Restore Text First, Enhance Image Later: Two-Stage Scene Text Image Super-Resolution with Glyph Structure Guidance

Supplementary Material

7. Reproducibility Statement

To ensure reproducibility, we have made the following efforts:

1. We will release our code and dataset.
2. We provide implementation details in Sec. 5.1 and Sec. 11, including the training process and selection of hyper-parameters.
3. We provide details on evaluation metrics and dataset preparation in Sec. 4 and Sec. 10, and the code and data will be made available along with it.

8. Ethics Statement

This work focuses on improving scene text image super-resolution to support beneficial applications such as enhancing accessibility, document restoration, and navigation assistance. However, we acknowledge potential risks, including misuse in privacy-sensitive contexts (e.g., recovering text from personal or social media images) or unintended deployment in surveillance. To mitigate such risks, users are encouraged to combine our methods with privacy-preserving techniques such as watermarking or selective inpainting. Our UZ-ST dataset was collected from public, non-sensitive scenes under ethical guidelines, without including personally identifiable or private information. We believe the societal benefits of improved text image restoration outweigh potential risks, provided that the technology is applied responsibly.

9. LLM Acknowledgments

We thank ChatGPT (GPT-5) and DeepSeek for the help in refining the language and improving clarity. All wording and factual content were reviewed and approved by the authors.

10. Dataset Collection

We use VIVO X200 Ultra to collect images for 4 separate focal lengths (14 mm, 35 mm, 84 mm, and 200 mm). We first use PP-OCRV5 [7] for the rough annotation, then we manually filter the images and annotations. During the filtering and annotation, each image undergoes the following rules:

- Width or height of the image should be no less than 256.
- Height of the text should not be less than 32 pixels.
- Score of OCR recognition of the text should not be lower than 0.9.
- Content of the text should not be empty or consist solely of whitespace.

11. More Implementation Details

Dataset Settings. For training, we combine a synthetic dataset with real paired datasets (Real-CE [31] and UZ-ST). Our synthetic dataset builds upon LSDIR [26], containing 27,000 triplets (x_H, x_L, x_m). We render text on the GT of LSDIR and the corresponding text mask using the LBTS [38], then apply Real-ESRGAN degradation Wang et al. [44] to generate LR images. Since there are misaligned image pairs in Real-CE [19, 58] and text lines not annotated, we filter out misaligned image pairs and reannotate the images manually. In the end, we obtain 337 training pairs and 188 testing pairs from the Real-CE dataset. Images under 13mm and 52mm focal lengths are considered as LR x_L and GT x_H, respectively. Following Hu et al. [19], we use SAM-TS [49] to obtain the text mask x_m from x_H. UZ-ST triplets (x_H, x_L, x_m) are processed identically. For stage 1 training, we use cropped text regions from the dataset, while for stage 2, we use full images. Following [39], we use PP-OCRv3 [23] to extract strings from the text region of predicted SR for evaluation of OCR-A.

Training Details. For stage 1, we build our model based on the IDM baseline of DiffTSR [58]. We set λ_{td}, λ_{Seg}, λ_{Focal}, and λ_{Dice} as 1, 0.1, 20, and 1 respectively. In stage 2, our model is built upon Stable Diffusion 3.5 medium [16] and follows a similar tile-based inference strategy to TADiSR [19] and SUPIR [50]. The timestep t is set to 150. The conditioning scale of the ControlNet is set to 1.0. λ_{l2}, λ_{LPIPS}, and λ_{edge} are set to 1, 5, 100 when using the synthetic dataset for pretraining. Then, λ_{edge} is set to 0 and the remaining coefficients are kept unchanged when training on real paired datasets. We train both stages of the model using the AdamW [28] optimizer and set the learning rate to 5×10^{-5} and 5×10^{-6} for stages 1 and 2 separately. All experiments are conducted on NVIDIA H20 GPUs. For stage 1, we train the model on synthetic and real datasets for 8 epochs, then finetune it on synthetic data for 2 epochs. For stage 2, we first pretrain the model on the synthetic dataset for 50 epochs. Then we train on the real paired datasets for 50 epochs.

12. More details of UZ-ST

In Tab. 7, we provide detailed statistics on the composition of the UZ-ST dataset. Additionally, in Fig. 10, we present

Table 7. Statistics of dataset size and line count in subsets of UltraZoom-ST.

Subset	**image count**	**line count**	**mean lines/img**	**img > 5 lines**
14 mm	1,439	15,073	10.47	754
35 mm	1,798	17,263	9.60	867
85 mm	1,799	17,339	9.64	869

Table 8. Alignment comparison.

Alignment Method	PSNR ↑	MSE ↓	SSIM ↑	NCC ↑	AKD ↓
RealSR	13.87	4002.65	0.5072	0.3285	546.78
SIFT Alignment (Single time)	12.77	6332.79	0.4896	0.4313	546.78
Cascade Coarse-to-Fine Alignment (Ours)	**23.00**	**441.52**	**0.7279**	**0.9184**	**241.44**

some example images from the dataset.

13. Alignment Comparison

We compare our proposed Cascade Coarse-to-Fine alignment strategy with the alignment method proposed by RealSR [2] and single-time alignment using SIFT. During the comparison, we implement our strategy using SIFT and utilize the raw images from the 35mm dataset of our proposed UZ-ST dataset for evaluation. For metrics, we adopt PSNR, Mean Squared Error (MSE), SSIM [42], Normalized Cross-Correlation (NCC), and Average Keypoint Distance (AKD) to assess the alignment quality of the aligned image.

As shown in Tab. 8 and Fig. 6, pixel-based optimization methods break when handling heavily degraded images due to significant loss of details. Keypoint-based alignment like SIFT suffers from the lack of valid keypoint for matching. While our alignment method outperforms all other methods and achieves the best results, this is due to the progressive cascade alignment strategy, which breaks the hard alignment process into simpler tasks, facilitating the bridging between images taken under different focal lengths.

14. Additional results

We provide detailed quantitative results of UZ-ST in Tab. 11. Additional qualitative results are provided in Fig. 7 and Fig. 8. Fig. 9 also shows the ability of our method in handling other languages like Japanese and Korean.

We also conduct additional ablation experiment on how the randomness of OCR output affect our method. As shown in Tab. 10, the text accuracy gains as the randomness of the OCR output declines. However, even under complete random OCR prediction, our method still manage to outperform TADiSR thanks to the LR condition still guiding the stage 1 to generate text structures. This proves that the reliance of our method on the OCR model is limited.

15. Efficiency Analysis

As shown in Tab. 9, stage 1 is a standard diffusion process that takes multiple steps in inference, the efficiency of our model may be suboptimal compared to one-step methods. However, our method achieves state-of-the-art performance in OCR-A, which cannot be easily obtained by extending inference time.

Table 9. Efficiency analysis.

Methods	Flops (GFLOPs)	Speed (ms)	OCR-A
HAT	6670.32	1086.68	37.9%
DiffTSR	58502.14	8610.59	31.7%
DiT4SR w/o llava	160787.14	17385.14	23.7%
DreamClear w/o llava	412843.67	83193.81	21.4%
TADiSR	4497.96	342.32	36.6%
TiGeSR (ours) Stage 1	6215.85	663.5	43.0%
TiGeSR (ours) Stage 2	3734.01	360.06	

Table 10. Ablation on the effect of stochastic OCR outputs on the performance of our method. The performance gains as the randomness of the OCR output drops. Even under 100% random OCR output, our method still outperforms TADiSR, proving low reliance on OCR.

Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	OCR-A↑
TADiSR	24.68	0.795	0.362	0.227	52.13	35.5%
TiGeSR (Ours) performance under different randomness of OCR output.						
100% Random Text	25.60	0.833	0.188	**0.150**	28.92	40.3%
50% Random Text	25.60	0.833	0.189	0.151	28.92	40.6%
20% Random Text	**25.63**	**0.834**	**0.185**	**0.150**	29.3	43.7%
0% Random Text (Ours)	**25.63**	**0.834**	**0.185**	**0.150**	**28.82**	**44.6%**

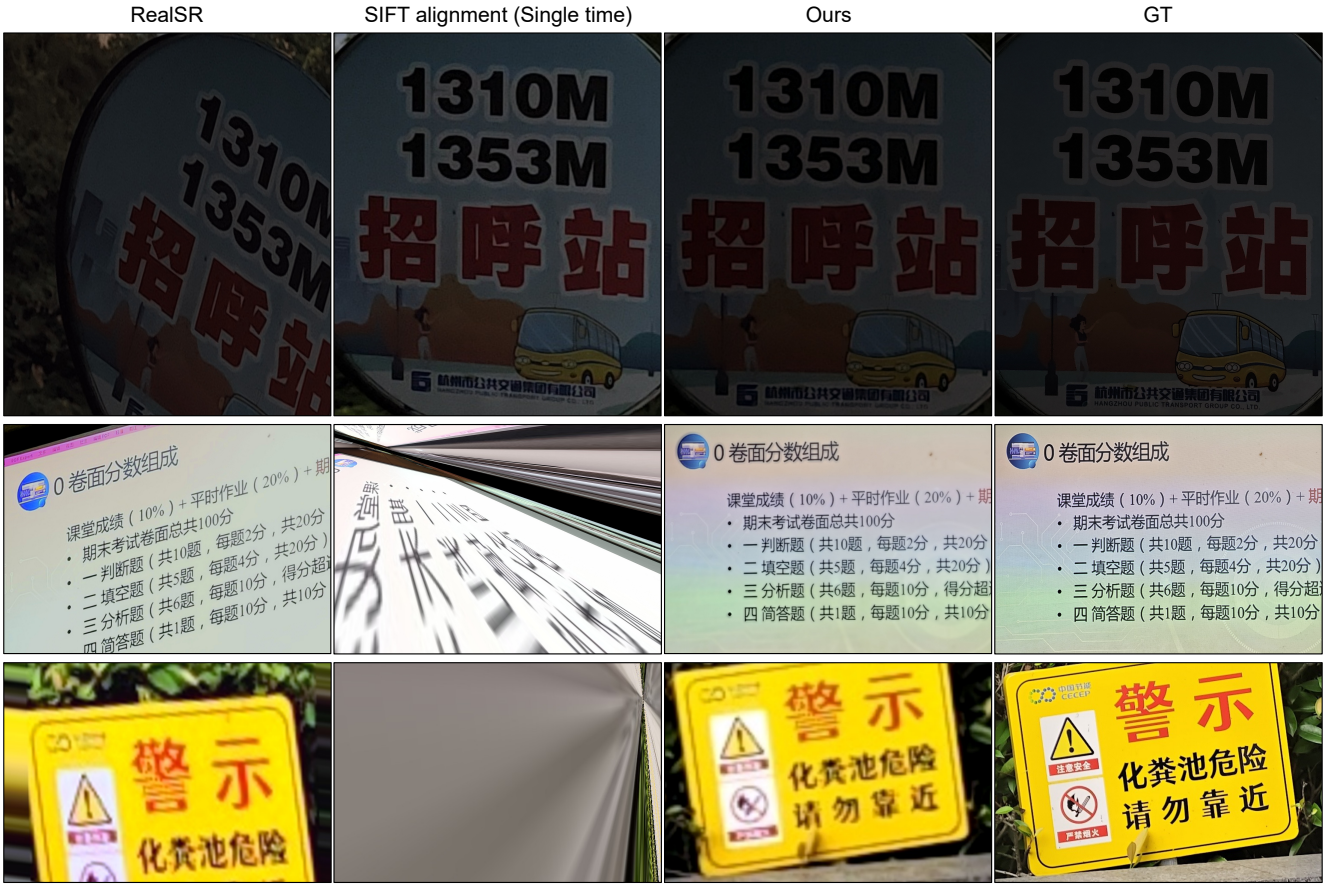

Figure 6. Alignment method comparison.

Table 11. Evaluation Results on Both Image Quality and Text Accuracy on UltraZoom-ST.

	Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	$PSNR_{cr}$↑	$SSIM_{cr}$↑	$LPIPS_{cr}$↓	$DISTS_{cr}$↓	OCR-A↑
85mm (×2.35)	Real-ESRGAN	25.84	0.854	0.139	0.130	28.25	23.59	0.843	0.170	0.168	60.8%
	HAT	**27.33**	**0.873**	0.128	0.133	27.21	24.76	**0.870**	0.139	0.146	62.0%
	MACRONet	22.90	0.804	0.234	0.173	40.19	17.92	0.693	0.515	0.360	54.2%
	SeeSR	25.02	0.826	0.167	0.159	32.31	21.62	0.796	0.245	0.207	28.2%
	SupIR	25.58	0.823	0.202	0.160	29.33	21.91	0.808	0.239	0.214	33.3%
	DiffTSR	23.48	0.805	0.230	0.157	35.77	17.92	0.688	0.392	0.285	48.2%
	DiffBIR	25.25	0.793	0.175	0.152	24.57	22.26	0.803	0.218	0.190	36.8%
	OSEDiff	26.38	0.852	0.153	0.137	23.92	22.84	0.833	0.200	0.204	41.0%
	DreamClear	25.31	0.811	0.169	0.151	24.19	21.44	0.790	0.253	0.221	26.8%
	TSD-SR	23.99	0.800	0.161	0.165	30.40	20.62	0.786	0.232	0.218	37.8%
	DiT4SR	24.71	0.810	0.173	0.143	29.15	21.25	0.800	0.221	0.193	48.2%
	TADiSR	26.31	0.858	0.177	0.149	30.47	23.55	0.856	0.175	0.178	55.7%
	Ours	27.00	0.871	**0.123**	**0.120**	**22.34**	**24.31**	0.860	**0.135**	**0.142**	**63.2%**
35mm (×5.71)	Real-ESRGAN	23.70	0.785	0.218	0.184	42.62	20.86	0.782	0.272	0.235	34.1%
	HAT	25.05	0.819	0.207	0.182	41.32	21.98	0.816	0.250	0.241	34.6%
	MACRONet	22.28	0.774	0.284	0.199	49.14	17.43	0.694	0.523	0.361	31.5%
	SeeSR	23.76	0.795	0.198	0.175	37.92	20.55	0.778	0.271	0.222	25.2%
	SupIR	23.49	0.746	0.302	0.204	41.80	20.00	0.757	0.335	0.282	24.0%
	DiffTSR	22.57	0.773	0.281	0.184	45.78	16.99	0.671	0.443	0.312	31.2%
	DiffBIR	23.76	0.731	0.245	0.196	35.99	20.82	0.770	0.275	0.234	28.6%
	OSEDiff	25.35	0.827	0.192	0.166	29.06	21.62	0.809	0.263	0.263	29.9%
	DreamClear	24.20	0.778	0.209	0.177	32.08	20.66	0.779	0.283	0.251	23.9%
	TSD-SR	22.93	0.757	0.200	0.199	37.61	19.64	0.760	0.282	0.246	27.3%
	DiT4SR	23.45	0.774	0.204	0.158	31.73	20.12	0.779	0.250	0.208	26.5%
	TADiSR	24.68	0.795	0.362	0.227	52.13	21.57	0.798	0.366	0.347	35.5%
	Ours	**25.63**	**0.834**	**0.185**	**0.150**	**28.82**	**22.14**	**0.823**	**0.204**	**0.194**	**44.6%**
14mm (×14.29)	Real-ESRGAN	22.09	0.718	0.423	0.284	89.22	18.87	0.721	0.503	0.427	11.5%
	HAT	22.66	0.738	0.451	0.299	89.58	19.27	0.738	0.533	0.484	12.2%
	MACRONet	21.00	0.716	0.422	0.254	79.67	17.01	0.680	0.553	0.382	10.1%
	SeeSR	21.79	0.731	0.310	0.228	64.31	18.87	0.717	0.396	0.308	13.2%
	SupIR	21.38	0.681	0.447	0.268	70.09	17.99	0.690	0.487	0.383	11.0%
	DiffTSR	20.88	0.713	0.412	0.233	74.07	15.87	0.630	0.538	0.348	12.0%
	DiffBIR	21.61	0.630	0.392	0.255	69.64	18.61	0.685	0.438	0.328	11.3%
	OSEDiff	23.11	0.768	0.274	0.214	49.76	19.44	0.740	0.385	0.334	12.7%
	DreamClear	22.45	0.719	0.364	0.260	62.95	19.14	0.725	0.440	0.377	11.7%
	TSD-SR	21.14	0.705	**0.273**	0.225	57.07	17.67	0.686	0.393	0.293	11.9%
	DiT4SR	20.88	0.707	0.282	**0.181**	**47.45**	17.36	0.687	**0.367**	**0.253**	11.7%
	TADiSR	22.43	0.721	0.555	0.320	98.44	19.20	0.730	0.585	0.518	14.4%
	Ours	**23.41**	**0.774**	0.301	0.209	54.74	**19.73**	**0.748**	0.374	0.322	**16.0%**
Total	RealEsrGAN	23.99	0.790	0.248	0.194	30.60	21.25	0.786	0.302	0.266	37.1%
	HAT	25.17	0.815	0.249	0.198	30.12	22.18	0.813	0.291	0.276	37.9%
	MARCONet	22.13	0.768	0.306	0.205	34.28	17.48	0.690	0.529	0.366	33.4%
	SeeSR	23.64	0.788	0.219	0.184	26.73	20.45	0.767	0.297	0.241	22.8%
	SupIR	23.62	0.754	0.308	0.207	27.91	20.10	0.756	0.344	0.287	23.6%
	DiffTSR	22.41	0.767	0.300	0.189	31.25	17.00	0.665	0.452	0.313	31.7%
	DiffBIR	23.67	0.724	0.262	0.197	23.10	20.70	0.757	0.302	0.245	26.6%
	OSEDiff	25.07	0.819	0.201	0.169	20.53	21.43	0.798	0.276	0.263	28.9%
	DreamClear	24.10	0.773	0.238	0.191	21.75	20.50	0.768	0.317	0.276	21.4%
	TSD-SR	22.79	0.757	0.207	0.194	24.08	19.42	0.748	0.296	0.249	26.6%
	DiT4SR	23.16	0.767	0.215	0.159	20.58	19.73	0.760	0.273	0.216	23.7%
	TADiSR	24.61	0.796	0.203	0.160	36.61	21.59	0.799	0.360	0.336	36.6%
	Ours	**25.48**	**0.830**	**0.196**	**0.156**	**20.01**	**22.22**	**0.814**	**0.228**	**0.212**	**43.0%**

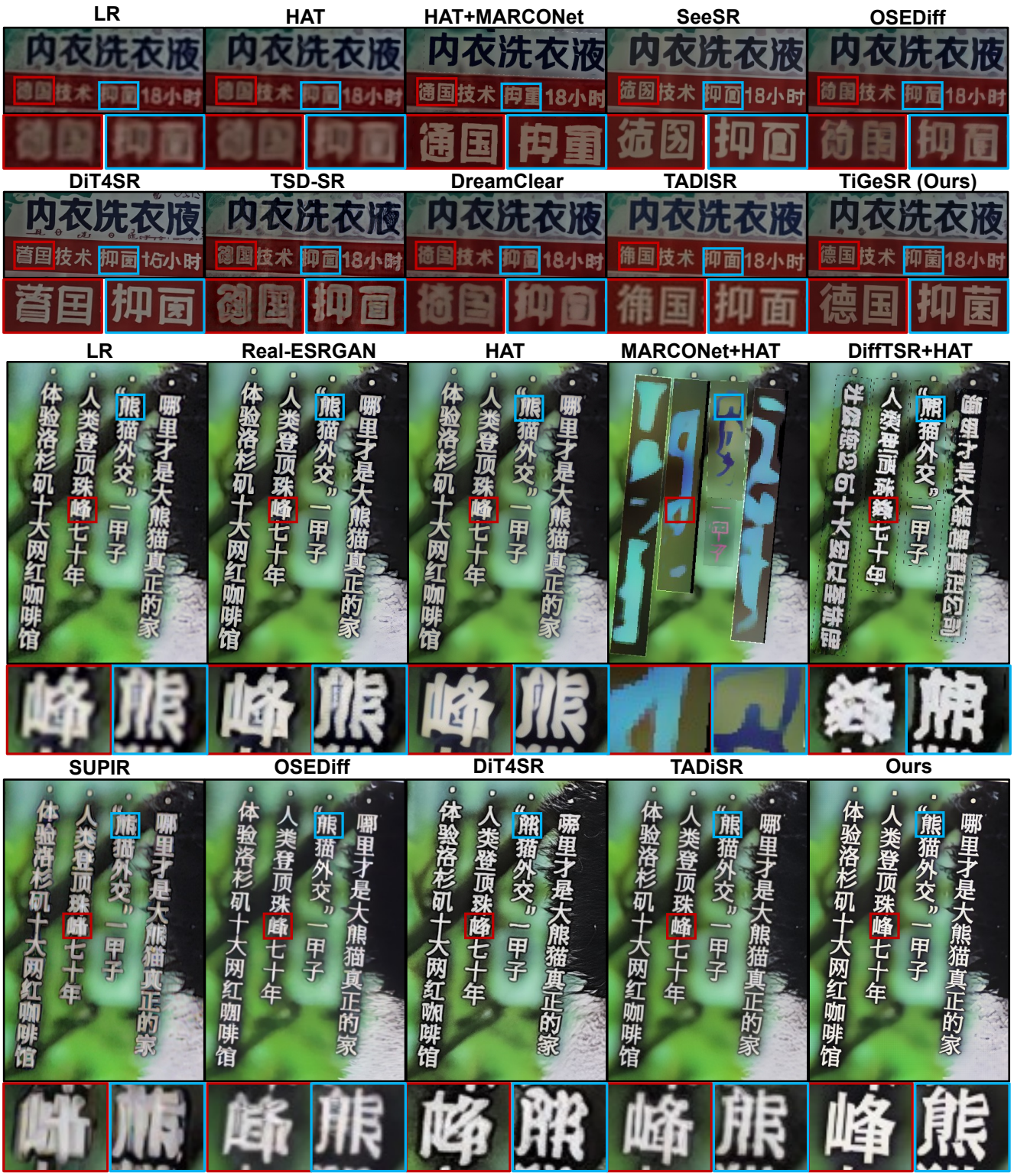

Figure 7. Additional qualitative results of TiGeSR.

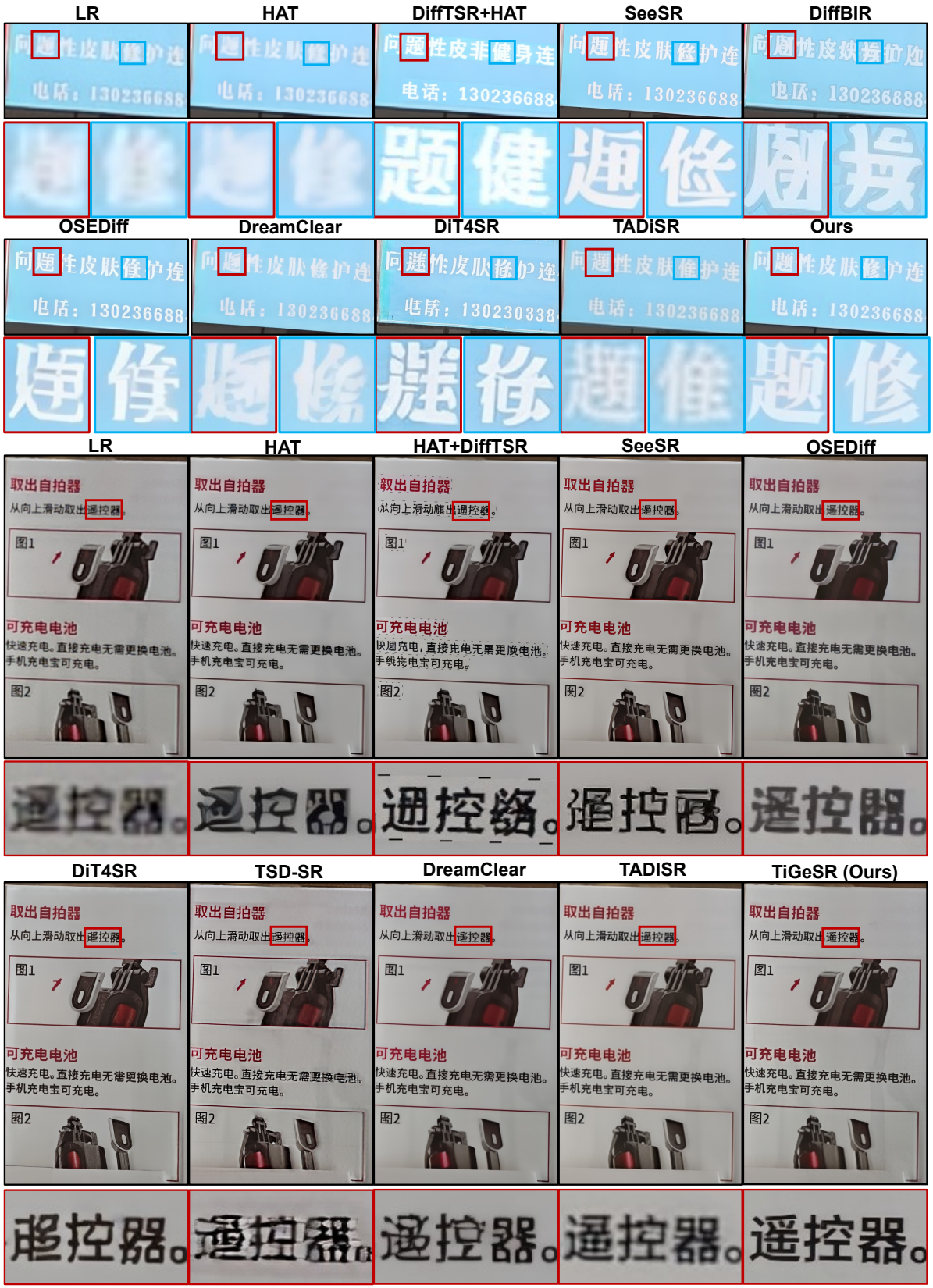

Figure 8. Additional qualitative results of TiGeSR.

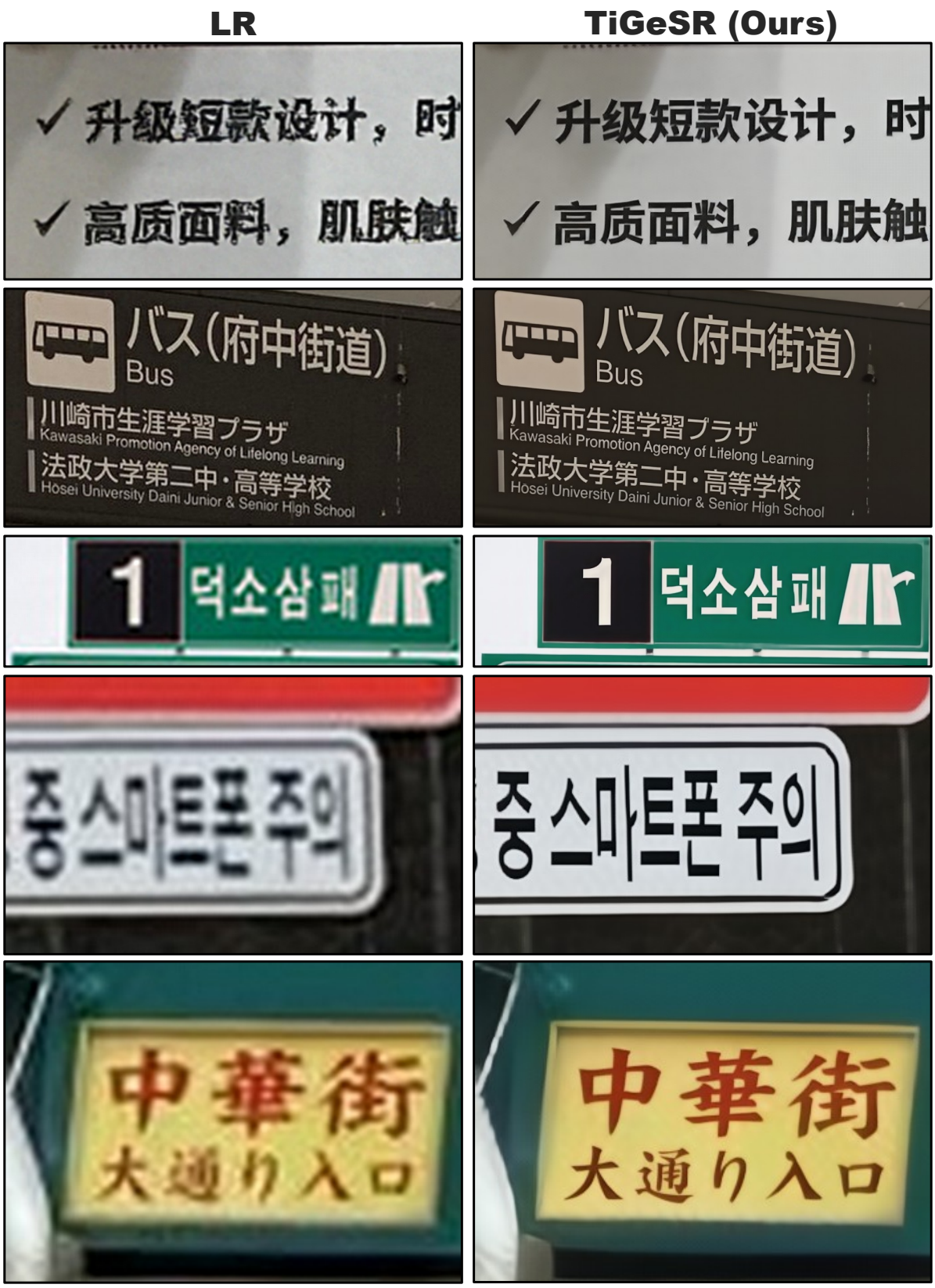

Figure 9. Demonstration of our method under different languages. (Zoom in for more details.)

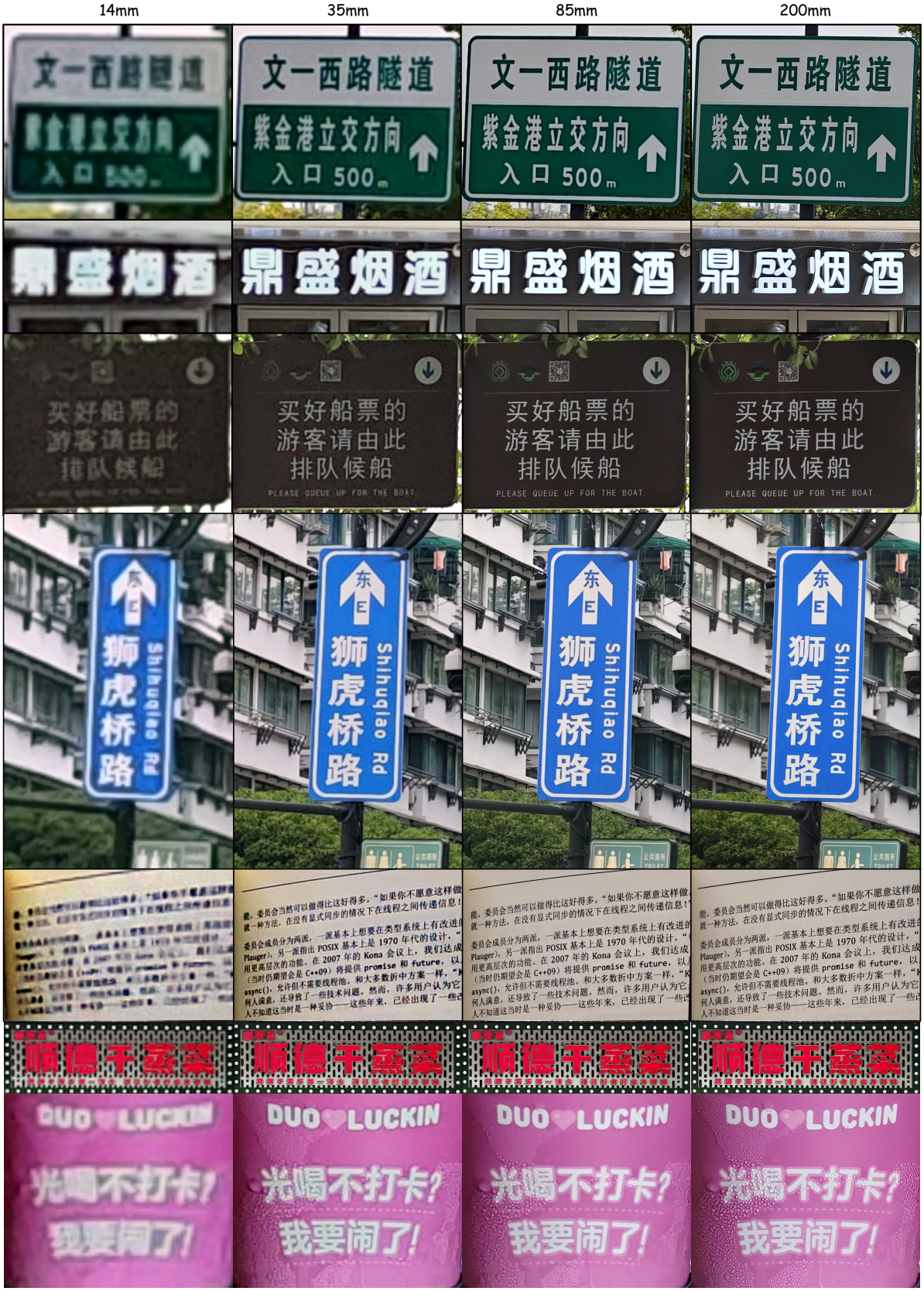

Figure 10. Detailed examples of UZ-ST.