Bridging the Gap Between End-to-End and Two-Step Text Spotting

Supplementary Materials

A. Performance on TextOCR and HierText

We conduct experiments on more challenging benchmarks on TextOCR [7] and HierText [4] to verify the effectiveness of our methods. For TextOCR, we adopt a detector similar to the GLASS [6] and use the DiG [8] as the recognizer. For HierText, we utilize the DBNet++ [3] as the detector and MAERec [1] as the recognizer. As shown in Tab. 1, the results demonstrate the effectiveness of our method.

Table 1. End-to-end text spotting results on TextOCR and HierText.

Method	TextOCR	HierText
MaskTextSpotter v3 [2]	50.8	–
GLASS [6]	67.1	–
HTS [5]	–	75.6
Ours	68.5	76.1

References

[1] Qing Jiang, Jiapeng Wang, Dezhi Peng, Chongyu Liu, and Lianwen Jin. Revisiting scene text recognition: A data perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20543–20554, 2023. 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] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[4] 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. 1

[5] Shangbang Long, Siyang Qin, Yasuhisa Fujii, Alessandro Bissacco, and Michalis Raptis. Hierarchical text spotter for joint text spotting and layout analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 903–913, 2024. 1

[6] 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. 1

[7] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8802–8812, 2021. 1

[8] Mingkun Yang, Minghui Liao, Pu Lu, Jing Wang, Shenggao Zhu, Hualin Luo, Qi Tian, and Xiang Bai. Reading and writing: Discriminative and generative modeling for self-supervised text recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4214–4223, 2022. 1