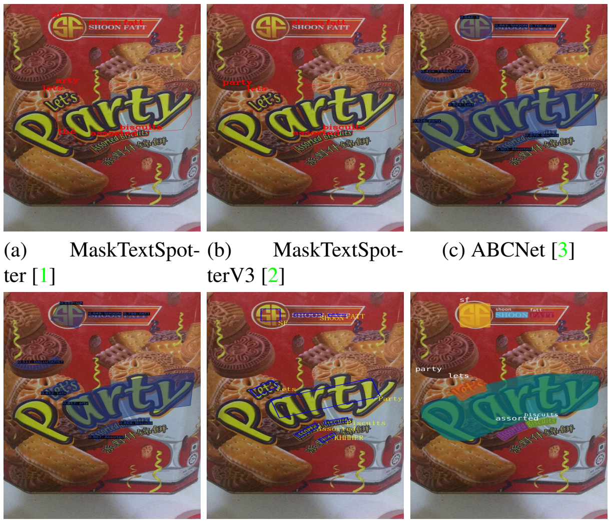
SwinTextSpotter: Scene Text Spotting via Better Synergy between Text Detection and Text Recognition (Supplementary Material)

Mingxin Huang[1†] Yuliang Liu[2†] Zhenghao Peng[2] Chongyu Liu[1] Dahua Lin[2]
Shenggao Zhu[3] Nicholas Yuan[3] Kai Ding[4] Lianwen Jin[1,5*]
[1]South China University of Technology [2]Chinese University of Hong Kong [3]Huawei Cloud AI
[4]IntSig Information Co., Ltd [5]Peng Cheng Laboratory
eelwjin@scut.edu.cn

1. Qualitative Comparisons

We make some qualitative analysis with previous method which is shown in Fig 1. It can be seen that previous methods failed on the difficult text instance such as "Party", while SwinTextSpotter can handle such case by exploiting the synergy of text detection and recognition.

Intuitively, the detection result of SwinTextSpotter is more accurate.



(a) MaskTextSpotter [1] (b) MaskTextSpotterV3 [2] (c) ABCNet [3]

(d) ABCNetv2 [4] (e) MANGO [5] (f) SwinTextSpotter

Figure 1. Qualitative analysis of SwinTextSpotter and other existing methods. Best view in screen.

2. Ablation study of Recognition Conversion

We verify the effectiveness of other components without RC which helps to better reveal the effectiveness of RC.

Method	Total-Text	
	Det-Hmean	E2E-Hmean
SwinTextSpotter-withou RC	82.8	63.4
SwinTextSpotter	83.2	66.9

Table 1. Ablation study on Recognition Conversion.

From Table 1, the performance, that using other components without RC, drops from 83.2% to 82.8% for detection and 66.9% to 63.4% for end-to-end scene text spotting.

3. Comparison different backbone in different frameworks

We also try to replace ResNet50 with Swin-Transformer on ABCNet [3]. From Table 2, the result can be improved by 1.8% with Swin-Transformer in text spotting. But there is no improvement for detection. It is similar to the case in SwinTextSpotter.

Table 2. Comparison different backbone on different architectures on Total-Text. ABC-R50 means ABCNet with ResNet50. ABC-Swin means ABCNet with SwinTransformer. Det. means detection result. E2E means end-to-end text spotting result.

ABC-R50		ABC-Swin		Our-R50		Our-Swin	
Det.	E2E	Det.	E2E	Det.	E2E	Det.	E2E
86.0	67.1	86.0	68.9	87.2	72.4	87.3	74.0

[†]Equal contribution.
[*]Corresponding author.

References

[1] M. Liao, P. Lyu, M. He, C. Yao, W. Wu, and X. Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):532–548, 2021. 1

[2] Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, and Xiang Bai. Mask textspotter v3: Segmentation proposal network

for robust scene text spotting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 706–722. Springer, 2020. 1

[3] 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. 1

[4] 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. *arXiv preprint arXiv:2105.03620*, 2021. 1

[5] 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 Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 2467–2476, 2021. 1