposed TransDLANet on our dataset, trained separately on Docbank, PubLayNet, and DocLayNet. The fourth and fifth columns present the results obtained using Mask R-CNN and our TransDLANet, both trained on our dataset.

(b) First row, we use Faster R-CNN pre-trained on the DocBank dataset to predict our subset of scientific articles.
Second row, we use Faster R-CNN pre-trained on our subset of scientific articles to predict our subset of scientific articles.

(c) First row, we useTransDLANet pre-trained on the DocLayNet dataset to predict our dataset.
Second row, we use TransDLANet pre-trained on our dataset to predict the DocLayNet dataset.

Figure 3. Visualization results. Zoom in for better view.

Table 3. Performance comparisons on M^6Doc.

Method	Backbone	Object Detection				Instance Segmentation		
		mAP	AP50	AP75	Recall	mAP	AP50	AP75
RetinaNet [19]	ResNet-101	21.4	33.1	23.3	37.4	21.0	33.0	22.6
YOLOv3 [31]	DarkNet-53	59.8	75.6	68.1	72.4	-	-	-
GFL [18]	ResNet-101	34.7	50.8	38.7	48.7	33.8	50.6	37.0
FCOS [35]	ResNet-101	40.6	59.3	45.9	59.5	39.3	58.9	43.1
FoveaBox [14]	ResNet-101	45.1	66.1	51.7	59.4	43.7	65.8	49.2
Faster R-CNN [32]	ResNet-101	49.0	67.8	57.2	57.2	47.8	67.8	55.2
Cascade R-CNN [3]	ResNet-101	54.1	70.4	62.3	61.4	52.7	70.2	60.1
Mask R-CNN [9]	ResNet-101	40.1	58.4	46.2	50.8	39.7	58.4	45.6
Cascade Mask R-CNN [3]	ResNet-101	54.4	70.5	62.9	62.1	52.9	70.4	60.6
HTC [5]	ResNet-101	58.2	74.3	67.2	68.1	57.1	74.4	65.7
SCNet [36]	ResNet-101	56.1	73.5	65.1	67.3	55.3	73.3	63.6
SOLO [37]	ResNet-101	38.7	56.0	42.7	54.9	38.7	56.3	43.0
SOLOv2 [38]	ResNet-101	46.8	67.5	51.4	61.5	48.3	67.5	53.4
Deformable DETR [45]	ResNet-101	57.2	76.8	63.4	**75.2**	55.6	76.5	61.1
QueryInst [8]	ResNet-101	51.0	67.1	58.1	71.0	50.6	67.4	57.5
ISTR [11]	ResNet-101	62.7	80.8	70.8	73.2	62.0	80.7	70.2
Ours	ResNet-101	**64.5**	**82.7**	**72.7**	74.9	**63.8**	**82.6**	**71.9**

based methods are low. The reasons behind this include: (1) It is difficult to set an aspect ratio that can match all the instances. As shown in Figure 3 (a), the fourth and fifth columns present the results obtained using Mask R-CNN and our TransDLANet, both trained on our dataset. Even though we set eight anchor ratio scales, the experimental results show that Mask R-CNN still cannot correctly detect the advertisement instances (the bottom half of the newspaper page in the last row of column 4 of Figure 3 (a)) with large ratio scales but can only detect the paragraphs inside. (2) Anchor-based methods use non-maximum suppression to filter candidate bounding boxes. Therefore, if the overlapped area of the candidate bounding boxes of skewed neighboring document instances is large, they may be filtered out. This leads to detection errors and low recall.

Our approach has achieved a remarkable mean average precision (mAP) of 64.5% on the M^6Doc dataset, surpassing the current state-of-the-art results. TransDLANet eliminates the need for complex anchor design by automatically learning to use a pre-set number of query vectors to encode and decode document instances in images. Additionally, the iterative refinement mechanism of TransDLANet helps overcome the challenges posed by dense arrangement, thereby reducing instance segmentation bias and achieving superior accuracy.

Table 4. Performance comparisons on DocLayNet dataset.

Method	Backbone	Caption	Footnote	Formula	List-item	Page-footer	Page-header	Picture	Section-header	Table	Text	Title	mAP
Faster R-CNN [32]	R101	70.1	73.7	63.5	81.0	58.9	72.0	72.0	68.4	82.2	85.4	79.9	73.4
Mask R-CNN [9]	R50	68.4	70.9	60.1	81.2	61.6	71.9	71.7	67.6	82.2	84.6	76.7	72.4
Mask R-CNN [9]	R101	71.5	71.8	63.4	80.8	59.3	70.0	72.7	69.3	82.9	85.8	80.4	73.5
YOLOv5 [12]	v5x6	**77.7**	**77.2**	**66.2**	**86.2**	**61.1**	**67.9**	**77.1**	**74.6**	**86.3**	**88.1**	**82.7**	**76.8**
Ours	R101	68.2	74.7	61.6	81.0	54.8	68.2	68.5	69.8	82.4	83.8	81.7	72.3

Table 5. Performance comparisons on PubLayNet dataset.

Method	Backbone	Text	Title	List	Table	Figure	mAP
Faster R-CNN [32]	X101	91.0	82.6	88.3	95.4	93.7	90.2
Mask R-CNN [9]	X101	91.6	84.0	88.6	96.0	94.9	91.0
VSR [43]	X101	**96.7**	**93.1**	94.7	**97.4**	96.4	95.7
Ours	R101	94.3	89.21	**95.2**	97.2	**96.6**	94.5

6.5. Performance of the TransDLANet in other datasets

We also conducted experiments on the existing layout dataset to explore the performance of TransDLANet. Tables 4, 5 show the performance of our model on DocLayNet and PubLayNet.

Table 4 displays the performance of our model on the DocLayNet dataset. As evident from the results, our model's performance was comparatively lower than those of other models. Upon further investigation of the visualization results (available in the Supplementary Material), we identified the primary reason for the low accuracy as the fact that we set a fixed number of queries in advance. This design caused our model to miss some instances in the images when multiple queries corresponded to a single instance.

Table 5 demonstrates that TransDLANet achieves comparable or even superior performances to the VSR model for AP in the list, table, and figure categories in the PubLayNet dataset. However, the performance in the text and title categories is inferior to that of VSR. This disparity could be attributed to the fact that VSR exploits both visual and semantic features. The text and title categories exhibit considerable differences in semantic features, so semantic branching can better recognize them. However, TransDLANet does not exploit this distinct feature, so performance is a bit lower compared to VSR.

6.6. Discussion of Failure Cases

The first row of Figure 3 (a) demonstrates the deficiency of both existing models and TransDLANet in detecting handwritten documents due to the unique characteristics of notes. Unlike published documents, handwritten notes are not standardized, and each person may use their own writing style, making them difficult to understand. In addition, the images and tables within the notes subset are not as visually prominent as in other documents, making detection even more challenging. Furthermore, the performance of the current model is unsatisfactory when dealing with real scenario files with significant distortion. Therefore, future researches can explore the use of document rectification [22, 42] as a preliminary step ahead of current methods to solve this challenge. Whatmore, both the current models and the TransDLANet face difficulties in detecting instances that are either densely packed or skewed. Although TransDLANet tries to mitigate this problem by using a transformer encoder to learn the relevance of queries, the problem of missing instance objects still exists. We can solve this problem by training more epochs, but this model converges very slowly. Therefore, future research should further accelerate the convergence rate of TransDLANet and think about how to improve the model's recall.

7. Conclusion

In this paper, we introduce the new M^6Doc dataset, consisting of seven subsets that were acquired using various methods, such as PDF to image conversion, document scanning, and photographing. To our knowledge, M^6Doc is the first dataset that includes real-world scenario files, diverse formats, types, languages, layouts, and comprehensive definitions of logical labels. It can serve as a valuable benchmark for studying logical layout analysis, generic layout analysis, multi-modal layout analysis, formula identification, and table analysis.

We carried out a comprehensive benchmark evaluation of M^6Doc using multiple baselines and conducted detailed analyses. Our findings demonstrate the challenging nature of the M^6Doc dataset and the effectiveness of the detailed label annotations.

For future work, we aim to design specialized models based on the M^6Doc dataset to address the issue of generic layout analysis. Additionally, we plan to explore the challenges of different languages for multi-modal models and consider how to unify visually and semantically consistent annotation formats. Furthermore, we aim to enhance the diversity of our dataset by including further document layouts and types, if possible, to enrich the layout and type diversity.

Acknowledgement

This research is supported in part by NSFC (Grant No.: 61936003), Zhuhai Industry Core and Key Technology Research Project (no. 2220004002350), GD-NSF (No.2021A1515011870), and the Science and Technology Foundation of Guangzhou Huangpu Development District (Grant 2020GH17)

References

[1] Apostolos Antonacopoulos, David Bridson, Christos Papadopoulos, and Stefan Pletschacher. A realistic dataset for performance evaluation of document layout analysis. In *ICDAR*, pages 296–300, 2009. 2, 3

[2] Dario Augusto Borges Oliveira and Matheus Palhares Viana. Fast CNN-Based Document Layout Analysis. In *ICCV*, pages 1173–1180, 2017. 3

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving Into High Quality Object Detection. In *CVPR*, pages 6154–6162, 2018. 6, 7

[4] Samuele Capobianco, Leonardo Scommegna, and Simone Marinai. Historical Handwritten Document Segmentation by Using a Weighted Loss. In *IAPR*, pages 395–406, 2018. 3

[5] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid Task Cascade for Instance Segmentation. In *CVPR*, pages 4974–4983, 2019. 6, 7

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[7] Randa Elanwar, Wenda Qin, Margrit Betke, and Derry Wijaya. Extracting text from scanned Arabic books: a large-scale benchmark dataset and a fine-tuned Faster-R-CNN model. *IJDAR*, 24(4):349–362, 2021. 3

[8] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances As Queries. In *ICCV*, pages 6910–6919, 2021. 6, 7

[9] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2961–2969, 2017. 3, 6, 7, 8

[10] Jie Hu, Liujuan Cao, Yao Lu, ShengChuan Zhang, Yan Wang, Ke Li, Feiyue Huang, Ling Shao, and Rongrong Ji. ISTR: End-to-End Instance Segmentation with Transformers. *arXiv preprint arXiv:2105.00637*, 2021. 5

[11] Jie Hu, Liujuan Cao, Yao Lu, ShengChuan Zhang, Yan Wang, Ke Li, Feiyue Huang, Ling Shao, and Rongrong Ji. ISTR: End-to-End Instance Segmentation with Transformers. *arXiv preprint arXiv:2105.00637*, 2021. 6, 7

[12] Glenn Jocher, Alex Stoken, Ayush Chaurasia, J Borovec, T Xie, Y Kwon, K Michael, L Changyu, J Fang, V Abhiram, et al. ultralytics/yolov5: v6. 0-YOLOv5n "Nano" models. *Roboflow integration, TensorFlow export, OpenCV DNN support (v6. 0)[Computer software]. Zenodo*, 2021. 8

[13] Koichi Kise, Akinori Sato, and Motoi Iwata. Segmentation of Page Images Using the Area Voronoi Diagram. *Computer Vision and Image Understanding*, 70(3):370–382, 1998. 3

[14] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. FoveaBox: Beyound Anchor-Based Object Detection. *IEEE TIP*, 29:7389–7398, 2020. 6, 7

[15] Joonho Lee, Hideaki Hayashi, Wataru Ohyama, and Seiichi Uchida. Page Segmentation using a Convolutional Neural Network with Trainable Co-Occurrence Features. In *ICDAR*, pages 1023–1028, 2019. 2, 3

[16] Kai Li, Curtis Wigington, Chris Tensmeyer, Handong Zhao, Nikolaos Barmpalios, Vlad I. Morariu, Varun Manjunatha, Tong Sun, and Yun Fu. Cross-Domain Document Object Detection: Benchmark Suite and Method. In *CVPR*, pages 12915–12924, 2020. 3

[17] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. DocBank: A Benchmark Dataset for Document Layout Analysis. In *ICCL*, pages 949–960, 2020. 2, 3, 5

[18] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. In *NeurIPS*, volume 33, pages 21002–21012, 2020. 6, 7

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *ICCV*, pages 2980–2988, 2017. 6, 7

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, pages 740–755, 2014. 5

[21] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019. 6

[22] Ke Ma, Sagnik Das, Zhixin Shu, and Dimitris Samaras. Learning From Documents in the Wild to Improve Document Unwarping. In *ACM SIGGRAPH*, pages 1–9, 2022. 8

[23] Logan Markewich, Hao Zhang, Yubin Xing, Navid Lambert-Shirzad, Zhexin Jiang, Roy Ka-Wei Lee, Zhi Li, and Seok-Bum Ko. Segmentation for document layout analysis: not dead yet. *IJDAR*, page 67–77, 2022. 3

[24] G. Nagy, S. Seth, and M. Viswanathan. A prototype document image analysis system for technical journals. *Computer*, 25(7):10–22, 1992. 3

[25] Anoop M. Namboodiri and Anil K. Jain. Document Structure and Layout Analysis. In *Digital Document Processing: Major Directions and Recent Advances*, pages 29–48, 2007. 2

[26] L. O'Gorman. The document spectrum for page layout analysis. *IEEE TPAMI*, 15(11):1162–1173, 1993. 3

[27] S. Ares Oliveira, B. Seguin, and F. Kaplan. dhSegment: A Generic Deep-Learning Approach for Document Segmentation. In *ICFHR*, pages 7–12, 2018. 3

[28] Nazih Ouwayed and Abdel Belaïd. A general approach for multi-oriented text line extraction of handwritten documents. *IJDAR*, 15(4):297–314, 2012. 3

[29] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter W J Staar. DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis. In *ACM SIGKDD*, page 3743–3751, 2022. 2, 3, 5

[30] Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. CascadeTabNet: An Approach for End to End Table Detection and Structure Recognition From Image-Based Documents. In *CVPR*, pages 572–573, 2020. 3

[31] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 6, 7

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, volume 28, 2015. 6, 7, 8

[33] Rana SM Saad, Randa I Elanwar, NS Abdel Kader, Samia Mashali, and Margrit Betke. BCE-Arabic-v1 dataset: Towards interpreting Arabic document images for people with visual impairments. In *ACM*, pages 1–8, 2016. 3

[34] Wang Shaoqiang. *Page Design: New Layout & Editorial Design.* Sandu Publishing, 2019. 4

[35] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully Convolutional One-Stage Object Detection. In *ICCV*, pages 9627–9636, 2019. 6, 7

[36] Thang Vu, Haeyong Kang, and Chang D. Yoo. SCNet: Training Inference Sample Consistency for Instance Segmentation. *AAAI*, 35(3):2701–2709, 2021. 6, 7

[37] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: Segmenting Objects by Locations. In *ECCV*, page 649–665, 2020. 6, 7

[38] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. SOLOv2: Dynamic and Fast Instance Segmentation. In *NeurIPS*, volume 33, pages 17721–17732, 2020. 6, 7

[39] K. Y. Wong, R. G. Casey, and F. M. Wahl. Document Analysis System. *IBM Journal of Research and Development*, 26(6):647–656, 1982. 3

[40] Yue Xu, Fei Yin, Zhaoxiang Zhang, and Cheng-Lin Liu. Multi-Task Layout Analysis for Historical Handwritten Documents Using Fully Convolutional Networks. In *IJCAI*, page 1057–1063, 2018. 3

[41] Xiao Yang, Ersin Yumer, Paul Asente, Mike Kraley, Daniel Kifer, and C. Lee Giles. Learning to Extract Semantic Structure From Documents Using Multimodal Fully Convolutional Neural Networks. In *CVPR*, pages 5315–5324, 2017. 2, 3

[42] Jiaxin Zhang, Canjie Luo, Lianwen Jin, Fengjun Guo, and Kai Ding. Marior: Margin Removal and Iterative Content Rectification for Document Dewarping in the Wild. In *ACM MM*, pages 2805–2815, 2022. 8

[43] 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. In *ICDAR*, pages 115–130, 2021. 3, 8

[44] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: Largest dataset ever for document layout analysis. In *ICDAR*, pages 1015–1022, 2019. 2, 3, 5

[45] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *ICLR*, pages 2988–2997, 2021. 6, 7