Effect of Training and Test Datasets on Image Restoration and Super-Resolution by Deep Learning

Ogun Kirmemis and A. Murat Tekalp
Department of Electrical and Electronics Engineering, Koc University, 34450 Istanbul, Turkey
{okirmemis16,mtekalp}@ku.edu.tr

***Abstract*—Many papers have recently been published on image restoration and single-image super-resolution (SISR) using different deep neural network architectures, training methodology, and datasets. The standard approach for performance evaluation in these papers is to provide a single "average" mean-square error (MSE) and/or structural similarity index (SSIM) value over a test dataset. Since deep learning is data-driven, performance of the proposed methods depends on the size of the training and test sets as well as the variety and complexity of images in them. Furthermore, the performance varies across different images within the same test set. Hence, comparison of different architectures and training methods using a single average performance measure is difficult, especially when they are not using the same training and test sets. We propose new measures to characterize the variety and complexity of images in the training and test sets, and show that our proposed dataset complexity measures correlate well with the mean PSNR and SSIM values obtained on different test data sets. Hence, better characterization of performance of different methods is possible if the mean and variance of the MSE or SSIM over the test set as well as the size, resolution and complexity measures of the training and test sets are specified.**

***Index Terms*—Image restoration, super-resolution, convolutional nets, deep learning, complexity of training and test datasets**

I. Introduction

Various linear, adaptive, and nonlinear filters have been developed for image restoration and single-image super-resolution (SISR) over the years [1]. It is well-known that linear filters are limited by their ability to trade-off noise amplification with regularization artifacts [2]. Adaptive restoration filters can control the amount of ringing artifacts by avoiding filtering across sharp edges (high spatial frequencies). Methods to avoid ringing originating from model-misfit at image boundaries were also discussed. Traditionally, different classical image restoration and SISR methods have been tested on a few standard images, such as Cameraman and Lena, and the mean square error (MSE) or peak-signal-to-noise ratio (PSNR) scores have been reported for evaluation and comparison of methods.

Recently, deep neural networks have proven successful in learning deblurring and super-resolution models in a supervised manner from a large number of example sharp and degraded image pairs. These are non-linear filters, which are not limited in their ability to suppress both noise amplification and regularization artifacts; hence, they produce much better results compared with the traditional filters. Various different deep network architectures and training methods exist, which are briefly reviewed in Section II. However, the performance of deep learning based image restoration and SISR not only depends on the network architecture and training methods, but also on the characteristics of the training and test image datasets, such as the size of the training set, as well as the resolution, variety, and complexity of the images in the training and test sets, which are not well-analyzed in the literature.

The main contributions of this paper are: We propose

- a new measure of image complexity in the frequency domain
- to measure the variety of images in a dataset by the variance of the proposed image complexity measure
- to measure the "difficulty" of the training and test sets by the mean and variance of the proposed image complexity measure over the respective datasets
- to characterize the performance of deep restoration or SISR methods by providing the mean and variance of the MSE and/or SSIM [3] over the test set, as well as providing the mean and variance of the complexity measures of the training and test datasets.

The rest of the paper is organized as follows: Related works are summarized in Section II. Deep network architectures and training methods used in this work are detailed in Section III. The proposed methodology to evaluate the effect of datasets is presented in Section IV. Experimental results are shown in Section V, and conclusions are provided in Section VI.

II. Related Works

Several image restoration and SISR methods using different deep network architectures and training methods exist in the literature. Early work using deep neural networks for SR by Dong et al. [4] used a three-layer network. Kim et al. [5] used residual learning for SISR. An auto-encoder with skip connections was proposed for deblurring and SISR [6], and residual learning with adversarial training for SISR was proposed in [7]. Twenty different SR methods that have competed in the first open challenge (NTIRE 2017) for SR using deep learning are described in [8]. A cascade of two convolutional networks [9] was proposed to solve deblurring and denoising sequentially. For blind space-varying deblurring a conditional adversarial network [10] and DenseNet based adversarial network [11] were proposed.

There is some work on interpretability of deep learning models in the literature [12]. However, no prior work exists on the evaluation of complexity and difficulty of training and test data sets for image restoration and super-resolution.

ISBN 978-90-827970-1-5 © EURASIP 2018

III. Architectures and Training Methods

This section presents the deep network architectures and the training methods that are used in this paper.

A. Architectures

For SISR, we employ the SRResNet architecture [7] in our experiments. The network has a convolutional layer and a nonlinearity at the input stage. Next, there are 16 residual blocks followed by 2 upsampling blocks. The upsampling blocks consist of one convolutional layer and a pixel shuffler layer. Finally, there is also a convolutional layer at the output stage. The input and output of the network are RGB images.

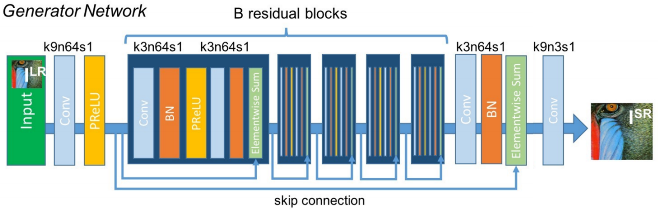

Fig. 1. Deep image restoration network.

For image restoration, we only excluded the upsampling blocks from SRResNet architecture that is used in our SISR experiments. The architecture of this deblurring network, which also has 16 residual blocks, is depicted in Figure 1.

B. Training Methods

We train both networks using the DIV2K training dataset [13]. Non-overlapping 96 × 96 crops are taken from all 800 training images, which yields 235,809 training patches. Degraded image patches are created according to the distortion model:

$$y(n_1, n_2) = D\{h(n_1, n_2) * x(n_1, n_2)\} + \eta(n_1, n_2) \qquad (1)$$

where $x(n_1, n_2)$ is the original patch, $h(n_1, n_2)$ is the known blurring filter, $*$ is the linear convolution operator, D is the down-sampling operator, $\eta(n_1, n_2)$ is white Gaussian noise with variance σ^2 such that the signal-to-noise ratio is 40 dB, and $y(n_1, n_2)$ is the degraded input patch. For the restoration task, D is the identity operator and h is 11×11 box filter. For the super-resolution task, D is the $4\times$ down-sampling operator and h is the 4×4 box filter.

The intensity of the degraded image pixels are scaled to the $(-1, 1)$ interval and random horizontal flip is applied as standard data augmentation technique. We train the networks on the MSE using Adam optimizer [14] with learning rate 10^{-4} as suggested in SRResNet. Mini-batch size is 32 for both networks. We let the training algorithm run for a minimum number of 200 epochs. We stop the training algorithm if no further improvement is observed after 200 epochs.

IV. Effect of Datasets on Performance

This section introduces the proposed measures to quantify image complexity and variety, which are used to evaluate the ”difficulty” of different training and test datasets.

A. Image Complexity Measures

Inspection of restoration and SISR results reveals that the performance of a method varies significantly, both visually and in terms of MSE, between images in a single test set, as well as across different test sets. We have observed that this variation is related to the complexity (or frequency content) of images. That is, if an image has high complexity, i.e. significant high frequency content, then the performance of restoration or SISR degrades. This leads us to propose two image complexity measures to quantify the difficulty of an image for restoration or SISR task. The first measure M_h given by

$$M_h = \frac{1}{N^2} \sum_{k_1} \sum_{k_2} |X[k_1, k_2] F[k_1, k_2]|^2 \qquad (2)$$

quantifies how much power an image has in high frequencies, where $X[k_1, k_2]$ is the discrete Fourier transform (DFT) of the Y channel of the image $x(n_1, n_2)$, $F[k_1, k_2]$ is an ideal high-pass filter with a specified cutoff frequency, and N is the DFT size in each dimension. Since the largest images we work with have 2K resolution, we set the DFT size to 2048 in both dimensions with zero-padding. We arbitrarily set the cut of frequency in both dimensions to 0.5π.

The second measure, M_r, defined by

$$M_r = \frac{M_h}{M_l} \qquad (3)$$

quantifies the ratio of the power in the high frequency band to that in the low frequency band, where M_h is defined in equation 2, and M_l is defined similar to M_h but instead of a high-pass filter, we use a low pass filter $1 - F[k_1, k_2]$.

B. Training Performance

Now that we defined measures to quantify how difficult an image is for image restoration and SISR tasks, it is natural to ask: Is it possible to assign difficulty measures to datasets?

Does the training set have sufficient complexity and variety? We propose to quantify the complexity of a dataset by the mean of M_h and/or the mean of M_r over a dataset. Since neural networks solve image restoration and SISR problems by learning an image manifold, there should also be sufficient variety in the training dataset so that the network learns a manifold that is representative of the test dataset. We propose to quantify the variety in a dataset by the variance of the image complexity measure M_h and/or M_r over a dataset. Hence, a good training set should have a large mean M_h and/or M_r, and a large variance of M_h and/or M_r values to ensure that the network learns a representative image manifold. Thus, we recommend that, while selecting a subset of large datasets such as ImageNet as a training set, the selection should not be completely random but should take the mean and variance of the metrics M_h and/or M_r of selected images into account.

It is important to note that the number of images in the training set is an independent parameter that is related to the depth (hence, the degrees of freedom) of the neural network and should be chosen to avoid overfitting. Although two distinct datasets with different number of images in them may

ISBN 978-90-827970-1-5 © EURASIP 2018

have similar mean and variance values for M_h and/or M_r, the larger dataset still contains more information and variety to prevent model overfitting. Hence, the proposed measures should be used to compare datasets of the same size.

Another important question is: Has the training converged? We propose to consider the variance of the loss function, i.e., the MSE, in addition to the mean MSE over the training set. The variance of MSE stems from the fact that different training crops have different complexity. We have observed that as the mean MSE decreases during iterations, the variance of the MSE also decreases. Hence, the variance of the MSE is a good indicator of convergence of training.

C. Test Performance

A typical test dataset contains tens of images. The performance of restoration or SISR networks over individual images in a test dataset can vary significantly. In the experimental results in Section V, we show that this performance variation depends on the complexity and variety of images in a test dataset. Hence, providing a single mean MSE or PSNR value without considering the complexity and variety (variation of complexity) of images in a test dataset can be quite deceiving.

Similar to difficulty of a training set, we propose to define difficulty of a test dataset in terms of the mean and variance of complexity of images in a test dataset. The higher the mean and variance of the complexity measures M_h and/or M_r, the more difficult is a test dataset; hence, the mean and the variance of the MSE will be higher. The mean complexity measure will determine the mean MSE performance of the network over the test dataset, whereas the variance of the complexity will determine the variance of the MSE. This allows us to compare the performance of a network on various test datasets in a predictable manner. That is, a network performs better on a test set, both visually and in terms of mean MSE, compared to another, if the former test set has lower mean and variance of complexity measure than the latter set.

An important consequence of this observation is that we can now explain why a network sometimes performs better on a test dataset compared to the training dataset. Our results show that this can happen when the test data set has lower complexity (difficulty) compared to the training dataset. This is also an indicator of that we do not overfit.

V. Experimental Results

This section demonstrates that the mean and variance of MSE and SSIM scores obtained for various test datasets correlate well with the proposed dataset difficulty measures.

In our experiments, we use the training dataset of DIV2K (denoted DIV2K train) with 800 images for training our networks. For the test sets, we used the validation set of DIV2K (denoted DIV2K val), GoPro [15], Sun-Hays80 [16], BSD100 [17], Set5 [18], Set14 [19], Urban100 [20], Kodak datasets. We only pick the first and the last frame of every sequence from the GoPro dataset yielding a total of 66 images. The mean and variance of complexity measures M_h and M_r for these datasets are provided in Table I.

Next, we show the performance of the image restoration and SISR networks on different test data sets in Table II and Table III, respectively. In these tables, we provide the mean MSE, PSNR and SSIM scores for both RGB and Y (luminance) only images, where the mean PSNR is computed as $10log_{10}(\frac{1}{MSE})$ such that MSE denotes the mean MSE score over the dataset, since the image intensities are normalized to the range (0,1). We also provided the variances of the MSE and SSIM scores for the Y only images. Note that the mean and variances of MSE and SSIM are computed for intensity-normalized images. Note that the standard deviations of MSE scores are in the same order of magnitude with the MSE scores, and the standard deviations of SSIM scores vary between 0.03 and 0.04, which both indicate significant variation of performance within the same test set. The visual performance of deblurring $11x11$ blur and SR by a factor of 4 for the images with the lowest and highest M_h scores are depicted in Figure 2.

In order to demonstrate that the performance variation between different images is related to the variation of image complexities, we present correlation coefficients between the MSE and SSIM scores of Y channel and the M_h scores with the corresponding p-values in Table IV. We observe that M_h shows stronger correlation compared to M_r on every dataset except Set5, which has only 5 images. P-values are substantially small, which means correlation coefficients are statistically significant for both metrics except Set5 and Set14, which have small number of images. It should be noted here that p-values of M_h are always smaller than those of M_r except for Set5 and Set14. These results demonstrate that both of proposed complexity metrics are statistically meaningful to predict performance of restoration and SISR networks over each test set, which contains images with similar resolutions.

On the other hand, M_r predicts the performance of a network more accurately, when we compare its performance over two different datasets, which contain low and high resolution images, respectively. High resolution images can preserve higher frequencies from the analog domain compared to low resolution images due to the Nyquist sampling theorem. Because of this, although we use the same size DFT, the analog frequencies that correspond to 0.5π are different for low and high resolution images. Since the measure M_r considers the ratio of energy at the high and low frequency bands, it becomes a more suitable measure for comparing performance over test sets containing images with different resolutions. In fact, we observe that the correlation coefficient between mean M_r and mean MSE on the all test sets is 0.96 with p-value 2.42e-05. This observation is also valid for the SISR results. It can be verified from Table II and Table I that the the network performs better on datasets with smaller mean M_r. Mean M_r values also explain why the network performs better on a test set compared to the training set with the only exception of DIV2K Val set. However, if we compare the mean M_h values for the DIV2K Train and DIV2K Val datasets, we see that both the mean and variance of M_h for DIV2K Val are smaller than those of DIV2K Train. Hence, the measure M_h is a better predictor of performance when comparing two

TABLE I
MEAN AND VARIANCE OF COMPLEXITY MEASURES FOR DIFFERENT DATASETS

	Dataset	Resolution	Mean M_h	Var. M_h	Mean M_r	Var. M_r
Training set	DIV2K Train	Full HD	2251.5	5.4738e+06	3.8617e-3	2.975e-05
Test set	DIV2K Val	Full HD	2226.7	4.6755e+06	4.530e-3	5.4387e-05
	GoPro	HD	274.25	8456.5	1.262e-3	1.4337e-07
	Sun-Hays80	HD	520.86	1.3307e+05	2.991e-3	5.4475e-06
	BSD100	SD	226.8	23911	7.189e-3	3.1625e-05
	Set5	SD	86.842	4305.7	3.561e-3	2.1294e-06
	Set14	SD	291.91	53827	5.278e-3	1.3152e-05
	Urban100	SD	459.93	95847	1.010e-2	4.7518e-05
	Kodak	SD	392.86	82041	4.952e-3	1.2454e-05

TABLE II
QUANTITATIVE IMAGE RESTORATION PERFORMANCE RESULTS ON DIFFERENT DATASETS.

		computed on RGB			computed on Y channel				
	Dataset	MSE	PSNR	SSIM	MSE	Var. MSE	PSNR	SSIM	Var. SSIM
Training set	DIV2K Train	8.65e-4	30.63	0.9439	8.09e-4	5.04e-07	30.92	0.9478	8.558e-4
Test set	DIV2K Val	8.30e-4	30.81	0.9437	7.83e-4	4.39e-07	31.06	0.9471	8.184e-4
	GoPro	3.22e-4	34.92	0.9745	3.04e-4	4.81e-08	35.17	0.9763	1.457e-4
	Sun-Hays80	7.73e-4	31.12	0.9420	7.46e-4	3.90e-07	31.27	0.9444	7.712e-4
	BSD100	1.20e-3	29.20	0.9232	1.17e-3	8.10e-07	29.29	0.9256	1.214e-3
	Set5	6.12e-4	32.13	0.9461	5.08e-4	2.70e-08	32.94	0.9568	6.086e-4
	Set14	1.20e-3	29.20	0.9185	1.00e-3	6.06e-07	29.98	0.9315	1.484e-3
	Urban100	1.91e-3	27.19	0.9234	1.77e-3	1.50e-06	27.52	0.9288	9.581e-4
	Kodak	7.76e-4	31.10	0.9399	7.5e-4	3.09e-07	31.25	0.9424	6.444e-4

datasets containing images with the same resolution. Another observation is that the mean M_r of DIV2K Val is lower than that of Kodak dataset, but the network performs better on the Kodak dataset. This can be explained by checking the variance of M_r in these two datasets. Kodak dataset has a lower variance of complexity, which explains the slightly better mean MSE results despite having larger mean complexity measure.

VI. CONCLUSIONS

Experimental results validate the following points:

(i) *Performance variation between individual images within the same test set:* Using a single mean PSNR or mean SSIM value to evaluate results may not be appropriate since the performance of networks for different images within the same test set varies. This variation quantified by the variance of MSE can be as high as 1.5×10^{-3} when image intensity values are scaled between 0 and 1. Note from Tables II and III that the mean MSE is on the order of 10^{-3}. Hence, this variance implies that the standard deviation is in the same order of magnitude as the mean. This variation of performance can be related to the variation of complexity of images in the dataset. We propose image complexity measures and show that there is significant correlation between individual image PSNR values and the complexity metric for images.

(ii) *Performance variation across different test sets:* The mean PSNR performance over a test set is correlated with the mean complexity metric for a test dataset. The PSNR performance across different test sets varies between 27 and 34 dB, which correlates well with the complexity of the respective datasets.

(iii) *Goodness of a training set:* Naturally, we would like to train networks with the most difficult and representative dataset. A good training set should have high mean complexity measure M_h and/or M_r with a high variance, which indicates that we have training samples with sufficient variation of image complexity.

(iv) *Image resolution:* The training set should contain images that has the same or higher spatial resolution as the test set. The capture resolution of images matters rather than their size. Smaller image crops taken from high resolution images carry higher spatial frequency information compared to images with the same size but with smaller resolution.

REFERENCES

[1] A. M. Tekalp, *Digital Video Processing*, Prentice Hall, 2015.

[2] A. M. Tekalp and M. I. Sezan, "Quantitative analysis of artifacts in linear space-invariant image restoration," *Multidimensional Systems and Signal Processing*, vol. 1, no. 2, pp. 143–177, Jun 1990.

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[4] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Patt. Anal. Mach. Intell*, vol. 38, no. 2, pp. 295–307, Feb 2015.

[5] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *IEEE Conf. on Comp. Vision and Patt. Recog. (CVPR)*, Jun 2016, pp. 1646–1654.

[6] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2016.

[7] 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 *eprint arXiv:abs/1609.04802*, May 2017.

[8] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, et al., "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *IEEE Conf. on Comp. Vision and Patt. Recog. (CVPR)*, July 2017.

[9] L. Xu, J. S. Ren, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1790–1798.

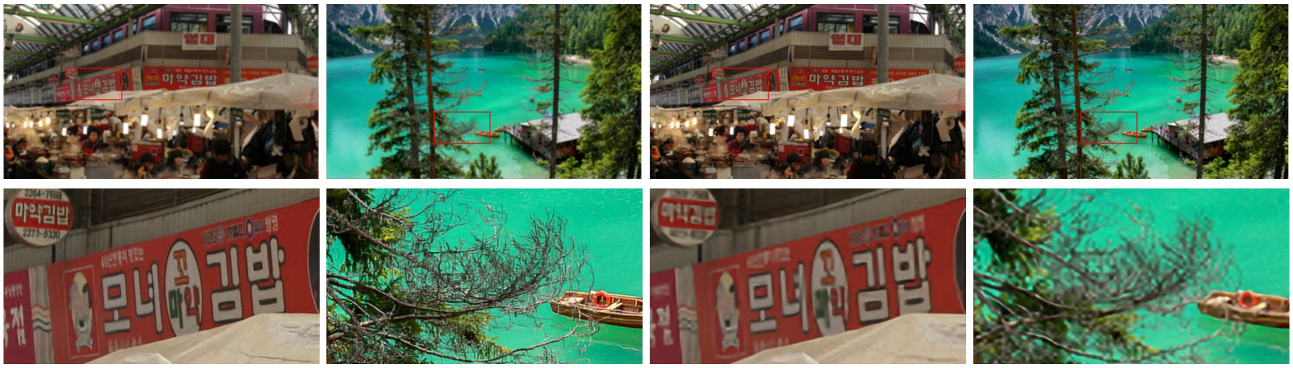

Fig. 2. Images from from DIVK2K Val set with low and high M_h values. Top row from left to right: Image 831 blurred by 11x11 box-filter; Image 807 blurred by 11x11 box-filter; Image 831 linearly interpolated by a factor of 4 after low-pass filtering by 4x4 box-filter and downsampled by 4; Image 807 linearly interpolated by a factor of 4 after low-pass filtering by 4x4 box-filter and downsampled by 4. Bottom row from left to right: The first two images are zoomed restored patches corresponding to the red window from the blurred images on the top; The last two images are the corresponding patches from the output of the SISR network. Note that M_h for Image 831 is 948 and its PSNR for restoration is 35 dB and for SISR is 29.3 dB. The M_h for Image 807 is 8499, and its PSNR for restoration is 24.6 dB and for SISR is 19.8 dB, which demonstrate that M_h values correlate well with PSNR. Observe that both restoration and SISR results for Image 831 do not contain visual artifacts, while there are annoying artifacts in both restoration and SISR results for Image 807 near the tree branches.

TABLE III
QUANTITATIVE SISR PERFORMANCE RESULTS ON DIFFERENT DATASETS.

		computed on RGB			computed on Y channel				
	Dataset	MSE	PSNR	SSIM	MSE	Var. MSE	PSNR	SSIM	Var. SSIM
Training set	DIV2K Train	2.69e-3	25.70	0.8648	2.62e-3	6.88e-06	25.82	0.8688	6.693e-3
Test set	DIV2K Val	2.59e-3	25.86	0.8640	2.54e-3	6.45e-06	25.95	0.8679	6.620e-3
	GoPro	1.02e-3	29.91	0.9366	9.94e-4	5.66e-07	30.02	0.9386	1.238e-3
	Sun-Hays80	2.19e-3	26.60	0.8603	2.15e-3	3.45e-06	26.67	0.8633	5.368e-3
	BSD100	3.79e-3	24.21	0.7956	3.78e-3	8.50e-06	24.23	0.7988	1.057e-2
	Set5	1.72e-3	27.64	0.8993	1.58e-3	1.64e-06	28.01	0.9134	5.164e-3
	Set14	3.52e-3	24.54	0.8173	3.23e-3	4.95e-06	24.91	0.8305	1.021e-2
	Urban100	7.61e-3	21.19	0.7633	7.39e-3	3.86e-05	21.32	0.7678	7.456e-3
	Kodak	2.91e-3	25.37	0.8356	2.89e-3	5.15e-06	25.39	0.8381	6.811e-3

TABLE IV
CORRELATION COEFFICIENTS BETWEEN MSE AND THE PROPOSED COMPLEXITY METRICS WITH THEIR P-VALUES ON DIFFERENT DATASETS FOR BOTH IMAGE RESTORATION AND SISR EXPERIMENTS.

		Image Restoration				SISR			
		M_h Metric		M_r Metric		M_h Metric		M_r Metric	
	Dataset	Corr	p-value	Corr	p-value	Corr	p-value	Corr	p-value
Training set	DIV2K Train	0.86	3.64e-234	0.65	2.24e-97	0.94	0	0.78	2.32e-163
Test set	DIV2K Val	0.82	5.88e-26	0.55	3.09e-09	0.97	3.80e-63	0.82	1.95e-25
	GoPro	0.80	4.76e-16	0.79	1.79e-15	0.84	4.63e-19	0.82	1.53e-17
	Sun-Hays80	0.92	2.56e-34	0.84	4.38e-22	0.94	7.28e-38	0.89	5.53e-29
	BSD100	0.94	1.20e-48	0.75	2.56e-19	0.93	5.77e-45	0.83	6.58e-27
	Set5	0.17	7.84e-4	0.42	4.87e-1	0.04	0.95	0.66	2.22e-1
	Set14	0.60	2.29e-2	0.97	2.18e-08	0.63	1.48e-2	0.80	5.58e-4
	Urban100	0.58	1.97e-10	0.57	4.26e-10	0.83	5.39e-27	0.80	7.74e-24
	Kodak	0.92	1.54e-10	0.88	1.39e-08	0.98	3.72e-18	0.90	3.31e-09

[10] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, “Deblurgan: Blind motion deblurring using conditional adversarial networks,” in *eprint arXiv:abs/1711.07064*, Nov 2017.

[11] S. Ramakrishnan, S. Pachori, A. Gangopadhyay, and S. Raman, “Deep generative filter for motion deblurring,” in *arxiv:abs/1709.03481*, 2017.

[12] Z. C. Lipton, “The mythos of model interpretability,” in *eprint arXiv:abs/1606.03490*, June 2016.

[13] E. Agustsson and R. Timofte, “NTIRE 2017 challenge on single image super-resolution: Dataset and study,” in *IEEE Conf. on Comp. Vision and Patt. Recog. (CVPR)*, July 2017.

[14] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *3rd Int. Conf. on Learning Representations (ICLR)*, May 2015.

[15] S. Nah, T. H. Kim, and K. M. Lee, “Deep multi-scale convolutional neural network for dynamic scene deblurring,” in *IEEE Conf. on Comp. Vision and Patt. Recog. (CVPR)*, July 2017.

[16] L. Sun and J. Hays, “Super-resolution from internet-scale scene matching,” in *IEEE Int. Conf. on Computational Photography (ICCP)*, April 2012, pp. 1–12.

[17] 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 *Proc. IEEE Int. Conf. Comp. Vis.*, 2001, vol. 2, pp. 416–423.

[18] M. Bevilacqua1, A. Roumy, C. Guillemot, and M.-L. A. Morel, “Low-complexity single-image super-resolution based on nonnegative neighbor embedding,” in *Proc. of the British Machine Vision Conf.* 2012, BMVA.

[19] R. Zeyde, M. Elad, and M. Protter, “Single image scale-up using sparse-representations,” in *Curves and Surfaces*. 2012, pp. 711–730, Springer.

[20] J.-B. Huang, A. Singh, and N. Ahuja, “Single image super-resolution from transformed self-exemplars,” in *IEEE Conf. on Comp. Vision and Patt. Recog. (CVPR)*, June 2015.