StyleSRN: Scene Text Image Super-Resolution with Text Style Embedding

Supplementary Material

In this supplementary file, we provide:

1. Detailed descriptions of five manually degraded scene text recognition datasets.
2. More ablation experiment results.
3. More comparison with State-of-the-arts.

1. Scene Text Recognition Datasets

1.1. Description of Scene Text Recognition Datasets

The ICDAR13 [5] test dataset predominantly inherits data from IC03 [7], comprising 1,015 ground-truth cropped word images. ICDAR15 [6] consists of scene images and is available in two versions: 1,811 images (IC15$_S$) and 2,077 images (IC15$_L$). For our experiments, we utilized the IC15$_S$ subset. This dataset is characterized by images that are noisy, blurred, contain complex backgrounds, and often feature low resolution, posing significant challenges even for human text recognition. CUTE80 [14] comprises 288 images with heavily curved text. The SVTP (Street View Text Perspective) [13] dataset contains 645 images with texts captured in perspective views. IIIT5K [11] includes 5,000 cropped word images, divided into 2,000 training images and 3,000 testing images. The test set was used for our experiments.

1.2. Degradation Settings

Following previous studies [4, 9], we manually degraded the images from the five datasets to simulate various real-world degradation processes. Initially, we applied pre-blurring using randomly selected Gaussian kernels of size 3×3 or 5×5 with σ uniformly sampled within the range $[5, 6]$. Next, with an 80% probability, random Gaussian noise was added to the images. Subsequently, a Gaussian blur was applied with a 70% probability, using σ uniformly sampled between $[2, 3]$ for noise reduction; otherwise, bilateral filtering was employed. Finally, the images were sharpened using Gaussian kernels of size 3×3 or 5×5 with σ uniformly sampled within the range $[2, 3]$ to produce the degraded results.

2. More Ablation Studies

In this section, we conduct a comprehensive series of ablation studies to further analyze the effectiveness of the proposed components and design choices in our method. Specifically, we begin by evaluating the impact of balancing parameters in the overall loss function, focusing on the contribution of the proposed Text Style Loss to image quality and text recognition accuracy. Next, we investigate the influence of multi-scale 1D convolution kernels, comparing fixed-size kernels with various multi-scale combinations to highlight their importance in capturing diverse text features. Finally, we examine the role of different text prior generators, demonstrating how varying text priors influence the model's ability to reconstruct both the structural and stylistic features of text. These studies provide a deeper understanding of the key factors contributing to the performance of our StyleSRN model and validate the robustness of our design choices.

β	Quality Metric		Accuracy(%)		
	PSNR	SSIM	CR [15]	MO [8]	AS [16]
0	20.34	0.7611	54.9	60.8	63.8
0.1	20.96	0.7663	55.6	62.5	65.7
0.5	21.82	0.7778	56.1	63.5	66.3
5	21.71	0.7768	55.0	61.6	65.0
1	**21.87**	**0.7784**	**57.4**	**64.1**	**67.3**

Table 1. Ablation studies on different β values. AS, MO and CR refer to ASTER [16], MORAN [8] and CRNN [15], respectively.

2.1. Selection on the Balancing Parameters of the Text Style Loss Function

In the Overall Loss section of the main text, the overall loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_2 + \alpha \mathcal{L}_{TP} + \beta \mathcal{L}_{style} \tag{1}$$

where α and β are balancing parameters.

Following previous research[4, 9, 17], we initially fixed α at 0.5. We then varied the coefficient β of our proposed Text Style Loss to identify the optimal value in terms of recognition accuracy and image quality metrics. As shown in Table 1, both the average recognition accuracy and SSIM improve as β increases. However, when β exceeds 1, all evaluation metrics decline, indicating that an excessive emphasis on text style loss can detrimentally affect the quality of text reconstruction. Consequently, we set β to 1.

2.2. Impact of Multi-Scale 1D Convolution Kernels

To evaluate the impact of different 1D convolution kernel configurations on the performance of our model, we conducted an ablation study using fixed-size kernels (3×3, 5×5, 7×7) and various multi-scale combinations. The results are presented in Table 2, which reports PSNR, SSIM, and recognition accuracy using CRNN [15] on the TextZoom dataset [17]. The findings demonstrate that single fixed-size kernels achieve acceptable performance, with the 3×3 kernel obtaining a PSNR of 21.70 and an SSIM of 0.7751,

Kernel Size	Kernel Type	Quality Metric		Accuracy
		PSNR	SSIM	
3×3	fixed	21.70	0.7751	56.5
5×5	fixed	21.63	0.7755	56.3
7×7	fixed	21.55	0.7750	56.1
3×3, 5×5	multi-scale	21.78	0.7754	56.4
3×3, 7×7	multi-scale	21.71	0.7757	56.7
5×5, 7×7	multi-scale	21.75	0.7765	56.9
3×3, 5×5, 7×7	multi-scale	**21.82**	**0.7778**	**57.4**

Table 2. Ablation study on multi-scale 1D convolution kernels. Accuracy represents the accuracy of CRNN [15] on TextZoom [17].

outperforming the larger 5×5 and 7×7 kernels slightly. This suggests that smaller kernels may better capture fine-grained features relevant to text reconstruction. However, the performance improves when combining multiple kernel sizes. For instance, the combination of 5×5 and 7×7 kernels increases SSIM to 0.7765 and recognition accuracy to 56.9%. The best results are achieved using all three kernel sizes (3×3, 5×5, and 7×7) simultaneously, yielding a PSNR of 21.82, an SSIM of 0.7778, and a recognition accuracy of 57.4%. These results highlight the effectiveness of multi-scale 1D convolution kernels in capturing diverse text features and improving both image quality and text recognition accuracy. The improvement achieved by multi-scale kernels can be attributed to their ability to model cross-channel dependencies at different scales, thereby enhancing the representation of stylistic and structural features in scene text images. These results validate the design choice of incorporating multi-scale 1D convolution kernels into our StyleSRN architecture.

2.3. Impact of Different Text Prior Generators

We assessed the effect of generating text priors on model performance using six widely used scene text recognizers, including CRNN [16], MORAN [8], ASTER [16], ABINet [3], MATRN [12], and PARSeq [1]. The impact of different text prior generators on our method is presented in Table 3. The results indicate that more powerful text recognizers lead to a significant improvement in recognition accuracy, which is intuitive since they can generate more accurate text priors. However, the image quality metrics do not exhibit a corresponding significant improvement. Additionally, more powerful text recognizers such as ABINet, MATRN, and PARSeq have higher parameter counts and computational complexity, which reduces the inference speed of the STISR model. Therefore, we selected ASTER as the text prior generator to better balance the trade-off between image quality, recognition accuracy, and inference speed.

TPG	Quality Metric		Accuracy(%)		
	PSNR	SSIM	CR [15]	MO [8]	AS [16]
CRNN [15]	21.65	0.7757	55.1	62.6	65.4
MORAN [8]	21.77	0.7765	56.5	63.6	66.4
ASTER [16]	**21.82**	**0.7778**	57.4	64.3	67.3
ABINet [3]	21.75	0.7766	57.9	65.0	68.1
MATRN [12]	21.80	0.7763	58.0	65.1	68.4
PARSeq [1]	21.69	0.7770	**58.5**	**66.6**	**69.2**

Table 3. Comparison on Image Quality Metric and Recognition Accuracy using different Text Prior Generator(TPG) to produce text prior. AS, MO and CR refer to ASTER [16], MORAN [8] and CRNN [15], respectively.

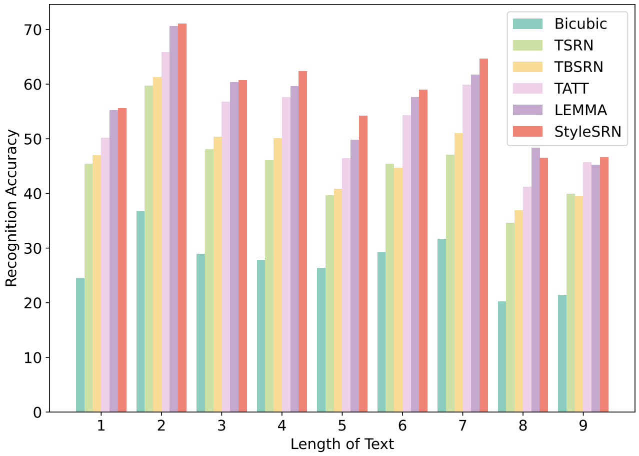

Figure 1. Recognition accuracy with different text lengths on TextZoom.

3. More comparison with State-of-the-arts

3.1. Experiment on Scene Text Images of Different Lengths

In real-world scenarios, the length of scene texts varies, and recognizing longer texts can be particularly challenging. Therefore, evaluating the ability of STISR methods to handle scene texts of different lengths is crucial for measuring their overall performance. We compared the performance of different STISR methods on texts of varying lengths using the TextZoom [17] dataset. As illustrated in Figure 1, the proposed method performs well across almost all text lengths, demonstrating its robustness in dealing with scene texts of varying lengths.

3.2. Comparison Results on Image Quality on Five Scene Text Recognition Datasets

The comparative results of PSNR and SSIM are presented in Table 4. Our method demonstrates competitiveness in SSIM and outperforms other methods in PSNR. This performance may be attributed to our method's emphasis on recovering text style information, which involves making

Method	ICDAR13 [5]		ICDAR15 [6]		CUTE80 [14]		SVTP [13]		IIIT5K [11]	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
TSRN [17]	21.74	**0.8366**	22.89	**0.8338**	20.48	**0.8167**	22.21	**0.8086**	19.07	**0.7567**
TBSRN [2]	21.62	0.8301	22.31	0.8164	20.73	0.8151	21.41	0.7877	19.00	0.7587
TPGSR [10]	21.47	0.8211	21.97	0.8124	19.54	0.7990	21.45	0.7861	18.50	0.7396
TATT [9]	21.08	0.8304	21.46	0.8212	18.77	0.8096	20.87	0.7961	18.13	0.7533
LEMMA [4]	21.46	0.8159	21.90	0.8038	18.80	0.7868	21.28	0.7809	17.95	0.7334
StyleSRN	**22.71**	0.8230	**23.77**	0.8125	**21.21**	0.8033	**23.01**	0.7886	**19.58**	0.7490

Table 4. Comparison of Image Quality on Six Scene Text Recognition datasets. **Bold text** and underlined text indicate the best and the second-best performances, respectively.

Method	Accuracy of ABINet [3](%)				Accuracy of MATRN [12](%)				Accuracy of PARSeq [1](%)			
	Easy	Medium	Hard	Average	Easy	Medium	Hard	Average	Easy	Medium	Hard	Average
BICUBIC	77.9	57.0	42.7	60.3	80.9	59.0	45.1	62.8	**90.2**	**75.5**	**56.8**	**75.2**
TBSRN [2]	79.8	65.0	48.5	65.4	81.7	66.0	50.1	66.9	83.7	66.7	51.8	68.4
TATT [9]	80.7	65.8	50.3	66.5	81.1	66.6	51.7	67.4	82.2	65.9	52.1	67.7
C3-STISR [18]	81.4	66.9	49.9	67.0	81.9	68.0	51.1	68.0	84.3	68.3	50.9	68.8
LEMMA [4]	82.2	69.2	50.6	68.5	82.8	70.4	51.7	69.3	83.6	69.2	52.3	69.3
StyleSRN	**86.2**	**69.5**	**53.8**	**70.9**	**86.5**	**70.8**	**54.1**	**71.5**	87.5	71.5	55.0	72.4

Table 5. Comparison with the existing methods in terms of the recognition accuracy on TextZoom [17]. **Bold text** and underlined text indicate the best and the second-best performances, respectively.

certain trade-offs in the preservation of character structure integrity. To qualitatively illustrate the superiority of our method in restoring image style, we also provide visualizations from the five STR datasets in Figure 2 and Figure 3. These results clearly show that our method produces outputs with sharper, more realistic edges, textures, and shadows. Moreover, for texts with distinctive font styles, our method achieves superior restoration of the font style.

3.3. Comparison with More Advanced Recognizers

In the Experiment section of the main text, we utilized three classic text recognizers to evaluate the impact of STISR methods on STR performance. Here, we extend this comparison by employing more advanced text recognizers, including ABINet [3], MATRN [12], and PARSeq [1], to further assess the performance of STISR methods. As shown in Table 5, when ABINet and MATRN are used as text recognizers, all STISRs exhibit significant performance improvements over BICUBIC. Notably, our method surpasses previous approaches in performance. However, when PARSeq is employed, the results differ slightly, with all STISR methods experiencing a certain degree of performance decline. We attribute this to the fact that some state-of-the-art text recognizers are capable of processing low-resolution images, rendering the existing STISR frameworks less effective with such advanced recognizers. Nevertheless, as illustrated in Table 5, our method produces fewer misleading results compared to previous methods.

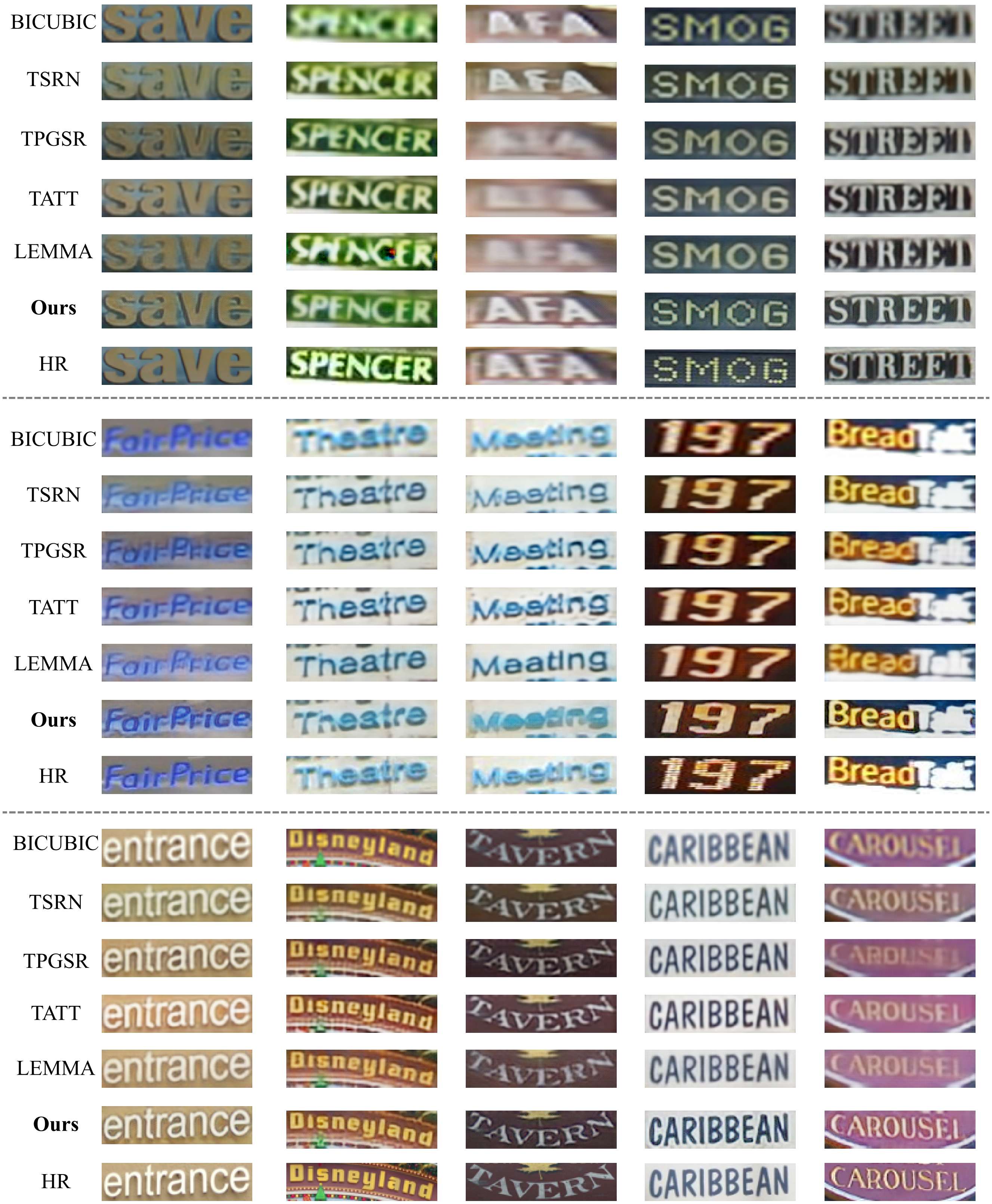

Figure 2. Visualized results on IC13[5] ,IC15[6] and CUTE80[14]

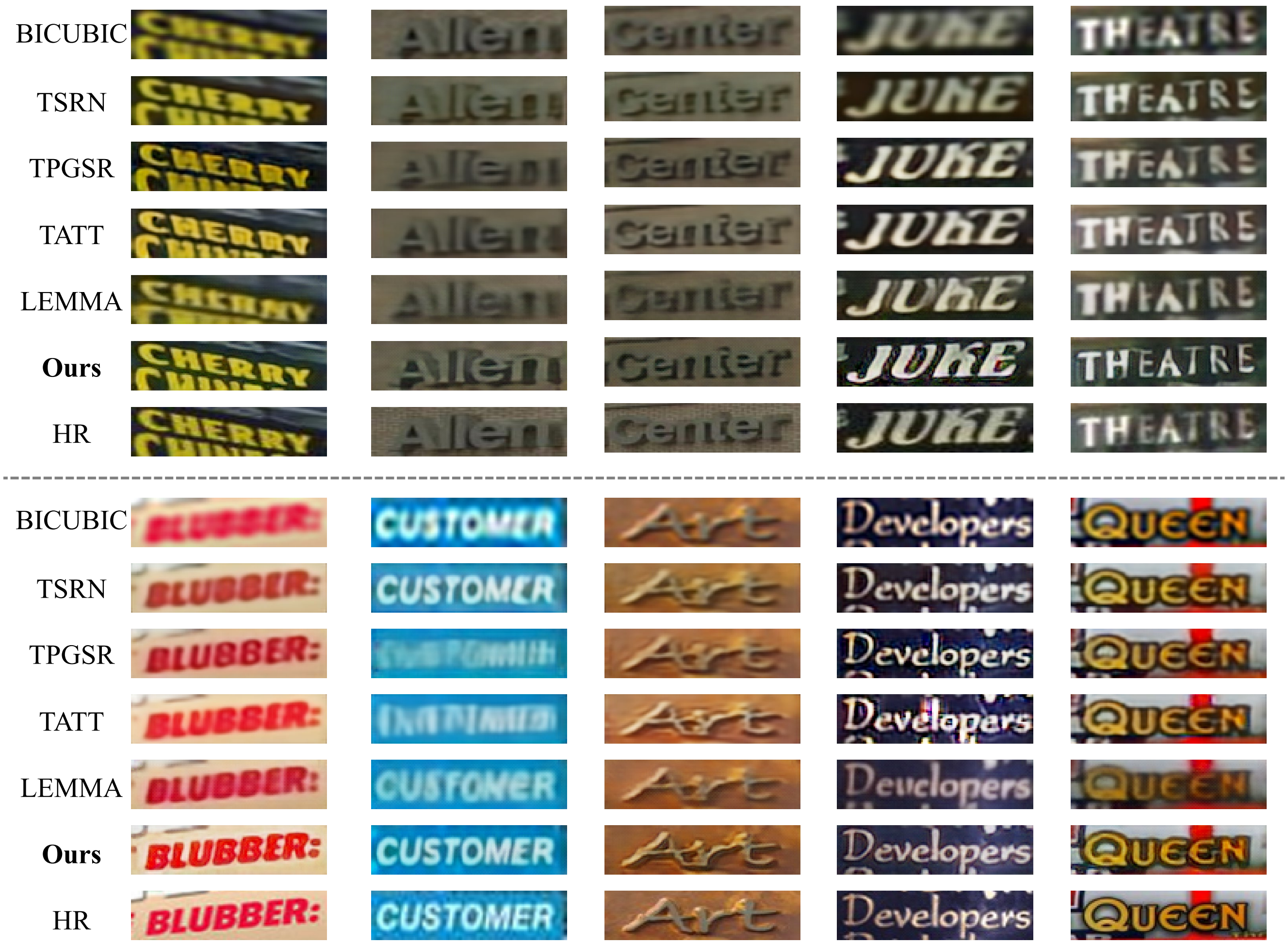

Figure 3. Visualized results on SVTP[13] and IIIT5K[11]

References

[1] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *Proceedings of the European Conference on Computer Vision*, pages 178–196, 2022. 2, 3

[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. 3

[3] 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, 3

[4] Hang Guo, Tao Dai, Guanghao Meng, and Shu-Tao Xia. Towards robust scene text image super-resolution via explicit location enhancement. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 782–790, 2023. 1, 3

[5] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *Proceedings of the 12th International Conference on Document Analysis and Recognition*, pages 1484–1493, 2013. 1, 3, 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 *Proceedings of the 13th International Conference on Document Analysis and Recognition*, pages 1156–1160, 2015. 1, 3, 4

[7] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, and R. Young. Icdar 2003 robust reading competitions. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*, 2003. 1

[8] Canjie Luo, Lianwen Jin, and Zenghui Sun. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90:109–118, 2019. 1, 2

[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, 3

[10] Jianqi Ma, Shi Guo, and Lei Zhang. Text prior guided scene text image super-resolution. *IEEE Transactions on Image Processing*, 32:1341–1353, 2023. 3

[11] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *Proceedings of the British Machine Vision Conference*, pages 1–11, 2012. 1, 3, 5

[12] Byeonghu Na, Yoonsik Kim, and Sungrae Park. Multi-modal text recognition networks: Interactive enhancements between visual and semantic features. In *Proceedings of the European Conference on Computer Vision*, pages 446–463, 2022. 2, 3

[13] 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. 1, 3, 5

[14] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18):8027–8048, 2014. 1, 3, 4

[15] 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

[16] 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. 1, 2

[17] Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *Proceedings of the European Conference on Computer Vision*, pages 650–666, 2020. 1, 2, 3

[18] Minyi Zhao, Miao Wang, Fan Bai, Bingjia Li, Jie Wang, and Shuigeng Zhou. C3-STISR: scene text image super-resolution with triple clues. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 1707–1713, 2022. 3