Supplementary Material *for* ICF-SRSR: Invertible scale-Conditional Function *for* Self-Supervised Real-world Single Image Super-Resolution

Reyhaneh Neshatavar[1*] Mohsen Yavartanoo[1*] Sanghyun Son[1] Kyoung Mu Lee[1,2]
[1]Dept. of ECE & ASRI, [2]IPAI, Seoul National University, Seoul, Korea
{reyhanehneshat,myavartanoo,thstkdgus35,kyoungmu}@snu.ac.kr

S1. Details of network architecture

As described in Sec. 3.4 of our main manuscript, our ICF-SRSR adopts EDSR [34] as a baseline. However, to handle both up-sampling and down-sampling operations with the same network, we slightly modify the tail part of the original EDSR architecture for each scaling factor, *e.g.*, ×2 and ×4, and their inverses. Fig. S1 shows the original EDSR (Fig. S1a) and our modified EDSR (Fig. S1b). We use the pixel-unshuffle operator to down-sample an input image and generate the corresponding LLR image. For more stable optimization, we use the detach operator of PyTorch before passing the first outputs to the network again.

S2. Details of multi-scale augmentation strategy

As we mention in Sec. 4.4 of our main manuscript, we can generate images with various scaling factors, *e.g.*, ×2, ×4, and ×8 and their corresponding inverses from a single LR input. Fig. S2a shows our multi-tail architecture, which introduces a tail for each of the scale conditions. Then, we pass the generated output images of different scales to the model f_θ with their inverse scaling factors. By doing so, we reconstruct the input LR image as shown in Fig. S2b. Accordingly, to train our model f_θ under such a configuration, we minimize the loss functions $\mathcal{L}^{\text{Cons}}$ and $\mathcal{L}^{\text{Color}}$ defined in Sec. 3.3 of our main manuscript between the generated images and the input LR image.

S3. Evaluation by SSIM

We quantitatively show the results of our ICF-SRSR and EDSR (LLR,LR) methods compared to other supervised and unsupervised methods trained on DIV2K [1] dataset and tested on the five standard benchmarks [4,64,38,24,39] by SSIM metric in Tab. S1. According to the results, our method outperforms unsupervised method [24] on both scaling factors ×2 and ×4 and supervised method [9] on scaling factor ×2 and is comparable with other methods.

S4. Ablation on baseline model

We employ different models LIIF [9], EDSR [34], RDN [71], and RCAN [70] as the baseline of our ICF-SRSR framework. In the case of EDSR, RDN, and RCAN, we develop the original network architecture to generate multi-scale images by applying a tail for each scaling factor s and its inverse $1/s$, individually. In the case of LIIF, we leverage its continuous attribute to generate any scale of images by sub-sampling from the reconstructed continuous image. Tab. S2 shows the results of our ICF-SRSR with different baselines. We illustrate that our method is model-agnostic and can leverage different state-of-the-art (SOTA) baseline models. We note that our method can achieve better performance using advanced baselines except LIIF, which is not trained with continuous scales due to the limitation of the color loss $\mathcal{L}^{\text{Color}}$. We select the model EDSR as our baseline due to its training time efficiency.

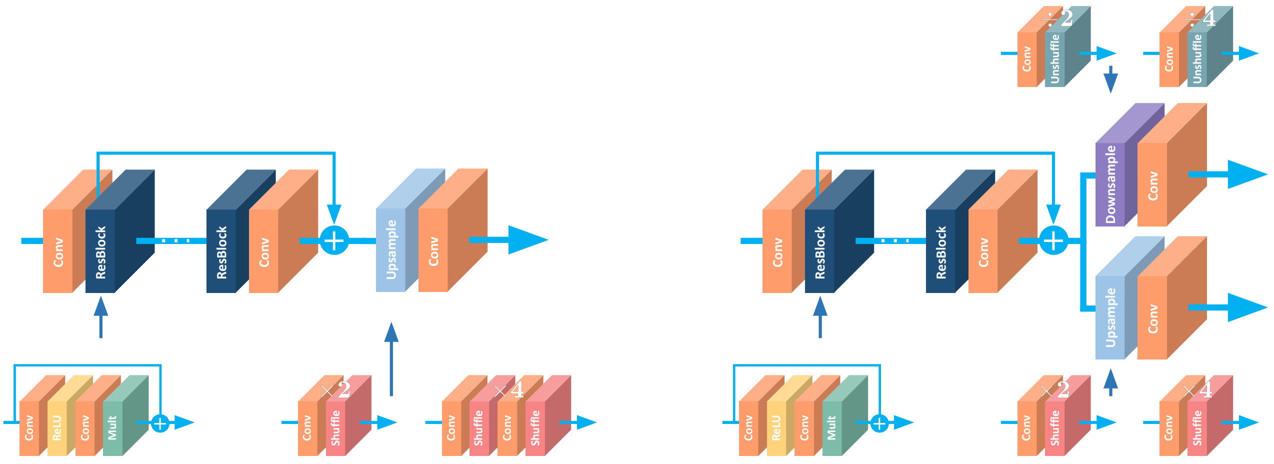
S5. Ablation on the hyperparameter λ_{Color}.

We conduct an ablation study to investigate the importance of our color loss $\mathcal{L}^{\text{Color}}$ defined in Sec. 3.3 by changing its weight λ_{Color}. Specifically, We increase the weight from 0.1 to 10 and report the performance of our ICF-SRSR trained on the scale ×2 of test sets of both real-world dataset RealSR [6] and synthetic datasets Set5 [4] and DIV2K [1] validation in Tab. S3. The results indicate that $\lambda_{\text{Color}} = 0.2$ achieves the best performance on different datasets.

S6. Comparison with DASR

We follow the official implementation of DASR [53] and train it using 1) HR images of DIV2K, 2) HR images of RealSR-V3, and 3) LR images of RealSR-V3 (self-supervised) and compare the results with our self-supervised method ICF-SRSR in Tab. S4. The results demonstrate the superiority of our method to effectively learn from LR images compared to the DASR method.

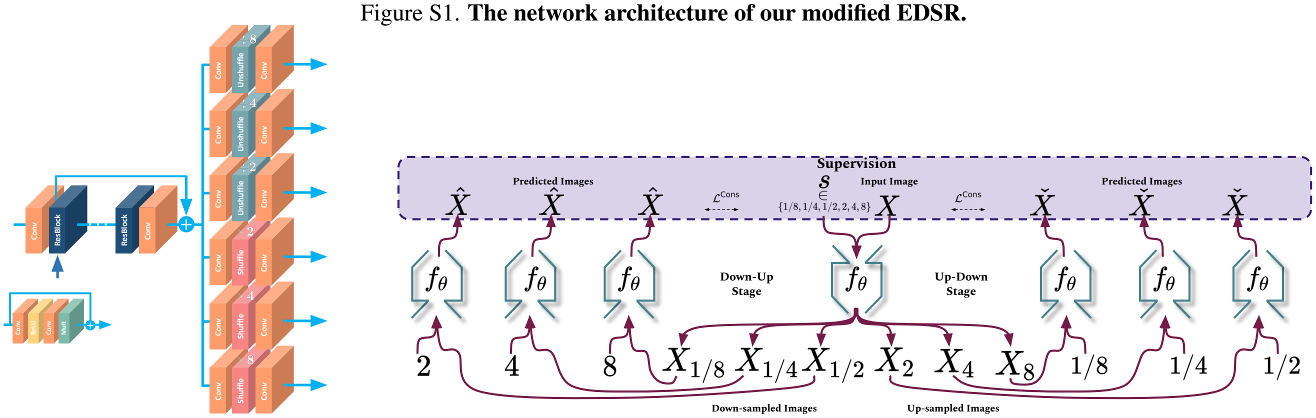
[*] equal contribution



(a) Original EDSR.

(b) Our modified EDSR for our ICF-SRSR.

Figure S1. **The network architecture of our modified EDSR.**



(a) Multi-tail modified EDSR.

(b) Multi-scale augmentation strategy.

Figure S2. **The overview of our multi-scale augmentation strategy.** (a) Our multi-tail EDSR for ×2, ×4, ×8 and their inverse scaling factors. (b) An overview of the proposed multi-scale augmentation.

S7. Noise-free results

In Sec. 4.2 of our main manuscript, we note that the ground-truth images of Set5 [4] and Set14 [64] datasets are noisy while our SR images are noise-free. We show the difference between our SR images and the noisy ground-truth images in Fig. S3. The results prove our claim and show that we can restore SR images without any noise.

S8. Complicated down-sampling degradations

As we show in Sec. 4.3 of our main manuscript, the proposed method can learn from real-world datasets with unknown degradations (real LR usually includes complicated degradations). For example, we can train our model f_θ on images from RealSR-V3 [6] and DRealSR [59] datasets directly and achieve promising results. Furthermore, we train and test our method ICF-SRSR on a dataset with more complicated degradations generated by the Real-ESRGAN [55] down-sampling strategy. We note that the generated LR images by the Real-ESRGAN [55] down-sampling model are synthesized by a sequence of classical degradations such as blur, resize, noise, JPEG compression, and artifacts to simulate more practical degradations. Fig. S4 demonstrates that our method ICF-SRSR can perform ×2 SR faithfully even on images with mild noise and artifacts.

S9. Visualization of the generated images

In Fig. S5 and Fig. S6, we visualize the generated down-sampled (LLR) and up-sampled (SR) images by our ICF-SRSR framework for different scaling factors ×2 and ×4, respectively on various benchmark datasets Set14 [64], BSD100 [38], and Urban100 [24] and also real-world dataset RealSR-V3 [6]. We further restore the down-sampled LR images given HR images for scaling factor ×2 of Canon and Nikon sets from the RealSR-V3 [6] dataset as shown in Fig. S7. The comparison demonstrates that the generated down-sampled LR images by our self-supervised method ICF-SRSR look similar to the real LR images, validating the ability of our method to synthesize realistic LR-HR image pairs. Such generated paired images LR-HR are useful to

Supervision	Method	Set5 ×2/×4	Set14 ×2/×4	BSD100 ×2/×4	Urban100 ×2/×4	Manga109 ×2/×4
	Bicubic	0.929/0.810	0.868/0.702	0.843/0.667	0.840/0.657	0.933/0.789
Supervised	VDSR [28]	0.959/0.884	0.912/0.768	0.896/0.725	0.914/0.752	0.975/0.887
	EDSR [34]	0.960/0.898	0.919/0.787	0.901/0.742	0.935/0.803	0.977/0.915
	CARN [2]	0.959/0.894	0.916/0.781	0.897/0.735	0.925/0.784	0.976/0.908
	RCAN [70]	0.961/0.900	0.921/0.788	0.902/0.743	0.938/0.806	0.978/0.917
	RDN [71]	0.961/0.899	0.921/0.787	0.901/0.741	0.935/0.802	0.978/0.915
	DRN-S [20]	0.960/0.901	0.910/0.790	0.900/0.744	0.920/0.807	**0.980**/0.919
	LIIF [9]	0.933/0.898	0.882/0.788	0.871/0.742	0.905/0.805	- / -
	ELAN [69]	**0.962/0.902**	**0.922/0.791**	**0.903/0.745**	**0.939/0.816**	0.979/**0.922**
Unsupervised	SelfExSR [24]	0.953/0.861	0.903/0.751	0.885/0.710	0.897/0.740	**0.968**/0.718
	ZSSR [46]	**0.957/0.879**	**0.910/0.765**	**0.892/0.721**	0.894/0.682	0.957/**0.813**
	MZSR [48]	0.956/ -	- / -	**0.892/** -	**0.909/** -	- / -
Self-supervised	**ICF-SRSR** (Ours)	0.956/0.874	0.908/0.760	0.888/0.715	0.910/0.740	0.970/0.872
	EDSR (LLR,LR) (Ours)	**0.957/0.876**	**0.909/0.763**	**0.889/0.717**	**0.911/0.745**	**0.971/0.876**

Table S1. **Quantitative comparisons of different methods on synthetic datasets by SSIM.** We compare our ICF-SRSR with several supervised and unsupervised methods on the five standard benchmark datasets [4, 64, 38, 24, 39] on scales ×2 and ×4. ICF-SRSR refers to our self-supervised method, while EDSR (LLR,LR) is the model EDSR trained on our generated pairs (LLR,LR) of the DIV2K dataset. We also note that MZSR does not report SSIM for ×4 SR in the original paper.

Baseline	Set5	Set14	BSD100	Urban100	DIV2K
ICF-SRSR (LIIF)	36.46	32.39	31.18	29.74	34.52
ICF-SRSR (EDSR)	37.01	32.86	31.54	30.39	35.19
ICF-SRSR (RDN)	37.03	32.87	31.56	30.42	35.18
ICF-SRSR (RCAN)	**37.12**	**32.92**	**31.59**	**30.50**	**35.21**

Table S2. **Evaluation of our ICF-SRSR with different baselines by PSNR metric on scale ×2.**

λ_{Color}	Canon	Nikon	Set5	DIV2K
0.1	30.62	29.97	36.24	**35.03**
0.2	**30.67**	29.99	**36.41**	35.02
1	30.63	**30.02**	36.38	34.93
10	30.61	29.98	36.35	34.82

Table S3. **Ablation on the hyperparameter λ_{Color}.**

Method	Self-Supervised	Set	Canon(×2)	Canon(×4)	Nikon(×2)	Nikon(×4)
DASR [53]	✗	DIV2K (HR)	30.66	25.98	29.74	25.25
DASR [53]	✗	RealSR-V3 (HR)	30.76	26.09	30.15	**25.94**
DASR [53]	✓	RealSR-V3 (LR)	30.68	25.38	30.08	25.13
ICF-SRSR	✓	RealSR-V3 (LR)	**30.98**	**26.26**	**30.31**	25.89

Table S4. **Quantitative comparison with DASR [53] method trained on different training datasets.**

train other off-the-shelf supervised methods, as evident in Tab. 6 of our main manuscript.

S10. Training on a single image

In Sec. 4.4 of our main manuscript, we show that our method ICF-SRSR can learn to restore SR images by training on a small dataset and even a single image as shown in Fig. 1. We show more samples to illustrate the ability of our method to learn from only a single image. Therefore, we train and evaluate our ICF-SRSR model on a single LR image from the test set of the RealSR-V3 [6] dataset captured by the Nikon camera for scaling factor ×2. Our results in Fig. S8 demonstrate that our method can restore an SR image by training the model on only the same image. Furthermore, our result for the single-image case is not only on par with the multi-image case but also shows better performance for some samples in terms of PSNR metric and visual appearance. This attribute makes our method more practical in real-world scenarios where there are not many sample images for training. Moreover, we train and evaluate our self-supervised method ICF-SRSR on a single real-world smartphone photo and show the results in Fig. S9.

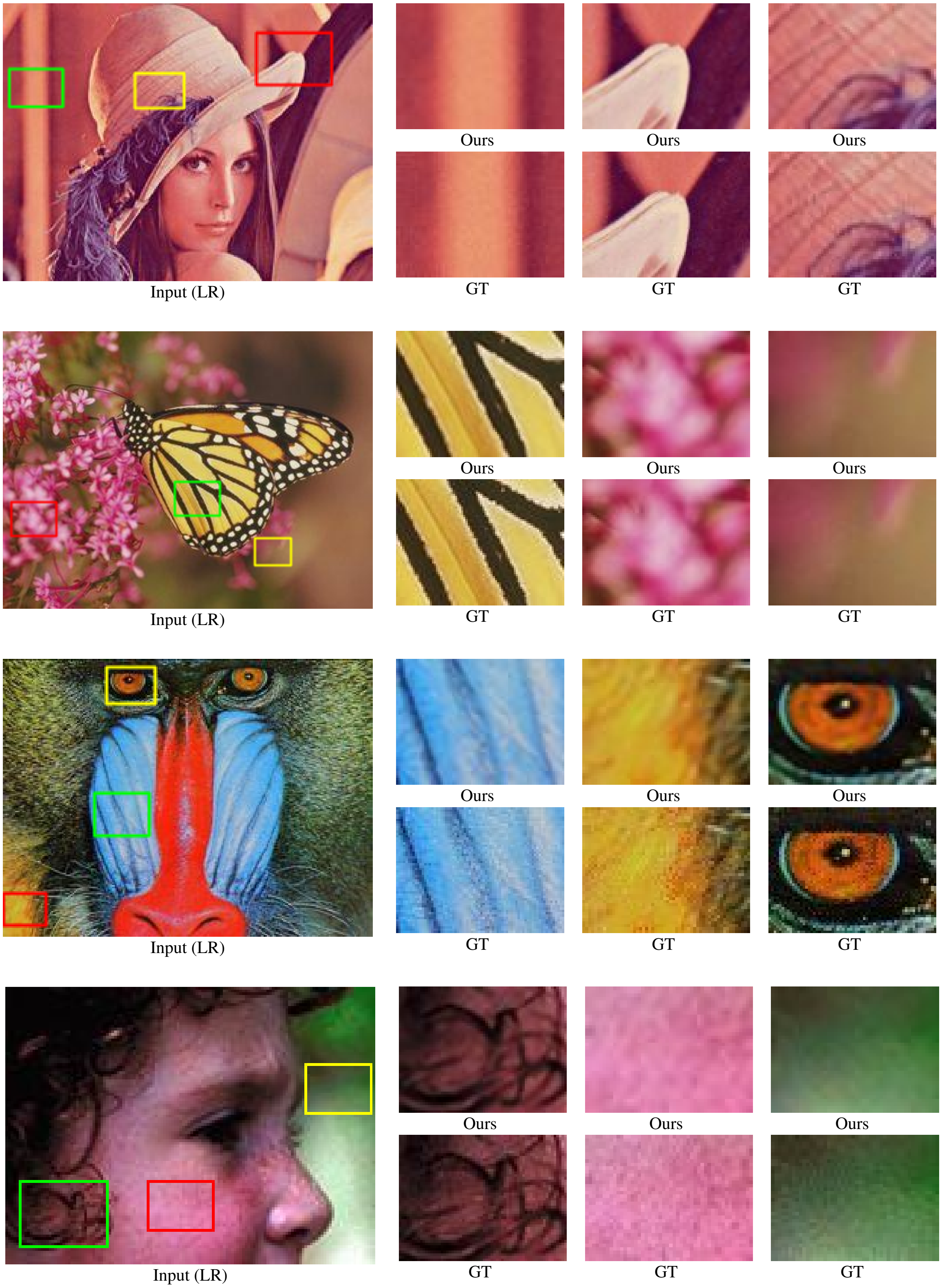


Figure S3. **Visualization of noise-free super-resolved images on scale $\times 2$.**

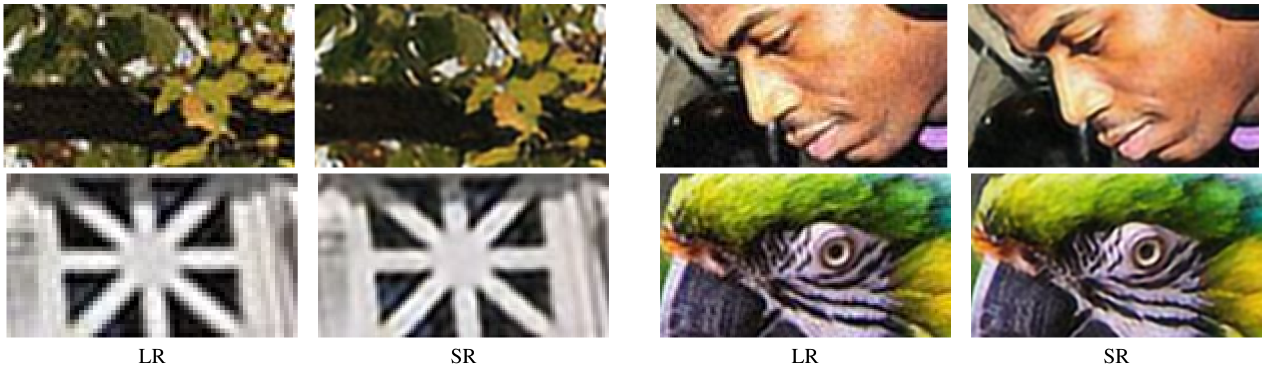


Figure S4. **Visualization of SR performance on images with more complicated down-sampling degradations.**

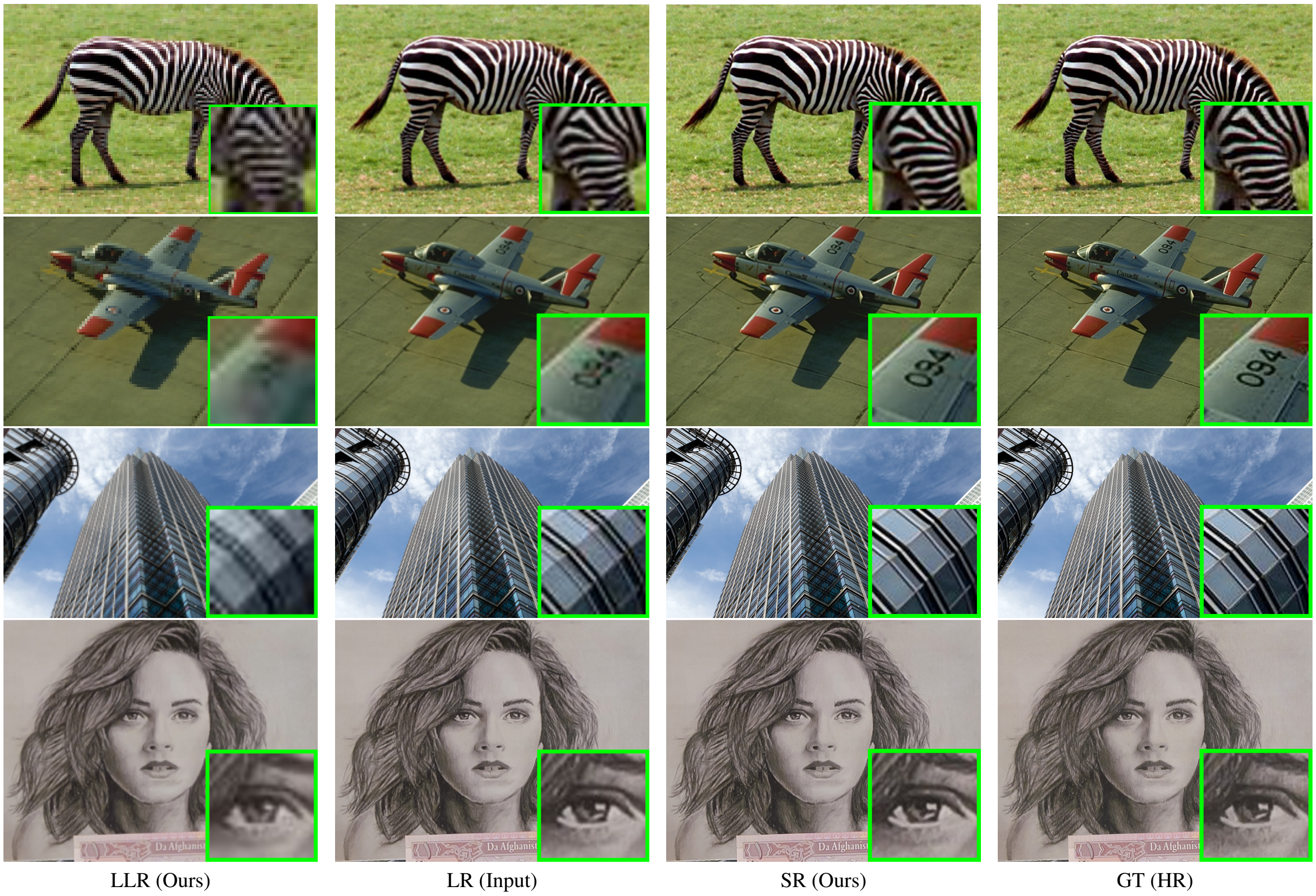
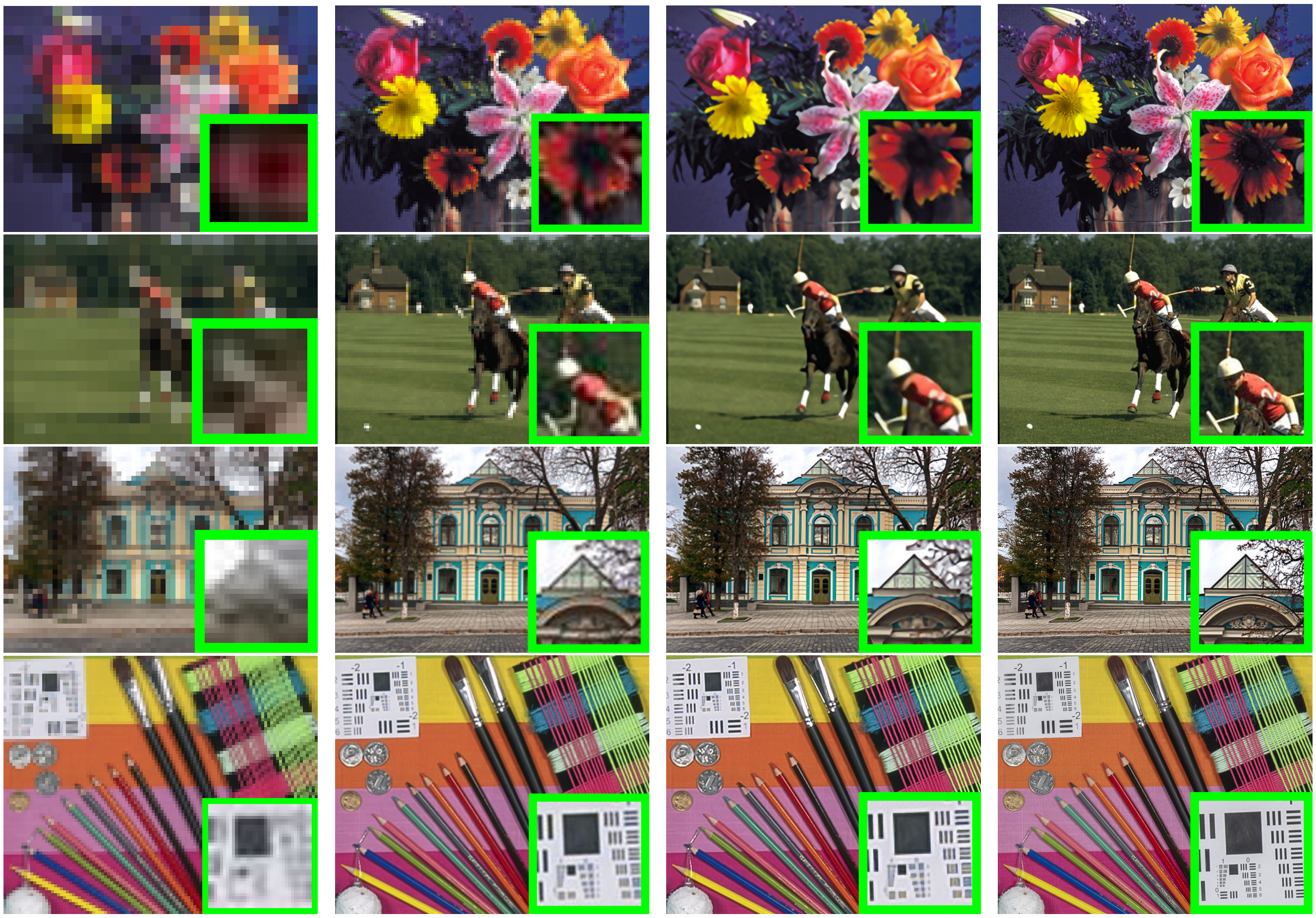
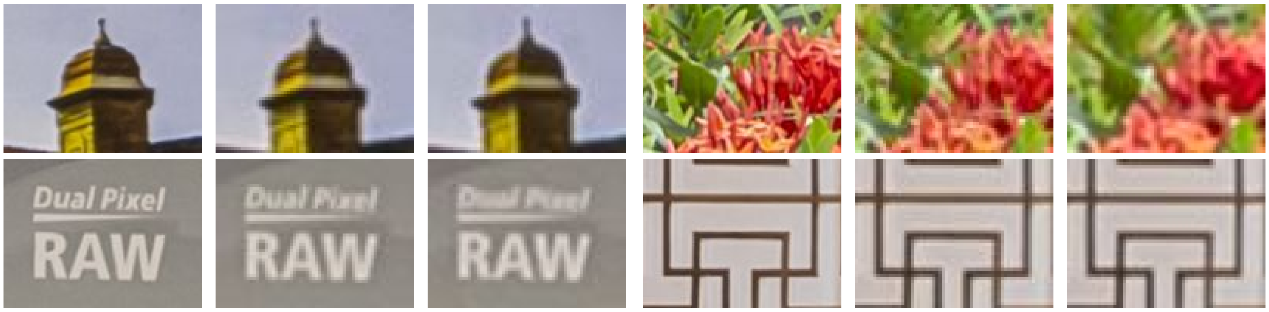


Figure S5. **Qualitative comparisons of the generated images (LLR and SR) by ICF-SRSR for scale $\times 2$.**



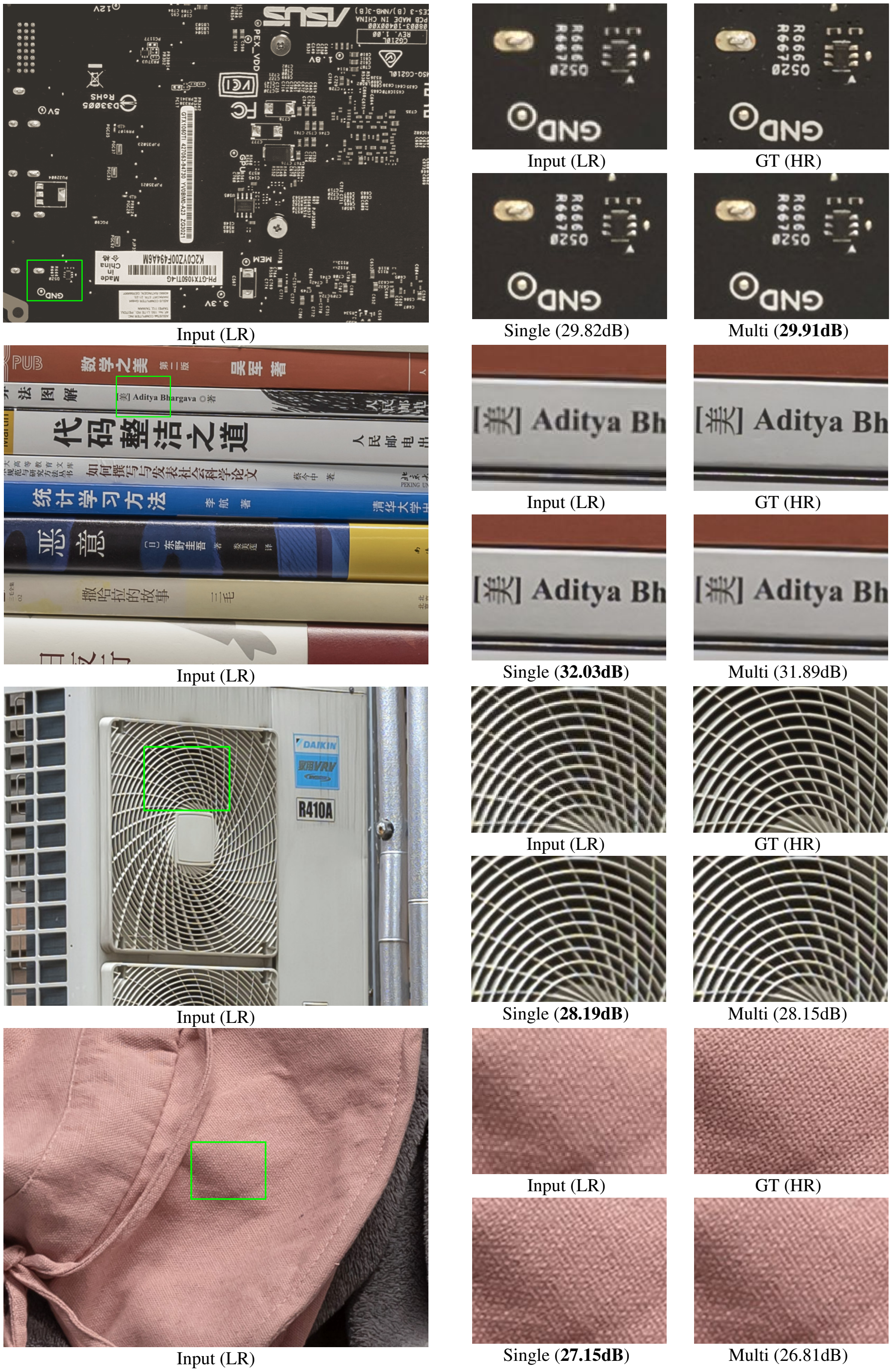
LLR (Ours) LR (Input) SR (Ours) GT (HR)

Figure S6. **Qualitative comparisons of the generated images (LLR and SR) by ICF-SRSR for scale $\times 4$.**



HR Our LR Real LR HR Our LR Real LR

Figure S7. **Qualitative comparisons of the real LR images and our generated LR images given HR images.**



Input (LR)

Input (LR)

GT (HR)

Single (29.82dB)

Multi (**29.91dB**)

Input (LR)

Input (LR)

GT (HR)

Single (**32.03dB**)

Multi (31.89dB)

Input (LR)

Input (LR)

GT (HR)

Single (**28.19dB**)

Multi (28.15dB)

Input (LR)

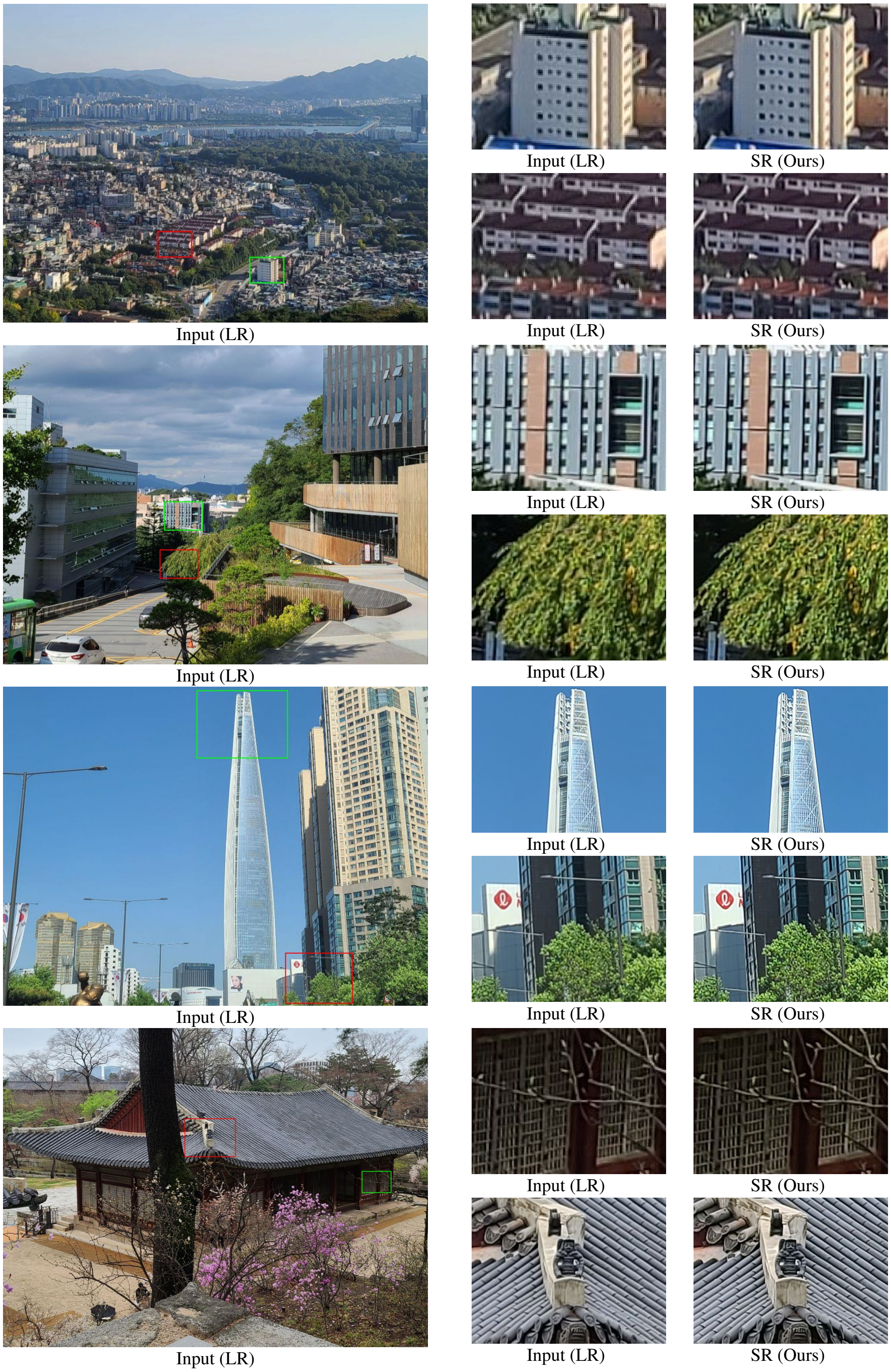
Input (LR)

GT (HR)

Single (**27.15dB**)

Multi (26.81dB)

Figure S8. **Qualitative SR comparisons on single and multiple training images for scale ×2.**



Input (LR)

SR (Ours)

Input (LR)

Input (LR)

SR (Ours)

Input (LR)

SR (Ours)

Input (LR)

Input (LR)

SR (Ours)

Input (LR)

SR (Ours)

Input (LR)

Input (LR)

SR (Ours)

Input (LR)

SR (Ours)

Input (LR)

Input (LR)

SR (Ours)

Input (LR)

SR (Ours)

Figure S9. **SR results on single training images from our captured images with scale** $\times 2$**.**