UNIVERSAL RATE-DISTORTION-PERCEPTION REPRESENTATIONS FOR LOSSY COMPRESSION

George Zhang
University of Toronto
gq.zhang@mail.utoronto.ca

Jun Chen
McMaster University
chenjun@mcmaster.ca

Ashish Khisti
University of Toronto
akhisti@ece.utoronto.ca

ABSTRACT

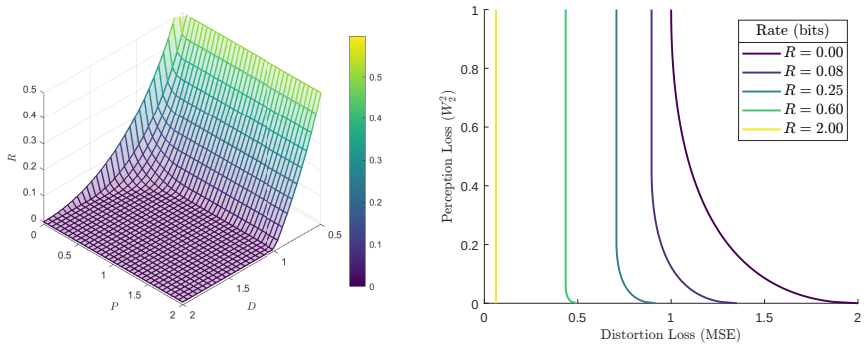
In the context of lossy compression, Blau & Michaeli (2019) adopt a mathematical notion of perceptual quality and define the rate-distortion-perception function, generalizing the classical rate-distortion tradeoff. We consider the notion of (approximately) universal representations in which one may fix an encoder and vary the decoder to (approximately) achieve any point along the perception-distortion tradeoff. We show that the penalty for fixing the encoder is zero in the Gaussian case, and give bounds in the case of arbitrary distributions, under MSE distortion and $W_2^2(\cdot,\cdot)$ perception losses. In principle, a small penalty refutes the need to design an end-to-end system for each particular objective. We provide experimental results on MNIST and SVHN to suggest that there exist practical constructions that suffer only a small penalty, i.e. machine learning models learn representation maps which are approximately universal within their operational capacities.

1 INTRODUCTION

A lossy compression system consists of an encoder producing a low-rate representation of a source and a decoder reconstructing an approximation. Unlike lossless compression, the decoder in a lossy compression system has flexibility in how it would like to reconstruct the source. It is conventional to optimize for the reconstruction minimizing some distortion measure between the original and the reconstruction such as mean squared error, PSNR or SSIM/MS-SSIM (Wang et al., 2003; 2004). Accordingly, lossy compression algorithms are analyzed through rate-distortion theory, wherein the objective is to minimize the amount of distortion for a given rate. However, low distortion is not necessarily synonymous with high perceptual quality; indeed, deep learning based image compression has inspired works in which authors have noted that increased perceptual quality may come at the cost of increased distortion (Blau & Michaeli, 2018; Agustsson et al., 2019). This culminated in the work of (Blau & Michaeli, 2019) who propose the rate-distortion-perception theoretical framework.

The main idea was to introduce a third *perception* axis which more closely mimics what humans would deem to be visually pleasing. Unlike distortion, judgement of perceptual quality is inherently no-reference; indeed, the mathematical proxy for perceptual quality is defined by a constraint between the source and reconstruction distributions. Leveraging generative adversarial networks (Goodfellow et al., 2014) in the training procedure has made such a task possible for complex data-driven settings, with efficacy even at very low rates (Tschannen et al., 2018). Naturally, this induces a tradeoff between optimizing for perceptual quality and optimizing for distortion. But in designing a lossy compression system, one may wonder where exactly this tradeoff lies: is the objective tightly coupled with optimizing the representations generated by the encoder, or can most of this tradeoff be achieved by simply changing the decoding scheme?

Our contributions are as follows. We prove that there is no loss in fixing the representation map across the perception-distortion tradeoff for the Gaussian distribution, then provide a bound on the penalty incurred for arbitrary source distributions. We perform experiments on MNIST and SVHN to demonstrate that learned representation maps can be approximately universal within their operational



(a) (b)

Figure 1: (a) $R(D, P)$ for a standard Gaussian source X. (b) Perception-distortion cross-sections across multiple rates. The tension between perception and distortion is most visible at low rates. Perception-distortion universality of Gaussian sources implies the existence of a universal representation which can be decoded to achieve any point along a perception-distortion curve.

rate-distortion-perception tradeoffs. This has important practical implications as flexibility is a highly desirable property in designing representations. It may not be plausible to design an entire end-to-end system for each potential objective in advance; our results show that representations from universal encoders are reusable across a family of lossy compression objectives without much loss.

2 Universal Representations

The process of quantizing a source $X \sim p_X$ leads to imperfect reconstruction as measured by some distortion function $\Delta : \mathbb{R} \to \mathbb{R}$. Minimizing an appropriate choice of $\Delta(\cdot, \cdot)$ should cause the output to look more like the input as perceived by humans. This methodology however does not explicitly capture how the distribution of the reconstructions compares to the distribution of the source, which Blau & Michaeli (2019) argue is a better proxy for what humans would perceive to be natural (disregarding the original source image). Accordingly, the conventional rate-distortion function is augmented with some "distance" between probability measures $d(\cdot, \cdot)$ as follows[1].

Definition 1. *The information rate-distortion-perception function for a source $X \sim p_X$ is defined as*

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

Popular choices for $d(\cdot, \cdot)$ include Wasserstein distances due to their empirical success, which we will assume henceforth unless otherwise stated. We first give a solution to the Gaussian case to provide an example of a distribution with the perception-distortion universality property.

Theorem 1. *For $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, the rate-distortion-perception function under squared error distortion and squared W_2 distance is achieved by some $\hat{X}$ jointly Gaussian with X and is given by*

$$R(D, P) = \begin{cases} \frac{1}{2} \log \frac{\sigma_X^2 (\sigma_X - \sqrt{P})^2}{\sigma_X^2 (\sigma_X - \sqrt{P})^2 - (\frac{\sigma_X^2 + (\sigma_X - \sqrt{P})^2 - D}{2})^2} \\ \qquad \text{if } \sqrt{P} < \sigma_X - \sqrt{|\sigma_X^2 - D|}, \\ \max\{\frac{1}{2} \log \frac{\sigma_X^2}{D}, 0\} \qquad \text{if } \sqrt{P} \geq \sigma_X - \sqrt{|\sigma_X^2 - D|}. \end{cases}$$

Due to joint Gaussianity of $(X, \hat{X})$, this expression admits the equivalent form

$$R(D, P) = \frac{1}{2} \log \frac{\sigma_X^2}{\mathbb{E}[(X - \mathbb{E}[X|\hat{X}])^2]}, \tag{1}$$

which describes the rate-distortion-perception function in terms of the minimum mean squared error estimator for X given $\hat{X}$. In principle, different (D, P) pairs are associated with different encoding schemes for X. One may use (1) to show that different (D, P) pairs associated with the same R can be achieved by linearly scaling the representation $\hat{X}$ produced by a single encoder. This representation

[1] We assume some rudimentary properties about $\Delta(\cdot, \cdot)$ and $d(\cdot, \cdot)$ given in the appendix.

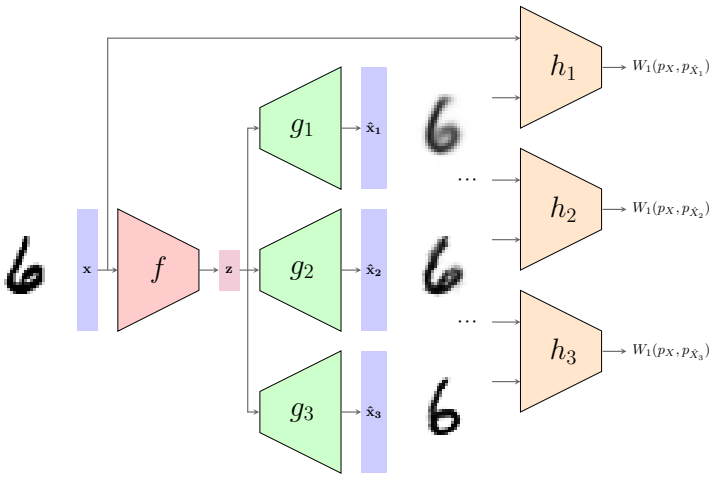


Figure 2: Schematic of the universal model. A single encoder f is trained with an initial tradeoff λ and has weights frozen. Subsequently other decoders $\{g_i\}$ are trained for different λ_i using the representations z produced by f. Separate critic networks $\{h_i\}$ are trained along with each decoder to promote perceptual quality. The top decoder places most weight on distortion loss whereas the bottom decoder places most weight on perception loss. This reduces the blurriness but produces a less faithful reconstruction of the original. Perception losses are estimated using the critics $\{h_i\}$ by replacing the expectations in Equation (2) with test samples.

is *universal* in the sense that it can achieve optimality across all perception and distortion tradeoffs for a given rate; to move between the points, it is sufficient to modify only the decoding scheme.

It turns out that general distributions satisfy an approximate version of universality. Let $X \sim p_X$ be a source and $\hat{X}^{(1)}$ be an optimal representation in the conventional rate-distortion framework (i.e., $I(X; \hat{X}^{(1)}) = R(D_1, P_1)$, $\mathbb{E}[(X - \hat{X}^{(1)})^2] \leq D_1$ for some $D_1 \in (0, \sigma_X^2)$ and $P_1 = \infty$).

Theorem 2. *(Approximate universality for general sources) Under MSE and $W_2^2(\cdot, \cdot)$, $\hat{X}^{(1)}$ can meet distortion constraint $D_2 = 2D_1$ and any perception constraint P_2 via suitable decoding.*

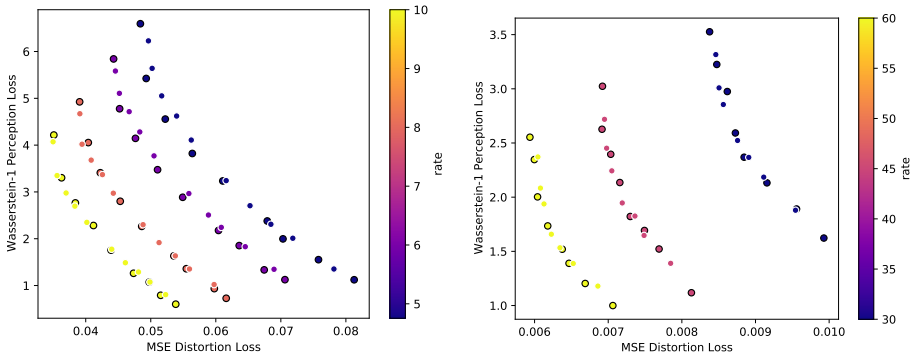
This result implies that the optimal representation in the conventional rate-distortion sense can be transformed to meet any perception constraint possibly with a distortion penalty of no more than $2D_1 - D_2$. It is worth noting that this special case gives rise to Theorem 2 ($R(D, \infty) \geq R(2D, 0)$) in Blau & Michaeli (2019). Namely, the numerical connection between $R(D, \infty)$ and $R(2D, 0)$ is a manifestation of the existence of approximately universal representations.

3 Experimental Results

The rate-distortion-perception tradeoff was observed in the context of GAN-enhanced image compression (Blau & Michaeli, 2019), while training an entire model for each desired setting over distortion and perception. In practice, it is undesirable to develop an entire system from scratch for each objective. Our theoretical results have already established a bound on the loss of optimality for using universal representations. We provide experimental evidence to show that the loss is small in practice.

Concretely, in an end-to-end model the encoder and decoder are trained jointly for an objective so as to utilize the flexibility in learning a representation map whereas in a universal model, the representation map is fixed so only the decoder can be trained. The encoder for a universal model may be re-used across many models and be designed separately from the decoders; we refer to it as a universal encoder. The architecture used in the experiments is a stochastic autoencoder augmented with a Wasserstein GAN and follows closely the design of (Blau & Michaeli, 2019).

The task of generating both the end-to-end and universal models is broken into two stages. The initial stage trains an end-to-end model, consisting of an encoder f producing latent representation $z = f(x)$ and decoder g producing reconstruction $\hat{x} = g(z)$. Here f incorporates both the deterministic quantized encoding step and the addition of stochastic noise u to support high perceptual quality. In practice, the quantized versions of the representations are used for compression then u can be added before decoding. For the distortion metric we use MSE and for the perception metric we use



(a) (b)

Figure 3: (a) MNIST. (b) SVHN. Rate-distortion-perception tradeoffs along various rates. Points with black outline are losses reported for the end-to-end encoder-decoder pairs trained jointly for a particular perception-distortion objective. The other points are the losses for universal models, in which decoders are trained over a frozen encoder optimized for small P (MNIST: $\lambda = 0.015$, SVHN: $\lambda = 0.002$). Universal model performance is very close to performance of end-to-end models across all tradeoffs $\{\lambda_i\}$. Example images are provided in the supplementary.

Wasserstein-1 distance. The objective is a weighted sum between the terms given by

$$L = \mathbb{E}[\|X - \hat{X}\|^2] + \lambda W_1(p_X, p_{\hat{X}}),$$

for some initial λ to be chosen. W_1 is estimated using the Kantorovich-Rubinstein dual form

$$W_1(p_X, p_{\hat{X}}) = \max_{h \in \mathcal{F}} \mathbb{E}[h(X)] - \mathbb{E}[h(g(f(X)))]. \tag{2}$$

Here, $\mathcal{F}$ is the set of all bounded 1-Lipschitz functions. In practice, we use a critic network as h and the Lipschitz condition is approximated with a gradient penalty (Gulrajani et al., 2017). Optimization alternates between minimizing f, g with h fixed and maximizing h with f, g fixed. In essence, g is optimized for both low distortion and high perception, so it acts as both a decoder and a generator.

Afterwards, the encoders of end-to-end models can be used to construct universal models. To do the latter, we freeze the parameters of f and introduce a new decoder g_1 and critic h_1 trained to minimize

$$L_1 = \mathbb{E}[\|X - \hat{X}_1\|^2] + \lambda_1 W_1(p_X, p_{\hat{X}_1}),$$

where λ_1 is a different tradeoff parameter and $p_{\hat{X}_1}$ is the distribution induced by pushing X through $g_1 \circ f$. The rest of the training procedure follows that of the first stage. This is repeated over many different parameters $\{\lambda_i\}$ to generate a tradeoff curve.

Figure 3 shows rate-distortion-perception curves at multiple rates on MNIST and SVHN. Note that the rate for each individual curve is fixed through using the same quantizer across all models. The rates are chosen to be low so that the tension between distortion and perception is most visible. Points outlined in black are losses for end-to-end models, while the other points correspond to the universal models sharing an encoder. The universal models are able to achieve a tradeoff which is very close to the end-to-end models (with outputs that are visually comparable).

For any fixed rate, decreasing the perception loss P reduces the output blur, but the reconstruction is less faithful to the original input. As noted by (Blau & Michaeli, 2019), this is especially evident at very low rates in which the compression system appears to act as a generative model. However, this does not necessarily come at the expense of the representation map learned by the encoder, i.e. the embeddings generated by an encoder trained for small P can also be used to produce a low-distortion reconstruction by training a new decoder. Conversely, training a decoder to produce reconstructions with high perceptual quality on top of an encoder trained only for distortion loss is also possible as the decoder is sufficiently expressive to act purely as a generative model.

4 Discussion

The use of deep generative models in data compression has highlighted the tradeoff between optimizing for low distortion and high perceptual quality. With a suitable encoding scheme, we have shown that modifying the decoder is sufficient to achieve this tradeoff in practice. Future directions include one-shot theoretical analysis, and employing the scheme to high-resolution images and videos.

REFERENCES

Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 221–231, 2019.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Yochai Blau and Tomer Michaeli. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pp. 675–685, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. volume 27, pp. 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pp. 5767–5777, 2017.

Dong Liu, Haochen Zhang, and Zhiwei Xiong. On the classification-distortion-perception tradeoff. In *Advances in Neural Information Processing Systems*, pp. 1206–1215, 2019.

Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4394–4402, 2018.

Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Lucas Theis and Eirikur Agustsson. On the advantages of stochastic encoders. *arXiv preprint arXiv:2102.09270*, 2021.

Michael Tschannen, Eirikur Agustsson, and Mario Lucic. Deep generative models for distribution-preserving lossy compression. In *Advances in Neural Information Processing Systems*, pp. 5929–5940, 2018.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402. IEEE, 2003.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

A THEORETICAL RESULTS

Definition 1. Let $\Delta(\cdot,\cdot)$ be a non-negative distortion measure satisfying $\Delta(x,y)=0$ if and only if $x=y$, and $d(\cdot,\cdot)$ be a (non-negative) divergence between probability measures, likewise with $d(p,q)=0$ if and only if $p=q$. The information rate-distortion-perception function for a source $X\sim p_X$ is defined as

$$\begin{aligned} R(D,P) = &\min_{p_{\hat{X}|X}} \ I(X;\hat{X}) \\ &\text{s.t. } \ \mathbb{E}[\Delta(X,\hat{X})]\le D, \quad d(p_X,p_{\hat{X}})\le P. \end{aligned}$$

We let $R(D)=R(D,\infty)$ denote the conventional rate-distortion function. We assume the use of MSE distortion loss and squared Wasserstein-2 perception loss in the theoretical results unless otherwise stated. Recall that

$$W_2^2(p_X,p_{\hat{X}}) = \inf \mathbb{E}[\|X-\hat{X}\|^2],$$

where the infimum is over all joint distributions of $(X,\hat{X})$ with marginals p_X and $p_{\hat{X}}$. When p_X and $p_{\hat{X}}$ are both Gaussian, $W_2^2(p_X,p_{\hat{X}}) = (\mu_X-\mu_{\hat{X}})^2+(\sigma_X-\sigma_{\hat{X}})^2$.

A.1 PROOF OF THEOREM 1

Theorem 1. For $X\sim\mathcal{N}(\mu_X,\sigma_X^2)$, the rate-distortion-perception function under squared error distortion and squared W_2 distance is achieved by some $\hat{X}$ jointly Gaussian with X and is given by

$$R(D,P) = \begin{cases} \frac{1}{2}\log \frac{\sigma_X^2(\sigma_X-\sqrt{P})^2}{\sigma_X^2(\sigma_X-\sqrt{P})^2-(\frac{\sigma_X^2+(\sigma_X-\sqrt{P})^2-D}{2})^2} \\ \qquad\qquad\qquad\qquad \text{if } \sqrt{P}<\sigma_X-\sqrt{|\sigma_X^2-D|}, \\ \max\{\frac{1}{2}\log\frac{\sigma_X^2}{D},0\} \qquad \text{if } \sqrt{P}\ge\sigma_X-\sqrt{|\sigma_X^2-D|}. \end{cases}$$

We first restate a useful result from estimation theory. Let $\hat{X}$ be a random variable with $\mathbb{E}[\hat{X}]=\mu_{\hat{X}}$, $\mathrm{Var}(\hat{X})=\sigma_{\hat{X}}^2$ and $\mathrm{Cov}(X,\hat{X})=\theta$. Let $\hat{X}_G$ be a random variable jointly Gaussian with X with the same first and second order statistics as $\hat{X}$.

Lemma 1. *Given $\mu_{\hat{X}}$, $\sigma_{\hat{X}}^2$, and θ, we have that*

$$\mathbb{E}[(X-\mathbb{E}[X|\hat{X}_G])^2] \ge \mathbb{E}[(X-\mathbb{E}[X|\hat{X}])^2].$$

Proof of Theorem 1. We shall show that there is no loss of optimality in assuming that $\hat{X}$ is jointly Gaussian with X. It is clear that $\mathbb{E}[(X-\hat{X})^2]=\mathbb{E}[(X-\hat{X}_G)^2]$, as the first and second order statistics are all given. Moreover, since every coupling of p_X and $p_{\hat{X}}$ induces a Gaussian coupling of p_X and $p_{\hat{X}_G}$ with the same covariance, it follows that

$$W_2^2(p_X,p_{\hat{X}}) \ge W_2^2(p_X,p_{\hat{X}_G}).$$

Finally, we have

$$\begin{aligned} I(X;\hat{X}) &= h(X)-h(X|\hat{X}) \\ &\ge h(X)-h(X-\mathbb{E}[X|\hat{X}]) \\ &\overset{(a)}{\ge} h(X)-\frac{1}{2}\log(2\pi e\mathbb{E}[(X-\mathbb{E}[X|\hat{X}])^2]) \\ &\overset{(b)}{\ge} h(X)-\frac{1}{2}\log(2\pi e\mathbb{E}[(X-\mathbb{E}[X|\hat{X}_G])^2]) \\ &= h(X)-h(X-\mathbb{E}[X|\hat{X}_G])) \\ &\overset{(c)}{=} h(X)-h(X|\hat{X}_G) \\ &= I(X;\hat{X}_G), \end{aligned}$$

where (a) is because the Gaussian distribution maximizes differential entropy for a given variance, (b) follows from Lemma 1 and (c) is because the estimation error is independent of $\hat{X}_G$. Thus, it suffices to solve the problem

$$R(D,P) = \min_{p_{\hat{X}_G|X}} I(X;\hat{X}_G)$$
$$\text{s.t.} \quad \mathbb{E}[(X-\hat{X}_G)^2] \leq D, \quad W_2^2(p_X, p_{\hat{X}_G}) \leq P.$$

Note that

$$\mathbb{E}[(X-\hat{X}_G)^2] = (\mu_X - \mu_{\hat{X}})^2 + \sigma_X^2 + \sigma_{\hat{X}}^2 - 2\theta,$$

so there is no loss of optimality in assuming $\mu_{\hat{X}} = \mu_X$. Accordingly,

$$I(X;\hat{X}_G) = \frac{1}{2}\log\frac{\sigma_X^2\sigma_{\hat{X}}^2}{\sigma_X^2\sigma_{\hat{X}}^2 - \theta^2},$$
$$W_2^2(p_X, p_{\hat{X}_G}) = (\sigma_X - \sigma_{\hat{X}})^2.$$

When $\sqrt{P} < \sigma_X - \sqrt{|\sigma_X^2 - D|}$, both P and D are active, and consequently we have $\sigma_{\hat{X}}^2 = (\sigma_X - \sqrt{P})^2$ and $\theta = \frac{\sigma_X^2 + \sigma_{\hat{X}}^2 - D}{2}$. Noting that the other case is simply the solution to $R(D)$, this concludes the proof. □

Remark. The proof of Theorem 1 can be easily modified to handle to the case $d(p_X, p_{\hat{X}}) = \mathrm{KL}(p_X, p_{\hat{X}})$, where $\mathrm{KL}(p_X, p_{\hat{X}}) = \int p_{\hat{X}}(x)\log\frac{p_{\hat{X}}(x)}{p_X(x)}\mathrm{d}x$ is the KL-divergence between p_X and $p_{\hat{X}}$. Note that given $(\mu_{\hat{X}}, \sigma_{\hat{X}}^2)$, $\mathrm{KL}(p_X, p_{\hat{X}})$ is minimized when $p_{\hat{X}}$ is a Gaussian distribution. Given p_X, one may see that there is a one-to-one correspondence between

$$\mathrm{KL}(p_X, p_{\hat{X}_G}) = \frac{\sigma_{\hat{X}}^2 - \sigma_X^2}{2\sigma_X^2} + \frac{1}{2}\log\frac{\sigma_X^2}{\sigma_{\hat{X}}^2},$$
$$W_2^2(p_X, p_{\hat{X}_G}) = (\sigma_X - \sigma_{\hat{X}})^2$$

which implies the rate-distortion-perception functions under $\mathrm{KL}(p_X, \cdot)$ and $W_2^2(p_X, \cdot)$ also share a one-to-one correspondence in P.

Remark. One may see that the Gaussian distribution is universally representable as follows. Let $\hat{X}_1$ and $\hat{X}_2$ be the induced outputs associated with (D_1, P_1) and (D_2, P_2), respectively. Note that we have equality of first moments, $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|\hat{X}_1]] = \mathbb{E}[\mathbb{E}[X|\hat{X}_2]]$. Since X is jointly Gaussian with $\hat{X}_1$, it follows from standard facts that (a) $\mathbb{E}[X|\hat{X}_1]$ is independent of the estimation error $X - \mathbb{E}[X|\hat{X}_1]$, and (b) $\mathbb{E}[X|\hat{X}_1]$ is a linear function of $\hat{X}_1$ (and thus is itself Gaussian). The same is true for $\hat{X}_2$. Furthermore, after simplifying $R(D_1, P_1) = R(D_2, P_2)$ we get that

$$\mathbb{E}[(X - \mathbb{E}[X|\hat{X}_1])^2] = \mathbb{E}[(X - \mathbb{E}[X|\hat{X}_2])^2]. \tag{3}$$

One may expand (3) and apply (a) to obtain equality of the cross terms, i.e. $\mathbb{E}[X\mathbb{E}[X|\hat{X}_1]] = \mathbb{E}[X\mathbb{E}[X|\hat{X}_2]]$. Plugging this back into (3) shows that $\mathbb{E}[\mathbb{E}[X|\hat{X}_1]^2] = \mathbb{E}[\mathbb{E}[X|\hat{X}_2]^2]$. Since Gaussians are uniquely parameterized by their first and second moments, $(X, \mathbb{E}[X|\hat{X}_1])$ and $(X, \mathbb{E}[X|\hat{X}_2])$ must be identically distributed. From (b), $\mathbb{E}[X|\hat{X}_1]$ and $\mathbb{E}[X|\hat{X}_2]$ are linear in $\hat{X}_1$ and $\hat{X}_2$, respectively, so $\hat{X}_1$ and $\hat{X}_2$ must themselves be linearly related.

A.2 Proof of Theorem 2

Recall that $X \sim p_X$ is an information source and $\hat{X}^{(1)}$ is an optimal representation in the conventional rate-distortion framework (i.e., $I(X;\hat{X}^{(1)}) = R(D_1, P_1)$ and $\mathbb{E}[(X - \hat{X}^{(1)})^2] \leq D_1$ for some $D_1 \in (0, \sigma_X^2)$ and $P_1 = \infty$).

Theorem 2. (Approximate universality for general sources) Via suitable decoding, representation $\hat{X}^{(1)}$ can meet distortion constraint $D_2 = 2D_1$ and any perception constraint P_2.

In fact, we first prove the following more general result. Representation $\hat{X}^{(1)}$ can meet distortion constraint $D_2 + \delta_D$ and perception constraint P_2 for any (D_2, P_2), where $\delta_D = D_1 - D_2 + \inf_{p_{\hat{X}^{(2)}}: d(p_X, p_{\hat{X}^{(2)}}) \leq P_2} W_2^2(p_{\hat{X}^{(1)}}, p_{\hat{X}^{(2)}})$.

Proof of Theorem 2. Let $\hat{X}^{(2)}$ be an arbitrary representation jointly distributed with $(X, \hat{X}^{(1)})$ such that $X - \hat{X}^{(1)} - \hat{X}^{(2)}$ form a Markov chain and $\mathbb{E}[(\hat{X}^{(1)} - \hat{X}^{(2)})^2] = W_2^2(p_{\hat{X}^{(1)}}, p_{\hat{X}^{(2)}})$. Note that $\hat{X}^{(2)}$ can be obtained from $\hat{X}^{(1)}$ via a determinstic/stochastic transformation. We have

$$\begin{aligned}\mathbb{E}[(X - \hat{X}^{(2)})^2] &\overset{\text{(a)}}{=} \mathbb{E}[(X - \hat{X}^{(1)})^2] + \mathbb{E}[(\hat{X}^{(1)} - \tilde{X}^{(2)})^2] \\ &\leq D_1 + W_2^2(p_{\hat{X}^{(1)}}, p_{\hat{X}^{(2)}}),\end{aligned}$$

where (a) is because $\hat{X}^{(1)} = \mathbb{E}[X|\hat{X}^{(1)}]$ almost surely and the MMSE estimation error is orthogonal to any function of $\hat{X}^{(1)}$. Minimizing $W_2^2(p_{\hat{X}^{(1)}}, p_{\hat{X}^{(2)}})$ subject to the constraint $d(p_X, p_{\hat{X}^{(2)}}) \leq P_2$ finishes the proof.

Observe that

$$\inf_{p_{\hat{X}^{(2)}}: d(p_X, p_{\hat{X}^{(2)}}) \leq P_2} W_2^2(p_{\hat{X}^{(1)}}, p_{\hat{X}^{(2)}}) \leq W_2^2(p_{\hat{X}^{(1)}}, p_X) \leq D_1,$$

because we may choose $X^{(2)}$ to have the same distribution as X. This yields the original statement. □

B Experimental Details

Training lasted 30 epochs for MNIST and 80 epochs for SVHN, and alternates between training the encoder and decoder with the critic fixed and training the critic with the encoder and decode fixed. The learning rate was decayed by a factor of 5 after 20 epochs for MNIST, and likewise after 25 epochs for SVHN. All models were trained with the Adam optimizer. The batch size used was 64. Training a single model takes about 30 minutes and 100 minutes for MNIST and SVHN, respectively. All training was performed on an RTX 2070 GPU.

The architectures used for the experiments are given as follows. Here each row represents a group of layers. Noise is added for stochasticity after the output of the encoder. d denotes the latent dimension and L the number of quantization levels per dimension, with $R = d \log L$ used to upper bound the true rate (for MNIST, we found that this was within 15% of optimality). The quantizer performs hard nearest-neighbour quantization on the forward pass and uses a soft relaxation given by Equation (3) in (Mentzer et al., 2018) during the backward pass. The bin centers for quantization are spaced evenly in $[-1, 1]$ for each dimension. The type of compression systems are denoted by E for end-to-end and U for (perception-distortion) universal. We found the results to be robust against model architectures when equipped with hyperparameters with good training performance. Moreover, we found that using universal encoders trained with larger λ performed best in practice, though using small λ is nearly as effective also.

B.1 MNIST

The universality experiments build off of the encoders produced by the end-to-end experiments of the same rate with $\lambda = 0.015$.

Table 1: Network and quantizer settings for MNIST for models shown in Figure 3(a).

System	R	d	L
E+U	4.75	3	3
E+U	6	3	4
E+U	8	4	4
E+U	10	5	4

Table 2: The tradeoff coefficients used across all rates in each experiment for MNIST.

System	Tradeoff coefficients
E (Figure 3(a))	$\lambda = 0,\ 0.0033,\ 0.005,\ 0.0066,\ 0.008,$ 0.01, 0.011, 0.013, 0.015
U (Figure 3(a))	$\lambda_i = 0,\ 0.0025,\ 0.004,\ 0.005,\ 0.006,$ 0.008, 0.009, 0.01, 0.011, 0.013

Table 3: Model architectures for MNIST. l-ReLU denotes Leaky ReLU.

Encoder

Input
Flatten
Linear, BatchNorm2D, l-ReLU
Linear, BatchNorm2D, l-ReLU
Linear, BatchNorm2D, l-ReLU
Linear, BatchNorm2D, l-ReLU
Linear, BatchNorm2D, Tanh
Quantizer

Decoder

Input
Linear, BatchNorm1D, l-ReLU
Linear, BatchNorm1D, l-ReLU
Unflatten
ConvT2D, BatchNorm2D, l-ReLU
ConvT2D, BatchNorm2D, l-ReLU
ConvT2D, BatchNorm2D, Sigmoid

Critic

Input
Conv2D, l-ReLU
Conv2D, l-ReLU
Conv2D, l-ReLU
Linear

Table 4: Hyperparameters used for training MNIST models across all rates, including for universal/refining encoders. α is the learning rate, (β_1, β_2) are the parameters for Adam, and λ_{GP} is the gradient penalty coefficient.

	α	β_1	β_2	λ_{GP}
Encoder	10^{-2}	0.5	0.9	10
Decoder	10^{-2}	0.5	0.9	10
Critic	2×10^{-4}	0.5	0.9	10

B.2 SVHN

The experiments are similar to MNIST, with the main difference being in the encoder architecture. The universality experiments build off of the encoders produced by the end-to-end experiments of the same rate with $\lambda = 0.002$.

Table 5: Network and quantizer settings for SVHN for models shown in Figure 3(b).

System	R	d	L
E+U	30	10	8
E+U	45	15	8
E+U	60	20	8

Table 6: The tradeoff coefficients used across all rates in each experiment for SVHN.

System	Tradeoff coefficients
E (Figure 3(b))	$\lambda = 0$, 0.00025, 0.0005, 0.00075, 0.001, 0.00125, 0.0015, 0.002
U (Figure 3(b))	$\lambda_i = 0$, 0.0003, 0.0005, 0.0008, 0.001, 0.0012, 0.0017

Table 7: Model architectures for SVHN.

Encoder
Input
Conv2D, l-ReLU
Conv2D, l-ReLU
Conv2D, l-ReLU
Flatten
Linear, Tanh
Quantizer

Decoder
Input
Linear, BatchNorm1D, l-ReLU
Linear, BatchNorm1D, l-ReLU
Unflatten
ConvT2D, BatchNorm2D, l-ReLU
ConvT2D, BatchNorm2D, l-ReLU
ConvT2D, BatchNorm2D, l-ReLU
ConvT2D, BatchNorm2D, Sigmoid

Critic
Input
Conv2D, l-ReLU
Conv2D, l-ReLU
Conv2D, l-ReLU
Linear

Table 8: Hyperparameters used for training. α is the learning rate, (β_1, β_2) are the parameters for Adam, and λ_{GP} is the gradient penalty coefficient.

	α	β_1	β_2	λ_{GP}
Encoder	10^{-4}	0.5	0.999	10
Decoder	10^{-4}	0.5	0.999	10
Critic	10^{-4}	0.5	0.999	10

B.3 Experimental Samples

As the emphasis on perception loss λ_i increases, the output blurriness is reduced, but the reconstruction is less faithful to the original (in extreme cases even changing the identity of the digit). The visual quality of both the end-to-end and universal models are on average comparable for each λ_i.

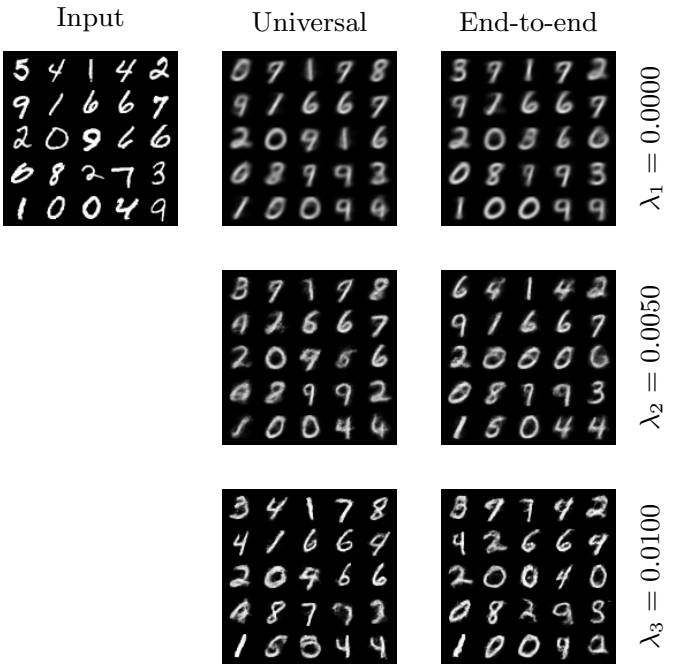


Figure 4: Outputs of selected universal and end-to-end models on MNIST ($R = 4.75$).

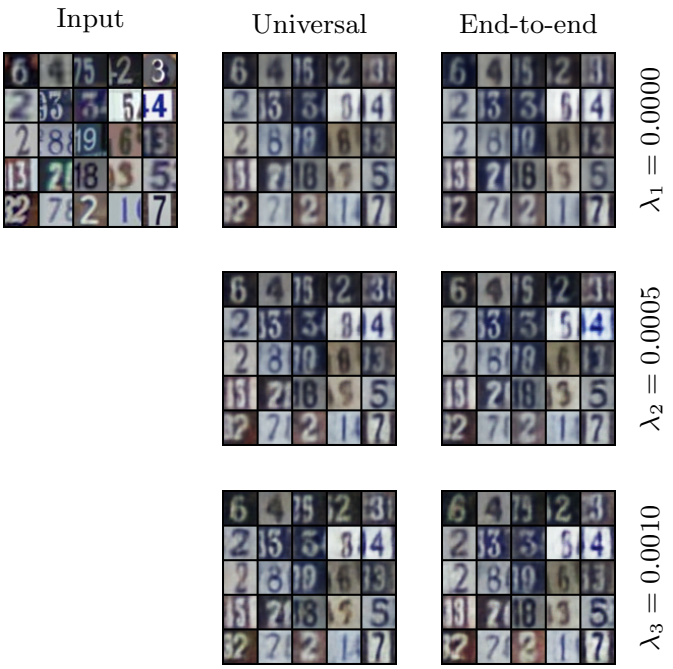


Figure 5: Outputs of selected universal and end-to-end models on SVHN ($R = 60$).