Supplementary Material for Towards Robust Scene Text Image Super-resolution via Explicit Location Enhancement

Hang Guo[1], Tao Dai[2], Guanghao Meng[1,3], and Shu-Tao Xia[1,3]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]College of Computer Science and Software Engineering, Shenzhen University
[3]Peng Cheng Laboratory, Shenzhen, China
{cshguo, daitao.edu}@gmail.com, mgh19@mails.tsinghua.edu.cn, xiast@sz.tsinghua.edu.cn

A Complexity Analysis in Feature Selection

Recall that we use the feature selection in the Location Enhancement Module. In this section, we analyze the computational complexity reduction from this scheme. Let the ratio of the number of selected K features to the number of all HW features be $\frac{K}{HW} = r \leq 1$. Consider that the computational complexity in the vanilla attention mechanism [11] is $\mathcal{O}(HWTC)$, where HW is the length of the flatten image sequence, T is the length of the text sequence, and C is the feature dimension. When we apply feature selection to turn the image sequence from $\mathbb{R}^{NCHW}$ into $\mathbb{R}^{NCK}$, the computational complexity is thus reduced to r. It is worth noting that the area occupied by the characters in one scene text image is usually less than half of the whole image. Therefore, the proposed feature selection can reduce the computational complexity by more than half in the attention mechanism. In addition, this scheme can also improve downstream recognition accuracy (See Table 3).

B Additional Experimental Results

B.1 Comparison with Other Recognizers

In the main paper, we followed previous works to compare only on CRNN, MORAN and ASTER with other SoTA STISR methods, without comparing on some recent text recognizers. Here we give the comparison results of the proposed method with other STISR models on ABINet [5], MATRN [7] and PARSeq [1](see Table 1). It can be seen that using the recent text recognizers, the proposed LEMMA still achieves the state-of-the-art performance in terms of average recognition accuracy.

Table 1. Recognition Accuracy with other STISR methods on ABINet, MATRN and PARSeq.

Method	ABINet [5]				MATRN [7]				PARSeq [1]			
	Easy	Medium	Hard	avgAcc	Easy	Medium	Hard	avgAcc	Easy	Medium	Hard	avgAcc
TBSRN	79.80%	64.99%	48.47%	65.40%	81.66%	65.98%	50.11%	66.91%	83.69%	66.69%	51.75%	68.40%
TATT	80.67%	65.77%	50.26%	66.52%	81.10%	66.62%	51.68%	67.39%	82.21%	65.91%	52.12%	67.71%
C3-STISR	81.35%	66.90%	49.89%	67.03%	81.90%	68.04%	51.08%	67.96%	**84.25%**	68.25%	50.86%	68.83%
Ours	**82.64%**	**69.24%**	**50.56%**	**68.46%**	**82.83%**	**70.38%**	**51.68%**	**69.25%**	83.63%	**69.17%**	**52.27%**	**69.33%**

B.2 Comparison Results on Fidelity

We present the comparison results on fidelity in Table 2. It can be seen that our method is comparable in terms of image fidelity (i.e., PSNR and SSIM), while outperforming other methods in terms of accuracy for downstream recognition task. As discussed in Section 4.6, the reason for the fidelity results is that we use location enhancement to emphasize character regions and we do not use the fidelity-related loss (e.g., SSIM-related loss). The excessive fidelity tends to over-emphasize the background restoration as well as smooth character boundaries, which is detrimental to the subsequent recognition tasks. Therefore, we mainly focus on the effectiveness on recognition accuracy improvement in this work.

Method	Metric		
	PSNR	SSIM	avgAcc
Bicubic	20.35	0.6961	26.8%
TBSRN [3]	20.91	0.7603	48.1%
TG [4]	21.40	0.7456	48.9%
TATT[6]	**21.52**	**0.7930**	52.6%
C3-STISR [14]	21.51	0.7721	53.7%
Ours	20.95	0.7729	**56.3%**

Table 2. Fidelity and recognition comparison with major existing methods. The results are obtained by averaging all of the three settings (Easy, Medium and Hard), CRNN [9] is used as text recognizer.

B.3 Ablation on Number of Features Selected

In the Location Enhancement Module, we use the feature selection to achieve a focused highlight on the foreground character regions while reducing the usage of expensive attention mechanisms. In order to investigate how many features should be selected to achieve a trade-off between performance and efficiency, we conducted this experiment. Table 3 shows the experimental results. It can be seen that using all features increases the image quality (higher PSNR and SSIM). As fewer features are selected, we obtain higher recognition accuracy. Considering the main purpose of the STISR task is to boost accuracy. We adopt a moderate features number 500 as the final choice.

Num of	Metric					
Feature	PSNR	SSIM	Easy	Medium	Hard	avgAcc
300/1024	20.4	0.7568	65.5%	54.8%	40.4%	54.3%
500/1024	20.9	**0.7729**	**67.1%**	**58.8%**	**40.6%**	**56.3%**
700/1024	20.9	0.7633	64.8%	56.3%	40.2%	54.5%
1024/1024	**21.0**	0.7667	65.0%	55.3%	39.8%	54.1%

Table 3. How many feature should be selected. K/N means selecting Top K according to the confidence from image feature with N element.

B.4 Ablation on Hyper-parameters Setting

We perform a study on the loss hyper-parameters in Eq.10. Since the fine-tune parameter α_2 varies with different text recognizers, meanwhile we find it has little effect on the results, we thus focus on the effect of the recognition parameter α_1. The result is given in Table 4. It can be seen that when α_1 is low, increasing α_1 can improve both image quality and recognition accuracy. However, when α_1 continues to increase, the model faces a trade-off between accuracy and fidelity: higher recognition accuracy usually implies clearer boundaries and this leads to a detriment of fidelity. Considering the goal of STISR task is to improve recognition accuracy, we take $\alpha_1 = 0.5$ as the final choice.

α_1	Fideliy		Accuracy			
	PSNR	SSIM	Easy	Meidium	Hard	Average
0.0	20.8	**0.7744**	58.7%	50.7%	36.1%	49.2%
0.2	20.8	0.7670	63.7%	54.5%	39.3%	53.2%
0.5	**20.9**	0.7729	**67.1%**	**58.8%**	**40.6%**	**56.3%**
0.7	**20.9**	0.7736	65.5%	55.6%	40.2%	54.5%
1.0	20.8	**0.7744**	64.4%	53.9%	39.5%	53.3%

Table 4. Ablation study on hyper-parameters setting in loss function.

C Supplement for Scene Text Recognition Benchmarks

In addition to using Textzoom [13], which is specifically designed for the STISR task, we also use four scene text recognition datasets to evaluate the generalizability of the model.

C.1 Description of STR Datasets

ICDAR2015 [2] consists of images taken from scenes and also has two versions: 1,811 images (IC15_S) and 2,077 images (IC15_L). In this work, we follow the previous method [3] and use the IC15_S for experiments. Many images from this dataset are noisy, blurred, contain complex background, and some are of low resolution which are difficult to text recognizers even for humans. CUTE80 [10] consists of 288 images of which texts are heavily curved. SVT (Street View Text) [12] has 647 images collected from Google Street View. SVTP (Street View Text Perspective) [8] contains 645 images of which texts are captured in perspective views.

C.2 Details of Synthetic LR

Following previous work [6], we first selected images with pixel size less than 16×64, and then we used manual degradation for pre-processing. This includes random GaussianBlur, random GaussianNoise, BilateralFilter, and Image Sharpening.

References

1. Bautista, D., Atienza, R.: Scene text recognition with permuted autoregressive sequence models. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII. pp. 178–196. Springer (2022)
2. i Bigorda, L.G., Karatzas, D., Nicolaou, A., Ghosh, S.K., Bagdanov, A.D., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V., Lu, S., Shafait, F., Uchida, S., Valveny, E.: Icdar 2015 competition on robust reading. international conference on document analysis and recognition (2015)
3. Chen, J., Li, B., Xue, X.: Scene text telescope: Text-focused scene image super-resolution. computer vision and pattern recognition (2021)
4. Chen, J., Yu, H., Ma, J., Li, B., Xue, X.: Text gestalt: Stroke-aware scene text image super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 285–293 (2022)
5. Fang, S., Xie, H., Wang, Y., Mao, Z., Zhang, Y.: Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7098–7107 (2021)
6. Ma, J., Liang, Z., Zhang, L.: A text attention network for spatial deformation robust scene text image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5911–5920 (2022)
7. Na, B., Kim, Y., Park, S.: Multi-modal text recognition networks: Interactive enhancements between visual and semantic features. In: European Conference on Computer Vision. pp. 446–463. Springer (2022)
8. Phan, T.Q., Shivakumara, P., Tian, S., Tan, C.L.: Recognizing text with perspective distortion in natural scenes. international conference on computer vision (2013)
9. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence **39**(11), 2298–2304 (2016)
10. Shivakumara, P., Risnumawan, A., Chan, C.S., Tan, C.L.: A robust arbitrary text detection system for natural scene images. Expert Systems With Applications (2014)
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
12. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. international conference on computer vision (2011)
13. Wang, W., Xie, E., Liu, X., Wang, W., Liang, D., Shen, C., Bai, X.: Scene text image super-resolution in the wild. european conference on computer vision (2020)
14. Zhao, M., Wang, M., Bai, F., Li, B., Wang, J., Zhou, S.: C3-stisr: Scene text image super-resolution with triple clues. international joint conference on artificial intelligence (2022)