NCAP: Scene Text Image Super-Resolution with Non-CAtegorical Prior Supplementary Material

Dongwoo Park and SUK PIL KO
THINKWARE Corporation, Republic of Korea
{infinity7428, spko}@thinkware.co.kr

A. Diagram for Character-Level Matching

Figure 1 illustrates the computation process for character-level matching. We calculate the character error rate (CER) score through the following steps. Initially, we conduct sequence alignment of the ground truth and predicted characters using Levenshtein distance, a technique for determining the alignment path. Levenshtein distance consists of three types of single-character edits: insertions, deletions, and substitutions. In our character-level matching process, the roles of each edit are slightly modified. For insertions, a space is added to the prediction; for deletions, a space is added to the ground truth; and for substitutions, no operation is performed. These operations collectively enable character matching. After aligning the predicted and ground truth words, we compute the character-level reliability score by comparing the character confidence score with correctness.

Prediction: an()imles
Ground Truth: animales

(a) Recognition Result

insert *space* at position 2

Position	0	1	2	3	4	5	6	7
Prediction	a	n		m	a	l	e	s
	↓	↓	↓	↓	↓	↓	↓	↓
Ground Truth	a	n	i	m	a	l	e	s

(b) Matching Process

Figure 1. An example of the calculation process for character-level matching. (a) represents the recognition results (Prediction) of super-resolved images using CRNN [12] and the ground truth. (b) illustrates the character-level matching process. Blue represents an example of a missing character.

B. Detailed Descriptions of Degradations

We conduct manual degradation in two degrees: one with light degradation and the other with severe degradation. Each degradation also comprises two cases: one includes Gaussian blur and Gaussian noise, and the other adds JPEG compression degradation to them. Table 1 presents the hyperparameters for blur, noise, and JPEG compression. For the added noise values, we report the average results of 5 experiments conducted with random seeds from 1 to 5, accounting for their inherent randomness.

Method	Component	Light Degradation	Severe Degradation
Gaussian Blur	kernel size	3	5
	sigma	5	6
Gaussian Noise	mean	0.0025	0.005
	std	0.0075	0.015
JPEG Compression	image quality	40	40

Table 1. Hyperparameters for manual degradation. We conduct experiments with two levels of degradation: light and severe.

Method	PSNR	SSIM
Bicubic	20.35	0.6961
TSRN [15]	21.42	0.7690
TBSRN [2]	20.91	0.7603
TG [3]	21.10	0.7341
TPGSR [8]	20.97	0.7719
TPGSR-3 [8]	21.18	0.7774
DPMN (+TATT) [17]	21.49	0.7925
C3-STISR [16]	21.51	0.7721
TATT [9]	21.52	0.7930
TATT [9] w/ Ours	21.53	0.7925
LEMMA [5]	20.90	0.7792
LEMMA [5] w/ Ours	20.55	0.7707

Table 2. Experimental results of PSNR and SSIM compared with mainstream STISR methods.

C. Limitations of Evaluation Metrics

Recently, metrics that evaluate STISR are changing from PSNR and SSIM, which evaluate visual quality, to text recognition accuracy. An inherent drawback of existing PSNR and SSIM metrics is their tendency to yield low scores, even when the image is better suited to represent text, owing to the presence of noise or degradation in the ground truth image. In particular, for the above two metrics, there is an inherent problem because the results are produced through comparison with HR images. Table 2

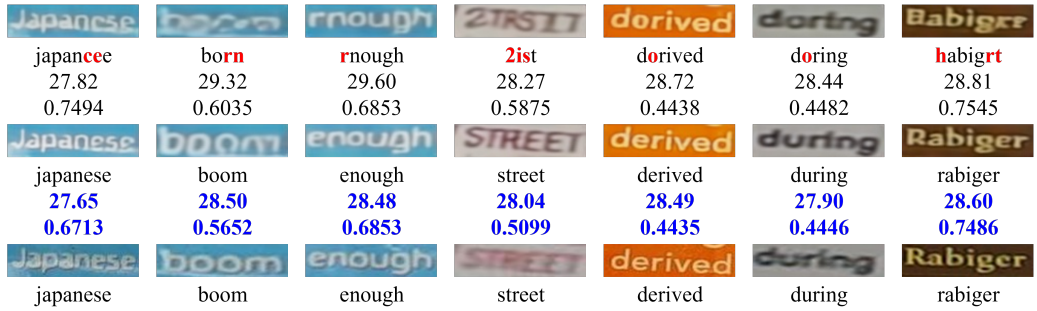

Figure 2. Limitations of visual metrics. Despite showing significantly better qualitative results due to noise present in the ground truth images, it exhibits low PSNR and SSIM values. This demonstrates the difficulty of using visual metrics as indicators of good performance. The experiment is based on TATT [9], where the super-resolved output is recognized using CRNN [12].

Parameter	**TATT [9]**	**LEMMA [5]**
Prior generator	PARSeq [1]	ABINet [4]
Batch size	64	64
Learning rate	10^{-3}	10^{-3}
Optimizer	Adam	Adam
Weight decay factor	-	0.5 after 400 epochs
Training epochs	500	500
Embedding dims	384	512
Alphabet set	'a' to 'z', 'A' to 'Z', '0' to '9', and punctuation symbols	'a' to 'z' and '0' to '9'
Image loss	$\|x_{SR} - x_{HR}\|_2$	$\|x_{SR} - x_{HR}\|_2$
Structure loss	$1 - TSSIM(DF(Y), F(DY), DX)$	-
Recognizer loss	-	$\|A_{SR} - A_{HR}\|_1 + WCE(p_{SR}, y_{gt})$
Distillation loss (removed)	$\|p^t_{HR} - p^s_{LR}\|_1 + KL(p^s_{LR} - p^t_{HR})$	-
Fine-tuning loss (removed)	-	$CE(p_{LR}, y_{gt})$
Loss function	$(1-\alpha) \cdot CE(p^s_{LR}, y_{gt}) + \alpha \cdot KL(p^s_{LR}(\tau), p^t_{HR}(\tau))$	

Table 3. Implementation details. We replaced the conventional distillation loss and fine-tuning loss used for training the prior generator in TATT [9] and LEMMA [5] with a mixed loss comprising cross-entropy loss and softened Kullback-Leibler divergence loss. In the image loss, x_{SR} represents the output image from the super-resolution network, and x_{HR} represents the high-resolution ground truth image.

displays low PSNR and SSIM values when our proposed method is applied. However, Figure 2 is an example of an image that is clearer or contains more structural detail despite lower performance in PSNR or SSIM metrics. We demonstrate high generalization performance by placing a greater emphasis on text recognition accuracy and evaluating across diverse scene text recognition datasets.

D. More Details for Implementation

We conduct experiments by applying our proposed method to two baseline methods, TATT [9] and LEMMA [5]. Therefore, we would like to introduce the implementation details used for each method. Table 3 summarizes the details for each method, including the character recognizer used, batch size, learning rate, optimizer, weight decay factor, epochs, embedding dimension of penultimate layer representations, alphabet set, and loss function. In the loss function, we replaced it with a unified loss that mixes the distillation loss and fine-tuning loss used in each method.

E. Hyperparameter Analysis

We analyze the impact of the hyperparameters τ and α, which are associated with our proposed loss function. We conduct experiments on the hyperparameters using LEMMA [5] with our method, and the same parameters are applied to TATT [9] with our method.

τ is a parameter that adjusts the smoothness of the Kullback-Leibler (KL) divergence loss in the proposed combined softened KL divergence loss and cross-entropy loss. When τ is greater than 1, the larger it gets, the smoother the distribution becomes. Table 4 shows the evaluation results on the TextZoom [15] dataset according to different τ values. The results indicate that a τ value of 3 generally yields the best performance.

τ	**ASTER [13]**	**MORAN [7]**	**CRNN [12]**	**Average**
1	66.1%	64.5%	56.8%	62.5%
3	**67.9%**	65.0%	**58.1%**	**63.7%**
5	66.9%	**65.1%**	58.0%	63.3%
10	67.4%	64.2%	**58.1%**	63.3%

Table 4. Recognition accuracies when adjusting the smoothness of the KL divergence loss. Average refers to average accuracy.

α is a parameter that adjusts the ratio between hard labels and soft labels. As α increases, the influence of soft labels becomes greater; conversely, a decrease in α increases the influence of hard labels. Table 5 shows the evaluation results on the TextZoom [15] dataset according to different α values. The results indicate that an α value of 0.5 generally yields the best performance.

α	ASTER [13]	MORAN [7]	CRNN [12]	Average
0.0	66.0%	63.2%	56.3%	61.8%
0.3	66.7%	64.7%	57.6%	63.0%
0.5	**67.9%**	65.0%	58.1%	**63.7%**
0.7	67.5%	**65.2%**	**58.2%**	63.6%
1.0	66.6%	64.1%	57.1%	62.6%

Table 5. Recognition accuracies adjusting the ratio between hard labels and soft labels.

F. Visualizations on Scene Text Recognition Datasets

To compare how well generalization is achieved, we provide visualizations and results for the scene text recognition dataset. Figure 3 shows visualization results for the scene text recognition dataset. Table 6 presents the experimental results of adding JPEG compression to manual degradation using existing Gaussian blur and Gaussian noise.

References

[1] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022. 2

[2] Jingye Chen, Bin Li, and Xiangyang Xue. Scene text telescope: Text-focused scene image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2021. 1

[3] Jingye Chen, Haiyang Yu, Jianqi Ma, Bin Li, and Xiangyang Xue. Text gestalt: Stroke-aware scene text image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 285–293, 2022. 1

[4] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. 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*, pages 7098–7107, 2021. 2

[5] Hang Guo, Tao Dai, Guanghao Meng, and Shu-Tao Xia. Towards robust scene text image super-resolution via explicit location enhancement. *arXiv preprint arXiv:2307.09749*, 2023. 1, 2, 4

[6] 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 *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 4

[7] Canjie Luo, Lianwen Jin, and Zenghui Sun. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90:109–118, 2019. 2, 3

[8] Jianqi Ma, Shi Guo, and Lei Zhang. Text prior guided scene text image super-resolution. *IEEE Transactions on Image Processing*, 32:1341–1353, 2023. 1

[9] Jianqi Ma, Zhetong Liang, and Lei Zhang. 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*, pages 5911–5920, 2022. 1, 2, 4

[10] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA, 2012. 4

[11] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 569–576, 2013. 4

[12] Baoguang Shi, Xiang Bai, and Cong Yao. 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. 1, 2, 3, 4

[13] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2035–2048, 2018. 2, 3

[14] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *2011 International conference on computer vision*, pages 1457–1464. IEEE, 2011. 4

[15] Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 650–666. Springer, 2020. 1, 2, 3

[16] Minyi Zhao, Miao Wang, Fan Bai, Bingjia Li, Jie Wang, and Shuigeng Zhou. C3-stisr: Scene text image super-resolution with triple clues. *arXiv preprint arXiv:2204.14044*, 2022. 1

[17] Shipeng Zhu, Zuoyan Zhao, Pengfei Fang, and Hui Xue. Improving scene text image super-resolution via dual prior modulation network. *arXiv preprint arXiv:2302.10414*, 2023. 1

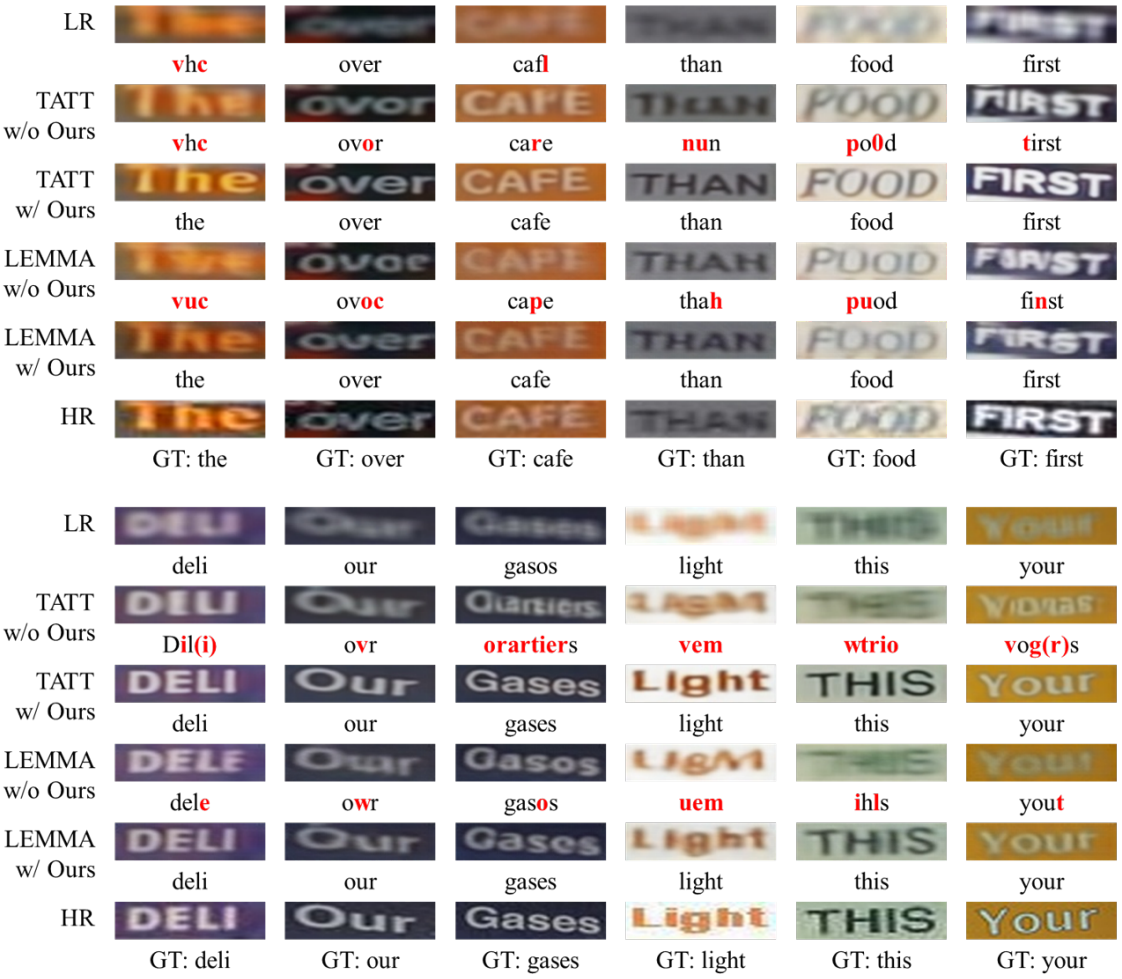

Figure 3. Visualization of SR images and their recognition results on scene text recognition datasets by CRNN [12], based on TATT [9] and LEMMA [5]. Red characters indicate wrong recognition results.

Method	Light Degradation				Severe Degradation			
	IIIT5K [10]	IC15 [6]	SVT [14]	SVTP [11]	IIIT5K [10]	IC15 [6]	SVT [14]	SVTP [11]
BICUBIC	21.0%	19.4%	0.5%	52.0%	1.1%	0.1%	0.0%	9.2%
TATT [9]	45.3%	35.4%	20.0%	72.3%	11.9%	9.5%	2.8%	28.8%
TATT [9] w/ Ours	**51.0%**	**46.4%**	**49.1%**	**76.3%**	**18.0%**	**12.7%**	**3.2%**	**31.4%**
Δ	**+5.8%**	**+10.9%**	**+29.1%**	**+4.1%**	**+6.0%**	**+3.3%**	**+0.3%**	**+2.6%**
LEMMA [5]	16.8%	15.5%	16.6%	32.1%	6.4%	2.7%	0.0%	17.3%
LEMMA [5] w/ Ours	**49.4%**	**46.5%**	**41.8%**	**72.3%**	**15.5%**	**12.2%**	**0.3%**	**34.4%**
Δ	**+32.6%**	**+31.0%**	**+25.2%**	**+45.1%**	**+9.1%**	**+9.5%**	**+0.3%**	**+17.1%**

Table 6. Recognition accuracies by CRNN [12] for the manual degradation scene recognition datasets. Manual degradation includes Gaussian blur, Gaussian noise, and JPEG compression.