Supplementary Materials to “A Benchmark for Chinese-English Scene Text Image Super-resolution”

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

In this supplementary file, we provide the following materials:

- The role of edge features $\mathcal{F}(\mathcal{C}_H)$ in $\mathcal{L}_{EA}^F$ (please refer to Section 4.2 in the manuscript).
- Ablation study on hyperparameter selection (please refer to Section 5 in the manuscript).
- Comparison on STISR models trained on synthetic LR-HR pairs and our Real-CE data (please refer to Section 5.1 in the main paper).
- More visual comparisons among the STISR models trained on different text image datasets (please refer to Section 5.1 in the main paper).
- More visual comparisons among the STISR models trained with different loss combinations (please refer to Section 5.2 in the main paper).
- Experiments on images out of the Real-CE dataset.

A. The Role of Edge Features $\mathcal{F}(\mathcal{C}_H)$ in $\mathcal{L}_{EA}^F$

The feature level edge-aware loss $\mathcal{L}_{EA}^F$ provides high-level supervision on the learning process so that the estimated HR images can be perceptually closer to the ground truths in structural areas. Let $\mathcal{V} = \mathcal{F}(\hat{\mathcal{I}}_H) \cdot \mathcal{F}(\hat{\mathcal{C}}_H) - \mathcal{F}(\mathcal{I}_H) \cdot \mathcal{F}(\mathcal{C}_H)$. The gradient back-propagated from $\mathcal{L}_{EA}^F$ to the estimated HR text image $\hat{\mathcal{I}}_H$ can be calculated as:

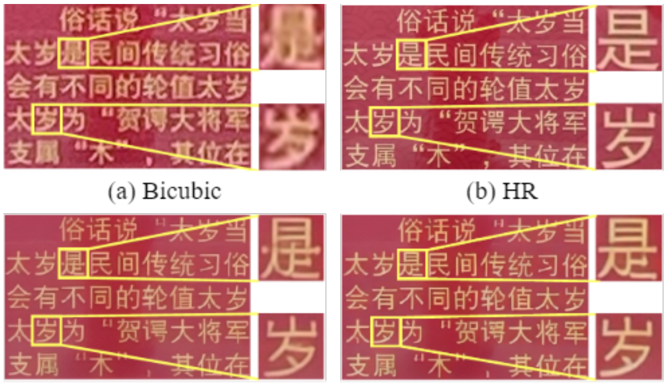
$$\frac{\partial \mathcal{L}_{EA}^F}{\partial \hat{\mathcal{I}}_H} = \left\{ \begin{array}{ll} \mathcal{F}(\hat{\mathcal{C}}_H) \cdot \frac{\partial \mathcal{F}(\hat{\mathcal{I}}_H)}{\partial \hat{\mathcal{I}}_H}, & \text{if } \mathcal{V} > 0 \\ -\mathcal{F}(\hat{\mathcal{C}}_H) \cdot \frac{\partial \mathcal{F}(\hat{\mathcal{I}}_H)}{\partial \hat{\mathcal{I}}_H}, & \text{else if } \mathcal{V} < 0 \\ 0, & \text{otherwise} \end{array} \right\} \tag{1}$$

From Eq. (1), one can see that the edge information $\mathcal{F}(\hat{\mathcal{C}}_H)$ is incorporated into the gradient to reconstruct the HR image. In this way, we can enforce the model to be aware of the edge information in the text areas. Figure 1 shows an example. Though supervision from the image feature $\mathcal{F}(\hat{\mathcal{I}}_H)$ can improve much the text recovery over bicubic interpolation, as visualized in Figure 1 (c), the text edges and strokes remain unclear, and some local structures of the characters are incorrectly recovered. By incorporating the edge feature $\mathcal{F}(\hat{\mathcal{C}}_H)$ in training, the STISR model can recover text characters with significantly better edge boundaries and structures, as can be seen in Figure 1 (d).

Quantitative results can also be viewed in Table 1. One can see that, using image feature only in supervision leads to limited enhancement on text recognition. Together with edge feature supervision, the SR recovery performance is largely enhanced from 0.6492 to 0.6622 in NED, demonstrating the effectiveness of edge feature information in SR text recovery.

B. Ablation Study on Hyperparameter Selection

We perform ablation experiments to determine the balancing parameters α and β in our loss function (see Eq. (4) in the main paper). We employ the RRDB [2] network as STISR model backbone. In the experiments, we first fix β to 0, and search for the best α. By setting α to 0, 0.01, 0.1, 1 and 10, we obtain the results as in the upper panel of Table 2. One can see that all the evaluation metrics increase when α increases from 0 to 1. However, when α is larger than 1, the LPIPS and recognition performance decrease. To keep the best recognition performance, we set $\alpha = 1$ in the experiments.



(a) Bicubic (b) HR

(c) w/o edge feature (d) w edge feature

Figure 1. (a) Bicubic LR image; (b) ground-truth HR; (c) and (d) show the STISR results of RRDB models trained on Real-CE without and with edge features $\mathcal{F}(\mathcal{C}_H)$ in edge-aware loss $\mathcal{L}^F_{EA}$, respectively.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	ACC ↑	NED ↑
LR	19.65	0.6684	0.3987	0.2759	0.6173
Baseline	**20.42**	**0.7303**	0.2630	0.2914	0.6399
w/o $\mathcal{F}(\mathcal{C}_H)$	20.09	0.7203	0.2108	0.3011	0.6492
w/ $\mathcal{F}(\mathcal{C}_H)$	20.14	0.7210	**0.2031**	**0.3093**	**0.6622**
HR	-	-	-	0.4807	0.8342

Table 1. STISR results of RRDB models trained on Real-CE with and without edge features $\mathcal{F}(\mathcal{C}_H)$ in edge-aware (EA) loss $\mathcal{L}^F_{EA}$. The baseline model is trained with only $\mathcal{L}_1$ and pixel-level EA loss.

	$\beta = 0$				
α	PSNR ↑	SSIM ↑	LPIPS ↓	ACC ↑	NED ↑
0	20.23	0.7231	0.2710	0.2865	0.6320
0.01	20.28	0.7247	0.2706	0.2908	0.6349
0.1	20.30	0.7278	0.2699	0.2887	0.6372
1	20.36	0.7299	**0.2630**	**0.2914**	**0.6399**
10	**20.39**	**0.7310**	0.2721	0.2902	0.6347
	$\alpha = 1$				
β	PSNR ↑	SSIM ↑	LPIPS ↓	ACC ↑	NED ↑
0.01	19.67	0.7156	**0.1984**	**0.3152**	**0.6707**
0.001	19.93	0.7199	0.2013	0.3127	0.6685
0.0005	20.14	0.7210	0.2031	0.3093	0.6622
0.0001	20.21	0.7227	0.2354	0.2982	0.6526
0	**20.36**	**0.7299**	0.2630	0.2914	0.6399

Table 2. Experimental results with different balancing parameters.

To determine the value of β, we first fix α to 1, and then conduct experiments on different values of β. In the lower panel of Table 2, we show the experimental results by setting β to 0.01, 0.001, 0.0005, 0.0001 and 0, respectively. One can see that the LPIPS and recognition performance of the STISR model show significant improvements as β increases; however, the PSNR/SSIM indices drop abruptly when β is larger than 0.0005. To balance all indices, we choose β to 0.0005 in the experiments.

C. STISR Models Trained on Synthetic Data *vs*. Real-CE Data

The image degradation modeling methods proposed in BSRGAN [3] and Real-ESRGAN [1] are widely used to synthesize LR-HR image pairs, yet they are still limited in approximating the real-world degradations. In Table 3, we show the $4\times$ STISR results of the RRDB models trained on directly down-sampled LR-HR pairs, Real-ESRGAN degraded LR-HR pairs and our Real-CE data. One can see that direct down-sampling and Real-ESRGAN degradation show marginal improvement over the

Train input	PSNR ↑	SSIM ↑	LPIPS ↓	ACC ↑	NED ↑
Bicubic	19.65	0.6684	0.3987	0.2759	0.6173
LR w/ down-sampling	18.98	0.6711	0.2346	0.2756	0.6285
LR w/ Real-ESRGAN	18.17	0.6919	0.2285	0.2864	0.6399
Real-CE	**20.14**	**0.7210**	**0.2031**	**0.3093**	**0.6622**

Table 3. Experimental results of RRDB trained on synthetic data generated by Real-ESRGAN [1] and Real-CE real data.

Bicubic interpolation. While using Real-CE, the trained model shows significant improvement over Bicubic, verifying the necessity and effectiveness of our Real-CE benchmark.

D. More Visual Comparisons among STISR Models Trained on Different Datasets

To better examine the effectiveness of our Real-CE dataset on recovering Chinese text, we present more visual comparisons and text recognition results in Figure 2. One can see that the estimated HR text images by models trained on Real-CE exhibit much better image quality and recognition accuracy than models trained on TextZoom and RealSR.

E. More Visual Comparisons among STISR Models Trained with Different Loss Combinations

To better illustrate the effectiveness of our proposed edge-aware losses $\mathcal{L}_{EA}^{P}$ and $\mathcal{L}_{EA}^{F}$ (refer to Section 4.2 in the main paper), we present more visualization examples with more STISR models, including RRDB [2], RCAN [5] and ELAN [4]. The results are shown in Figures 3, 4 and 5, correspondingly. In each figure, different rows show the results generated in $4\times$ (13mm to 52mm) and $2\times$ (26mm to 52mm and 13mm to 26mm) zooming modes, and different columns show the results generated by different loss combinations ($\mathcal{L}_1$, $\mathcal{L}_1 + \mathcal{L}_{EA}^{P}$ and $\mathcal{L}_1 + \mathcal{L}_{EA}^{P} + \mathcal{L}_{EA}^{F}$).

From these figures, similar conclusions to Figure 8 in the main paper can be drawn, *i.e.*, employing the EA losses into model training can significantly enhance the text structure recovery. In addition, our loss works well with different backbone networks and for different zooming modes.

F. Experiments on Images out of the Real-CE Dataset

We collect some test images using Redmi, OPPO, VIVO and Samsung smartphone sensors, which are out of the Real-CE dataset, and perform STISR on them using RRDB [2] models trained on TextZoom, RealSR and our Real-CE. Visual results can be found in Figure 6. One can see that the STISR model trained on our Real-CE dataset, which is built by iPhone device, performs well, showing superior generalization capability to other sensors.

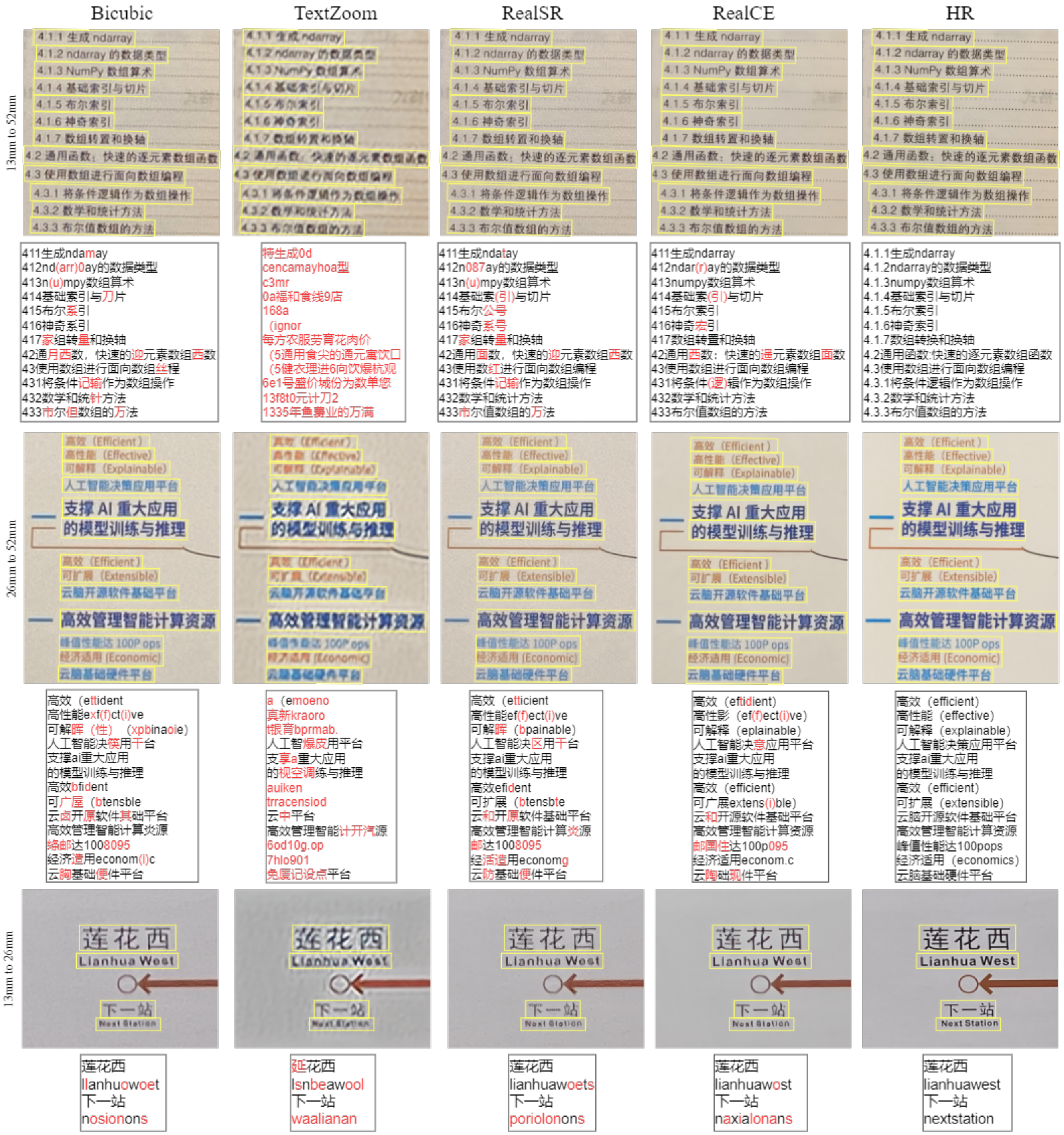


Figure 2. More visualizations and recognition results of RRDB models trained on different training datasets. The test images are from the Real-CE test set. From top row to bottom row are visual and recognition results with zooming factors 4× (13mm to 52mm) and 2× (26mm to 52mm and 13mm to 26mm). Wrong results are in red. Please zoom in for more details.

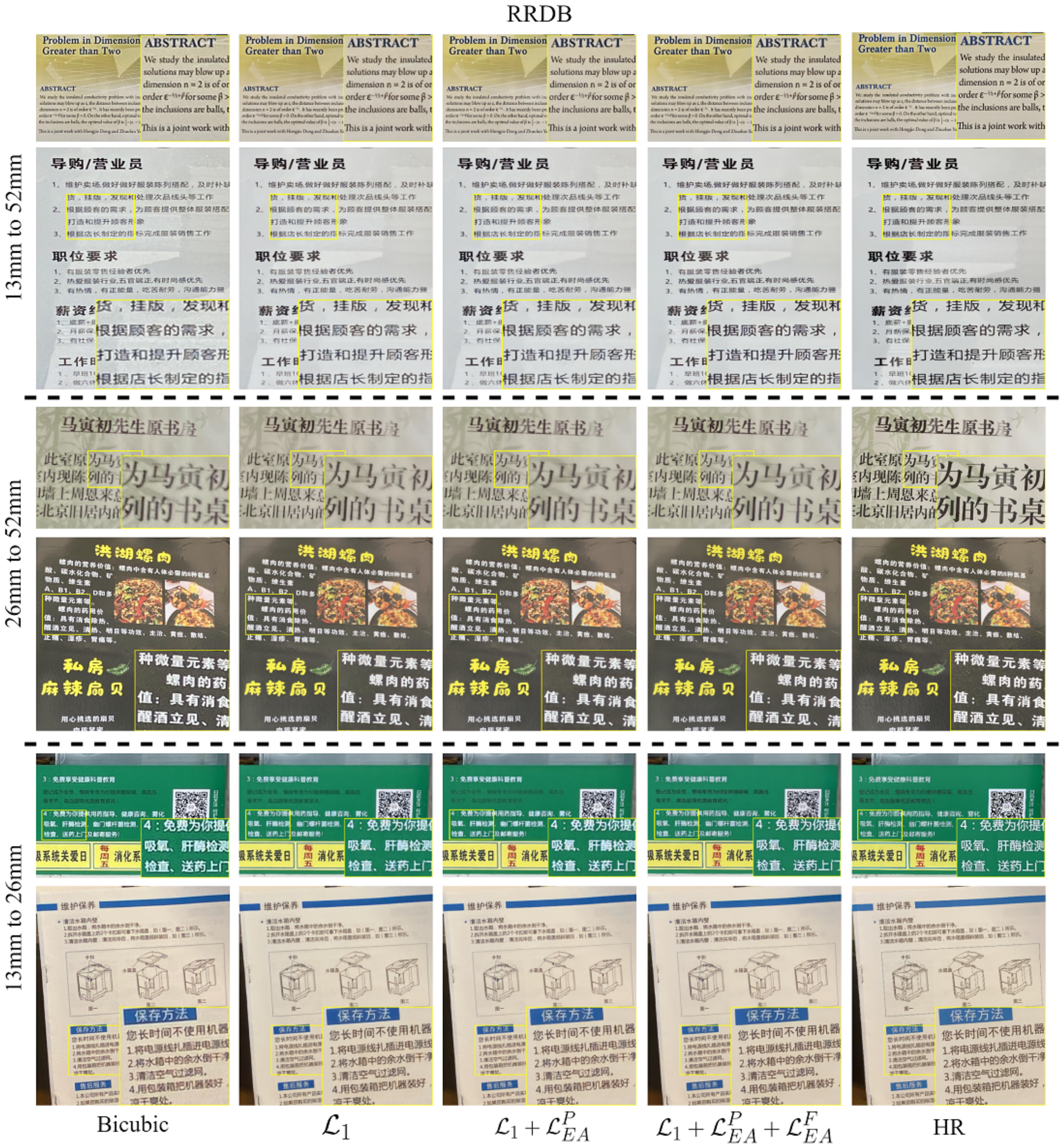

RRDB
13mm to 52mm
26mm to 52mm
13mm to 26mm
Bicubic
$\mathcal{L}_1$
$\mathcal{L}_1 + \mathcal{L}_{EA}^P$
$\mathcal{L}_1 + \mathcal{L}_{EA}^P + \mathcal{L}_{EA}^F$
HR

Figure 3. STISR results of RRDB models trained on Real-CE with different combination of losses. From top row to bottom row are results with zooming factors 4× (13mm to 52mm) and 2× (26mm to 52mm and 13mm to 26mm). From left column to right column are the results of bicubic interpolation, models trained with $\mathcal{L}_1$, $\mathcal{L}_1 + \mathcal{L}_{EA}^P$ and $\mathcal{L}_1 + \mathcal{L}_{EA}^P + \mathcal{L}_{EA}^F$ and HR. Please zoom in for more details.

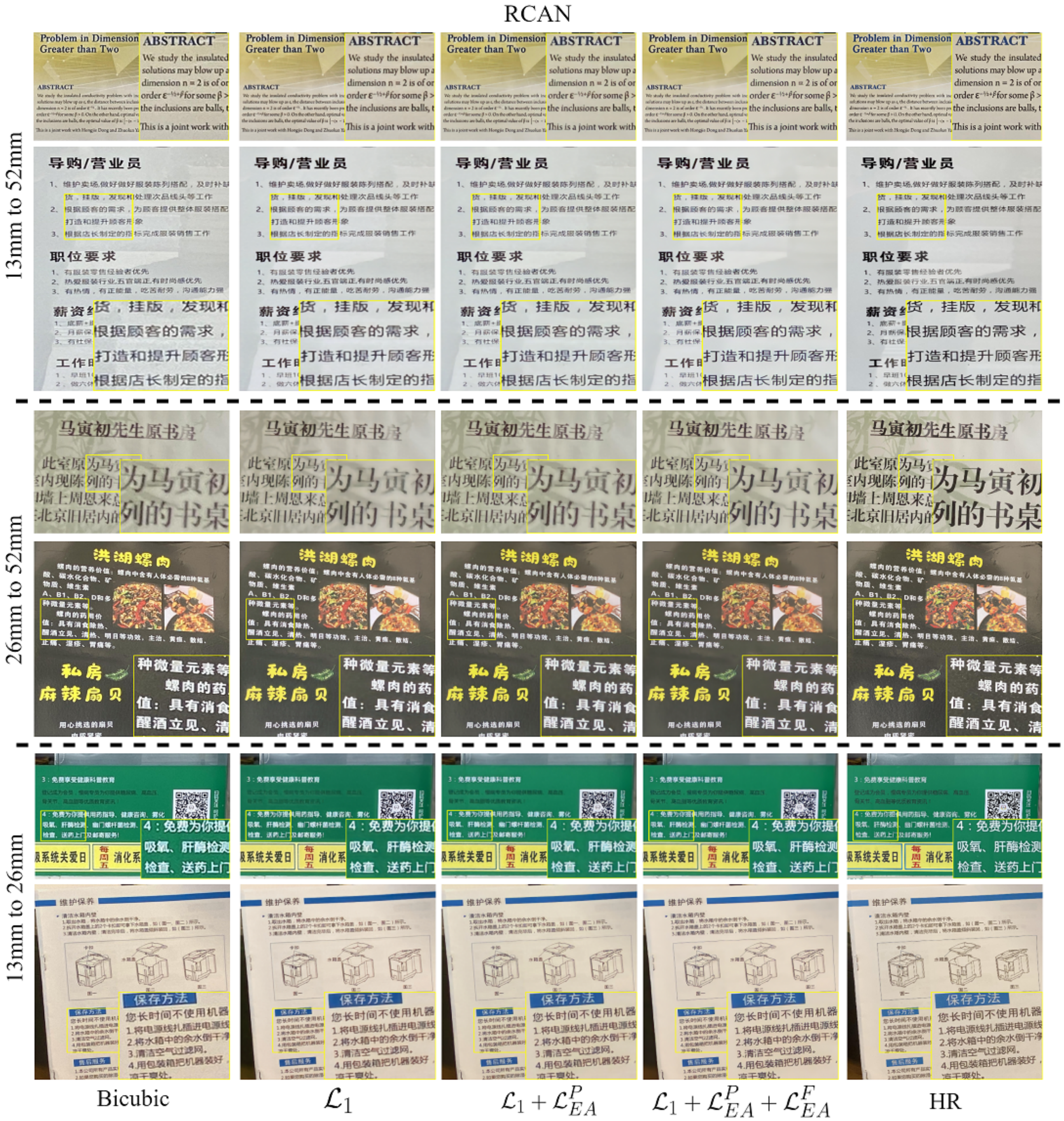

RCAN
13mm to 52mm
26mm to 52mm
13mm to 26mm
Bicubic
$\mathcal{L}_1$
$\mathcal{L}_1 + \mathcal{L}_{EA}^P$
$\mathcal{L}_1 + \mathcal{L}_{EA}^P + \mathcal{L}_{EA}^F$
HR

Figure 4. STISR results of RCAN models trained on Real-CE with different combination of losses. From top row to bottom row are results with zooming factors 4× (13mm to 52mm) and 2× (26mm to 52mm and 13mm to 26mm). From left column to right column are the results of bicubic interpolation, models trained with $\mathcal{L}_1$, $\mathcal{L}_1 + \mathcal{L}_{EA}^P$ and $\mathcal{L}_1 + \mathcal{L}_{EA}^P + \mathcal{L}_{EA}^F$ and HR. Please zoom in for more details.

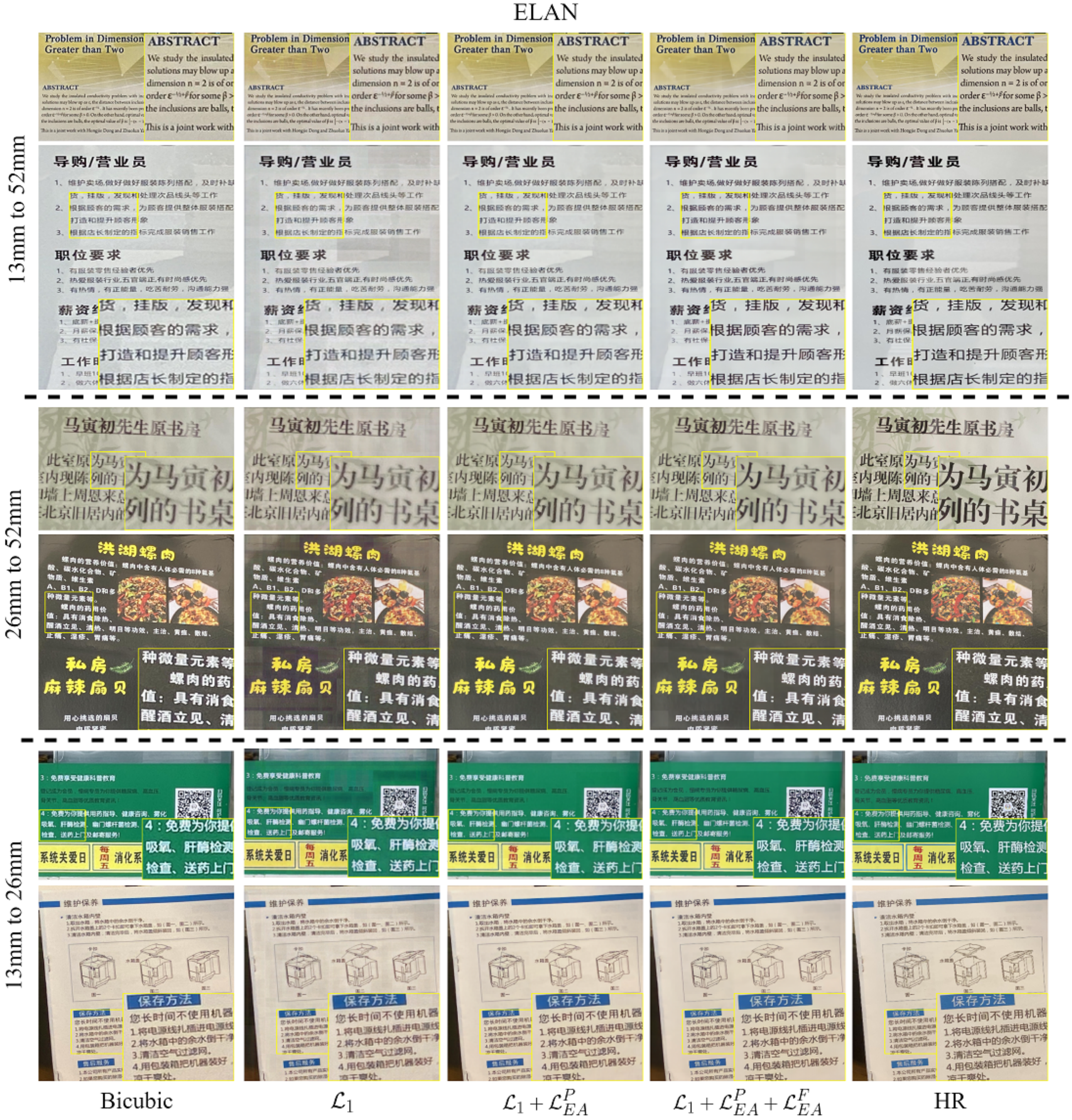

ELAN
13mm to 52mm
26mm to 52mm
13mm to 26mm
Bicubic
$\mathcal{L}_1$
$\mathcal{L}_1 + \mathcal{L}_{EA}^P$
$\mathcal{L}_1 + \mathcal{L}_{EA}^P + \mathcal{L}_{EA}^F$
HR

Figure 5. STISR results of ELAN models trained on Real-CE with different combination of losses. From top row to bottom row are results with zooming factors 4× (13mm to 52mm) and 2× (26mm to 52mm and 13mm to 26mm). From left column to right column are the results of bicubic interpolation, models trained with $\mathcal{L}_1$, $\mathcal{L}_1 + \mathcal{L}_{EA}^P$ and $\mathcal{L}_1 + \mathcal{L}_{EA}^P + \mathcal{L}_{EA}^F$ and HR. Please zoom in for more details.

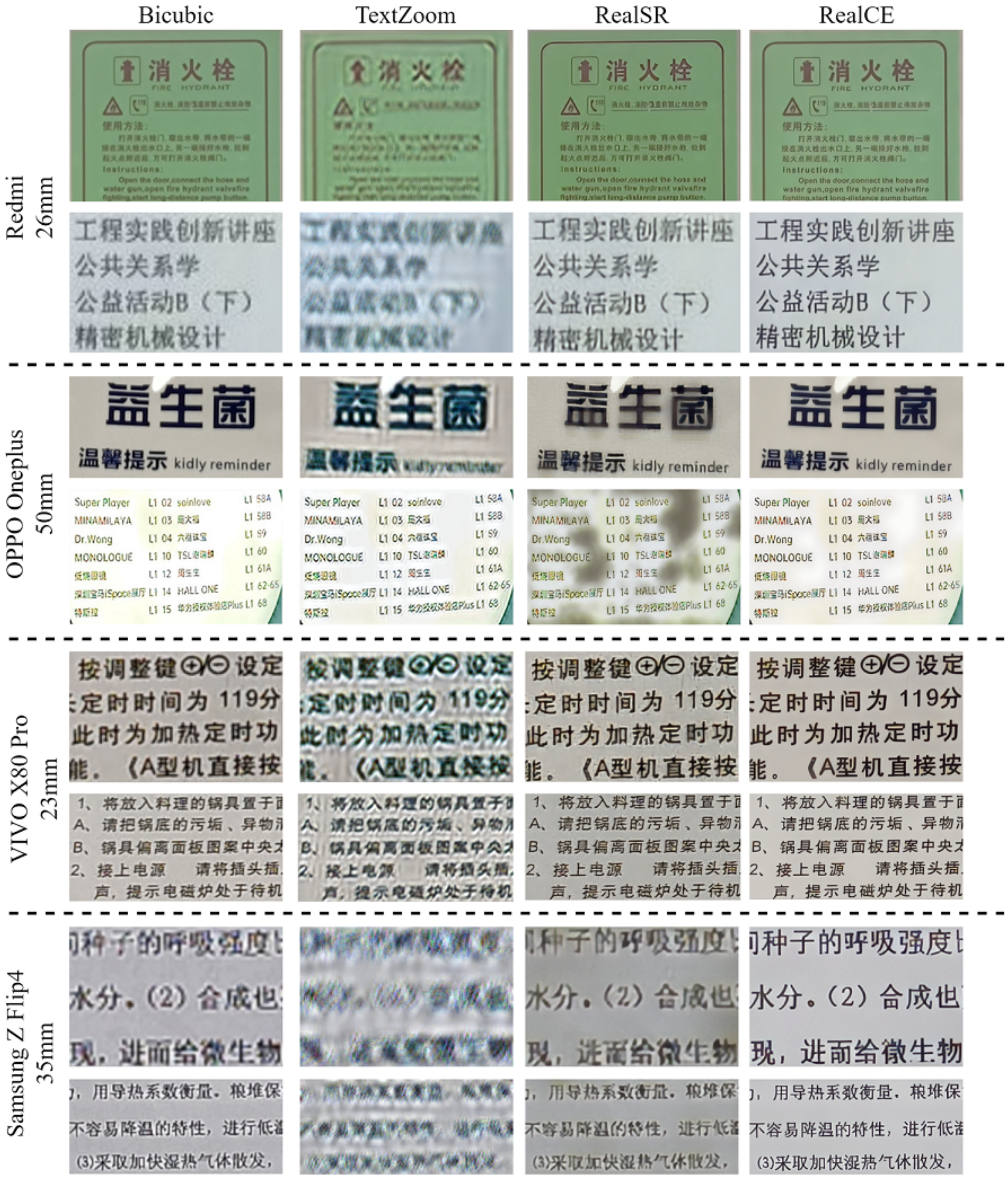


Figure 6. Comparison of STISR models trained on different datasets on samples captured with different smartphones, which are out of our Real-CE dataset. Please zoom in for more details.

References

[1] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Int. Conf. Comput. Vis.*, pages 1905–1914, 2021. 2, 3

[2] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Eur. Conf. Comput. Vis. Worksh.*, pages 0–0, 2018. 1, 3

[3] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Int. Conf. Comput. Vis.*, pages 4791–4800, 2021. 2

[4] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. *Eur. Conf. Comput. Vis.*, 2022. 3

[5] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Eur. Conf. Comput. Vis.*, pages 286–301, 2018. 3