Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff

Yochai Blau [1] **Tomer Michaeli** [1]

Abstract

Lossy compression algorithms are typically designed and analyzed through the lens of Shannon's rate-distortion theory, where the goal is to achieve the lowest possible distortion (e.g., low MSE or high SSIM) at any given bit rate. However, in recent years, it has become increasingly accepted that "low distortion" is not a synonym for "high perceptual quality", and in fact optimization of one often comes at the expense of the other. In light of this understanding, it is natural to seek for a generalization of rate-distortion theory which takes perceptual quality into account. In this paper, we adopt the mathematical definition of perceptual quality recently proposed by Blau & Michaeli (2018), and use it to study the three-way tradeoff between rate, distortion, and perception. We show that restricting the perceptual quality to be high, generally leads to an elevation of the rate-distortion curve, thus necessitating a sacrifice in either rate or distortion. We prove several fundamental properties of this triple-tradeoff, calculate it in closed form for a Bernoulli source, and illustrate it visually on a toy MNIST example.

1. Introduction

Lossy compression techniques are ubiquitous in the modern-day digital world, and are regularly used for communicating and storing images, video and audio. In recent years, lossy compression is seeing a surge of research, due in part to the advancements in deep learning and their application in this domain (Toderici et al., 2016; 2017; Ballé et al., 2016; 2017; 2018; Agustsson et al., 2017; 2018; Rippel & Bourdev, 2017; Minnen et al., 2018; Li et al., 2018; Mentzer et al., 2018; Johnston et al., 2018; Galteri et al., 2017; Tschannen et al., 2018; Santurkar et al., 2018; Rott Shaham & Michaeli, 2018). The theoretical foundations of lossy compression are rooted in Shannon's seminal work on rate-distortion theory (Shannon, 1959), which analyzes the fundamental tradeoff between the bit rate used for representing data, and the distortion incurred when reconstructing the data from its compressed representation (Cover & Thomas, 2012).

The premise in rate-distortion theory is that reduced distortion is a desired property. However, recent works demonstrate that minimizing distortion alone does not necessarily drive the decoded signals to have good perceptual quality. For example, incorporating generative adversarial type losses has been shown to lead to significantly better perceptual quality, but at the cost of *increased* distortion (Tschannen et al., 2018; Agustsson et al., 2018; Santurkar et al., 2018). This behavior has also been studied in the context of signal restoration (Blau & Michaeli, 2018), where it was shown that minimizing distortion causes the distribution of restored signals to deviate from that of the ground-truth signals (indicating worse perceptual quality). In light of this understanding, it is natural to seek for a generalized rate-distortion theory, which also accounts for perception. In particular, it is of key importance to understand how the best achievable rate depends not only on the distortion, but also on the perceptual quality of the algorithm. A preliminary attempt to incorporate perceptual quality into rate-distortion theory was briefly reported in (Matsumoto, 2018a;b). Yet, no theoretical characterization nor practical demonstration of its effect on the rate-distortion tradeoff was presented.

In this paper, we adopt the mathematical definition of perceptual quality used in (Blau & Michaeli, 2018), and prove that there is a triple tradeoff between rate, distortion *and* perception. Our key observation is that the rate-distortion function elevates as the perceptual quality is enforced to be higher (see Fig. 1). In other words, to obtain good perceptual quality, it is necessary to make a sacrifice in either the distortion or the rate of the algorithm.

Our analysis is based on the definition of a *rate-distortion-perception function* $R(D, P)$, which characterizes the minimal achievable rate R for any given distortion D and perception index P. We begin by deriving a closed form for this function in the classical case study of a Bernoulli source, a simple example which nonetheless nicely illustrates the typical behavior of the tradeoff. We then prove several general properties of $R(D, P)$, showing that it is monotone

[1]Technion–Israel Institute of Technology, Haifa, Israel. Correspondence to: Yochai Blau <yochai@campus.technion.ac.il>, Tomer Michaeli <tomer.m@ee.technion.ac.il>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

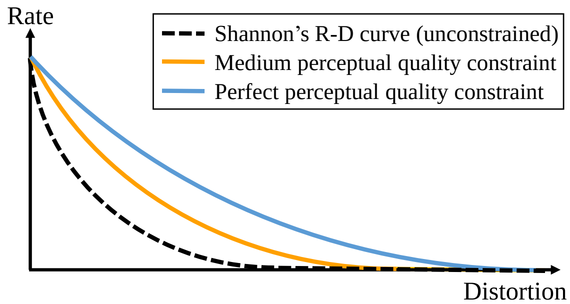


Figure 1. **The rate-distortion function under a perceptual quality constraint.** When perceptual quality is unconstrained, the tradeoff is characterized by Shannon's classic rate-distortion function (black line). However, as the constraint on perceptual quality is tightened to ensure perceptually pleasing reconstructions, the function elevates (colored lines). Thus, the improvement in perceptual quality comes at the cost of a higher rate and/or distortion.

and convex for any full-reference distortion measure (under minor assumptions), and that there is a range of P values for which it necessarily does not coincide with the traditional rate-distortion function. For the specific case of the squared-error distortion, we also provide an upper bound on the increase in distortion that has to be incurred in order to achieve perfect perceptual quality, at any given rate.

Our observations have important implications for the design and evaluation of practical compression methods. In particular, they suggest that comparing between algorithms only in terms of their rate-distortion curves can be misleading. We demonstrate this in the context of image compression using a toy MNIST example, by systematically exploring the visual effect of improvement in each of the three properties (rate, distortion, perception) on the expense of the others. We do this by training an encoder-decoder net utilizing a generative model, similarly to (Tschannen et al., 2018; Agustsson et al., 2018). As we show, the phenomena we discuss are dominant at low bit rates, where the classical approach of optimizing distortion alone leads to unacceptable perceptual quality. This is perhaps not surprising when using the MSE distortion, which is known to be inconsistent with human perception. But our theory shows that *every* distortion measure (excluding pathological cases) must have a tradeoff with perceptual quality. This includes e.g., the popular SSIM/MS-SSIM (Wang et al., 2003; 2004), the L_2 distance between deep features (Johnson et al., 2016), and any other full reference criterion. To illustrate this, we repeat our toy experiment with the distortion measure of (Johnson et al., 2016), which has been used as a means for enhancing perceptual quality in low-level vision tasks (Ledig et al., 2017). As we show, minimizing this distortion *does not* lead to good perceptual quality at low bit rates, just like our theory predicts. Moreover, when enforcing high perceptual quality, this distortion rather increases.

2. Background

2.1. Rate-Distortion Theory

Rate-distortion theory analyzes the fundamental tradeoff between the rate (bits per sample) used for representing samples from a data source $X \sim p_X$, and the expected distortion incurred in decoding those samples from their compressed representations. Formally, the relation between the input X and output $\hat{X}$ of an encoder-decoder pair, is a (possibly stochastic) mapping defined by some conditional distribution $p_{\hat{X}|X}$, as visualized in Fig. 2. The expected distortion of the decoded signals is thus defined as

$$\mathbb{E}[\Delta(X, \hat{X})], \tag{1}$$

where the expectation is with respect to the joint distribution $p_{X,\hat{X}} = p_{\hat{X}|X} p_X$, and $\Delta : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$ is any full-reference distortion measure such that $\Delta(x, \hat{x}) = 0$ if and only if $x = \hat{x}$ (e.g., squared error, L_2 distance between deep features (Johnson et al., 2016; Zhang et al., 2018), SSIM/ MS-SSIM[1] (Wang et al., 2003; 2004), PESQ (Rix et al., 2001), etc.).

A key result in rate-distortion theory states that for an iid source X, if the expected distortion is bounded by D, then the lowest achievable rate R is characterized by the (information) rate-distortion function

$$R(D) = \min_{p_{\hat{X}|X}} I(X, \hat{X}) \quad \text{s.t.} \quad \mathbb{E}[\Delta(X, \hat{X})] \leq D, \tag{2}$$

where I denotes mutual information (Cover & Thomas, 2012). Closed form expressions for the rate-distortion function $R(D)$ are known for only a few source distributions and under quite simple distortion measures (e.g., squared error or Hamming distance). However several general properties of this function are known, including that it is always monotonically non-increasing and convex.

2.2. Perceptual Quality

The perceptual quality of an output sample $\hat{x}$ refers to the extent to which it is perceived by humans as a valid (natural) sample, regardless of its similarity to the input x. In various domains, perceptual quality has been associated with the deviation of the distribution $p_{\hat{X}}$ of output signals from the distribution p_X of natural signals, which, as discussed in (Blau & Michaeli, 2018), is linked to the common practice of quantifying perceptual quality via real-vs.-fake user studies (Isola et al., 2017; Salimans et al., 2016; Zhang et al., 2016; Denton et al., 2015). In particular, deviation from natural scene statistics is the basis for many no-reference image quality measures (Mittal et al., 2013; 2012; Wang

[1]Measures like SSIM, which quantify similarity rather than dissimilarity and are not necessarily positive, need to be negated and shifted to qualify as valid distortion measures.

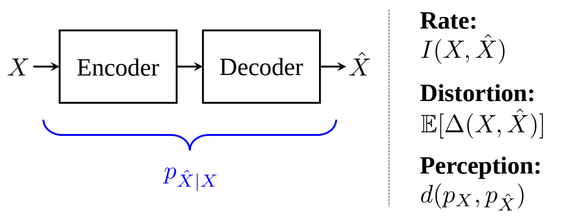


Figure 2. **Lossy compression.** A source signal $X \sim p_X$ is mapped into a coded sequence by an encoder and back into an estimated signal $\hat{X}$ by the decoder. Three desired properties are: (i) the coded sequence be compact (low bit rate); (ii) the reconstruction $\hat{X}$ be similar to the source X on average (low distortion); (iii) the distribution $p_{\hat{X}}$ be similar to p_X, so that decoded signals are perceived as genuine source signals (good perceptual quality).

& Simoncelli, 2005), which have been shown to correlate well with human opinion scores. It is also the principle underlying GAN-based image restoration schemes, which achieve enhanced perceptual quality by directly minimizing some divergence $d(p_X, p_{\hat{X}})$ (Ledig et al., 2017; Pathak et al., 2016; Isola et al., 2017; Wang et al., 2018). Based on these works, and following (Blau & Michaeli, 2018), we define the perceptual quality index (lower is better) of an algorithm as

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

where $d(\cdot,\cdot)$ is some divergence between distributions[2] (e.g., Kulback-Leibler, Wasserstein, etc.). Note that the divergence function $d(\cdot,\cdot)$ which best relates to human perception is a subject of ongoing research. Yet, our results below hold for (nearly) any divergence.

Obviously, perceptual quality, as defined above, is very different from distortion. In particular, minimizing the perceptual quality index does not necessarily lead to low distortion. For example, if the decoder disregards the input, and outputs random samples from the source distribution p_X, it will achieve perfect perceptual quality but very poor distortion. It turns out that this is true also in the other direction. That is, minimizing distortion does not necessarily lead to good perceptual quality. This observation has been studied in (Blau & Michaeli, 2018) in the specific context of signal restoration (e.g. denoising, super-resolution). In particular, perception and distortion are fundamentally at odds with each other (for non-invertible degradations), in the sense that optimizing one always comes on the expense of the other. This behavior, coined *the perception-distortion tradeoff*, was shown to hold true for *any* distortion measure.

[2]We assume that $d(p,q) \geq 0, d(p,q) = 0 \Leftrightarrow p = q$.

3. The Rate-Distortion-Perception Tradeoff

Since both perceptual quality and distortion are typically important, here we extend the rate-distortion function (2) to take into account the perception index[3] (3).

Definition 1 *The (information) rate-distortion-perception function is defined as*

$$R(D,P) = \min_{p_{\hat{X}|X}} \; I(X, \hat{X}) \quad \text{s.t. } \mathbb{E}[\Delta(X, \hat{X})] \leq D, \; d(p_X, p_{\hat{X}}) \leq P. \tag{4}$$

Unfortunately, closed form solutions for (4) are even harder to obtain than for (2). Yet, one notable exception is the classical case study of a binary source, as we show next. While of limited applicability, this example illustrates the typical behavior of (4), which we analyze in Sec. 3.2.

3.1. Bernoulli Source

Consider the problem of encoding a binary source $X \sim \mathrm{Bern}(p)$, where the decoder's output $\hat{X}$ is also constrained to be binary. Let us take the distortion measure $\Delta(\cdot,\cdot)$ to be the Hamming distance, and the perception index[4] to be the total-variation (TV) distance $d_{\mathrm{TV}}(\cdot,\cdot)$. Without loss of generality, we assume that $p \leq \frac{1}{2}$. When perception is not constrained (i.e., $P = \infty$), the solution to (4) reduces to the rate-distortion function (2) of a binary source, which is known to be given by

$$R(D,\infty) = \begin{cases} H_b(p) - H_b(D) & D \in [0,p) \\ 0 & D \in [p,\infty) \end{cases} \tag{5}$$

where $H_b(\alpha)$ is the entropy of a Bernoulli random variable with probability α (Cover & Thomas, 2012).

In the Supplementary Material, we derive the solution for arbitrary P. It turns out that as long as the perceptual quality constraint is sufficiently loose, the solution remains the same. However, when $P \leq p$, the perception constraint in (4) becomes active whenever the distortion constraint is loose enough, from which point the function $R(\cdot,P)$ departs from $R(\cdot,\infty)$. Specifically, for $P \leq p$, we have

$$R(D,P) = \tag{6}$$

$$\begin{cases} H_b(p) - H_b(D) & D \in \mathcal{S}_1 \\ 2H_b(p) + H_b(p-P) - H_t(\frac{D-P}{2}, p) - H_t(\frac{D+P}{2}, q) & D \in \mathcal{S}_2 \\ 0 & D \in \mathcal{S}_3 \end{cases}$$

[3]Similarly to (2), $R(D,P)$ in (4) lower bounds the best achievable rate for an iid source (see Supplementary Material). We do not prove achievability of $R(D,P)$ in general. However for the MSE distortion, we show an achievable upper bound (see Theorem 2).

[4]The term "perception" is somewhat inappropriate for a Bernoulli source, as it is not *perceived* by humans (contrary to images, audio). Yet, we keep this terminology here for consistency.

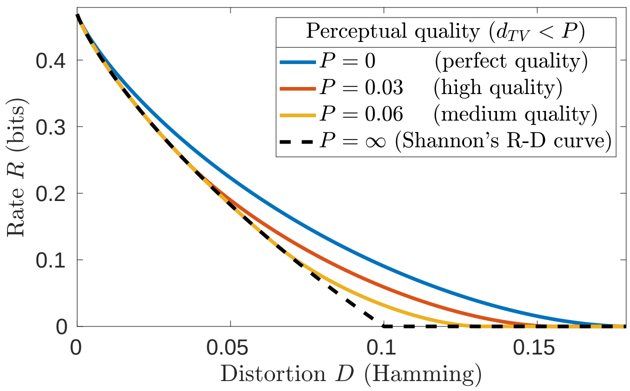


Figure 3. **Perception constrained rate-distortion curves for a Bernoulli source.** Shannon's rate-distortion function (dashed curve) characterizes the best achievable rate under any prescribed distortion level, yet does not ensure good perceptual quality. When constraining the perceptual quality index $d_{\mathrm{TV}}(p_X, p_{\hat{X}})$ to be low (good quality), the rate-distortion function elevates (solid curves). This indicates that good perceptual quality must come at the cost of a higher rate and/or a higher distortion. Here $X \sim \text{Bern}(\frac{1}{10})$.

where $q = 1 - p$ and $H_t(\alpha, \beta)$ denotes the entropy of a ternary random variable with probabilities α, β, $1 - \alpha - \beta$. Here, $\mathcal{S}_1 = [0, D_1)$, $\mathcal{S}_2 = [D_1, D_2)$, and $\mathcal{S}_3 = [D_2, \infty)$, where $D_1 = \frac{P}{1-2(p-P)}$ and $D_2 = 2pq - (q - p)P$.

Figure 3 plots $R(D, P)$ as a function of D for several values of P. As can be seen, at $D = 0$, all the curves merge. This is because at this point $\hat{X} = X$ (lossless compression), so that $p_{\hat{X}} = p_X$, and thus the perceptual quality is perfect. Yet, as the allowed distortion D grows larger, the curves depart. This illustrates that achieving the classical rate-distortion curve (black dashed line) does not generally lead to good perceptual quality. The more stringent our prescribed perceptual quality constraint (lower P), the more the rate-distortion curve elevates (colored curves). In particular, the tradeoff becomes severe at the low bit rate regime, where good perceptual quality comes at the cost of a significantly higher distortion and/or bit rate. Notice that it is possible to achieve *perfect* perceptual quality at every rate (blue curve) by compromising the distortion to some extent. In Sec. 3.2 we provide an upper-bound on the increase in distortion required for obtaining perfect perceptual quality.

While Fig. 3 displays cross-sections of $R(D, P)$ along rate-distortion planes, in Fig. 4 we plot $R(D, P)$ as a surface in 3 dimensions, as well as its cross-sections along the other planes. The equi-rate level sets shown on the surface in Fig. 4(a), provide another visualization for the phenomenon described above. That is, at high bit-rates, it is possible to achieve good perceptual quality (low P) without a significant sacrifice in the distortion D. However, as the bit-rate becomes lower, the equi-rate level sets substantially curve towards the low P values, illuminating the exacerbation in the tradeoff between distortion and perception in this regime. Figure 4(b) provides an additional viewpoint, by showing perception-distortion curves for different bit rates. Notice again that the tradeoff between distortion and perceptual quality becomes stronger at low bit-rates. Finally, Fig. 4(c) shows the somewhat counter-intuitive tradeoff between rate and perceptual quality as a function of distortion. Specifically, we see that at every *constant* distortion level, the perceptual quality can be improved by increasing the rate.

3.2. Theoretical Properties

For general source distributions, it is usually impossible to solve (4) analytically. However, it turns out that the behavior we saw for a Bernoulli source is quite typical. We next prove several general properties of the function (4), which hold under rather mild assumptions. Specifically, we assume:

A1 The divergence $d(\cdot, \cdot)$ in (4) is convex in its second argument. That is, for any $\lambda \in [0, 1]$ and for any three distributions p_0, q_1, q_2,

$$d(p_0, \lambda q_1 + (1-\lambda) q_2) \leq \lambda d(p_0, q_1) + (1-\lambda) d(p_0, q_2). \quad (7)$$

A2 The function $k(z) = \mathbb{E}_{X \sim p_X}[\Delta(X, z)]$ is not constant[5] over the entire support of p_X.

Assumption **A1** is not very limiting. For instance, any f-divergence (e.g. KL, TV, Hellinger, $\mathcal{X}^2$) as well as the Renyi divergence, satisfies this assumption (Csiszár et al., 2004; Van Erven & Harremos, 2014). Assumption **A2** holds in any setting where the mean distance between a "valid" signal z and all other "valid" signals is not constant[6]. In particular, it holds for any distortion function $\Delta(\cdot, \cdot)$ with a *unique* minimizer, such as the squared-error distortion and the SSIM index (under some assumptions) (Brunet, 2012). Using these assumptions, we are able to qualitatively characterize the general shape of the function $R(D, P)$.

Theorem 1 *The rate-distortion-perception function (4):*

1. *is monotonically non-increasing in D and P;*
2. *is convex if **A1** holds;*
3. *satisfies $R(\cdot, 0) \neq R(\cdot, \infty)$ if **A2** holds.*

The proof of Theorem 1 can be found in the Supplementary Material. Note that when assumption **A2** holds, properties 1 and 3 indicate that there exists some D_0 for which $R(D_0, 0) > R(D_0, \infty)$, showing that the rate-distortion curve necessarily elevates when constraining for perfect perceptual quality. In any case, assumption **A2** is a *sufficient* condition for property 3, so that even if it does not hold, this does not necessarily imply that $R(\cdot, 0) = R(\cdot, \infty)$.

[5]In fact, we only need the weaker condition that $k(z)$ do not attain its minimum over the entire support of p_X.

[6]A valid signal is any $x : p_X(x) > 0$. Also, we use "distance" here for clarity, although $\Delta(\cdot, \cdot)$ is not necessarily a metric.

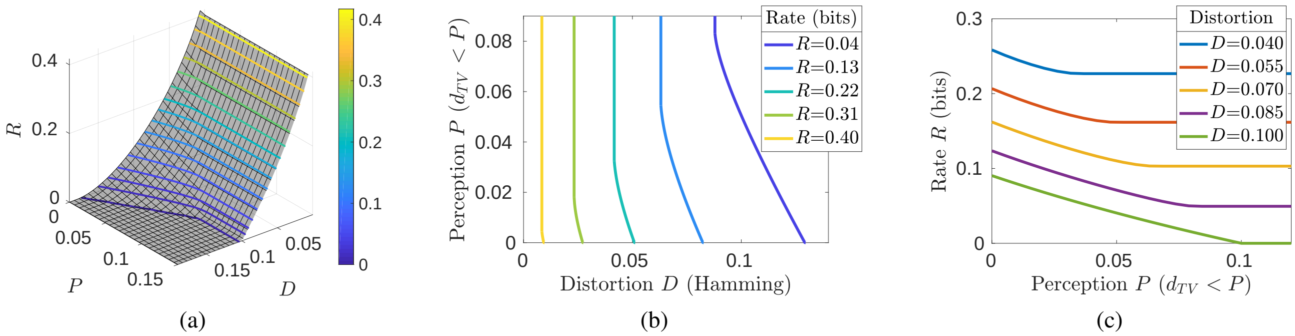


Figure 4. **The rate-distortion-perception function of a Bernoulli source.** (a) Equi-rate level sets depicted on the rate-distortion-perception function $R(D, P)$. At low bit-rates, the equi-rate lines curve substantially when approaching $P = 0$, displaying the increasing tradeoff between distortion and perceptual quality. (b) Cross sections of $R(D, P)$ along perception-distortion planes. Notice the tradeoff between perceptual quality and distortion, which becomes stronger at low bit-rates. (c) Cross sections of $R(D, P)$ along rate-perception planes. Note that at *constant* distortion, the perceptual quality can be improved by increasing the rate.

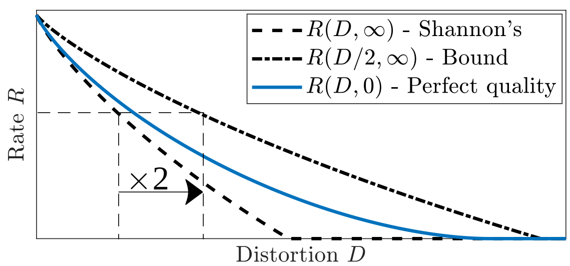


Figure 5. **Illustration of Theorem 2.** When using the MSE distortion, the rate-distortion curve for compression with perfect perceptual quality (blue) is higher than Shannon's rate-distortion function (black dashed line) but is necessarily lower than the $2\times$ scaled version of Shannon's function (dotted line).

How much does the rate-distortion curve elevate when constraining for *perfect* perceptual quality? The next theorem upper-bounds this elevation for the MSE distortion (see proof in the Supplementary Material).

Theorem 2 *When using the squared-error distortion, the function $R(\cdot, 0)$ (rate-distortion at perfect perceptual quality) is bounded by*

$$R(D, 0) \leq R(\tfrac{1}{2}D, \infty). \tag{8}$$

Theorem 2 shows that it is possible to attain perfect perceptual quality without increasing the rate, by sacrificing no more than a 2-fold increase in the mean squared-error (MSE). More specifically, attaining perfect perceptual quality at distortion D does not require a higher bit rate than that necessary for compression at distortion $\frac{1}{2}D$ with no perceptual quality constraint. This is illustrated in Fig. 5, where the perfect-quality curve $R(\cdot, 0)$ shown in blue is bounded by the scaled version of Shannon's unconstrained quality curve $R(\cdot, \infty)$ shown as a black dashed line. In image restoration scenarios, such a 2-fold increase in the MSE (3dB decrease in PSNR) has been shown to enable a substantial improvement in perceptual quality by practical algorithms (Blau et al., 2018; Ledig et al., 2017). Note that this bound is generally not tight. Thus, in some settings, perfect perceptual quality can be obtained with an even smaller increase in distortion.

4. Experimental Illustration

We now turn to demonstrate the visual implications of the rate-distortion-perception tradeoff in lossy image compression on a toy MNIST example. We make no attempt to propose a new state-of-the-art compression method. Our sole goal is to systematically explore the effect of the balance between rate, distortion, and perception. To this end, we utilize a net-based encoder-decoder pair trained in an end-to-end fashion, similarly to recent works. By tuning the influence of each of the different terms of the loss, we can easily control the balance between these three quantities.

More concretely, we use an encoder f and a decoder g, both parametrized by deep neural nets (DNNs). The encoder maps the input x into a latent feature vector $f(x)$, whose entries are then uniformly quantized to L levels to obtain the representation $\hat{f}(x)$. The decoder outputs a reconstruction $\hat{x} = g(\hat{f}(x))$. To enable back-propagation through the quantizer, we use the differentiable relaxation of (Mentzer et al., 2018). Note that this relaxation affects only the gradient computation through the quantizer during back-propagation, but not the forward-pass "hard" quantization.

As in recent perceptual-quality driven lossy compression schemes (Tschannen et al., 2018; Agustsson et al., 2018), the rate is controlled by the dimension dim of the encoder's output $f(x)$, and the number of levels L used for quantizing each of its entries, such that $R \leq dim \times \log_2(L)$. Note that

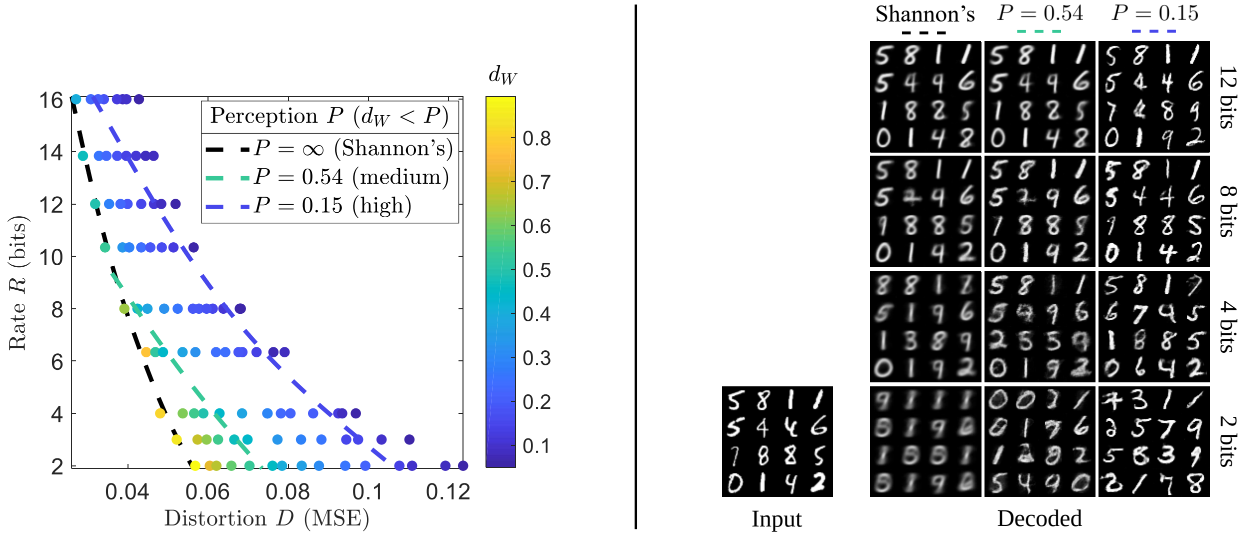


Figure 6. **Perceptual lossy compression of MNIST digits.** *Left*: Shannon's rate-distortion curve (black) describes the lowest possible rate (bits per digit) as a function of distortion, but leads to low perceptual quality (high d_W values), especially at low rates. When constraining the perceptual quality to be good (low P values), the rate-distortion curve elevates, indicating that this comes at the cost of a higher rate and/or distortion. *Right*: Encoder-decoder outputs along Shannon's rate-distortion curve and along two equi-perceptual-quality curves. As the rate decreases, the perceptual quality along Shannon's curve degrades significantly. This is avoided when constraining the perceptual quality, which results in visually pleasing reconstructions, even at extremely low bit-rates. Notice that increased perceptually quality does not imply increased accuracy, as most reconstructions fail to preserve the digits' identities at a 2-bit rate.

this only upper-bounds the best achievable rate, as lossless compression of $\hat{f}(x)$ would potentially further reduce the representation's size. However, it significantly simplifies the scheme, and was found to be only slightly sub-optimal (Agustsson et al., 2018).

For any fixed rate, we train the encoder-decoder to minimize a loss comprising a weighted combination of the expected distortion and the perception index,

$$\mathbb{E}[\Delta(X, g(\hat{f}(X)))] + \lambda d_{\mathrm{W}}(p_X, p_{\hat{X}}). \tag{9}$$

Here, $\hat{X} = g(\hat{f}(X))$, $d_{\mathrm{W}}(\cdot, \cdot)$ is the Wasserstein distance, and λ is a tuning parameter, which we use to control the balance between perception and distortion. The perceptual quality term can be optimized with the framework of generative adversarial networks (GAN) (Goodfellow et al., 2014) by introducing an additional (discriminator) DNN $h : \mathcal{X} \to \mathbb{R}$ and minimizing

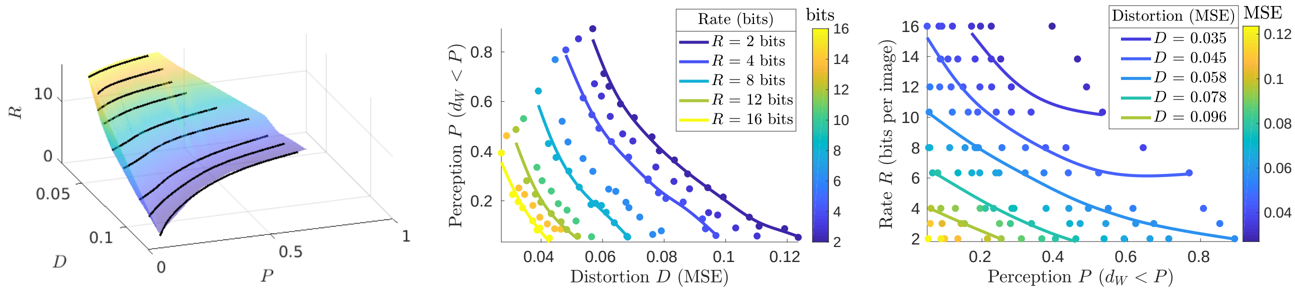
$$\mathbb{E}[\Delta(X, g(\hat{f}(X)))] + \lambda \max_{h \in \mathcal{F}} \Big(\mathbb{E}[h(X)] - \mathbb{E}[h(g(\hat{f}(X)))] \Big) \tag{10}$$

where in our specific case of a Wasserstein GAN (Arjovsky et al., 2017), $\mathcal{F}$ denotes the class of bounded 1-Lipschitz functions. As usual, all expectations are replaced by sample means, the constraint $h \in \mathcal{F}$ is replaced by a gradient penalty (Gulrajani et al., 2017), and the loss is minimized by alternating between minimization w.r.t. f, g while holding h fixed and maximization w.r.t. h while holding f, g fixed.

To achieve good perceptual quality, especially at low rates, it is essential that the decoder be stochastic (Tschannen et al., 2018). This is commonly carried out by an additional random noise input. Yet, deep generative models in the conditional setting tend to ignore this type of stochasticity (Zhu et al., 2017a;b; Mathieu et al., 2016). Tschannen et al. (2018) remedy this by applying a two-stage training scheme, which indeed promotes the use of stochasticity within the decoder, but can lead to sub-optimal results. Here, instead of concatenating a noise vector n to the encoder's output $\hat{f}(x)$, we *add* it, so that the decoder in fact operators on the noisy representation $\hat{f}(x) + n$. This *does not* lead to loss of information, as the noise n is drawn from a uniform distribution $U(-\frac{\alpha}{2}, \frac{\alpha}{2})$, with α smaller than the quantization bin size. Thus, different coded representations $\hat{f}(x)$ do not "mix-up", and can always be distinguished from one another. This scheme urges the decoder to utilize the stochastic input, while allowing end-to-end training in a one-step manner.

4.1. Squared-Error Distortion

We begin by experimenting with the squared-error distortion $\Delta(x, \hat{x}) = \|x - \hat{x}\|^2$. We train 98 encoder-decoder pairs on the MNIST handwritten digit dataset (LeCun et al., 1998), while varying the encoder's output dimension dim and number of quantization levels L to control the rate R, and the tuning coefficient λ to achieve different balances between distortion and perceptual quality. A list of all combinations



(a) (b) (c)

Figure 7. **The rate-distortion-perception function of MNIST images.** (a) Equi-rate lines plotted on $R(D, P)$ highlight the tradeoff between distortion and perceptual quality at any constant rate. (b) Cross sections of $R(D, P)$ along perception-distortion planes show that this tradeoff becomes stronger at low bit-rates. (c) Cross-sections of $R(D, P)$ along rate-perception planes highlght that at any *constant* distortion, the perceptual quality can be improved by increasing the rate.

of (dim, L, λ) used, along with all other training details can be found in the Supplementary Material.

The left side of Fig. 6 plots the 98 trained encoder-decoder pairs on the rate-distortion plane, with the perceptual quality indicated by color coding and rate measured in bits per digit. The perceptual quality is quantified by the final discriminator loss[7], which approximates the Wasserstein distance $d_{\text{W}}(p_X, p_{\hat{X}})$. We plot an approximation of Shannon's rate-distortion function (obtained with $\lambda = 0$), and two additional rate-distortion curves with (approximately) constant perceptual quality[8]. As can be seen, the rate-distortion curve elevates when constraining the perceptual quality to be good. This demonstrates once again that we can improve the perceptual quality w.r.t. that obtained on Shannon's rate-distortion curve, yet this must come at the cost of a higher rate and/or distortion. Notice that the perception index is not constant along Shannon's function; it increases (worse quality) towards lower bit-rates.

On the right side of Fig. 6, we depict the outputs of encoder-decoder pairs along Shannon's rate-distortion function, and along the two equi-perception curves shown on the left. It can be seen that as the rate decreases, the perceptual quality of the reconstructions along Shannon's function degrades. However, this is avoided when constraining the perceptual quality, which results in visually pleasing reconstructions even at extremely low bit-rates. Notice that this increased perceptual quality does not imply increased accuracy, as at low bit rates (e.g., 2 bits), most reconstructions fail to preserve even the identity of the digit. Yet, while the encoder-decoder pairs on Shannon's rate-distortion curve are more accurate on average, no doubt that the perceptually-constrained encoder-decoder pairs are favorable in terms of perceptual quality. Also, notice that at a rate of 2 bits, the outputs of the perceptually-constrained encoder-decoder pairs are all distinct, even though there are only 4 code words (as can be seen for Shannon's encoder-decoder). This shows that the decoder effectively utilizes the noise.

Figure 7 depicts the function $R(D, P)$ in 3-dimensions, as well as its cross sections along the other axis aligned planes. In Fig. 7(a), the *curved* equi-rate lines show the tradeoff between distortion and perceptual quality. This is also apparent in Fig. 7(b), which shows cross sections along perception-distortion planes at different rates. As can be seen, the tradeoff becomes stronger at low bit-rates. Figure 7(c) shows the counter-intuitive tradeoff between rate and perception. That is, at *constant* distortion, the perceptual quality can be improved by increasing the rate.

4.2. Advanced Distortion Measures

The peak-signal-to-noise ratio (PSNR), which is a rescaling of the MSE, is still the most common quality measure in image compression. Yet, it is well-known to be inadequate for quantifying distortion as perceived by humans (Wang & Bovik, 2009). Over the past decades, there has been a constant search for better distortion criteria, ranging from the simple SSIM/MS-SSIM (Wang et al., 2003; 2004) to the recently popular deep-feature based distortion (Johnson et al., 2016; Zhang et al., 2018). Interestingly, the perceptual quality along Shannon's classical rate-distortion function is not perfect for nearly any distortion measure (see property 3 in Theorem 1). This implies that perfect perceptual quality cannot be achieved by merely switching to more advanced distortion criteria, but rather requires directly optimizing the perception index (e.g. using GAN-based schemes). This is not to say that the function $R(D, P)$ is the same for all distortion measures. The strength of the tradeoff can

[7] $\mathcal{L}_{\text{dis}} = \frac{2}{N}\left(\sum_{i=1}^{N/2} h(x_i) - \sum_{i=N/2+1}^{N} h(g(\hat{f}(x_i)))\right)$, where $\{x_i\}_{i=1}^{N}$ are the test samples.

[8] We plot a smoothing spline calculated over the set of points which satisfy the constraint $d_{\text{W}}(p_X, p_{\hat{X}}) \leq P$ and have the minimal distortion among all points with the same rate.

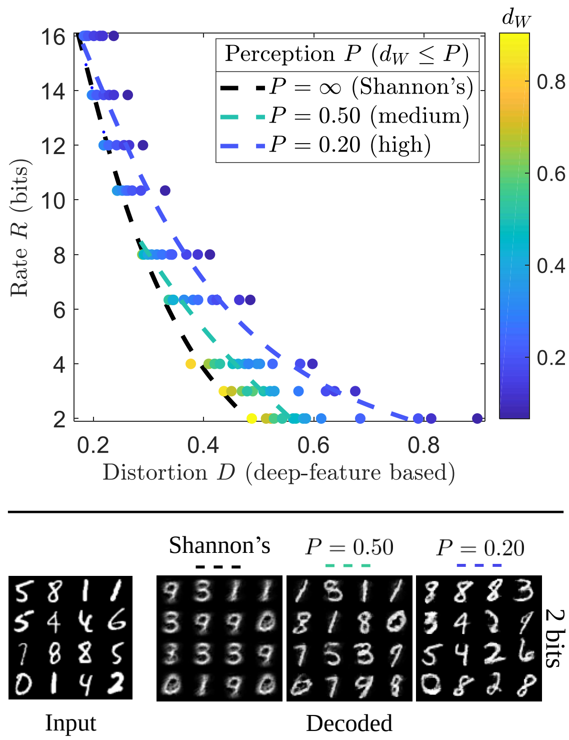


Figure 8. **Replacing the MSE with the deep feature based distortion.** We repeat the experiment of Fig. 6, while replacing the squared-error distortion with the deep-feature based distortion of Johnson et al. (2016). *Top*: The rate-distortion curves elevate when constraining the perceptual quality, demonstrating that the use of this advanced distortion measure *does not* eliminate the tradeoff. *Bottom*: Even with this popular advanced distortion, is it still beneficial to compromise distortion and constrain for improved perceptual quality. Nevertheless, the tradeoff does appear to be a bit weaker than in Fig. 6, as minimizing distortion alone (Shannon's curve) is now somewhat more pleasing visually.

certainly decrease for distortion criteria which capture more semantic similarities (Blau & Michaeli, 2018).

We demonstrate this by repeating the experiment of Sec. 4.1, while replacing the squared-error distortion by the deep-feature based distortion of (Ledig et al., 2017), i.e.,

$$\Delta(x, \hat{x}) = \|x - \hat{x}\|^2 + \alpha \|\Psi(x) - \Psi(\hat{x})\|^2, \qquad (11)$$

where $\Psi(x)$ is the output of an intermediate DNN layer for input x. Here we take the second conv-layer output of a 4-layer DNN, which we pre-trained to achieve over 99% classification accuracy on the MNIST test set. All training details appear in the Supplementary Material.

Figure 8 plots 98 encoder-decoder pairs on the rate-distortion plane, trained exactly as in Fig. 6, but this time with the loss (11) instead of MSE. As can be seen, here too the rate-distortion curves elevate when constraining the perceptual quality, demonstrating that the use of advanced distortion measures *does not* eliminate the tradeoff. From the decoded outputs, however, it is evident that the tradeoff here is somewhat weaker, as minimizing distortion alone (Shannon's) appears a bit more visually pleasing compared to Fig. 6 (though still with reduced variability and more blur than the perception constrained reconstruction).

4.3. Related Work

Our theoretical analysis and experimental validation help explain some of the observations reported in the recent literature. Specifically, a lot of research efforts have been devoted to optimizing the rate-distortion function (2) using deep nets (Toderici et al., 2016; 2017; Agustsson et al., 2017; Ballé et al., 2017; Minnen et al., 2018; Li et al., 2018). Some papers explicitly targeted high perceptual quality. One line of works did so by choosing the distortion criterion to be some advanced full-reference measure, like SSIM/MS-SSIM (Ballé et al., 2018; Mentzer et al., 2018; Johnston et al., 2018), normalized Laplacian pyramid (Ballé et al., 2016) and deformation-aware sum of squared differences (DASSD) (Rott Shaham & Michaeli, 2018). While beneficial, these methods could not demonstrate high perceptual quality at very low bit rates, which aligns with our theory. Another line of works incorporated generative models, which explicitly encourage the distribution of outputs to be similar to that of natural images (decreasing the divergence in (3)). This was done on an image patch level (Rippel & Bourdev, 2017), on reduced-size (thumbnail) images (Tschannen et al., 2018; Santurkar et al., 2018), on a full-image scale (Agustsson et al., 2018), and as a post-processing step (Galteri et al., 2017). In particular, Tschannen et al. (2018) propose a practical method for distribution-preserving compression ($P = 0$ in our terminology). These methods managed to obtain impressive perceptual quality at very low bit rates, but not without a substantial sacrifice in distortion, as predicted by our theory. Finally, we note that rate-distortion analysis (with a specific distortion) has also been used in the context of generative models (Alemi et al., 2018), which target $p_{\hat{X}} = p_X$ (i.e., $P = 0$). Our results hold for arbitrary distortions and arbitrary P.

5. Conclusion

We proved that in lossy compression, perceptual quality is at odds with rate and distortion. Specifically, any attempt to keep the statistics of decoded signals similar to that of source signals, will result in a higher distortion or rate. We characterized the triple tradeoff between rate, distortion and perception, and empirically illustrated its manifestation in image compression. Our observations suggest that comparing methods based on their rate-distortion curves alone may be misleading. A more informative evaluation must also include some (no-reference) perceptual quality measure.

Acknowledgements

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

References

Agustsson, E., Mentzer, F., Tschannen, M., Cavigelli, L., Timofte, R., Benini, L., and Gool, L. V. Soft-to-hard vector quantization for end-to-end learning compressible representations. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2017.

Agustsson, E., Tschannen, M., Mentzer, F., Timofte, R., and Van Gool, L. Generative adversarial networks for extreme learned image compression. *arXiv preprint arXiv:1804.02958*, 2018.

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. Fixing a broken elbo. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Ballé, J., Laparra, V., and Simoncelli, E. P. End-to-end optimization of nonlinear transform codes for perceptual quality. In *Proceedings of the Picture Coding Symposium (PCS)*, 2016.

Ballé, J., Laparra, V., and Simoncelli, E. P. End-to-end optimized image compression. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. Variational image compression with a scale hyperprior. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Blau, Y. and Michaeli, T. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., and Zelnik-Manor, L. The 2018 PIRM challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2018.

Brunet, D. A study of the structural similarity image quality measure with applications to image processing. 2012.

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

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

Denton, E. L., Chintala, S., Szlam, A., and Fergus, R. Deep generative image models using a laplacian pyramid of adversarial networks. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2015.

Galteri, L., Seidenari, L., Bertini, M., and Del Bimbo, A. Deep generative adversarial compression artifact removal. In *Proceedings of the Internatioanl Conference on Computer Vision (ICCV)*, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2017.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Johnston, N., Vincent, D., Minnen, D., Covell, M., Singh, S., Chinen, T., Hwang, S. J., Shor, J., and Toderici, G. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A. P., Tejani, A., Totz, J., Wang, Z., and Shi, W. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Li, M., Zuo, W., Gu, S., Zhao, D., and Zhang, D. Learning convolutional networks for content-weighted image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Mathieu, M., Couprie, C., and LeCun, Y. Deep multi-scale video prediction beyond mean square error. In *Proceedings of the International Conference on Learning Representation (ICLR)*, 2016.

Matsumoto, R. Introducing the perception-distortion tradeoff into the rate-distortion theory of general information sources. *IEICE Communications Express*, 7(11):427–431, 2018a.

Matsumoto, R. Rate-distortion-perception tradeoff of variable-length source coding for general information sources. *IEICE Communications Express*, 2018b.

Mentzer, F., Agustsson, E., Tschannen, M., Timofte, R., and Van Gool, L. Conditional probability models for deep image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Minnen, D., Ballé, J., and Toderici, G. D. Joint autoregressive and hierarchical priors for learned image compression. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2018.

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

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

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Rippel, O. and Bourdev, L. Real-time adaptive image compression. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

Rott Shaham, T. and Michaeli, T. Deformation aware image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2016.

Santurkar, S., Budden, D., and Shavit, N. Generative compression. In *Proceedings of the Picture Coding Symposium (PCS)*, 2018.

Shannon, C. E. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.

Toderici, G., O'Malley, S. M., Hwang, S. J., Vincent, D., Minnen, D., Baluja, S., Covell, M., and Sukthankar, R. Variable rate image compression with recurrent neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Toderici, G., Vincent, D., Johnston, N., Jin Hwang, S., Minnen, D., Shor, J., and Covell, M. Full resolution image compression with recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Tschannen, M., Agustsson, E., and Lucic, M. Deep generative models for distribution-preserving lossy compression. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2018.

Van Erven, T. and Harremos, P. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., and Tang, X. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2018.

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

Wang, Z. and Simoncelli, E. P. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *Human Vision and Electronic Imaging X*, volume 5666, pp. 149–160, 2005.

Wang, Z., Simoncelli, E. P., and Bovik, A. C. Multiscale structural similarity for image quality assessment. In *Proceedings of the Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003.

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

Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017a.

Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., and Shechtman, E. Toward multimodal image-to-image translation. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2017b.