Universal Representations for Classification-enhanced Lossy Compression

Nam Nguyen
Oregon State University, Oregon, United States
{nguynam4}@oregonstate.edu

Abstract

In lossy compression, the classical tradeoff between compression rate and reconstruction distortion has traditionally guided algorithm design. However, Blau and Michaeli [5] introduced a generalized framework, known as the rate-distortion-perception (RDP) function, incorporating perceptual quality as an additional dimension of evaluation. More recently, the rate-distortion-classification (RDC) function was investigated in [19], evaluating compression performance by considering classification accuracy alongside distortion. In this paper, we explore universal representations, where a single encoder is developed to achieve multiple decoding objectives across various distortion and classification (or perception) constraints. This universality avoids retraining encoders for each specific operating point within these tradeoffs. Our experimental validation on the MNIST dataset indicates that a universal encoder incurs only minimal performance degradation compared to individually optimized encoders for perceptual image compression tasks, aligning with prior results from [23]. Nonetheless, we also identify that in the RDC setting, reusing an encoder optimized for one specific classification-distortion tradeoff leads to a significant distortion penalty when applied to alternative points.

1. Introduction

Lossy compression involves reconstructing data from compressed representations with acceptable levels of distortion, typically measured using metrics such as mean squared error (MSE), PSNR, and SSIM [20, 21]. While classical rate-distortion theory seeks to minimize distortion for a given compression rate, recent studies have demonstrated that lower distortion does not necessarily imply higher perceptual quality [1, 3]. Addressing this discrepancy, Blau and Michaeli introduced the rate-distortion-perception (RDP) framework [5], which incorporates perceptual quality—measured through the divergence between source and reconstruction distributions—as an additional dimension of evaluation. The emergence of generative adversarial networks (GANs) [9] has facilitated this perceptual enhancement even at low bitrates [17], highlighting a clear tradeoff between perceptual quality and distortion.

Additionally, recent work has explored the integration of classification performance into compression objectives, resulting in the rate-distortion-classification (RDC) tradeoff [19, 24]. Here, improved classification performance is often achieved at the cost of increased distortion or reduced perceptual quality.

In this paper, we investigate the concept of *universal representations*, where a single encoder is used to support multiple decoding objectives across various distortion-classification (or distortion-perception) tradeoff points. We demonstrate through experiments on the MNIST dataset that universal encoders achieve performance close to specialized encoders in perceptual image compression tasks, confirming prior findings [23]. However, we also highlight that reusing encoders trained for a particular classification-distortion tradeoff introduces a notable distortion penalty when applied to other points. Our findings suggest practical ways to simplify and streamline the training of deep-learning-based compression systems by reducing the need for specialized encoders. Throughout this study, we focus on scenarios where the compression rate is fixed.

2. Related Works

Image quality assessment typically involves full-reference metrics such as MSE and SSIM [20], or no-reference metrics like BRISQUE and Fréchet Inception Distance [11]. Broadly, full-reference metrics quantify distortion, whereas no-reference metrics measure perceptual quality. Recently, GAN-based methods have enhanced perceptual realism by leveraging discriminators trained to estimate statistical divergences [2].

Rate-distortion theory provides the foundational framework for analyzing lossy compression [8], influencing both representation learning and generative modeling [18]. Distribution-preserving lossy compression has also been studied within classical information theory [22]. Recent approaches have integrated GAN regularization into compressive autoencoders to reduce artifacts, significantly improving perceptual quality even at low bitrates [14].

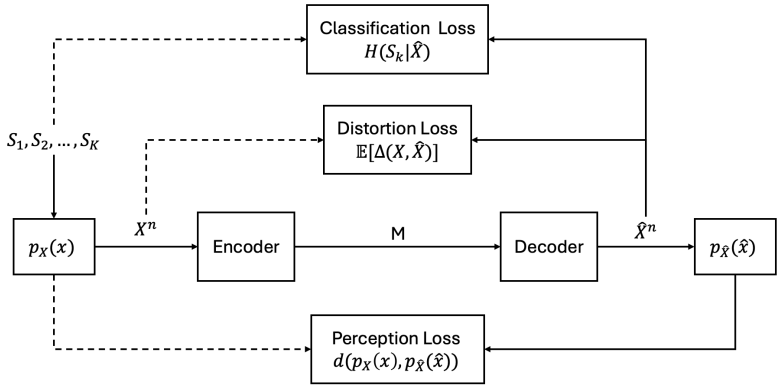


Figure 1. Schematic representation of a task-oriented lossy compression framework.

Incorporating classification performance into compression objectives has resulted in notable tradeoffs between distortion, perception, and classification accuracy [12, 19]. Such studies highlight that improved classification accuracy frequently leads to increased distortion or lower perceptual quality.

3. Problem Formulation

Consider a source generating observable data $X \sim p_X(x)$ associated with several underlying but unobserved labels $S_1, \ldots, S_K$. These labels and the observation X follow a joint distribution $p_{X,S}(x, s_1, \ldots, s_K)$. For instance, X could represent speech data, with labels such as spoken content, speaker identity, or speaker gender.

As depicted in Fig. 1, a lossy compression scheme consists of an encoder-decoder pair. Given an independent and identically distributed (i.i.d.) sequence $X_1, X_2, \ldots, X_n \sim p_X(x)$: The encoder maps the source sequence X^n into a compressed representation $M \in \{1, 2, \ldots, 2^{nR}\}$ at rate R bits. The decoder reconstructs data $\hat{X}^n$ from the compressed message M.

Distortion constraint. The reconstruction must satisfy a distortion limit:

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

where $\Delta(\cdot, \cdot)$ measures distortion, e.g., MSE or Hamming distance [8], and the expectation is taken over the joint distribution of X and $\hat{X}$.

Classification constraint. The reconstructed data should preserve sufficient information for accurate inference of labels:

$$H(S_k|\hat{X}) \leq C_k, \quad \forall k \in [K], \tag{2}$$

where $H(S_k|\hat{X})$ denotes conditional entropy, limiting uncertainty about each label given the reconstructed data [19].

Perception constraint. Perceptual quality measures how natural and realistic reconstructed images appear to human observers, distinguishing them from artificially generated or processed images [15]. To formally quantify perceptual quality, we impose a constraint based on a divergence measure between the distributions of original and reconstructed data, following established practices [4, 6]:

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

where $d(\cdot, \cdot)$ denotes a statistical divergence measure such as total-variation (TV) divergence or Kullback-Leibler (KL) divergence.

To characterize achievable rates under these constraints, we define the information rate-distortion-perception-classification (RDPC) function [12]:

$$\begin{aligned} R(D, P, \mathbf{C}) = \min_{p_{\hat{X}|X}} \quad & I(X; \hat{X}) && \text{(4a)} \\ \text{s.t.} \quad & \mathbb{E}[\Delta(X, \hat{X})] \leq D, && \text{(4b)} \\ & d(p_X, p_{\hat{X}}) \leq P, && \text{(4c)} \\ & H(S_k|\hat{X}) \leq C_k, \quad \forall k \in [K], && \text{(4d)} \end{aligned}$$

where $\mathbf{C} = (C_1, \ldots, C_K)$ represents the constraints on classification uncertainty.

4. Rate-Distortion-Perception/Classification Representations

In lossy compression, an encoder-decoder pair optimizes the tradeoffs among compression rate, distortion, and additional constraints like perceptual quality or classification accuracy. We formally define two essential functions for these tradeoffs: the information rate-distortion-perception and information rate-distortion-classification.

Information rate-distortion-perception function. Incorporating perceptual quality into the classical rate-distortion

framework, the RDP function is defined as:

$$R(D,P) = \min_{p_{\hat{X}|X}} \quad I(X;\hat{X}) \tag{5a}$$

$$\text{s.t.} \quad \mathbb{E}[\Delta(X,\hat{X})] \le D, \tag{5b}$$

$$d(p_X, p_{\hat{X}}) \le P. \tag{5c}$$

where $\Delta(\cdot,\cdot)$ measures reconstruction distortion, and $d(p_X, p_{\hat{X}})$ is the divergence between original and reconstructed distributions. The RDP function is monotonically non-increasing and convex [5].

Information rate-distortion-classification function. For tasks emphasizing classification accuracy, the RDC function characterizes the optimal tradeoff:

$$R(D,C) = \min_{p_{\hat{X}|X}} \quad I(X;\hat{X}) \tag{6a}$$

$$\text{s.t.} \quad \mathbb{E}[\Delta(X,\hat{X})] \le D, \tag{6b}$$

$$H(S|\hat{X}) \le C, \tag{6c}$$

where $H(S|\hat{X})$ limits uncertainty about labels. Similar to RDP, the RDC function is monotonically non-increasing and convex [19].

5. Universal Representations for RDC

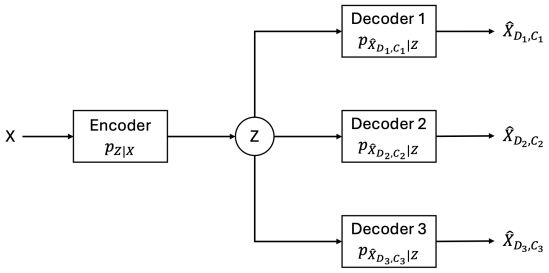


Figure 2. Illustration of the universal representation framework.

The concept of universal representations, initially introduced in [23] for the rate-distortion-perception framework, aims to design a single encoder that supports multiple decoding objectives. Here, we extend this concept to the rate-distortion-classification scenario.

Consider a fixed encoder that generates a universal representation, from which various decoders can reconstruct data satisfying multiple distortion-classification constraints (D,C) as shown in Figure 2. Formally, the information universal rate-distortion-classification function is defined as:

$$R(\Theta) = \inf_{p_{Z|X} \in \mathcal{P}_{Z|X}(\Theta)} I(X;Z), \tag{7}$$

where $\mathcal{P}_{Z|X}(\Theta)$ includes all encoding schemes allowing decoders to satisfy constraints $(D,C) \in \Theta$.

The *rate penalty* from using a single universal encoder instead of specialized encoders is:

$$A(\Theta) = R(\Theta) - \sup_{(D,C)\in\Theta} R(D,C). \tag{8}$$

Ideally, if we define $\Omega(R) = \{(D,C) : R(D,C) \le R\}$, the rate penalty $A(\Omega(R))$ should be minimal (preferably zero). This implies that one encoder can efficiently achieve multiple classification-distortion constraints without incurring additional encoding cost, significantly simplifying the system design and reducing computational overhead.

6. Experiments

7. Universal Representation for Lossy Compression

The rate-distortion-classification tradeoff arises naturally when integrating classification objectives into deep-learning-based image compression [19]. Typically, each desired RDC setting requires training an individual encoder-decoder pair. However, retraining entire models for every specific objective is inefficient. Therefore, we consider reusing a pre-trained encoder and adjusting only the decoder—termed as *universal* models. In contrast, models fully retrained for each objective are called *end-to-end* models. Our goal is to quantify the distortion and classification performance penalties when reusing encoders compared to specialized training.

Setup and Training. We employ a stochastic autoencoder architecture incorporating GAN and pre-trained classifier regularization. Specifically, the model comprises an encoder (f), decoder (g), critic (h), and a pre-trained classifier (c). Detailed network configurations are provided in the supplementary material.

Given an input image x, the encoder's final layer applies a tanh activation, producing a continuous representation $f(x) \in [-1,1]^d$. This representation is quantized into L evenly spaced length intervals $2/(L-1)$ across each dimension. Employing the shared randomness assumption, we utilize dithered quantization [10, 16], where both sender and receiver share a common random vector $u \sim U[-1/(L-1), +1/(L-1)]^d$. The sender computes:

$$z = \text{Quantize}(f(x) + u), \tag{9}$$

and transmits z. The receiver reconstructs the input by decoding $z - u$. During training, the soft gradient estimator from [13] is used to propagate gradients through the quantization step.

Our training approach closely follows [5]. Initially, *end-to-end* models are trained where the encoder, decoder, and critic are jointly optimized. We use mean squared error (MSE) as the distortion metric, Wasserstein-1 distance for perceptual realism, and classification accuracy derived from

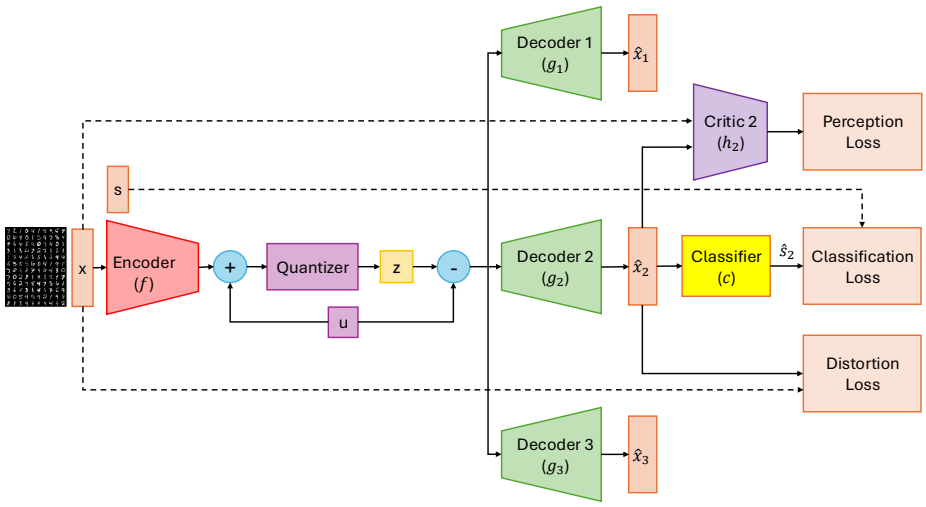
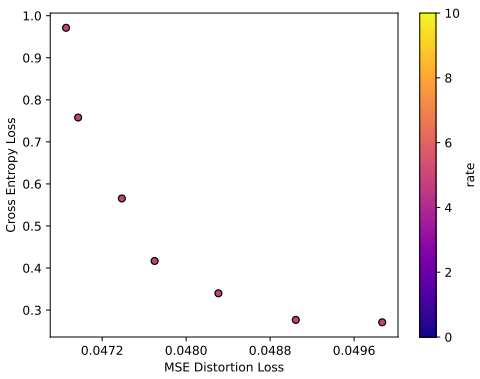
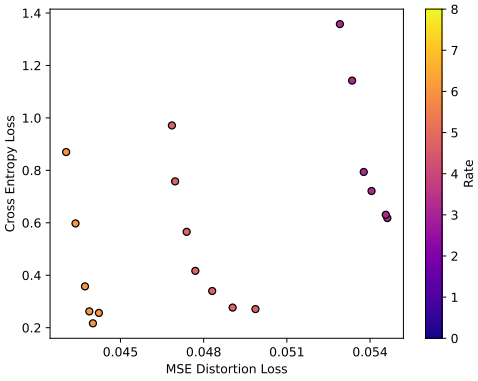


Figure 3. Diagram of the experimental framework for the universal representation model. Initially, an encoder network f is trained to achieve a predetermined balance between classification accuracy and reconstruction distortion (alternatively, perception and distortion). After training, the encoder's parameters are fixed. Subsequently, multiple specialized decoders $\{g_i\}$ are independently optimized, each targeting distinct trade-off criteria using the fixed representation z generated by encoder f. A shared source of randomness u is accessible to both sender and receiver to facilitate universal quantization. Additionally, dedicated critic networks $\{h_i\}$ are concurrently trained alongside each decoder to enhance perceptual quality. A pre-trained classifier network (C) is utilized for evaluating classification performance.



(a) CDR for MNIST.



(b) RDC with multiple rates for MNIST.

Figure 4. Classification-distortion-rate functions along various rates for the MNIST dataset, illustrating the tradeoff between rate, distortion, and classification.

the pre-trained classifier c. Specifically, the reconstructed image $\hat{x}$ is classified by c, yielding predicted class probabilities $\hat{\mathbf{s}}$. Classification accuracy is determined using the true label s and $\hat{\mathbf{s}}$.

Conditional entropy constraints are approximated via the cross-entropy (CE) loss with a pre-trained classifier parameterized by ψ, following [7, 19]:

$$H(S|\hat{X}) \leq \text{CE}(s, \hat{\mathbf{s}}).$$

The compression rate R is set as $\dim \times \log_2(L)$, the product of the representation dimension and quantization levels, simplifying the implementation with minimal suboptimality [1, 5].

The overall loss functions for the RDC and RDP cases are:

$$\mathcal{L}_{RDC} = \mathbb{E}[\|X - \hat{X}\|^2] + \lambda_c \text{CE}(s, \hat{\mathbf{s}}), \quad (10)$$

$$\mathcal{L}_{RDP} = \mathbb{E}[\|X - \hat{X}\|^2] + \lambda_p W_1(p_X, p_{\hat{X}}), \quad (11)$$

where λ_c and λ_p control the tradeoffs between distortion, classification, and perception.

To build universal models, we freeze encoders from previously trained end-to-end models and retrain only new decoders (g_1) and critics (h_1) using modified parameters (λ_c^1,

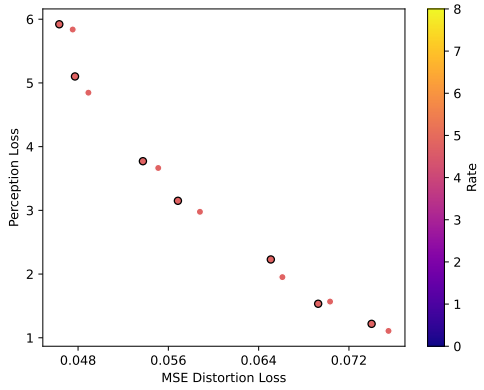


Figure 5. Perception-distortion-rate functions evaluated at a fixed rate of $R = 4.75$ on the MNIST dataset. Points highlighted with black outlines indicate results obtained from end-to-end trained encoder-decoder models tailored specifically to particular perception-distortion targets. All other points represent outcomes from universal models, in which decoders are trained separately using representations from an encoder fixed at low perceptual distortion ($\lambda_p = 0.015$). The universal models closely match the performance of the jointly trained models across the entire range of trade-offs.

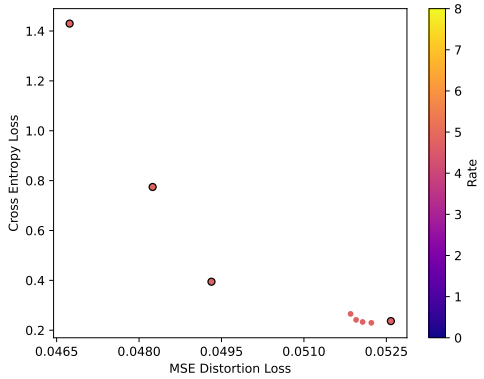


Figure 6. Classification-distortion-rate curves at a fixed rate $R = 4.75$ for the MNIST dataset. Points with black outlines correspond to results from encoder-decoder models trained jointly end-to-end for specific classification-distortion objectives. All other points depict universal model outcomes, where decoders are optimized using representations from an encoder fixed at low classification loss ($\lambda_c = 0.015$). The universal models show cross-entropy losses nearly identical to the frozen encoder baseline; however, a notable distortion gap remains between universal and end-to-end models. It is suggested to utilize λ_c^1 at a different scaling factor from λ_c for better alignment.

λ_p^1):

$$\mathcal{L}_{RDC}^1 = \mathbb{E}[\|X - \hat{X}_1\|^2] + \lambda_c^1 \mathrm{CE}(s, \hat{\mathbf{s}}_1), \tag{12}$$

$$\mathcal{L}_{RDP}^1 = \mathbb{E}[\|X - \hat{X}_1\|^2] + \lambda_p^1 W_1(p_X, p_{\hat{X}_1}). \tag{13}$$

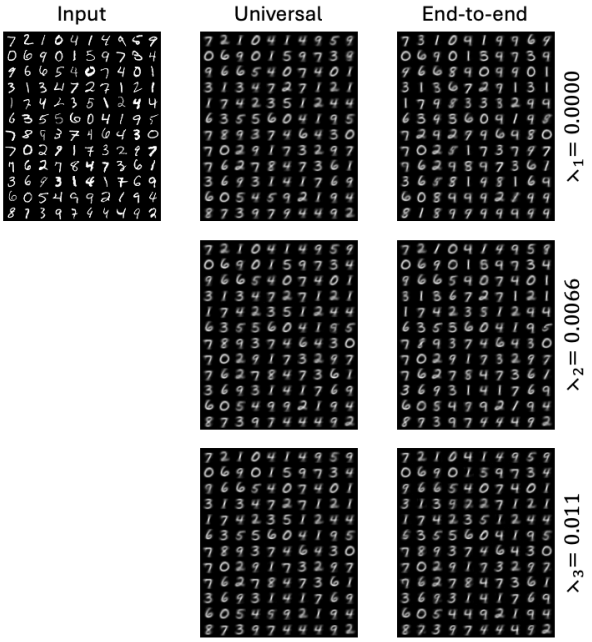


Figure 7. Decompression outputs of selected models for classification-distortion-rate at rate $R = 4.75$ on MNIST dataset. As the emphasis on classification loss (λ_c increases), the decompression images become sharper.

Weights of the new decoders and critics are initialized randomly or from the previously trained critic. This process is repeated over multiple parameters (λ_c^i, λ_p^i) to produce comprehensive RDC and RDP tradeoff curves. Additional details are included in the supplementary material.

7.1. Results

Figure 4 presents the classification-distortion-rate tradeoffs for MNIST using several rates defined by $R = \dim \times \log_2(L)$ with dimension-level pairs $(3, 3)$, $(3, 4)$, and $(4, 4)$. Each point corresponds to a trained encoder-decoder pair at specific combinations of R and λ_c. Points sharing the same color indicate identical rates, clearly illustrating the inherent tradeoff: enhancing classification accuracy generally leads to increased distortion. Additionally, higher rates enable simultaneous improvements in distortion and classification accuracy.

Figure 5 demonstrates the performance comparison between end-to-end and universal models on the rate-distortion-perception curve at a fixed rate of $R = 4.75$. Universal models, which reuse an encoder trained from an end-to-end model, closely approach the performance of their end-to-end counterparts across the perception-distortion tradeoff, validating prior observations [23].

Similarly, Figure 7 presents the rate-distortion-classification curve for the same fixed rate ($R = 4.75$). Notably, we observe a significant distortion gap when

reusing encoders trained at a particular distortion-classification tradeoff for other tradeoff points. To address this, we recommend selecting the tradeoff parameter λ_c^1 at a scale different from the original λ_c, allowing universal models to effectively span the entire distortion-classification tradeoff spectrum.

8. Conclusion

In this work, we examined the inherent tradeoffs among rate, distortion, classification accuracy, and perceptual quality within lossy compression frameworks. While previous approaches typically involved training end-to-end systems separately for each desired tradeoff point, our findings suggest that fixing a carefully optimized representation (encoder) and only adapting the decoder is sufficient to achieve comparable performance. This significantly simplifies system design by reducing computational costs and model redundancy. Future research may explore extending this universal representation concept to more complex scenarios, including high-resolution image and video compression tasks.

References

[1] 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*, pages 221–231, 2019. 1, 4

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017. 1

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 1

[4] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 2

[5] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pages 675–685, 2019. 1, 3, 4

[6] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 675–685. PMLR, 09–15 Jun 2019. 2

[7] Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. *arXiv preprint*, 2021. arXiv 2003.08983. 4

[8] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 1999. 1, 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. volume 27, pages 2672–2680, 2014. 1

[10] Robert M Gray and Thomas G Stockham. Dithered quantizers. *IEEE Transactions on Information Theory*, 39(3):805–812, 1993. 3

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 1

[12] Dong Liu, Haochen Zhang, and Zhiwei Xiong. On the classification-distortion-perception tradeoff. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[13] 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*, pages 4394–4402, 2018. 3

[14] 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. 1

[15] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364, 2011. 2

[16] Lucas Theis and Eirikur Agustsson. On the advantages of stochastic encoders. *arXiv preprint arXiv:2102.09270*, 2021. 3

[17] Michael Tschannen, Eirikur Agustsson, and Mario Lucic. Deep generative models for distribution-preserving lossy compression. In *Advances in Neural Information Processing Systems*, pages 5929–5940, 2018. 1

[18] Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018. 1

[19] Yuhan Wang, Youlong Wu, Shuai Ma, and Ying-Jun Angela Zhang. Lossy compression with data, perception, and classification constraints. In *2024 IEEE Information Theory Workshop (ITW)*, pages 366–371, 2024. 1, 2, 3, 4

[20] 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. 1

[21] 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, pages 1398–1402. IEEE, 2003. 1

[22] Ram Zamir and Kenneth Rose. Natural type selection in adaptive lossy compression. *IEEE Transactions on information theory*, 47(1):99–111, 2001. 1

[23] George Zhang, Jingjing Qian, Jun Chen, and Ashish Khisti. Universal rate-distortion-perception representations for lossy compression. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages

11517–11529. Curran Associates, Inc., 2021. 1, 3, 5
[24] Yuefeng Zhang. A rate-distortion-classification approach for lossy image compression. *Digital Signal Processing*, 141:104163, Sept. 2023. 1

A. Architecture

The architecture employed in our experiments is a stochastic autoencoder combined with GAN and pre-trained classifier regularization. Each model is composed of an encoder, decoder, critic, and classifier. The detailed network architectures for these components are summarized in Table 1, with each row indicating a sequence of layers.

Table 1. Detailed architecture for encoder, decoder, critic, and pre-trained classifier used in MNIST experiments. l-ReLU denotes Leaky ReLU activation.

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

Pre-trained Classifier

Input
Conv2D (10 filters, kernel=5), ReLU
MaxPool2D (kernel size=2)
Conv2D (10 filters, kernel=5), ReLU
MaxPool2D (kernel size=2)
Flatten
Linear, ReLU
Linear, Softmax

Table 2 summarizes the hyperparameters used in training each component of the model across all experiments.

Table 2. Training hyperparameters used across all experiments.

	α	β_1	β_2	λ_{GP}
Encoder	10^{-2}	0.5	0.9	-
Decoder	10^{-2}	0.5	0.9	-
Critic	2×10^{-4}	0.5	0.9	10
Classifier	10^{-3}	0.9	0.999	-