Multi-Branch Network with Ensemble Learning for Text Removal in the Wild Supplementary Material

Yujie Hou[1], Jiwei Chen[1], and Zengfu Wang[1,2]✉

[1] School of Information Science and Technology, University of Science and Technology of China, Hefei, China
[2] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, China
{houyj1219,cjwbdw06}@mail.ustc.edu.cn, zfwang@ustc.edu.cn

1 Introduction

This document presents the supplementary material of the main paper due to the space limitation. Sec. 2 demonstrates the Detection-Eval results about our MBE and other scene text removal (STR) methods. Sec. 3 presents examples of text removal on the SCUT-Syn dataset. Sec. 4 shows the results of our method by training with different size images. Sec. 5 introduces the limitation of our method.

2 Comparison With State-of-the-Art Methods in Detection-Eval

We use the metrics which is denoted as Detection-Eval in [2, 4] to evaluate the quality of the text-erased results on SCUT-EnsText [2]. In our experiment, we denote Precision, Recall, and F-score as “P”, “R”, and “F”, respectively, with “TP”, “TR”, and “TF” for TIoU-Precision, TIoU-Recall, and TIoU-F-score. The results on SCUT-EnsText are shown in Tab. 1. It indicates that our method achieves better performance than the previous STR methods. Though SceneTextEraser [3] has lower P and TP than our proposed model, it breaks the integrity of the whole image by processing the text removal and background restoration on 64×64 patches. Thus, our method has higher results in other metrics and image quality than the SceneTextEraser.

3 The visualization of erasure results on SCUT-Syn

Fig. 1 shows the qualitative results of our method on SCUT-Syn. Our method can remove synthetic texts in images while reserving the integrity of backgrounds.

4 Generalization of MBE

It is worth noting that our method has strong generalization when training with part of original inputs. We randomly crop the input image into different sizes for training, such as the middle resolution image (416 × 416) and the quarter size image (256 × 256), and then we test the model with full resolution image (512 × 512) in SCUT-EnsText. Tab. 2 indicates that our method outperforms existing STR methods in PSNR only training with a quarter size image and gets better performance as increasing the image size. Further, MBE with quarter size inputs has 4× fewer memory in GPU consumption while being 4× faster than MBE with original resolution input.

Table 1: Quantitative comparison of our method and start-of-art methods on SCUT-EnsText. Best and second best scores are highlighted and underlined.

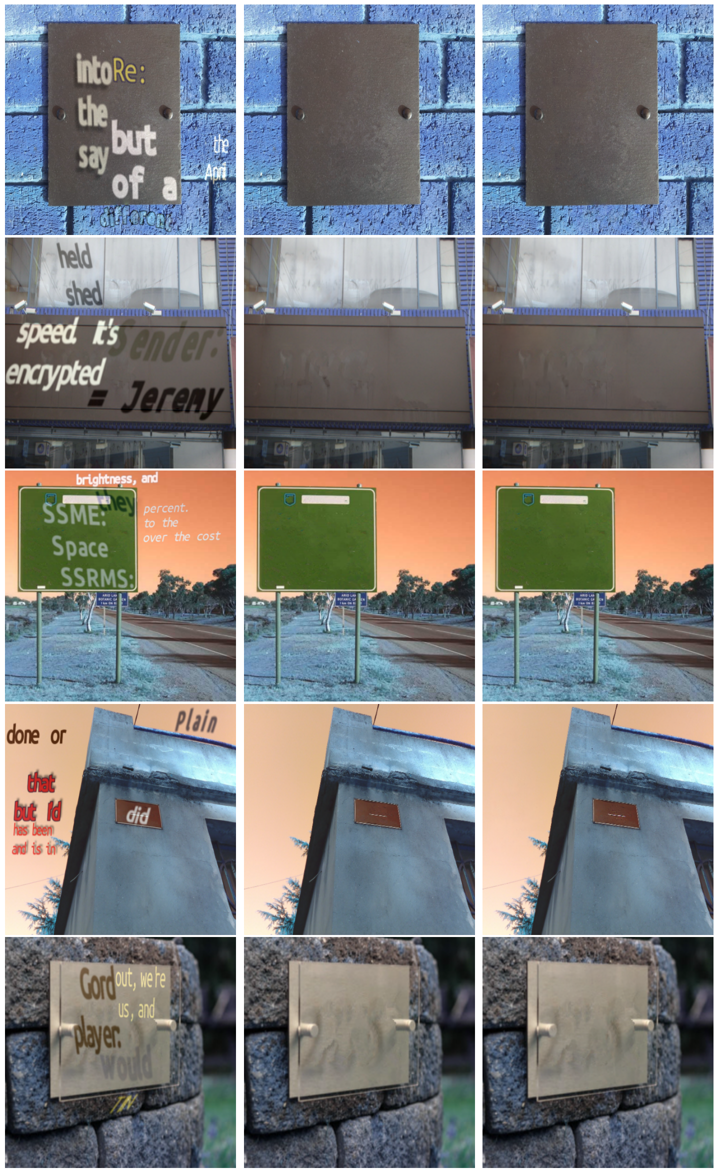
Method	P(↓)	R(↓)	F(↓)	TP(↓))	TR(↓)	TF(↓)
Original images	79.4	69.5	74.1	61.4	50.9	55.7
Pix2Pix [1]	69.7	35.4	47.0	52.0	24.3	33.1
SceneTextEraser [3]	**40.9**	5.9	10.2	**28.9**	3.6	6.4
EnsNet [5]	68.7	32.8	44.4	50.7	22.1	30.8
Erasenet [2]	53.2	4.6	8.5	37.6	2.9	5.4
PERT [4]	52.7	<u>2.9</u>	<u>5.4</u>	38.7	<u>1.8</u>	<u>3.5</u>
MBE	<u>42.5</u>	**1.7**	**3.3**	30.7	**1.2**	**2.3**

5 Limitation

Our method fails when the scene text is large since the simple segmentation head can not capture the whole text region (see the second row in Fig. 2).

Table 2: Experimental results of MBE with input image randomly cropped from quarter size (256 × 256) to original resolution (512 × 512) on SCUT-EnsText. Best and second best scores are highlighted and underlined.

Method	Training Image Size	PSNR(↑)	MSSIM(↑)	AGE(↓))	pEPs(↓)	pCEPs(↓)
EraseNet [2]	512 × 512	32.2976	0.954	3.1264	0.0192	0.0110
PERT [4]	512 × 512	33.2492	0.9695	2.1833	0.0136	**0.0088**
MBE	256 × 256	33.8445	0.9700	2.2046	0.0144	0.0098
MBE	320 × 320	34.6093	0.9701	2.1858	0.0142	0.0097
MBE	352 × 352	34.6761	0.9705	2.1661	0.0148	0.0097
MBE	416 × 416	34.7574	<u>0.9709</u>	<u>2.1064</u>	<u>0.0133</u>	<u>0.0090</u>
MBE	480 × 480	<u>34.8175</u>	0.9707	2.1144	0.0135	0.0092
MBE	512 × 512	**35.0304**	**0.9731**	**2.0594**	**0.01282**	**0.0088**



(a) Original image (b) Ground-truth (c) MBE (Ours)

Fig. 1: Qualitative results of our method on SCUT-Syn.

Fig. 2: Failure cases of MBE in the real world SCUT-EnsText dataset.

References

1. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 2
2. Liu, C., Liu, Y., Jin, L., Zhang, S., Luo, C., Wang, Y.: Erasenet: End-to-end text removal in the wild. IEEE Transactions on Image Processing **29**, 8760–8775 (2020) 1, 2
3. Nakamura, T., Zhu, A., Yanai, K., Uchida, S.: Scene text eraser. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 832–837. IEEE (2017) 1, 2
4. Wang, Y., Xie, H., Fang, S., Qu, Y., Zhang, Y.: A simple and strong baseline: Progressively region-based scene text removal networks. arXiv preprint arXiv:2106.13029 (2021) 1, 2
5. Zhang, S., Liu, Y., Jin, L., Huang, Y., Lai, S.: Ensnet: Ensconce text in the wild. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 801–808 (2019) 2