Progress Report Ph.D.
Curriculum Smart Computing

Andrea Gemelli

September 2022

Contents

1 Document Layout Analysis **2**

2 First Year - 2020/2021 **4**
2.1 Achievements 4
2.2 Plans and Future Work 5
2.3 Conferences and Summer Schools 5

3 Second Year - 2021/2022 **7**
3.1 Achievements 7
3.2 Plans and Future Work 8
3.3 Conferences and Summer Schools 8

4 Third Year - 2022/2023 **9**
4.1 Achievements 9
4.2 Conferences and Summer Schools 9

5 Visiting Research Period **10**
5.1 Other Activities 11

6 Publications **13**

7 Credits **14**

UNIVERSITÀ
DEGLI STUDI
FIRENZE

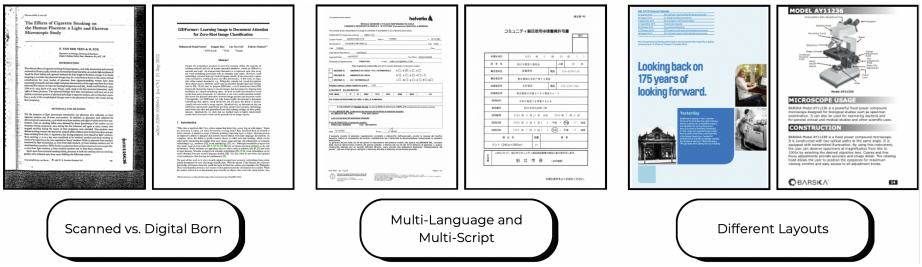

Figure 1: Document AI challenges. (images taken from RVL-CDIP, XFUND and DocLayNet. (cite second paper shown)

1 Document Layout Analysis

The research interest about documents has increased significantly in recent years: for industries, it has become essential to speed up processes that otherwise would cost a lot of resources. For academics, develop document understanding frameworks, it means solving several challenging tasks, including vision, language and geometry. In a recent survey [5], the term Document AI (DAI) has been used to refer at the process of automatic understanding, classifying and extracting information from different kinds documents exploiting artificial intelligence. DAI it is an hard task to solve, due to high variability in several aspects (summarized in Fig. 1): (i) data quality, e.g. scanned vs. digital born documents or training vs. in-the-wild samples; (ii) contents, due to different languages and/or scripts; (iii) different document layouts, e.g. magazines, scientific papers and invoices.

Among others, one of the first and most important task regarding documents has been Document Layout Analysis (DLA): it aims at automatically finding region of interest, such as text or figures, and, if needed, recognizing and classifying them, e.g. discriminating two blocks of text into title or paragraph. Physically speaking, the objective is to identify homogeneous contents (regions) boundaries in terms of coordinates, for the majority bounding boxes, through different page layouts, such as rectangular, Manhattan,, non-Manhattan, Multi-column Manhattan, Arbitrary Complex and overlapping (horizontally and diagonally) [16, 2]. During the years several methods have been proposed trying to solve DLA, following the application of novel techniques and the gathering of larger collections of annotated data. Raging from the early '90s up to nowadays, it is possible to broadly divide the different techniques into three groups: heuristics, statistical machine learning and deep learning based methods. Marinai S. largely describe the first two groups in [21], dividing the different approaches depending on two criteria. The first one refer to *how* the document is analyzed, either using bottom-up, top-down or hybrid techniques. Bottom-up techniques start gathering information at pixel-level, then iteratively grouping them into larger areas, from connected components (CCs) up to larger meaningful areas of text or non-text (e.g. figures). Representative algorithms from this group are

RLSA [33], Docstrum [23] and Voronoi diagrams [17]. On the contrary, top-down techniques start from the whole document until basic components are found in subsequent steps, line in the X-Y cut algorithm [22]. Finally, hybrid methods are compositions of the aforementioned ones. The second categorization criteria discriminates based on *what* to analyze, either the physical or logical document layout. The first one aims at the identification of homogeneous region in the page while the latter at assigning a functional information, a label, to these regions. Methods are categorized on these terms depending on the downstream task they are used for: Strouthopoulos and Papamarkos [30] propose an ANN to classify 8x8 document regions text, graph or halftones; Wu et al. [34] segment text regions using a series of split-or-merge operations guided by a binary SVM classifier. For further details and insights over other methods revised by the author, refer to [21]. Once the page objects are segmented and / or classified, some post-processing techniques could be consider to generalize the results over different layouts [2]

More recently, deep learning techniques have been started to be used also for DLA, also thanks to larger available document collections. In [5] the most important ones are listed and divided into three broad categories: Convolutional Neural Networks, Graph Neural Neural Networks. Faster R-CNN [26] and Mask R-CNN [11] based architectures have been widely use across several benchmarks to detect different page objects, while LayoutLM [35] has been the first multi-model architecture transformer-based applied to DIAR. Even if graph have been mainly use for information extraction, Gemelli et al. [**gemelli2022**] proposed a GNN to tackle DLA and Table Understanding at the same time. In the last competition of ICDAR2021 [14], Zhang P. et al. [37] achieved the SOTA on the Document Layout Recognition track, proposing a multi-modal Mask-RCNN-based object detection framework that make use of vision, language and geometry.

It is possible to observe that the majority of methods used so far for DLA strongly rely on supervised learning. This is why the amount of labelled data has been always at the same time an important and complicated point to address in DIAR. There are two main problems related to annotated collections of documents: (i) first of all not all kind of documents are publicly available due to policy issue and the majority of benchmarks are composed by scientific papers; (ii) furthermore, not all available data comes with structured information for automatic annotation, forcing to chose either to manually supervise a small amount of data (not desired for DL) or excluding a vast amount of unlabeled documents. These problems highly affect the develop DLA frameworks, both in terms of robustness and generalization. Following what is outlined in [31], a solution to fill the gap between expensive annotation procedures and large automatically labeled collections could be the generation of synthetic data that, by definition, comes with annotation. Even this third solution open to new possibilities, comes with the difficult to be as much faithful as possible to real data to do not let the trained algorithms drop drastically on the wild. In the next chapters we describe some generative methods, discussing how their usage is important with traditional annotation procedures.

2 First Year - 2020/2021

2.1 Achievements

In the first year of my PhD I initially continued my master thesis work whose main objective has been to perform administrative document understanding exploiting geometric deep learning methods. In particular, the main goal has been to build a pipeline capable to extract the graph document structures in order to find key information inside invoices, such as sender, receiver and total amount to be paid. Graphs proved themself to be well suited for that, helping to bring also structural features in the deep learning phase. Starting from the work described in [27], we proposed and compared different techniques in the graph construction pipeline and a novel graph convolutional architecture. We have been able to reproduce and obtain slightly better results on their provided subset of RVL-CDIP (518 annotated documents), mainly due to some structural information added to the graph edges.

One important limitation of the research in administrative document understanding is the lack of large publicly available datasets, with meaningful/useful labels and high quality documents. One possibility to continue working in that direction would be to explore new data augmentation or auto-labelling techniques, or involving methods that can deal with zero or few examples, e.g. zero/few shot learning. Therefore, in order to continue studying and using graphs in the document analysis domain, we extended the subject of study to the important field of table understanding in scientific papers. The reasons have been mainly two: in the domain of scientific papers large datasets such as PubLayNet [40] and DocBank [19] are available and, an in-depth study on objects such as tables does not exclude being able to reapply it again in the future to administrative documents.

The graphs previously created to represent an entire document have been reused to represent table structures. The first experiments conducted involved Detectron2 to automatically find cell objects inside tables: subsequently, the graph has been constructed connecting their bounding boxes using a visibility approach. In the nearly future I will try also to enrich nodes feature vectors with content information (with NLP-based techniques) and deep visual features: bringing different approaches together have already been reached great results [38].

In the first year I also extensively studied the state of the art of topics of interest for my research, also thanks to the conferences I attended (Sec. 2.3). I first deepened my knowledge about graph neural networks starting from the very first works [29] until the most recent discoveries [3, 4]. In the last three years graph have gathered an astonishing and increasing attention, allowing the publication of a growing number of papers in a wide range of different topics, along with new specialized libraries and benchmark datasets. Among others domain like chemistry, also researchers from the document analysis community brought graphs in their works [25, 27, 38]

In parallel, I also studied in depth the problem of table understanding, meaning both detection among other document objects and recognition of their structure in the meaning of cells, rows and columns. [10]. While the first task is nowadays a mostly solved task, the second one is still a challenging problem: simply applying an OCR is not enough to make them completely understandable and editable. Once the structure has been properly recognized, there are several consequent applications that can be performed, such as keep track of the state-of-the-art models among the literature exploiting results tables [15]. In the future, I would also like trying to apply my acquired knowledge in order to help visually impaired people access document contents in a smarter and easier way.

Most recently I also started to study the latest discoveries in the natural language processing domain, mainly connected to the document analysis one.

2.2 Plans and Future Work

As my main purpose for the next research year, I would like to continue and improve what I started during this one, focusing on tables and scientific papers. Among the steps I would like to take in the near future there is to improve and continue the work already done with tables:

- adding more relevant information to the nodes of table graph structures, also using NLP-based approaches;
- enriching table content finding in the paragraphs where the table is cited additional information exploiting in-paper references, e.g. information about datasets, models and metrics proposed;

Moreover, looking forward to the next two years, it would be interesting find a complete and summarised representation of an entire scientific paper in order to perform information extraction, topic classification and other related tasks.

2.3 Conferences and Summer Schools

I have attended the following courses, soft and complementary skills and conferences:

- NEURIPS 2020. 6-12 Dec 2020.
- ICPR 2020. 10-15 Jan 2021.
- SSPR 2020. 21-22 Jan 2021.
- "Learning Symbolic Equations with Deep Learning", ACM. 7 Jun 2021.
- ICDAR 2021. 5-10 Sep 2021.
- UCA Deep Learning School 2021, Université Côte D'Azur: Graph neural networks and neural-symbolic computation, (Marco Gori)

- 4th IAPR SSDA 2021 (on-site), Luleå University of Technology, 23-27 Aug: recent developments in document analysis.

3 Second Year - 2021/2022

3.1 Achievements

During my second year my research kept focusing on the application of Geometric Deep Learning to the domain of Document Analysis. I have mainly addressed the "plans and future work" outlined at the end of the first one (Sec. 2.2). We found these limitations interesting to be further explored:

1. scarcity of data for administrative documents;
2. usually table extraction task is tackled do not considering also other contextual information.
3. lack of information carried out throughout the graph structure beyond layout positioning, such as language and visual features;

To address the scarcity of data, we published two papers with novel data augmentation techniques. The first one, "Data augmentation on graphs for table type classification" [**del2023data**], proposes data augmentation directly on the graph structure. We have shown that, applying a combination of node /edge removal and column/row inversion techniques, we were able to re-balance the dataset in our hand and increase the performances for the downstream task. On the contrary in "Automatic generation of scientific papers for data augmentation in document layout analysis" [**pisaneschi2023automatic**], we have exploited a LayoutTransformer to generate scientific paper pages, both single and double column layouts. Using the new generated data, helped improving AP scores for Document Layout Analysis over two small collection of papers, ICDAR 2019 and of the ICPR 2021 workshops, in particular for double column layouts.

Regarding the table extraction task, we found out that several work carry out it in different separate steps. In "Graph neural networks and representation embedding for table extraction in PDF documents" [**gemelli2022graph**], we propose a geometric approach to tackle the Contextualized Table Extraction problem, addressing table extraction and document layout analysis at once. We also enriched nodes with novel "representation embeddings" and our ablation studies have shown that they are a good alternative to language models, such as SpaCy [8] or SciBERT [1], to distinguish tables from the rest of the document objects.

During my research stay at the Computer Vision Center (CVC) I deepened my knowledge about GNNs applied over documents related tasks, working to a new project with experts in this field. I developed a new library called "Doc2Graph", to create a task-agnostic pipeline overcoming the limitations that usually are applied during the extraction of graph structures from documents. We validate this novel proposal for four different task over two challenging benchmarks for Information Extraction and Document Understanding in our paper "Doc2Graph: a Task Agnostic Document Under- standing Framework based on Graph Neural Networks" [**gemelli2023doc2graph**]. We compared with other SOTA models, both graphs and transformers, obtaining good results

with a small amount of trainable parameter; furthermore, classifying nodes and edges in a end-to-end manner, we brought interpretability over the graph structure (in particular for the Entity Linking task on FUNSD). Doc2Graph helped also to address the third point left as to be done from the first year: we used visual features applying and comparing different backbones, and we have proposed novel relative positioning features over edges. In the paper we extensively discuss their usage and usefulness through ablation studies.

Beyond the work carried on, the experiences I have made positively helped me to increase my knowledge and networking. As a visiting student to the CVC I could work with expert colleagues and improve my research work, also with other topics of interest of machine learning in general. Furthermore, being in presence at ICPR conference and S+SSPR workshop, gave me the opportunity to talk about my work over a larger audience of expertise all over the world.

3.2 Plans and Future Work

For my next and last year of Ph.D. i would like to further explore these points:

- extend the data augmentation works over business documents. A new dataset "DocLayNet" [24] has been just released, and could be interesting to explore it to try the already developed techniques;
- try and propose novel "anonymous properties" for business documents as an alternative to language models, given the interest of industry over privacy issues;
- Expanding "Doc2Graph" for multi-lingual purposes. The new dataset "XFUND" [36] is an extension of "FUNSD" with forms coming from seven different languages, and it would be interesting try self-supervision / continual learning techniques exploiting my library to tackle this task.

3.3 Conferences and Summer Schools

From March 2022 I have been collaborating as a visiting researcher with the Computer Vision Center (CVC), Barcelona. In addition, I have attended the following conferences and summer school:

- Ellis Machine Learning Summer School (on-site), Cambridge University, 11-15 Jul 2022.
- ICPR 2022 (on-site), 21-25 Aug 2022.
- S+SSPR 2022 (on-site), 26-27 Aug 2022.
- Annual Catalan Meeting on Computer Vision ACMCV (on-site), 19 Sep 2022.

4 Third Year - 2022/2023

4.1 Achievements

During my third year my research kept focusing on the application of Geometric Deep Learning to the domain of Document Analysis. I am still actively working on an extension of our journal publication on PRL regarding document generation to address both scarcity of data and practical business needs and applications. In particular, I have worked in order to generate multi-page documents, conditioning their creation with the number of page and previous pages. This includes both novelty from a research perspective and a practical application need on everyday business needs. Moreover, I am still working on proposing a metric based on graph neural networks to meet the lack and necessity of evaluating the quality of generated documents in a quantitative way in order to be able to compare different methods. In addition to this main topic, I have also worked and published in slightly different domains that helped me applying my expertise in other interesting scenarios. I released, with my lab collegues, a dataset for dysgraphia detection in children handwriting and a trained algorithm, publishing the final work in the DocEng 2023 conference. The paper has been accepted for oral presentation that I personally exposed on site at the University of Limerick. A second and equally interesting research I have worked on required the application of Graph Neural Networks into a video forensic scenario. In particular, me and my collegues constructed a graph representation over videos following the H.264/AVI codec and trained a GNN to classify those into five different social network of origin, surpassing previous state-of-the-art methods on three different benchmarks. This is particular helpful to trace back malicious content published and shared over different social networks. Finally, we worked on collecting and summarising months of research in a survey currently under review for a journal. I described different datasets for Layout Analysis over Scientific Articles, a central task on Document Layout Analysis. In particular, we created a taxonomy of different annotations methods, describing pros and cons of the different techniques. Each dataset has been described in details to help scholars find the best benchmarks to try their algorithm on, classifying them by complexity and reporting the state-of-the-art methods trained over them.

4.2 Conferences and Summer Schools

I have attended the following conferences and summer school:

- Italian Research Conference on Digital Libraries (IRCDL) (on-site), Bari, 23-24 Feb 2023.
- International Computer Vision Summer School (ICVSS) (on-site), 09-15 Jul 2023.
- The 23rd ACM Symposium on Document Engineering (DocEng 2023) (on-site), 22-25 Aug 2023.

5 Visiting Research Period

I can broadly divide the whole year visiting into two main periods: a first half used to design the GNN-based framework we published as Doc2Graph; and a second one used to explore on how to enrich the developed project to be further extended.

In March 2022 I started my collaboration with the Computer Vision Center (CVC) of Universitat Autònoma de Barcelona. As a visiting researcher I moved there to start a new project on Entity Detection and Entity Linking tasks within rich documents such as forms. We started our joint research by considering two papers mainly: (i) Pau Riba et al. [28], where authors have shown GNNs being an effective solution for document information extraction and (ii) FUDGE [7] where a GNN has been exploited as a light-weight alternative to Transformers. On top of these previous inspiring solutions, we have built *"Doc2Graph"* [**gemelli2023doc2graph**], following this schedule:

- the first month, I revised the state-of-the-art to find methods to compare with, the tasks we were interested in to explore and the relevant datasets to be used for benchmarking.
- the second month, I have started to develop a first approach to tackle entity linking on FUNSD [13], comparing several GNNs architectures and reporting different ablation studies; during this period we chose DGL[1] as the development framework, after a preliminary study on different possibilities.
- during the third and fourth months, the Doc2Graph framework has been started to be developed, with user customisation on document preprocessing, network adaptation for the proposed message passing algorithm and first results visualisation.
- on the last month, before the submission on July 22, the method has been refined to meet the numbers presented in the paper, a second dataset and three new tasks have been added to the experimentation, and the paper has been written along with the production of all the additional material.

All these steps have been followed for five months on a weekly update schedule with my supervisor and CVC colleagues, where I always received useful feedbacks that made the work every time better.

The final work published in ECCV is a novel framework based on GNNs that we called *task-agnostic*, since it can generalise over four different tasks such as Entity Recognition, Entity Linking, Layout Analysis and Table Detection. Among other contributions, I would like to outline that the solution propsed it is light-weighted compared to large and main-stream language models and it does not rely on huge pretraining. At the time of publication, the method still had some limitations, such as the fixed visual features and more benchmarks

[1] https://www.dgl.ai

could have been included: we met some requirements on the second part of my stay, while others are still under investigation.

In the second period, starting from September 2022, we began to think on which direction we could have driven Doc2Graph for further extensions. With no specific order, we mainly worked on the following important points:

- the official source code has been released on github[2] to be shared with the community; in particular, after the proceedings publication, the code has attracted an increasing interest, that also helped to fix some problems left for reproducibility;
- the network has been updated to meet more recent trends [6, 12]; both the visual and positional encodings have been updated to be learnt at training time. Preliminary results already have shown interesting improvements, but nothing has been released yet;
- given the brief but successful history on pretraing approaches [41] and, to the best of my knowledge, the almost absolute abscence of successful approaches for graph structures [20], we started to explore a pretrained version of Doc2Graph over RVL-CDIP[9]. During the submission for the next ICDAR we found out that another excellent work [39] had already been shared (at the time) on arXiv following this direction: even if we found room of improvement, we thought for the sake of a good publication, to skip this deadline to have time including and comparing with their results and approach too;
- an important prior knowledge included in Transformers is the encoding of token order, that usually is lost on GNNs that are permutation invariant [32]; an interesting work proposing ROPE [18], made us start working in this direction as well, updating the message passing algorithm to take into consideration the reading order in the neighbouring nodes aggregation; preliminary experiments do not still meet our expectations and so this path still require further investigations.

5.1 Other Activities

Outside of my main project, at CVC I could join other important activities for my personal and professional growth. Among others, I would like to point out:

- weekly reading groups with other researchers to keep being updated on the latest discoveries on the field;
- internal seminars and poster sessions that helped me sharing my research also with experts outside of my research field to get insightful feedbacks on how to get better at lots of aspects;

[2] https://github.com/andreagemelli/doc2graph

- despite attending ECCV and presenting my work at one of the best conferences in computer science, I also had the chance to share projects in the The Annual Catalan Meeting on Computer Vision[3] (ACMCV) 2022 poster sessions;

[3] http://acmcv.cat/

6 Publications

Here I list the publication made as a result of my work, accepted by conferences, workshops and journals:

- "Doc2Graph: a Task Agnostic Document Understanding Framework based on Graph Neural Networks", TiE @ ECCV 2022 [**gemelli2023doc2graph**] - Workshop Poster, Tel Aviv (Israel)
- "Graph neural networks and representation embedding for table extraction in PDF documents", ICPR 2022 [**gemelli2022graph**] - Conference Poster, Montréal (Canada)
- "Data augmentation on graphs for table type classification", S+SSPR 2022 [**del2023data**] - Workshop Oral, Montréal (Canada)
- "Automatic generation of scientific papers for data augmentation in document layout analysis", ANDARDA Special Issue (Patter Recognition Letter) [**pisaneschi2023automatic**] - Journal
- "CTE: A Dataset for Contextualized Table Extraction", IRCDL 2023 [**gemelli2023cte**] - Conference, Bari (Italy)
- "Structure Matters: Analyzing Video Via Graph Neural Networks for Social Media Platform Attribution", *(under review)* [**gemelli2023structure**]
- "Datasets and Annotations for Layout Analysis in Scientific Articles", *(under review)* [**gemelli2023datasets**] - Journal

7 Credits

During my Ph.D. I have attended the following exams and complementary skills courses:

Exams

- Probabilistic Graphical Models(Manfred Jaeger). Oct-Nov 2020, (3 CFU)
- Sequence Learning (Paolo Frasconi). Apr 2021, (3 CFU)
- Memory Networks (Federico Becattini). Apr-May 2021, (1 CFU, no exam held)
- Explainable AI (Paolo Frasconi). May 2021, (3 CFU)
- 4th IAPR SSDA 2021 (on-site), Luleå University of Technology, 23-27 Aug 2021: recent developments in document analysis. (5 CFU)
- Towards Developmental Learning (Marco Gori). 20-27 Jun 2022 (5 CFU, no exam held)
- Ellis Machine Learning Summer School (on-site), Cambridge University, 11-15 Jul 2022. (5 CFU)
- Learning with multiple data distributions (Paolo Frasconi). July 2023, (3 CFU)

Complementary Skills

- Incontri potenziamento competenze trasversali. Nov-Dec 2020, (1.5 CFU).
- Impresa Campus Univi: Final presentation of works, Dec 2020, (0.5 CFU).
- Writing, Publishing, Presenting and Searching Scientific Literature, including Journalology. 26-29 January, (3 CFU).
- Impresa Campus Unifi, first call 2021. Feb-Jul 2021. (no credits got)
- La comunicazione scientifica, Dott.ssa E. Jafrancesco, 11, 18, 25 May 2023, (1 CFU)

References

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. "SciBERT: A Pretrained Language Model for Scientific Text". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui et al. Association for Computational Linguistics, 2019, pp. 3613–3618. DOI: 10.18653/v1/D19-1371. URL: https://doi.org/10.18653/v1/D19-1371.

[2] Galal M. Binmakhashen and Sabri A. Mahmoud. "Document Layout Analysis: A Comprehensive Survey". In: *ACM Comput. Surv.* 52.6 (Oct. 2019). ISSN: 0360-0300. DOI: 10.1145/3355610. URL: https://doi.org/10.1145/3355610.

[3] Michael Bronstein. *Deep learning on graphs: successes, challenges, and next steps*. https://tinyurl.com/deep-learning-on-graphs.

[4] Michael M. Bronstein et al. "Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges". In: *CoRR* abs/2104.13478 (2021). arXiv: 2104.13478. URL: https://arxiv.org/abs/2104.13478.

[5] Lei Cui et al. "Document AI: Benchmarks, Models and Applications". In: *arXiv preprint arXiv:2111.08609* (2021).

[6] Brian Davis et al. "End-to-end document recognition and understanding with Dessurt". In: *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*. Springer. 2023, pp. 280–296.

[7] Brian Davis et al. "Visual fudge: Form understanding via dynamic graph editing". In: *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*. Springer. 2021, pp. 416–431.

[8] Explosion. *spaCy: Industrial-strength NLP*. https://spacy.io/. 2016.

[9] Adam W Harley, Alex Ufkes, and Konstantinos G Derpanis. "Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval". In: *International Conference on Document Analysis and Recognition (ICDAR)*. 2015.

[10] Khurram Azeem Hashmi et al. "Current Status and Performance Analysis of Table Recognition in Document Images with Deep Neural Networks". In: *CoRR* abs/2104.14272 (2021). arXiv: 2104.14272. URL: https://arxiv.org/abs/2104.14272.

[11] Kaiming He et al. "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[12] Yupan Huang et al. "Layoutlmv3: Pre-training for document ai with unified text and image masking". In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 4083–4091.

[13] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. "Funsd: A dataset for form understanding in noisy scanned documents". In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. Vol. 2. IEEE. 2019, pp. 1–6.

[14] Antonio Jimeno Yepes, Peter Zhong, and Douglas Burdick. "ICDAR 2021 competition on scientific literature parsing". In: *International Conference on Document Analysis and Recognition*. Springer. 2021, pp. 605–617.

[15] Marcin Kardas et al. "AxCell: Automatic Extraction of Results from Machine Learning Papers". In: *CoRR* abs/2004.14356 (2020). arXiv: 2004.14356. URL: https://arxiv.org/abs/2004.14356.

[16] Koichi Kise. "Page Segmentation Techniques in Document Analysis". In: *Handbook of Document Image Processing and Recognition*. Ed. by David S. Doermann and Karl Tombre. Springer, 2014, pp. 135–175. DOI: 10.1007/978-0-85729-859-1_5.

[17] Koichi Kise, Akinori Sato, and Motoi Iwata. "Segmentation of page images using the area Voronoi diagram". In: *Computer Vision and Image Understanding* 70.3 (1998), pp. 370–382.

[18] Chen-Yu Lee et al. "ROPE: reading order equivariant positional encoding for graph-based document information extraction". In: *arXiv preprint arXiv:2106.10786* (2021).

[19] Minghao Li et al. *DocBank: A Benchmark Dataset for Document Layout Analysis*. 2020. arXiv: 2006.01038 [cs.CL].

[20] Yixin Liu et al. "Graph self-supervised learning: A survey". In: *IEEE Transactions on Knowledge and Data Engineering* (2022).

[21] Simone Marinai. "Learning algorithms for document layout analysis". In: *Handbook of Statistics*. Vol. 31. Elsevier, 2013, pp. 400–419.

[22] George Nagy and Sharad C Seth. "Hierarchical representation of optically scanned documents". In: (1984).

[23] Lawrence O'Gorman. "The document spectrum for page layout analysis". In: *IEEE Transactions on pattern analysis and machine intelligence* 15.11 (1993), pp. 1162–1173.

[24] Birgit Pfitzmann et al. "DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis". In: *arXiv preprint arXiv:2206.01062* (2022).

[25] Shah Rukh Qasim, Hassan Mahmood, and Faisal Shafait. "Rethinking Table Parsing using Graph Neural Networks". In: *CoRR* abs/1905.13391 (2019). arXiv: 1905.13391. URL: http://arxiv.org/abs/1905.13391.

[26] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (2015).

[27] Pau Riba et al. “Table Detection in Invoice Documents by Graph Neural Networks”. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 2019, pp. 122–127. DOI: 10.1109/ICDAR.2019.00028.

[28] Pau Riba et al. “Table detection in invoice documents by graph neural networks”. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2019, pp. 122–127.

[29] Franco Scarselli et al. “The Graph Neural Network Model”. In: *IEEE Transactions on Neural Networks* 20.1 (2009), pp. 61–80. DOI: 10.1109/TNN.2008.2005605.

[30] Charalambos Strouthopoulos and Nikos Papamarkos. “Text identification for document image analysis using a neural network”. In: *Image and Vision Computing* 16.12-13 (1998), pp. 879–896.

[31] Ernest Valveny. “Datasets and Annotations for Document Analysis and Recognition”. In: *Handbook of Document Image Processing and Recognition*. Ed. by David Doermann and Karl Tombre. London: Springer London, 2014, pp. 983–1009. ISBN: 978-0-85729-859-1. DOI: 10.1007/978-0-85729-859-1_32. URL: https://doi.org/10.1007/978-0-85729-859-1_32.

[32] Petar Veličković. “Everything is connected: Graph neural networks”. In: *Current Opinion in Structural Biology* 79 (2023), p. 102538.

[33] Friedrich M Wahl, Kwan Y Wong, and Richard G Casey. “Block segmentation and text extraction in mixed text/image documents”. In: *Computer graphics and image processing* 20.4 (1982), pp. 375–390.

[34] Chung-Chih Wu, Chien-Hsing Chou, and Fu Chang. “A machine-learning approach for analyzing document layout structures with two reading orders”. In: *Pattern Recognition* 41.10 (2008), pp. 3200–3213.

[35] Yiheng Xu et al. “Layoutlm: Pre-training of text and layout for document image understanding”. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 1192–1200.

[36] Yiheng Xu et al. “XFUND: A Benchmark Dataset for Multilingual Visually Rich Form Understanding”. In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3214–3224. DOI: 10.18653/v1/2022.findings-acl.253. URL: https://aclanthology.org/2022.findings-acl.253.

[37] Peng Zhang et al. “VSR: A Unified Framework for Document Layout Analysis combining Vision, Semantics and Relations”. In: *ICDAR*. Vol. 12821. 2021, pp. 115–130.

[38] Peng Zhang et al. “VSR: A Unified Framework for Document Layout Analysis combining Vision, Semantics and Relations”. In: *CoRR* abs/2105.06220 (2021). arXiv: 2105.06220. URL: https://arxiv.org/abs/2105.06220.

[39] Zhenrong Zhang et al. “Multimodal Pre-training Based on Graph Attention Network for Document Understanding”. In: *IEEE Transactions on Multimedia* (2022), pp. 1–13. DOI: 10.1109/TMM.2022.3214102.

[40] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. “PubLayNet: largest dataset ever for document layout analysis”. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. Sept. 2019, pp. 1015–1022. DOI: 10.1109/ICDAR.2019.00166.

[41] Ce Zhou et al. “A comprehensive survey on pretrained foundation models: A history from bert to chatgpt”. In: *arXiv preprint arXiv:2302.09419* (2023).