SciA11y: Converting Scientific Papers to Accessible HTML

Lucy Lu Wang*
lucyw@allenai.org
Allen Institute for AI
Seattle, WA, USA

Isabel Cachola*†
icachola@cs.jhu.edu
The Johns Hopkins University
Baltimore, MD, USA

Jonathan Bragg
jbragg@allenai.org
Allen Institute for AI
Seattle, WA, USA

Evie Yu-Yen Cheng
eviec@allenai.org
Allen Institute for AI
Seattle, WA, USA

Chelsea Haupt
chealseah@allenai.org
Allen Institute for AI
Seattle, WA, USA

Matt Latzke
mattl@allenai.org
Allen Institute for AI
Seattle, WA, USA

Bailey Kuehl
baileyk@allenai.org
Allen Institute for AI
Seattle, WA, USA

Madeleine van Zuylen
madeleinev@allenai.org
Allen Institute for AI
Seattle, WA, USA

Linda Wagner
lindaw@allenai.org
Allen Institute for AI
Seattle, WA, USA

Daniel S. Weld
danw@allenai.org
Allen Institute for AI and University
of Washington
Seattle, WA, USA

ABSTRACT

We present SciA11y, a system that renders inaccessible scientific paper PDFs into HTML. SciA11y uses machine learning models to extract and understand the content of scientific PDFs, and reorganizes the resulting paper components into a form that better supports skimming and scanning for blind and low vision (BLV) readers. SciA11y adds navigation features such as tagged headings, a table of contents, and bidirectional links between inline citations and references, which allow readers to resolve citations without losing their context. A set of 1.5 million open access papers are processed and available at https://scia11y.org/. This system is a first step in addressing scientific PDF accessibility, and may significantly improve the experience of paper reading for BLV users.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in accessibility**; **Accessibility systems and tools**; *HCI design and evaluation methods*; *Accessibility design and evaluation methods.*

KEYWORDS

accessibility, accessible reader, scientific documents, blind and low vision users

ACM Reference Format:
Lucy Lu Wang, Isabel Cachola, Jonathan Bragg, Evie Yu-Yen Cheng, Chelsea Haupt, Matt Latzke, Bailey Kuehl, Madeleine van Zuylen, Linda Wagner, and Daniel S. Weld. 2021. SciA11y: Converting Scientific Papers to Accessible HTML. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21), October 18–22, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3441852.3476545

1 INTRODUCTION

Scientific papers are primarily available in PDF, a format designed for faithful visual representation that is difficult to make web accessible. Though PDFs can be made accessible,[1] the process is rife with challenges [3]. Though signals suggest that scientific PDF accessibility is improving over time [10, 17], the vast majority of previously and newly published paper PDFs remain inaccessible. For example, Wang et al. [17] find in an analysis of papers sampled across diverse scientific disciplines that only 2.4% of scientific PDFs published since 2010 satisfy five defined accessibility criteria, with only around 15% of papers being properly tagged, and 7% of papers including detectable figure alt-text (though the actual percentage of usable alt-text is much lower).

Science publishers are shifting towards dual publishing or alternate publishing schemes[2] that yield accessible HTML or XML versions of papers in addition to PDF. However,it is unclear how

[*] Denotes equal contribution
[†] Work done while at the Allen Institute for AI

[1] https://helpx.adobe.com/acrobat/using/creating-accessible-pdfs.html
[2] The ACM (https://www.acm.org/publications/authors/submissions) and eLife (https://reviewer.elifesciences.org/author-guide/journal-policies) are examples.

ASSETS '21, October 18–22, 2021, Virtual Event, USA
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8306-6/21/10.
https://doi.org/10.1145/3441852.3476545

quickly broad adoption will be reached; in the meantime, the significant back catalog of existing research works continues to languish in their inaccessible forms. Tools are available for extracting content from PDFs (e.g. PDF parsers such as pdf2text, pdfalto, PDFBox), converting between document formats (e.g. Pandoc, pdf2html), or making PDFs accessible (e.g. Adobe Acrobat Pro). However, these tools 1) may only extract textual components (in the case of many PDF parsers), 2) may not handle PDFs as input (in the case of Pandoc, which can convert *to* but not *from* PDF), 3) may not produce a satisfactory representation of the scientific document that contains the document's original structural semantics (e.g. most PDF parsers produce a flat stream of tokens and locations as output), or 4) may not do this in an automated fashion (e.g. Adobe Acrobat Pro requires significant manual input to resolve reading order and tag headings and objects appropriately, and is proprietary). In other words, there is currently no suitable system that can understand the content of scientific PDFs en masse and present this content accessibly to users of screen readers.

As an interim solution to this problem, we introduce SciA11y, a system that employs document parsing methods to extract the semantic content of scientific PDFs. The system renders this content in HTML and introduces accessibility features such as navigational headings, tagged objects, table of contents, and within-document navigational links. An example document highlighting various features provided by SciA11y is shown in Figure 1. In this paper, we describe the features available in SciA11y (https://scia11y.org/), and the methods used to produce these features. An intrinsic evaluation of HTML quality revealed that around 86% of papers in our sample had reasonable extractions (good or okay readability per criteria described in Wang et al. [17]), and a preliminary user study with 6 BLV researchers was also positive, with all users stating they would be likely to use the system in the future were it to have high coverage of papers. We refer readers to Wang et al. [17] for further details on evaluation.

2 THE SCIA11Y DEMO

A demo of SciA11y is available at https://scia11y.org/. We process 1.5 million open access scientific PDFs using our pipeline and make these available in the demo. Processing steps for deriving a document with improved logical reading order are described in Section 2.1 and the addition of navigational features are described in Section 2.2.

2.1 Inferring logical reading order

SciA11y integrates the output of two scientific PDF extraction systems: S2ORC [11] (which leverages Grobid [12] to process PDFs) and DeepFigures [15], to identify and extract textual elements and figure/table elements respectively. The Grobid PDF processing library [12] generates an XML representation of the document content with labeled text spans corresponding to metadata fields like title, authors, affiliation, and venue; textual fields like section headings, body text paragraphs, and figure and table captions; and bibliographic fields like reference entries. S2ORC [11] then implements additional logic to improve the accuracy of Grobid output and merge this output with metadata from the Semantic Scholar literature corpus.[3] For figure/table elements, we employ DeepFigures [15], which identifies bounding boxes for figures and tables in scientific PDFs and extracts these as images along with their caption text.

We combine textual elements from S2ORC and figure/table elements from DeepFigures to create the HTML representation. We attempt to preserve the logical reading order intended by the authors. To do this, we begin with metadata fields like title, author, and abstract, followed by paragraphs organized under section headings, placed in intended reading order based on columnar organization, and finally, the references section with itemized bibliography entries. Figures and tables are inserted at paragraph breaks immediately following their first mention (heuristically detected). The logical order of figures and tables is also preserved, i.e., if "Figure 2" is mentioned in text before "Figure 1", we insert Figure 1 before Figure 2 immediately following this paragraph.

2.2 Navigation features

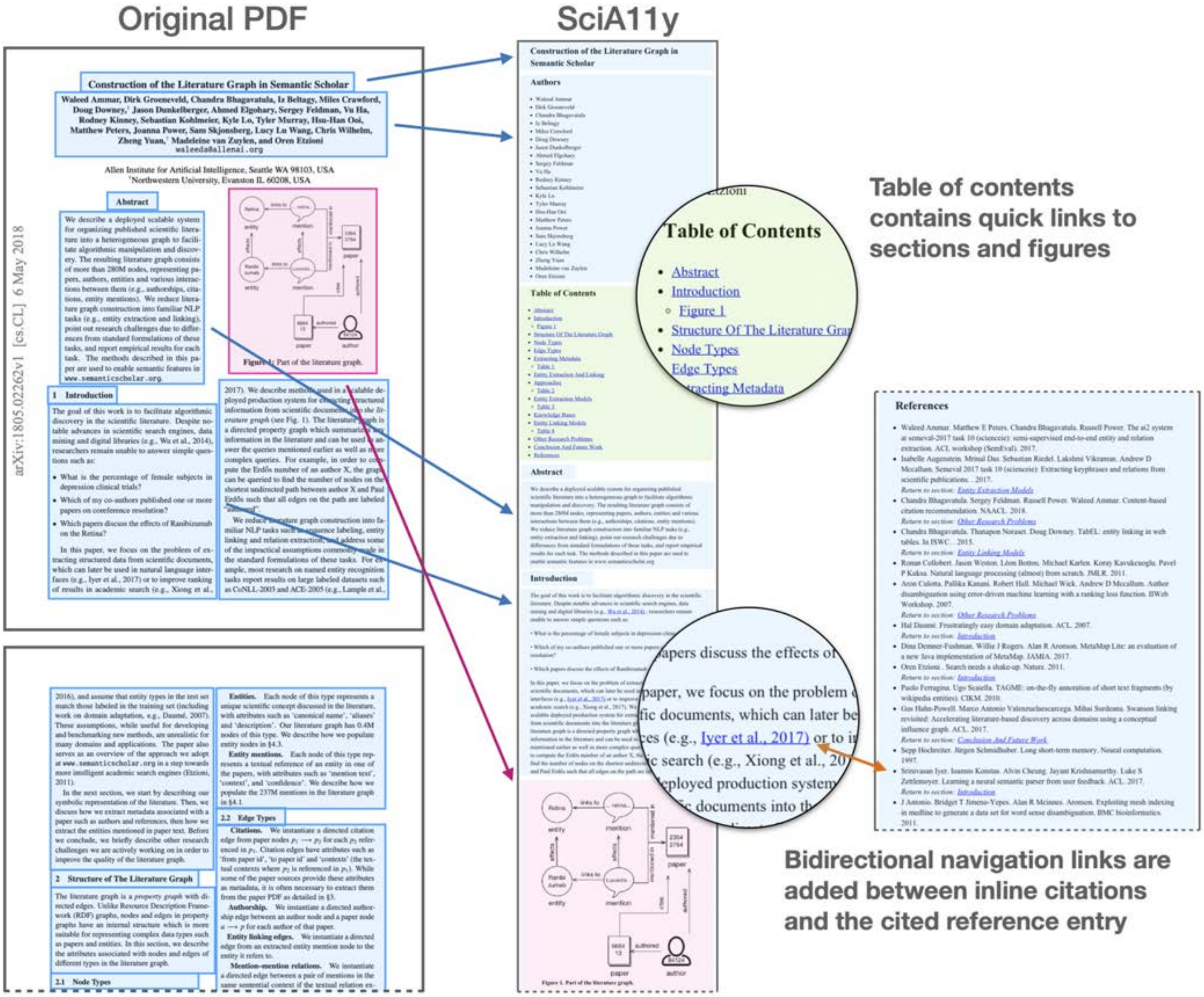
To the basic document structure, we add navigational features, focusing on features to support skimming and scanning for users of screen readers (examples shown in Figure 1). We add a table of contents near the beginning of the document, with links resolving to all section headers as well as figures and tables nested under these headers. This exposes users to the overall structure of the paper, and allows readers to more quickly skip to intended sections. We also insert bidirectional links between inline citations and the cited reference entry in the bibliography. When multiple inline citations exist for the same reference entry, multiple back links are provided to the first occurrence of the citation in each section of the paper. This allows the user to resolve an inline citation and return to their original reading context with minimal key strokes.

3 SUMMARY

Scientific PDFs are difficult to make accessible for screen readers, and most paper PDFs are consequently inaccessible and difficult to navigate. We propose and demonstrate a pipeline for extracting the semantic content of paper PDFs and rendering this content as an accessible HTML document. Our hope is that this system and framework resolves some of the challenges of reading papers using screen readers, and can be a starting platform for further improvements. Both an intrinsic and preliminary user evaluation of the system had positive results (see Wang et al. [17] for details). In a user study conducted with 6 BLV researchers who primarily engage with scientific papers using screen readers, all users responded positively to our introduced navigation features, as they simplified skimming the document and navigating to specific sections. All users said that they would be likely to use the system in the future if available, indicating a need for systems like SciA11y. The user studies also revealed a host of other challenges, such as difficulties accessing the content of figures, tables, and math equations, which pave the way for future work in this direction.

The current iteration of SciA11y focuses on improving screen reader accessibility in terms of navigation within the HTML document. To build upon this framework, we intend to improve PDF

[3] https://semanticscholar.org/


Original PDF
SciA11y
Table of contents contains quick links to sections and figures
Bidirectional navigation links are added between inline citations and the cited reference entry
arXiv:1805.02262v1 [cs.CL] 6 May 2018
Table of Contents
Abstract
Introduction
Figure 1
Structure Of The Literature Gra
Node Types
Edge Types
tracting Metadata
apers discuss the effects of
paper, we focus on the problem o
fic documents, which can later be
ces (e.g., Iyer et al., 2017) or to i
c search (e.g., Xiong et al., 20
eployed production system
c documents into th

Construction of the Literature Graph in Semantic Scholar

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey,† Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan,† Madeleine van Zuylen, and Oren Etzioni
waleeda@allenai.org

Allen Institute for Artificial Intelligence, Seattle WA 98103, USA
†Northwestern University, Evanston IL 60208, USA

Abstract

We describe a deployed scalable system for organizing published scientific literature into a heterogeneous graph to facilitate algorithmic manipulation and discovery. The resulting literature graph consists of more than 280M nodes, representing papers, authors, entities and various interactions between them (e.g., authorships, citations, entity mentions). We reduce literature graph construction into familiar NLP tasks (e.g., entity extraction and linking), point out research challenges due to differences from standard formulations of these tasks, and report empirical results for each task. The methods described in this paper are used to enable semantic features in www.semanticscholar.org.

Figure 1: Part of the literature graph.

1 Introduction

The goal of this work is to facilitate algorithmic discovery in the scientific literature. Despite notable advances in scientific search engines, data mining and digital libraries (e.g., Wu et al., 2014), researchers remain unable to answer simple questions such as:

- What is the percentage of female subjects in depression clinical trials?

References

- Waleed Ammar. Matthew E Peters. Chandra Bhagavatula. Russell Power. The ai2 system at semeval-2017 task 10 (scienceie): semi-supervised end-to-end entity and relation extraction. ACL workshop (SemEval). 2017.
- Isabelle Augenstein. Mrinal Das. Sebastian Riedel. Lakshmi Vikraman. Andrew D Mccallum. Semeval 2017 task 10 (scienceie): Extracting keyphrases and relations from scientific publications. . 2017.
  Return to section: Entity Extraction Models
- Chandra Bhagavatula. Sergey Feldman. Russell Power. Waleed Ammar. Content-based citation recommendation. NAACL. 2018.
  Return to section: Other Research Problems
- Chandra Bhagavatula. Thanapon Noraset. Doug Downey. TabEL: entity linking in web tables. In ISWC. . 2015.
  Return to section: Entity Linking Models
- Ronan Collobert. Jason Weston. Léon Bottou. Michael Karlen. Koray Kavukcuoglu. Pavel P Kuksa. Natural language processing (almost) from scratch. JMLR. 2011.
- Aron Culotta. Pallika Kanani. Robert Hall. Michael Wick. Andrew D Mccallum. Author disambiguation using error-driven machine learning with a ranking loss function. IIWeb Workshop. 2007.
  Return to section: Other Research Problems
- Hal Daumé. Frustratingly easy domain adaptation. ACL. 2007.
  Return to section: Introduction
- Dina Demner-Fushman. Willie J Rogers. Alan R Aronson. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. JAMIA. 2017.
- Oren Etzioni . Search needs a shake-up. Nature. 2011.
  Return to section: Introduction
- Paolo Ferragina. Ugo Scaiella. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). CIKM. 2010.
- Gus Hahn-Powell. Marco Antonio Valenzuelaescarcega. Mihai Surdeanu. Swanson linking revisited: Accelerating literature-based discovery across domains using a conceptual influence graph. ACL. 2017.
  Return to section: Conclusion And Future Work
- Sepp Hochreiter. Jürgen Schmidhuber. Long short-term memory. Neural computation. 1997.
- Srinivasan Iyer. Ioannis Konstas. Alvin Cheung. Jayant Krishnamurthy. Luke S Zettlemoyer. Learning a neural semantic parser from user feedback. ACL. 2017.
  Return to section: Introduction
- J Antonio. Bridget T Jimeno-Yepes. Alan R Mcinnes. Aronson. Exploiting mesh indexing in medline to generate a data set for word sense disambiguation. BMC bioinformatics. 2011.

Figure 1: SciA11y converts PDF to HTML, introducing navigation features such as a table of contents and links between inline citations and cited reference entries. An example of several pages of Ammar et al. [1] rendered into HTML by SciA11y.

extraction, e.g., by adopting better scientific document understanding models like those proposed in Shen et al. [14], and provide better representations of paper elements like figures, tables, and equations. For example, we intend to integrate features for reading graphs and charts [4–6], mathematical equations [2, 7, 13, 16], and further processing table images into HTML [8, 18, 19]. Regarding figures specifically, the vast majority of figures from scientific papers lack alt-text, and methods for inferring or generating alt-text could improve current screen reader user experience when reading figures. As for UI/UX considerations, interactions such as those proposed in Head et al. [9] may allow us to provide additional information at the point of interest (surfacing bibliography entries at the inline citation rather than navigating away), further reducing loss of reading context, though the accessibility of these interactions for users of screen readers must be explored independently.

In summary, we introduce the SciA11y system for rendering scientific PDFs as HTML, which can increase the accessibility of these documents for screen readers. A set of 1.5 million open access papers are available to read in HTML format at our demo site: https://scia11y.org/.

ACKNOWLEDGMENTS

This work was supported in part by ONR grant N00014-18-1-2193, NSF RAPID grant 2040196, and the University of Washington WRF/Cable Professorship. We thank Sam Skjonsberg for help deploying the demo. We thank the Semantic Scholar team for assisting with data access and system infrastructure. Finally, we thank the user study participants and many others who offered invaluable feedback and suggestions to improve the prototype.

REFERENCES

[1] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu A. Ha, Rodney Michael Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler C. Murray, Hsu-Han Ooi, Matthew E. Peters, Joanna L. Power, Sam Skjonsberg, Lucy Lu Wang, Christopher Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the Literature Graph in Semantic Scholar. In *NAACL-HLT*.

[2] E. Bates and D. Fitzpatrick. 2010. Spoken Mathematics Using Prosody, Earcons and Spearcons. In *ICCHP*.

[3] Jeffrey P. Bigham, E. Brady, Cole Gleason, Anhong Guo, and D. Shamma. 2016. An Uninteresting Tour Through Why Our Research Papers Aren't Accessible. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2016).

[4] Stephanie Elzer, E. J. Schwartz, S. Carberry, D. Chester, Seniz Demir, and Peng Wu. 2008. Accessible bar charts for visually impaired users.

[5] Christin Engel, David Gollasch, Meinhardt Branig, and G. Weber. 2017. Towards Accessible Charts for Blind and Partially Sighted People. In *Mensch & Computer*.

[6] Christin Engel, E. Müller, and G. Weber. 2019. SVGPlott: an accessible tool to generate highly adaptable, accessible audio-tactile charts for and from blind and visually impaired people. *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (2019).

[7] S. Flores, M. Andrade-Aréchiga, Alfonso Flores-Barriga, and Juan Lazaro-Flores. 2010. MathML to ASCII-Braille and Hierarchical Tree Converter. In *ICCHP*.

[8] L. Gao, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and E. Lang. 2019. ICDAR 2019 Competition on Table Detection and Recognition (cTDaR). *2019 International Conference on Document Analysis and Recognition (ICDAR)* (2019), 1510–1515.

[9] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2021. Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021).

[10] J. Lazar, E. Churchill, T. Grossman, G. V. D. Veer, Philippe A. Palanque, J. Morris, and Jennifer Mankoff. 2017. Making the field of computing more inclusive. *Commun. ACM* 60 (2017), 50 – 59.

[11] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4969–4983. https://doi.org/10.18653/v1/2020.acl-main.447

[12] P. Lopez and Laurent Romary. 2015. GROBID - Information Extraction from Scientific Publications. *ERCIM News* 2015 (2015).

[13] M. Mackowski, P. Brzoza, M. Zabka, and D. Spińczyk. 2017. Multimedia platform for mathematics' interactive learning accessible to blind people. *Multimedia Tools and Applications* 77 (2017), 6191–6208.

[14] Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S. Weld, and Doug Downey. 2021. Incorporating Visual Layout Structures for Scientific Text Classification. *ArXiv* abs/2106.00676 (2021).

[15] N. Siegel, Nicholas Lourie, R. Power, and Waleed Ammar. 2018. Extracting Scientific Figures with Distantly Supervised Neural Networks. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (2018).

[16] V. Sorge, C. Chen, T. Raman, and David Tseng. 2014. Towards making mathematics a first class citizen in general screen readers. In *W4A*.

[17] Lucy Lu Wang, Isabel Cachola, Jonathan Bragg, Evie Yu-Yen Cheng, Chelsea Hess Haupt, Matt Latzke, Bailey Kuehl, Madeleine van Zuylen, Linda M. Wagner, and Daniel S. Weld. 2021. Improving the Accessibility of Scientific Documents: Current State, User Needs, and a System Solution to Enhance Scientific PDF Accessibility for Blind and Low Vision Users. *ArXiv* abs/2105.00076 (2021).

[18] Jiaquan Ye, X. Qi, Yelin He, Yihao Chen, Dengyi Gu, Peng Gao, and Rong Xiao. 2021. PingAn-VCGroup's Solution for ICDAR 2021 Competition on Scientific Literature Parsing Task B: Table Recognition to HTML. *ArXiv* abs/2105.01848 (2021).

[19] Xinyi Zheng, D. Burdick, Lucian Popa, and N. Wang. 2021. Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2021), 697–706.