Supplementary Materials to "A Text Attention Network for Spatial Deformation Robust Scene Text Image Super-resolution"

Jianqi Ma[1] Zhetong Liang[2] Lei Zhang[1]
[1]The Hong Kong Polytechnic University; [2]OPPO Research
{csjma, cslzhang}@comp.polyu.edu.hk, zhetongliang@163.com

In this supplementary file, we provide:

1. Detailed descriptions on the proposed recurrent positional encoding method (refer to Section 3.2 in the main paper).
2. Derivation of the triplex SSIM formula (refer to Section 3.3 in the main paper).
3. Best choices for the balancing parameters α and β for our loss function (refer to Section 3.4 in the main paper).
4. Degradation settings and visual comparisons on the three recognition datasets (refer to Section 4.4 in the main paper).

1. Descriptions on Recurrent Positional Encoding

We adopt a learnable recurrent positional encoding (RPE) to encode the sequential bias along the width dimension of the image feature f_I. Firstly, we initialize a set of learnable sequential parameters with length w and channel size hc, denoted as $\theta \in \mathbb{R}^{w \times hc}$. Then θ is fed to a BiGRU network [1] to learn the position encoding. The BiGRU network is composed of two modules that recurrently process a sequence in different directions. The forward module G_{lr} processes the input sequence θ in a left-to-right direction (along the length dimension), while G_{rl} takes the right-to-left direction. Hidden states are maintained in the two modules to record the sequential dependency. The computation for the forward module G_{lr} is described as:

$$\overrightarrow{\theta_i}, \overrightarrow{h_i} = G_{lr}\left(\theta_i, \overrightarrow{h_{i-1}}\right), \quad i = 1, ..., w \tag{1}$$

where i denotes the index in the length dimension. $\overrightarrow{\theta_i}$, $\overrightarrow{h_i}$ and $\overrightarrow{h_{i-1}}$ denote the output, current hidden state and previous hidden state of G_{lr}, respectively. The symbol → indicates the left-to-right direction. The initial hidden state $\overrightarrow{h_0}$ is set to an array with zero values. The backward module G_{rl} has a similar formulation, which is described as:

$$\overleftarrow{\theta_i}, \overleftarrow{h_i} = G_{rl}\left(\theta_i, \overleftarrow{h_{i+1}}\right), \quad i = 1, ..., w \tag{2}$$

Then, the position encoding θ_R is obtained by concatenating $\overrightarrow{\theta} \in \mathbb{R}^{w \times \frac{hc}{2}}$ and $\overleftarrow{\theta} \in \mathbb{R}^{w \times \frac{hc}{2}}$ in the channel dimension, denoted as:

$$\theta_R = \overrightarrow{\theta} \odot \overleftarrow{\theta} \tag{3}$$

Finally, the position encoding $\theta_R \in \mathbb{R}^{w \times hc}$ is reshaped to the size $hw \times c$ and added to the image feature f_I in an element-wise manner.

2. Derivation of Triplex SSIM

Similar to the structural similarity index measure (SSIM) [6], the proposed Triplex SSIM (TSSIM) calculates three similarity measurements on a triplex of image patches (X, Y and Z), which is described as:

$$\textit{TSSIM}(X, Y, Z) = l(X, Y, Z) \cdot c(X, Y, Z) \cdot s(X, Y, Z) \tag{4}$$

α	β	avg	PSNR	SSIM	β	α	avg	PSNR	SSIM
1	0.001	51.6%	21.50	0.7822	0.1	0.001	48.5%	21.22	0.7886
	0.01	52.0%	21.45	0.7886		0.01	50.4%	**21.60**	0.7899
	0.1	**52.6%**	**21.52**	0.7930		0.1	52.0%	21.45	0.7910
	1	52.4%	21.12	**0.7947**		1	**52.6%**	21.52	**0.7930**
	10	52.0%	20.97	0.7929		10	51.2%	21.01	0.7864

Table 1. Ablation studies on different β values. The evaluation results on SR text recognition (avg), PSNR and SSIM are in average of three splits of TextZoom [5].

where $l(X,Y,Z)$, $c(X,Y,Z)$ and $s(X,Y,Z)$ denote the similarity measurements on luminance, contrast and structure, respectively, which are formulated as:

$$\begin{aligned} l(X,Y,Z) &= \frac{\mu_X\mu_Y + \mu_Y\mu_Z + \mu_X\mu_Z + C_1}{\mu_X^2 + \mu_Y^2 + \mu_Z^2 + C_1} \\ c(X,Y,Z) &= \frac{\sigma_X\sigma_Y + \sigma_Y\sigma_Z + \sigma_X\sigma_Z + C_2}{\sigma_X^2 + \sigma_Y^2 + \sigma_Z^2 + C_2} \\ s(X,Y,Z) &= \frac{\sigma_{XY} + \sigma_{YZ} + \sigma_{XZ} + C_3}{\sigma_X\sigma_Y + \sigma_Y\sigma_Z + \sigma_X\sigma_Z + C_3} \end{aligned} \tag{5}$$

In (5), C_1, C_2 and C_3 denote the coefficients that stabilize the division. μ_X, μ_Y and μ_Z denote the mean values of X, Y and Z, respectively. σ_X, σ_Y and σ_Z denote the standard deviations of X, Y and Z, respectively. σ_{XY} denotes the correlation operations on X and Y, which is described as:

$$\sigma_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y) \tag{6}$$

where i and N denote the pixel index and the total number of pixels in a patch, respectively. Similar formulations can be derived for σ_{YZ} and σ_{XZ}. Lastly, by combining Eqs. (4) and (5) and setting $C_3 = C_2$, we obtain the formula of TSSIM as follows:

$$\mathit{TSSIM} = \frac{(\mu_X\mu_Y + \mu_Y\mu_Z + \mu_X\mu_Z + C_1)(\sigma_{XY} + \sigma_{YZ} + \sigma_{XZ} + C_2)}{(\mu_X^2 + \mu_Y^2 + \mu_Z^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + \sigma_Z^2 + C_2)} \tag{7}$$

where C_1 and C_2 are set to 0.01 and 0.03, respectively, in our experiments.

3. Selection on the Balancing Parameters of the Loss Function

Firstly, we fix β to 0.1 and vary the values of α to choose the best α in terms of recognition accuracy and PSNR/SSIM metrics. From the results in the right part of Tab. 1, one can see that the average recognition accuracy and SSIM are improved when α increases. However, all the evaluation metrics decline when α is larger than 1. This indicates that a too large weight for the text prior loss will impair the text reconstruction quality. Thus, we set α to 1. Then we fix α to 1 and evaluate the effect of β. From the results in the left part of Tab. 1, we can see that a larger β leads to improved recognition accuracy and SSIM/PSNR. However, the recognition accuracy and PSNR decline when β is larger than 0.1. Thus, we set β to 0.1.

4. Degradation Settings in the Three Recognition Datasets

We add contrast variation, blurring and noise to the images in the three recognition datasets, including ICDAR2015 [2], SVTP [3] and CUTE80 [4]. The contrast variation is formulated as $\hat{Y} = k_1 Y + k_2$, where Y and $\hat{Y}$ denote the text image before and after the contrast degradation, respectively. The perturbation parameters k_1 and k_2 are set to 1.9 and 0.44, respectively. To blur the image, we convolve the image with a 5×5 sized Gaussian kernel with σ=1. As for the noise corruption, we add Gaussian noise with σ=50 to the original image. The super-resolution results by different methods are shown in Fig. 1. We can see that the proposed TATT network achieve superior image quality over the compared methods across all types of degradations.

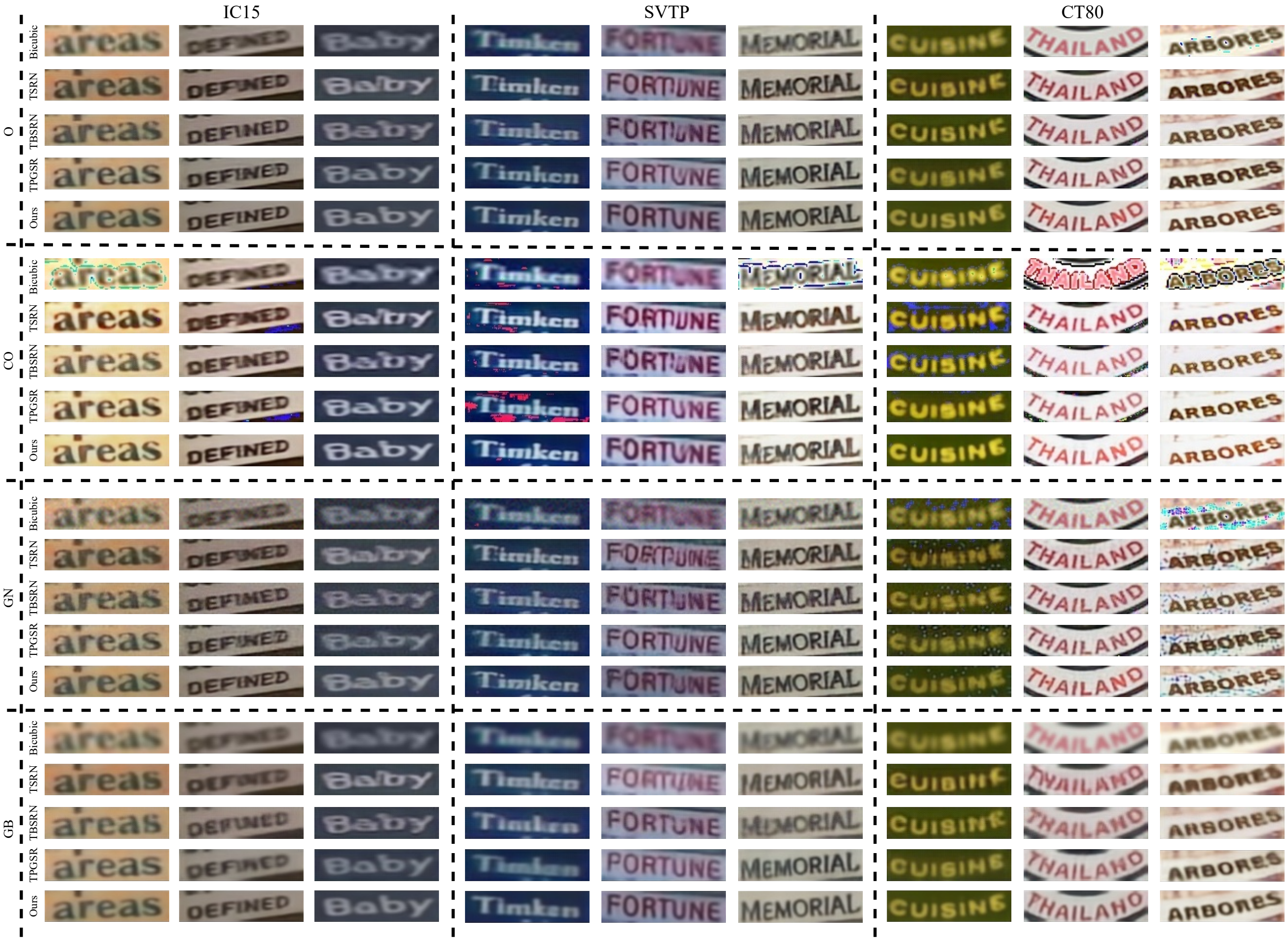

Figure 1. Recovered samples by different STISR models on ICDAR2015 [2] (IC15), SVTP [3] and CUTE80 [4] (CT80). ‘O’, ‘GB’, ‘GN’ and ‘CO’ refer to original, Gaussian blurring, Gaussian noise and contrast variation. Zoom in for more details.

References

[1] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 1

[2] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. ICDAR 2015 competition on robust reading. In *Int. Conf. Doc. Anal. Recog.*, pages 1156–1160. IEEE, 2015. 2, 3

[3] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *Int. Conf. Comput. Vis.*, pages 569–576, 2013. 2, 3

[4] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Syst. Appl*, 41(18):8027–8048, 2014. 2, 3

[5] Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *Eur. Conf. Comput. Vis.*, 2020. 2

[6] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 1