The Perception-Distortion Tradeoff

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

Abstract

*Image restoration algorithms are typically evaluated by some distortion measure (*e.g.* PSNR, SSIM, IFC, VIF) or by human opinion scores that quantify perceived perceptual quality. In this paper, we prove mathematically that distortion and perceptual quality are at odds with each other. Specifically, we study the optimal probability for correctly discriminating the outputs of an image restoration algorithm from real images. We show that as the mean distortion decreases, this probability must increase (indicating worse perceptual quality). As opposed to the common belief, this result holds true for any distortion measure, and is not only a problem of the PSNR or SSIM criteria. However, as we show experimentally, for some measures it is less severe (*e.g.* distance between VGG features). We also show that generative-adversarial-nets (GANs) provide a principled way to approach the perception-distortion bound. This constitutes theoretical support to their observed success in low-level vision tasks. Based on our analysis, we propose a new methodology for evaluating image restoration methods, and use it to perform an extensive comparison between recent super-resolution algorithms.*

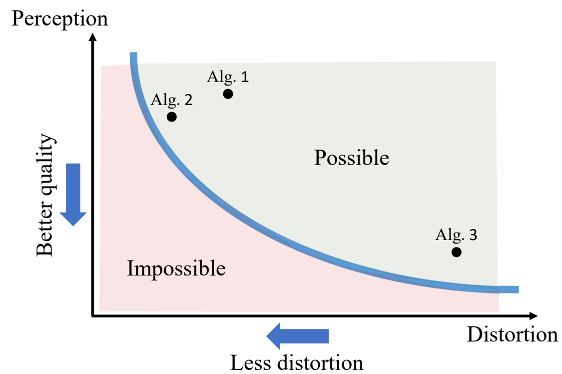

Figure 1. **The perception-distortion tradeoff.** Image restoration algorithms can be characterized by their average distortion and by the perceptual quality of the images they produce. We show that there exists a region in the perception-distortion plane which cannot be attained, regardless of the algorithmic scheme. When in proximity of this unattainable region, an algorithm can be potentially improved only in terms of its distortion *or* in terms of its perceptual quality, one at the expense of the other.

1. Introduction

The last decades have seen continuous progress in image restoration algorithms (*e.g.* for denoising, deblurring, super-resolution) both in visual quality and in distortion measures like peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [45]. However, in recent years, it seems that the improvement in reconstruction accuracy is not always accompanied by an improvement in visual quality. In fact, and perhaps counter-intuitively, algorithms that are superior in terms of perceptual quality, are often inferior in terms of *e.g.* PSNR and SSIM [22, 16, 6, 38, 51, 49]. This phenomenon is commonly interpreted as a shortcoming of the existing distortion measures [44], which fuels a constant search for alternative "more perceptual" criteria.

In this paper, we offer a complementary explanation for the apparent tradeoff between perceptual quality and distortion measures. Specifically, we prove that there exists a region in the perception-distortion plane, which cannot be attained regardless of the algorithmic scheme (see Fig. 1). Furthermore, the boundary of this region is monotone. Therefore, in its proximity, it is only possible to improve *either perceptual quality or distortion*, one at the expense of the other. The perception-distortion tradeoff exists for *all distortion measures*, and is not only a problem of the mean-square error (MSE) or SSIM criteria. However, for some measures, the tradeoff is weaker than others. For example, we find empirically that the recently proposed distance between deep-net features [16, 22] has a weaker tradeoff with perceptual quality than MSE. This aligns with the observation that this measure is "more perceptual" than MSE.

Let us clarify the difference between distortion and perceptual quality. The goal in image restoration is to estimate an image x from its degraded version y (*e.g.* noisy, blurry, etc.). Distortion refers to the dissimilarity between the reconstructed image $\hat{x}$ and the original image x. Perceptual quality, on the other hand, refers only to the visual quality of $\hat{x}$, regardless of its similarity to x. Namely, it is the extent to which $\hat{x}$ looks like a valid natural image. An increasingly popular way of measuring perceptual quality is by using real-vs.-fake user studies, which examine the ability of human observers to tell whether $\hat{x}$ is real or the output of an algorithm [15, 53, 39, 8, 6, 14, 54, 11] (similarly to the idea underlying generative adversarial nets [10]). Therefore, perceptual quality can be defined as the best possible probability of success in such discrimination experiments, which as we show, is proportional to the distance between the distribution of $\hat{x}$ and that of natural images.

Based on these definitions of perception and distortion, we follow the logic of rate-distortion theory [4]. That is, we seek to characterize the behavior of the best attainable perceptual quality (minimal deviation from natural image statistics) as a function of the maximal allowable average distortion, for any estimator. This perception-distortion function (wide curve in Fig. 1) separates between the attainable and unattainable regions in the perception-distortion plane and thus describes the fundamental tradeoff between perception and distortion. Our analysis shows that algorithms cannot be simultaneously very accurate *and* produce images that fool observers to believe they are real, no matter what measure is used to quantify accuracy. This tradeoff implies that optimizing distortion measures can be not only ineffective, but also potentially damaging in terms of visual quality. This has been empirically observed *e.g.* in [22, 16, 38, 51, 6], but was never established theoretically.

From the standpoint of algorithm design, we show that generative adversarial nets (GANs) provide a principled way to approach the perception-distortion bound. This gives theoretical support to the growing empirical evidence of the advantages of GANs in image restoration [22, 38, 35, 51, 36, 15, 55].

The perception-distortion tradeoff has major implications on low-level vision. In certain applications, reconstruction accuracy is of key importance (*e.g.* medical imaging). In others, perceptual quality may be preferred. The impossibility of simultaneously achieving both goals calls for a new way for evaluating algorithms: By placing them on the perception-distortion plane. We use this new methodology to conduct an extensive comparison between recent super-resolution (SR) methods, revealing which SR methods lie closest to the perception-distortion bound.

2. Distortion and perceptual quality

Distortion and perceptual quality have been studied in many different contexts, and are sometimes referred to by different names. Let us briefly put past works in our context.

2.1. Distortion (full-reference) measures

Given a distorted image $\hat{x}$ and a ground-truth reference image x, full-reference distortion measures quantify the quality of $\hat{x}$ by its discrepancy to x. These measures are often called full reference image quality criteria because of the reasoning that if $\hat{x}$ is similar to x and x is of high quality, then $\hat{x}$ is also of high quality. However, as we show in this paper, this logic is not always correct. We thus prefer to call these measures distortion or dissimilarity criteria.

The most common distortion measure is the MSE, which is quite poorly correlated with semantic similarity between images [44]. Many alternative, more perceptual, distortion measures have been proposed over the years, including SSIM [45], MS-SSIM [47], IFC [41], VIF [40], VSNR [3] and FSIM [52]. Recently, measures based on the ℓ_2-distance between deep feature maps of a neural-net have been shown to capture more semantic similarities. These measures were used as loss functions in super-resolution and style transfer applications, leading to reconstructions with high visual quality [16, 22, 38].

2.2. Perceptual quality

The perceptual quality of an image $\hat{x}$ is the degree to which it looks like a natural image, and has nothing to do with its similarity to any reference image. In many image processing domains, perceptual quality has been associated with deviations from natural image statistics.

Human opinion based quality assessment Perceptual quality is commonly evaluated empirically by the mean opinion score of human subjects [31, 29]. Recently, it has become increasingly popular to perform such studies through real vs. fake questionnaires [15, 53, 39, 8, 6, 14, 54, 11]. These test the ability of a human observer to distinguish whether an image is real or the output of some algorithm. The probability of success p_{success} of the optimal decision rule in this hypothesis testing task is known to be

$$p_{\text{success}} = \tfrac{1}{2} d_{\text{TV}}(p_X, p_{\hat{X}}) + \tfrac{1}{2}, \tag{1}$$

where $d_{\text{TV}}(p_X, p_{\hat{X}})$ is the total-variation (TV) distance between the distribution $p_{\hat{X}}$ of images produced by the algorithm in question, and the distribution p_X of natural images [32]. Note that p_{success} decreases as the deviation between $p_{\hat{X}}$ and p_X decreases, becoming $\frac{1}{2}$ (no better than a coin toss) when $p_{\hat{X}} = p_X$.

No-reference quality measures Perceptual quality can also be measured by an algorithm. In particular, no-reference measures quantify the perceptual quality of an image $\hat{x}$ *without* depending on a reference image. These measures are commonly based on estimating deviations from natural image statistics. For example, [46, 48, 23] proposed

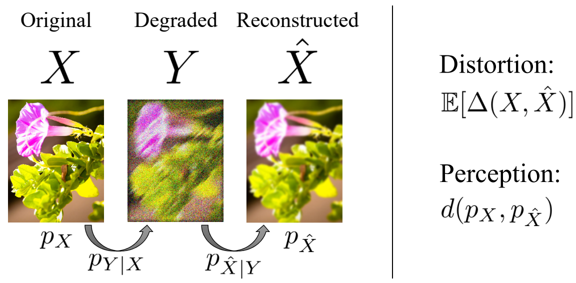

Figure 2. **Problem setting.** Given an original image $x \sim p_X$, a degraded image y is observed according to some conditional distribution $p_{Y|X}$. Given the degraded image y, an estimate $\hat{x}$ is constructed according to some conditional distribution $p_{\hat{X}|Y}$. Distortion is quantified by the mean of some distortion measure between $\hat{X}$ and X. The perceptual quality index corresponds to the deviation between $p_{\hat{X}}$ and p_X.

a perceptual quality index based on the Kullback-Leibler (KL) divergence between the distribution of the wavelet coefficients of $\hat{x}$ and that of natural scenes. This idea was further extended by the popular methods DIIVINE [31], BRISQUE [29], BLIINDS-II [37] and NIQE [30], which quantify perceptual quality by various measures of deviation from natural image statistics in the spatial, wavelet and DCT domains.

GAN-based image restoration Most recently, GAN-based methods have demonstrated unprecedented perceptual quality in super-resolution [22, 38], inpainting [35, 51], compression [36] and image-to-image translation [15, 55]. This was accomplished by utilizing an adversarial loss, which minimizes some distance $d(p_X, p_{\hat{X}_{\text{GAN}}})$ between the distribution $p_{\hat{X}_{\text{GAN}}}$ of images produced by the generator and the distribution p_X of images in the training dataset. A large variety of GAN schemes have been proposed, which minimize different distances between distributions. These include the Jenson-Shannon divergence [10], the Wasserstein distance [1], and any f-divergence [34].

3. Problem formulation

In statistical terms, a natural image x can be thought of as a realization from the distribution of natural images p_X. In image restoration, we observe a degraded version y relating to x via some conditional distribution $p_{Y|X}$ (corresponding to noise, blur, down-sampling, etc.). Given y, we produce an estimate $\hat{x}$ according to some distribution $p_{\hat{X}|Y}$. This description is quite general in that it does not restrict the estimator $\hat{x}$ to be a deterministic function of y. This problem setting is illustrated in Fig. 2.

Given a full-reference dissimilarity criterion $\Delta(x, \hat{x})$, the average distortion of an estimator $\hat{X}$ is given by

$$\mathbb{E}[\Delta(X, \hat{X})], \tag{2}$$

where the expectation is over the joint distribution $p_{X,\hat{X}}$. This definition aligns with the common practice of evaluating average performance over a database of degraded natural images. Note that some distortion measures, *e.g.* SSIM, are actually *similarity* measures (higher is better), yet can always be inverted to become dissimilarity measures.

As discussed in Sec. 2.2, the perceptual quality of an estimator $\hat{X}$ (as quantified *e.g.* by real vs. fake human opinion studies) is directly related to the distance between the distribution of its reconstructed images $p_{\hat{X}}$, and the distribution of natural images p_X. We thus define the perceptual quality index (lower is better) of an estimator $\hat{X}$ as

$$d(p_X, p_{\hat{X}}), \tag{3}$$

where $d(\cdot,\cdot)$ is some divergence between distributions, *e.g.* the KL divergence, TV distance, Wasserstein distance, etc.

Notice that the best possible perceptual quality is obtained when the outputs of the algorithm follow the distribution of natural images (*i.e.* $p_{\hat{X}} = p_X$). In this situation, by looking at the reconstructed images, it is impossible to tell that they were generated by an algorithm. However, not every estimator with this property is necessarily accurate. Indeed, we could achieve perfect perceptual quality by randomly drawing natural images that have nothing to do with the original "ground-truth" images. In this case the distortion would be quite large.

Our goal is to characterize the tradeoff between (2) and (3). But let us first exemplify why minimizing the average distortion (2), does not necessarily lead to a low perceptual quality index (3). We illustrate this with the square-error distortion $\Delta(x, \hat{x}) = \|x - \hat{x}\|^2$ and the $0-1$ distortion $\Delta(x, \hat{x}) = 1 - \delta_{x,\hat{x}}$ (where δ is Kronecker's delta).

3.1. The square-error distortion

The minimum mean square-error (MMSE) estimator is given by the posterior-mean $\hat{x}(y) = \mathbb{E}[X|Y=y]$. Consider the case $Y = X + N$, where X is a discrete random variable with probability mass function

$$p_X(x) = \begin{cases} p_1 & x = \pm 1, \\ p_0 & x = 0, \end{cases} \tag{4}$$

and $N \sim \mathcal{N}(0,1)$ is independent of X (see Fig. 3). In this setting, the MMSE estimate is given by

$$\hat{x}_{\text{MMSE}}(y) = \sum_{n \in \{-1,0,1\}} n \; p(X = n|y), \tag{5}$$

where

$$p(X = n|y) = \frac{p_n \exp\{-\frac{1}{2}(y-n)^2\}}{\sum_{m \in \{-1,0,1\}} p_m \exp\{-\frac{1}{2}(y-m)^2\}}. \tag{6}$$

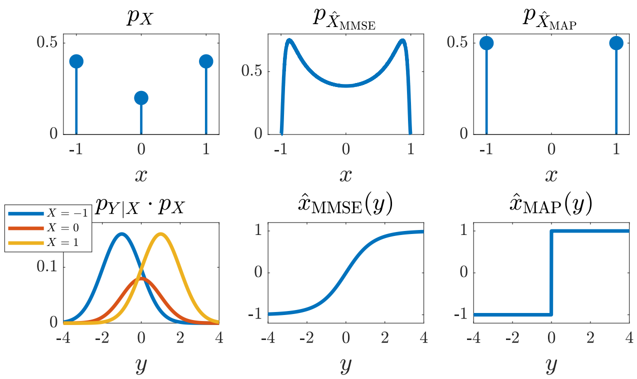

Figure 3. **The distribution of the MMSE and MAP estimates.** In this example, $Y = X + N$, where $X \sim p_X$ and $N \sim \mathcal{N}(0, 1)$. The distributions of both the MMSE and the MAP estimates deviate significantly from the distribution p_X.

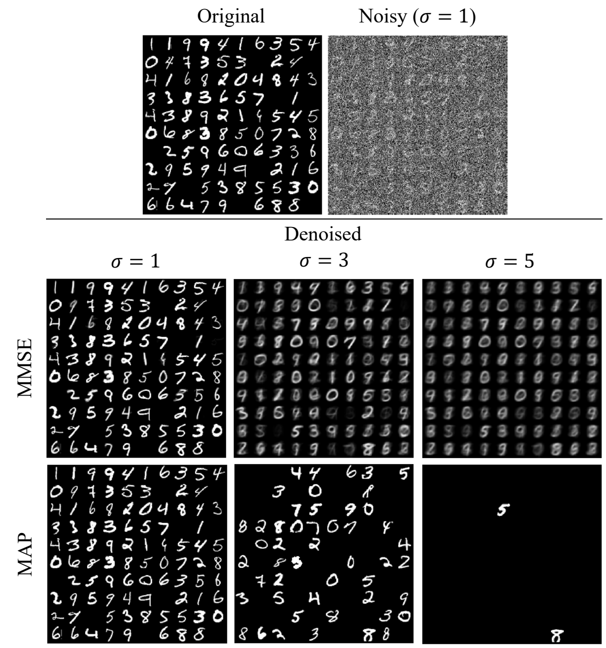

Figure 4. **MMSE and MAP denoising.** Here, the original image consists of 100 smaller images, chosen uniformly at random from the MNIST dataset enriched with blank images. After adding Gaussian noise ($\sigma = 1, 3, 5$), the image is denoised using the MMSE and MAP estimators. In both cases, the estimates significantly deviate from the distribution of images in the dataset.

Notice that $\hat{x}_{\text{MMSE}}$ can take any value in the range $(-1, 1)$, whereas x can only take the discrete values $\{-1, 0, 1\}$. Thus, clearly, $p_{\hat{X}_{\text{MMSE}}}$ is very different from p_X, as illustrated in Fig. 3. This demonstrates that minimizing the MSE distortion *does not* generally lead to $p_{\hat{X}} \approx p_X$.

The same intuition holds for images. The MMSE estimate is an average over all possible explanations to the measured data, weighted by their likelihoods. However the average of valid images is not necessarily a valid image, so that the MMSE estimate frequently "falls off" the natural image manifold [22]. This leads to unnatural blurry reconstructions, as illustrated in Fig. 4. In this experiment, x is a 280×280 image comprising 100 smaller 28×28 digit images. Each digit is chosen uniformly at random from a dataset comprising 54K images from the MNIST dataset [21] and an additional 5.4K blank images. The degraded image y is a noisy version of x. As can be seen, the MMSE estimator produces blurry reconstructions, which do not follow the statistics of the (binary) images in the dataset.

3.2. The $0 - 1$ distortion

The discussion above may give the impression that unnatural estimates are mainly a problem of the square-error distortion, which causes averaging. One way to avoid averaging, is to minimize the binary $0 - 1$ loss, which restricts the estimator to choose $\hat{x}$ only from the set of values that x can take. In fact, the minimum mean $0 - 1$ distortion is attained by the maximum-a-posteriori (MAP) rule, which is very popular in image restoration. However, as we exemplify next, the distribution of the MAP estimator also deviates from p_X. This behavior has also be studied in [33].

Consider again the setting of (4). In this case, the MAP estimate is given by

$$\hat{x}_{\text{MAP}}(y) = \underset{n \in \{-1,0,1\}}{\arg\max} \; p(X = n|y), \qquad (7)$$

where $p(X = n|y)$ is as in (6). Now, it can be easily verified that when $\log(p_1/p_0) > 1/2$, we have $\hat{x}_{\text{MAP}}(y) = \text{sign}(y)$. Namely, the MAP estimator never predicts the value 0. Therefore, in this case, the distribution of the estimate is

$$p_{\hat{X}_{\text{MAP}}}(\hat{x}) = \begin{cases} 0.5 & \hat{x} = +1, \\ 0.5 & \hat{x} = -1, \end{cases} \qquad (8)$$

which is obviously different from p_X of (4) (see Fig. 3).

This effect can also be seen in the experiment of Fig. 4. Here, the MAP estimator is increasingly dominated by blank images as the noise level rises, and thus clearly deviates from the underlying prior distribution.

4. The perception-distortion tradeoff

We saw that low distortion does not generally imply good perceptual-quality. An interesting question, then, is: What is the best perceptual quality that can be attained by an estimator with a prescribed distortion level?

Definition 1. The perception-distortion function of a signal restoration task is given by

$$P(D) = \min_{p_{\hat{X}|Y}} d(p_X, p_{\hat{X}}) \quad \text{s.t.} \quad \mathbb{E}[\Delta(X, \hat{X})] \leq D, \qquad (9)$$

where $\Delta(\cdot, \cdot)$ is a distortion measure and $d(\cdot, \cdot)$ is a divergence between distributions.

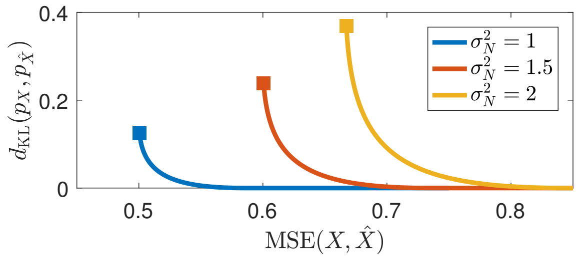

Figure 5. **Plot of Eq. (9) for the setting of Example 1.** The minimal attainable KL distance between p_X and $p_{\hat{X}}$ subject to a constraint on the maximal allowable MSE between X and $\hat{X}$. Here, $Y = X + N$, where $X \sim \mathcal{N}(0,1)$ and $N \sim \mathcal{N}(0, \sigma_N)$, and the estimator is linear, $\hat{X} = aY$. Notice the clear trade-off: The perceptual index (d_{KL}) drops as the allowable distortion (MSE) increases. The graphs cut-off at the MMSE (marked by a square).

In words, $P(D)$ is the minimal deviation between the distributions p_X and $p_{\hat{X}}$ that can be attained by an estimator with distortion D. To gain intuition into the typical behavior of this function, consider the following example.

Example 1. Suppose that $Y = X+N$, where $X \sim \mathcal{N}(0,1)$ and $N \sim \mathcal{N}(0, \sigma_N)$ are independent. Take $\Delta(\cdot,\cdot)$ to be the square-error distortion and $d(\cdot,\cdot)$ to be the KL divergence. For simplicity, let us restrict attention to estimators of the form $\hat{X} = aY$. In this case, we can derive a closed form solution to Eq. (9) (see Supplementary), which is plotted for several noise levels σ_N in Fig. 5. As can be seen, the minimal attainable $d_{\mathrm{KL}}(p_X, p_{\hat{X}})$ drops as the maximal allowable distortion (MSE) increases. Furthermore, the tradeoff is convex and becomes more severe at higher noise levels σ_N.

In general settings, it is impossible to solve (9) analytically. However, it turns out that the behavior seen in Fig. 5 is typical, as we show next (see proof in the Supplementary).

Theorem 1 (The perception-distortion tradeoff)**.** *Assume the problem setting of Section 3. If $d(p,q)$ of (3) is convex in its second argument*[1]*, then the perception-distortion function $P(D)$ of (9) is*

1. *monotonically non-increasing;*
2. *convex.*

Note that Theorem 1 requires no assumptions on the distortion measure $\Delta(\cdot,\cdot)$. This implies that a tradeoff between perceptual quality and distortion exists for *any distortion measure*, including *e.g.* MSE, SSIM, square error between VGG features [16, 22], etc. Yet, this does not imply that all distortion measures have the same perception-distortion function. Indeed, as we demonstrate in Sec. 6, the tradeoff tends to be less severe for distortion measures that capture semantic similarities between images.

[1] $d(p, \lambda q_1 + (1-\lambda) q_2) \le \lambda d(p, q_1) + (1-\lambda) d(p, q_2),\ \forall \lambda \in [0,1]$

The convexity of $P(D)$ implies that the tradeoff is more severe at the low-distortion and at the high-perceptual-quality extremes. This is particularly important when considering the TV divergence which is associated with the ability to distinguish between real vs. fake images (see Sec. 2.2). Since $P(D)$ is steeper at the low-distortion regime, any *small* improvement in distortion for an algorithm whose distortion is already low, must be accompanied by a *large* degradation in the ability to fool a discriminator. Similarly, any *small* improvement in the perceptual quality of an algorithm whose perceptual index is already low, must be accompanied by a *large* increase in distortion. Let us comment that the assumption that $d(p,q)$ is convex, is not very limiting. For instance, any f-divergence (*e.g.* KL, TV, Hellinger, $\mathcal{X}^2$) as well as the Renyi divergence, satisfy this assumption [5, 43]. In any case, the function $P(D)$ is monotonically non-increasing even without this assumption.

4.1. Connection to rate-distortion theory

The perception-distortion tradeoff is closely related to the well-established rate-distortion theory [4]. This theory characterizes the tradeoff between the bit-rate required to communicate a signal, and the distortion incurred in the signal's reconstruction at the receiver. More formally, the rate-distortion function of a signal X is defined by

$$R(D) = \min_{p_{\hat{X}|X}} I(X;\hat{X}) \quad \text{s.t.} \quad \mathbb{E}[\Delta(X,\hat{X})] \le D, \quad (10)$$

where $I(X;\hat{X})$ is the mutual information between X and $\hat{X}$.

There are, however, several key differences between the two tradeoffs. First, in rate-distortion the optimization is over all conditional distributions $p_{\hat{X}|X}$, *i.e.* given the *original* signal. In the perception-distortion case, the estimator has access only to the degraded signal Y, so that the optimization is over the conditional distributions $p_{\hat{X}|Y}$, which is more restrictive. In other words, the perception-distortion tradeoff depends on the degradation $p_{Y|X}$, and not only on the signal's distribution p_X (see Example 1). Second, in rate-distortion the rate is quantified by the mutual information $I(X;\hat{X})$, which depends on the joint distribution $p_{X,\hat{X}}$. In our case, perception is quantified by the similarity between p_X and $p_{\hat{X}}$, which does not depend on their joint distribution. Lastly, mutual information is inherently convex, while the convexity of the perception-distortion curve is guaranteed only when $d(\cdot,\cdot)$ is convex.

5. Traversing the tradeoff with a GAN

There exists a systematic way to design estimators that approach the perception-distortion curve: Using GANs. Specifically, motivated by [22, 35, 51, 38, 36, 15], restoration problems can be approached by modifying the loss of

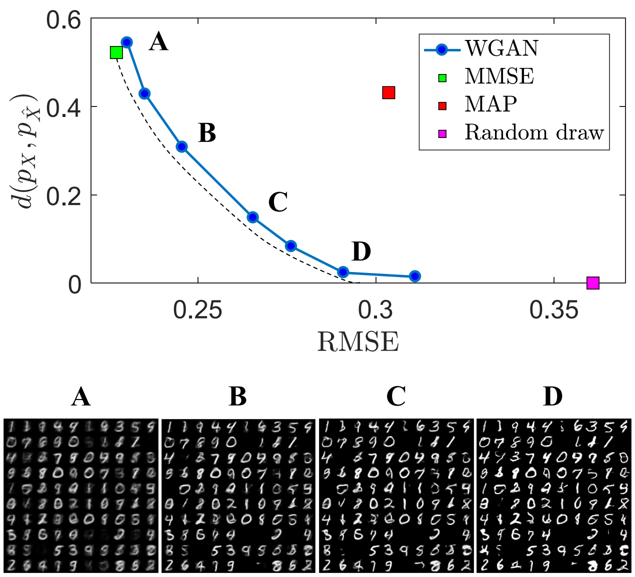

Figure 6. **Image denoising utilizing a GAN.** A Wasserstein GAN was trained to denoise the images of the experiment in Fig. 4. The generator loss $l_{\text{gen}} = l_{\text{MSE}} + \lambda\, l_{\text{adv}}$ consists of a perceptual quality (adversarial) loss and a distortion (MSE) loss, where λ controls the trade-off between the two. For each $\lambda \in [0, 0.3]$, the graph depicts the distortion (MSE) and perceptual quality (Wasserstein distance between p_X and $p_{\hat{X}}$). The curve connecting the estimators is a good approximation to the theoretical perception-distortion trade-off (illustrated by a dashed line).

the generator of a GAN to be

$$\ell_{\text{gen}} = \ell_{\text{distortion}} + \lambda\, \ell_{\text{adv}}, \tag{11}$$

where $\ell_{\text{distortion}}$ is the distortion between the original and reconstructed images, and ℓ_{adv} is the standard GAN adversarial loss. It is well known that ℓ_{adv} is proportional to some divergence $d(p_X, p_{\hat{X}})$ between the generator and data distributions [10, 1, 34] (the type of divergence depends on the loss). Thus, (11) in fact approximates the objective

$$\ell_{\text{gen}} \approx \mathbb{E}[\Delta(x, \hat{x})] + \lambda\, d(p_X, p_{\hat{X}}). \tag{12}$$

Viewing λ as a Lagrange multiplier, it is clear that minimizing ℓ_{gen} is equivalent to minimizing (9) for some D. Varying λ correspond to varying D, thus producing estimators along the perception-distortion function.

Let us use this approach to explore the perception-distortion tradeoff for the digit denoising example of Fig. 4 with $\sigma = 3$. We train a Wasserstein GAN (WGAN) based denoiser [1, 12] with an MSE distortion loss $\ell_{\text{distortion}}$. Here, ℓ_{adv} is proportional to the Wasserstein distance $d_W(p_X, p_{\hat{X}})$ between the generator and data distributions. The WGAN has the valuable property that its discriminator (critic) loss is an accurate estimate (up to a constant factor) of $d_W(p_X, p_{\hat{X}})$ [1]. This allows us to easily compute the perceptual quality index of the trained denoiser. We obtain a set of estimators with several values of $\lambda \in [0, 0.3]$. For each denoiser, we evaluate the perceptual quality by the final discriminator loss. As seen in Fig. 6, the curve connecting the estimators on the perception-distortion plane is monotonically decreasing. Moreover, it is associated with estimates that gradually transition from blurry and accurate to sharp and inaccurate. This curve obviously does not coincide with the analytic bound (9) (illustrated by a dashed line). However, it seems to be adjacent to it. This is indicated by the fact that the left-most point of the WGAN curve is very close to the left-most point of the theoretical bound, which corresponds to the MMSE estimator. See the Supplementary for the WGAN training details and architecture.

Besides the MMSE estimator, Figure 6 also includes the MAP estimator and an estimator which randomly draws images from the dataset (denoted "random draw"). The perceptual quality of those three estimators is evaluated, as above, by the final loss of the WGAN discriminator [1], trained (without a generator) to distinguish between the estimators' outputs and images from the dataset. Note that the denoising WGAN estimator (D) achieves the same distortion as the MAP estimator, but with far better perceptual quality. Furthermore, it achieves nearly the same perceptual quality as the random draw estimator, but with a significantly lower distortion.

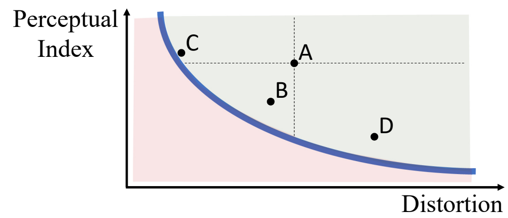

Figure 7. **Dominance and admissibility.** Algorithm A is dominated by Algorithm B, and is thus inadmissible. Algorithms B, C and D are all admissible, as they are not dominated by any algorithm.

6. Practical method for evaluating algorithms

Certain applications may require low-distortion (*e.g.* in medical imaging), while others may prefer superior perceptual quality. How should image restoration algorithms be evaluated, then?

Definition 2. We say that Algorithm A *dominates* Algorithm B if it has better perceptual quality *and* less distortion.

Note that if Algorithm A is better than B in only one of the two criteria, then neither A dominates B nor B dominates A. Therefore, among a group of algorithms, there may be a large subset which can be considered equally good.

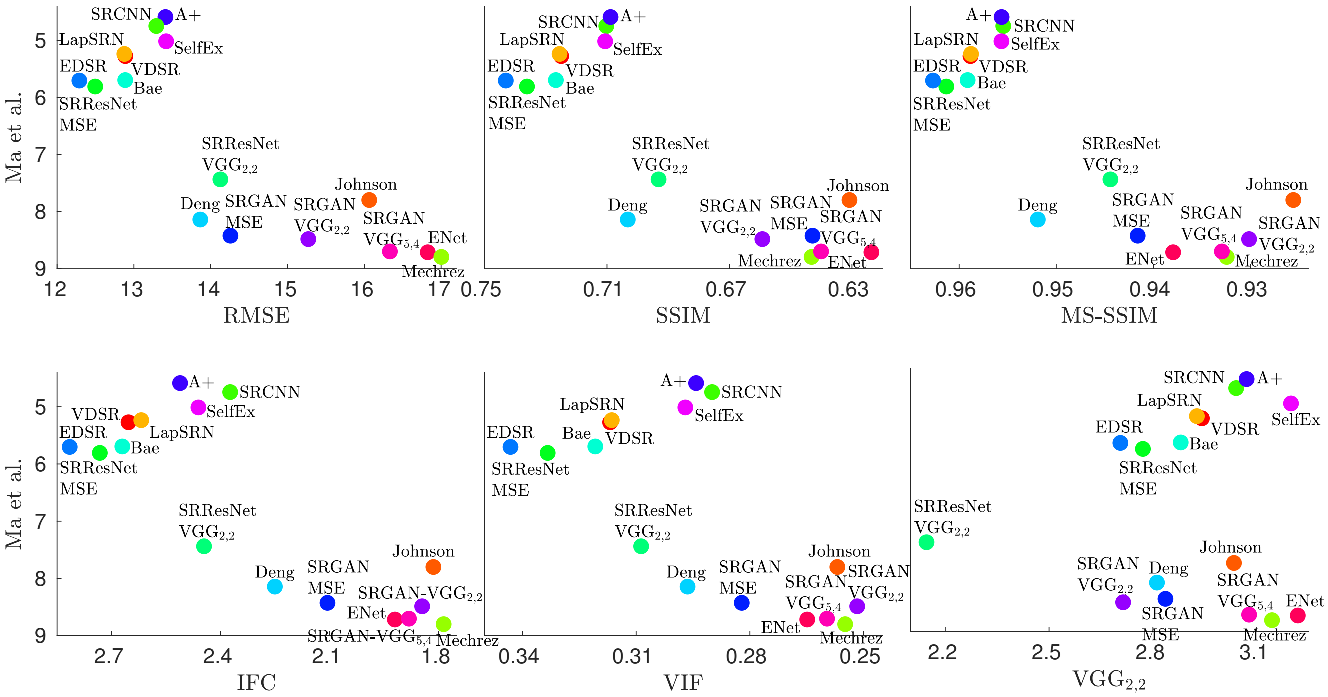

Figure 8. **Perception-distortion evaluation of SR algorithms.** We plot 16 algorithms on the perception-distortion plane. Perception is measured by the recent NR metric by Ma *et al*. [26] which is specifically designed for SR quality assessment. Distortion is measured by the common full-reference metrics RMSE, SSIM, MS-SSIM, IFC, VIF and $VGG_{2,2}$. In all plots, the lower left corner is blank, revealing an unattainable region in the perception-distortion plane. In proximity of the unattainable region, an improvement in perceptual quality comes at the expense of higher distortion.

Definition 3. We say that an algorithm is *admissible* among a group of algorithms, if it is not dominated by any other algorithm in the group.

As shown in Figure 7, these definitions have very simple interpretations when plotting algorithms on the perception-distortion plane. In particular, the admissible algorithms in the group, are those which lie closest to the perception-distortion bound.

As discussed in Sec. 2, distortion is measured by *full*-reference (FR) metrics, *e.g*. [45, 47, 41, 40, 3, 52, 16]. The choice of the FR metric, depends on the type of similarities we want to measure (per-pixel, semantic, etc.). Perceptual quality, on the other hand, is ideally quantified by collecting human opinion scores, which is time consuming and costly [31, 37]. Instead, the divergence $d(p_X, p_{\hat{X}})$ can be computed, for instance by training a discriminator net (see Sec. 5). However, this requires *many* training images and is thus also time consuming. A practical alternative is to utilize *no*-reference (NR) metrics, *e.g*. [29, 30, 37, 31, 50, 17, 26], which quantify the perceptual quality of an image *without* a corresponding original image. In scenarios where NR metrics are highly correlated with human mean-opinion-scores (*e.g*. $4\times$ super-resolution [26]), they can be used as a fast and simple method for approximating the perceptual quality of an algorithm[2].

We use this approach to evaluate 16 SR algorithms in a $4\times$ magnification task, by plotting them on the perception-distortion plane (Fig. 8). We measure perceptual quality using the recent NR metric by Ma *et al*. [26] which is specifically designed for SR quality assessment (see Supplementary for experiments with the NR metrics BRISQUE [29], NIQE [30] and BLIINDS-II [37]). We measure distortion by the five common FR metrics RMSE, SSIM [45], MS-SSIM [47], IFC [41] and VIF [40], and additionally by the recent $VGG_{2,2}$ metric (the distance in the feature space of a VGG net) [22, 16]. To conform to previous evaluations, we compute all metrics on the y-channel after discarding a 4-pixel border (except for $VGG_{2,2}$, which is computed on RGB images). Comparisons on color images can be found in the Supplementary. The algorithms are evaluated on the BSD100 dataset [27]. The evaluated algorithms include: A+ [42], SRCNN [9], SelfEx [13], VDSR [18], Johnson *et al*. [16], LapSRN [19], Bae *et al*. [2] ("primary" variant), EDSR [24], SRResNet variants which optimize MSE

[2]In scenarios where NR metrics are inaccurate (*e.g*. blind deblurring with large blurs [20, 25]), the perceptual metric should be human-opinion-scores or the loss of a discriminator trained to distinguish the algorithms' outputs from natural images.

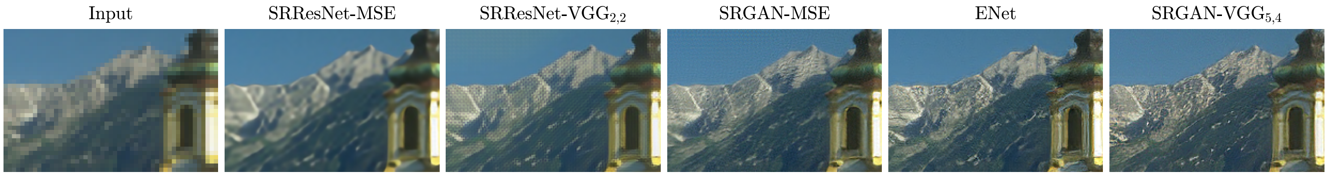

Figure 9. **Visual comparison of algorithms closest to the perception-distortion bound.** The algorithms are ordered from low to high distortion (evaluated by IFC). Notice the co-occurring increase in perceptual quality.

and VGG$_{2,2}$ [22], SRGAN variants which optimize MSE, VGG$_{2,2}$, and VGG$_{5,4}$, in addition to an adversarial loss [22], ENet [38] ("PAT" variant), Deng [7] ($\gamma = 0.55$), and Mechrez *et al*. [28].

Interestingly, the same pattern is observed in all plots: (i) The lower left corner is blank, revealing an unattainable region in the perception-distortion plane. (ii) In proximity of this blank region, NR and FR metrics are *anti-correlated*, indicating a tradeoff between perception and distortion. Notice that the tradeoff exists even for the IFC and VIF measures, which are considered to capture visual quality better than MSE and SSIM. The tradeoff is evident also for the VGG$_{2,2}$ measure, but is somewhat weaker than for MSE. This may indicate that VGG$_{2,2}$ is a more "perceptual" metric. It should be noted, however, that when using other NR metrics to measure perceptual quality, the tradeoff for VGG$_{2,2}$ does not appear to be weaker (see Supplementary). This is due to the sensitivity of some of the NR metrics to the periodic artifacts that arise when minimizing the VGG$_{2,2}$ distortion[3] (see Fig. 9).

Figure 9 depicts the outputs of several algorithms lying closest to the perception-distortion bound in the IFC graph. While the images are ordered from low to high distortion (according to IFC), their perceptual quality clearly improves from left to right.

Both FR and NR measures are commonly validated by calculating their correlation with human opinion scores, based on the assumption that both should be correlated with perceptual quality. However, as Fig. 10 shows, while FR measures can be well-correlated with perceptual quality when distant from the unattainable region, this is clearly not the case when approaching the perception-distortion bound. In particular, all tested FR methods are inconsistent with human opinion scores which found the SRGAN to be superb in terms of perceptual quality [22], while NR methods successfully determine this. We conclude that image restoration algorithms should always be evaluated by a pair of NR and FR metrics, constituting a reliable, reproducible and simple method for comparison, which accounts for both perceptual quality and distortion.

[3]Minimizing VGG$_{2,2}$ (as done by SRResNet-VGG$_{2,2}$), leads to sharper images (compared to minimizing MSE) but with periodic artifacts [16]. Different NR metrics have different sensitivities to these artifacts.

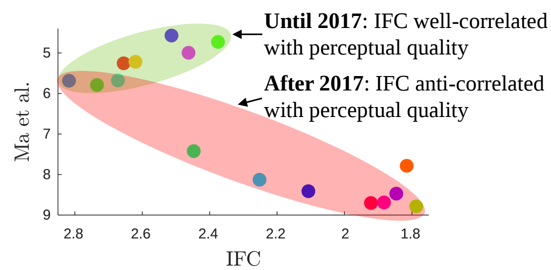

Figure 10. **Correlation between distortion and perceptual quality.** In proximity of the perception-distortion bound, distortion and perceptual quality are *anti-correlated*. However, correlation is possible at distance from the bound.

Up until 2016, SR algorithms occupied only the upper-left section of the perception-distortion plane. Nowadays, emerging techniques are exploring new regions in this plane. The SRGAN, ENet, Deng, Johnson *et al*. and Mechrez *et al*. methods are the first (to our knowledge) to populate the high perceptual quality region. In the near future we will most likely witness continued efforts to approach the perception-distortion bound, not only in the low-distortion region, but throughout the entire plane.

7. Conclusion

We proved and demonstrated the counter-intuitive phenomenon that distortion and perceptual quality are at odds with each other. Namely, the lower the distortion of an algorithm, the more its distribution must deviate from the statistics of natural scenes. We showed empirically that this tradeoff exists for many popular distortion measures, including those considered to be well-correlated with human perception. Therefore, any distortion measure alone, is unsuitable for assessing image restoration methods. Our novel methodology utilizes a pair of NR and FR metrics to place each algorithm on the perception-distortion plane, facilitating a more informative comparison of image restoration methods.

Acknowledgements This research was supported in part by an Alon Fellowship, by the Israel Science Foundation (grant no. 852/17), and by the Ollendorf Foundation.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 214–223, 2017. 3, 6

[2] W. Bae, J. Yoo, and J. Chul Ye. Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 145–153, 2017. 7

[3] D. M. Chandler and S. S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16(9):2284–2298, 2007. 2, 7

[4] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012. 2, 5

[5] I. Csiszár, P. C. Shields, et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004. 5

[6] R. Dahl, M. Norouzi, and J. Shlens. Pixel recursive super resolution. In *International Conference on Computer Vision (ICCV)*, pages 5439–5448, 2017. 1, 2

[7] X. Deng. Enhancing image quality via style transfer for single image super-resolution. *IEEE Signal Processing Letters*, 2018. 8

[8] E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a Laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1486–1494, 2015. 2

[9] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 184–199, 2014. 7

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014. 2, 3, 6

[11] S. Guadarrama, R. Dahl, D. Bieber, M. Norouzi, J. Shlens, and K. Murphy. PixColor: Pixel recursive colorization. *British Machine Vision Conference (BMVC)*, 2017. 2

[12] 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. 6

[13] J.-B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2015. 7

[14] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (TOG)*, 35(4):110, 2016. 2

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017. 2, 3, 5

[16] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 694–711, 2016. 1, 2, 5, 7, 8

[17] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1733–1740, 2014. 7

[18] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1646–1654, 2016. 7

[19] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep Laplacian pyramid networks for fast and accurate super-resolution. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 624–632, 2017. 7

[20] W.-S. Lai, J.-B. Huang, Z. Hu, N. Ahuja, and M.-H. Yang. A comparative study for single image blind deblurring. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1709, 2016. 7

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 4

[22] C. Ledig, L. Theis, F. Huszár, 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017. 1, 2, 3, 4, 5, 7, 8

[23] Q. Li and Z. Wang. Reduced-reference image quality assessment using divisive normalization-based image representation. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):202–211, 2009. 2

[24] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 7

[25] Y. Liu, J. Wang, S. Cho, A. Finkelstein, and S. Rusinkiewicz. A no-reference metric for evaluating the quality of motion deblurring. *ACM Transactions on Graphics (TOG)*, 32(6):175–1, 2013. 7

[26] 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. 7

[27] 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 *International Conference on Computer Vision (ICCV)*, volume 2, pages 416–423, 2001. 7

[28] R. Mechrez, I. Talmi, F. Shama, and L. Zelnik-Manor. Learning to maintain natural image statistics. *arXiv preprint arXiv:1803.04626*, 2018. 8

[29] 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. 2, 3, 7

[30] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 3, 7

[31] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364, 2011. 2, 3, 7

[32] F. Nielsen. Hypothesis testing, information divergence and computational geometry. In *Geometric Science of Information*, pages 241–248. 2013. 2

[33] M. Nikolova. Model distortions in Bayesian MAP reconstruction. *Inverse Problems and Imaging*, 1(2):399, 2007. 4

[34] S. Nowozin, B. Cseke, and R. Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 271–279, 2016. 3, 6

[35] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016. 2, 3, 5

[36] O. Rippel and L. Bourdev. Real-time adaptive image compression. In *International Conference on Machine Learning (ICML)*, pages 2922–2930, 2017. 2, 3, 5

[37] 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. 3, 7

[38] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch. EnhanceNet: Single image super-resolution through automated texture synthesis. In *International Conference on Computer Vision (ICCV)*, pages 4491–4500, 2017. 1, 2, 3, 5, 8

[39] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2234–2242, 2016. 2

[40] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006. 2, 7

[41] H. R. Sheikh, A. C. Bovik, and G. De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, 2005. 2, 7

[42] R. Timofte, V. De Smet, and L. Van Gool. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian Conference on Computer Vision*, pages 111–126, 2014. 7

[43] T. Van Erven and P. Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. 5

[44] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009. 1, 2

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1, 2, 7

[46] Z. Wang and E. P. Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *Human Vision and Electronic Imaging*, volume 5666, pages 149–159, 2005. 2

[47] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *Conference on Signals, Systems and Computers*, volume 2, pages 1398–1402, 2003. 2, 7

[48] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik. Quality-aware images. *IEEE Transactions on Image Processing*, 15(6):1680–1689, 2006. 2

[49] C.-Y. Yang, C. Ma, and M.-H. Yang. Single-image super-resolution: A benchmark. In *European Conference on Computer Vision (ECCV)*, pages 372–386, 2014. 1

[50] P. Ye, J. Kumar, L. Kang, and D. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1098–1105, 2012. 7

[51] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with deep generative models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5485–5493, 2017. 1, 2, 3, 5

[52] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. 2, 7

[53] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, pages 649–666, 2016. 2

[54] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros. Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics (TOG)*, 9(4), 2017. 2

[55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 2, 3