Learning Generative Structure Prior for Blind Text Image Super-resolution (Supplementary Material)

Xiaoming Li[1] Wangmeng Zuo[2] Chen Change Loy[1](✉)
[1]S-Lab, Nanyang Technological University [2]Harbin Institute of Technology
csxmli@gmail.com wmzuo@hit.edu.cn ccloy@ntu.edu.sg

This supplemental material mainly contains:

- Analyses of model complexity in Section I
- Architecture details in Section II
- Broader impact in Section III
- Structure images of 182 font families that are used in synthesizing the HR text images in Figure A
- More visual comparison on synthetic and real-world LR text images in Figures B, C, D and E
- Analyses of recognition prior in competing methods and our style vector w in Figure F
- Analyses of failure cases in Figure G.

I. Model Complexity

The whole framework of our MARCONet consists of three parts, 1) learnable codebook and StyleGAN with 27.97 M parameters, 2) Transformer encoder with 59.66 M parameters, and 3) SR network with 16.86 M parameters. It takes 0.21s on average to super-resolve a 32×512 LR input to a 128×2048 SR result with one Tesla V100 GPU.

II. Architecture Details

The Transformer encoder follows the default settings of the vanilla Transformer [5]. Here we mainly show the architectural details of the structure prior transform module and final reconstruction module. The StyleGAN used for text images in this work has $i \in \{8, 16, 32, 64, 128\}$ scales in total and each scale i represents the size of intermediate features. We adopt scales $i \in \{32, 64\}$ on the multi-scale structure prior transform modules. Details are presented in Table A. SNConv (d, k, s) represents a convolutional layer followed by a spectral normalization [3] , where d, k and s are output dimension, kernel size and stride, respectively. LReLU (c) is leaky ReLU with a negative slope c. Swish is a self-gated activation [4] and is defined as $f(x) = x \cdot \text{sigmoid}(x)$.

III. Broader Impact

Our text super-resolution framework is of great value in real-world scenarios, *e.g.*, the restoration of old printed media for digitalization and preservation, and the enhancement of captions in digital media, license plates, and invoices. It can also be employed in image acquisition devices to enhance the visual quality of text regions and further improve optical character recognition (OCR). As with other computational imaging techniques, the text SR technology may be misused, causing information leakage or other issues in privacy. We note that the accuracy of text SR will be compromised under severe degradation, causing errors in the final character output. Such an error may be unacceptable for forensic document analysis. It is worth noting that the positive impact of text SR outweighs the potential problems. We call on people to use this technology without harming personal information and take the SR result as a reference.

Table A. Architecture details of our multi-scale structure prior transform modules and final reconstruction module.

Structure Prior Transform on 32×32		
Input	LR Feature	Character Structure Prior
RoIAlign		
For Each Character	F_{LR}^c	$\mathcal{P}^c$
	AdaIN $(F_{LR}^c, \mathcal{P}^c)$ Concatenate output with F_{LR}^c	
ResBlock	GroupNorm, Swish, SNConv (256, 3, 1) GroupNorm, Swish, SNConv (256, 3, 1) Element-wise Addition	
Spatial Feature Transform	SNConv (256, 3, 1) LReLU (0.2) SNConv (256, 3, 1) Scale α	SNConv (256, 3, 1) LReLU (0.2) SNConv (256, 3, 1) Shift β
	$\alpha \cdot F_{LR}^c + \beta$	
Reverse RoIAlign		
UpSample (32→64)	Bilinear UpSample SNConv (256, 3, 1), LReLU (0.2) ResBlock, SNConv (256, 3, 1)	
Structure Prior Transform on 64×64 (Kindly refer to the former steps)		
Reconstruction (64→128)	SNConv (128, 3, 1), LReLU (0.2) Bilinear Upsample SNConv (64, 3, 1), LReLU (0.2) ResBlock SNConv (3, 3, 1) Tanh	
Output	SR Result	

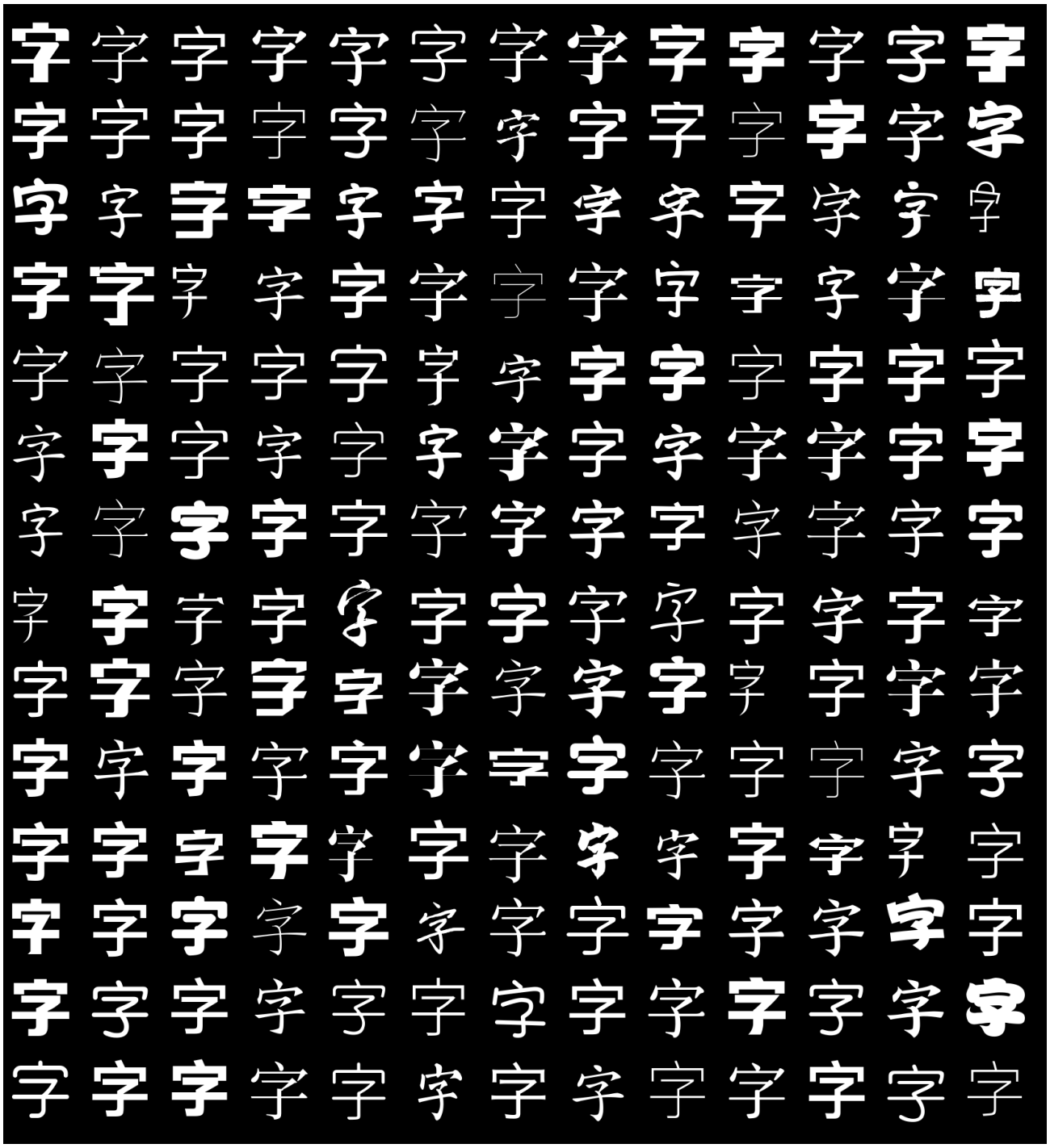
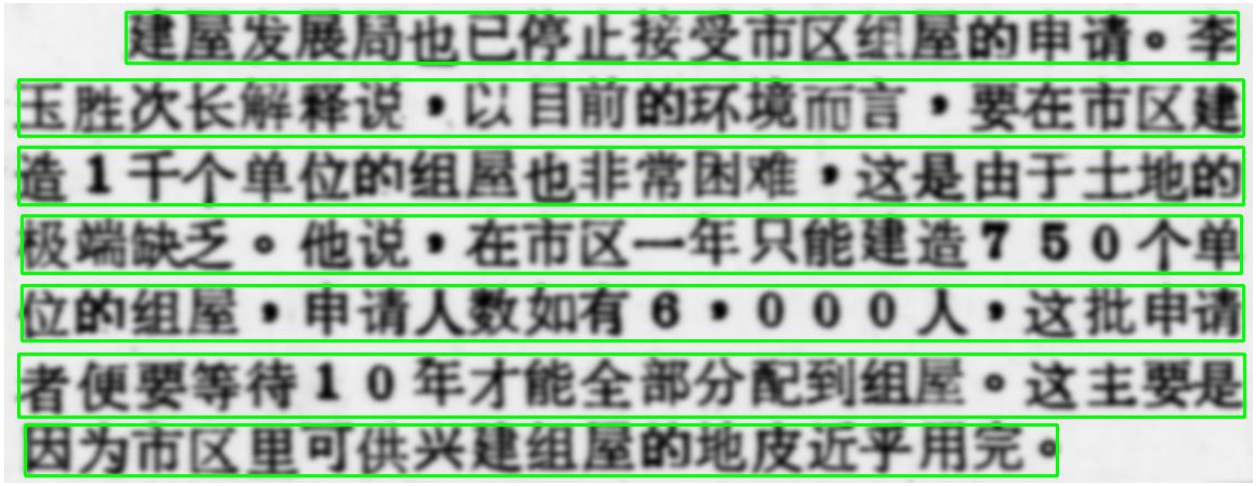


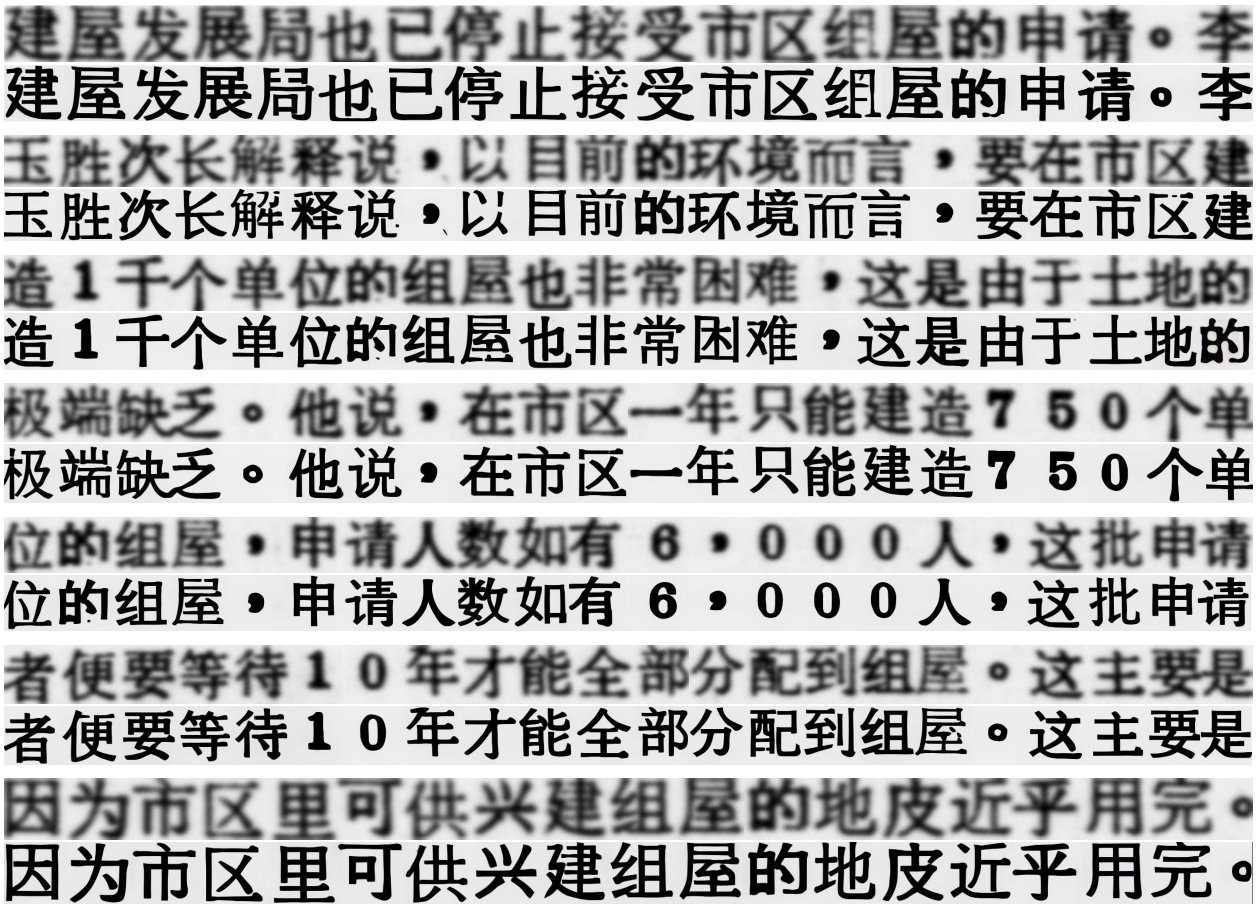
Figure A. Structure images of the 182 font families used in synthesizing our training data.

Figure B. More visual comparison with competing methods (*i.e.*, retrained TBSRN [1] and TATT [2]) on our synthetic LR text images.

Figure C. Results obtained with retrained TBSRN [1], TATT [2] and our approach on several segments of real-world LR text images.



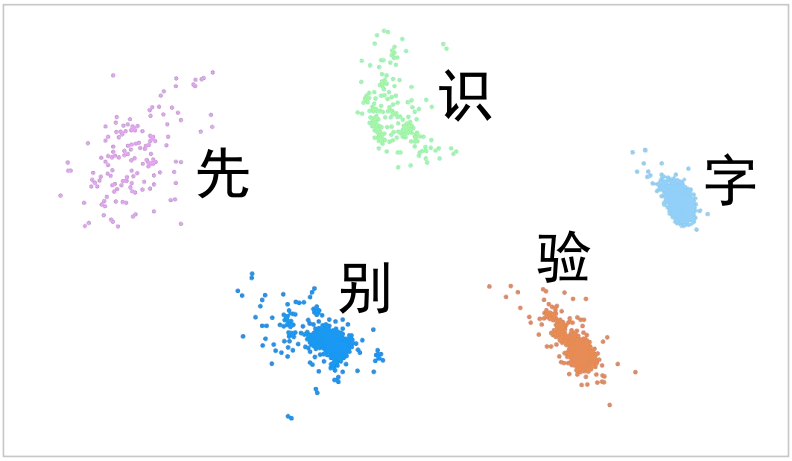
(a) Real-world LR segment from an old newspaper



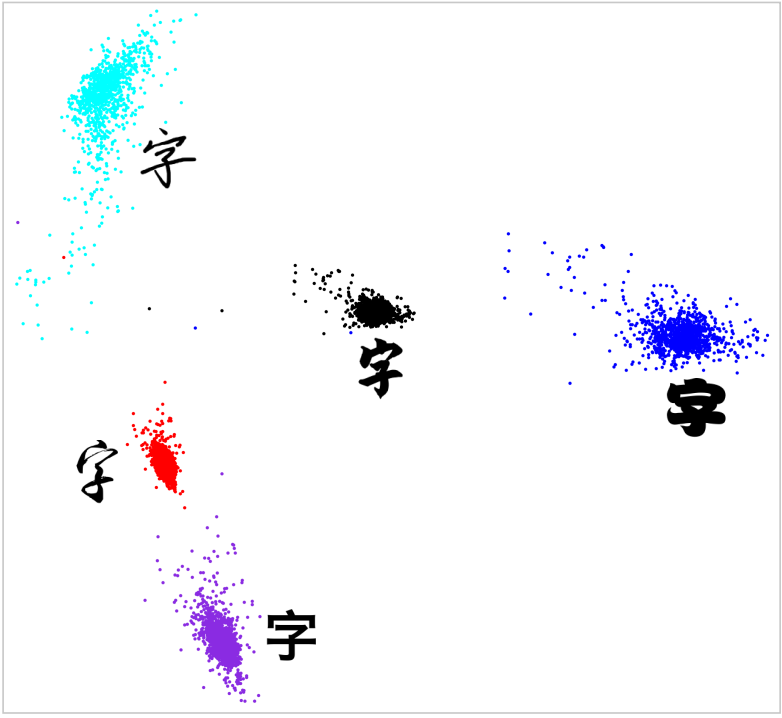
(b) Restoration results of our MARCONet

Figure D. Results of our approach on super-resolving the segments from a real-world LR newspaper.

Figure E. More results of our approach on super-resolving the segments from different LR sources.

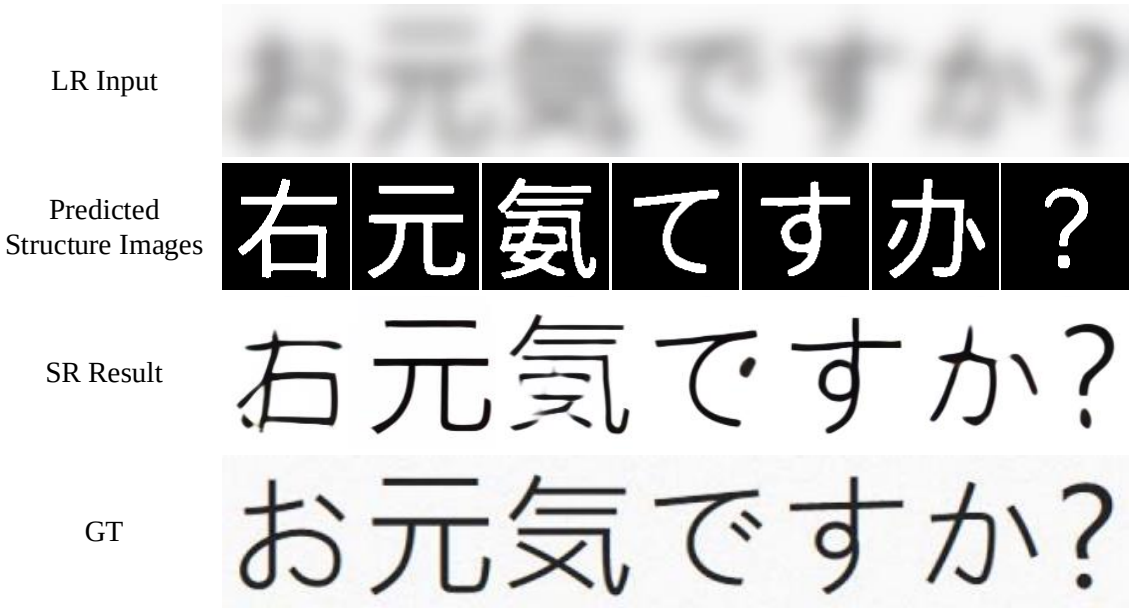


(a) The t-SNE results on recognition prior from the final layer of the character recognizer. We adopt 5 characters (see them in the plot) and synthesize their images with random degradation and 182 font families, respectively.

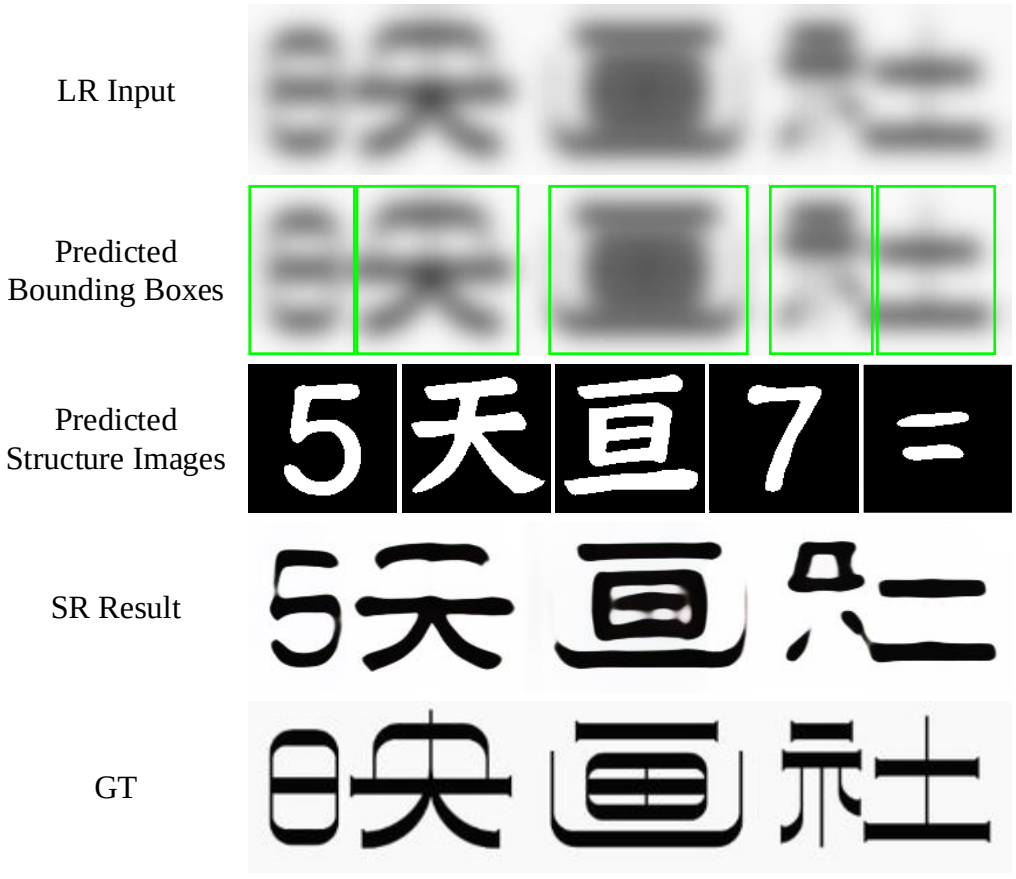


(b) The t-SNE results on our style space W. We build 5 groups of text images and each group has the same font family but is synthesized with random text and degradation (see their font style examples in the plot).

Figure F. Analyses of recognition prior used in competing methods and our style vector w. We can observe that (a) recognition prior can well distinguish each character but is not enough to aid SR task because it easily ignores the font styles which are crucial for super-resolving a faithful and accurate structure. So these recognition prior based methods (*i.e.*, TBSRN [1] and TATT [2]) have limited performance on this task; (b) in contrast, the style vector w in our framework can easily capture the styles of different font families, no matter what text and degradation they have. By combining style vector w with the codebook, our MARCONet can well benefit the SR task by providing accurate structure-level guidance.



(a) Results of our MARCONet which is trained on Chinese but is adopted on Kanji text image SR



(b) Results of our MARCONet on LR text image which has unusual font style

Figure G. Analyses of failure cases. We can observe that (a) when directly using our MARCONet on another language (*i.e.*, Kanji), the error in predicting character classification easily leads to poor guidance for the SR results, (b) when the LR input has an unusual font style, our MARCONet easily has an inaccurate prediction in character classification or bounding box regression, which easily results in the inconsistent structures on the SR results. For this case, we will add more font families in our training phase to improve the generalization performance.

References

[1] Jingye Chen, Bin Li, and Xiangyang Xue. Scene text telescope: Text-focused scene image super-resolution. In *CVPR*, 2021. 3, 4, 7

[2] Jianqi Ma, Zhetong Liang, and Lei Zhang. A text attention network for spatial deformation robust scene text image super-resolution. In *CVPR*, 2022. 3, 4, 7

[3] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 1

[4] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 1

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 1