DocSAM: Unified Document Image Segmentation via Query Decomposition and Heterogeneous Mixed Learning

Supplementary Material

6. Dataset Statistics

Statistics of datasets involved in this paper are listed in Tab. 8. Datasets with underline (15 datasets) are used for ablation study, mixed pre-training and dataset specific fine-tuning, then all datasets(48 datasets) are used for training the final DocSAM model. Please note that some datasets may contain multiple subsets. These datasets cover various domains and tasks and exhibit great heterogeneity in document types, annotation formats and many other aspects. Typical examples of these datasets can be found in Fig. 1. In the following, we briefly introduce the 15 datasets used in our experiments, and for other datasets which are only used to train the final DocSAM model, we recommend the readers to read their original papers for more details.

PubLayNet [106] is a large-scale dataset for layout analysis of English scientific papers. It contains over 364,000 pages, which are divided into training, validation, and test sets containing 340,391, 11,858, and 11,983 pages, respectively. Five classes of page regions are annotated in this dataset including *text*, *title*, *list*, *table*, and *figure*. Though large-scale it is, the diversity of this dataset is limited.

DocLayNet [60] is a large-scale dataset designed for document layout analysis and understanding. It contains over 80,000 annotated pages from diverse document types, including scientific papers, reports, and forms. Each page is labeled with detailed layout information, such as text blocks, figures, tables, and captions. The dataset supports tasks like document image segmentation, object detection, and layout recognition.

D^{4}LA [13] is a diverse and detailed dataset for document layout analysis which contains 12 types of documents and defines 27 document layout categories. It contains over 11,000 annotated pages which are divided into training and validation sets containing 8,868 and 2,224 pages, respectively.

M^{6}Doc [9] is by far the most diverse dataset for document layout analysis which contains 9 types of documents and defines 74 document layout categories. It contains over 9,000 annotated pages of different languages which are divided into training, validation and test sets containing 5,448, 908 and 2,724 pages, respectively.

SCUT-CAB [8] is a large-scale dataset for layout analysis of complex ancient Chinese books. It contains 4,000 annotated images, encompassing 31,925 layout elements that vary in binding styles, fonts, and preservation conditions. To support various tasks in document layout analysis, the dataset is divided into two subsets: SCUT-CAB-Physical for physical layout analysis, with four categories, and SCUT-CAB-Logical for logical layout analysis, comprising 27 categories.

HJDataset [69] is a large dataset of historical Japanese documents with complex layouts. It contains 2,271 document image scans and over 250,000 layout element annotations of seven types. In addition to bounding boxes and masks of the con- tent regions, it also includes the hierarchical structures and reading orders for layout elements.

CASIA-HWDB [41] is a large-scale handwritten dataset for Chinese text recognition. It contains ovwe 6,000 pages which are split into training and test sets containing 4875 and 1215 pages, respectively. Since it also contains bounding boxes annotations for characters and text lines, we can use it to train our DocSAM.

SCUT-HCCDoc [95] is a large-scale handwritten Chinese dataset containing 12,253 camera-captured document images of diverse styles with 116,629 text lines and 1,155,801 characters. The dataset can used for text detection, recognition or end-to-end text spotting.

TableBank [35] is a large-scale dataset for table detection and recognition which contains over 278,000 latex or word pages for table detection and over 145,000 cropped table images for table recognition. In this paper,we only use the detection subset of TableBank since the recognition subset doesn't contain cell bounding box annotations.

PubTabNet [107] is a large-scale dataset for table structure recognition, containing over 619,000 table images. Originally designed for end-to-end table recognition, PubTabNet 2.0.0 added bounding box annotations for non-empty cells, enabling cell region detection. It provides instance annotations for two classes: *table* and *cell*. However, since the images are already cropped to focus on tables, making table detection a trivial task. Therefore, we only report results for the *cell* class.

FinTabNet [104] is a real-world and complex scientific and financial datasets with detailed annotations which can be used for both table detection and recognition. It contains table and cell bounding boxes annotations for over 76,000 pages which are divided into training, validation and test sets containing 61,801, 7,191 and 7,085 pages, respectively.

MSRA-TD500 [86] is a dataset for multi-oriented scene text detection. It contains 500 natural scene images with multi-oriented scene texts annotated with quadrilateral points, among which 300 are used for training and 200 are used for testing.

ICDAR2015 [28] incidental scene text dataset com-

Task	Dataset	#Images			#Classes	Language	Dataset	#Images			#Classes	Language
		Train	Val	Test				Train	Val	Test		
DLA	BaDLAD [71]	20,365	–	13,328†	4	Bengali	CDLA [3]	5,000	1,000	–	10	Chinese
	D^{4}LA [13]	8,868	2,224	–	27	English	DocBank [36]	40,000	5,000	5,000	13	English
	DocLayNet [60]	69,375	6,489	4,999	11	English	ICDAR2017-POD [16]	1,600	–	817	3	English
	IIIT-AR-13K [55]	9,333	1,955	2,120	5	English	M^{6}Doc [9]	5,448	908	2,724	74	Multilingual
	PubLayNet [106]	340,391	11,858	11,983	5	English	RanLayNet [2]	6,998	500	–	5	English
AHDS	CASIA-AHCDB-style1 [85]	5,854	–	1,679	2	Chinese	CASIA-AHCDB-style2 [85]	3,215	–	1,068	2	Chinese
	CHDAC-2022 [31]	2,000	–	1,000†	1	Chinese	ICDAR2019-HDRC [68]	11,715	–	1,135†	2	Chinese
	SCUT-CAB-physical [8]	3,200	–	800	4	Chinese	SCUT-CAB-logical [8]	3,200	–	800	27	Chinese
	MTHv2 [53]	2,399	–	800	2	Chinese	HJDataset [69]	1,433	307	308	7	Japanese
	CASIA-HWDB [41]	4,875	–	1,215	2	Chinese	SCUT-HCCDoc [95]	9,801	–	2,452	1	Chinese
TSR	FinTabNet [104]	61,801	7,191	7,085	2	English	PubTabNet [107]	500,777	9,115	9,138†	2	English
	ICDAR2013 [18]	–	–	156	2	English	ICDAR2017-POD [16, 38]	549	–	243	2	English
	cTDaR-modern [17, 38]	600	–	340	2	English	cTDaR-archival [17]	600	–	499	2	English
	NTable-cam [109]	11,904	3,408	1,696	1	Multilingual	NTable-gen [109]	11,984	3,424	1,712	1	Multilingual
	PubTables-1M-TD [72]	460,589	57,591	57,125	2	English	PubTables-1M-TSR [72]	758,849	94,959	93,834	6	English
	TableBank-latex [35]	187,199	7,265	5,719	1	English	TableBank-word [35]	73,383	2,735	2,281	1	English
	TNCR [1]	4,634	1,015	1,000	5	English	STDW [19]	7470	–	–	1	English
	WTW [47]	10,970	–	3,611	1	Multilingual						
STD	CASIA-10k [22]	7,000	–	3,000	1	Chinese	COCO-Text [76]	43,686	10,000	10,000†	1	English
	CTW1500 [43]	1,000	–	500	1	English	CTW-Public [92]	24,290	1,597	3,270	1	Chinese
	HUST-TR400 [87]	–	–	400	1	English	ICDAR2015 [28]	1,000	–	500	1	English
	ICDAR2017-RCTW [70]	8,034	–	4,229†	1	Chinese	ICDAR2017-MLT [56]	7200	1800	9,000†	1	Multilingual
	ICDAR2019-ArT [12]	5,603	–	4,563†	1	English	ICDAR2019-LSVT [74]	30,000	–	20,000†	1	Chinese
	ICDAR2019-MLT [57]	10,000	–	10,000†	1	Multilingual	ICDAR2019-ReCTS [100]	20,000	–	5,000†	2	Chinese
	ICDAR2023-HierText [49]	8,281	1,724	1,634†	3	English	ICDAR2023-ReST [91]	5,000	–	5,000†	1	Chinese
	ICPR2018-MTWI [21]	10,000	–	10,000†	1	Multilingual	MSRA-TD500 [86]	300	–	200	1	Multilingual
	ShopSign [94]	1265	–	–	1	Multilingual	Total-Text [11]	1,255	–	300	1	English
	USTB-SV1K [90]	500	–	500	1	English						

Table 8. Dataset statistics. Numbers with "†" means the datasets or their ground-truth annotations are not public available.

prises 1,670 images and 17,548 annotated regions, and 1,500 of the images have been made publicly available, among which 1,000 images are used for training and 500 images are used for testing. The remaining 170 images comprise a sequestered, private set.

CTW1500 [43] is a dataset for scene text detection and recognition, containing 1,500 images collected from real-world scenes. The dataset is divided into a training set with 1,000 images and a testing set with 500 images. Each image is annotated with text bounding boxes and transcriptions, making it suitable for evaluating text detection and recognition algorithms in complex scenes.

Total-Text [11] is a dataset for scene text detection and recognition, consisting of 1,255 natural scene images. The dataset is divided into a training set with 750 images and a testing set with 505 images. Each image is annotated with word-level irregular text instances, including curved and multi-oriented text, making it suitable for evaluating advanced text detection and recognition algorithms.

7. Train Details

Due to the significant differences in the size of various datasets, directly combining them to build a mixed heterogeneous dataset would lead to serious imbalance among the datasets. Training directly on such an imbalanced heterogeneous dataset would degrade the overall performance of DocSAM. Therefore, we propose a more reasonable strategy to address this issue. Specifically speaking, for each iteration during training we randomly sample B samples from all datasets to constitute a batch, with the sampling probability of each dataset proportional to $\sqrt{C_i}$, where $\sqrt{C_i}$ is the number of classes in the ith dataset. This adjusted sampling probability ensures that more complex datasets, which typically contain a greater number of classes, receive more attention during training.

Considering that some datasets may contain hundreds or even thousands of instances, such as characters, words, or cells, directly training and testing on entire images could result in low recall. To mitigate this issue, we adopt a cropped training and testing strategy. During training, we first scale the input images so that the shorter side is within the range of [704, 896] pixels, and then randomly crop them into patches of size 640×640 pixels. Alternatively, with a probability of 0.2, we resize the entire image to 640×640 pixels. During testing, we initially process the resized whole images (640×640 pixels) and then combine these results with those obtained from patches. For the patch-based approach, we first scale the entire image so that the shorter side is 800 pixels, and then crop it into patches using a sliding window method. Low-resolution whole images are used to detect larger objects or objects that span across patches, while high-resolution patches focus on smaller objects. When combining results, we reduce the confidence

Task	Dataset	Instance					Semantic	Dataset	Instance					Semantic
		AP50	AP75	mAP	mAP_b	mAF	mIoU		AP50	AP75	mAP	mAP_b	mAF	mIoU
DLA	BaDLAD [71]	0.686	0.478	0.459	0.468	0.560	0.682	CDLA [3]	0.948	0.878	0.781	0.769	0.804	0.860
	D^{4}LA [13]	0.660	0.590	0.516	0.504	0.557	0.476	DocBank [36]	0.631	0.479	0.445	0.434	0.522	0.655
	DocLayNet [60]	0.772	0.616	0.556	0.539	0.623	0.703	ICDAR2017-POD [16]	0.900	0.847	0.800	0.783	0.816	0.922
	IIIT-AR-13K [55]	0.796	0.618	0.568	0.581	0.618	0.626	M^{6}Doc [9]	0.590	0.492	0.434	0.416	0.448	0.319
	PubLayNet [106]	0.951	0.900	0.848	0.840	0.884	0.918	RanLayNet [2]	0.922	0.887	0.838	0.833	0.857	0.854
AHDS	CASIA-AHCDB-style1 [85]	0.958	0.920	0.846	0.821	0.884	0.940	CASIA-AHCDB-style2 [85]	0.951	0.918	0.813	0.799	0.864	0.913
	CHDAC-2022 [31]	0.845	0.645	0.558	0.489	0.603	0.905	ICDAR2019-HDRC [68]	0.947	0.801	0.753	0.681	0.815	0.909
	SCUT-CAB-physical [8]	0.950	0.871	0.805	0.774	0.849	0.948	SCUT-CAB-logical [8]	0.726	0.605	0.526	0.512	0.552	0.473
	MTHv2 [53]	0.928	0.804	0.677	0.657	0.703	0.913	HJDataset [69]	0.967	0.935	0.894	0.883	0.905	0.822
	CASIA-HWDB [41]	0.948	0.840	0.784	0.708	0.838	0.945	SCUT-HCCDoc [95]	0.867	0.663	0.559	0.567	0.635	0.855
TSR	FinTabNet [104]	0.885	0.809	0.718	0.698	0.799	0.870	PubTabNet [107]	0.972	0.803	0.662	0.650	0.739	0.860
	ICDAR2013 [18]	0.942	0.564	0.612	0.520	0.566	0.844	ICDAR2017-POD [16, 38]	0.941	0.854	0.764	0.735	0.799	0.897
	cTDaR-modern [17, 38]	0.919	0.575	0.646	0.601	0.706	0.878	cTDaR-archival [17]	0.897	0.717	0.672	0.627	0.691	0.956
	NTable-cam [109]	0.893	0.803	0.714	0.727	0.770	0.875	NTable-gen [109]	0.951	0.920	0.861	0.862	0.909	0.947
	PubTables-1M-TD [72]	0.968	0.915	0.829	0.797	0.855	0.931	PubTables-1M-TSR [72]	0.826	0.689	0.637	0.582	0.702	0.806
	TableBank-latex [35]	0.966	0.953	0.922	0.912	0.945	0.953	TableBank-word [35]	0.886	0.848	0.845	0.829	0.864	0.857
	TNCR [1]	0.607	0.545	0.526	0.514	0.473	0.386	STDW [19]	0.956	0.941	0.908	0.878	0.930	0.972
	WTW [47]	0.949	0.897	0.795	0.788	0.813	0.975							
STD	CASIA-10k [22]	0.652	0.408	0.386	0.385	0.428	0.807	COCO-Text [76]	0.538	0.248	0.270	0.275	0.300	0.642
	CTW1500 [43]	0.800	0.518	0.469	0.438	0.564	0.822	CTW-Public [92]	0.365	0.101	0.145	0.122	0.183	0.563
	HUST-TR400 [87]	0.850	0.746	0.632	0.601	0.682	0.863	ICDAR2015 [28]	0.688	0.302	0.340	0.346	0.381	0.630
	ICDAR2017-RCTW [70]	0.611	0.301	0.318	0.335	0.381	0.805	ICDAR2017-MLT [56]	0.685	0.476	0.427	0.425	0.477	0.840
	ICDAR2019-ArT [12]	0.761	0.480	0.442	0.457	0.496	0.799	ICDAR2019-LSVT [74]	0.630	0.384	0.368	0.370	0.423	0.816
	ICDAR2019-MLT [57]	0.721	0.510	0.456	0.454	0.508	0.851	ICDAR2019-ReCTS [100]	0.737	0.533	0.478	0.470	0.527	0.846
	ICDAR2023-HierText [49]	0.558	0.287	0.293	0.282	0.335	0.669	ICDAR2023-ReST [91]	0.949	0.870	0.743	0.825	0.774	0.827
	ICPR2018-MTWI [21]	0.649	0.390	0.380	0.384	0.445	0.843	MSRA-TD500 [86]	0.832	0.617	0.532	0.570	0.574	0.763
	ShopSign [94]	0.666	0.272	0.320	0.332	0.392	0.814	Total-Text [11]	0.783	0.483	0.443	0.456	0.493	0.782
	USTB-SV1K [90]	0.839	0.428	0.450	0.442	0.492	0.718							

Table 9. Performance of DocSAM on heterogeneous datasets and tasks.

scores of objects detected near the boundaries of patches, as these detections are more likely to be fragmented. Finally, after combining the results, we apply non-maxima suppression to eliminate duplicate predictions arising from different patches and whole images.

8. Additional Results

We train the final DocSAM model using Swin-Large [45] as the vision backbone on all 48 datasets listed in Tab. 8 and report the testing results of DocSAM on these datasets in Tab. 9. If the ground-truth annotations for the test set or validation set of a specific dataset are publicly available, we test and report the results of DocSAM on the standard test set or validation set. Otherwise, we randomly split the original training set into a new training set and a validation set at a ratio of 9:1 and use these new sets for training and evaluation. Please note that this is intended to provide an intuitive sense of DocSAM's performance on these datasets and is not suitable for direct comparison with the results of other works.

From Tab. 9, we can see that as a single all-in-one model, DocSAM provides fairly good results across all datasets with various tasks and heterogeneous document types, despite variations in performance due to differing levels of difficulty. This demonstrates the superiority and effectiveness of DocSAM. As a single-modal model, DocSAM may underperform on datasets like D^{4}LA [13], DocLayNet [60], M^{6}Doc [9], and SCUT-CAB-Logical [8], which often contain more classes and require multi-modal information for fine-grained logical layout analysis. This is also indirectly verified by the relatively low performance of semantic segmentation on these datasets. Additionally, DocSAM achieved lower performance on scene text detection datasets, likely due to the greater diversity in shapes and backgrounds of scene texts, which require more carefully designed strategies to ensure model performance. Despite these challenges, DocSAM is quite successful in achieving its goal of being a simple and unified document segmentation model applicable to a wide variety of datasets and tasks. It shows decent performance across various datasets and tasks and holds great potential for downstream applications, both as a versatile segmenter and as a pre-trained model. We believe that DocSAM can greatly benefit from more sophisticated model design and better data augmentation and training strategies to further accelerate its convergence and improve its performance.

9. Qualitative results

Finally, we present some qualitative results of DocSAM on representative datasets and tasks in Fig. 4, Fig. 5, Fig. 6, and Fig. 7. From these figures, it is evident that DocSAM produces reliable predictions across a wide range of datasets and tasks, including modern and historical document layout analysis, table structure decomposition, handwritten text

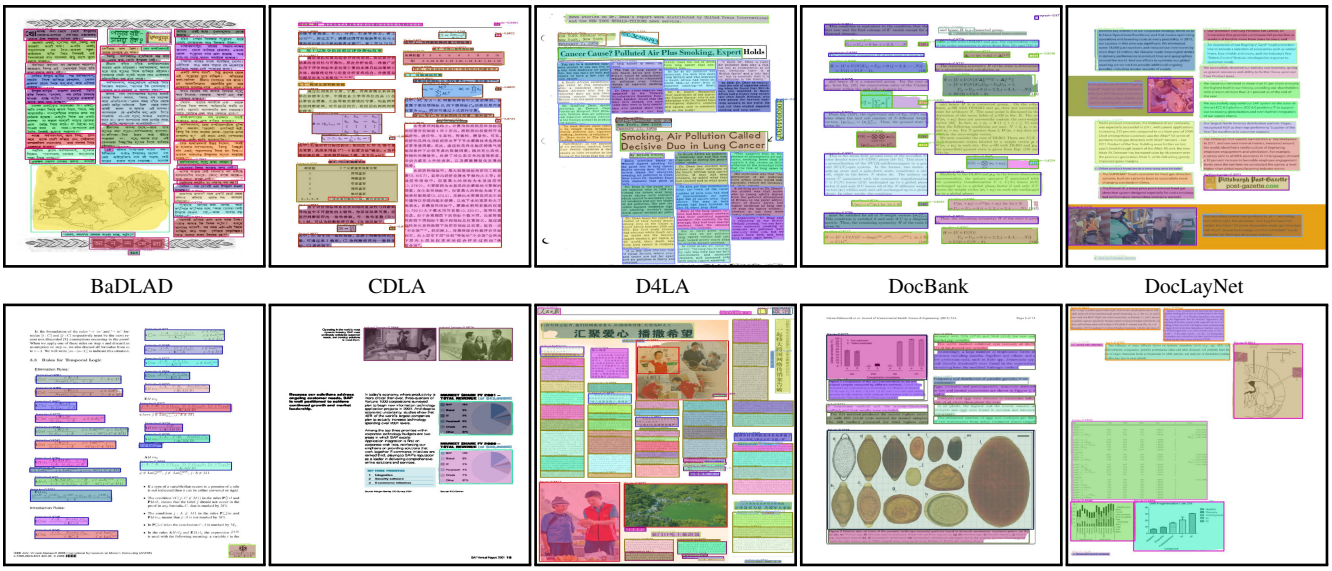

Figure 4. Qualitative results on public document layout analysis benchmarks produced by our DocSAM model.

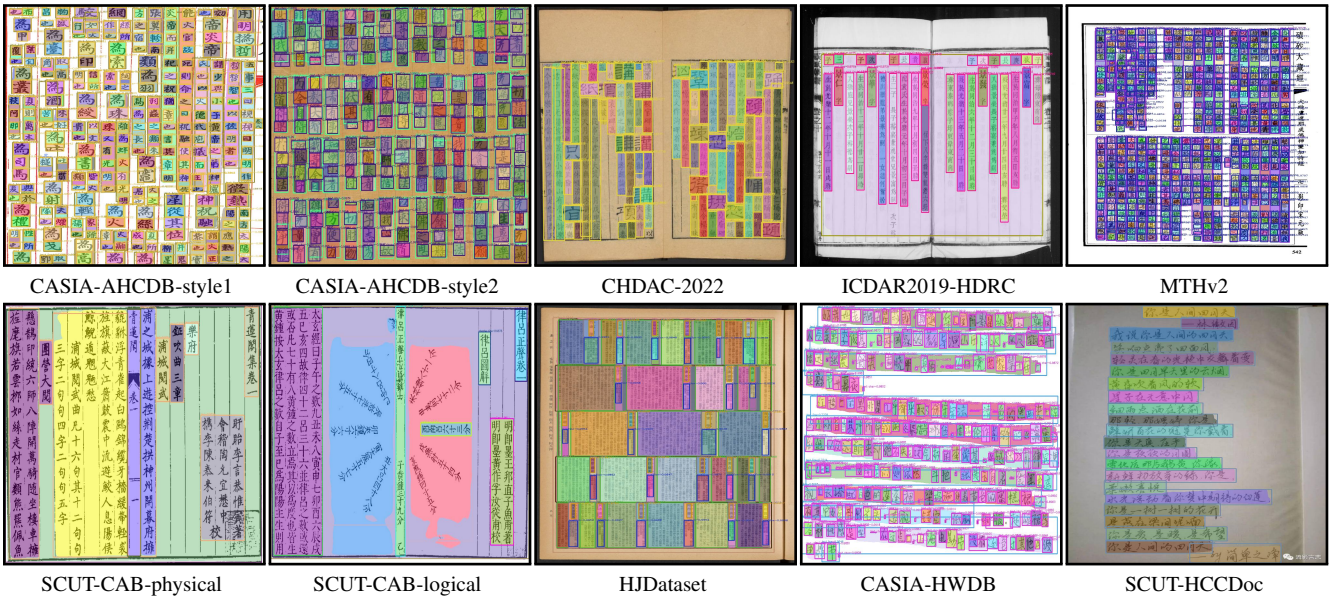

Figure 5. Qualitative results on public ancient and handwritten document segmentation benchmarks produced by our DocSAM model.

detection, scene text detection, and more. Specifically, DocSAM demonstrates robust performance in modern and historical document layout analysis, where it accurately identifies and segments various elements such as figures, tables, and text blocks. In table structure decomposition, DocSAM effectively recognizes and separates table cells, even in complex layouts with dense rows and columns. For handwritten text detection, the model successfully identifies and localizes individual characters and lines, even in challenging scripts and varying handwriting styles. Additionally, in scene text detection, DocSAM shows strong capabilities in detecting text in real-world images, handling diverse scenarios such as curved and multilingual texts. These results underscore the versatility and effectiveness of DocSAM across a wide range of document processing tasks, highlighting its potential for practical applications in various domains.

We also highlight some failure cases in Fig. 8. Typical failure cases for document layout analysis primarily involve over-segmentation, which is often due to annotation

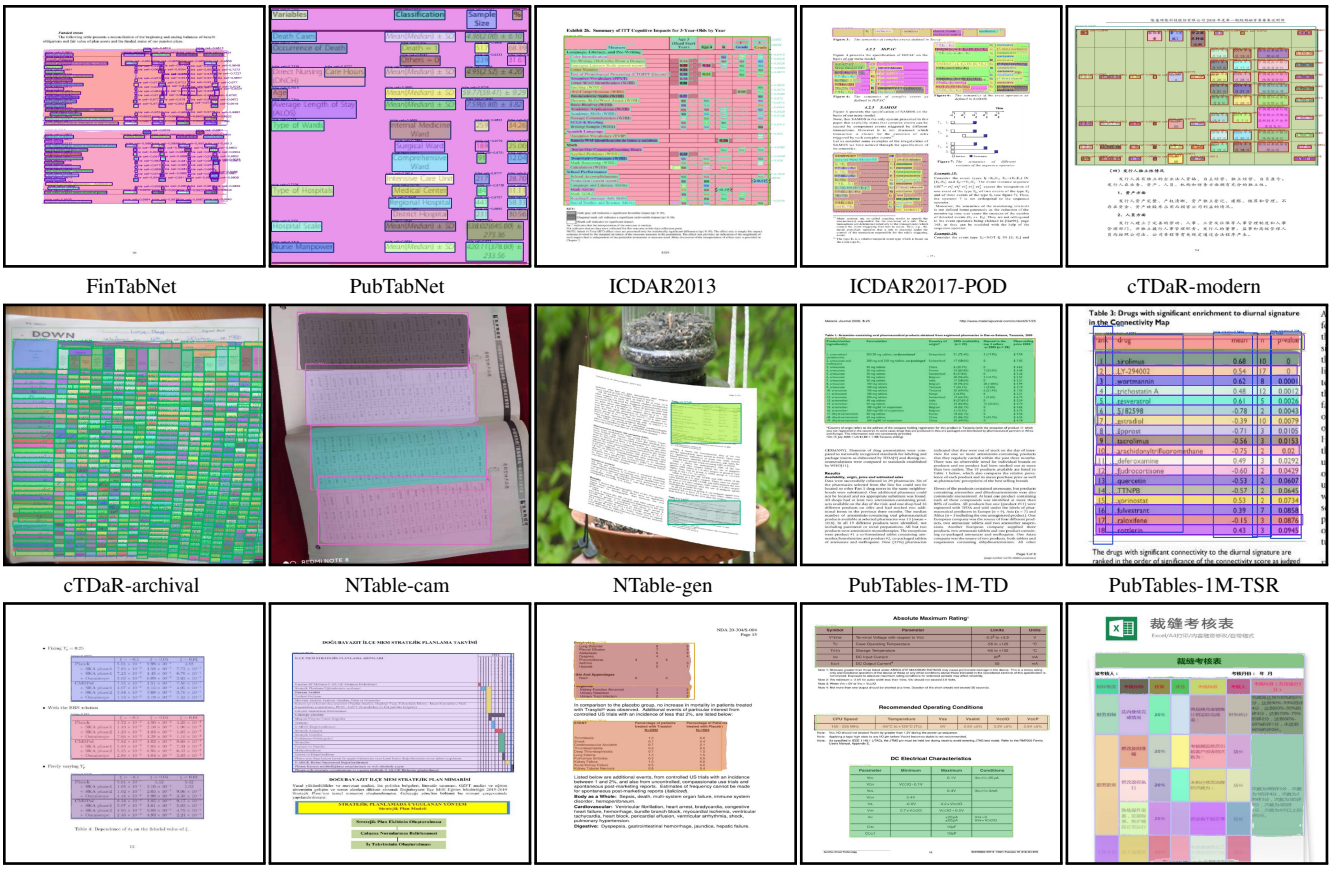

Figure 6. Qualitative results on public table detection and structure recognition benchmarks produced by our DocSAM model.

ambiguity across different datasets. Over-segmentation is also particularly common in large table cells that contain numerous lines and paragraphs. Another frequent issue in layout analysis and table structure recognition is the imprecise prediction of bounding boxes for dense and curved text lines and cells. For scene text detection, typical failure cases mainly involve dense, curved, blurred, tiny, and occluded texts. These challenging scenarios can significantly impact the accuracy of the model, highlighting areas where further improvements are needed. By identifying these failure cases, we can better understand the limitations of DocSAM and guide future research and development efforts to enhance its performance in these challenging scenarios.

Figure 7. Qualitative results on public scene text detection benchmarks produced by our DocSAM model.

Figure 8. Failure cases produced by our DocSAM model. “GT” means ground-truth and “DT” means detection results.