The Perception-Distortion Tradeoff
Supplementary Material

Yochai Blau and Tomer Michaeli
Technion–Israel Institute of Technology, Haifa, Israel
{yochai@campus,tomer.m@ee}.technion.ac.il

In this supplemental, we first provide a proof for Theorem 1, the derivation of Example 1, and the derivation of the MMSE and MAP estimators which appear in Sections 3.1 and 3.2. We then briefly discuss the real-vs.-fake study setting and its relation to Bayesian hypothesis testing. In the following sections, we specify all training and architecture details for the WGAN experiment which was presented in Section 5, and also include details regarding the super-resolution algorithms comparison in Section 6. Finally, we include additional comparisons between super-resolution algorithms using two extra no-reference methods, and perform a comparison on RGB images as well.

I. Proof of Theorem 1

The proof of Theorem 1 follows closely that of the rate-distortion theorem from information theory [1]. The value $P(D)$ is the minimal distance $d(p_X, p_{\hat{X}})$ over a constraint set whose size increases with D. This implies that the function $P(D)$ is non-increasing in D. Now, to prove the convexity of $P(D)$, we will show that

$$\lambda P(D_1) + (1-\lambda)P(D_2) \geq P(\lambda D_1 + (1-\lambda)D_2), \tag{S1}$$

for all $\lambda \in [0,1]$. First, by definition, the left hand side of (S1) can be written as

$$\lambda d(p_X, p_{\hat{X}_1}) + (1-\lambda) d(p_X, p_{\hat{X}_2}), \tag{S2}$$

where $\hat{X}_1$ and $\hat{X}_2$ are the estimators defined by

$$p_{\hat{X}_1|Y} = \underset{p_{\hat{X}|Y}}{\arg\min}\, d(p_X, p_{\hat{X}}) \;\text{ s.t. }\; \mathbb{E}\left[\Delta(X,\hat{X})\right] \leq D_1, \tag{S3}$$

$$p_{\hat{X}_2|Y} = \underset{p_{\hat{X}|Y}}{\arg\min}\, d(p_X, p_{\hat{X}}) \;\text{ s.t. }\; \mathbb{E}\left[\Delta(X,\hat{X})\right] \leq D_2. \tag{S4}$$

Since $d(\cdot,\cdot)$ is convex in its second argument,

$$\lambda d(p_X, p_{\hat{X}_1}) + (1-\lambda) d(p_X, p_{\hat{X}_2}) \geq d(p_X, p_{\hat{X}_\lambda}), \tag{S5}$$

where $\hat{X}_\lambda$ is defined by

$$p_{\hat{X}_\lambda|Y} = \lambda p_{\hat{X}_1|Y} + (1-\lambda)\, p_{\hat{X}_2|Y}. \tag{S6}$$

Denoting $D_\lambda = \mathbb{E}[\Delta(X, \hat{X}_\lambda)]$, we have that

$$d(p_X, p_{\hat{X}_\lambda}) \geq \min_{p_{\hat{X}|Y}} \left\{ d(p_X, p_{\hat{X}}) \,:\, \mathbb{E}[\Delta(X,\hat{X})] \leq D_\lambda \right\} = P(D_\lambda), \tag{S7}$$

because $\hat{X}_\lambda$ is in the constraint set. Below, we show that

$$D_\lambda \leq \lambda D_1 + (1-\lambda) D_2. \tag{S8}$$

Therefore, since $P(D)$ is non-increasing in D, we have that

$$P(D_\lambda) \geq P(\lambda D_1 + (1-\lambda)D_2). \tag{S9}$$

Combining (S2),(S5),(S7) and (S9) proves (S1), thus demonstrating that $P(D)$ is convex.

To justify (S8), note that

$$\begin{aligned} D_\lambda &= \mathbb{E}\left[\Delta(X,\hat{X}_\lambda)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\Delta(X,\hat{X}_\lambda)|Y\right]\right] \\ &= \mathbb{E}\left[\lambda\mathbb{E}\left[\Delta(X,\hat{X}_1)|Y\right] + (1-\lambda)\mathbb{E}\left[\Delta(X,\hat{X}_2)|Y\right]\right] \\ &= \lambda\mathbb{E}\left[\Delta(X,\hat{X}_1)\right] + (1-\lambda)\,\mathbb{E}\left[\Delta(X,\hat{X}_2)\right] \\ &\leq \lambda D_1 + (1-\lambda)D_2, \end{aligned} \tag{S10}$$

where the second and fourth transitions are according to the law of total expectation and the third transition is justified by

$$\begin{aligned} p(x,\hat{x}_\lambda|y) &= p(\hat{x}_\lambda|x,y)p(x|y) = p(\hat{x}_\lambda|y)p(x|y) = (\lambda p(\hat{x}_1|y) + (1-\lambda)p(\hat{x}_2|y))p(x|y) \\ &= \lambda p(\hat{x}_1|y)p(x|y) + (1-\lambda)p(\hat{x}_2|y))p(x|y) = \lambda p(x,\hat{x}_1|y) + (1-\lambda)p(x,\hat{x}_2|y)). \end{aligned} \tag{S11}$$

Here we used (S6) and the fact that given Y, X is independent of $\hat{X}_\lambda$, $\hat{X}_1$, and $\hat{X}_2$.

II. Derivation of Example 1

Since $\hat{X} = aY = a(X+N)$, it is a zero-mean Gaussian random variable. Now, the Kullback-Leibler distance between two zero-mean normal distributions is given by

$$d_{\text{KL}}(p_X\|p_{\hat{X}}) = \ln\left(\frac{\sigma_{\hat{X}}}{\sigma_X}\right) + \frac{\sigma_X^2}{2\sigma_{\hat{X}}^2} - \frac{1}{2}, \tag{S12}$$

and the MSE between X and $\hat{X}$ is given by

$$\text{MSE}(X,\hat{X}) = E[(X-\hat{X})^2] = \sigma_X^2 - 2\sigma_{X\hat{X}} + \sigma_{\hat{X}}^2. \tag{S13}$$

Substituting $\hat{X} = aY$ and $\sigma_X^2 = 1$, we obtain that $\sigma_{\hat{X}} = |a|\sqrt{1+\sigma_N^2}$ and $\sigma_{X\hat{X}} = a$, so that

$$d_{\text{KL}}(a) = \ln\left(|a|\sqrt{1+\sigma_N^2}\right) + \frac{1}{2a^2(1+\sigma_N^2)} - \frac{1}{2}, \tag{S14}$$

$$\text{MSE}(a) = 1 + a^2(1+\sigma_N^2) - 2a, \tag{S15}$$

and

$$P(D) = \min_a d_{\text{KL}}(a) \quad \text{s.t.} \quad \text{MSE}(a) \leq D. \tag{S16}$$

Notice that d_{KL} is symmetric, and $\text{MSE}(|a|) \leq \text{MSE}(a)$ (see Fig. S1). Thus, for any negative a, there always exists a positive a with which d_{KL} is the same and the MSE is not larger. Therefore, without loss of generality, we focus on the range $a \geq 0$.

For $D < D_{\min} = \frac{\sigma_N^2}{1+\sigma_N^2}$ the constraint set of $\text{MSE}(a) < D$ is empty, and there is no solution to (S16). For $D \geq D_{\min}$, the constraint is satisfied for $a_- \leq a \leq a_+$, where

$$a_\pm(D) = \frac{1}{(1+\sigma_N^2)}\left(1 \pm \sqrt{D(1+\sigma_N^2) - \sigma_N^2}\right). \tag{S17}$$

For $D = D_{\min}$, the optimal (and only possible) a is

$$a = a_+(D_{\min}) = a_-(D_{\min}) = \frac{1}{(1+\sigma_N^2)}. \tag{S18}$$

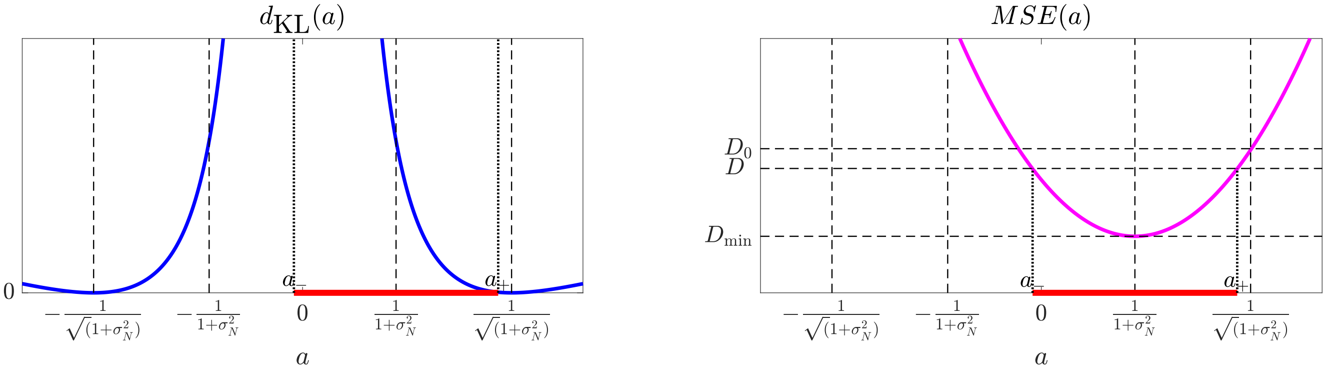

Figure S1. Plots of (S14) and (S15). D defines the range (a_-, a_+) of a values complying with the MSE constraint (marked in red). The objective d_{KL} is minimized over this range of possible a values.

For $D > D_{\min}$, a_+ monotonically increases with D, broadening the constraint set. The objective $d_{\mathrm{KL}}(a)$ monotonically decreases with a in the range $a \in (0, 1/\sqrt{(1+\sigma_N^2)})$ (see Fig. S1 and the mathematical justification below). Thus, for $D_{\min} < D \leq D_0$, the optimal a is always the largest possible a, which is $a = a_+(D)$, where D_0 is defined by $a_+(D_0) = 1/\sqrt{(1+\sigma_N^2)}$ (see Fig. S1). For $D > D_0$, the optimal a is $a = 1/\sqrt{(1+\sigma_N^2)}$, which achieves the global minimum $d_{\mathrm{KL}}(a) = 0$. The closed form solution is therefore given by

$$P(D) = \begin{cases} d_{\mathrm{KL}}(a_+(D)) & D_{\min} \leq D < D_0 \\ 0 & D_0 \leq D \end{cases} \tag{S19}$$

To justify the monotonicity of $d_{\mathrm{KL}}(a)$ in the range $a \in (0, 1/\sqrt{(1+\sigma_N^2)})$, notice that for $a > 0$,

$$\frac{d}{da} d_{\mathrm{KL}}(a) = \frac{1}{a} - \frac{1}{(1+\sigma_N^2)} \frac{1}{a^3}, \tag{S20}$$

which is negative for $a \in (0, 1/\sqrt{(1+\sigma_N^2)})$.

III. Derivation of MMSE and MAP estimators (Sec. 3.1,3.2)

In these sections, X which is a 280×280 binary image is denoised from its noisy counterpart $Y = X + N$, where $N \sim \mathcal{N}(0, \sigma^2 I)$ is independent from X. Thus, the conditional distribution of Y given X is $p(y|X = x) \sim \mathcal{N}(x, \sigma^2 I)$. The MMSE estimator is given by posterior-mean

$$\hat{x}_{\text{MMSE}}(y) = \mathbb{E}[X|Y = y] = \sum_x x p(x|y) = \sum_x x \frac{p(y|x)p(x)}{\sum_{x'} p(y|x')p(x')} = \sum_x x \frac{\exp(-\frac{1}{2\sigma^2}\|y - x\|^2)p(x)}{\sum_{x'} \exp(-\frac{1}{2\sigma^2}\|y - x'\|^2)p(x')}, \tag{S21}$$

where $p(x) = 1/59400$ for non-blank images and $p(x) = 1/11$ for the blank image. The MAP estimator is given by

$$\hat{x}_{\text{MAP}}(y) = \arg\max_x p(x|y) = \arg\min_x -\log(p(y|x)p(x)) = \arg\min_x \frac{1}{2\sigma^2}\|y - x\|^2 - \log(p(x)). \tag{S22}$$

Notice that since the noise N is i.i.d., subparts of y can be denoised separately. Specifically, denoising the whole 280×280 image is equivalent to denoising each sub-image containing one MNIST digit separately.

The MMSE and MAP estimators for the simple example of the discrete distribution in (4), are reported in Sec. 3.1,3.2. We calculate the distribution of the MMSE estimator in this simple example (Fig. 3) by

$$p_{\hat{X}_{\text{MMSE}}}(\hat{x}) = p_Y(\hat{x}^{-1}_{\text{MMSE}}(\hat{x})) \left| \frac{d}{d\hat{x}} \hat{x}^{-1}_{\text{MMSE}}(\hat{x}) \right| \tag{S23}$$

where the inverse of the MMSE estimator $\hat{x}_{\text{MMSE}}(y)$ (see (5)) and its derivative are calculated numerically, and $p_Y(y) = \sum_x p(y|x)p(x)$ with $p(y|x) \sim \mathcal{N}(x, 1)$ and $p(x)$ is given by (4).

Table S1. Generator and discriminator architecture. FC is a fully-connected layer, BN is a batch-norm layer, and l-ReLU is a leaky-ReLU layer.

Discriminator	
Size	Layer
$28 \times 28 \times 1$	Input
$12 \times 12 \times 32$	Conv (stride=2), l-ReLU (slope=0.2)
$4 \times 4 \times 64$	Conv (stride=2), l-ReLU (slope=0.2)
1024	Flatten
1	FC
1	Output

Generator	
Size	Layer
$28 \times 28 \times 1$	Input
784	Flatten
$4 \times 4 \times 128$	FC, unflatten, BN, ReLU
$7 \times 7 \times 64$	transposed-Conv (stride=2), BN, ReLU
$14 \times 14 \times 32$	transposed-Conv (stride=2), BN, ReLU
$28 \times 28 \times 1$	transposed-Conv (stride=2), sigmoid
$28 \times 28 \times 1$	Output

IV. Real-vs.-fake study setting

We assume the setting where an observer is shown a real image (realization of p_X) or an algorithm output (realization of $p_{\hat{X}}$), with a prior probability of 0.5 each. The task is to identify which distribution the image was drawn from (p_X or $p_{\hat{X}}$) with a maximal success rate. This is the setting of the Bayesian hypothesis testing problem, for which the maximum a-posteriori (MAP) decision rule minimizes the probability of error (see Section 1 in [7]). When there are two possible hypotheses with equal priors (as is our setting), the relation between the probability of error and the total-variation distance between p_X and $p_{\hat{X}}$ in (1) can be easily derived (see Section 2 in [7]).

V. WGAN architecture and training details (Sec. 5)

The architecture of the WGAN trained for denoising the MNIST images is detailed in Table S1. The training algorithm and adversarial losses are as proposed in [2]. The generator loss was modified to include a content loss term, *i.e.* $\ell_{\text{gen}} = \ell_{\text{MSE}} + \lambda\,\ell_{\text{adv}}$, where ℓ_{MSE} is the standard MSE loss. For each λ the WGAN was trained for 35 epochs, with a batch size of 64 images. The ADAM optimizer [3] was used, with $\beta_1 = 0.5, \beta_2 = 0.9$. The generator/discriminator initial learning rate is $10^{-3}/10^{-4}$ respectively, where learning rate of both decreases by half every 10 epochs. The filter size of the discriminator convolutional layers is 5×5, and these are performed without padding. The filter size in the generator transposed-convolutional layers is $5 \times 5/4 \times 4$, and these are performed with $2/1$ pixel padding for the first/second and third transposed-convolutional layers, respectively. The stride of each convolutional layer and the slope for the leaky-ReLU layers appear in Table S1. Note that the perception-distortion curve in Fig. 6 is generated by training on single digit images, which in general may deviate from the perception-distortion curve of whole images containing i.i.d. sub-blocks of digits.

VI. Super-resolution evaluation details and additional comparisons (Sec. 6)

The no-reference (NR) and full-reference (FR) methods BRISQUE, BLIINDS-II, NIQE, SSIM, MS-SSIM, IFC and VIF were obtained from the LIVE laboratory website[1], the NR method of Ma *et al.* was obtained from the project webpage[2], and the pretrained VGG-19 network was obtained through the PyTorch torchvision package[3]. The low-resolution images were obtained by factor 4 downsampling with a bicubic kernel. The super-resolution results on the BSD100 dataset of the SRGAN and SRResNet variants were obtained online[4], and the results of EDSR, Deng, Johnson *et al.* and Mechrez *et al.* were kindly provided by the authors. The algorithms for testing the other SR methods were obtained online: A+[5], SRCNN[6], SelfEx[7], VDSR[8], LapSRN[9], Bae *et al.*[10] and ENet[11]. All NR and FR metrics and all SR algorithms were used with the default

[1] http://live.ece.utexas.edu/research/Quality/index.htm

[2] https://github.com/chaoma99/sr-metric

[3] http://pytorch.org/docs/master/torchvision/index.html

[4] https://twitter.box.com/s/lcue6vlrd01ljkdtdkhmfvk7vtjhetog

[5] http://www.vision.ee.ethz.ch/~timofter/ACCV2014_ID820_SUPPLEMENTARY/

[6] http://mmlab.ie.cuhk.edu.hk/projects/SRCNN.html

[7] https://github.com/jbhuang0604/SelfExSR

[8] http://cv.snu.ac.kr/research/VDSR/

[9] https://github.com/phoenix104104/LapSRN

[10] https://github.com/iorism/CNN

[11] https://webdav.tue.mpg.de/pixel/enhancenet/

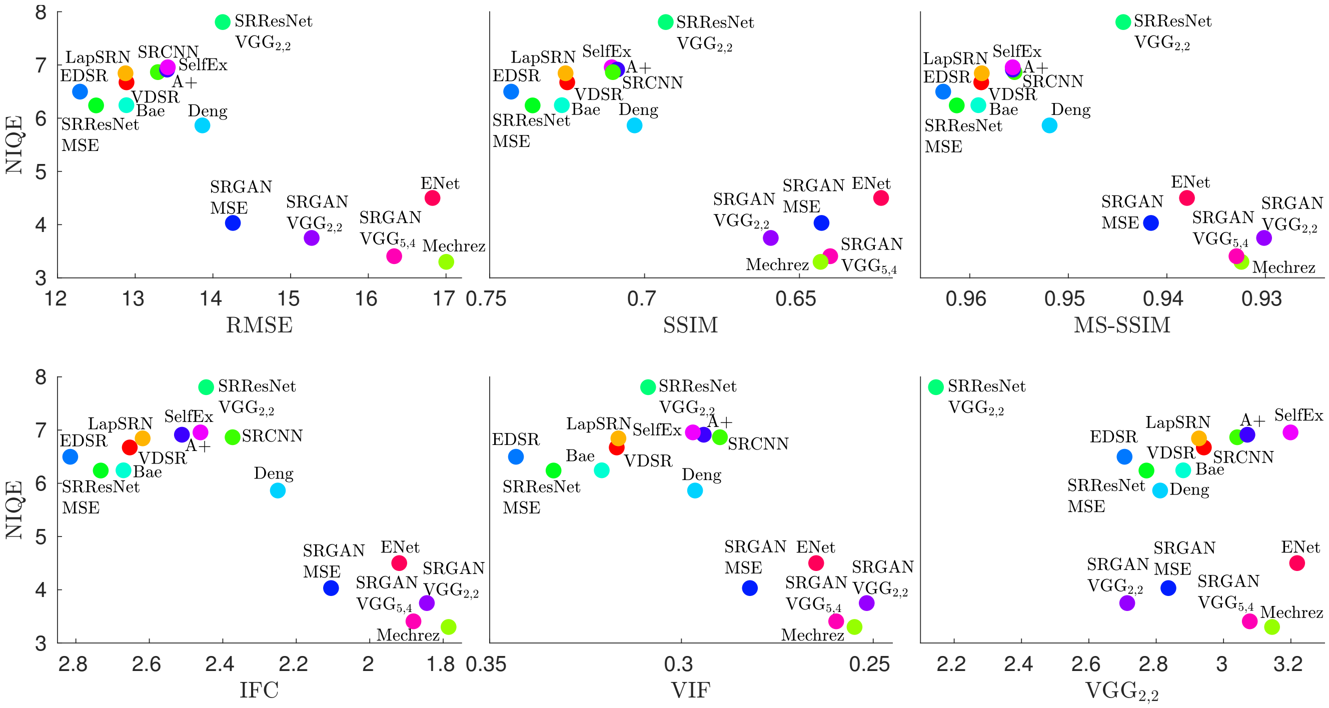

Figure S2. Plot of 15 algorithms on the perception-distortion plane, where perception is measured by the NR metric NIQE, and distortion is measured by the common full-reference metrics RMSE, SSIM, MS-SSIM, IFC, VIF and VGG$_{2,2}$. All metrics were calculated on the **y-channel** alone.

parameters and models. In the paper, we reported comparisons on the y-channel (except for the VGG$_{2,2}$ measure). Below, we report results with additional NR metrics on the y-channel, as well as results on color images. When comparing color images, for SR algorithms which treat the y-channel alone, the Cb and Cr channels are upsampled by bicubic interpolation.

The general pattern appearing in Fig. 8 will appear for any NR method which accurately predicts the perceptual quality of images. We show here three additional popular NR methods NIQE [6], BRISQUE [5],and BLIINDS-II [8] in Figs. S2,S3,S4, where the same conclusions as for Ma *et al*. [4] (see Sec. 6) are apparent. The same pattern appears for RGB images as well, as shown in Figs. S5,S6. Note that the perceptual quality of Johnson *et al*. and SRResNet-VGG$_{2,2}$ is inconsistent between NR metrics, likely due to varying sensitivity to the cross-hatch pattern artifacts which are present in these method's outputs. For this reason, Johnson *et al*. does not appear in the NIQE plots, as its NIQE score is 13.55 (far off the plots).

References

[1] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012. 1

[2] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5769–5779, 2017. 4

[3] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2014. 4

[4] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 5

[5] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 5

[6] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 5

[7] F. Nielsen. Hypothesis testing, information divergence and computational geometry. In *Geometric Science of Information*, pages 241–248. 2013. 4

[8] M. A. Saad, A. C. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE transactions on Image Processing*, 21(8):3339–3352, 2012. 5

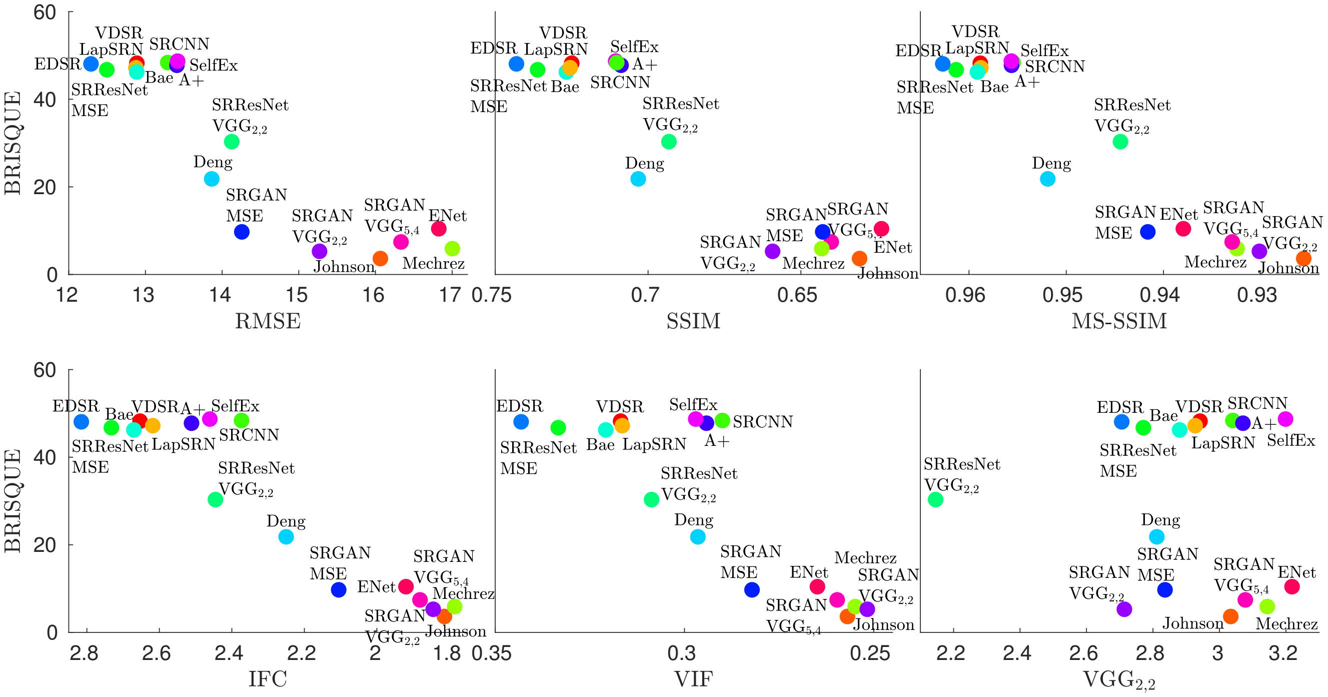

Figure S3. Plot of 16 algorithms on the perception-distortion plane, where perception is measured by the NR metric BRISQUE, and distortion is measured by the common full-reference metrics RMSE, SSIM, MS-SSIM, IFC, VIF and VGG$_{2,2}$. All metrics were calculated on the **y-channel** alone.

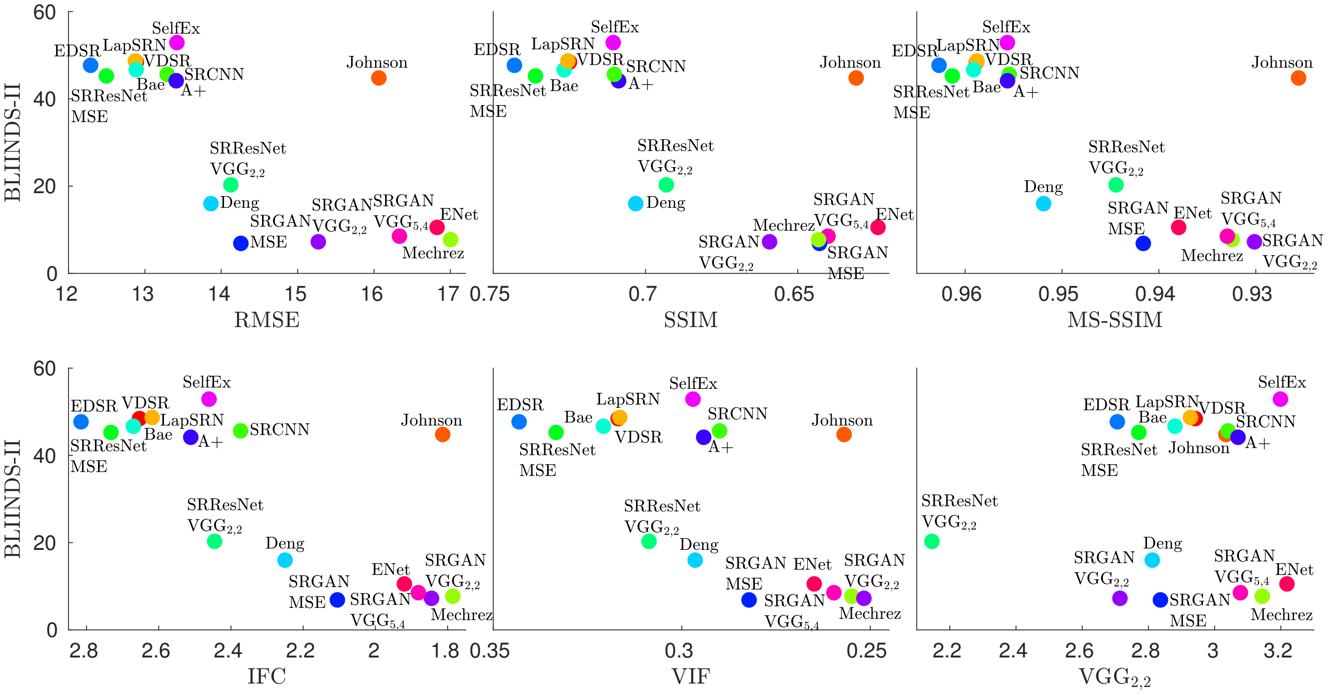

Figure S4. Plot of 16 algorithms on the perception-distortion plane, where perception is measured by the NR metric BLIINDS-II, and distortion is measured by the common full-reference metrics RMSE, SSIM, MS-SSIM, IFC, VIF and VGG$_{2,2}$. All metrics were calculated on the **y-channel** alone.

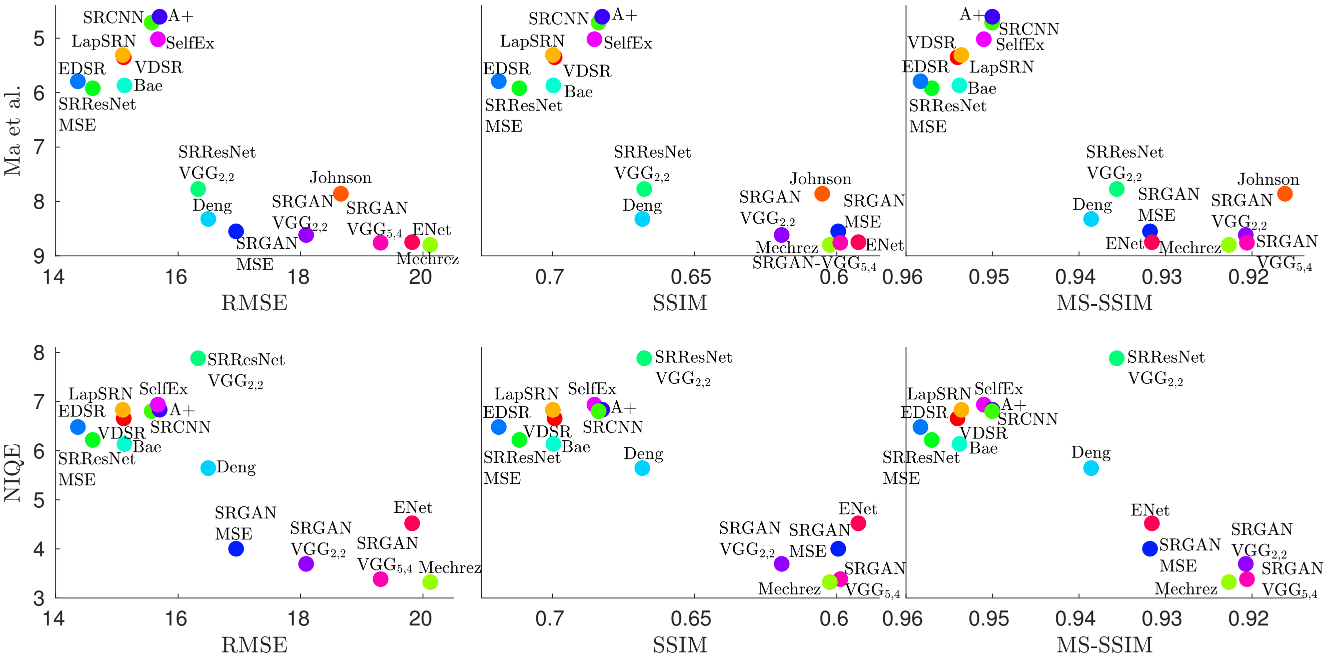

Figure S5. Plot of 16 algorithms on the perception-distortion plane. Perception is measured by the the NR metrics of Ma *et al*. and NIQE, and distortion is measured by the common full-reference metrics RMSE, SSIM and MS-SSIM. All metrics were calculated on **three channel RGB** images.

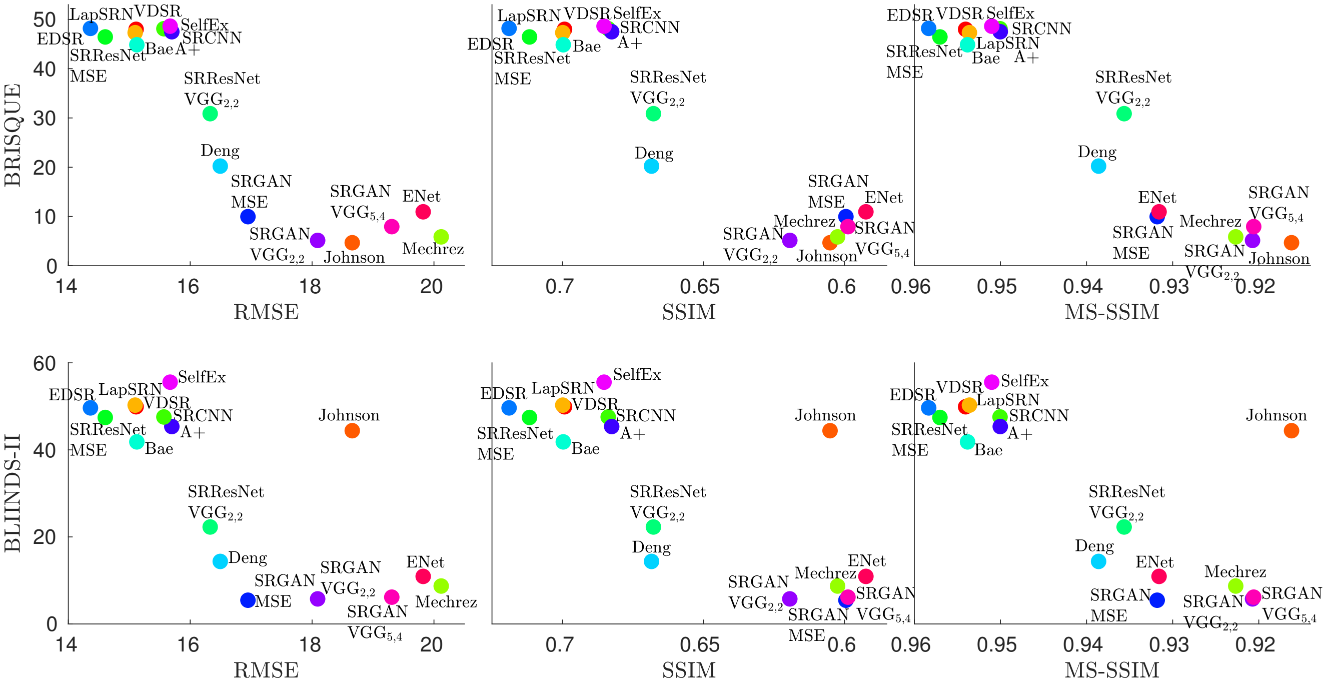

Figure S6. Plot of 16 algorithms on the perception-distortion plane. Perception is measured by the the NR metrics BRISQUE and BLIINDS-II, and distortion is measured by the common full-reference metrics RMSE, SSIM and MS-SSIM. All metrics were calculated on **three channel RGB** images.