Supplementary for SwinIR: Image Restoration Using Swin Transformer

Jingyun Liang[1] Jiezhang Cao[1] Guolei Sun[1] Kai Zhang[1] Luc Van Gool[1,2] Radu Timofte[1]
[1]Computer Vision Lab, ETH Zurich, Switzerland [2] KU Leuven, Belgium
{jinliang, jiezcao, guosun, kaizhang, vangool, timofter}@vision.ee.ethz.ch
https://github.com/JingyunLiang/SwinIR

1. Training and Evaluation Details

Training. For classical and lightweight image SR, following [29, 18, 17], we train SwinIR on 800 training images of DIV2K [1]. Some compared methods (*e.g.*, [7], [23]) further use 2560 images from Flickr2K [20] for training, so we also train SwinIR on larger datasets (DIV2K+Flickr2K) to investigate whether SwinIR can further improve its performance. For fair comparison, we use 48×48 and 64×64 LQ image patches respectively in above two cases following the common settings. The HQ-LQ image pairs are obtained by the MATLAB bicubic kernel. The total training iterations and mini-batch size are set to 500K and 32, respectively. The learning rate is initialized as 2e-4 and reduced by half at [250K,400K,450K,475K]. For $\times 3$, $\times 4$ and $\times 8$ classical image SR, we initialize the model with $\times 2$ weights and halve the learning rate as well as total training iterations. Unlike other Transformer-based models that often uses AdamW [13] optimizer with cosine learning rate decay strategy, we find that using Adam [10] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ leads to better performance.

For real-world image SR, we use the same image degradation model as BSRGAN [28] and train it on a combination of DIV2K, Flickr2K and OST [22]. The model is trained for 1,000K iterations for the PSNR training stage. The learning rate is halved at [500K,800K,900K,950K]. For the GAN training stage, we train it for 600K iterations and the learning rate is halved at [400K,500K,550K,575K]. Weighting parameters between L_1 pixel loss, perceptual loss and GAN loss are 1, 1 and 0.1, respectively. Note that we use the same EMA strategy, USM strategy, perceptual loss and GAN loss as [21].

For denoising and compression artifact reduction, following [30, 27], we use random crops from the combination of 800 DIV2K images, 2650 Flickr2K images, 400 BSD500 images [2] and 4744 WED images [14]. The batch size is 8. The patch sizes are 128×128 (window size is 8×8) and 126×126 (window size is 7×7), respectively. We obtain noisy images by adding additive white Gaussian noises (AWGN) with noise level σ, and compressed images by the MATLAB JPEG encoder with JPEG level q. The total training iterations and mini-batch size are set to 1600K and 8, respectively. The learning rate is halved at [800K,1200K,1400K,1500K]. When $\sigma = 15$ or $q = 40$, we train the model from scratch. When $\sigma = 25/50$ or $q = 10/20/30$, we fine-tune from $\sigma = 15$ or $q = 40$. Other details are the same as classical SR.

Evaluation. Following the tradition of image SR, we report PSNR and SSIM [24] on the Y channel of the YCbCr space. For image denoising, we report the PSNR on the RGB channel and Y channel for color and grayscale denoising, respectively. For compression artifact reduction, in addtion to the Y channel PSNR and SSIM, we also report PNSR-B [25] that is specially designed for deblocking quality assessment. Particularly, we pad the image in testing so that the image size is a multiple of window size. We also find that using a sliding window strategy [4] to crop the image into patches can further improve the PSNR by $0.02 \sim 0.03$dB at the cost of longer testing time, so we do not use it for comparison.

2. Results on image SR ($\times 8$)

We show the comparison on classical image SR ($\times 8$) in Table 1.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 1

[2] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010. 1

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie line Alberi Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *British Machine Vision Conference*, pages 135.1–135.10, 2012. 2

Table 1: Quantitative comparison (average PSNR/SSIM) with state-of-the-art methods for **classical image SR** ($\times 8$) on benchmark datasets. Best and second best performance are in red and blue colors, respectively.

Method	Scale	Training Dataset	Set5 [3]		Set14 [26]		BSD100 [15]		Urban100 [8]		Manga109 [16]	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SRCNN [6]	×8	DIV2K	25.33	0.6900	23.76	0.5910	24.13	0.5660	21.29	0.5440	22.46	0.6950
VDSR [9]	×8	DIV2K	25.93	0.7240	24.26	0.6140	24.49	0.5830	21.70	0.5710	23.16	0.7250
LapSRN [11]	×8	DIV2K	26.15	0.7380	24.35	0.6200	24.54	0.5860	21.81	0.5810	23.39	0.7350
MemNet [19]	×8	DIV2K	26.16	0.7414	24.38	0.6199	24.58	0.5842	21.89	0.5825	23.56	0.7387
EDSR [12]	×8	DIV2K	26.96	0.7762	24.91	0.6420	24.81	0.5985	22.51	0.6221	24.69	0.7841
RCAN [29]	×8	DIV2K	27.31	0.7878	25.23	0.6511	24.98	0.6058	23.00	0.6452	25.24	0.8029
SAN [5]	×8	DIV2K	27.22	0.7829	25.14	0.6476	24.88	0.6011	22.70	0.6314	24.85	0.7906
HAN [18]	×8	DIV2K	27.33	0.7884	25.24	0.6510	24.98	0.6059	22.98	0.6347	25.20	0.8000
SwinIR (Ours)	×8	DIV2K	27.37	0.7877	25.26	0.6523	24.99	0.6063	23.03	0.6457	25.26	0.8005
SwinIR+ (Ours)	×8	DIV2K	27.47	0.7907	25.34	0.6546	25.03	0.6078	23.12	0.6499	25.42	0.8047
DBPN [7]	×8	DIV2K+Flickr2K	27.21	0.7840	25.13	0.6480	24.88	0.6010	22.73	0.6312	25.14	0.7987
SwinIR (Ours)	×8	DIV2K+Flickr2K	27.55	0.7941	25.46	0.6568	25.04	0.6092	23.17	0.6547	25.55	0.8132
SwinIR+ (Ours)	×8	DIV2K+Flickr2K	27.59	0.7952	25.51	0.6588	25.08	0.6104	23.27	0.6581	25.73	0.8167

[4] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 1

[5] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019. 2

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pages 184–199, 2014. 2

[7] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1664–1673, 2018. 1, 2

[8] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015. 2

[9] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016. 2

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[11] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 624–632, 2017. 2

[12] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017. 2

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[14] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2016. 1

[15] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE Conference on International Conference on Computer Vision*, pages 416–423, 2001. 2

[16] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017. 2

[17] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3517–3526, 2021. 1

[18] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European Conference on Computer Vision*, pages 191–207, 2020. 1, 2

[19] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *IEEE International Conference on Computer Vision*, pages 4539–4547, 2017. 2

[20] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 114–125, 2017. 1

[21] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. *arXiv preprint arXiv:2107.10833*, 2021. 1

[22] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 606–615, 2018. 1

[23] 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 *European Conference on Computer Vision Workshops*, pages 701–710, 2018. 1

[24] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1

[25] Changhoon Yim and Alan Conrad Bovik. Quality assessment of deblocked images. *IEEE Transactions on Image Processing*, 20(1):88–98, 2010. 1

[26] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, pages 711–730, 2010. 2

[27] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[28] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE Conference on International Conference on Computer Vision*, 2021. 1

[29] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision*, pages 286–301, 2018. 1, 2

[30] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2480–2495, 2020. 1