Supplementary Material for The Surprisingly Straightforward Scene Text Removal Method With Gated Attention and Region of Interest Generation: A Comprehensive Prominent Model Analysis

Hyeonsu Lee[1 [0000−0002−6317−9883]] and Chankyu Choi[1 [0000−0002−9166−2100]]

NAVER Corp {hyeon-su.lee, chankyu.choi}@navercorp.com

A Appendix

A.1 Comparison with general image inpainting

We compared our proposed method with general inpainting methods. LBAM [3], Gated Convolution [4] are adopted for comparison. We trained LBAM [3], which is pre-trained on Paris Street View [1], for 4 epochs on our combined dataset. For general inpainting methods, the text-stroke which is located outside of the box mask can affect the model's performance. To solve this issue, we provided the bounding box information, which was dilated 4 times with a 3x3 kernel, to the inpainting model. However, due to time constraints, we could not train GC [4] on our combined dataset. Instead, we evaluated model, pre-trained on Places2 [5], which has Contextual Attention module [2]. The quantitative results are shown in Tab. 6. For comparison, we used composited images generated by using box masks. Table 1 shows that our proposed method outperforms the existing inpainting method in all metrics. The qualitative results are shown in Figure 1. As shown in Figure 1, the results of LBAM [3] are blurry and incomplete. When masked regions contain complex backgrounds, the model can not reconstruct non-text information in masked regions properly, while our methods can reconstruct non-text regions and only erase text-stroke regions.

We acknowledge the lack of comparison with GC [4]. We planned to trained GC [4] on our combined dataset in the future. In addition, we planned to apply the Gated Convolution module to our model to compare performance between Gated Convolution and our proposed module without the influence of Contextual Attention [2].

Fig. 1. Comparison of the quality of images. Image from left to right: input image, ground truth image, LBAM pre-trained on Paris Street View, GC pre-trained on Places2, LBAM fine-tuned on our dataset, Ours.

Table 1. Comparison for SCUT-EnsText (Image Eval). For a fair comparison, we provided box masks, which were dilated four times with a 3x3 kernel, to our methods.

Method	data	Input size	SCUT-EnsText PSNR	SSIM	AGE
LBAM [3]	pre-trained	512	34.20	96.13	1.6670
	pre-trained + Ours	512	36.76	97.55	1.1404
GC [4]	pre-trained	512	34.24	96.46	1.5049
Ours	Ours	512	39.20	98.11	0.8302

References

1. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.A.: What makes paris look like paris? Communications of the ACM **58**(12), 103–110 (2015)
2. Song, Y., Yang, C., Lin, Z., Liu, X., Huang, Q., Li, H., Kuo, C.C.J.: Contextual-based image inpainting: Infer, match, and translate. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018)
3. Xie, C., Liu, S., Li, C., Cheng, M.M., Zuo, W., Liu, X., Wen, S., Ding, E.: Image inpainting with learnable bidirectional attention maps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8858–8867 (2019)
4. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4471–4480 (2019)
5. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)