Enhanced Table Structure Recognition with Multi-Modal Approach

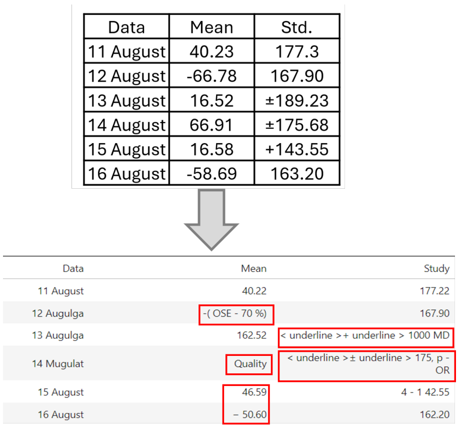
Huichen Yang[1]**, Andrew Hellicar**[1]**, Maciej Rybinski**[2]**, Sarvnaz Karimi**[1]
[1]CSIRO Data61, Australia
[2]ITIS, University of Málaga, Málaga, Spain
{Huichen.Yang, Andrew.Hellicar, Sarvnaz.Karimi}@data61.csiro.au, maciek.rybinski@uma.es

Abstract

Tables are fundamental for presenting information in research articles, technical documents, manuals, and reports. One key challenge is accessing the information in tables that are embedded in Portable Document Format (PDF) files or scanned images. It requires accurately recognising table structures in diverse table layouts and complex tables. Table Structure Recognition (TSR) task aims to recognise the internal structure of table images and convert them into a machine-readable format. We propose a flexible multi-modal framework for image-based TSR. Our approach utilises two-stream transformer encoders in conjunction with task-specific decoders for extracting table structures and detecting cell bounding boxes. Experiments on benchmark datasets demonstrate that our model achieves highly competitive results compared to strong baselines, outperforming single-modality approaches by 5.4% on the FinTabNetd dataset.

1 Introduction

Tables commonly present and summarise information in a structured format. They are widely used in various texts, such as scientific literature, books, business documents, manuals, and technical documents, due to their easier readability in presenting data. Managing, understanding, and analysing table data have become increasingly important, especially with the rapid growth of digitised data and the demand for intelligent document processing (Cui et al., 2021; Yu et al., 2023). However, table data are often restricted to digitised documents or images. While humans can easily interpret them, they are not readily processed by machines. The digitised table can be easily converted into a table image, but recognising its structure is challenging due to the complex styles. Therefore, extracting table data while preserving its structure in a machine-readable format is a fundamental step in table understanding.



Data	Mean	Std.
11 August	40.23	177.3
12 August	-66.78	167.90
13 August	16.52	±189.23
14 August	66.91	±175.68
15 August	16.58	+143.55
16 August	-58.69	163.20

Data	Mean	Study
11 August	40.22	177.22
12 Augulga	-(OSE - 70 %)	167.90
13 Augulga	162.52	< underline >+ underline > 1000 MD
14 Mugulat	Quality	< underline >± underline > 175, p - OR
15 August	46.59	4 - 1 42.55
16 August	– 50.60	162.20

Figure 1: Examples of failures in an end-to-end method include cases where the model identifies the correct table structure but incorrect content (Ly and Takasu, 2023).

Table Structure Recognition (TSR) is the task of automatically recognising table structures and extracting table content as free text for machine processing, which is a key step in table understanding. The table structure could follow pre-defined formats, such as HTML or JSON. Once the table structure is recognised, the table content can be extracted by any optical character recognition (OCR) tool, allowing the reorganisation of data into a table as it was originally presented in the table image. The structured table data, consisting of free text, enables machine processing and analysis of table data, and it is a crucial step for table-related downstream tasks, such as table-based question answering (TQA) (Iyyer et al., 2017; Chen et al., 2020b; Gupta et al., 2023), table-based fact verification (Chen et al., 2020a; Xie et al., 2022), information retrieval (Chen et al., 2020c; Engelmann et al., 2023), and text mining (Xie et al., 2020).

Tables have diverse structures and styles, which pose significant challenges for accurate recognition. For instance, tabular data is often organised with cells spanning multiple rows and columns. Such ta-

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 201–207
December 23, 2025 ©2025 Association for Computational Linguistics

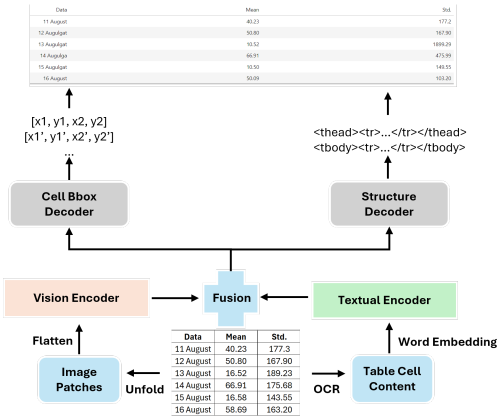
bles may include complex headers, cells containing multi-line text, empty cells, and varying line sizes or shapes used to separate cell contents. Moreover, table size introduces an additional challenge for the TSR task, as large tables may extend across multiple pages, particularly in certain scientific domains or technical documents.

Models based on deep neural networks have been proposed to address challenges in the TSR task. Recent methods for TSR can be divided into two strategies: the end-to-end and the non-end-to-end approach. The end-to-end method aims to use a single pipeline to process a given table image and output all table information, including the table structure, table cell bounding boxes, and table cell content (Schreiber et al., 2017; Ly and Takasu, 2023). This method is straightforward to understand, but its effectiveness is often unsatisfactory, especially when complex characters are present in the cell content. For example, as shown in Figure 1, the table structure can be identified well, but some content may be lost or incorrectly recognised.

On the other hand, the non-end-to-end method divides the TSR task into two sub-tasks: (1) recognising the table structure and table cell bounding boxes; and (2) extracting the cells' contents (Qiao et al., 2021; Nassar et al., 2022). Table cell content recognition can be considered an OCR task, which means we only need to extract the content rather than understand its semantic meaning. Many off-the-shelf OCR tools can be utilised instead of being integrated into the model training process to increase the training complexity.

We explore the efficacy of pre-training a multi-modal model for TSR. We propose a novel multi-modal approach for the TSR task, which differs from previous studies that only consider single modality pixel-based images (e.g., (Chi et al., 2019; Xing et al., 2023)). Our approach uses both the table image and its content as inputs for two transformer-based encoders, followed by separate decoders to generate the table structure and bounding boxes for non-empty table cells. This method aims to enhance the accuracy and robustness of TSR by integrating multiple data modalities, addressing the limitations of single-modality models for the task. Our main contributions are summarised as follows:

- Exploring and comparing the effectiveness of multi-modal models compared to vision-based models for the TSR task.



Data	Mean	Std.
11 August	40.23	177.2
12 Augulgat	50.80	167.90
13 Augulgat	10.52	1899.29
14 Augulga	66.91	475.99
15 Augulgat	10.50	149.55
16 August	50.09	103.20

Data	Mean	Std.
11 August	40.23	177.3
12 August	50.80	167.90
13 August	16.52	189.23
14 August	66.91	175.68
15 August	16.58	143.55
16 August	58.69	163.20

Figure 2: Our proposed two-stream multi-modal model architecture.

- Proposing a novel multi-modal approach for the TSR task, and the experimental results demonstrate that the approach is efficient.

2 Related Work

Early work on the TSR task relied on heuristic rule-based methods. These approaches required hand-crafted features and designed rules or templates to cover specific table layouts for structure recognition. For example, ruling lines were used to detect horizontal and vertical lines in tables, and the arrangement of text components followed a top-down approach to recognise table structures (Ramel et al., 2003; Hassan and Baumgartner, 2007). These approaches work well with simple tables, but struggle with complex table structures.

Machine learning-based methods are widely used for the TSR task. Early methods involved statistical machine learning techniques, such as using Support Vector Machines (SVM) to classify tables based on line information (Kasar et al., 2013), or clustering word segments in a bottom-up manner (Kieninger and Dengel, 1999). Recently, with the availability of large datasets, deep learning methods have been preferred. One common approach considers TSR as an object detection task, employing well-known detection frameworks such as Faster R-CNN (Girshick, 2015), Mask R-CNN (He et al., 2017), and YOLO (Redmon and Farhadi, 2018). Another approach frames TSR as an image-to-sequence task using transformer-based encoder-decoder methods (Khang and Hong, 2024), for example, applying Convolutional Neural Networks (CNN) as the encoder for image feature representation and Recurrent Neural Networks

(RNN) as the decoder for structure sequence generation (Li et al., 2020a), or using vision transformers for TSR (Nassar et al., 2022; Chen et al., 2023). Graph Neural Networks (GNN) have also been applied to TSR, leveraging text cells as graph vertices and employing graph attention mechanisms to generate their representations (Xue et al., 2019; Chi et al., 2019). More recently, Vision Large Language Models (VLLMs) (Zhou et al., 2025) have been explored for TSR as well.

3 Methodology

We consider the TSR task as an image-to-sequence generation task. We propose a framework that uses vision and text transformer as two-stream encoders, with the fused multi-modal feature representation for sequence generation through two decoders. The model generates a machine-processable sequence **S** from a given table image **I**. The generated sequence **S** includes the table structure **T** = [t_1, ..., t_n], and the non-empty table cell bounding box **B** = [b_1, ..., b_m]. The table cell contents **C** = [$c_1, ...c_m$] are obtained using an off-the-shelf OCR (Smith, 2007). The table cell contents correspond to the table bounding boxes, but may differ from the table structure sequence due to empty cells in the table. The table structure is represented using HTML tags, which can be converted into various formats depending on the requirements.

3.1 Encoder

We use two stream encoders to extract visual and textual features, aiming to obtain better cross-modal representations from table images. For the visual encoder, inspired by ViT (Dosovitskiy et al., 2021), the input table image is resized and split into non-overlapping P x P patches, which are then reshaped into flattened 2D patches. These patches are linearly projected into a D-dimensional sequence, serving as the input to a stack of transformer encoder layers. The final output is encoded visual sequence features of the table image. The textual encoder follows the approach of Roberta (Liu et al., 2019). It takes word embeddings of the table's textual content as input. The global tokens [CLS] and [SEP] are added at the beginning and the end of each text sequence, and [PAD] tokens are appended to the end to match the maximum sequence length L. The textual encoder outputs the textual representation. Finally, the outputs of both encoders are integrated using an element-wise sum. This allows the model to learn the complex relationships between visual and textual features to obtain contextual text-and-image representations.

3.2 Decoder

The decoder is built on a standard transformer decoder that takes embedded features from the fused encoder outputs. It consists of a stack of four decoder layers, each containing multi-head attention and feed-forward layers. We employ separate decoders with the same architecture to decode the table structure and the table cell bounding boxes. The structure decoder generates HTML tags representing the table structure, including starting tags such as <thead>, <tbody>, <tr>, etc. The bounding box decoder generates coordinates for each non-empty table cell in the format $[x_{min}, y_{min}, x_{max}, y_{max}]$. We apply teacher forcing during model training and use beam search for inference.

Since the pre-trained vision encoder is not trained on table images, we continue to train it with the TableBank dataset (Li et al., 2020b), along with the aligned table text encoder, to enhance table feature representation. Masked image modeling (Bao et al., 2022) is applied to the visual encoder during pre-training. We fine-tune the entire TSR model during the fine-tuning process.

4 Experimental Setup

The pre-trained Swin-tiny transformer (Liu et al., 2021) is used for visual embedding initialisation, and the text embedding is initialised from Roberta (Liu et al., 2019). We use Adam optimiser (Kingma and Ba, 2015) with an initial learning rate of $2e{-}5$, which decays by 0.02 after the 3rd epoch. We trained the encoder for 10 epochs with a batch size of 16. The decoder includes 4 layers with an input feature size of 512 and 4 attention heads for table structure and cell bounding box decoding. Similar to the encoder, the decoder uses the Adam optimiser but with an initial learning rate of $2e{-}4$, trained for 10 epochs with a batch size of 16. We use Tesseract OCR [1] to obtain table cell content from the table image.

4.1 Datasets

We evaluate our approach on three benchmark datasets for the TSR task.

PubTabNet (Zhong et al., 2020) contains 509k table images extracted from scientific literature and

[1] https://github.com/tesseract-ocr/tesseract

provides annotation for table structure in HTML format, table cell bounding boxes, and table cell content. This dataset also provides evaluation metrics such as Tree-edit-distance-based similarity (TEDS) for both table structure and table cell content evaluation. We use the validation dataset as the test dataset since the test dataset is not available.

FinTabNet (Zheng et al., 2021) is created from the annual reports of the S&P 500 companies in PDF format. It includes 113k table images from 1,600 different types of financial tables and is annotated for table structure (in HTML), table cell bounding box, and table cell content. This dataset is reviewed manually, making it more reliable.

SciTSR (Chi et al., 2019) contains 15k tables extract from scientific PDF files. It provides corresponding structure labels obtained from LaTeX source files. The dataset is split into 12k tables for training and 3k tables for testing. Because SciTSR does not provide tables in HTML format, we convert the structure labels into HTML for S-TEDS evaluation. We use the bounding box coordinates to recover the logical row and column layout, place each cell into the correct position in a two-dimensional grid, and produce an HTML table that reflects the original structure.

4.2 Evaluation Metrics

For evaluation, we use Intersection over Union (IoU) with COCO average precision (AP) (Lin et al., 2014) to measure the overlap between ground truth and predicted bounding boxes. The AP_{50} is reported as the evaluation result for table cell bounding box detection. The structure-only Tree-Edit-Distance-Based Similarity or S-TEDS (**?**) is used for table structure-based evaluation. It converts table HTML tags into a tree structure and measures the edit distance between the prediction and ground-truth tree structures. Higher similarity corresponds to a shorter edit distance, leading to a higher TEDS score.

5 Experimental Results

We compared our models with six baselines—Cascade R-CNN (Cai and Vasconcelos, 2018), Deformable-DETR (Zhu et al., 2021), TSRDet (Xiao et al., 2025), VAST (**?**), TABLET (Hou and Wang, 2025), and NGTR (Zhou et al., 2025)—on three TSR task-related benchmark datasets (PubTabNet, FinTabNet, and SciTSR), using AP_{50} and S-TEDS metrics. We utilised structure-based S-TEDS as the primary evaluation metric to avoid the noise of table cell content that is generated by OCR. Our multi-modal approach outperformed almost all visual-only baseline methods and achieved highly competitive results on table structure recovery, as shown in Table 1. In particular, the multi-modal approach showed a clear improvement in S-TEDS compared with the vision-only Deformable-DETR, which suggests that using text information helps the model better handle confusing layouts and cells that look similar in table images. The ablation study on FinTabNet (Table 2) demonstrates that incorporating the visual modality leads to a significant gain in S-TEDS (+5.4), indicating that visual and textual features work together and complement each other for TSR. We note that our approach also outperforms the VLLM approach (NGTR) (Zhou et al., 2025) (Table 3) as per reported results on the same datasets.

Model	**Dataset**	**AP_{50}**	**S-TEDS(%)**
Cascade R-CNN	PubTabNet	95.38	83.78
Deformable-DETR	PubTabNet	97.43	95.73
TSRDet	PubTabNet	**98.26**	96.58
VAST	PubTabNet	94.80	97.23
TABLET	PubTabNet	—	97.67
Ours	PubTabNet	97.90	**97.69**
Cascade R-CNN	FinTabNet	97.53	87.49
Deformable-DETR	FinTabNet	98.42	97.81
TSRDet	FinTabNet	98.33	**99.05**
VAST	FinTabNet	96.20	98.63
TABLET	FinTabNet	—	98.99
Ours	FinTabNet	**98.97**	98.96
Cascade R-CNN	SciTSR	95.27	79.09
Deformable-DETR	SciTSR	97.39	97.30
TSRDet	SciTSR	96.79	98.41
Ours	SciTSR	**98.32**	**98.52**

Table 1: Comparing our method with baselines on PubTanNet, FinTanNet, and SciTSR datasets.

Model	**AP_{50}**	**S-TEDS(%)**
Swin-T	92.36	93.56
Ours	**98.97**	**98.96**

Table 2: Ablation results for vision-only and multi-modal approaches on the FinTabNet dataset.

6 Conclusions

We present a multi-modal approach with two stream encoders and separate decoders for the Table Structure Recognition (TSR) task. The pro-

Model	Dataset	S-TEDS(%)
NGTR	PubTabNet	92.31
Ours	PubTabNet	**97.69**
NGTR	SciTSR	95.78
Ours	SciTSR	**98.52**

Table 3: Comparing our TSR method and reported results on VLLMs (Zhou et al., 2025).

posed model integrates features from both visual and textual modalities, generating table structure and table cell bounding boxes simultaneously. Our experimental results on three different datasets from scientific and financial domains show that the effectiveness of the proposed model is competitive compared to visual-only approaches.

7 Limitations

The proposed multi-modal approach demonstrated its effectiveness with regular table images, but it is worthwhile to further explore irregular table images in real-world scenarios, such as table images from scanned books, wired tables in the wild, and handwritten tables. Meanwhile, training a unified framework to integrate all sub-tasks of TSR (table structure, table cell bounding boxes, and table cell content) also presents opportunities for exploration.

References

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. Beit: BERT pre-training of image transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162.

Leiyuan Chen, Chengsong Huang, Xiaoqing Zheng, Jinshu Lin, and Xuan-Jing Huang. 2023. "tablevlm: Multi-modal pre-training for table structure recognition". In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 2437–2449.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020a. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Zhiyu Chen, Mohamed Trabelsi, Jeff Heflin, Yinan Xu, and Brian D Davison. 2020c. Table search using a deep contextualized language model. In *SIGIR*, pages 589–598.

Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. 2019. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document ai: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Björn Engelmann, Timo Breuer, and Philipp Schaer. 2023. Simulating users in interactive web table retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3875–3879.

Ross Girshick. 2015. "fast r-cnn". In *ICCV*, pages 1440–1448.

Vivek Gupta, Pranshu Kandoi, Mahek Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Srikumar. 2023. TempTabQA: Temporal question answering for semi-structured tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2431–2453, Singapore. Association for Computational Linguistics.

Tamir Hassan and Robert Baumgartner. 2007. Table recognition and understanding from pdf files. In *Ninth International Conference on Document Analysis and Recognition*, volume 2, pages 1143–1147.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. "mask r-cnn". In *ICCV*, pages 2961–2969.

Qiyu Hou and Jun Wang. 2025. Tablet: Table structure recognition using encoder-only transformers. In *Proceedings of the 19th International Conference on Document Analysis and Recognition*.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1821–1831.

Thotreingam Kasar, Philippine Barlas, Sebastien Adam, Clément Chatelain, and Thierry Paquet. 2013. Learning to detect tables in scanned document images using line information. In *12th International Conference on Document Analysis and Recognition*, pages 1185–1189.

Minsoo Khang and Teakgyu Hong. 2024. Tflop: table structure recognition framework with layout pointer mechanism. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 947–955.

Thomas Kieninger and Andreas Dengel. 1999. The t-recs table recognition and analysis system. In *Document Analysis Systems: Theory and Practice: Third IAPR Workshop, DAS'98 Nagano, Japan, November 4–6, 1998 Selected Papers 3*, pages 255–270.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2020a. Tablebank: Table benchmark for image-based table detection and recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1918–1925.

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2020b. TableBank: Table benchmark for image-based table detection and recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1918–1925, Marseille, France. European Language Resources Association.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022.

Nam Tuan Ly and Atsuhiro Takasu. 2023. An end-to-end multi-task learning model for image-based table recognition. *arXiv preprint arXiv:2303.08648*.

Ahmed Nassar, Nikolaos Livathinos, Maksym Lysak, and Peter Staar. 2022. Tableformer: Table structure understanding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4614–4623.

Liang Qiao, Zaisheng Li, Zhanzhan Cheng, Peng Zhang, Shiliang Pu, Yi Niu, Wenqi Ren, Wenming Tan, and Fei Wu. 2021. LGPMA: complicated table structure recognition with local and global pyramid mask alignment. In *International conference on document analysis and recognition*, pages 99–114.

J-Y Ramel, Michel Crucianu, Nicole Vincent, and Claudie Faure. 2003. Detection, extraction and representation of tables. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, pages 374–378.

Joseph Redmon and Ali Farhadi. 2018. "yolov3: An incremental improvement". In *In Computer vision and pattern recognition (Vol. 1804)*, pages 1–6.

Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. 2017. Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *The 14th IAPR international conference on document analysis and recognition*, volume 1, pages 1162–1167.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633, Washington, DC, USA. IEEE Computer Society.

Bin Xiao, Murat Simsek, Burak Kantarci, and Ala Abu Alkheir. 2025. Rethinking detection based table structure recognition for visually rich document images. *Expert Systems with Applications*, 269:126461.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 602–631, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xia Xie, Yu Fu, Hai Jin, Yaliang Zhao, and Wenzhi Cao. 2020. A novel text mining approach for scholar information extraction from web content in Chinese. *Future Generation Computer Systems*, 111:859–872.

Hangdi Xing, Feiyu Gao, Rujiao Long, Jiajun Bu, Qi Zheng, Liangcheng Li, Cong Yao, and Zhi Yu. 2023. Lore: logical location regression network for table structure recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2992–3000.

Wenyuan Xue, Qingyong Li, and Dacheng Tao. 2019. Res2tim: Reconstruct syntactic structures from table images. In *The International Conference on Document Analysis and Recognition*, pages 749–755.

Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2023. Structextv2: Masked visual-textual prediction for document image pre-training. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Xinyi Zheng, Doug Burdick, Lucian Popa, Peter Zhong, and Nancy Xin Ru Wang. 2021. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. *Winter Conference for Applications in Computer Vision (WACV)*.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: Data, model, and evaluation. In *ECCV*, page 564–580.

Yitong Zhou, Mingyue Cheng, Qingyang Mao, Jiahao Wang, Feiyang Xu, and Xin Li. 2025. Enhancing table recognition with vision llms: A benchmark and neighbor-guided toolchain reasoner. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 2503–2511. International Joint Conferences on Artificial Intelligence Organization.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.