Appendix: Invertible Image Rescaling

Mingqing Xiao[1⋆], Shuxin Zheng[2†][2423], Chang Liu[2†], Yaolong Wang[3⋆], Di He[1], Guolin Ke[2], Jiang Bian[2], Zhouchen Lin[1], and Tie-Yan Liu[2]

[1] Peking University
[2] Microsoft Research Asia
[3] University of Toronto
{mingqing_xiao, di_he, zlin}@pku.edu.cn, {shuz, changliu, guoke, jiabia, tyliu}@microsoft.com, yaolong.wang@mail.utoronto.ca

1 Details of the distribution loss

According to the main text, we choose the Jensen-Shannon (JS) divergence as the distribution metric and minimize the difference between ${f_\theta^{-1}}_{\#}\left[{f_\theta^y}_{\#}[q(x)]\ p(z)\right]$ and $q(x)$:

$$
\begin{aligned}
& L_{\text{distr}}(\theta) = \text{JS}({f_\theta^{-1}}_{\#}\left[{f_\theta^y}_{\#}[q(x)]\ p(z)\right], q(x)) \\
= & \frac{1}{2}\max_T \Big\{\mathbb{E}_{q(x)}\left[\log\sigma(T(x))\right] \\
& + \mathbb{E}_{x'\sim {f_\theta^{-1}}_{\#}\left[{f_\theta^y}_{\#}[q(x)]\ p(z)\right]}\left[\log\left(1-\sigma(T(x'))\right)\right]\Big\} + \log 2 \\
= & \frac{1}{2}\max_T \left\{\mathbb{E}_{q(x)}\left[\log\sigma(T(x))\right]\right. \\
& \left. + \mathbb{E}_{(y,z)\sim {f_\theta^y}_{\#}[q(x)]\ p(z)}\left[\log\left(1-\sigma(T(f_\theta^{-1}(y,z)))\right)\right]\right\} + \log 2 \\
\approx & \frac{1}{2N}\max_T \sum_n \Big\{\log\sigma(T(x^{(n)})) \\
& + \log\left(1-\sigma(T(f_\theta^{-1}(f_\theta^y(x^{(n)}), z^{(n)})))\right)\Big\} + \log 2. \qquad (1)
\end{aligned}
$$

The first equality stems from the variational form of the JS divergence which is composed for training generative adversarial networks [3]. The second equality is a reformulation using the definition of pushed-forward distribution. The third approximate equality leads to a Monte Carlo estimation to the objective function using the corresponding samples: $\{z^{(n)}\}_{n=1}^N$ i.i.d. drawn from $p(z)$, and $\{x^{(n)}\}_{n=1}^N \sim q(x)$.

2 Detailed Training Strategies on DIV2K dataset

We train and compare our model in $2\times$ and $4\times$ downscaling scale with one and two downscaling modules respectively. Each downscaling module has 8 InvBlocks and

⋆ Work done during an internship at Microsoft Research Asia.
† Corresponding authors.

downscale the original image by $2\times$. We use Adam optimizer [5] with $\beta_1 = 0.9, \beta_2 = 0.999$ to train our model. We set weight decay as 1e-5 and gradient clipping as 10. The mini-batch size is set to 16. The input HR image is randomly cropped into 144×144 and augmented by applying random horizontal and vertical flips. In the pre-training stage, the total number of iteration is $50K$, and the learning rate is initialized as 2×10^{-4} where halved at $[10k, 20k, 30k, 40k]$ mini-batch updates. The hyper-parameters in Eqn.10 are set as $\lambda_1 = 1, \lambda_2 = 16, \lambda_3 = 1$. After pre-training, we finetune our model for another $20K$ iterations as described in Sec.3.3. The learning rate is initialized as 1×10^{-4} and halved at $[5k, 10k]$ iterations. We set $\lambda_1 = 0.01, \lambda_2 = 16, \lambda_3 = 1, \lambda_4 = 0.01$ in Eqn.11 and pre-train the discriminator for 5000 iterations. The discriminator is similar to [6], which contains eight convolutional layers with 3×3 kernels, whose numbers increase from 64 to 512 by a factor 2 each two layers, and two dense layers with 100 hidden units.

3 Quantitive results of IRN+

Table 1. Quantitative evaluation results (PSNR / SSIM) of different $4\times$ image downscaling and upscaling methods on benchmark datasets: Set5, Set14, BSD100, Urban100, and DIV2K validation set. For our model, differences on average PSNR / SSIM of different samples for z are less than 0.02. We report the mean result. The best result is in red, while the second is in blue.

Downscaling & Upscaling	Scale	Param	Set5	Set14	BSD100	Urban100	DIV2K
Bicubic & Bicubic	4×	/	28.42 / 0.8104	26.00 / 0.7027	25.96 / 0.6675	23.14 / 0.6577	26.66 / 0.8521
Bicubic & SRCNN [2]	4×	57.3K	30.48 / 0.8628	27.50 / 0.7513	26.90 / 0.7101	24.52 / 0.7221	–
Bicubic & EDSR [7]	4×	43.1M	32.62 / 0.8984	28.94 / 0.7901	27.79 / 0.7437	26.86 / 0.8080	29.38 / 0.9032
Bicubic & RDN [11]	4×	22.3M	32.47 / 0.8990	28.81 / 0.7871	27.72 / 0.7419	26.61 / 0.8028	–
Bicubic & RCAN [10]	4×	15.6M	32.63 / 0.9002	28.87 / 0.7889	27.77 / 0.7436	26.82 / 0.8087	30.77 / 0.8460
Bicubic & ESRGAN [9]	4×	16.3M	32.74 / 0.9012	29.00 / 0.7915	27.84 / 0.7455	27.03 / 0.8152	30.92 / 0.8486
Bicubic & SAN [1]	4×	15.7M	32.64 / 0.9003	28.92 / 0.7888	27.78 / 0.7436	26.79 / 0.8068	–
TAD & TAU [4]	4×	–	31.81 / –	28.63 / –	28.51 / –	26.63 / –	31.16 / –
CAR & EDSR [8]	4×	52.8M	33.88 / 0.9174	30.31 / 0.8382	29.15 / 0.8001	29.28 / 0.8711	32.82 / 0.8837
IRN (ours)	4×	4.35M	36.19 / 0.9451	32.67 / 0.9015	31.64 / 0.8826	31.41 / 0.9157	35.07 / 0.9318
IRN+ (ours)	4×	4.35M	33.59 / 0.9147	29.97 / 0.8444	28.94 / 0.8189	28.24 / 0.8684	32.24 / 0.8921

IRN+ aims at producing more realistic images by minimizing the distribution difference, not exactly matching details of original images as IRN does. The difference will lead to lower PSNR and SSIM, which is the same as GAN-based super-resolution methods. Despite the difference, IRN+ still outperforms most methods in PSNR and SSIM as shown in Table.1, demonstrating the good similarity between the reconstructed images and original HR images.

4 Different samples of z

As shown in Fig. 1, there are only tiny noisy distinction in high-frequency areas without typical textures, which can hardly perceived when combined with low-frequency contents. Different samples lead to different but perceptually meaningless noisy distinctions.

Fig. 1. Difference between upscaled images by different samples of z. (a): Original image. (b-d): Residual of three randomly upscaled images with another sample (averaged over the three channels). (e-g): Detailed difference of (b-d). The darker the larger difference. To ensure the visual perception, we set rebalance factor by 20.

5 More qualitative results

As shown in Fig.2,3,4,5, images reconstructed by IRN and IRN+ significantly outperforms previous both PSNR-oriented and perceptual-driven methods in visual quality and similarity to original images. IRN is able to reconstruct rich details including detailed lines and textures, which contributes to the pleasing perception. IRN+ further produce sharper and more realistic images as a result of the distribution matching objective.

6 Evaluation on downscaled images

As shown in Fig. 6, images downscaled by IRN share a similar visual perception with images downscaled by bicubic.

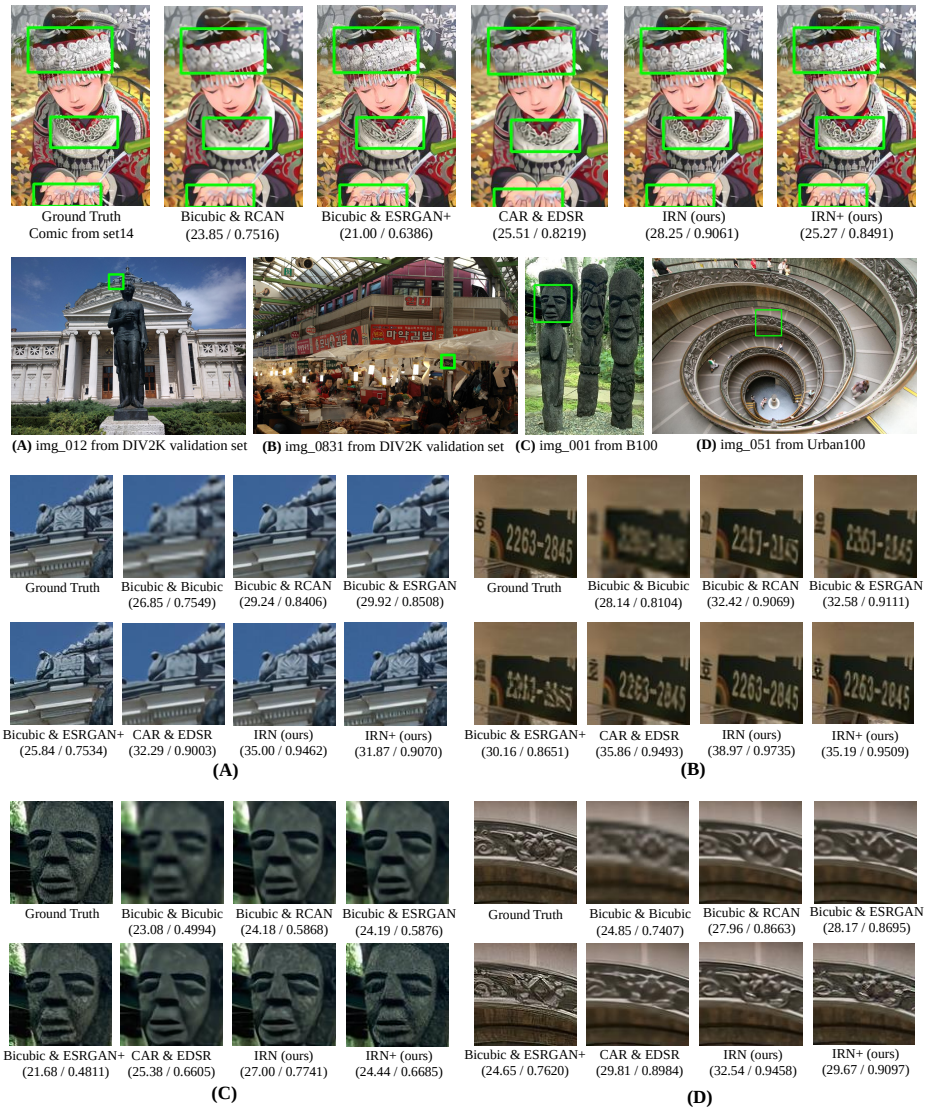

Fig. 2. More qualitative results of upscaling the $4\times$ downscaled images on Set14, BSD100, Urban100 and DIV2K validation datasets.

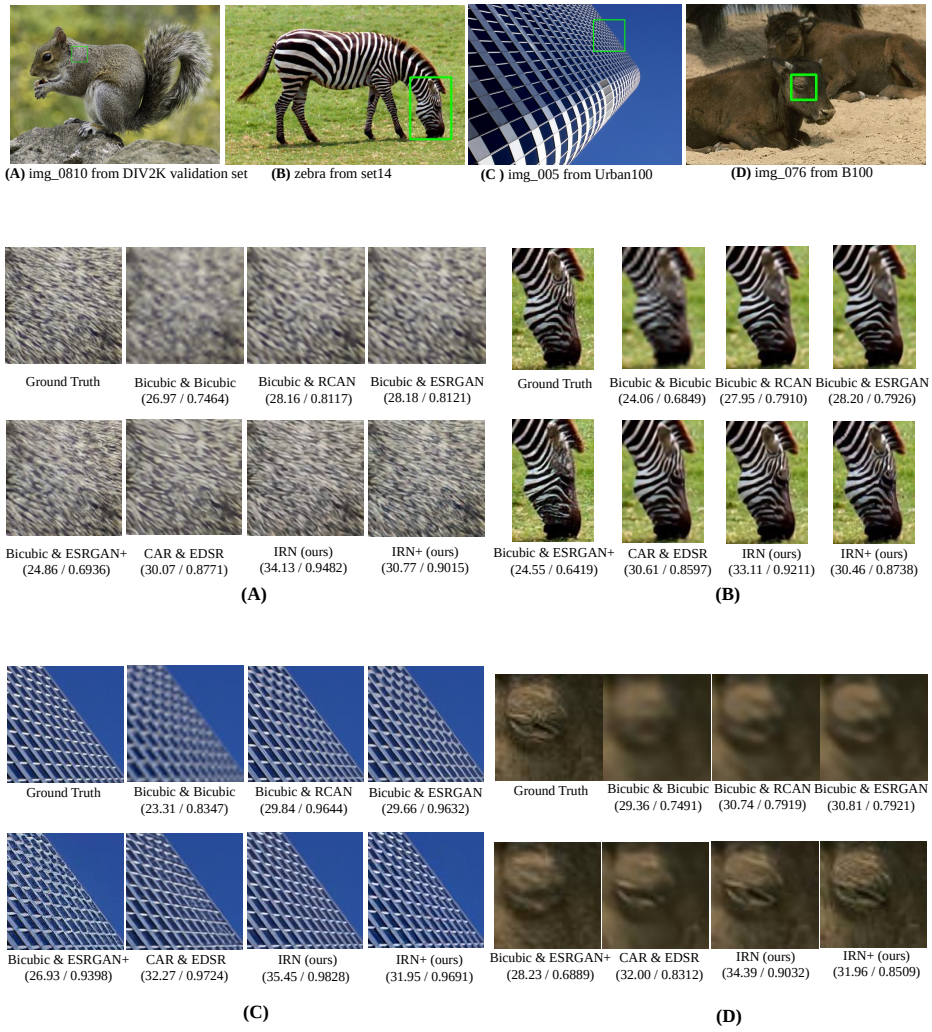

Fig. 3. More qualitative results of upscaling the $4\times$ downscaled images on Set14, BSD100, Urban100 and DIV2K validation datasets.

Fig. 4. More qualitative results of upscaling the $4\times$ downscaled images on DIV2K validation dataset.

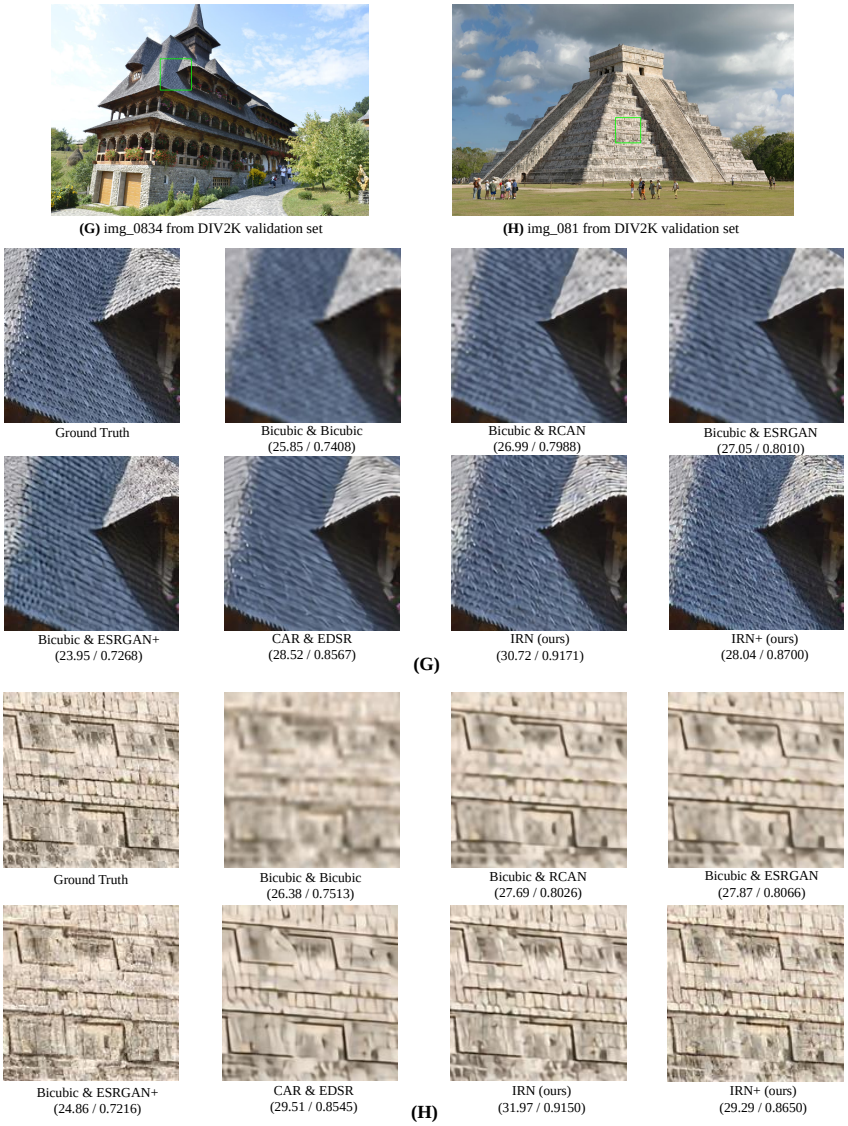

Fig. 5. More qualitative results of upscaling the $4\times$ downscaled images on DIV2K validation dataset.

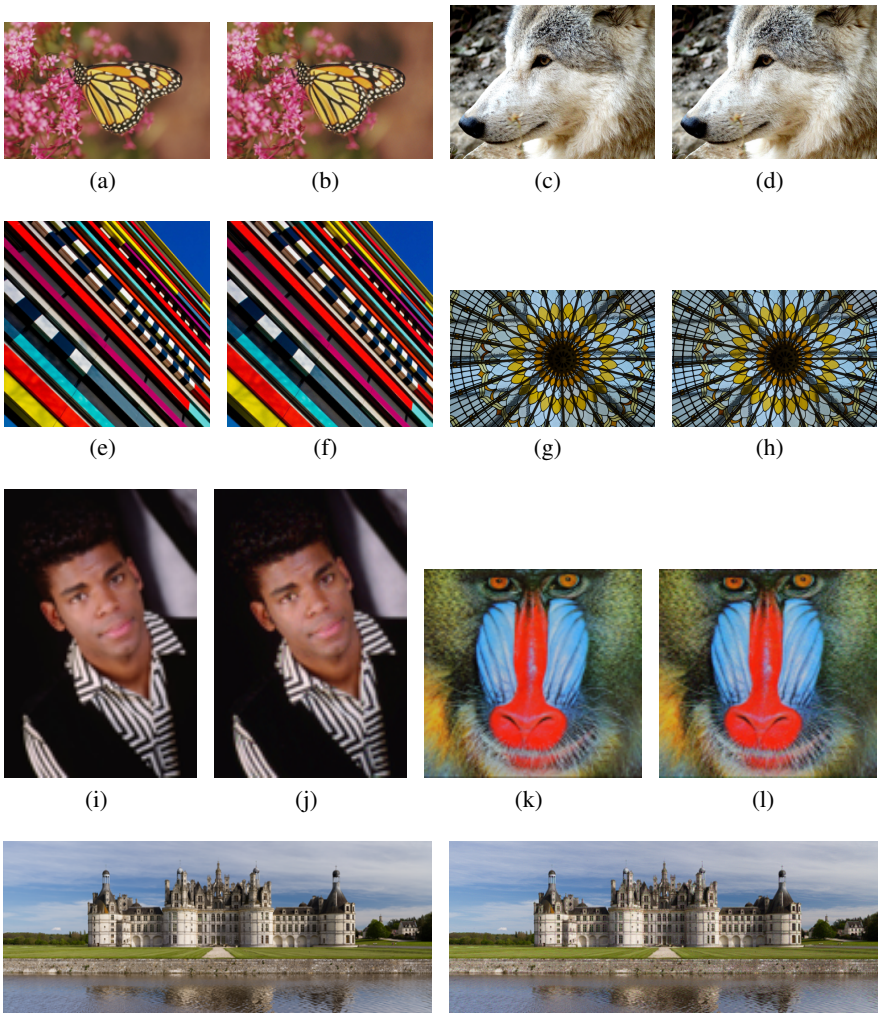

Fig. 6. Demonstration of the downscaled images from Set14, B100, Urban100, and DIV2K validation set. Left column (a,c,e,g,i,k,m): Image downscaled by Bicubic. Right column (b,d,f,h,j,l,n): Image downscaled by IRN. They share a similar visual perception.

References

1. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11065–11074 (2019) 2
2. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015) 2
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680. NIPS Foundation, Montréal, Canada (2014) 1
4. Kim, H., Choi, M., Lim, B., Mu Lee, K.: Task-aware image downscaling. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 399–414 (2018) 2
5. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 2
6. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017) 2
7. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017) 2
8. Sun, W., Chen, Z.: Learned image downscaling for upscaling using content adaptive resampler. IEEE Transactions on Image Processing **29**, 4027–4040 (2020) 2
9. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018) 2
10. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 286–301 (2018) 2
11. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2472–2481 (2018) 2