GloSAT Historical Measurement Table Dataset

Enhanced Table Structure Recognition Annotation for Downstream Historical Data Rescue

Juliusz Ziomek
School of Electronics and Computer Science
University of Southampton
Southampton, UK
jkz1g18@soton.ac.uk

Stuart E. Middleton
School of Electronics and Computer Science
University of Southampton
Southampton, UK
sem03@soton.ac.uk

ABSTRACT

Understanding and extracting tables from documents is a research problem that has been studied for decades. Table structure recognition is the labelling of components within a detected table, which can be detected automatically or manually provided. This paper presents the GloSAT historical measurement table dataset designed to train table structure recognition models for use in downstream historical data rescue applications. The dataset contains 500 scanned and manually annotated images of pages from meteorological measurement logbooks. We enhance standard full table and individual cell annotations by adding additional annotations for headings, headers, and table bodies. We also provide annotations for coarse segmentation cells consisting of multiple data cells logically grouped by ruling lines of ink or whitespace in the table, which often represent data cells that are semantically grouped. Our dataset annotations are provided in VOC2007 and ICDAR-2019 Competition on Table Detection and Recognition (cTDaR-19) XML formats, and our dataset can easily be aggregated with the cTDaR-19 dataset. We report results running a series of benchmark algorithms on our new dataset, concluding that post-processing is very important for performance, and that page style is not as significant a feature as table type on model performance.

CCS CONCEPTS

•Computing methodologies~Artificial intelligence~Natural language processing~Information extraction•Applied computing~Document management and text processing•Computing methodologies~Machine learning~Machine learning approaches~Neural networks

KEYWORDS

Document Layout Analysis, Table Structure Recognition, Image Processing, Deep Learning, Historical Documents, Measurements

ACM Reference format:

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

1 Introduction

Understanding and extracting tables from scanned document images is a research problem that has been studied for decades. Developments in deep learning techniques, coupled with the availability of larger annotated corpora has seen much progress in recent years and a shift away from heuristic-based approaches. *Table detection* is the detection of tables within an image of a document page, typically assigning bounding boxes around each whole table element. *Table structure recognition* is the labelling of components within a detected table.

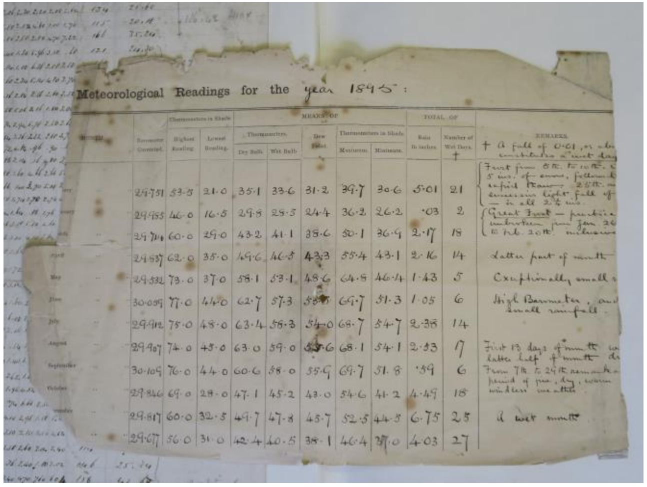


Figure 1: Example of a handwritten scanned logbook page of a measurement table from the GloSAT dataset. Scanned images can be of damaged pages, exhibit small skew and have sub-optimal backgrounds depending on the scanning conditions when image was taken.

An example of a downstream application for table structure recognition is data rescue, such as data rescue of measurement tables from historical ship logbooks or land station records (see Fig. 1). The GloSAT project[1] has access to 100,000's of scanned pages from measurement logbooks dating from the 1700's to the modern day. Data rescue from modern electronic documents in

[1] www.glosat.org

formats such as PDF is a relatively simple parsing process. However, data rescue from printed or handwritten historical documents needs table structure recognition applied to scanned page images to identify not just the full table cell regions, but also contextually relevant parts of the table and document (see Fig. 2) and coarse segmentation cells made up of semantically associated groups of individual cells indicated from ruling lines of ink or whitespace within tables (see Fig. 3).

Most current table structure recognition datasets with scanned historical documents focus on annotating full table and cell regions and do not have annotation labels for table headers or relevant elements outside of the table region such as page headers. These datasets are usually hand annotated and small in size as a result. The early table detection datasets UNLV and UW-3 [12] annotated scanned pages from news, magazines, and business reports and had 427 and 120 annotated documents respectively. More recent datasets from the International Conference on Document Analysis and Recognition (ICDAR) series competition on Table Detection and Recognition (cTDaR-19) [4] have 799 historical and 840 modern images, with a mixture of printed and handwritten pages in both English and Chinese.

Datasets with modern electronic documents, although not as useful for historical data rescue applications, are much larger in size and have richer annotations than just table and cell regions as they can extract this automatically from metadata in document formats such as PDF. The PubLayNet dataset [13] has 340,000 annotated image scans from pages of medical PDF's, containing annotations for text, title, list, table and figures. Another example is the TableBank dataset [7] with 417,000 images sourced from MS Word documents and latex documents.

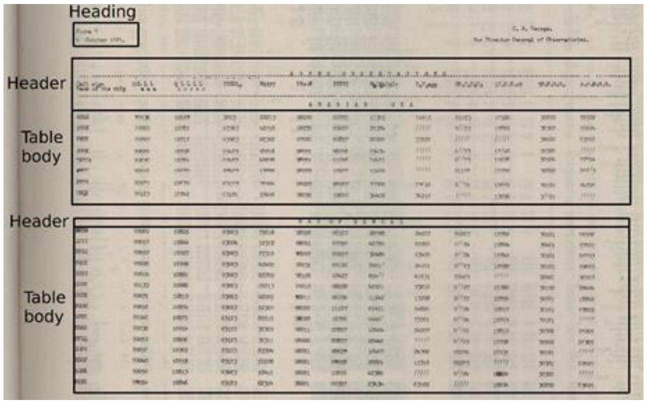
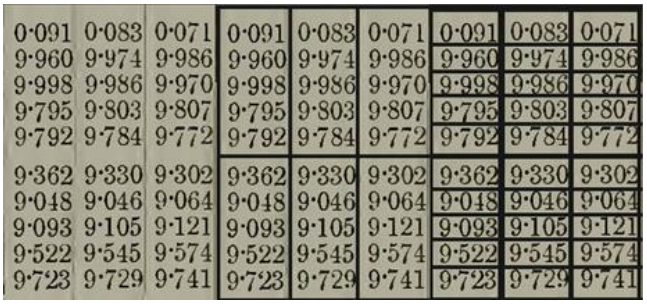

Heading
Header
Table body
Header
Table body

Figure 2: Example of GloSAT enhanced annotations for two tables with a heading, header and table_body. The full table annotation is the aggregation of header(s) and table_body.

Modern data driven approaches to table structure recognition include Faster Region-based Convolutional Neural Networks (F-RCNN) such as DeepDeSRT [10], and F-RCNN approaches [2]. Encoder-decoder architectures have also been used such as a Fully Convolutional Network (FCN) [8] and a hybrid FCN and Conditional Random Field (CRF) model [5]. A multistage extension of Mask R-CNN is used by CDeC-Net [1], and one of the latest Region-based Convolutional Neural Network (R-CNN) approaches is CascadeTabNet [9], which is a variation of a R-CNN using a cascade mask Region-based CNN.



0·091	0·083	0·071	0·091	0·083	0·071	0·091	0·083	0·071
9·960	9·974	9·986	9·960	9·974	9·986	9·960	9·974	9·986
9·998	9·986	9·970	9·998	9·986	9·970	9·998	9·986	9·970
9·795	9·803	9·807	9·795	9·803	9·807	9·795	9·803	9·807
9·792	9·784	9·772	9·792	9·784	9·772	9·792	9·784	9·772
9·362	9·330	9·302	9·362	9·330	9·302	9·362	9·330	9·302
9·048	9·046	9·064	9·048	9·046	9·064	9·048	9·046	9·064
9·093	9·105	9·121	9·093	9·105	9·121	9·093	9·105	9·121
9·522	9·545	9·574	9·522	9·545	9·574	9·522	9·545	9·574
9·723	9·729	9·741	9·723	9·729	9·741	9·723	9·729	9·741

Figure 3: Example of individual cell annotation (right) and coarse segmentation cell annotation (middle) in the GloSAT dataset. The original image without annotation is (left). Coarse segmentation closely follows ruling lines of ink or whitespace within the table that group related cells together.

This paper describes the GloSAT historical measurement table dataset of 500 images sampled from a dataset of 100,000's of scanned physical measurement log books using a maximum variation sampling strategy. This dataset uses cTDaR-19 formatted XML annotations for full table and individual cells, but also includes new enhanced annotation types for page headings, table headers, table bodies and coarse segmentation cells. We evaluate table structure recognition using a benchmark algorithm are based on the successful CascadeTabNet [9] model which uses a Region CNN model architecture. We also experiment with a novel post-processing technique which can boost cell detection performance using a DBScan extrapolation of missing cells. We release as open source our dataset, train/test splits, benchmark code and model checkpoints for future researchers to use.

The novel contributions of this paper are (a) a new dataset for table structure recognition with enhanced annotation of contextual table elements and (b) results running benchmark algorithms including a novel post-processing method.

3 Dataset

The GloSAT dataset contains 500 images sampled from nine different sources of historical archive and represents a mixture of land station measurement journals and ship logs over the last few hundred years of meteorological record. These are summarized in Table 1. The dataset is released via Github archive[2] under a BSD open source license. We worked with climate scientists from the UK Met Office, US National Oceanic and Atmospheric Administration (NOAA) and University of Reading to help ensure we had a representative selection of measurement logbook archives to work from. Dataset images were sampled from the 100,000's of scanned document pages available using a maximum

[2] https://github.com/stuartemiddleton/glosat_table_dataset

variation sampling strategy, capturing as many different page styles (printed, handwritten or mixed), and measurement table formats (borderless using whitespace, semi-bordered and bordered using ink) and historical periods as possible. A table is considered bordered when every cell is separated from others by a ruling line. If some, but not all cells are separated or if all cells are separated but not from every side, then the table is considered semi-bordered. There can be multiple tables per page and tables can appear alongside with other text, maps and figures on the page.

Source ID; **Region, Timeframe**	**Images / Tables / Headers**	**Page Style;** **Table Style**
20cr_DWR_MO; India, 1970's	24 / 31 / 31	Printed; Borderless
20cr_DWR_NOAA; India, 1930's	24 / 24 / 24	Printed; Semi-bordered
20cr_Kubota; Philippines, 1900's	24 / 28 / 28	Printed; Semi-bordered
20cr_Natal_Witness Africa, 1870's	26 / 26 / 26	Printed; Semi-bordered
Ben Nevis UK, 1890's	97 / 137 / 82	Printed; Semi-bordered
DWR UK and world 1900's	93 / 139 / 139	Mixed; Semi-bordered
WesTech Rodgers Arctic 1880's	82 / 164 / 82	Mixed; Semi-bordered
WR_10_years UK, 1830's to 1930's	97 / 129 / 129	Mixed; Bordered
WR_Devon_Extern UK, 1890's to 1940's	33 / 33/ 33	Mixed; Bordered
Total	500 / 710 / 573	

Table 1: Summary of GloSAT historical measurement table dataset sources, images and page/table styles.

There are two available versions of the dataset, one with individual cell annotations and one with coarse segmentation cell annotations (see Fig. 2). For comparison, the cTDaR19 [4] dataset would be considered an individual cell dataset under our definition. Coarse segmentation cell annotation follows the original table row ruling lines closely, grouping multiple rows of values into a single cell if the original document groups them in this way. This is useful for downstream applications needing to preserve the spatial grouping of cells within tables and their associated semantic meaning (e.g. groups of weather station ID's cell entries from the same local region grouped by whitespace).

The GloSAT dataset follows the standard ICDAR annotation XML schema [4], extended with annotations for cells belong to headers, page type and table style (see Fig. 1). ICDAR annotations are limited to table components within the full table region only. Besides ICDAR annotations, we also provide VOC2007 format annotations [3], allowing easier integration with many deep learning frameworks. The VOC2007 format allows us to define an additional heading class, which is used to mark relevant information about the table, such as a caption or date and time mark for the table, which itself is not a part of the table. These heading and header class labels could be ignored for ICDAR-style table structure recognition competitions but are useful as they provide important semantic information for downstream applications.

Manual annotation of table and cells regions in each image was performed using a combination of TTruth[3] and Transkribus[4] tools. Individual grids of cells were first drawn as uniform rectangular grids using the fast and simple TTruth tool. These were then loaded into the slower but more powerful Transkribus tool, so that the annotations could be visualized, and row/column/cell region boundaries manually adjusted to provide a best fit for each image as well as adding table class labels to each region. The final annotations were exported in VOC2007 format and converted into the ICDAR annotation XML schema using a Python script.

4 Evaluation of Table Structure Recognition

We explored how table structure recognition algorithms perform using the GloSAT dataset, cTDaR19 dataset (both track B1 and B2 datasets) and an aggregation of them all. The cTDaR19 dataset track B1 is where the full table region is provided in the metadata and only cell detection is needed. The cTDaR19 dataset track B2 is where the full table region must be detected automatically. We used 75% of the dataset for training and 25% for testing, and all train/test splits are available from our dataset Github site. We trained our models using a Nvidia Tesla V100 enterprise GPU with 16GB of memory.

Our first benchmark algorithm was the CascadeTabNet [7] model (*CascadeTabNet model*). This model is representative of the latest class of Cascade Region CNN [6] model, which is a refined version of standard Faster Region CNN, and has been successfully applied to table structure recognition on the cTDaR19 dataset. As in a standard Region CNN, a “backbone” network is first used by the model to extract features from the input image and then features corresponding to regions in the image are separately fed as an input to a region proposal network, which can reject or accept and refine given region. Such refined regions are then fed as an input to a “head” fully connected network, which again can either reject or refine it. In Cascade RCNN there are multiple consecutive “heads” (see Fig. 4) and each of them can additionally refine or reject a region.

Following the same method as the CascadeTabNet authors, we add a cell class and re-train the model on our datasets. We found training the model to detect full table and cell classes at the same time extremely difficult, and through experimentation we found the best performing training strategy was to have two separate models (one for full table and one for cell) and combine the results at the end. The full table trained model was used for automated table region detection, the cell trained model for table

[3] http://www.iapr-tc11.org/dataset/TableGT_UW3_UNLV/t-truth.tar.gz
[4] https://readcoop.eu/transkribus/

structure recognition. We think this was due to class instance imbalance in the datasets, with the GloSAT dataset for example, having 35,555 instances of individual cells and 710 instances of full table. Even when using two separate models the detection of cells was weak (weighted avg. F1 score as low as 0.047 for GloSAT dataset). Our model had a high precision on cells (0.92 on 0.6 IoU level) but very low recall (0.042 on 0.6 IoU level). The CascadeTabNet authors [4] state that "high-end post-processing can improve the results significantly", so we developed our own simple but novel post-processing method which we describe next.

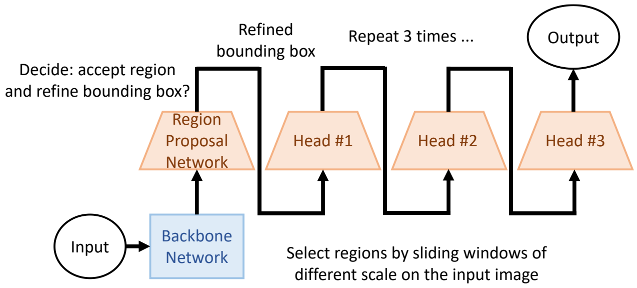

Refined bounding box
Repeat 3 times ...
Output
Decide: accept region and refine bounding box?
Region Proposal Network
Head #1
Head #2
Head #3
Input
Backbone Network
Select regions by sliding windows of different scale on the input image

Figure 4: Information Flow through the Cascade Region CNN model

Our second benchmark algorithm (*CascadeTabNet + post-processing model*) employs a post-processing step after CascadeTabNet. Since cells are ordered into rows and columns, if one knows the position of a subset of cells, the position of the remaining ones can be found by extrapolation with the assumption that the table is rectangular (which it almost always is in the datasets we have used). Therefore, to find all the cells in a table we only need to know one cell for every horizontal and vertical ruling line. We extrapolate vertical and horizontal ruling lines for a table region by applying a one-dimensional DBScan clustering algorithm [11], where each cell provides two points (their start and end) for this clustering procedure in each of the two dimensions. The advantage of DBScan is that it does not require us to specify the expected number of clusters, but rather the *minPts* and *eps* parameters. The former is used to determine minimum number of points that should lie in a cluster, and the latter is the distance threshold for points, below which they are considered members of the same cluster. The value of *eps* is heuristically set to half the average of the cell width/height for the horizontal/vertical clustering and the value of *minPts* is set to 2.

Where datasets have both coarse segmentation cells and individual cells (i.e. the GloSAT dataset) we use results from two models trained on each separately as input to this post-processing method. We found for tables where multiple rows were grouped together, as opposed to each row being clearly defined by a ruling line, the coarse segmentation model results were superior. Therefore, providing results from both models (individual cells and coarse segmentation cells) as input to the post-processing algorithm makes sense, and delivered better results. The cTDaR19 (track B1 and B2) datasets do not have coarse annotations so only results from the individual cell annotation model is used.

We trained our models using the Adam optimizer with a learning rate set to 0.0001 for table structure recognition and 0.0003 for table detection. The models were trained until the loss stabilized, which usually took between 60 to 100 epochs. The learning rate was divided by three after reaching 50 epochs. To help reduce memory footprint during training the backbone and region proposal network were frozen and only the final heads of CascadeTabNet were trained.

Dataset / *Model*	**Automated Table Detection**	**Weighted Average F1 Score**	**Row F1 Score**	**Col F1 Score**
GloSAT (coarse segmentation cells), 129 test, 371 train				
CascadeTabNet	No	0.385	-	-
	Yes	0.385	-	-
CascadeTabNet + Postprocesssing	No	0.602	0.88	0.95
	Yes	0.578	0.87	0.94
GloSAT (individual cells), 129 test, 371 train				
CascadeTabNet	No	0.047	-	-
	Yes	0.047	-	-
CascadeTabNet + Postprocesssing	No	0.284	0.46	0.91
	Yes	0.263	0.45	0.91
cTDaR19 (track B1), 150 test, 600 train				
CascadeTabNet	No	0.084	-	-
	Yes	0.084	-	-
CascadeTabNet + Postprocesssing	No	0.161	0.42	0.65
	Yes	0.156	0.41	0.63
cTDaR19 (track B2), 250 test, 600 train				
CascadeTabNet	No	0.082	-	-
	Yes	0.073	-	-
CascadeTabNet + Postprocessing	No	0.155	0.40	0.55
	Yes	0.129	0.32	0.50
Aggregated Dataset (GloSAT + B1 + B2), 529 test, 1,571 train				
CascadeTabNet	No	0.071	-	-
	Yes	0.071	-	-
CascadeTabNet + Postprocessing	No	0.181	0.40	0.63
	Yes	0.170	0.39	0.61

Table 2: Results for table structure recognition. The CascadeTabNet model does not explicitly return ruling line information, so we cannot compute row/column F1 for it without post-processing. We show results with and without automated table region detection from the full table trained model.

We used two metrics to evaluate model performance on our datasets. The first was weighted average F1 score used by the cTDaR-19 table detection competition, which is a weighted average of F1 scores at different IoU levels of {0.6, 0.7, 0.8, 0.9}, where the IoU level is the weight (equation 1). We use this metric to compare true and predicted bounding boxes for each cell regardless of the position of others or if they are blank or not. We do not use the cTDaR-19 track B competition special rules for cell adjacency and blank cells as we wanted to compare results for table and cells using an identical metric. We also present results for a novel metric which we call a row/col F1 score. The row/col F1 score evaluates the F1 score directly on ruling lines rather than cells, allowing us to explore a model's ability to identify table structural elements directly. A ruling line matches the ground truth if it is at most d pixels away from it in either direction (we set d to be 20% of the average width/height of table's cells). Row/col metrics can be better than cell IoU metrics for cell structure recognition. This is true for applications where ruling line accuracy (row or col) is paramount, such as applications doing optical character recognition that want to avoid clipping cells.

$$\frac{\sum_{i=1}^{4} F1@IoU_i * IoU_i}{\sum_{i=1}^{4} IoU_i} \tag{1}$$

The results can be seen in Table 2. We also ran additional experiments for the best performing model (*CascadeTabNet + post-processing model*) breaking down performance on the GloSAT dataset broken down by page style and table type. These results are shown in Table 3.

Table type	**Weighted Average F1 Score for cells**
Bordered	0.79
Semi-bordered	0.31
Borderless	0.023
Page style	**Weighted Average F1 Score for cells**
Printed	0.26
Mixed	0.42

Table 3: Results for the best performing CascadeTabNet + postprocessing model on the GloSAT (individual cells) dataset broken down by table style and page type

6 Discussion

From our results for table structure recognition results it is clear there is value in the post-processing method (e.g. post-processing boosted F1 score from 0.047 to 0.26 for the GloSAT individual cell dataset). It should be noted that the original CascadeTabNet authors used a pipeline which contained post-processing for both borderless and bordered tables, which is different from our post-processing method. However, the clear benefits of post-processing in our results shows the value of encoding regular structural knowledge about tables obvious to humans but not to a deep learning algorithm (at least not without many more examples).

When results are broken down by table type, it becomes clear that bordered tables are easiest to segment than the borderless ones. Although unsurprising, it does suggest that more work is needed focussing on borderless tables and tables where the border is faded or damaged, which is fairly common in historical documents.

The breakdown of results for different page styles show that F1 scores are higher for mixed documents, which is counter intuitive as a handwritten table will be less well formatted than a fully printed one. We think this result is due to a correlation in the GloSAT dataset between mixed documents and bordered tables, since a lot of documents contain both a printed border and handwritten values. This type of mixed document is very common in historical measurement logbooks. The fully printed pages were more likely not to bother with a printed border as the regular printed text was neatly aligned into borderless columns. We conclude that page style is not as significant a feature as table type on the performance of our models.

In conclusion, our paper has presented a new dataset for table structure recognition, enhancing the standard full table and cell annotations with annotations for headings, headers, table bodies and annotations for coarse segmentation cells consisting of groups of data cells logically grouped in a table. We expect that these enhanced annotations will be useful for downstream applications that need to interpret semantic associations between table components, such as data rescue pipelines that take spatial table structure components and convert them into a useful downstream dataset of semantically grounded measurements. There is room for improvement in our benchmark models, such as exploring end to end models to remove the need for post-processing, but our benchmark results provide a solid baseline for any future researchers using this dataset.

ACKNOWLEDGMENTS

This work was supported by the Natural Environment Research Council (NE/S015604/1). The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

REFERENCES

[1] Agarwal, M. Mondal, A. Jawahar, C.V. CDeC-Net: Composite Deformable Cascade Network for Table Detection in Document Images. arXiv:2008.10831v1 (2020)

[2] Arif, A. Shafait, F. Table Detection in Document Images using Foreground and Background Features. In Digital Image Computing: Techniques and Applications (DICTA), pp. 1-8. Canberra, Australia (2018)

[3] Everingham, M. Gool, L. Williams, C.K. Winn, J. Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. Int. J. Comput. Vision 88, 2 (June 2010), 303–338. (2010)

[4] Gao, L. Huang, Y. Déjean, H. Meunier, J. Yan, Q. Fang, Y. Kleber, F. Lang, E. ICDAR 2019 Competition on Table Detection and Recognition (cTDaR). In International Conference on Document Analysis and Recognition (ICDAR), pp. 1510-1515. Sydney, Australia (2019)

[5] He, D. Cohen, S. Price, B. Kifer, D. Giles, C.L. Multi-scale multi-task fcn for semantic page segmentation and table detection. In Document Analysis and Recognition (ICDAR), 2017 14th IAPR, International Conference on, vol. 1, pp. 254–261, IEEE. (2017)

[6] Koci, E. Thiele, M. Rehak, J. Romero, O. Lehner, W. DECO: A Dataset of Annotated Spreadsheets for Layout and Table Recognition. In International Conference on Document Analysis and Recognition (ICDAR), pp. 1280-1285. Sydney, Australia (2019)

[7] Li, M. Cui, L. Huang, S. Wei, F. Zhou, M. Li, Z. TableBank: Table Benchmark for Image-based Table Detection and Recognition. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC), ACL. Marseille, France (2020)

[8] Paliwal,S. D, V. Rahul, R. Sharma, M. Vig, L. TableNet: Deep Learning model for end-to-end Table detection and Tabular data extraction from Scanned Document Images. arXiv:2001.01469v1 (2020)

[9] Prasad, D. Gadpal, A. Kapadni, K. Visave, M. Sultanpure, K. CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents. arXiv:2004.12629v2 (2020)

[10] Schreiber, S. Agne, S. Wolf, I. Dengel, A. Ahmed, S. DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images. In 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 1162-1167. Kyoto (2017)

[11] Schubert, E. Sander, J. Ester, M. Kriegel, H.P. Xu, X. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 42(3), 19. (2017)

[12] Shahab, A. Shafait, F. Kieninger, T. Dengel, A. An open approach towards the benchmarking of table structure recognition systems. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems (DAS '10), pp. 113–120. Association for Computing Machinery, New York, NY, USA (2010)

[13] Zhong, X. Tang, J. Yepes, A.J. PubLayNet: largest dataset ever for document layout analysis. arXiv:1908.07836v1 (2019)