DeepSolo: Let Transformer Decoder with Explicit Points Solo for Text Spotting

Supplementary Materials

A. Performance on ICDAR 2013

On ICDAR 2013 (IC13) benchmark for horizontal scene text, DeepSolo is pre-trained on a mixture of Synth150K, MLT17, Total-Text, IC13, IC15, then fine-tuned on IC13 for 1K iterations. During evaluation, we resize the shorter sizes of images to 1,000 while keeping the longer ones shorter than 1,824 pixels. Tab. 1 compares the performance of our method with previous models. DeepSolo achieves the performance of 95.1%, 93.7%, and 90.1% on three metrics.

Method	E2E		
	S	W	G
MaskTextSpotter [8]	92.2	91.1	86.5
MaskTextSpotter v2 [3]	93.3	91.3	88.2
MANGO [10]	93.4	92.3	88.7
SPTS [9]	93.3	91.7	88.5
DeepSolo (ResNet-50)	**95.1**	**93.7**	**90.1**

Table 1. End-to-end text spotting results on ICDAR 2013.

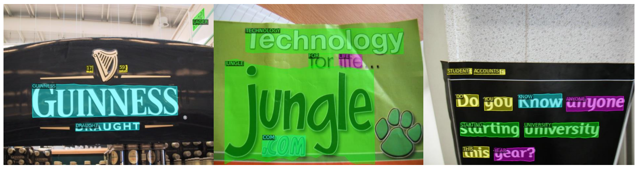

Figure 1. Qualitative results on ICDAR 2013.

B. Performance on Inverse-Text

Inverse-Text [11] is a newly proposed test set for verifying the robustness of scene text detectors and spotters on highly rotated texts. It consists of 500 testing images for arbitrary-shape scene text, with about 40% inverse-like instances. The DeepSolo models reported in Tab. 5 of the main paper are directly evaluated on Inverse-Text. Results are shown in Tab. 2. Note that we do not use a stronger rotation augmentation, *e.g.*, angle chosen from $[-90°, +90°]$ as in SwinTextSpotter. Surprisingly, when TextOCR is leveraged, the 'None' and 'Full' performance are additionally improved by 16.1% and 17.3%, respectively. While replacing ResNet-50 with ViTAEv2-S, there are 4.2% and 4.6% absolute gain on the two metrics.

Visual Analysis. Some qualitative results from DeepSolo (ViTAEv2-S) are illustrated in Fig. 2. In the first row, some inverse-like instances can be correctly recognized. However, as shown in the bottom row, some boundary predictions are not stable, resulting in invalid polygons. Adding boundary points matching may be helpful. Besides, some detection results of inverse instances are filtered because of low confidence score, which can be simply improved by adopting more extensive rotation augmentation.

Method	E2E	
	None	Full
MaskTextSpotter v2 [3]	39.0	43.5
ABCNet [5]	22.2	34.3
ABCNet v2 [6]	34.5	47.4
TESTR [13]	34.2	41.6
SwinTextSpotter [2]	55.4	67.9
SPTS [9]	38.3	46.2
DeepSolo (ResNet-50, data-1)	47.6	53.0
DeepSolo (ResNet-50, data-2)	48.5(+0.9)	53.9(+0.9)
DeepSolo (ResNet-50, data-3)	64.6(+17.0)	71.2(+18.2)
DeepSolo (ViTAEv2-S, data-3)	**68.8**(+21.2)	75.8(+22.8)
ABCNet v2 w/ Pos.Label ‡	62.2	**76.7**
TESTR w/ Pos.Label ‡	63.1	75.4
SwinTextSpotter ‡	62.9	74.7

Table 2. End-to-end text spotting results on Inverse-Text. 'data-1' denotes the external dataset is Synth150K as in Tab. 5 of the main paper. 'data-2': 'Synth150K+MLT17+IC13+IC15'. 'data-3': 'Synth150K+MLT17+IC13+IC15+TextOCR'. '‡': results from [11], using extensive rotation augmentation.

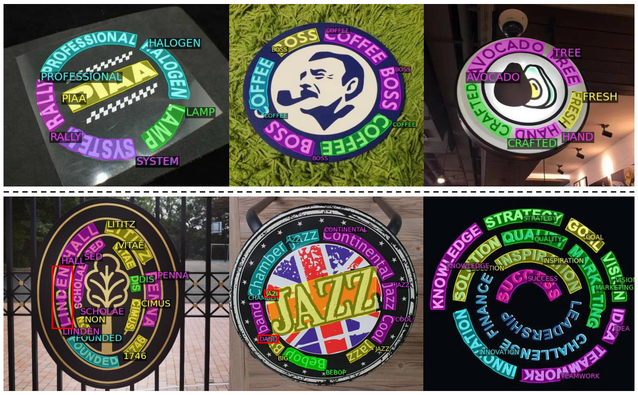

Figure 2. Qualitative results on Inverse-Text. Some failure cases are shown in the bottom row.

#Row	Backbone	Pre-training				Fine-tuning				Where in the Main Paper
		Training Data	lr (Backbone)	Iterations	Step	Training Data	lr (Backbone)	Iterations	Step	
1	ResNet-50	Synth150K+Total-Text	$1e^{-4}$ ($1e^{-5}$)	350K	300K	Total-Text	$1e^{-5}$ ($1e^{-6}$)	10K	–	Tab. 1, 2, 3, 5
2		Synth150K+Total-Text+MLT17+IC13+IC15	$1e^{-4}$ ($1e^{-5}$)	375K	320K	Total-Text	$1e^{-5}$ ($1e^{-6}$)	10K	–	Tab. 3, 4, 5
3		Synth150K+Total-Text+MLT17+IC13+IC15+TextOCR	$1e^{-4}$ ($1e^{-5}$)	435K	375K	Total-Text	$1e^{-5}$ ($1e^{-6}$)	2K	–	Tab. 3, 5
4		Synth150K+Total-Text+MLT17+IC13+IC15	$1e^{-4}$ ($1e^{-5}$)	375K	320K	IC15	$1e^{-5}$ ($1e^{-6}$)	3K	–	Tab. 6
5		Synth150K+Total-Text+MLT17+IC13+IC15+TextOCR	$1e^{-4}$ ($1e^{-5}$)	435K	375K	IC15	$1e^{-5}$ ($1e^{-6}$)	1K	–	Tab. 6
6		Total-Text	$1e^{-4}$ ($1e^{-5}$)	120K	80K	–	–	–	–	Fig. 3
7		Synth150K+MLT17+IC13+IC15+TextOCR	$1e^{-4}$ ($1e^{-5}$)	435K	375K	Total-Text+IC13+IC15	$2e^{-5}$ ($2e^{-6}$)	6K	–	Fig. 6
8		Synth150K+Total-Text+MLT17+IC13+IC15	$1e^{-4}$ ($1e^{-5}$)	375K	320K	CTW1500	$5e^{-5}$ ($5e^{-6}$)	12K	8K	Tab. 7
9	ResNet-101	Synth150K+Total-Text+MLT17+IC13+IC15	$1e^{-4}$ ($1e^{-5}$)	375K	320K	Total-Text	$1e^{-5}$ ($1e^{-6}$)	10K	–	Tab. 4

Table 3. Training details of DeepSolo with ResNet. “Step” denotes the iteration step where the learning rate is divided by 10.

C. More Details

The data augmentations include: 1) random rotation with angle chosen from $[-45°, +45°]$; 2) instance-aware random crop; 3) random resize and 4) color jitter. For inference on Total-Text and CTW1500, the image shorter sides are resized to 1,000. For IC15, following [4, 13], the shorter sizes are resized to 1,440.

C.1. Details of DeepSolo with ResNet

In this subsection, the training details of DeepSolo with ResNet [1] (ImageNet pre-trained weights from TORCHVISION) are listed in Tab. 3 with corresponding training data. In Fig. 3 of the main paper, the training schedule of DeepSolo is related to Row #6, *i.e.*, only the Total-Text training set is utilized. For SPTS [9], we only plot the final performance since it needs much more data and a longer training schedule to achieve ideal performance. The training setting of DeepSolo with line labels is provided in Row #7. Note that during fine-tuning, the line annotations are used and stronger rotation augmentation (angle randomly chosen from $[-90°, +90°]$) is adopted.

C.2. Details of DeepSolo with Swin Transformer

In Tab. 4 of the main paper, with Swin Transformer [7], we pre-train the model for 375K iterations and fine-tune it on Total-Text for 10K iterations. No part of the backbone is frozen. During pre-training, the initial learning rate for the backbone is $1e^{-4}$. The drop path rate of Swin-T and Swin-S is set to 0.2 and 0.3, respectively. During fine-tuning, we set the learning rate for the backbone to $1e^{-5}$, and the drop path rate to 0.2 and 0.3 for Swin-T and Swin-S. Other training schedules are the same as Row #2 in Tab. 3.

C.3. Details of DeepSolo with ViTAE

With ViTAE-v2-S [12], the drop path rate is set to 0.3 for pre-training and 0.2 for fine-tuning. Other schedules are the same as Swin-S backbone.

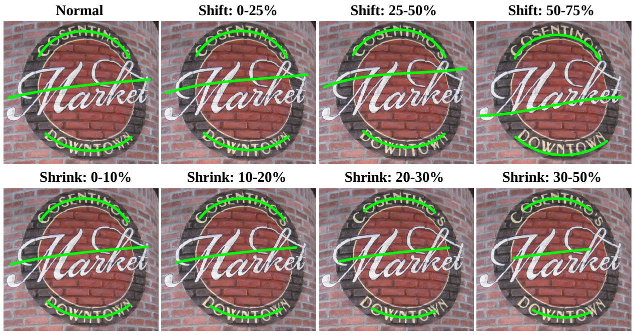

Figure 3. The illustration of line labels at different noisy levels.

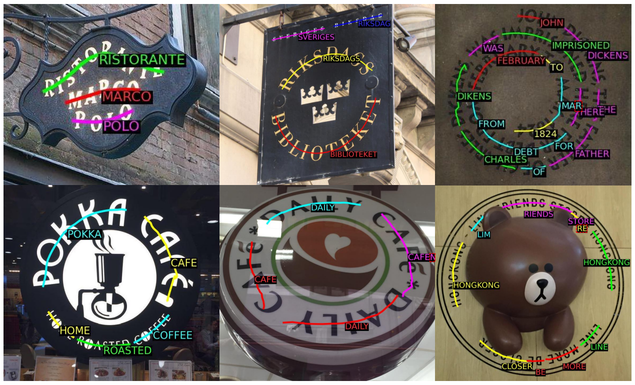

Figure 4. Spotting visualizations of DeepSolo using line annotations. The predicted center curve points in each text instance are connected, forming word lines.

D. More Visualizations

D.1. Visualizations Using Line Annotations

In Sec. 4.5 of the main paper, we study the model sensitivity to different line locations. Here, we provide a group of visualizations in Fig. 3 to intuitively show the noisy line locations. Moreover, some qualitative spotting results of DeepSolo using normal line annotations are presented in Fig. 4. Without special design, DeepSolo can correctly recognize most up-side-down text instances with stronger rotation augmentation.

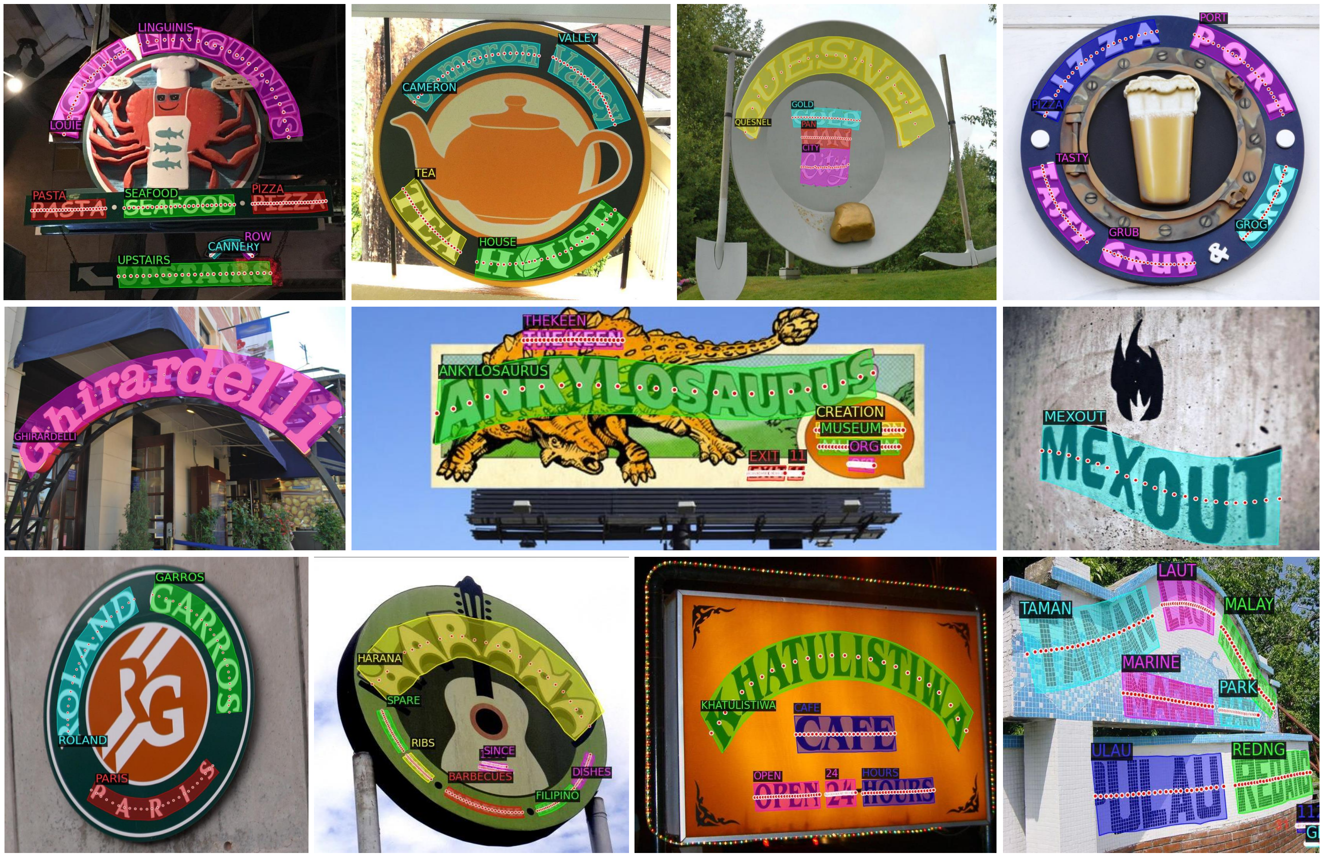

Figure 5. More qualitative results on Total-Text.

D.2. More Qualitative Results on Benchmarks

More qualitative results on Total-Text, ICDAR 2015, and CTW1500 are provided in Fig. 5, Fig. 6, and Fig. 7.

E. Limitation and Discussion

We adopt the label form which is in line with the reading order to implicitly guide DeepSolo to learn the text order. However, when the label form is not in line with the reading order or the predicted order is incorrect, how to get the accurate recognition result is worth further exploration. In this work, we do not utilize an explicit language-aware module to progressively refine recognition results. The combination of DETR-based DeepSolo and language modeling may be promising. Besides, we only study the English scene text spotting framework, we plan to develop a simple and unified multi-language scene text spotter based on DeepSolo.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[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 *CVPR*, 2022. 1

[3] Minghui Liao, Pengyuan Lyu, Minghang He, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. *PAMI*, 43(2):532–548, 2021. 1

[4] Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, and Xiang Bai. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In *ECCV*, 2020. 2

[5] 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 *CVPR*, 2020. 1

[6] Yuliang Liu, Chunhua Shen, Lianwen Jin, Tong He, Peng Chen, Chongyu Liu, and Hao Chen. Abcnet v2: Adaptive bezier-curve network for real-time end-to-end text spotting. *PAMI*, 2021. 1

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[8] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural

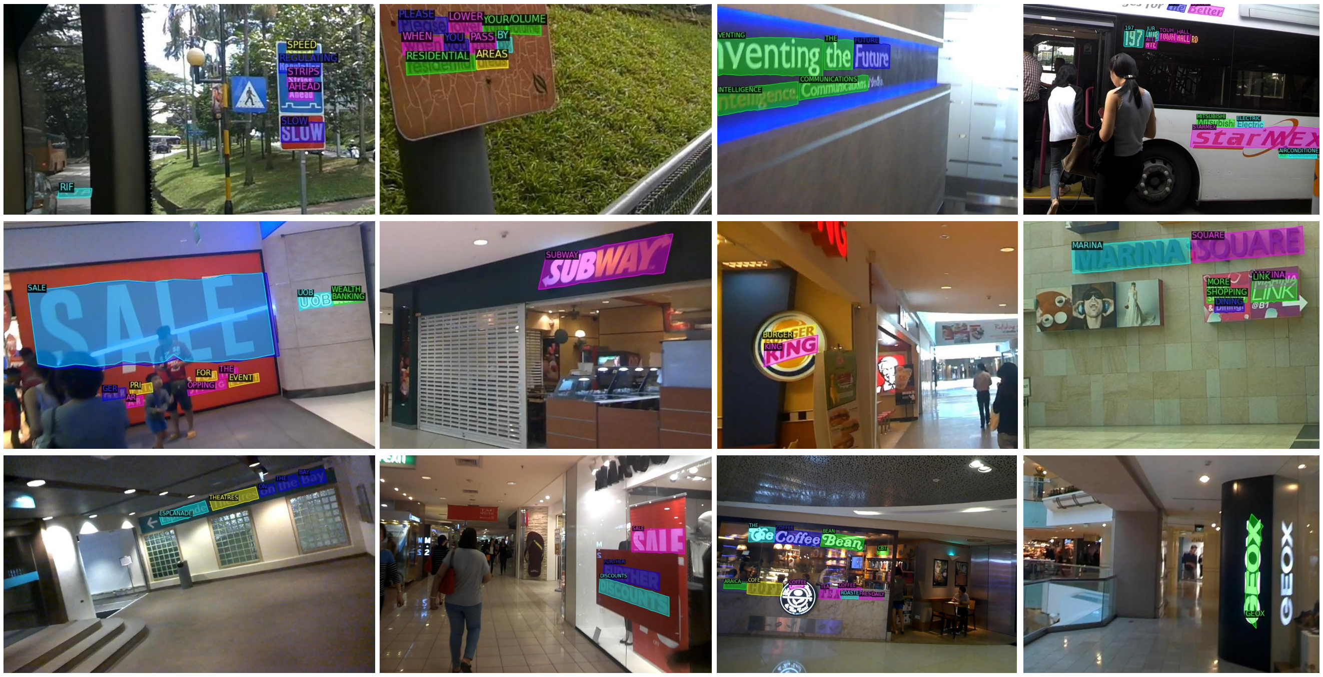

Figure 6. More qualitative results on ICDAR 2015. Center curve points are hidden for better view of small text instances.

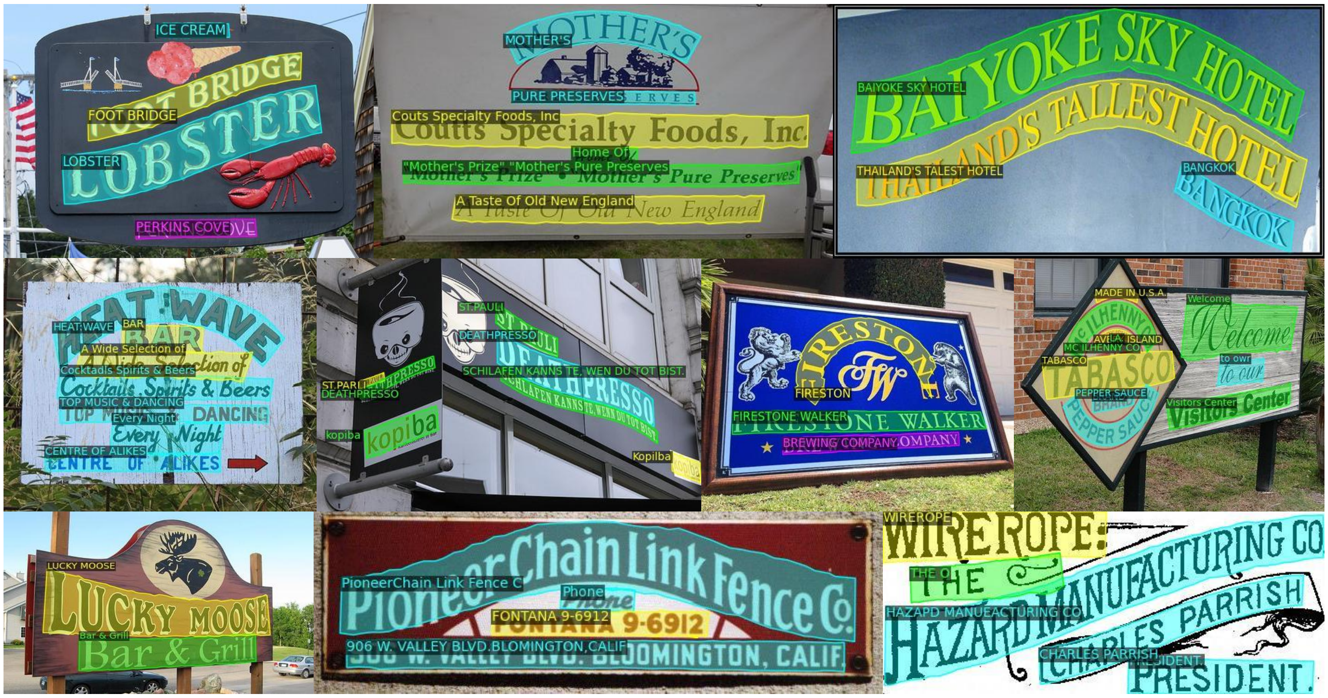

Figure 7. More qualitative results on CTW1500.

network for spotting text with arbitrary shapes. In *ECCV*, 2018. 1

[9] Dezhi Peng, Xinyu Wang, Yuliang Liu, Jiaxin Zhang, Mingxin Huang, Songxuan Lai, Jing Li, Shenggao Zhu, Dahua Lin, Chunhua Shen, et al. Spts: Single-point text spotting. In *ACM MM*, 2022. 1, 2

[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 *AAAI*, 2021. 1

[11] Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Bo

Du, and Dacheng Tao. Dptext-detr: Towards better scene text detection with dynamic points in transformer. In *AAAI*, 2023. 1

[12] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *IJCV*, 2022. 2

[13] Xiang Zhang, Yongwen Su, Subarna Tripathi, and Zhuowen Tu. Text spotting transformers. In *CVPR*, 2022. 1, 2