Range Scaling Global U-Net for Perceptual Image Enhancement on Mobile Devices

Jie Huang[1], Pengfei Zhu[1(✉)], Mingrui Geng[1], Jiewen Ran[1], Xingguang Zhou[1], Chen Xing[1], Pengfei Wan[2], and Xiangyang Ji[2]

[1] MTlab, Meitu Inc., Xiamen, China
zpf2@meitu.com
[2] Tsinghua University, Beijing, China

Abstract. Perceptual image enhancement on mobile devices—smart phones in particular—has drawn increasing industrial efforts and academic interests recently. Compared to digital single-lens reflex (DSLR) cameras, cameras on smart phones typically capture lower-quality images due to various hardware constraints. Without additional information, it is a challenging task to enhance the perceptual quality of a single image especially when the computation has to be done on mobile devices. In this paper we present a novel deep learning based approach—the Range Scaling Global U-Net (RSGUNet)—for perceptual image enhancement on mobile devices. Besides the U-Net structure that exploits image features at different resolutions, proposed RSGUNet learns a global feature vector as well as a novel range scaling layer that alleviate artifacts in the enhanced images. Extensive experiments show that the RSGUNet not only outputs enhanced images with higher subjective and objective quality, but also takes less inference time. Our proposal wins the 1st place by a great margin in track B of the Perceptual Image Enhancement on Smartphones Challenge (PRIM2018). Code is available at https://github.com/MTlab/ECCV-PIRM2018.

Keywords: Perceptual image enhancement · Global feature vector
Range scaling layer

1 Introduction

Nowadays, more and more people prefer taking photos using mobile phones due to the simplicity and portability. However, images taken by mobiles phones

J. Huang, P. Zhu, M. Geng and J. Ran—Equally contributed.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-11021-5_15) contains supplementary material, which is available to authorized users.

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11133, pp. 230–242, 2019.
https://doi.org/10.1007/978-3-030-11021-5_15

typically exhibit lower quality compared to those taken by high end digital single-lens reflex (DSLR) cameras. Besides smart phones, mobile devices like drones, tablets and sport cameras are also capable of taking photos yet suffering the same problem. Therefore there exist real and active needs for improving the perceptual quality of images taken on mobile devices.

Existing image enhancement methods [1,2] improve low-quality images in terms of brightness, color, contrast, details, noise suppression, etc. But few of them address the problem of perceptual image enhancement on mobile devices which casts new challenges in terms of computation and perceptual quality. Recently [2] achieves good perceptual image enhancement results, the slow processing speed and large memory consumption prevent it from being actual deployed on mobile applications.

To overcome the drawbacks of existing methods for perceptual image enhancement on mobile devices, we propose the Range Scaling Global U-Net (RSGUNet). With an efficient U-Net backbone, it exploits image feature maps in various resolutions. Besides, we conjecture that visual artifacts in the enhanced images are largely caused by lacking utilizing of global feature vector, so we introduce global feature vector into our network structure which turns out to greatly improve the enhancement performance. Instead of the traditional residual learning in the literature of deep-learning-based image processing, we propose to learn a range scaling layer that multiplies images rather than adds them. Contributions of this work include:

1. RSGUNet exploits features at different resolutions and achieves good tradeoff between speed and quality;
2. Incorporating global feature vector significantly alleviate the visual artifacts in the enhanced images;
3. Learning range scaling layer instead of residuals performs very well for perceptual image enhancement.

The rest of the paper is organized as follows: Sect. 2 discusses related works on perceptual image enhancement; Sect. 3 presents the network architecture; Sect. 4 demonstrates experimental results; and Sect. 5 concludes the paper.

2 Related Work

Image enhancement has been studied for a long time [2–5]. Existing approaches can be broadly divided into three categories, namely spatial domain methods, frequency domain methods, and hybrid domain methods. Spatial domain methods process pixel values directly, e.g. histogram equalization [6]. Frequency domain methods manipulate components in some transform domain, e.g. wavelet transform [7]. Hybrid domain methods combines spatial domain methods and frequency domain methods. For example, Fan et al. [8] convolved the input image with an optimal Gaussian filter, divided the original histogram into different areas by the valley values, and processed each area separately. Rajavel [9] combined curvelet transform and histogram matching technique to enhance image contrast while preserving image brightness.

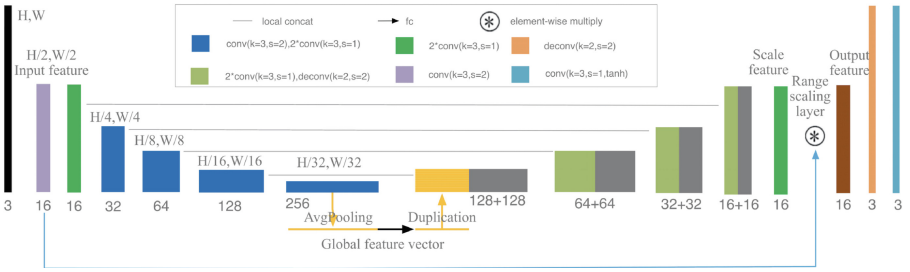


Fig. 1. Network architecture of proposed RSGUNet.

Recently, convolutional neural networks (CNNs) have made great progress in many low-level computer vision tasks, including super-resolution [10–13], deblurring [14], dehazing [4], denoising [15], and image enhancement [16]. Yan et al. [16] proposed a neural network to learn local color transform coefficient between the input and the enhanced images. Enhancenet [17] generated images with more realistic texture by using a perceptual loss. Inspired by bilateral grid processing and local affine color transforms, Gharbi et al. [18] proposed a novel neural network architecture that could process 1080p resolution video in real time on smart phones. Ignatov et al. [1] used a residual CNN to learn the translation function between ordinary photos and DSLR-quality photos, which improved both color rendition and image sharpness. Compared to previous methods, in this paper we propose a new deep learning based approach for better image enhancement performance on mobile devices.

3 Proposed Method

3.1 Network Architecture

Figure 1 illustrates the network architecture of proposed RSGUNet. The backbone is a U-Net [19] that progressively downsamples feature maps at different levels to accelerate the computation. An input RGB image of size $H * W$ is gradually downsampled till $\frac{H}{32} * \frac{W}{32}$ in the first half of the network. In particular, there are two normal convolution layers and four downsample blocks in the first half of the network. Each downsample block consists of one downsample convolution layer and two normal convolution layers. Afterwards, the *global feature vector* of size $256 * 1 * 1$ is extracted through average pooling on the $256 * \frac{H}{32} * \frac{W}{32}$ tensor. The global feature vector encodes the global characteristics of the input image, which proves to be important for perceptual image enhancement in our experiments.

In the second half of the network, the global feature vector is first mapped to size $128 * 1 * 1$ by a fully connected layer. After duplicating each element $\frac{H}{16} * \frac{W}{16}$ times, we obtain a $128 * \frac{H}{16} * \frac{W}{16}$ tensor which is further concatenated with the tensor of the same size in the first half of the network (symmetric skip connection with concatenation). After three upsample blocks with skip connections, we

arrive at the scale feature map that shares identical size with the input feature map. In the proposed *range scaling layer*, the scale feature map and input feature map are elementwise-multiplied to yield the output feature map. Finally, the network outputs the enhanced image of size $H * W$ after a deconvolution layer and another convolution layer.

Learning global feature vector and range scaling layer significantly alleviates the visual artifacts in enhanced images according to our experiments. Global feature vector serves as a regularizer to penalize any mishandling in low resolution features that could potentially lead to artifacts [18]. What's more, using average pooling to extract global feature vector requires much less parameters compared to fully connected layer as in [2]. Besides global featuer vector, the range scaling layer enables per-pixel scaling of pixel intensities. Due to the fact that a collection of simple local transformations suffices to approximate any complex image processing pipelines [20], proposed RSGUNet has much higher capacity than traditional residual-learning networks to learn the subtle and complex mappings from low-quality images to high-quality ones.

3.2 Loss Functions

Besides network architecture, loss function plays another key role in network design. In our experiments, we find that a combination of L_1 loss, MS-SSIM loss [21], VGG loss [22], GAN loss [3], and total variation loss [23] leads to the best performance of RSGUNet.

$$L = \rho_1 * L_1 + \rho_2 * L_{\text{MS-SSIM}} + \rho_3 * L_{\text{VGG}} + \rho_4 * L_{\text{GAN}} + \rho_5 * L_{\text{TV}}, \quad (1)$$

where ρ_1, ρ_2, ρ_3, ρ_4, and ρ_5 are tunable hyper-parameters.

L_1 + MS-SSIM loss has been shown to outperform L_2 loss in image reconstruction [21]. Advantage of L_1 loss is its ability to retain more image color and brightness information. Advantage of MS-SSIM loss is its ability to preserve more high frequency information. They are defined as follows:

$$L_1 = \left\| I_t - F_w(I_s) \right\|_1, \quad (2)$$

$$L_{\text{MS-SSIM}} = 1 - \text{MS-SSIM}(I_t, F_w(I_s)), \quad (3)$$

where I_t denotes the target image, I_s denotes the source image, and $F_w(I_s)$ denotes the enhanced image, respectively.

VGG loss encourages similar feature representations between the enhanced image and the target image. It is calculated on multiple layers of the pre-trained VGG network as follows:

$$L_{\text{VGG}} = \sum_{j=1,3,5} \frac{1}{C_j H_j W_j} \left\| \phi_j(I_t) - \phi_j(F_w(I_s)) \right\|_2^2, \quad (4)$$

where ϕ_j denotes feature map at the jth convolution layer of VGG-19. Scalars C_j, H_j, and W_j denote number of channels, height, and width of the corresponding layer, respectively.

Generative adversarial network (GAN) loss can approximate the perceptive distance between two images [24]. Therefore, minimizing the GAN loss leads to improved perceptual quality of the enhanced image. Our discriminator network D is pre-trained, so the GAN loss is defined on the generator F_w as follows:

$$L_{\text{GAN}} = -\sum \log D(I_t, F_w(I_s)). \tag{5}$$

Total variation (TV) loss is effective in suppressing high frequency noise [23], which is defined as follows:

$$L_{\text{TV}} = \frac{1}{CHW}(\left\|\nabla_x F_w(I_s)\right\|_2^2 + \left\|\nabla_y F_w(I_s)\right\|_2^2), \tag{6}$$

where C, H, and W denote number of channels, height, and weight of the enhanced image, respectively.

4 Experimental Results

4.1 Experiment Settings

We use the DPED [1] dataset to train our model. In the dataset, four photos are taken for each scene, including three photos by three different mobile phones and the fourth one by a DSLR camera. In our experiments, only the photos taken by the iPhone®3GS and the DSLR camera (Canon®EOS 70D) are included for training and validation. Photos taken by iPhone serve as the input, while the corresponding ones taken by DSLR serve as the ground-truth. Since it is difficult to align photos in full size, all the images provided in DPED dataset were cut into patches of size $100 * 100$ and then aligned. In total, 160000 training patches and 43000 validation patches are used in our experiment. To faithfully evaluate the objective and subjective performance, we use the 400 images provided by the PIRM2018 Challenge as test images. For objective evaluation, we use PSNR, SSIM [25] and inference time as metrics; for subjective evaluation, we use the full-size images (instead of the patches) as input to compare the enhanced output against the DSLR ground-truth.

We implement the proposed network using Tensorflow[1] 1.1.0. The network is trained on one single NVIDIA®GTX1080Ti GPU for 150000 iterations with batch size 32. Adam optimizer is used and the learning rate is set to $5e^{-4}$ without decay. Hyper-parameters ρ in the loss function (1) are set to 0.05 for L_1 loss, 500 for MS-SSIM loss, 0.001 for VGG loss, 10 for GAN loss, and 2000 for TV loss, respectively. The hyper-parameter values are determined such that all losses are of the same order of magnitude when multiplied with the corresponding ρ. The trained model are evaluated using Tensorflow 1.8.0 on a single NVIDIA®GTX1060 GPU as required by the PIRM2018 Challenge.

[1] https://www.tensorflow.org/.

4.2 Ablation Study

We conduct the following ablation experiments to demonstrate the effectiveness of different components of the proposed network.

Analysis of the Architecture. As described in Sect. 3, the RSGUNet improves the original U-Net with two major modifications: learning global feature (GF) vector and the range scaling (RS) layer. As shown in Table 1, either GF or RS leads to increased PSNR and SSIM value with negligible inference time increase; and combining them further improves the objective performance.

Table 1. Objective performance of different network architectures.

	U-Net	U-Net+RS	U-Net+GF	RSGUNet(U-Net+RS+GF)
PSNR (dB)	22.74	22.95	22.96	**23.01**
SSIM	0.9293	0.9309	0.9307	**0.9312**
Inference time (ms)	**486**	493	490	508

Besides the superior objective performance, RS and GF also significantly improve the subjective performance. As shown in Fig. 2, colors in enhanced image are more evenly distributed after adding GF, and RS contribute to natural brightness of the enhanced image.

Analysis of Loss Functions. We test different combinations of loss functions and the objective results are summarized in Table 2 and Fig. 3. The loss strategy of the DPED paper [1] is included as the baseline, which combines L_2 loss, vanilla VGG loss, GAN loss, and TV loss with parameters 0.5, 10, 1, and 2000, respectively. To study the effect of different losses on the enhancement performance, we train models using the following loss strategies respectively: 1. the baseline DPED loss (Loss-B); 2. replacing L_2 loss in Loss-B with L_1+MS-SSIM losses (Loss-L); 3. replacing the vanilla VGG in Loss-B with our proposed VGG loss (Loss-V); 4. replacing L_2 loss in Loss-V with L_1+MS-SSIM losses (Loss-P).

Table 2. Objective performance of RSGUNet trained using different loss strategies.

	Loss-B	Loss-L	Loss-V	Loss-P (proposed)
PSNR (dB)	22.74	22.85	22.68	**23.01**
SSIM	0.9196	0.9290	0.9261	**0.9312**

(a) Input

(b) U-Net

(c) U-Net+RS

(d) U-Net+GF

(e) RSGUNet

Fig. 2. Enhanced images by different network architectures, taking (a) as input. (Color figure online)

We see from Table 2 that Loss-L greatly increases the PSNR and SSIM values. Loss-V increases the SSIM value but not the PSNR value. In terms of subjective quality, Loss-L tends to make the resulted images a little darker as shown in Fig. 3(c), while Loss-V tends to result in brighter images as shown in Fig. 3(d). The Loss-P leads to the best PSNR and SSIM values as well as good visual quality of enhanced image, see Fig. 3(e).

(a) Input

(b) Loss-B

(c) Loss-L

(d) Loss-V

(e) Loss-P

Fig. 3. Enhanced images by RSGUNet trained with different loss strategies, taking (a) as input.

4.3 Comparison with the State-of-the-Art Methods

We compare our method with several state-of-the-art methods including SRCNN [10], DPED [1], and EDSR [12]. As shown in Table 3, proposed RSGUNet outperforms competing methods in all three objective metrics (PSNR, SSIM and inference time). In other words, proposed RSGUNet achieves better enhancement quality while being much faster. Besides the good objective performance,

RSGUNet also has outstanding subjective performance. As shown in Fig. 4, enhanced image by RSGUNet exhibits the least visual artifacts.

Though RSGUNet performs very well on most images, there exist few cases where the enhanced images look kind of darker or blurred than those of competing methods (Fig. 5). The most probable reason is the downsampling operations of U-Net. Nevertheless, we did not observe severe artifacts in images enhanced by RSGUNet in our experiments.

Table 3. Objective performance of competing image enhancement methods. m-n means the model has m convolution layers with each layer having n channels.

	SRCNN	DPED(8-32)	DPED(12-64)	EDSR	RSGUNet
PSNR (dB)	21.33	22.40	22.19	21.18	**23.01**
SSIM	0.9040	0.9166	0.9204	0.9067	**0.9312**
Inference time (ms)	2305	4357	14682	16141	**508**

Table 4. Results of the PIRM2018 Challenge (track B: image enhancement).

	PSNR (dB)	SSIM	MOS	CPU (ms)	GPU (ms)	Razer Phone (ms)	Huawei P20 (ms)	RAM (GB)
Mt.Phoenix	21.99	**0.9125**	**2.6804**	**682**	**64**	1472	2187	1.4
2nd	21.65	0.9048	2.6523	3241	253	5153	Out of memory	2.3
3rd	21.99	0.9079	2.6283	1620	111	1802	2321	1.6
4th	**22.22**	0.9086	2.6108	1461	138	2279	3459	1.8
5th	21.85	0.9067	2.5583	828	111	-	-	1.6
6th	21.56	0.8948	2.5123	2153	181	3200	4701	2.3
7th	22.03	0.9042	2.465	1448	81	1987	3061	1.6

4.4 Results of the PIRM2018 Challenge

We participated the track B (image enhancement) of the Perceptual Image Enhancement on Smartphones (PRIM2018) Challenge. Results of the top-8 teams are presented in Table 4. The proposed RSGUNet (team Mt.Phoenix) ranks first under almost all metrics and wins the championship by a great margin. Please find details of the competition on http://ai-benchmark.com/challenge.html. The mean opinion score (MOS) is a commonly-used metric of subjective performance, which indicates the perceived quality of the enhanced images.

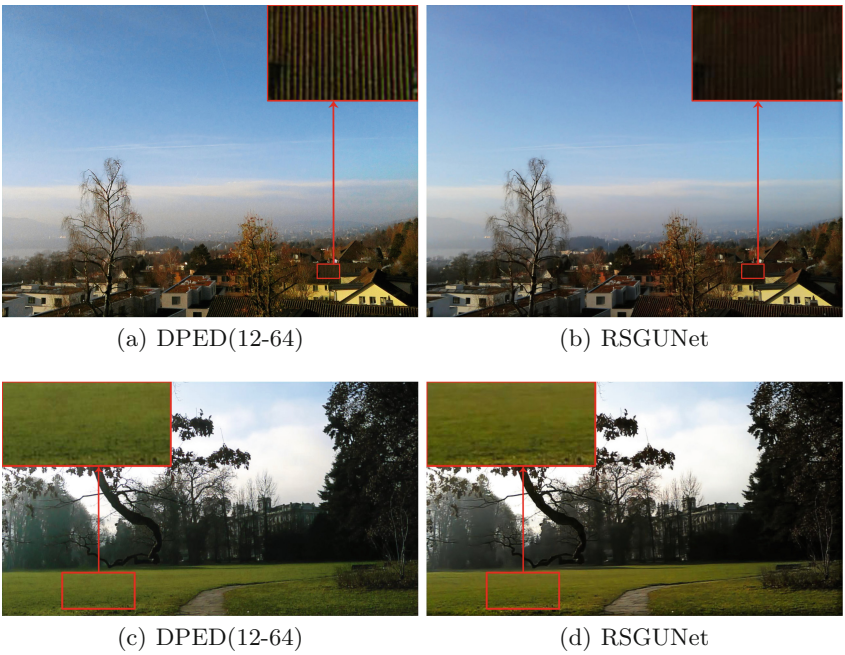
(a) Input (b) SRCNN

(c) DPED(8-32) (d) DPED(12-64)

(e) EDSR (f) RSGUNet

Fig. 4. Enhanced images by different methods using (a) as input. m-n means the model has m convolution layers with each layer having n channels.

We also experimented on super-resolution task using the proposed RSGUNet architecture but the performance was not as good. That is because super-resolution and enhancement are two tasks different in nature. For example, in enhancement global information is important for adjusting the overall appearance, while in super-resolution the interpolation depends heavily on local gradients.



(a) DPED(12-64) (b) RSGUNet

(c) DPED(12-64) (d) RSGUNet

Fig. 5. Some failure cases of RSGUNet, as compared to DPED (12 layers, 64 channels per layer).

5 Conclusion

We proposed the RSGUNet, a novel CNN-based approach for perceptual image enhancement. The outstanding objective and subjective enhancement performance as well as the low computational complexity make RSGUNet very suitable for perceptual image enhancement on mobile devices. In the future, we would like to investigate new network structures for real-time image enhancement.

References

1. Ignatov, A., Kobyshev, N., Timofte, R., Vanhoey, K., Van Gool, L.: DSLR-quality photos on mobile devices with deep convolutional networks. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
2. Chen, Y.S., Wang, Y.C., Kao, M.H., Chuang, Y.Y.: Deep photo enhancer: unpaired learning for image enhancement from photographs with GANS. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6306–6314 (2018)
3. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017)

4. Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., Yang, M.-H.: Single image dehazing via multi-scale convolutional neural networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 154–169. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_10
5. Ignatov, A., Timofte, R., et al.: PIRM challenge on perceptual image enhancement on smartphones: report. In: European Conference on Computer Vision Workshops (2018)
6. Divya, K., Roshna, K.: A survey on various image enhancement algorithms for naturalness preservation. Int. J. Comput. Sci. Inf. Technol. **6**(3), 2043–2045 (2015)
7. Bedi, S., Khandelwal, R.: Various image enhancement techniques-a critical review. Int. J. Adv. Res. Comput. Commun. Eng. **2**(3), 267–274 (2013)
8. Yang, F., Wu, J.: An improved image contrast enhancement in multiple-peak images based on histogram equalization. In: 2010 International Conference on Computer Design and Applications (ICCDA), vol. 1, pp. V1–346. IEEE (2010)
9. Rajavel, P.: Image dependent brightness preserving histogram equalization. IEEE Trans. Consum. Electron. **56**(2), 756–763 (2010)
10. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 184–199. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_13
11. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654 (2016)
12. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2017)
13. Kligvasser, I., Shaham, T.R., Michaeli, T.: xunit: learning a spatial activation function for efficient image restoration. In: CVPR (2018)
14. Noroozi, M., Chandramouli, P., Favaro, P.: Motion deblurring in the wild. In: Roth, V., Vetter, T. (eds.) GCPR 2017. LNCS, vol. 10496, pp. 65–77. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66709-6_6
15. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: residual learning of deep cnn for image denoising. IEEE Trans. Image Process. **26**(7), 3142–3155 (2017)
16. Yan, Z., Zhang, H., Wang, B., Paris, S., Yu, Y.: Automatic photo adjustment using deep neural networks. ACM Trans. Graph. (TOG) **35**(2), 11 (2016)
17. Sajjadi, M.S., Schölkopf, B., Hirsch, M.: Enhancenet: single image super-resolution through automated texture synthesis. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 4501–4510. IEEE (2017)
18. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. ACM Trans. Graph. (TOG) **36**(4), 118 (2017)
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Chen, J., Adams, A., Wadhwa, N., Hasinoff, S.W.: Bilateral guided upsampling. ACM Trans. Graph. (TOG) **35**(6), 203 (2016)
21. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Trans. Comput. Imaging **3**(1), 47–57 (2017)

22. Ustyuzhaninov, I., Brendel, W., Gatys, L., Bethge, M.: What does it take to generate natural textures? In: International Conference on Learning Representations (2017)
23. Aly, H.A., Dubois, E.: Image up-sampling using total-variation regularization with a new observation model. IEEE Trans. Image Process. **14**(10), 1647–1659 (2005)
24. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: CVPR (2018)
25. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)