Supplementary Material: Towards Unified Scene Text Spotting based on Sequence Generation

1. Implementation Details

1.1. The Starting-Point Prompting

Our proposed method applies starting-point prompting to enable the model to extract texts from an arbitrary starting point, allowing it to generate a longer sequence than the maximum decoding length. To apply starting-point prompting, we use raster scan order as the order of text instance extraction. The text extraction process is shown in Fig. 1. Text instances with central points within the search region are extracted in raster scan order. Furthermore, during testing, if the generated output sequence does not end with an <eos> token, our method sets the starting point as the last detected text position in the previous step, then continues and enables the re-generation of a text sequence corresponding to the remaining objects, as shown in Fig. 2.

1.2. Multi-way Decoder

Our model employs a multi-way transformer decoder, as shown in Fig. 3. The detection and recognition feedforward networks (FFNs) are separated, and attention layers are shared between them. When generating a detection token, the model is required to pass it through the detection FFN, and when it is time to generate a recognition token, the model must pass it through the recognition FFN. This separation enables the model to better learn multiple detection formats simultaneously, and the shared attention modules can learn from both tasks, improving the overall performance.

2. More Experiments

2.1. Word Spotting Evaluation on Benchmarks

In the main paper, we reported the end-to-end evaluation scores on the ICDAR 2015 and Total-Text datasets. In text spotting, there are two evaluation protocols: end-to-end and word-spotting. In this section, we report both word spotting and end-to-end scores on these two benchmark datasets in Tabs. 1 and 2. End-to-end results are repeated for comparison with word spotting. We confirm that the proposed method shows state-of-the-art performance for all vocabularies in both evaluation protocols.

2.2. Evaluation on CTW-1500, ICDAR 2013, and Rotated ICDAR 2013

To demonstrate the superiority of our method, further evaluation is conducted on other benchmark datasets: CTW-1500 [8], ICDAR 2013 [3], and Rotated ICDAR 2013 [6]. CTW-1500 includes arbitrary-shaped texts and contains 1,000 training and 500 testing images. Unlike the other datasets, CTW-1500 is annotated in not word-level but text line-level annotations. Since this setting is different from the unified model learned with word-level annotation, we fine-tune the model with the CTW-1500 dataset and only measure the performance of the fine-tuned model. In fine-tuning, we set the maximum length of each text transcription to 100. The model is fine-tuned for 20k steps with fixed learning rate $3e^{-5}$ from unified model UNITS$_{\text{Shared}}$. For evaluating CTW-1500, we predict 16-point polygons, and for ICDAR 2013, we use both bounding box and quadrilaterals. For Rotated ICDAR 2013, we use 4-point quadrilaterals. Tabs. 3 to 5 show that our method achieves competitive results with the existing methods. Some qualitative results are shown in Fig. 4.

2.3. Evaluation on TextOCR

The proposed method can detect and recognize more text instances than the maximum number of instances allowed by the decoder length. To demonstrate this effect, we evaluate on TextOCR dataset, which contains a relatively large number of texts. Since TextOCR does not provide annotations for the test sets, the performance evaluation is performed for validation sets. We did not use validation sets at all in training or model selection. Similar to the experiments in the main paper, we fine-tune the model for TextOCR and report the results of both the unified and fine-tuned models. The model is fine-tuned for 150k steps with a fixed learning rate of $3e^{-5}$ from the unified model UNITS$_{\text{Shared}}$. The detection and end-to-end scores are reported in Tab. 6. Some qualitative results are shown in Figs. 5 and 6

References

[1] Youngmin Baek, Seung Shin, Jeonghun Baek, Sungrae Park, Junyeop Lee, Daehyun Nam, and Hwalsuk Lee. Character

Method	Word Spotting				End-to-End			
	Strong	Weak	Generic	None	Strong	Weak	Generic	None
CRAFTS [1]	-	-	-	-	83.1	82.1	74.9	-
MaskTextSpotter v3 [6]	83.1	79.1	75.1	-	83.3	78.1	74.2	-
ABCNet v2 [7]	-	-	-	-	82.7	78.5	73.0	-
MANGO [10]	85.2	81.1	74.6	-	85.4	80.1	73.9	-
DEER [4]	-	-	-	-	82.7	79.1	75.6	71.7
SwinTextSpotter [2]	-	-	-	-	83.9	77.3	70.5	-
TESTR [12]	-	-	-	-	85.2	79.4	73.6	65.3
TTS [5]	85.0	81.5	77.3	-	85.2	81.7	77.4	-
GLASS [11]	86.8	82.5	78.8	-	85.3	79.8	74.0	-
UNITS$_{\text{Shared}}$	88.1	84.9	80.7	78.7	88.4	83.9	79.7	78.5
UNITS	**88.8**	**85.2**	**81.5**	**78.8**	**89.0**	**84.1**	**80.3**	**78.7**

Table 1. Experiment results on ICDAR 2015. "Strong", "Weak", "Generic" and "None" represent recognition with each lexicon respectively.

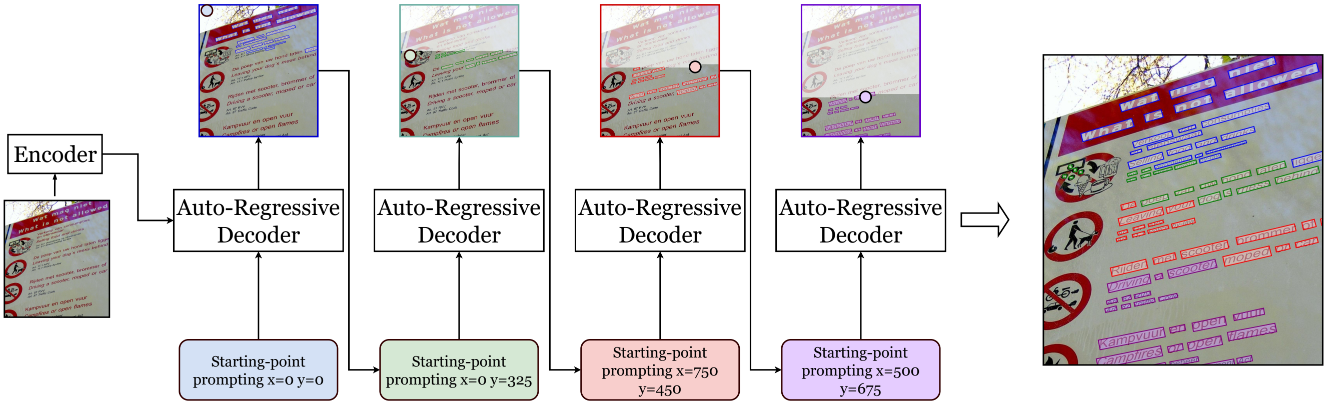

Encoder
Auto-Regressive Decoder
Auto-Regressive Decoder
Auto-Regressive Decoder
Auto-Regressive Decoder
Starting-point prompting x=0 y=0
Starting-point prompting x=0 y=325
Starting-point prompting x=750 y=450
Starting-point prompting x=500 y=675

Figure 1. Illustration of text extraction using the starting-point prompt.

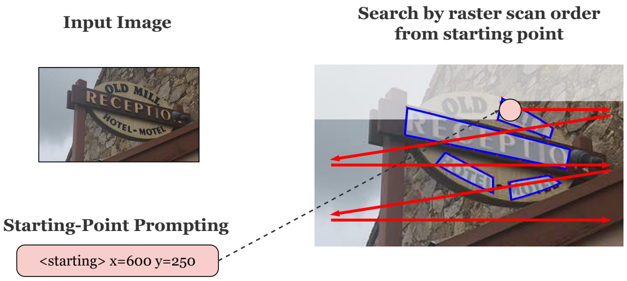
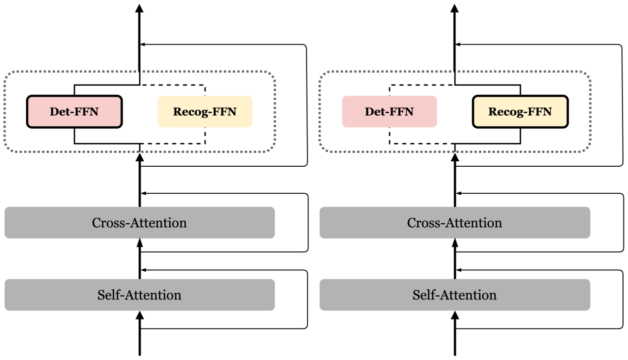


Figure 2. Illustration of the order of text extraction.



(a) Figure of the decoder when generating a detection token.

(b) Figure of the decoder when generating a recognition token.

Figure 3. Illustration of the multi-way transformer decoder.

region attention for text spotting. In *European Conference on Computer Vision*, pages 504–521. Springer, 2020. 2, 3

[2] Mingxin Huang, Yuliang Liu, Zhenghao Peng, Chongyu Liu, Dahua Lin, Shenggao Zhu, Nicholas Yuan, Kai Ding, and Lianwen Jin. Swintextspotter: Scene text spotting via better synergy between text detection and text recognition. In *Pro-*

Method	Word Spotting		End-to-End	
	None	Full	None	Full
CRAFTS [1]	-	-	**78.7**	-
MaskTextSpotter v3 [6]	75.1	81.8	71.2	78.4
ABCNet v2 [7]	-	-	70.4	78.1
MANGO [10]	72.9	83.6	68.9	78.9
DEER [4]	-	-	74.8	83.3
SwinTextSpotter [2]	-	-	74.3	84.1
TESTR [12]	-	-	73.3	83.9
TTS [5]	78.2	86.3	75.6	84.4
GLASS [11]	79.9	86.2	76.6	83.0
$\text{UNITS}_{\text{Shared}}$	81.2	87.0	77.3	85.0
UNITS	**82.2**	**88.0**	**78.7**	**86.0**

Table 2. Experiment results on Total-Text. "Full" and "None" represent recognition with each lexicon respectively.

Method	Detection	End-to-End	
	F-measure	None	Full
ABCNet v2 [7]	84.7	51.8	77.0
MANGO [10]	-	58.9	78.7
SwinTextSpotter [2]	88.0	51.8	77.0
TESTR [12]	87.1	56.0	81.5
UNITS	**88.6**	**66.4**	**82.3**

Table 3. Experiment results on CTW1500. "Full" and "None" represent recognition with each lexicon respectively.

Method	End-to-End		
	Strong	Weak	Generic
CRAFTS [1]	94.2	93.8	92.2
MaskTextSpotter v2 [9]	93.3	91.3	88.2
MANGO [10]	93.4	92.3	88.7
$\text{UNITS}_{\text{Shared}}$ – Box	**95.1**	**94.6**	92.9
$\text{UNITS}_{\text{Shared}}$ – Quad	**95.1**	**94.6**	**93.0**

Table 4. Experiment results on ICDAR 2013. "Strong", "Weak", and "Generic" represent recognition with each lexicon respectively.

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4593–4603, 2022. 2, 3

[3] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, pages 1484–1493. IEEE, 2013. 1

Method	45°		60°	
	DET	E2E	DET	E2E
MaskTextSpotter v3 [6]	84.2	76.1	84.7	76.6
SwinTextSpotter [2]	-	77.6	-	77.9
TTS [5]	88.8	80.4	87.6	**80.1**
$\text{UNITS}_{\text{Shared}}$	**91.8**	**80.6**	**90.3**	78.1

Table 5. Experiment results on Rotated ICDAR 2013. The end-to-end recognition task is evaluated without any lexicon.

Method	Detection			End-to-End
	R	P	F	
$\text{UNITS}_{\text{Shared}}$	63.1	83.2	71.8	61.4
UNITS	67.1	84.8	74.9	63.6

Table 6. Experiment results on TextOCR validation sets. "R", "P", and "F" represent recall, precision and F-measure respectively.

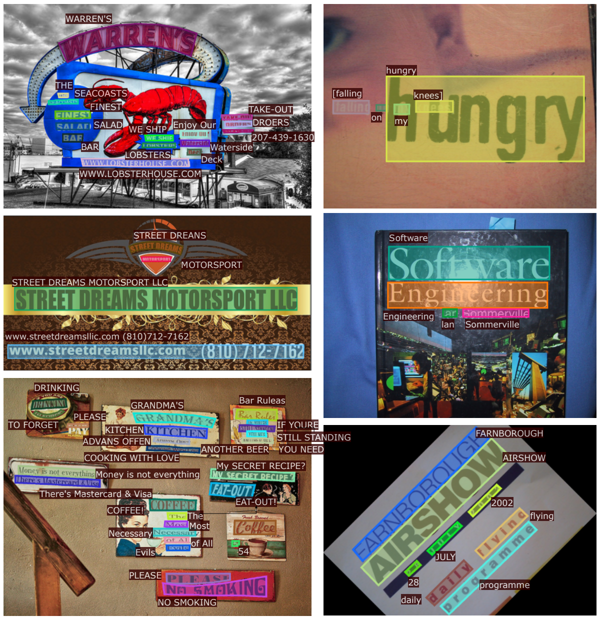


Figure 4. Qualitative results of our method on CTW-1500, ICDAR 2013 and Rotated ICDAR 2013.

[4] Seonghyeon Kim, Seung Shin, Yoonsik Kim, Han-Cheol Cho, Taeho Kil, Jaeheung Surh, Seunghyun Park, Bado Lee, and Youngmin Baek. Deer: Detection-agnostic end-to-end recognizer for scene text spotting. *arXiv preprint arXiv:2203.05122*, 2022. 2, 3

[5] Yair Kittenplon, Inbal Lavi, Sharon Fogel, Yarin Bar, R Manmatha, and Pietro Perona. Towards weakly-supervised text spotting using a multi-task transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

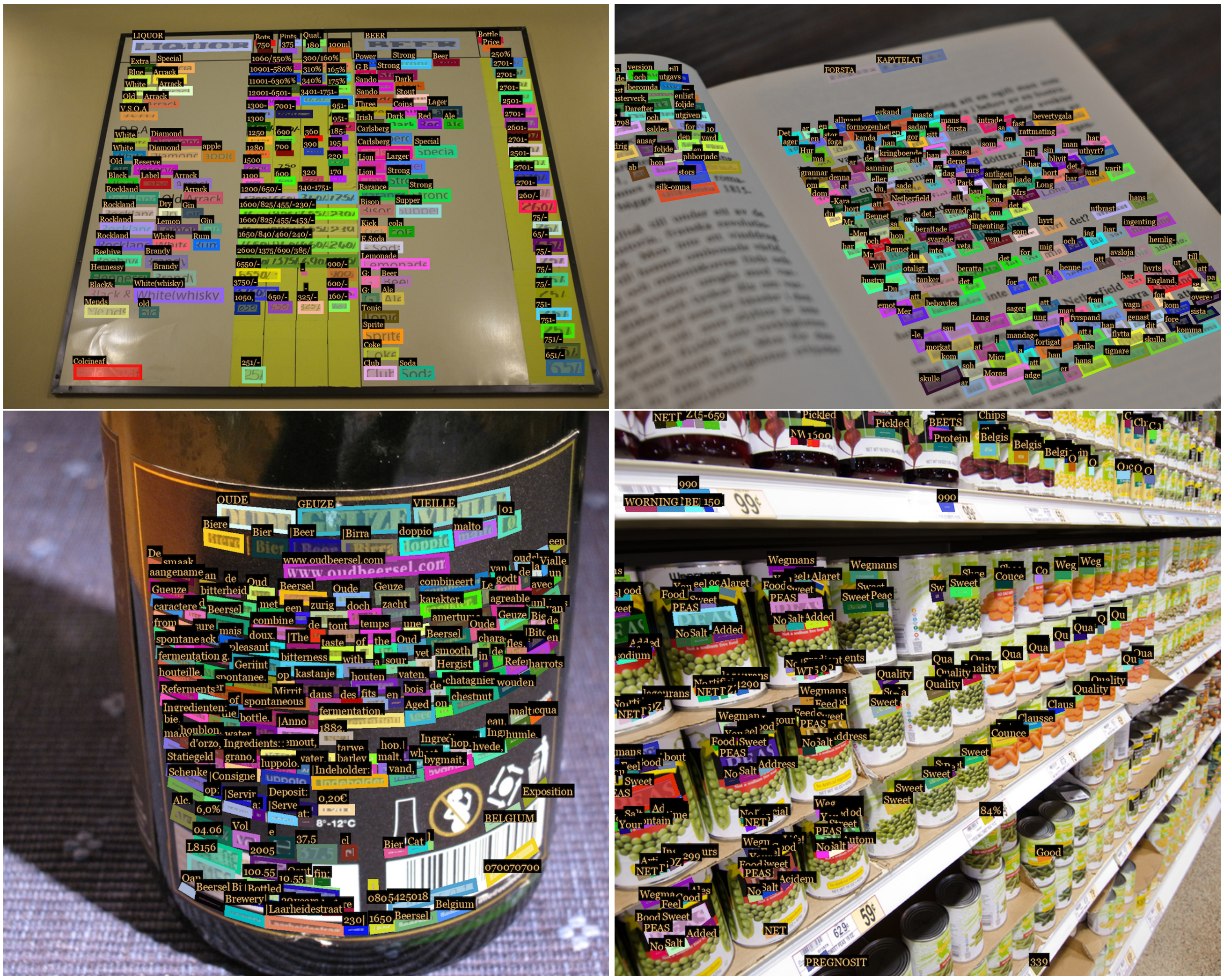


Figure 5. Qualitative results of our method on TextOCR.

Recognition, pages 4604–4613, 2022. 2, 3

[6] Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, and Xiang Bai. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In *European Conference on Computer Vision*, pages 706–722. Springer, 2020. 1, 2, 3

[7] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9809–9818, 2020. 2, 3

[8] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition*, 90:337–345, 2019. 1

[9] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–83, 2018. 3

[10] Liang Qiao, Ying Chen, Zhanzhan Cheng, Yunlu Xu, Yi Niu, Shiliang Pu, and Fei Wu. Mango: A mask attention guided one-stage scene text spotter. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2467–2476, 2021. 2, 3

[11] Roi Ronen, Shahar Tsiper, Oron Anschel, Inbal Lavi, Amir Markovitz, and R Manmatha. Glass: Global to local attention for scene-text spotting. In *European Conference on Computer Vision*, pages 249–266. Springer, 2022. 2, 3

[12] Xiang Zhang, Yongwen Su, Subarna Tripathi, and Zhuowen Tu. Text spotting transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9519–9528, 2022. 2, 3

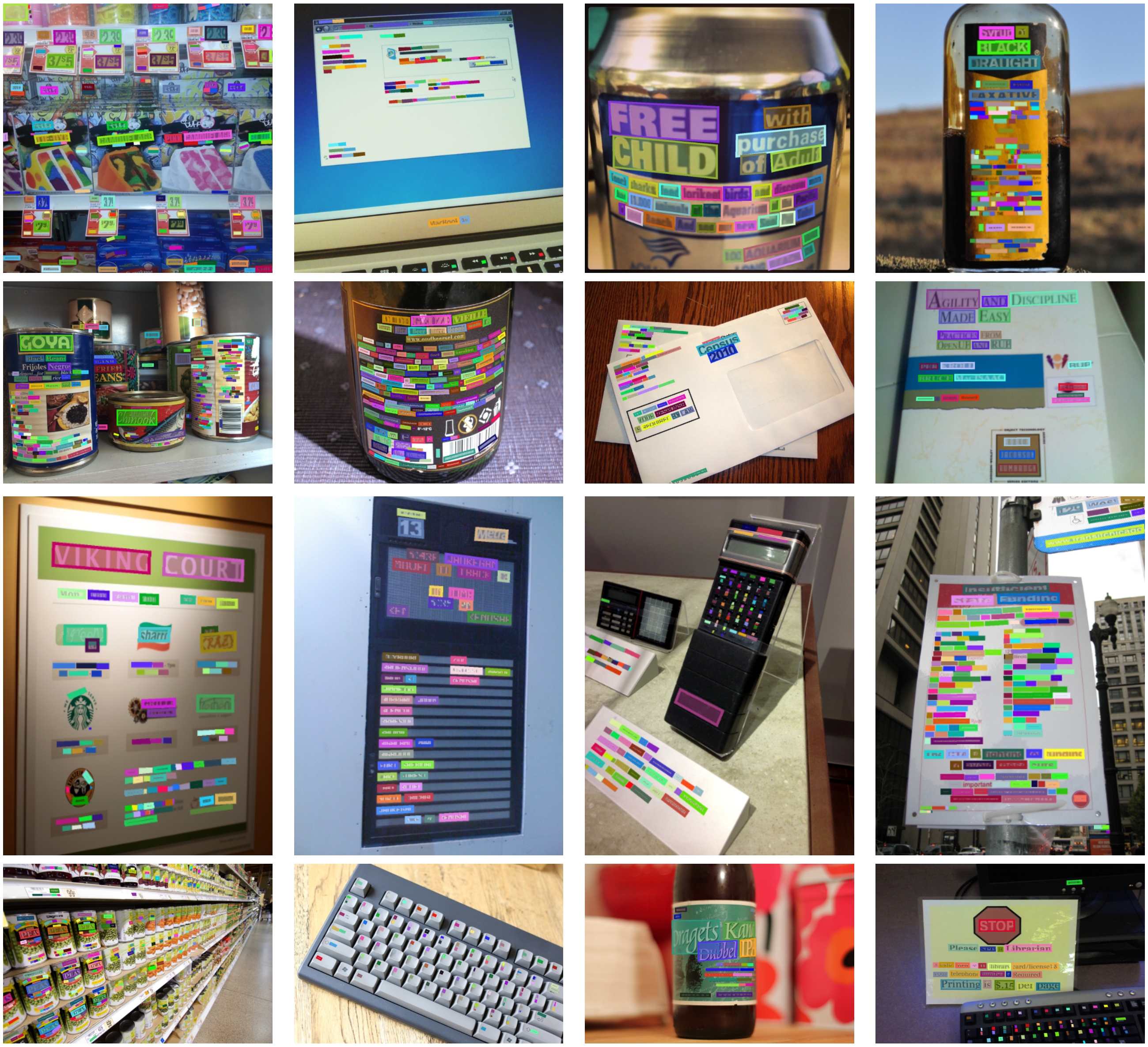


Figure 6. Qualitative detection results of our method on TextOCR.