iiTransformer for Image Restoration Supplementary Material

Soo Min Kang
sm2.kang@samsung.com
Youngchan Song
yc87.song@samsung.com
Hanul Shin
hsky.shin@samsung.com
Tammy Lee
tammy.lee@samsung.com

Samsung Research
Samsung Electronics Co., Ltd.
Seoul, Republic of Korea

1 Introduction

This supplemental material serves four purposes. First, it provides details on training and evaluation for each task in Sec. 2. Second, it provides various ways of incorporating local and non-local correlations within the iiTransformer architecture and its corresponding quantitative results in Sec. 3. Third, results across *all* degradation levels of the ablation study are provided in Sec. 4. Finally, qualitative results comparing iiTransformer to other methods across *all* degradation levels are provided in Sec. 5.

2 Experimental Details

The potential of iiTransformer as a general purpose backbone architecture for image restoration is demonstrated through experimentation on three restoration tasks: image denoising, compression artifact removal, and single image super-resolution. Below, we provide training and evaluation details associated with each task.

Image Denoising. Image denoising is the process of recovering a clean image from its noisy counterpart. Various types of noise are prevalent in the real-world (e.g., camera sensor noise, Poisson noise). For synthetic noise, we use the common assumption: additive white Gaussian noise (AWGN) to obtain LQ-HQ pairs. AWGN is added to clean images with a standard deviation $\sigma = \{15, 25, 50\}$ on patches of size 128×128 with a batch size of 8 for training. For real noise, we consider noise that appear *after* the images are processed in-camera that maps the sensor-dependent RGB colours to a device-independent colour space and use data collected specifically for this task (i.e., sRGB of SIDD [1]). We report PSNR and SSIM on the RGB channels evaluated on Kodak24 [6], BSDS68 [10], McMaster18 [16], and Urban100 [7] for synthetic noise and SIDD [1] for real noise.

CAR. Compression artifact reduction (CAR) is the process of removing artifacts caused by compression algorithms, such as blocking, ringing, and blurry artifacts. Among several compression algorithms for storage and/or bandwidth reduction (e.g, WebP, MPEG2), we consider one of the most widely used image compression algorithms, JPEG, on four quality factors: $q = \{10, 20, 30, 40\}$, where the compressed LQ images are obtained using the MATLAB JPEG encoder. We use patches of size 128×128 with a batch size of 32 for training. Following the protocol of other CAR methods, we report PSNR and SSIM on the Y channel for two benchmark datasets: Classic5 [5] and LIVE1 [17].

SISR. Single image super-resolution (SISR) aims to reconstruct a natural and sharp detailed high-resolution image given a low-resolution input. Following the tradition of SISR, we report PSNR and SSIM on the Y channel of the YCbCr space for five benchmark datasets: Set5 [2], Set14 [15], BSDS100 [10], Urban100 [7], and Manga109 [11]. We consider three scales: $s = \{2, 3, 4\}$, where the low-resolution images are obtained by the MATLAB bicubic interpolation method. We used patches of size $(LQ, HQ) = (64 \times 64,\ s64 \times s64)$, where s denotes the scale, and a batch size of 32.

3 Combining Local and Non-local Information

Local and non-local correlations can be combined in various ways. In this section, we explore the various design choices by replacing each component of the framework described in Sec. 3 and Fig. 3 of the main manuscript, which we refer to as the *baseline*. The hyperparameters for each design choice was chosen to ensure that the total number of MSAs remained consistent across the experiments. For quantitative evaluation of each fusion method, we provide PSNR / SSIM on SISR for scale $s = 2$ in Tab. 1.

(a) The baseline iiTB structure sequentially processes the input feature through aRTB followed by eRTB (see Fig. 3b of main). Since features from aRTB are generated by considering internal correlations within a patch and eRTB is based on correlations across patches, we tested the importance of sequential *order* by altering the order of aRTB and eRTB in iiTB, as in Fig. 1a. That is, there could be benefits of seeking structurally similar patches through eRTB, then finetuning the details through aRTB. However, our quantitative evaluation on SISR at scale $s = 2$ indicated similar or slight degradation in performance across benchmark datasets. This could imply that there are greater benefits of processing features based on its neighbouring pixels then seeking non-local similarities than vice versa.

(b) The baseline residual Transformer blocks in Fig. 3c of main are designed to consecutively apply intra (or inter) correlations using aTLs (or eTLs) to obtain deep features based on local (or non-local) information. We observed the effectiveness of aMSA and eMSA interaction frequency by interweaving aTLs and eTLs to build the residual Transformer block as shown in Fig. 1b. Our quantitative evaluation indicated deteriorating performance compared to the baseline signifying the necessity to apply multiple aMSAs (or eMSAs) consecutively for meaningful feature extraction.

(c, d) The baseline Transformer layer and iiTransformer block correlates local and non-local relationship of features in a *sequential* manner. That is, the baseline Transformer layer in Fig. 3d of the main manuscript considers either aMSA or eMSA to build aTL or eTL, resp., but not simultaneously. We considered processing the features through aMSA and eMSA in parallel within a Transformer layer by concatenating their outputs as in Fig. 1c. Alternatively, rather than sequentially processing aRTB followed by eRTB as in Fig. 3b of main, we considered concatenating the output features of aRTB and eRTB within a single

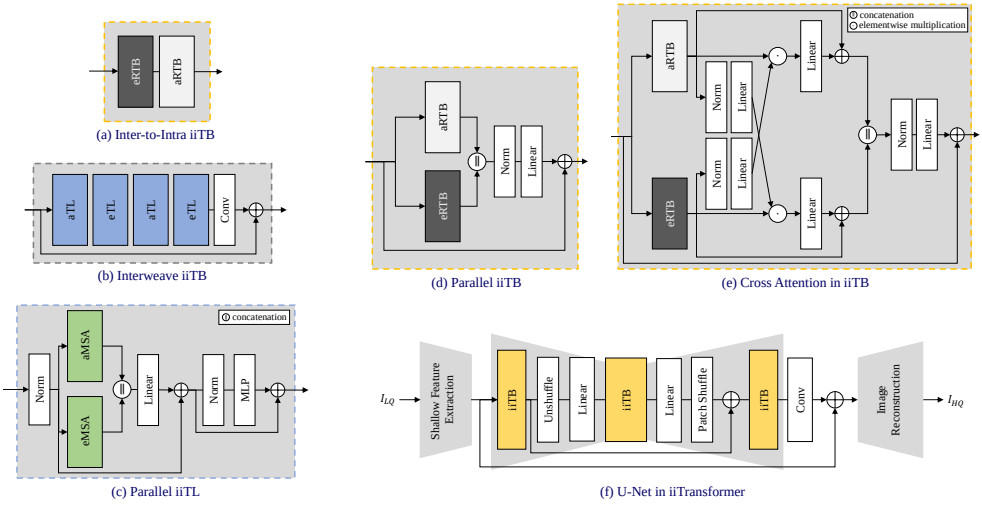


Figure 1: Alternative ways of combining local and non-local information.

Architecture	Set5 [2]	Set14 [35]	BSDS100 [21]	Urban100 [9]	Manga109 [20]
Baseline	**38.25 / 0.9611**	**34.08** / 0.9207	**31.81** / 0.8954	33.27 / 0.9378	**39.36 / 0.9781**
(a) Inter-to-Intra iiTB	**38.25 / 0.9611**	34.07 / **0.9214**	**31.81 / 0.8955**	33.23 / **0.9379**	39.32 / 0.9780
(b) Intertweave iiTB	38.24 / 0.9610	33.90 / 0.9207	31.83 / 0.8957	33.25 / 0.9377	39.34 / 0.9780
(c) Parallel iiTL	38.20 / 0.9609	33.89 / 0.9199	31.80 / 0.8953	33.09 / 0.9364	39.24 / 0.9775
(d) Parallel iiTB	38.19 / 0.9608	33.93 / 0.9210	31.78 / 0.8949	32.99 / 0.9356	39.22 / 0.9776
(e) Cross Attention in iiTB	38.22 / 0.9609	33.97 / 0.9206	31.79 / 0.8951	33.09 / 0.9362	39.12 / 0.9773
(f) U-Net in iiTransformer	38.24 / 0.9609	34.05 / 0.9211	31.81 / 0.8954	**33.28 / 0.9379**	39.35 / 0.9780

Table 1: Quantitative evaluation for SISR at scale $\times 2$ comparing various local and non-local information fusion methods. The best performing method is marked in **bold**.

iiTB as in Fig. 1d. We observed no quantitative benefits in parallel processing over sequential processing.

(e) We explored selective gating via cross-attention, as in [14], by using the output of aRTB as an attention map of eRTB and vice versa, as shown in Fig. 1e, such that the features from aRTB and eRTB have a more direct interaction with one another. However, quantitative analysis showed no benefit in applying such cross-attention approach compared to the baseline framework proposed in Sec. 3 of the main paper.

(f) We adapted the U-Net structure that is widely used in image restoration within iiTransformer by applying down/upscaling operations on the intermediate features output by iiTBs. Downscaling processes the data in a fine-to-coarse manner to consider larger scope of the image, while upscaling processes the data in a coarse-to-fine manner to enable generation of details at the finer scale. Correspondingly, we spatially downscaled the intermediate features using the unshuffle operation and used the linear layer to reduce the number of channels; and conversely upscaled the features spatially using the shuffle operation [33] and increased the number of channels using a linear layer (see Fig. 1f). Quantitative analysis showed very little benefits in applying the multi-scale approach compared to the baseline framework.

While we experimented with several sophisticated designs of local and non-local information fusion, we observed marginal benefits, if there were any. Indeed, the sequential process of aRTB followed by eRTB in iiTB presented in the main manuscript yielded the most optimal results.

4 Additional Ablation Results

We studied the effects of employing local and/or non-local information on *select* degradation levels in Sec. 4.1 of the main manuscript. Here, we provide ablation results on *all* degradation levels in Tab. 2. Note that intraTransformer shares a very similar structure as SwinIR [9] with the main difference in the hyperparameters (e.g., number of intra MSAs) along with the specifics on the training strategy (e.g., dataset); hence, there are slight deviations in PSNR and SSIM values presented in Tab. 2 of the supplemental and Tab. 2 of the main manuscript. While it is generally favourable to use both local and non-local information in unison as done in iiTransformer, some datasets at specific tasks and degradation levels indicate otherwise. For example, Set5 [2] and Set14 [15] for SISR display slight preference in intraTransformer (Transformer containing only of intra MSAs) over iiTransformer and a few datasets in AWGN denoising at $\sigma = 50$ indicate similar performance between inter-Transformer (Transformer containing only of inter MSAs) and iiTransformer. Exploration on degradation-specific gating mechanism that selects the more ideal sub-region level for long-range computation (i.e., pixel or patch-level) is a promising direction for future work.

Dataset	Noise Level	intraTransformer	interTransformer	iiTransformer
Kodak24 [8]	15	34.99 / 0.9264	35.04 / 0.9271	**35.09 / 0.9275**
BSDS68 [10]		33.90 / 0.9296	33.94 / 0.9300	**33.96 / 0.9302**
McMaster18 [14]		34.76 / 0.9081	34.78 / 0.9084	**34.85 / 0.9092**
Urban100 [6]		34.29 / 0.9317	34.48 / 0.9334	**34.53 / 0.9336**
Kodak24 [8]	25	32.25 / 0.8833	32.31 / 0.8846	**32.35 / 0.8848**
BSDS68 [10]		30.94 / 0.8793	30.98 / **0.8799**	**31.01** / 0.8798
McMaster18 [14]		31.64 / 0.8515	31.67 / 0.8522	**31.73 / 0.8523**
Urban100 [6]		31.42 / 0.8983	31.68 / **0.9021**	**31.74 / 0.9021**
Kodak24 [8]	50	28.01 / 0.7859	**28.09 / 0.7896**	**28.09** / 0.7882
BSDS68 [10]		26.64 / 0.7628	**26.70 / 0.7660**	**26.70** / 0.7642
McMaster18 [14]		26.45 / 0.7343	26.51 / **0.7376**	**26.53** / 0.7366
Urban100 [6]		26.59 / 0.8212	26.88 / **0.8330**	**26.91** / 0.8314

(a) AWGN image denoising for noise levels $\sigma \in \{15, 25, 50\}$

Dataset	Quality	intraTransformer	interTransformer	iiTransformer
Classic5 [5]	10	30.03 / 0.8189	30.02 / 0.8189	**30.06 / 0.8196**
LIVE1 [11]		28.87 / 0.8223	28.87 / 0.8228	**28.91 / 0.8232**
Classic5 [5]	20	32.27 / 0.8711	32.28 / 0.8712	**32.36 / 0.8722**
LIVE1 [11]		31.29 / 0.8871	31.27 / 0.8871	**31.33 / 0.8878**
Classic5 [5]	30	33.54 / 0.8938	33.56 / 0.8939	**33.61 / 0.8945**
LIVE1 [11]		32.73 / 0.9144	32.73 / 0.9144	**32.77 / 0.9149**
Classic5 [5]	40	34.37 / 0.9068	34.37 / 0.9068	**34.44 / 0.9074**
LIVE1 [11]		33.72 / 0.9291	33.72 / 0.9292	**33.78 / 0.9296**

(b) JPEG CAR for qualities $q \in \{10, 20, 30, 40\}$

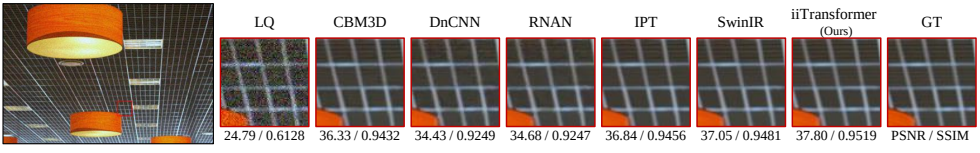
Dataset	Scale	intraTransformer	interTransformer	iiTransformer
Set5 [2]	2	38.22 / 0.9610	38.12 / 0.9603	**38.25 / 0.9611**
Set14 [15]		33.97 / **0.9208**	33.96 / 0.9201	**34.08** / 0.9207
BSDS100 [10]		31.79 / 0.8949	31.77 / 0.8947	**31.81 / 0.8954**
Urban100 [6]		33.07 / 0.9362	32.93 / 0.9356	**33.27 / 0.9378**
Manga109 [13]		39.28 / 0.9775	39.11 / 0.9773	**39.36 / 0.9781**
Set5 [2]	3	**34.78 / 0.9299**	34.59 / 0.9285	34.75 / 0.9298
Set14 [15]		30.69 / **0.8489**	30.54 / 0.8455	**30.70** / 0.8487
BSDS100 [10]		28.95 / 0.8046	28.89 / 0.8034	**28.96 / 0.8055**
Urban100 [6]		28.97 / 0.8677	28.71 / 0.8627	**29.12 / 0.8708**
Manga109 [13]		34.34 / 0.9489	33.96 / 0.9465	**34.43 / 0.9493**
Set5 [2]	4	**32.61 / 0.9007**	32.45 / 0.8976	32.58 / 0.9003
Set14 [15]		28.80 / 0.7869	28.67 / 0.7840	**28.90 / 0.7887**
BSDS100 [10]		27.37 / 0.7331	27.36 / 0.7324	**27.41 / 0.7339**
Urban100 [6]		26.64 / 0.8023	26.48 / 0.7972	**26.75 / 0.8049**
Manga109 [13]		31.13 / 0.9166	30.87 / 0.9129	**31.20 / 0.9168**

(c) SISR for scales $s = \{2, 3, 4\}$

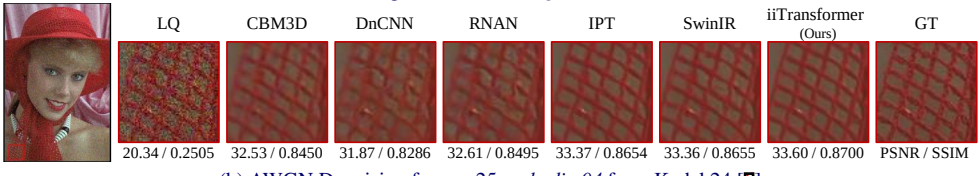
Table 2: Full ablation experimental results on using intra MSAs only (intraTransformer), inter MSAs only (interTransformer), and both (iiTransformer) for all considered degradation level across tasks. The best performing algorithm is in **bold**.

5 Additional Qualitative Results

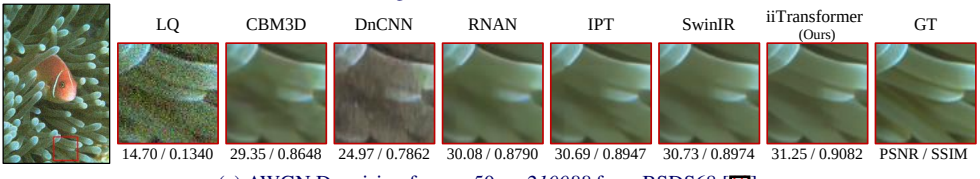
We provide comparative qualitative results on *all* degradation levels for AWGN image denoising in Fig. 2, JPEG CAR in Fig. 3, and SISR in Fig. 4. It can be seen that iiTransformer is able to restore distinct and sharp images from its degraded LQ inputs. It is particularly interesting to note that the stripes created by the book stack in Fig. 4c are restored similarly between local-based methods (i.e., bicubic, SRResNet [8], RCAN [17], and SwinIR [9]) in the northwest-southeast direction, while methods with a non-local correlation mechanism (i.e., SAN [4], IPT [3], and iiTransformer) super-resolve the book stack in the southwest-northeast direction.



(a) AWGN Denoising for $s = 15$ on *img_044* from Urban100 [7]

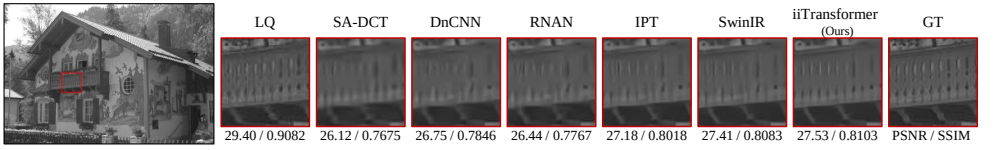


(b) AWGN Denoising for $s = 25$ on *kodim04* from Kodak24 [6]

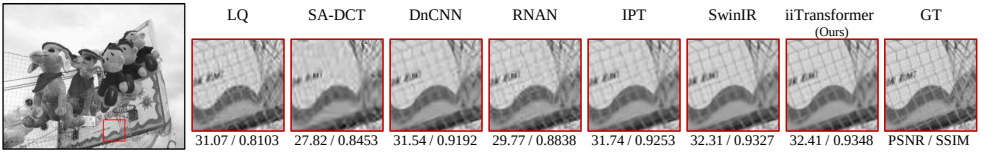


(c) AWGN Denoising for $s = 50$ on *210088* from BSDS68 [10]

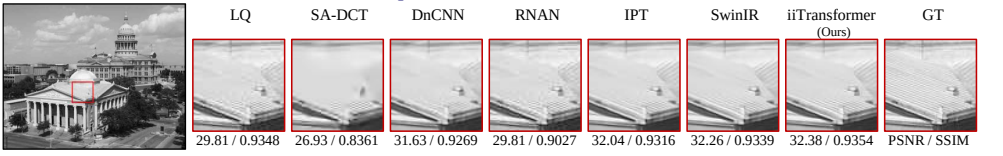
Figure 2: Qualitative comparison of AWGN denoising on various state-of-the-art methods.



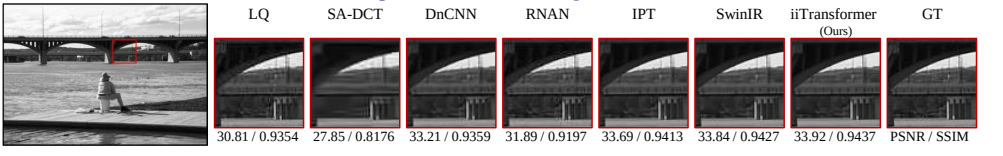
(a) JPEG CAR for $q = 10$ on *paintedhouse* from LIVE1 [ ]



(b) JPEG CAR for $q = 20$ on *carnivaldolls* from LIVE1 [ ]

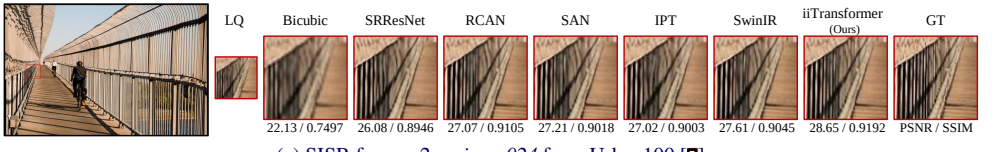


(c) JPEG CAR for $q = 30$ on *churchandcapitol* from LIVE1 [ ]

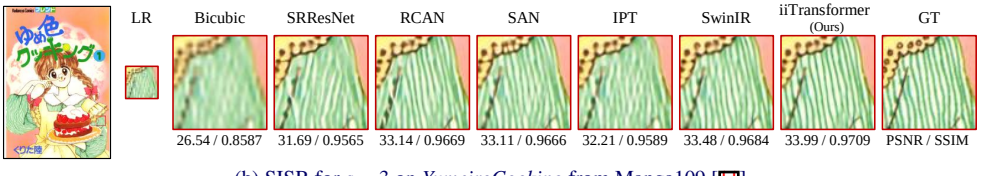


(d) JPEG CAR for $q = 40$ on *manfishing* from LIVE1 [ ]

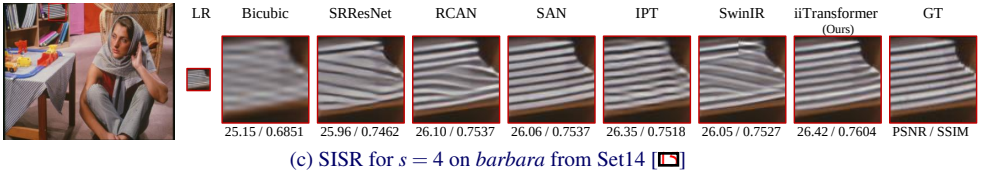
Figure 3: Qualitative comparison of JPEG CAR on various state-of-the-art methods.



(a) SISR for $s = 2$ on *img_024* from Urban100 [ ]



(b) SISR for $s = 3$ on *YumeiroCooking* from Manga109 [ ]



(c) SISR for $s = 4$ on *barbara* from Set14 [ ]

Figure 4: Qualitative comparison of SISR on various state-of-the-art methods.

References

[1] A. Abdelhamed, S. Lin, and M. S. Brown. A High-Quality Denoising Dataset for Smartphone Cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] M. Bevilacqua, A. Roumy, C. Guillemot, and M. Alberi Morel. Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding. In *British Machine Vision Conference (BMVC)*, 2012.

[3] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao. Pre-Trained Image Processing Transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[4] T. Dai, J. Cai, Y. Zhang, S. T. Xia, and L. Zhang. Second-order Attention Network for Single Image Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] A. Foi, V. Katkovnik, and K. Egiazarian. Pointwise Shape-Adaptive DCT for High-Quality Denoising and Deblocking of Grayscale and Color Images. *IEEE Transactions on Image Processing (TIP)*, 2007.

[6] R. Frazen. Kodak Lossless True Color Image Suite, 1999. URL http://r0k.us/graphics/kodak/.

[7] J. Huang, A. Singh, and N. Ahuja. Single Image Super-resolution from Transformed Self-Exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[8] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. SwinIR: Image Restoration using Swin Transformer. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021.

[10] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *IEEE International Conference on Computer Vision (ICCV)*, 2001.

[11] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa. Sketch-based Manga Retrieval using Manga109 Dataset. In *Multimedia Tools and Applications*, 2017.

[12] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. LIVE Image Quality Assessment Database Release 2, 2004. URL https://live.ece.utexas.edu/research/quality/subjective.htm.

[13] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li. MAXIM: Multi-Axis MLP for Image Processing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[15] R. Zeyde, M. Elad, and M. Protter. On Single Image Scale-Up Using Sparse-Representations. In *International Conference on Curves and Surfaces*, 2010.

[16] L. Zhang, X. Wu, A. Buades, and X. Li. Color Demosaicking by Local Directional Interpolation and Non-local Adaptive Thresholding. *Journal of Electronic Imaging*, 2011.

[17] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *European Conference on Computer Vision (ECCV)*, 2018.