OMNIPARSER: A Unified Framework for Text Spotting, Key Information Extraction and Table Recognition (Supplementary Materials)

1. Implementation Details

1.1. Spatial-Window Prompting

Spatial-window prompting comprises two components: fixed mode and random mode. In the fixed mode, the image is divided into grid blocks evenly, such as 3x3 or 2x2. Conversely, in the random mode, the starting point of spatial window is randomly determined. In order to encompass more texts within the random box, the area of the random box is established to be no less than 1/9 of the original image. To elaborate further, a 30% probability is assigned for selecting the fixed mode, another 30% probability for selecting the random mode, and a 40% probability for defaulting window to cover the entire image. Following [6], we set the bin size of coordinate vocab as 1000. The pseudo-code of spatial-window prompting is shown in the following.

```
import random

# prob for different mode
prob = random.uniform(0, 1)

# quantizing coordinates with n_bins
n_bins = 1000

if prob < 0.4:
    # default window
    start_x, start_y, end_x, end_y = [0, 0, n_bins - 1, n_bins - 1]
elif prob < 0.7:
    # x-axis and y-axis are partitioned into varying numbers of blocks.
    num_xs = [3, 3, 1, 3, 2, 2, 2, 1]
    num_ys = [3, 1, 3, 2, 3, 2, 1, 2]

    total_windows = []
    for num_x, num_y in zip(num_xs, num_ys):
        inter_x = min(int(n_bins / num_x), n_bins - 1)
        inter_y = min(int(n_bins / num_y), n_bins - 1)

        for i in range(num_x):
            for j in range(num_y):
                start_x = i*inter_x
                start_y = j*inter_y
                end_x = min(start_x + inter_x, n_bins - 1)
                end_y = min(start_y + inter_y, n_bins - 1)
                total_windows.append([start_x, start_y, end_x, end_y])

    start_x, start_y, end_x, end_y = random.choice(total_windows)
else:
    inter = int(n_bins / 3)
    start_x = random.randint(0, inter * 2)
    start_y = random.randint(0, inter * 2)
    rect_w, rect_h = random.randint(inter, n_bins - 1), random.randint(inter, n_bins - 1)
    end_x, end_y = min(start_x + rect_w, n_bins - 1), min(start_y + rect_h, n_bins - 1)

spatial_window_prompt = [start_x, start_y, end_x, end_y]
```

1.2. Table Recognition

Given a table image, we resize it to 1,024×1,024 pixels. The Structured Points Decoder, utilizing the feature vector from the Image Encoder, simultaneously generates pure HTML tags with structural cell point sequences in the same sequence representing the table's logical and physical structures. These structural cell point sequences serve as start-prompting input for the Content Decoder, which extracts table cell contents in parallel. The final output combines pure HTML tags with cell contents, forming complete HTML sequences faithfully representing the table's structure and content.

Datasets. Since our model predicts both the logical structure of tables with cell bounding box central points and cell content, datasets lacking cell content and corresponding bounding box annotations, such as TABLE2LATEX-450K [2], TableBank [8], UNLV [12], IC19B2H [3], WTW [9] and TUCD [11], are not suitable for our approach. Similarly, datasets like ICDAR2013Table [4], SciTSR [1], and PubTables-1M [13], which provide cell content and content box annotations, employ metrics based on box representations that are incompatible with our point-based format. Consequently, PubTabNet (PTN) [16] and FinTabNet (FTN) [15] are selected for our model evaluation.

GT Generation. The ground truth pure HTML tags of tables are tokenized into structural tokens. Following the previous works [5, 14], we use the merged labels to represent a non-spanning cell to reduce the length of the HTML tags. Specifically, we use *<td></td>* and *<td>[]</td>* to denote empty cells and non-empty cells, respectively. For a cell spanning multiple rows or columns, the original HTML tags are broken into four tokens: *<td*, *colspan="n"* or *rowspan="n"*, *>*, and *</td>*. We use the first token *<td* to represent a spanning cell. In addition, four special symbol categories need to be added: *<S>*, *</S>*, *<PAD>*, and *<UNK>*, which represents the beginning of a sequence, the end of a sequence, padding symbols, and unknown characters, respectively. For building the GT of Structured Points Decoder, we insert center points of each cell text box to corresponding HTML tags. For building the GT of Content Decoder, we combine each cell text with corresponding center points as a whole sequence where center points can be viewed as a start-prompting input for recognizing text, and each cell text is tokenized at the character level. An example of building a training sequence GT for the Structured Points Decoder and the Content Decoder in the table recognition task is illustrated in Fig. 1.

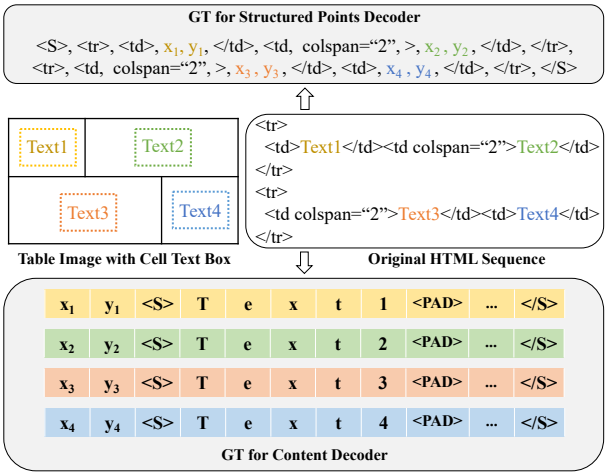


Figure 1. **An Example of building training GTs for table recognition task.** We use the center points of each cell text box to build GTs for the Structured Points Decoder and the Content Decoder. If the cell is empty text, the corresponding points in the GTs are left empty as well.

2. Comparisons with Donut on KIE Task

As shown in Fig. 2, OMNIPARSER can achieve entity extraction while predicting the location of each entity word. However, Donut only predicts the structured sequence for entity extraction without any localization ability. Thus, the absence of direct region supervision during both training and prediction stages often leads to inferior results for entities of same values (Row 1), repeated entities (Row 2) or entities with explicit trigger names (Row 3).

3. Training Donut on Table Recognition Task

We fine-tuned the OCR-free end-to-end model Donut [7] for table recognition on FinTabNet dataset. The ground truth sequence utilized combined HTML tags with table cell text, and we use different training hyper-parameters for adequate verification, as shown in Tab. 1. Due to GPU memory limitations, we constrained the decoder's max length in Donut to 4,000. Note that the original HTML sequence max lengths for PubTabNet and FinTabNet are 8,722 and 8,035, respectively. For long sequence prediction tasks such as table recognition, training an end-to-end model like Donut with combined HTML stages, including cell text, is non-trivial. There is a high probability of error accumulation and attention drift in long-sequence scenarios leading to the inferior performance of Donut for table recognition. An illustrative example of a failure case for Donut in table recognition task is shown in Fig. 3. Specifically, due to the lack of region supervision, the end-to-end model Donut has demonstrated an attention drift problem, resulting in the prediction of repeated tokens and leading to a high probability of error accumulation in long-sequence scenarios. In contrast, OMNIPARSER decomposes the location-aware structured points sequence and cell text recognition generation, alleviating the issues of attention drift and error accumulation.

Methods	LR	Epoch	S-TEDS	TEDS
Donut [7]	3e-5	20	22.2	17.2
	3e-5	40	26.2	20.0
	1e-4	40	30.7	29.1
	1e-3	40	41.7	40.5
	1e-3	100	41.9	41.2
OMNIPARSER (ours)	-	-	**91.55**	**89.75**

Table 1. **Comparisons of different training hyper-parameters of Donut on FinTabNet datasets.** LR is short for learning rate.

4. Generalization to Hierarchical Text Detection Task

Thanks to the flexible expression of structured sequence in OMNIPARSER, it is convenient for us to extend it to other OCR-related tasks, such as hierarchical text detection, which aims to group the text in the image into three levels, namely word, line, and paragraph, based on spatial position and semantic relationship. Previous methods [10] mainly achieved hierarchical results by clustering based on similarity. In our approach, we distinguish the text belonging to different hierarchical intervals by simply inserting *<LINE>* and *<PARA>* structural tags into the sequence of text center

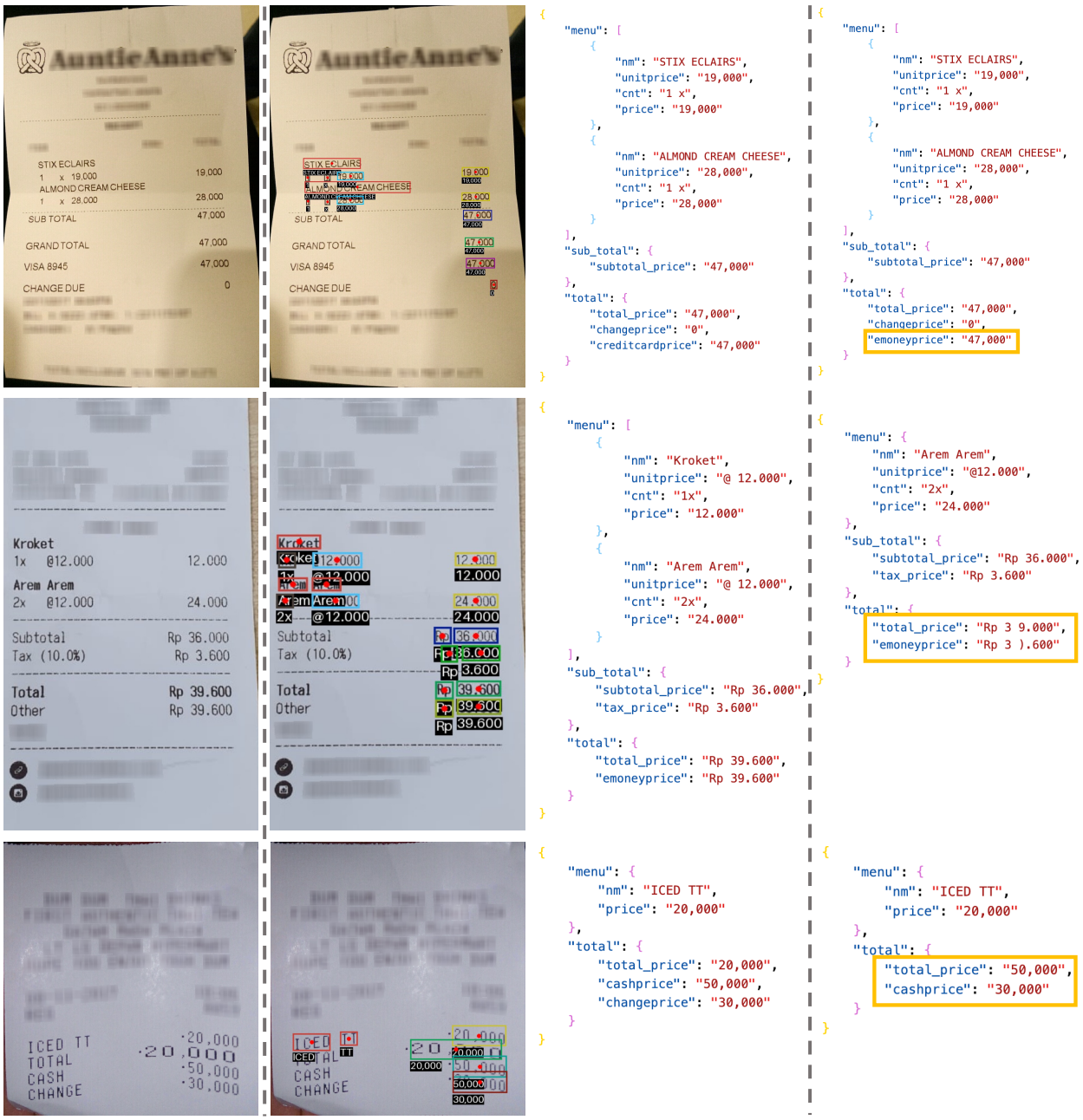


Figure 2. **A comparative analysis of partial results obtained from** OMNIPARSER **and Donut on CORD.** The first column depicts the original image, while columns 2 and 3 illustrate our detection results and the corresponding formatted output, respectively. Column 4 showcases the Donut's formatted output. Notably, our model demonstrates superior performance in entity extraction.

points, as shown in Fig. 4. The experiments are mainly conducted on the HierText dataset [10], which consists of 8,281 training images, 1,724 validation images, and 1,634 test images. We train the model on the training set and evaluate on the validation set. Partial visualization results are shown in Fig. 5. Without any task-specific architectural designs, our model achieves promising results, demonstrating its strong generalization ability.

5. More Visualizations

Fig. 6, Fig. 7, and Fig. 8 are more qualitative results of text spotting, key information extraction, and table recognition, respectively.

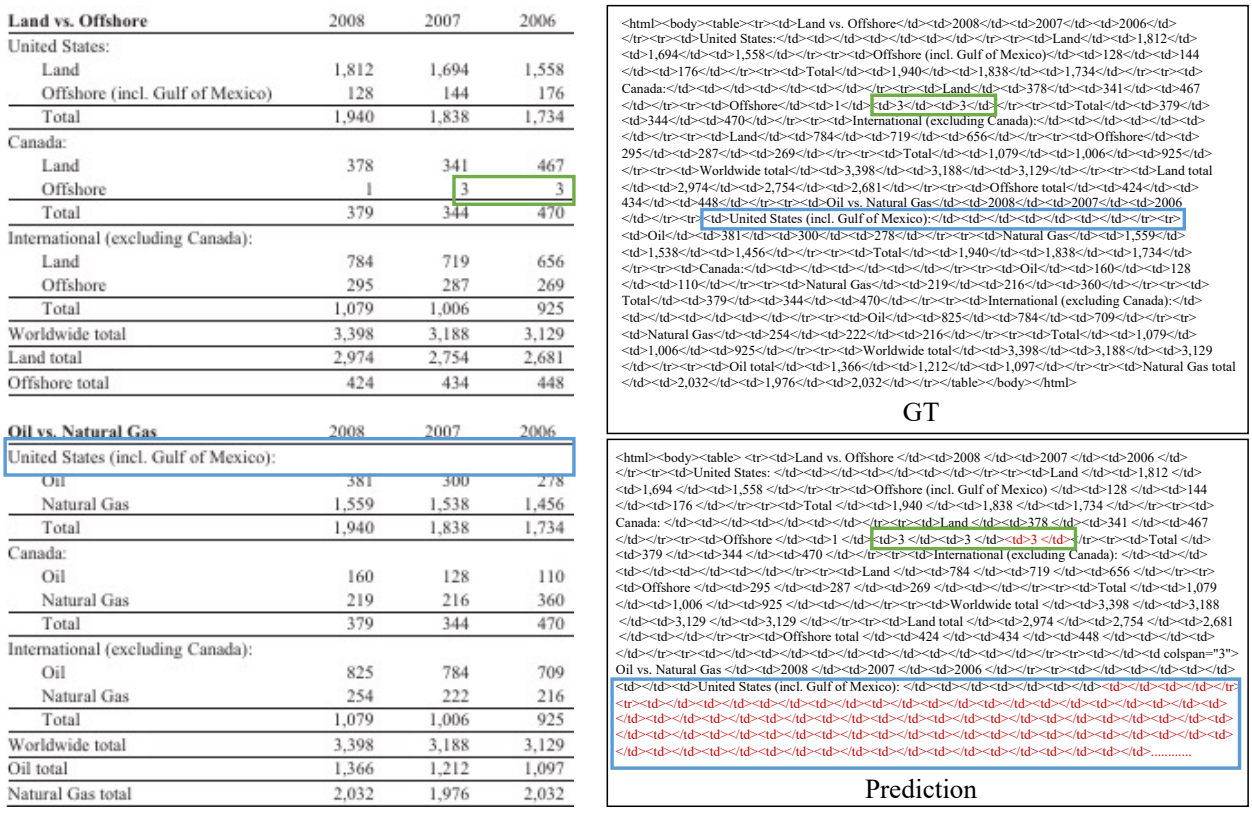


Figure 3. **Illustrative failure case of Donut in table recognition task.** Red text means error predictions. For readability, we only highlight two errors in this example. Due to the lack of point location information, Donut has an attention drift problem, resulting in the prediction of repeated tokens and leading to a high probability of error accumulation in long-sequence scenarios. (The figure is best viewed in color.)

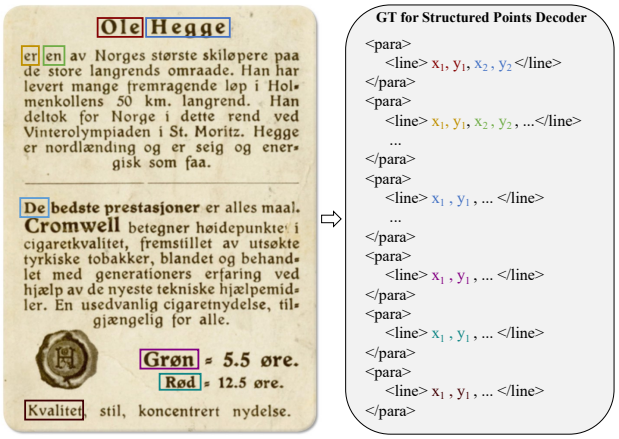


Figure 4. **An Example of building training GTs for hierarchical text detection task.**

References

[1] Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*, 2019. 1

[2] Yuntian Deng, David Rosenberg, and Gideon Mann. Challenges in end-to-end neural scientific table recognition. In *ICDAR*. IEEE, 2019. 1

[3] Liangcai Gao, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Lang. Icdar 2019 competition on table detection and recognition (ctdar). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1510–1515. IEEE, 2019. 1

[4] Max Göbel, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. Icdar 2013 table competition. In *ICDAR*, pages 1449–1453. IEEE, 2013. 1

[5] Yongshuai Huang, Ning Lu, Dapeng Chen, Yibo Li, Zecheng Xie, Shenggao Zhu, Liangcai Gao, and Wei Peng. Improving table structure recognition with visual-alignment sequential coordinate modeling. In *CVPR*, pages 11134–11143, 2023. 2

[6] Taeho Kil, Seonghyeon Kim, Sukmin Seo, Yoonsik Kim, and Daehee Kim. Towards unified scene text spotting based on sequence generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15223–15232, 2023. 1

[7] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sang-

Figure 5. **Visualization results of hierarchical text detection.** Columns 1-3 represent the detection results for word, line, and paragraph levels, respectively. Text instances belonging to the same hierarchical level are enclosed within rectangles of the same color. (The figure is best viewed in color.)

Figure 6. **Visualization results of text spotting.** Rows 1-2 depict the visual results on the Total-Text dataset, while rows 3 and 4 respectively illustrate the visual results on the ICDAR 2015 and CTW1500 datasets. (The figure is best viewed in color.)

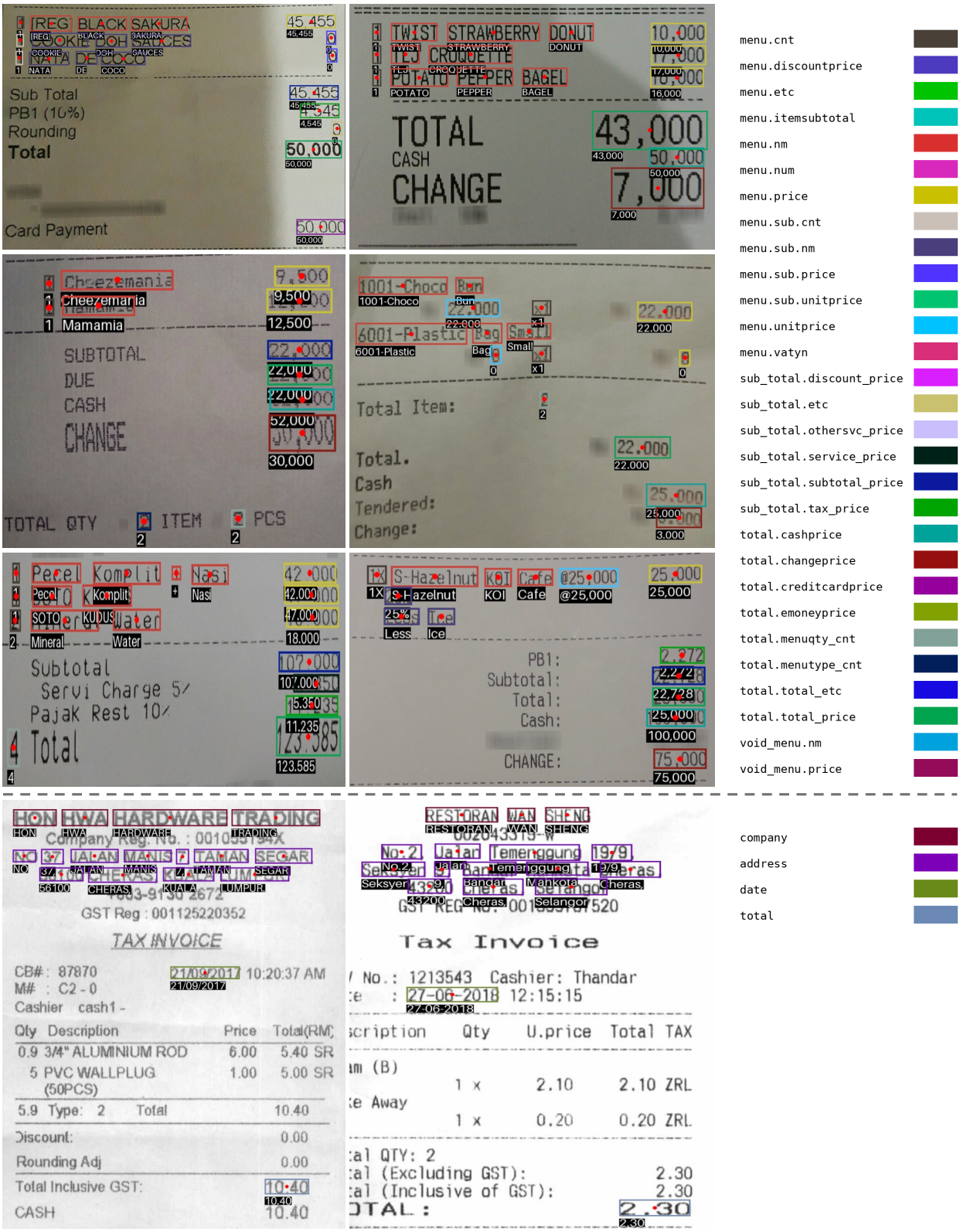
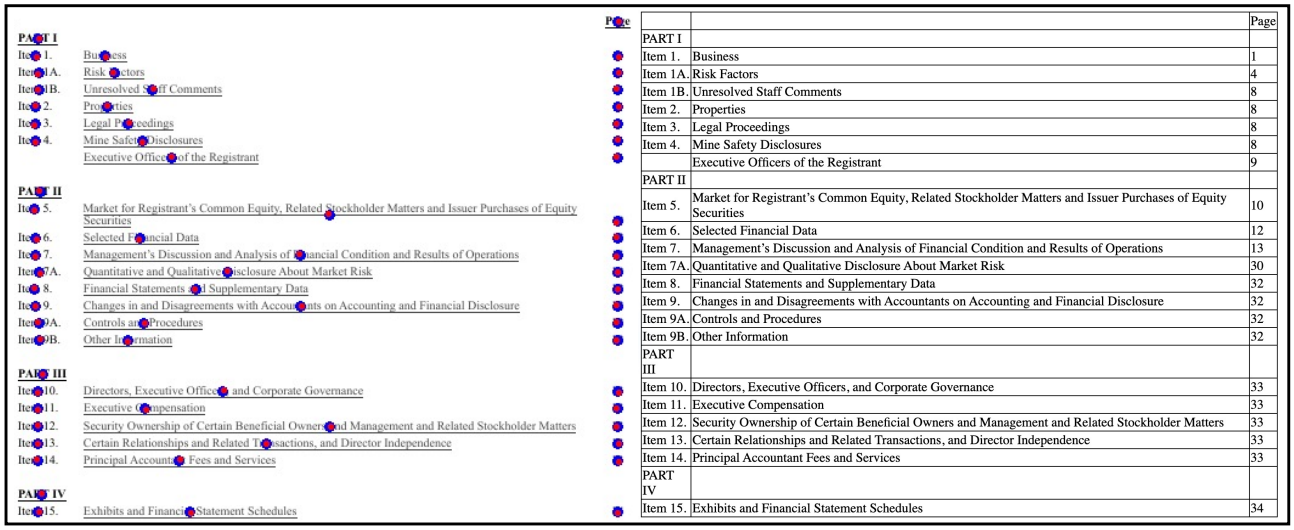


Figure 7. **Visualization results of key information extraction.** Rows 1-3 and row 4 demonstrate the visual results on the CORD and SROIE datasets respectively. In order to differentiate entities of different categories, we employ rectangles of varying colors. The correspondence between colors and categories can be seen in the legend on the right side. (The figure is best viewed in color.)



		Page
PART I		
Item 1.	Business	1
Item 1A.	Risk Factors	4
Item 1B.	Unresolved Staff Comments	8
Item 2.	Properties	8
Item 3.	Legal Proceedings	8
Item 4.	Mine Safety Disclosures	8
	Executive Officers of the Registrant	9
PART II		
Item 5.	Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities	10
Item 6.	Selected Financial Data	12
Item 7.	Management's Discussion and Analysis of Financial Condition and Results of Operations	13
Item 7A.	Quantitative and Qualitative Disclosure About Market Risk	30
Item 8.	Financial Statements and Supplementary Data	32
Item 9.	Changes in and Disagreements with Accountants on Accounting and Financial Disclosure	32
Item 9A.	Controls and Procedures	32
Item 9B.	Other Information	32
PART III		
Item 10.	Directors, Executive Officers, and Corporate Governance	33
Item 11.	Executive Compensation	33
Item 12.	Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters	33
Item 13.	Certain Relationships and Related Transactions, and Director Independence	33
Item 14.	Principal Accountant Fees and Services	33
PART IV		
Item 15.	Exhibits and Financial Statement Schedules	34

	2010	2009	2008	2007	2006
Cash and current marketable securities	4,380.1	2,954.8	2,195.6	2,410.8	1,414.8
Working capital	6,026.4	4,410.2	3,517.2	3,571.9	2,182.8
Current ratio	4.8	4.1	3.4	3.7	2.6
Property, plant and equipment—net	798.3	947.6	963.8	991.6	914.9
Capital expenditures	182.1	131.3	155.2	187.7	209.4
Depreciation and amortization	410.2	385.3	387.6	366.6	324.1
Total assets	10,895.1	9,071.3	7,603.3	7,354.0	5,873.8
Long-term debt, including current maturities	1,021.8	18.0	20.5	16.8	14.8
Shareholders' equity	7,173.6	6,595.1	5,406.7	5,378.5	4,191.0
Return on average equity	18.5%	18.5%	21.3%	21.3%	20.8%
Net cash provided by operating activities	1,547.4	1,460.7	1,175.9	1,028.3	867.3
Number of shareholders of record	4,586	4,607	4,500	4,373	4,091
Number of employees	20,036	18,582	17,594	16,026	18,806

Land vs. Offshore	2007	2006	2005
United States:			
Land	1,694	1,558	1,287
Offshore	73	90	93
Total	1,767	1,648	1,380
Canada:			
Land	341	467	454
Offshore	3	3	4
Total	344	470	458
International (excluding Canada):			
Land	719	656	593
Offshore	287	269	258
Total	1,006	925	851
Worldwide total	3,117	3,043	2,689
Land total	2,754	2,681	2,334
Offshore total	363	362	355
Oil vs. Gas	2007	2006	2005
United States:			
Oil	297	273	194
Gas	1,470	1,375	1,186
Total	1,767	1,648	1,380
Canada:			
Oil	128	110	100
Gas	216	360	358
Total	344	470	458
International (excluding Canada):			
Oil	784	709	651
Gas	222	216	200
Total	1,006	925	851
Worldwide total	3,117	3,043	2,689
Oil total	1,209	1,092	945
Gas total	1,908	1,951	1,744

	2014	2013	2012
Operating income	$603.0	$550.5	$578.3
Impact of special charges and loss on voluntary pension settlement	5.2	40.3	—
Adjusted operating income	$608.2	$590.8	$578.3
% increase versus prior year	2.9%	2.2%	7.0%
Net income	$437.9	$389.0	$407.8
Impact of special charges and loss on voluntary pension settlement	3.7	29.2	—
Adjusted net income	$441.6	$418.2	$407.8
% increase versus prior year	5.6%	2.6%	9.0%
Earnings per share—diluted	$3.34	$2.91	$3.04
Impact of special charges and loss on voluntary pension settlement	0.03	0.22	—
Adjusted earnings per share—diluted	$3.37	$3.13	$3.04
% increase versus prior year	7.7%	3.0%	9.0%

Figure 8. **Visualization results of table recognition.** We present point locations and a rendered table with an additional border for readability based on the prediction sequence in each group. Blue points and red points denote the GT and predicted points respectively. (The figure is best viewed in color.)

doo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, pages 498–517. Springer, 2022. 2

[8] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. Tablebank: Table benchmark for image-based table detection and recognition. In *Proceedings of*

the Twelfth Language Resources and Evaluation Conference, pages 1918–1925, 2020. 1

[9] Rujiao Long, Wen Wang, Nan Xue, Feiyu Gao, Zhibo Yang, Yongpan Wang, and Gui-Song Xia. Parsing table structures in the wild. In *ICCV*, pages 944–952, 2021. 1

[10] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2022. 2, 3

[11] Sachin Raja, Ajoy Mondal, and CV Jawahar. Visual understanding of complex table structures from document images. In *WACV*, pages 2543–2552, 2022. 1

[12] Asif Shahab, Faisal Shafait, Thomas Kieninger, and Andreas Dengel. An open approach towards the benchmarking of table structure recognition systems. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 113–120, 2010. 1

[13] Brandon Smock, Rohith Pesala, and Robin Abraham. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *CVPR*, pages 4634–4642, 2022. 1

[14] Jiaquan Ye, Xianbiao Qi, Yelin He, Yihao Chen, Dengyi Gu, Peng Gao, and Rong Xiao. Pingan-vcgroup's solution for icdar 2021 competition on scientific literature parsing task b: table recognition to html. *arXiv preprint arXiv:2105.01848*, 2021. 2

[15] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *WACV*, pages 697–706, 2021. 1

[16] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *ECCV*, pages 564–580. Springer, 2020. 1