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
{wusihang2, zhuqiyuan2, xiezecheng1, lijing260}@huawei.com
eelwjin@scut.edu.cn
danny_ding@intsig.net

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Paper

M^{6}Doc
Github Homepage

Overview

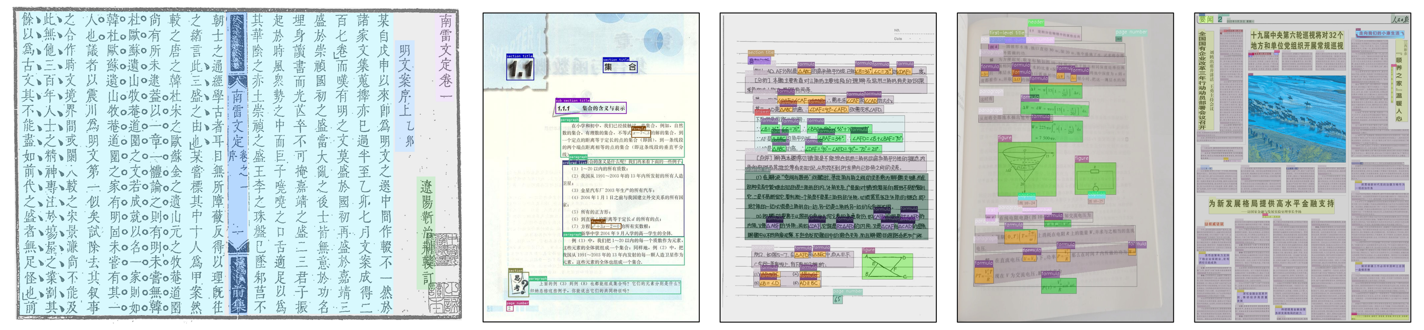
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Background

Document layout analysis (DLA): is a crucial step in digital systems that aims to decompose page images into homogeneous regions, such as text, images, and tables.

Difficulties in the task of document layout analysis include:

- **Document diversity** (including document variety in terms of format, type, layout, and language).
- **Document image quality diversity**.
- **Mutual influence between elements in documents**.
- **Ambiguity in document elements**.



Motivation

Limitations of currently available modern document datasets:

- **Small dataset scale**. Most datasets only containing a few hundred images.
- **Only PDF document format**, which poses a significant challenge for evaluating the effectiveness of DLA methods in real-world scenarios.
- **Severe lack of diversity**, which will affect the development of routine layout analysis in multiple fields.
- **Limited document languages.** DLA methods may encounter domain shift problems in different languages, which remain unexplored.
- **Few annotation categories,** which prevents more granular layout information extraction.

Contributions

In this paper, we presents a modern dataset and method for document layout analysis.

1. We have constructed and public the M⁶Doc dataset, which is a modern document dataset that supports multiple formats, types, layouts, languages, and annotation categories.
2. M⁶Doc is the first layout analysis dataset that contains both real-world (photographed and scanned) and born-digital document images. Additionally, it is the first dataset that includes Chinese documents.
3. M⁶Doc contains 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.
4. We propose the TransDLANet, a Transformer-based method for DLA.

Dataset Description

M⁶ designation represents six properties:

- **Multi-Format**: scanned, photographed, and PDF documents;
- **Multi-Type**: scientific articles, textbooks, books, test papers, magazines, newspapers, and notes;
- **Multi-Layout**: rectangular, Manhattan, non-Manhattan, and multi-column Manhattan;
- **Multi-Language**: Chinese and English;
- **Multi-Annotation Category**: 74 types of annotation labels with 237,116 annotation instances in 9,080 manually annotated pages;
- **Modern documents**.

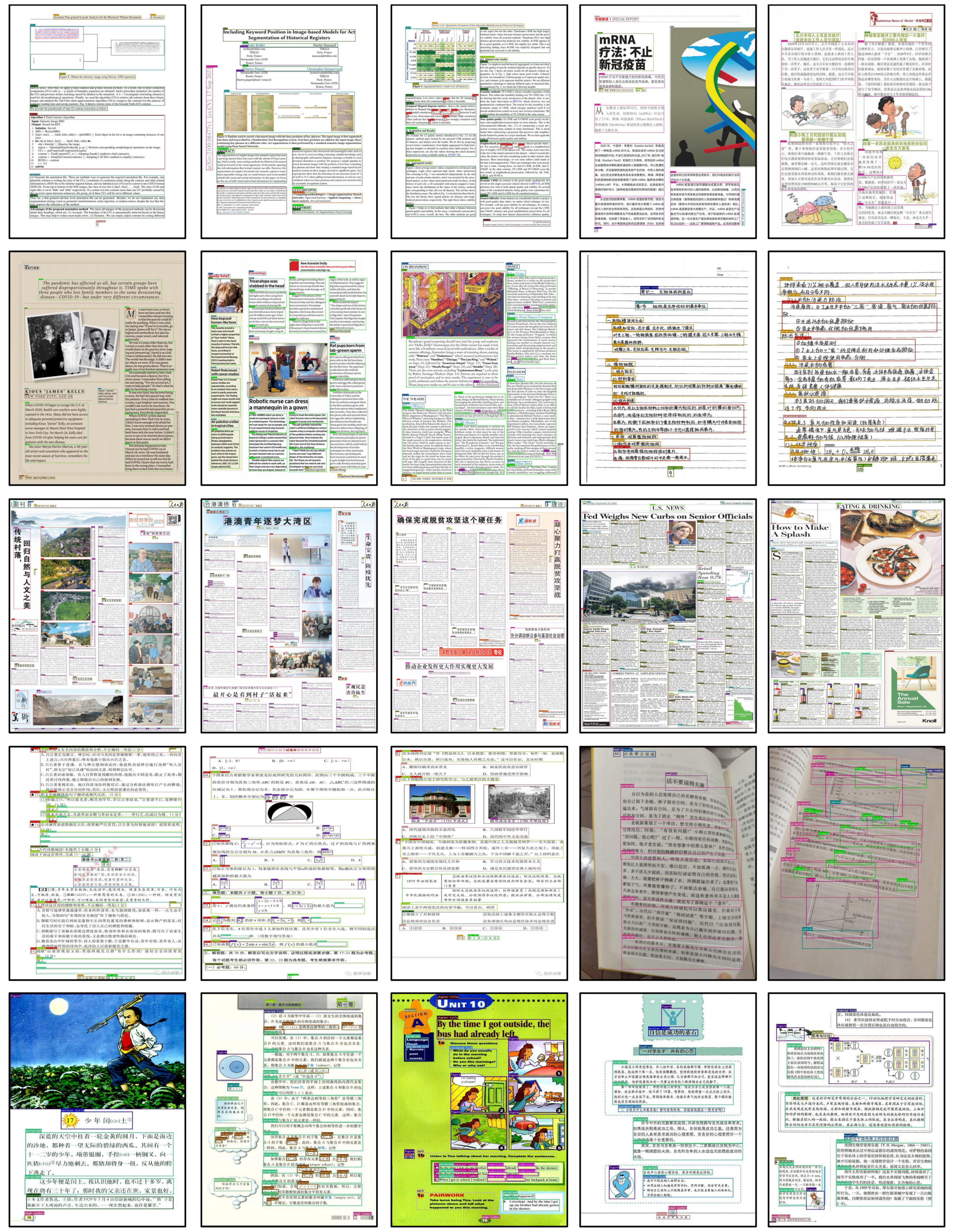


Figure1. Example annotations of the M⁶Doc.

Methodology

TransDLANet contains four main components:

1. **CNN-based backbone**: extract document image features.
2. **Transformer encoder**: performs self-attentive feature learning on query embedding vectors and uses an adaptive element matching mechanism to further enhance the association between document instances encoded by the query vectors.
3. **Dynamic decode**: fuses the query vector with the features of the bounding box image region obtained by the query vector using RoIAlign.
4. **Shared parameter MLP branches**: decodes the classification confidence, the bounding boxes' coordinate position, and the segmentation mask of the document instance region.

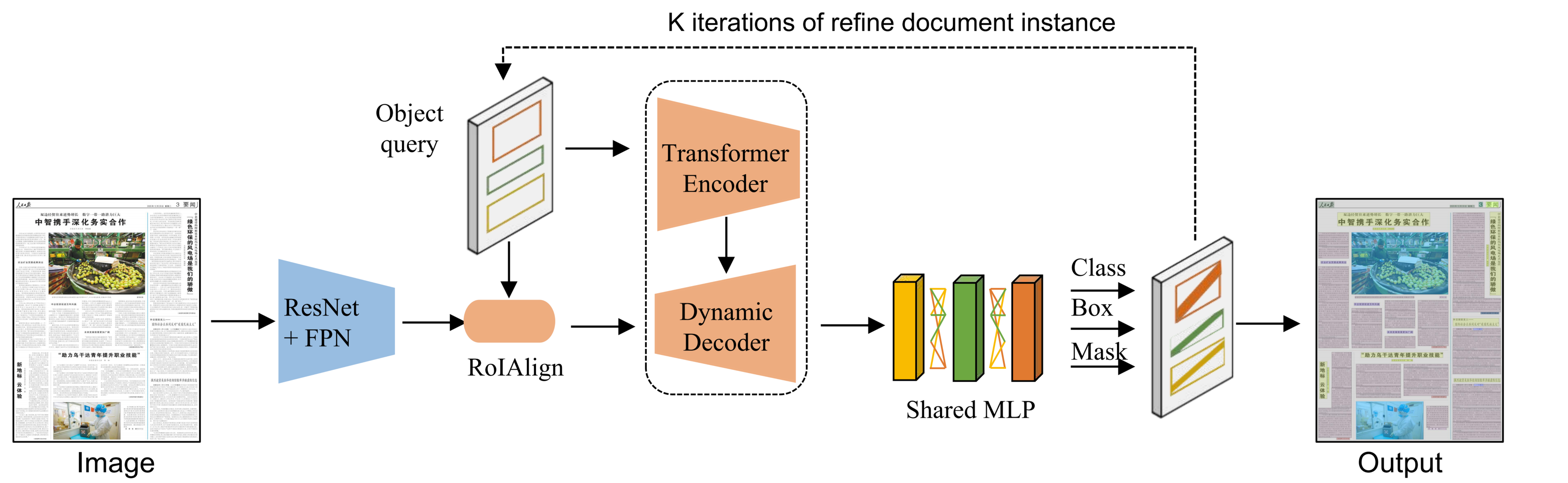


Figure2. The pipeline of TransDLANet.

Experimental Results

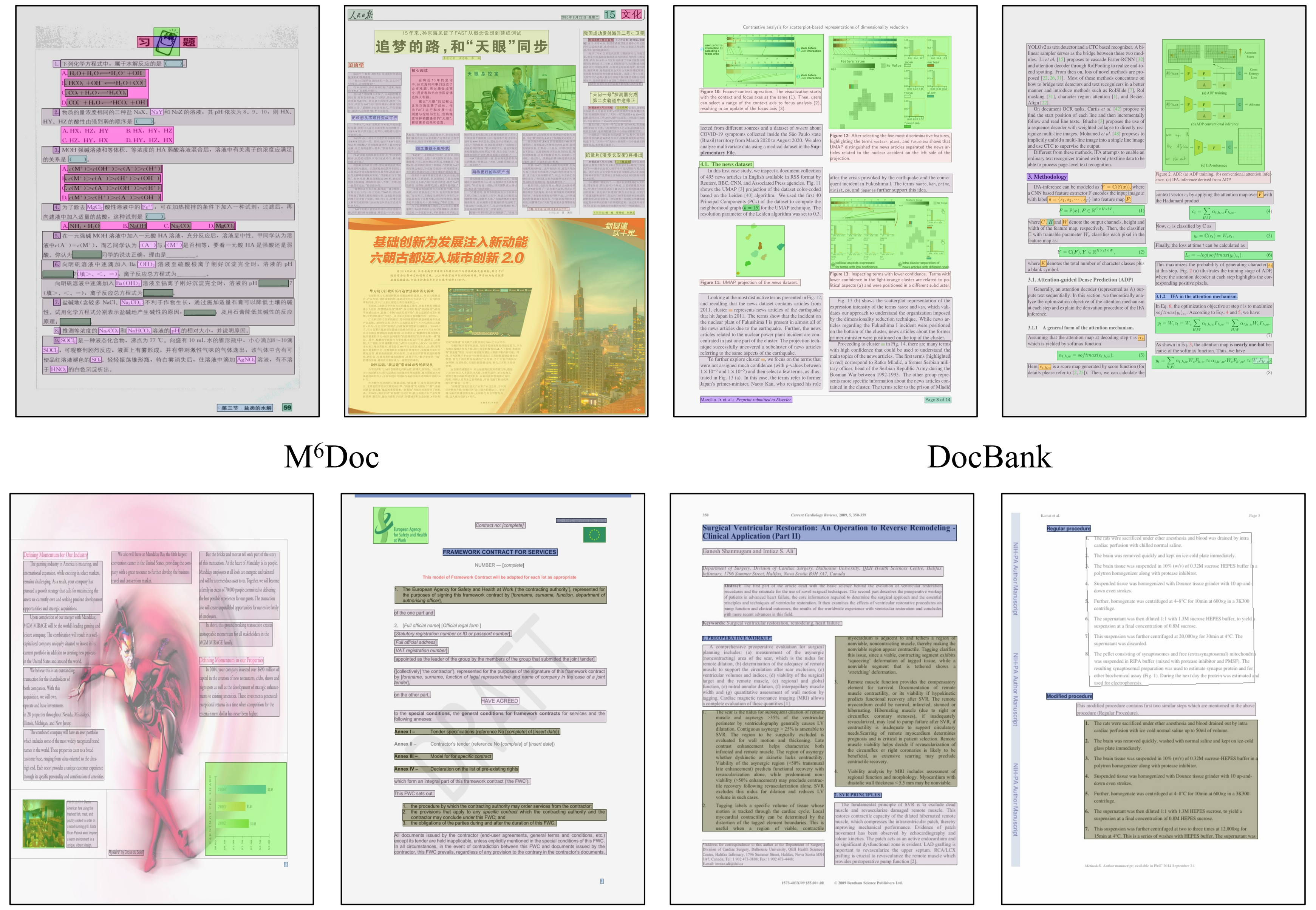
Performance comparisons on M⁶Doc

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**

Ablation study for different components

Ablation study	Object Detection				Instance Segmentation		
	mAP	AP50	AP75	Recall	mAP	AP50	AP75
Ours, w/o Transformer encoder	47.8	62.6	54.4	65.4	47.3	62.6	53.9
Ours, w/o Dynamic decoder	52.8	70.5	48.0	73.9	52.3	70.4	47.6
Ours, w/o Shared_MLP	64.2	82.3	72.1	74.1	63.6	82.2	71.2
Ours	**64.5**	**82.7**	**72.7**	**74.9**	**63.8**	**82.6**	**71.9**

TransDLANet results on different datasets



M⁶Doc DocBank DocLayNet PubLayNet

Background

Document Layout Analysis (DLA) is a key step in the digitization system, which aims to decompose page images into homogeneous regions, such as text, images, tables, etc.

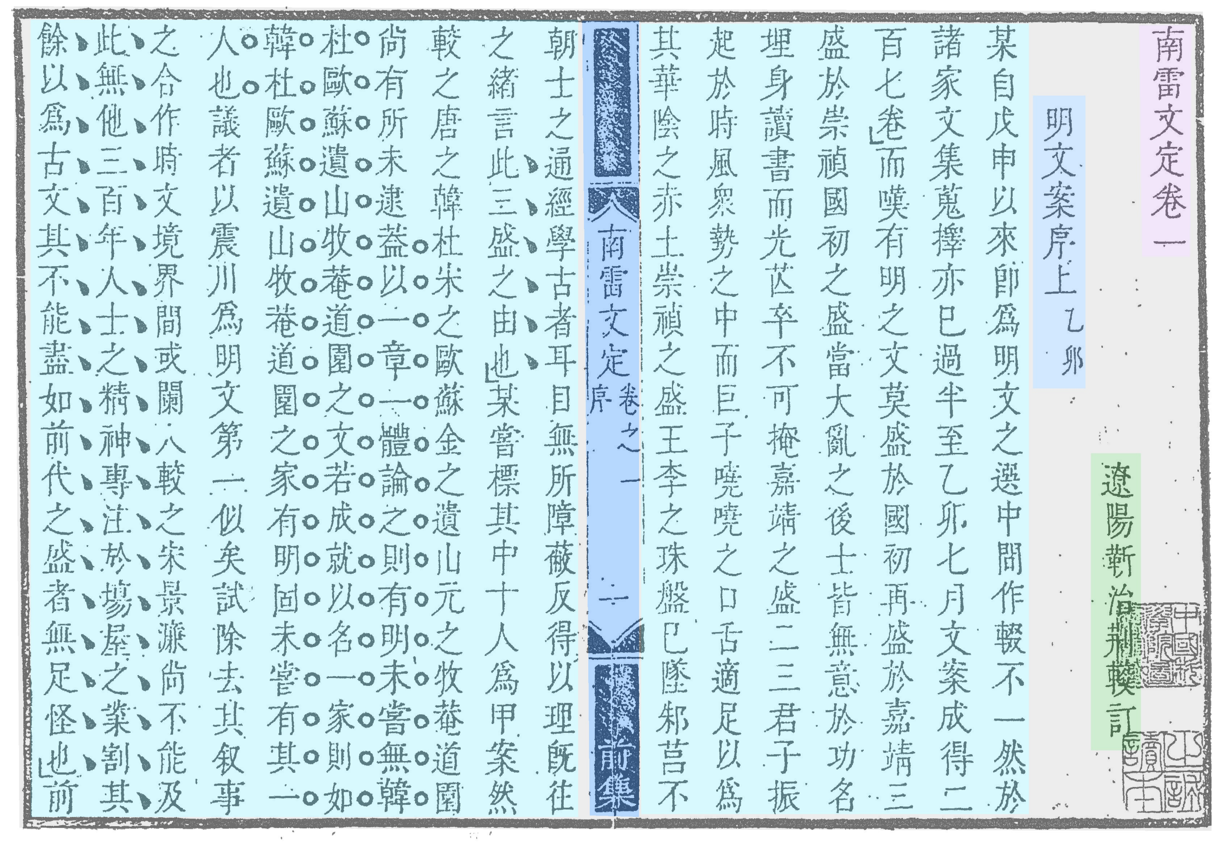
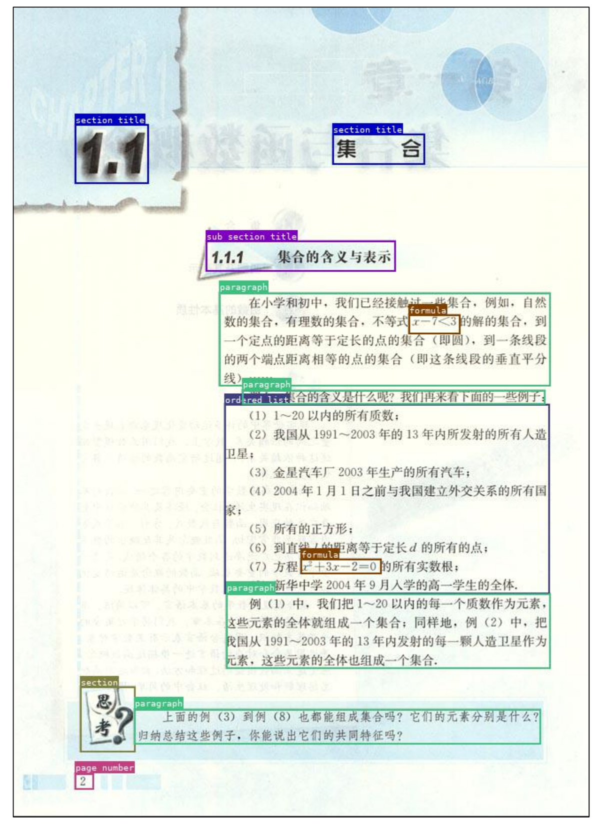
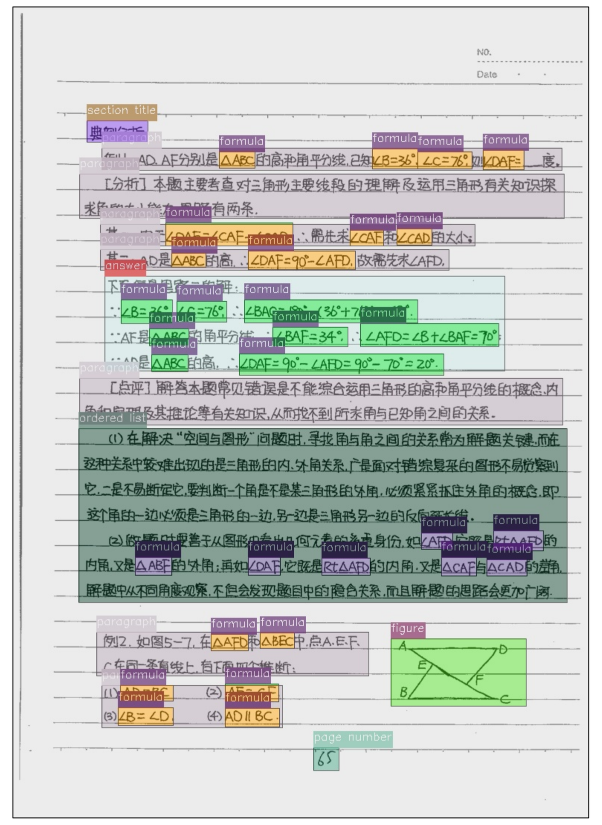
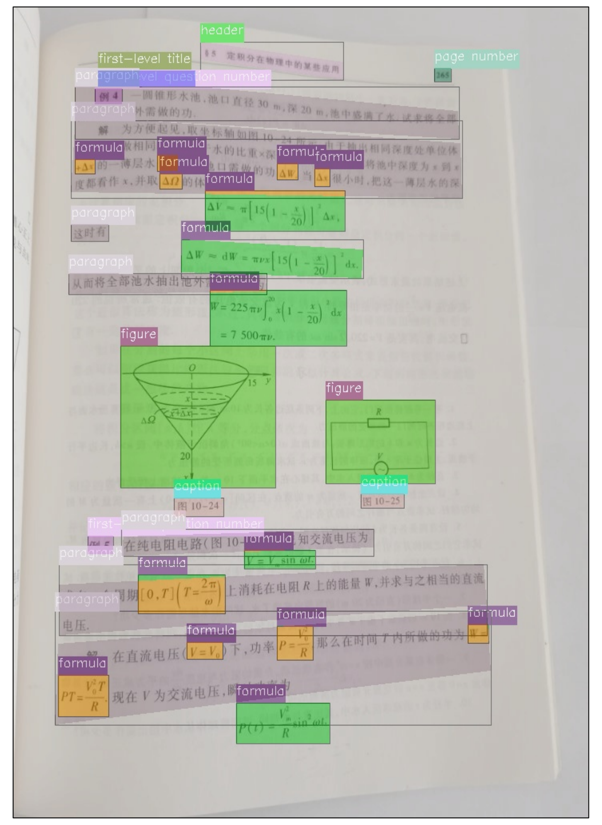
Difficulties in the task of document layout analysis include:

(1) Document diversity (including document variety in terms of format, type, layout, and language).

(2) Document image quality diversity (including distortion, varying illumination, and blur).

(3) Mutual influence between elements in documents.

(4) Ambiguity in document elements.


Motivation

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

There are certain limitations associated with existing datasets for layout analysis.

(1) Small size. Early DLA datasets were small-scale and contained only hundreds of images.

(2) Limited document format. The formats of current public large-scale datasets are all PDF documents. It presents a huge challenge to evaluate the effectiveness of different methods in realistic scenarios.

(3) Limited document diversity. The lack of style diversity would prejudice the development of multi-domain general layout analysis.

(4) Limited document languages. 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.

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^6Doc (**Ours**)	9,080	74	237,116	Manual	PDF, Scanned, Photographed	Scientific articles, Textbooks, Books, Test papers, Magazines, Newspapers, Notes	English, Chinese

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

M⁶Doc Dataset

The M⁶ 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);

(6) **Modern documents**.

M⁶DOC is the first DLA dataset to consider real-world documents and include the most detailed manual annotations, consisting of 9,080 document images and 237,116 annotated instances.

Data sources include but are not limited to:

[1]https://arxiv.org/

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

[3]https://vk.com/

M⁶Doc
Github Homepage

Label Definition

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**,

(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 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.

Category	Training		Validate		Test	
	Number	%	Number	%	Number	%
background	0	0.000	0	0.000	0	0.000
QR code	59	0.041	15	0.065	23	0.032
advertisement	257	0.180	45	0.194	145	0.205
algorithm	12	0.008	3	0.013	12	0.017
answer	165	0.115	30	0.129	77	0.109
author	2,424	1.695	403	1.736	1,188	1.676
barcode	10	0.007	1	0.004	3	0.004
bill	3	0.002	2	0.009	3	0.004
blank	189	0.132	58	0.250	90	0.127
bracket	863	0.603	164	0.707	273	0.385
breakout	411	0.287	72	0.310	188	0.265
byline	1,276	0.892	185	0.797	660	0.931
caption	3,508	2.452	605	2.607	1,766	2.492
catalogue	39	0.027	10	0.043	19	0.027
chapter title	245	0.171	33	0.142	124	0.175
code	62	0.043	7	0.030	31	0.044
correction	9	0.006	1	0.004	6	0.008
credit	1,523	1.065	255	1.099	728	1.027
dateline	901	0.630	140	0.603	482	0.680
drop cap	414	0.289	71	0.306	234	0.330
editor's note	39	0.027	4	0.017	9	0.013
endnote	35	0.024	4	0.017	19	0.027
examinee information	8	0.006	2	0.009	6	0.008
fifth-level title	13	0.009	2	0.009	20	0.028
figure	7,614	5.323	1,242	5.351	3,762	5.309
first-level question number	5,669	3.963	930	4.007	2,740	3.866
first-level title	586	0.410	81	0.349	292	0.412
flag	30	0.021	5	0.022	12	0.017
folio	1,442	1.008	213	0.918	685	0.967
footer	1,984	1.387	310	1.336	987	1.393
footnote	295	0.206	49	0.211	139	0.196
formula	1,3090	9.151	2,058	8.867	6,191	8.736
fourth-level section title	15	0.010	3	0.013	19	0.027
fourth-level title	70	0.049	13	0.056	66	0.093
header	1,877	1.312	297	1.280	969	1.367
headline	4,115	2.877	643	2.770	1,981	2.795
index	214	0.150	36	0.155	100	0.141
inside	16	0.011	1	0.004	5	0.007

Category	Training		Validate		Test	
	Number	%	Number	%	Number	%
institute	60	0.042	9	0.039	28	0.040
jump line	381	0.266	63	0.271	180	0.254
kicker	516	0.361	91	0.392	257	0.363
lead	664	0.464	109	0.470	285	0.402
marginal note	238	0.166	37	0.159	101	0.143
matching	7	0.005	1	0.004	8	0.011
mugshot	73	0.051	11	0.047	46	0.065
option	3,198	2.236	515	2.219	1,577	2.225
ordered list	1,012	0.707	172	0.741	510	0.720
other question number	42	0.029	3	0.013	31	0.044
page number	4,782	3.343	803	3.460	2,383	3.363
paragraph	65,642	45.891	10,575	45.562	33,069	46.664
part	524	0.366	89	0.383	283	0.399
play	10	0.007	3	0.013	2	0.003
poem	98	0.069	18	0.078	33	0.047
reference	149	0.104	23	0.099	62	0.087
sealing line	3	0.002	2	0.009	5	0.007
second-level question number	2,773	1.939	377	1.624	1,330	1.877
second-level title	273	0.191	48	0.207	140	0.198
section	2,508	1.753	408	1.758	1,228	1.733
section title	897	0.627	171	0.737	442	0.624
sidebar	54	0.038	10	0.043	27	0.038
sub section title	567	0.396	107	0.461	269	0.380
subhead	1,998	1.397	394	1.698	1,069	1.508
subsub section title	101	0.071	21	0.090	71	0.100
supplementary note	986	0.689	158	0.681	487	0.687
table	821	0.574	146	0.629	409	0.577
table caption	287	0.201	41	0.177	143	0.202
table note	8	0.006	2	0.009	5	0.007
teasers	32	0.022	7	0.030	7	0.010
third-level question number	240	0.168	36	0.155	102	0.144
third-level title	146	0.102	44	0.190	94	0.133
title	201	0.141	35	0.151	100	0.141
translator	73	0.051	11	0.047	38	0.054
underscore	3,687	2.578	590	2.542	1,717	2.423
unordered list	497	0.347	84	0.362	271	0.382
weather forecast	10	0.007	3	0.013	3	0.004
Total	**143,040**	**100**	**23,210**	**100**	**70,866**	**100**

TransDLANet Model

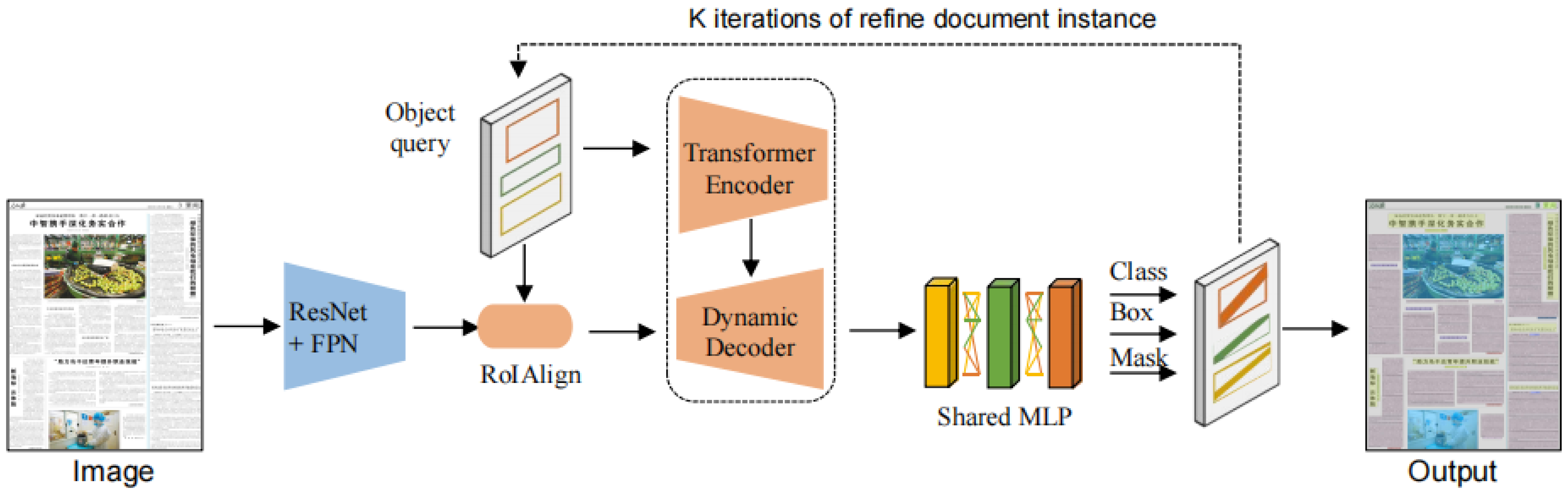


Figure 2. The pipeline of TransDLANet contains four main components: 1) a CNN-based backbone; 2) a transformer encoder; 3) a dynamic decoder that decodes the instance-level features; and 4) three shared multi-layer perceptron(MLP) branches that obtain the classification confidence, bounding boxes, and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

(1) Backbone: extracts features.

(2) Transformer encoder: used as a position-encoding free feature fusion method for learning the relationship between query vectors. Meanwhile, with an adaptive element matching mechanism, the query embedding is better matched with real annotations.

(3) Dynamic interactive: decoding module fuses and interacts the query vectors with the bounding box image region features.

(4) Shared parameter MLP: performs both detection and segmentation to achieve more accurate document instance segmentation.

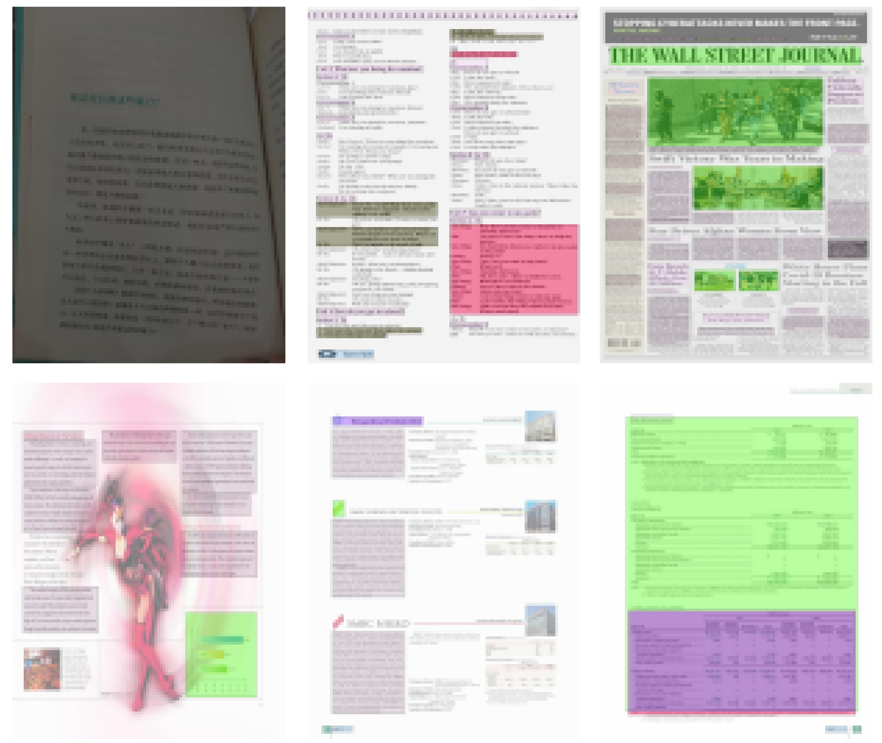
Experimental results and analysis

Datasets: M^6Doc, DocBank [1], PubLayNet [2], and DocLayNet [3].

Significance of M^6Doc

(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 use TransDLANet 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.

Cross-validation experiments

Training on	labels	Testing on M^6Doc	Testing on DocBank
M^6Doc	figure	69.77	42.67
	table	72.57	43.29
	title	58.16	36.47
	mAP	66.83	40.81
DocBank	figure	20.70	58.47
	table	18.01	62.98
	title	7.26	83.70
	mAP	15.32	68.38

Training on	labels	Testing on M^6Doc	Testing on PubLayNet
M^6Doc	Text	72.56	60.21
	Title	63.50	53.26
	List	38.95	59.15
	Table	74.83	79.66
	Figure	74.23	62.45
	mAP	64.81	62.94
PubLayNet	Text	20.46	94.26
	Title	12.92	89.20
	List	7.41	95.18
	Table	12.98	97.21
	Figure	8.39	96.62
	mAP	12.43	94.49

Training on	labels	Testing on M^6Doc	Testing on DocLayNet
M^6Doc	Caption	61.9	12.7
	Footnote	70.2	5.8
	Formula	47.7	9.0
	Page-footer	71.0	8.0
	Page-header	71.1	3.2
	Picture	75.4	30.0
	Section-header	73.2	5.0
	Table	78.0	34.2
	Text	80.0	26.2
	Title	71.1	0.4
	mAP	69.96	13.45
DocLayNet	Caption	13.2	68.2
	Footnote	7.0	74.7
	Formula	2.5	61.6
	Page-footer	8.2	54.8
	Page-header	0.8	68.2
	Picture	40.1	68.5
	Section-header	1.6	69.8
	Table	39.2	82.4
	Text	45.8	83.8
	Title	3.6	81.7
	mAP	16.20	71.37

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

Experimental results and analysis

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Ablation study for TransDLANet

Table 8. Ablation study for mask embedding dimension.

Ablation study	Object Detection				Instance Segmentation		
	mAP	AP50	AP75	Recall	mAP	AP50	AP75
embedding dimension = 20	63.2	81.0	72.0	74.0	62.7	80.9	71.3
embedding dimension = 40	**64.5**	**82.7**	**72.7**	**74.9**	**63.8**	**82.6**	**71.9**
embedding dimension = 60	63.4	81.1	74.6	72.3	62.8	81.0	71.3

Table 9. Ablation study for different components.

Ablation study	Object Detection				Instance Segmentation		
	mAP	AP50	AP75	Recall	mAP	AP50	AP75
Ours, w/o Transformer encoder	47.8	62.6	54.4	65.4	47.3	62.6	53.9
Ours, w/o Dynamic decoder	52.8	70.5	48.0	73.9	52.3	70.4	47.6
Ours, w/o Shared_MLP	64.2	82.3	72.1	74.1	63.6	82.2	71.2
Ours	**64.5**	**82.7**	**72.7**	**74.9**	**63.8**	**82.6**	**71.9**

Impact of Document Size

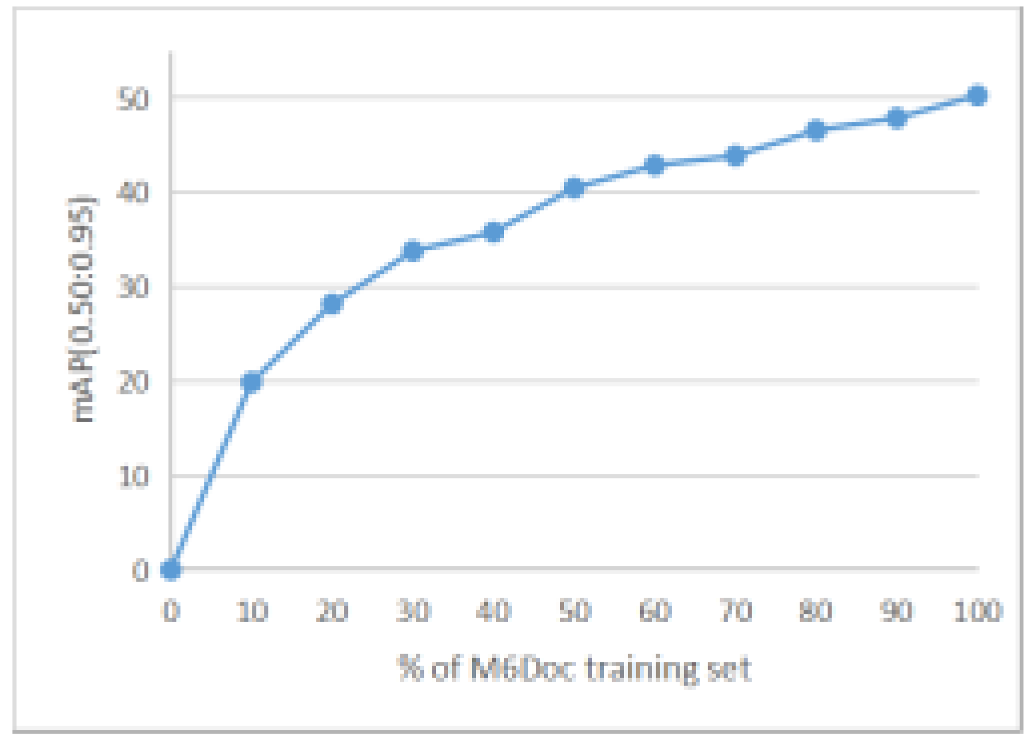


Figure 1. Mask R-CNN network with ResNet50 backbone trained on increasing fractions of the M^6Doc dataset.

Impact of Class Labels

Category	**note_v1**	**note_v2**
answer	8.1	5.8
bracket	0.0	-
caption	0.0	0.1
catalogue	19.2	14.3
chapter title	18.0	18.3
fifth-level title	2.4	paragraph
figure	0.4	0.7
first-level question number	0.0	-
first-level title	13.6	paragraph
footer	62.5	58.5
formula	1.5	2.6
fourth-level title	19.5	paragraph
option	0.0	0
ordered list	3.2	2.2
page number	55.3	55.3
paragraph	28.1	41.3
part	0.0	0
second-level question number	0.0	-
second-level title	0.0	paragraph
section	12.4	17
section title	9.3	7
sub section title	5.1	5.9
supplementary note	0.0	0
table	22.7	17.4
third-level title	25.8	paragraph
underscore	0.0	-
unordered list	28.5	25.9
mAP	12.4	15.1

Comparisons with object detection and instance segmentation methods

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**

TransDLANet results on different datasets

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

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 10. Performance comparisons on nine subsets of M^6Doc.

Method	Backbone	**scientific article**						**magazine_ch**						**magazine_en**					
		Object Detection			Instance Segmentation			Object Detection			Instance Segmentation			Object Detection			Instance Segmentation		
		mAP	AP50	AP75	mAP	AP50	AP75	mAP	AP50	AP75	mAP	AP50	AP75	mAP	AP50	AP75	mAP	AP50	AP75
FCOS	ResNet-101	26.3	45.1	27.2	25.9	44.9	26.5	40.1	57.3	45.8	39.7	57.3	45.1	38.4	60.5	42.5	37.8	60.5	41.8
FoveaBox	ResNet-101	29.8	52.7	30.7	29.4	52.4	30.8	43.4	59.7	50.4	43.1	59.7	50.0	41.5	66.7	44.0	41.1	66.9	42.8
Faster R-CNN	ResNet-101	41.5	62.0	46.8	40.9	61.7	45.4	49.0	63.5	58.3	48.4	63.5	57.1	47.9	66.7	55.9	47.1	66.7	54.5
Cascade R-CNN	ResNet-101	39.8	55.5	45.7	39.4	55.8	44.8	51.3	63.5	60.0	50.7	63.4	59.5	46.3	61.3	54.0	45.9	61.2	53.3
Mask R-CNN	ResNet-101	34.9	53.5	37.6	35.0	53.3	38.3	47.1	61.1	55.4	46.4	61.0	55.2	43.9	60.8	50.2	43.2	60.7	50.3
Cascade Mask R-CNN	ResNet-101	41.8	57.3	47.4	41.4	57.1	46.6	46.4	58.6	54.0	46.0	58.6	53.9	59.4	74.7	69.0	58.5	74.9	68.3
HTC	ResNet-101	49.2	66.0	55.2	48.8	65.9	54.3	**51.9**	**64.7**	**60.3**	**50.7**	**64.8**	**59.4**	60.4	77.3	71.7	59.6	77.3	70.9
SCNet	ResNet-101	36.0	51.4	40.9	35.5	51.3	39.4	49.0	62.2	57.2	48.3	62.2	56.9	49.1	66.3	57.3	48.2	66.2	56.2
SOLO	ResNet-101	32.1	51.1	33.9	32.8	53.5	33.9	35.6	53.0	39.9	37.1	54.6	42.1	34.4	59.9	32.8	36.1	59.6	34.5
SOLOv2	ResNet-101	33.5	54.0	35.9	33.0	54.5	36.0	33.8	51.8	36.5	35.8	53.7	39.5	45.3	71.1	49.4	47.9	72.7	54.0
Deformable DETR	ResNet-101	32.3	43.7	35.8	32.0	43.7	35.3	40.2	55.1	45.8	39.9	55.0	45.0	51.1	72.0	58.6	50.8	71.9	57.7
QueryInst	ResNet-101	32.0	46.2	36.3	31.6	45.8	35.5	37.4	49.7	43.2	37.6	50.4	43.5	44.8	60.6	53.8	44.5	61.1	53.2
ISTR	ResNet-101	**61.8**	**80.3**	**69.7**	**61.1**	**80.2**	**70.2**	50.5	63.4	58.4	50.5	63.5	58.4	66.3	83.0	75.6	65.7	83.0	75.0
Ours	ResNet-101	59.7	78.7	68.2	59.1	78.5	67.0	50.2	63.0	57.7	49.8	62.9	57.3	**68.2**	**85.0**	**77.2**	**68.2**	**85.0**	**77.2**

Method	Backbone	**note**						**newspaper_ch**						**newspaper_en**					
		Object Detection			Instance Segmentation			Object Detection			Instance Segmentation			Object Detection			Instance Segmentation		
		mAP	AP50	AP75	mAP	AP50	AP75	mAP	AP50	AP75	mAP	AP50	AP75	mAP	AP50	AP75	mAP	AP50	AP75
GFL	ResNet-101	11.1	19.1	11.7	11.0	19.1	12.1	22.1	35.9	24.7	21.8	35.9	23.9	20.5	30.1	22.8	20.3	30.0	22.7
FCOS	ResNet-101	19.1	36.7	18.7	18.9	36.5	17.8	22.7	41.7	21.1	22.5	41.6	21.7	17.8	32.6	17.4	17.5	32.4	16.8
FoveaBox	ResNet-101	19.8	36.2	21.5	19.5	36.3	20.3	21.5	37.6	22.3	21.3	37.5	22.1	35.5	56.4	39.6	34.9	56.3	37.7
Faster R-CNN	ResNet-101	29.3	46.1	33.9	28.9	46.1	32.9	32.2	50.6	33.9	32.3	50.8	33.9	34.3	50.7	39.4	34.0	50.7	39.3
Cascade R-CNN	ResNet-101	22.5	34.9	27.3	22.3	34.9	27.3	27.7	42.6	29.8	27.7	42.7	30.0	26.3	36.5	29.6	26.0	36.3	29.3
Mask R-CNN	ResNet-101	15.1	27.6	15.3	15.2	27.8	14.6	21.2	36.9	21.2	20.5	36.2	19.9	19.9	31.1	21.9	19.7	31.0	21.8
Cascade Mask R-CNN	ResNet-101	24.3	36.7	28.9	24.0	36.7	28.0	43.2	60.8	47.1	42.9	60.7	47.1	23.4	32.8	26.9	23.1	32.7	26.6
HTC	ResNet-101	36.7	53.4	43.4	36.7	53.7	42.3	36.3	53.3	38.8	5.6	53.1	37.9	43.7	57.3	48.4	43.4	57.1	48.1
SCNet	ResNet-101	27.9	41.9	33.8	27.9	41.6	33.0	20.0	33.0	20.7	19.9	32.8	20.7	19.3	27.2	22.3	19.2	27.2	21.9
SOLO	ResNet-101	22.2	38.0	22.8	22.1	39.3	23.8	30.5	48.0	33.1	30.9	48.5	34.2	14.5	32.7	11.6	14.9	31.8	14.1
SOLOv2	ResNet-101	26.9	44.1	28.5	29.0	44.7	32.7	24.5	40.2	26.1	26.2	42.5	28.0	30.7	50.1	31.5	32.7	51.8	34.9
Deformable DETR	ResNet-101	24.2	33.5	28.5	23.9	33.5	28.3	29.7	43.8	32.4	29.6	43.7	32.3	34.2	49.7	38.1	34.1	49.3	38.4
QueryInst	ResNet-101	23.3	35.5	27.1	23.3	35.5	26.5	28.9	43.6	31.5	29.3	44.7	31.8	36.5	48.2	41.4	38.4	51.2	43.4
ISTR	ResNet-101	**48.6**	**63.9**	**57.3**	**48.5**	**63.9**	**56.7**	52.7	68.3	58.9	52.3	68.4	58.0	61.0	73.9	68.8	60.7	73.9	68.1
Ours	ResNet-101	44.1	60.5	50.7	43.6	60.3	49.9	**59.4**	**78.1**	**65.9**	**59.0**	**78.1**	**65.3**	**64.0**	**78.4**	**73.3**	**63.6**	**78.2**	**72.6**

Method	Backbone	**test paper**						**textbook**						**book**					
		Object Detection			Instance Segmentation			Object Detection			Instance Segmentation			Object Detection			Instance Segmentation		
		mAP	AP50	AP75	mAP	AP50	AP75	mAP	AP50	AP75	mAP	AP50	AP75	mAP	AP50	AP75	mAP	AP50	AP75
GFL	ResNet-101	44.4	65.4	52.1	43.4	65.1	49.6	38.7	57.1	44.9	37.1	57.0	42.5	8.0	15.8	6.8	6.8	15.5	4.7
FCOS	ResNet-101	37.9	59.7	43.0	36.6	59.6	39.4	33.2	52.7	38.0	31.6	52.5	34.8	10.3	23.5	6.7	7.9	20.7	4.2
FoveaBox	ResNet-101	42.8	66.8	48.9	41.4	66.8	46.3	34.5	54.2	39.7	33.2	54.1	37.3	21.6	37.6	22.3	21.3	37.5	22.1
Faster R-CNN	ResNet-101	52.0	74.1	60.6	51.1	74.2	59.8	43.6	62.7	52.7	42.4	62.6	50.2	14.5	27.2	13.5	12.0	26.3	9.2
Cascade R-CNN	ResNet-101	54.3	73.4	63.0	53.8	73.5	62.7	47.1	64.7	55.4	45.8	64.6	54.2	9.2	16.3	8.9	7.7	16.2	6.0
Mask R-CNN	ResNet-101	49.1	70.2	56.9	48.1	70.0	55.7	40.3	58.1	48.3	39.8	58.0	47.5	7.1	14.9	5.7	7.6	15.0	5.9
Cascade Mask R-CNN	ResNet-101	52.6	71.7	61.1	52.3	71.8	60.4	45.7	62.8	54.1	44.4	62.7	52.3	10.8	18.8	11.5	8.8	18.4	7.1
HTC	ResNet-101	57.9	77.7	66.9	57.2	77.6	65.9	51.2	69.6	60.3	50.5	69.6	59.1	19.6	29.6	24.1	18.8	29.5	22.4
SCNet	ResNet-101	54.8	75.5	64.2	53.9	75.3	62.6	44.6	62.7	51.8	43.6	62.5	50.3	6.6	12.2	6.3	6.7	12.2	6.3
SOLO	ResNet-101	36.2	59.1	38.5	36.2	61.1	37.9	31.8	49.7	35.1	31.1	50.3	34.8	5.8	13.9	4.0	3.2	10.0	1.1
SOLOv2	ResNet-101	33.0	55.7	33.3	34.5	57.4	36.5	33.7	54.6	36.6	35.0	55.3	38.2	17.3	29.8	17.2	15.6	29.2	16.4
Deformable DETR	ResNet-101	53.7	75.2	60.8	53.5	75.3	60.2	46.6	64.6	53.8	45.0	64.4	51.5	14.0	21.7	15.5	10.1	20.0	9.0
QueryInst	ResNet-101	44.0	60.9	50.4	43.4	61.2	49.9	35.7	50.0	41.3	35.5	50.1	41.1	10.7	17.1	11.8	10.5	17.2	11.6
ISTR	ResNet-101	60.4	80.9	**68.5**	60.1	80.9	**67.9**	50.1	68.2	58.7	49.5	68.1	57.3	**29.0**	39.9	**35.4**	**28.4**	39.8	**34.4**
Ours	ResNet-101	**60.7**	**81.9**	68.0	**60.3**	**82.2**	66.9	**51.7**	**70.1**	**60.3**	**51.2**	**70.1**	**59.5**	28.3	**41.0**	33.0	27.9	**40.7**	33.0

Conclusion

We presents a modern dataset and method for document layout analysis.

1. We have constructed and publicized the M^6Doc dataset, which is a modern document dataset that supports multiple formats, types, layouts, languages, and annotation categories.
2. M^6Doc is the first layout analysis dataset that contains both real-world (photographed and scanned) and born-digital document images. Additionally, it is the first dataset that includes Chinese documents.
3. M^6Doc contains 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.
4. We propose the TransDLANet, a Transformer-based method for DLA.