Universal Rate-Distortion-Classification Representations for Lossy Compression

Nam Nguyen[1], Thuan Nguyen[2], Thinh Nguyen[1], and Bella Bose[1]
[1]School of Electrical and Computer Engineering, Oregon State University, Corvallis, OR, 97331
[2]Department of Engineering, Engineering Technology, and Surveying, East Tennessee State University, Johnson, TN, 37604
Emails: nguynam4@oregonstate.edu, nguyent11@etsu.edu, thinhq@eecs.oregonstate.edu, bella.bose@oregonstate.edu

***Abstract*—In lossy compression, Wang et al. [1] recently introduced the rate-distortion-perception-classification function, which supports multi-task learning by jointly optimizing perceptual quality, classification accuracy, and reconstruction fidelity. Building on the concept of a universal encoder introduced in [2], we investigate the universal representations that enable a broad range of distortion-classification tradeoffs through a single shared encoder coupled with multiple task-specific decoders. We establish, through both theoretical analysis and numerical experiment, that for a Gaussian source under mean-squared error (MSE) distortion, the distortion-classification tradeoff region can be achieved using a single universal encoder. For general sources, we characterize the achievable region and identify conditions under which a universal encoder can produce a small distortion penalty. The experimental result on the MNIST dataset further supports our theoretical findings. We show that universal encoders can obtain distortion performance comparable to task-specific encoders. These results demonstrate the practicality and effectiveness of the proposed universal framework in multi-task compression scenarios.**

***Keywords*—Lossy compression, rate-distortion-classification tradeoff, universal representations.**

I. Introduction

Rate-distortion theory has long served as the foundation for lossy compression, characterizing the minimum distortion achievable at a given bit rate [3]. Conventional systems are typically evaluated using full-reference distortion metrics such as MSE, PSNR, SSIM, and MS-SSIM [4], [5]. However, recent research has demonstrated that minimizing distortion alone is insufficient to yield perceptually convincing reconstructions. This limitation is particularly evident in deep learning-based image compression, where empirical evidence suggests that improvements in perceptual quality often come at the expense of increased distortion [6], [7].

To address this limitation, Blau and Michaeli [8] introduced the rate-distortion-perception (RDP) framework, which incorporates perceptual quality, measured via distributional divergence, as an another metric. The RDP formulation shows a fundamental tradeoff among rate, distortion fidelity, and perceptual realism. Practical implementations, particularly those using GANs [9], have demonstrated high perceptual quality at low bit rates [10]–[13]. Common no-reference perceptual metrics include FID [14], NIQE [15], PIQE [16], and BRISQUE [17].

This work was supported by the National Science Foundation Grant CCF-2417898.

Extending this approach, the classification-distortion-perception (CDP) framework was introduced in [18], incorporating classification accuracy as an additional metric. This work showed that the tradeoffs among distortion, perceptual quality, and classification accuracy are fundamentally irreconcilable; that is, improving one metric generally degrades the others.

Recent work has explored integrating classification into lossy compression frameworks to jointly optimize multiple downstream objectives. The rate-distortion-classification (RDC) framework, proposed by Zhang et al. [19], formalized the joint optimization of rate, distortion, and classification accuracy, and established desirable properties such as monotonicity and convexity under multi-distribution source models. Extending this, Wang et al. [1] introduced the rate-distortion-perception-classification (RDPC) function, which characterizes the tradeoffs among rate, distortion, perception, and classification. Their results show that improving classification typically increases distortion or degrades perceptual quality.

These tradeoffs raise a fundamental question: are they inherently tied to the encoder's representation, or can a single encoder flexibly support different task objectives by using different decoders? To address this, the universal RDP framework was introduced in [2], in which a fixed encoder is paired with multiple decoders to achieve various operating points in the distortion-perception space, without requiring encoder retraining.

Motivated by the idea in [2], we investigate the universal representations that enable a broad range of distortion-classification tradeoffs through a single shared encoder coupled with multiple task-specific decoders. We establish, through both theoretical analysis and numerical experiment, that for Gaussian source under mean-squared error (MSE) distortion, the distortion-classification tradeoff region can be achieved using a single universal encoder. For general sources, we characterize the achievable region and identify conditions under which a universal encoder can produce a small distortion penalty. The experimental result on the MNIST dataset further supports our theoretical findings. We show that universal encoders can obtain distortion performance comparable to task-specific encoders. These results demonstrate the practicality and effectiveness of the proposed universal framework in multi-task compression scenarios.

II. Rate-Distortion-Classification Representations

Consider a source generating observable data $X \sim p_X$ with multiple latent target labels $S_1, \dots, S_K \sim p_{S_1,\dots,S_K}$. These labels are correlated with X under the joint distribution $p_{X,S_1,\dots,S_K}$. While $S_1, \dots, S_K$ are not directly observed, they are often inferable from X. For example, if X is an image, classification tasks can be object recognition or scene understanding.

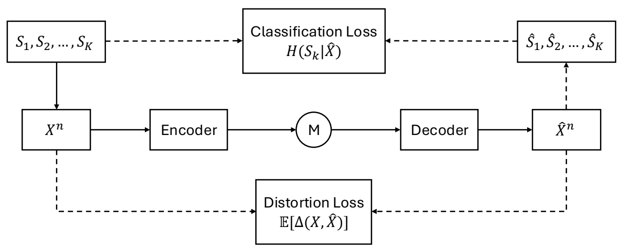

Figure 1: Task-oriented lossy compression framework.

As shown in Fig. 1, the lossy compression system consists of an encoder and a decoder. For a sequence $X^n \sim p_X^n$, the encoder $f : \mathcal{X}^n \to \{1, \dots, 2^{nR}\}$ compresses it to a message M at rate R, which is then decoded by $g : \{1, \dots, 2^{nR}\} \to \hat{\mathcal{X}}^n$ to produce $\hat{X}^n$. The goal is to retain task-relevant information while compressing efficiently.

Distortion constraint. The reconstructed signal $\hat{X}$ must satisfy:

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

where Δ is a distortion metric, e.g., Hamming or MSE.

Classification constraint. We impose the following classification constraint:

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

for some $C_k > 0$. This ensures that the uncertainty of the classification variable S_k given the reconstructed source $\hat{X}$ does not exceed C_k.

Information RDC function. To jointly capture distortion and classification constraints, the information rate-distortion-classification function is defined as follows:

Definition 1. [1] For a source $X \sim p_X$ and a single associated classification variable S, the *information rate-distortion-classification function* is defined as:

$$R(D, C) = \min_{p_{\hat{X}|X}} \quad I(X; \hat{X}) \tag{3a}$$

$$\text{s.t.} \quad \mathbb{E}[\Delta(X, \hat{X})] \leq D, \tag{3b}$$

$$H(S|\hat{X}) \leq C. \tag{3c}$$

III. Gaussian Source

This section analyzes the RDC tradeoff for a scalar Gaussian source, deriving closed-form expressions that highlight the relationship between compression, distortion, and classification.

For a scalar Gaussian source, the closed-form expression of $R(D, C)$ is given in the following theorem by Wang et al. [1].

Theorem 1. [1] Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ be a Gaussian source and $S \sim \mathcal{N}(\mu_S, \sigma_S^2)$ be an associated classification variable, with a covariance of $\text{Cov}(X, S) = \theta_1$. The problem (3) is feasible if $C \geq \frac{1}{2} \log \left(1 - \frac{\theta_1^2}{\sigma_S^2 \sigma_X^2}\right) + h(S)$. For the MSE distortion (i.e., $\mathbb{E}[\Delta(X, \hat{X})] = \mathbb{E}[(X - \hat{X})^2]$), the *information rate-distortion-classification function* is achieved by a jointly Gaussian estimator $\hat{X}$ and given by

$$R(D,C) = \begin{cases} \frac{1}{2}\log\frac{\sigma_X^2}{D}, & \\ \quad D \leq \sigma_X^2\left(1 - \frac{1}{\rho^2}\left(1 - e^{-2h(S)+2C}\right)\right) & \\ -\frac{1}{2}\log\left(1 - \frac{1}{\rho^2}\left(1 - e^{-2h(S)+2C}\right)\right), & \\ \quad D > \sigma_X^2\left(1 - \frac{1}{\rho^2}\left(1 - e^{-2h(S)+2C}\right)\right) & \\ 0, \quad C > h(S) \text{ and } D > \sigma_X^2. & \end{cases}$$

where $\rho = \frac{\theta_1}{\sigma_S \sigma_X}$ represents the correlation coefficient between X and S, while $h(\cdot)$ denotes the differential entropy.

It is noted that $R(D, C)$ is both convex and monotonically non-increasing in D and C, as established in [1]. Furthermore, we characterize the minimum achievable distortion as a function of C and R by the following definition.

Definition 2. For a source $X \sim p_X$ and classification variable S, the *information distortion-classification-rate (DCR) function* is defined as:

$$D(C, R) = \min_{p_{\hat{X}|X}} \quad \mathbb{E}[(X - \hat{X})^2] \tag{5a}$$

$$\text{s.t.} \quad I(X; \hat{X}) \leq R, \tag{5b}$$

$$H(S|\hat{X}) \leq C. \tag{5c}$$

Our first contribution is the derivation of a closed-form expression for $D(C, R)$ in the Gaussian source setting, as formally stated in Theorem 2.

Theorem 2. Consider a Gaussian source $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and an associated classification variable $S \sim \mathcal{N}(\mu_S, \sigma_S^2)$ with covariance $\text{Cov}(X, S) = \theta_1$. The problem (5) is feasible if the classification loss satisfies $C \geq \frac{1}{2} \log \left(1 - \frac{\theta_1^2}{\sigma_S^2 \sigma_X^2}\right) + h(S)$. Under the MSE distortion, the *information distortion-classification-rate function* is achieved by a jointly Gaussian estimator $\hat{X}$ and is given by (4).

Proof. The proof is provided in [20, Appendix A]. □

Following the results in [1], it can also be shown that the distortion-classification-rate function $D(C, R)$ is convex and monotonically non-increasing in both C and R. For any fixed R, as C increases from $\frac{1}{2} \log \left(1 - \frac{\theta_1^2}{\sigma_S^2 \sigma_X^2}\right) + h(S)$ to $\frac{1}{2} \log \left(1 - \frac{\theta_1^2 (\sigma_X^2 - \sigma_X^2 e^{-2R})}{\sigma_S^2 \sigma_X^4}\right) + h(S)$, the distortion $D(C, R)$ decreases monotonically from $\sigma_X^2 - \frac{\sigma_S^2 \sigma_X^4}{\theta_1^2} \left(1 - e^{-2h(S)+2C}\right)$ to the optimal value $\sigma_X^2 e^{-2R}$. Increasing C beyond this point does not yield further improvements in distortion.

$$
D(C,R)=\begin{cases}\sigma_X^2 e^{-2R}, & C>\frac{1}{2}\log\left(1-\frac{\theta_1^2(\sigma_X^2-\sigma_X^2 e^{-2R})}{\sigma_S^2\sigma_X^4}\right)+h(S)\\ \sigma_X^2-\frac{\sigma_S^2\sigma_X^4}{\theta_1^2}\left(1-e^{-2h(S)+2C}\right), & \\ & \frac{1}{2}\log\left(1-\frac{\theta_1^2}{\sigma_S^2\sigma_X^2}\right)+h(S)\le C\le\frac{1}{2}\log\left(1-\frac{\theta_1^2(\sigma_X^2-\sigma_X^2 e^{-2R})}{\sigma_S^2\sigma_X^4}\right)+h(S)\\ 0, & C>h(S)\text{ and } R>h(X).\end{cases} \tag{4}
$$

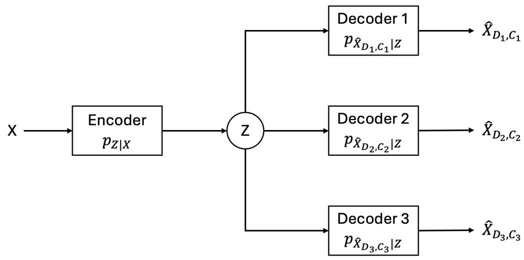

Figure 2: The universal representation framework.

IV. Universal Representations

Designing separate encoders for each distortion-classification constraint is often not desirable. This motivates the use of universal representations, where a single encoder supports multiple decoding constraints, each for a distinct task, as shown in Figure 2. This section introduces the universal RDC framework, quantifies the rate penalty, and presents theoretical results for both Gaussian and general sources.

A. Definitions

In the standard RDC setting, the minimum rate to satisfy a distortion-classification pair (D, C) is achieved by jointly optimizing the encoder and decoder. The proposed universal RDC framework extends this by fixing the encoder and allowing the decoder to adapt, thereby supporting all constraint pairs $(D, C) \in \Theta$, where Θ is a given set of multiple (D, C) pairs.

This raises a key question: What is the minimal additional rate required to satisfy all constraints in Θ with a single encoder? Ideally, this rate penalty is small, indicating near-optimal performance across tasks. We formalize this notion using the information universal rate-distortion-classification function, adapting the definition from [2], as follows.

Definition 3. Let $Z \sim p_{Z|X}$ be a representation of X. Define $\mathcal{P}_{Z|X}(\Theta)$ as the set of encoders such that for every $(D, C) \in \Theta$, there exists a decoder $p_{\hat{X}_{D,C}|Z}$ satisfying $\mathbb{E}[\Delta(X, \hat{X}_{D,C})] \le D, \quad H(S|\hat{X}_{D,C}) \le C$, with $X \leftrightarrow Z \leftrightarrow \hat{X}_{D,C}$. The universal RDC function is defined as

$$
R(\Theta) = \inf_{p_{Z|X} \in \mathcal{P}_{Z|X}(\Theta)} I(X; Z). \tag{6}
$$

A representation Z is Θ-universal if it satisfies all constraints in Θ via appropriate decoders.

Similarly, we adapt the definition from [2] to quantify the rate penalty under the classification constraint as follows:

Definition 4. The *rate penalty* is

$$
A(\Theta) = R(\Theta) - \sup_{(D,C)\in\Theta} R(D, C), \tag{7}
$$

which quantifies the extra rate required for universality.

Let $\Omega(R) = \{(D, C) : R(D, C) \le R\}$ denote the set of achievable distortion-classification pairs at rate R, and define

$$
\Omega(p_{Z|X}) = \left\{ (D, C) : \exists\, p_{\hat{X}_{D,C}|Z} \text{ s.t. } \begin{array}{l} \mathbb{E}[\Delta(X, \hat{X}_{D,C})] \le D, \\ H(S|\hat{X}_{D,C}) \le C \end{array} \right\}.
$$

If $I(X; Z) = R$ and $\Omega(p_{Z|X}) = \Omega(R)$, then Z achieves the maximal distortion-classification region at rate R.

B. Main Results

Theorem 3. Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ be a scalar Gaussian source and $S \sim \mathcal{N}(\mu_S, \sigma_S^2)$ a classification variable with $\mathrm{Cov}(X, S) = \theta_1$. Assume MSE distortion and classification loss measured by $H(S|\hat{X})$. Then, for any non-empty set Θ of distortion-classification pairs (D, C), the rate penalty is zero:

$$
A(\Theta) = 0. \tag{8}
$$

Moreover, any jointly Gaussian representation Z of X satisfying

$$
I(X; Z) = \sup_{(D,C)\in\Theta} R(D, C), \tag{9}
$$

achieves the maximal distortion-classification region:

$$
\Theta \subseteq \Omega(p_{Z|X}) = \Omega(I(X; Z)). \tag{10}
$$

Proof. The detailed proof is presented in [20, Appendix B]. □

In addition, we consider a general source $X \sim p_X$ and characterize the distortion-classification region induced by an arbitrary representation Z under MSE distortion.

Theorem 4. Let $X \sim p_X$ be a general source and S a classification variable with $\mathrm{Cov}(X, S) = \theta_1$. Assume distortion is measured by MSE and classification loss by $H(S|\hat{X})$. Let Z be any representation of X, and define $\tilde{X} = \mathbb{E}[X|Z]$ as the minimum mean square estimator. Then the closure of the achievable region, $\mathrm{cl}(\Omega(p_{Z|X}))$, satisfies

$$
\Omega(p_{Z|X}) \subseteq \left\{ (D, C) : D \ge \mathbb{E}\|X - \tilde{X}\|^2 + \begin{array}{c} \inf_{p_{\hat{X}}} W_2^2(p_{\tilde{X}}, p_{\hat{X}}) \\ \text{s.t. } H(S|\hat{X}) \le C \end{array} \right\}
$$
$$
\subseteq \mathrm{cl}(\Omega(p_{Z|X})),
$$

where the squared 2-Wasserstein distance is $W_2^2(p_X, p_{\hat{X}}) = \inf_{p_{X,\hat{X}}} \mathbb{E}[\|X - \hat{X}\|^2]$ with the infimum taken over all joint distributions with marginals p_X and $p_{\hat{X}}$.

Moreover, $\mathrm{cl}(\Omega(p_{Z|X}))$ contains the extreme points:

$$(D^{(a)}, C^{(a)}) = \left(\mathbb{E}[\|X - \tilde{X}\|^2], \sum_s \sum_{\tilde{x}} p_{\tilde{X}} p_{S|\tilde{X}} \log \frac{1}{p_{S|\tilde{X}}}\right),$$
$$(D^{(b)}, C^{(b)}) = \left(\mathbb{E}[\|X - \tilde{X}\|^2] + W_2^2(p_{\tilde{X}}, p_{\hat{X}^{C_{\min}}}), C_{\min}\right),$$

where

$$p_{\hat{X}^{C_{\min}}} = \arg\min_{p_{\hat{X}}} \quad H(S|\hat{X}) \tag{11a}$$
$$\text{s.t.} \quad \mathbb{E}[\|X - \hat{X}\|^2] \le D. \tag{11b}$$

The minimum classification loss is: $C_{\min} = \sum_s \sum_{\hat{x}^{C_{\min}}} p_{\hat{X}^{C_{\min}}} p_{S|\hat{X}^{C_{\min}}} \log \frac{1}{p_{S|\hat{X}^{C_{\min}}}}$.

Proof. A complete proof is provided in [20, Appendix C]. □

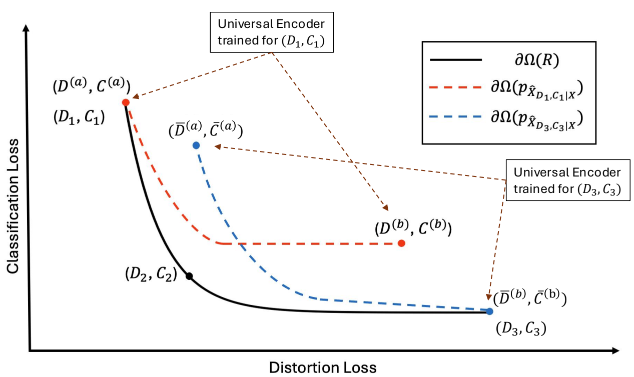

Figure 3: Universality for a general source. Shown are the boundaries of achievable distortion-classification regions for three representations: the minimal distortion point (D_1, C_1), where $R(D_1, C_1) = R(D_1, \infty)$; the midpoint (D_2, C_2); and the minimal classification loss point (D_3, C_3).

To further analyze the structure of the achievable region, consider a point (D, C) on the distortion-classification trade-off curve at a fixed rate R, with an associated optimal reconstruction $Z = \hat{X}_{D,C}$ satisfying $I(X; \hat{X}_{D,C}) = R$, $\mathbb{E}[\|X - \hat{X}_{D,C}\|^2] = D$, and $H(S|\hat{X}_{D,C}) = C$. Assuming that such an optimal reconstruction exists for every point on the trade-off curve and that any further reduction in either D or C would violate the rate constraint, it follows that the point (D, C) lies on the boundary of the closure $\mathrm{cl}(\Omega(p_{\hat{X}_{D,C}|X}))$.

By Theorem 4, the closure of the achievable region includes two extreme points: the upper-left point $(D^{(a)}, C^{(a)})$, which minimizes distortion, and the lower-right point $(D^{(b)}, C^{(b)})$, which minimizes classification loss. Both points are realized by the universal encoder trained at the operating point (D_1, C_1). The region is convex and contains all intermediate pairs. Figure 3 illustrates both $\Omega(R)$ and the achievable region $\Omega(p_{\hat{X}_{D,C}|X})$ for representative points on the trade-off curve. The following theorem quantifies this structure.

Theorem 5. Let $\hat{X}_{D_1,C_1}$ denote the optimal reconstruction at point (D_1, C_1) on the conventional RDC trade-off curve, satisfying $I(X; \hat{X}_{D_1,C_1}) = R(D_1, C_1)$. Then the upper-left extreme point of $\Omega(p_{\hat{X}_{D_1,C_1}|X})$ satisfies $(D^{(a)}, C^{(a)}) = (D_1, C_1)$. Now consider the lower-right extreme points: $(D^{(b)}, C^{(b)}) \in \Omega(p_{\hat{X}_{D_1,C_1}|X})$ and $(D_3, C_3) \in \Omega(R)$, where $C_3 = C_{\min}$ and $R(D_3, C_3) = R(D_1, \infty)$. The distortion gap between these points is bounded below by:

$$D_3 - D^{(b)} \ge \sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - 2\sigma_{\hat{X}_{D_3,C_3}} \sqrt{\sigma_X^2 - D_1} - 2D_1, \tag{12}$$

and the corresponding distortion ratio satisfies:

$$\frac{D_3}{D^{(b)}} \ge \frac{\sigma_X^2 + \sigma^2_{\hat{X}_{D_3,C_3}} - 2\sigma_{\hat{X}_{D_3,C_3}} \sqrt{\sigma_X^2 - D_1}}{2D_1}. \tag{13}$$

In the case where $W_2^2(p_X, p_{\hat{X}_{D_3,C_3}}) = 0$, i.e., $\sigma_X^2 = \sigma^2_{\hat{X}_{D_3,C_3}}$, the distortion gap becomes small under:

$$D_3 - D^{(b)} \approx 0 \ \text{ if } \ D_1 \approx 0 \text{ or } D_1 \approx \sigma_X^2, \tag{14}$$
$$\frac{D_3}{D^{(b)}} \approx 1 \ \text{ if } \ D_1 \approx \sigma_X^2. \tag{15}$$

Proof. The proof is provided in [20, Appendix D]. □

V. Experimental Results

A. Gaussian Zero-Rate Penalty

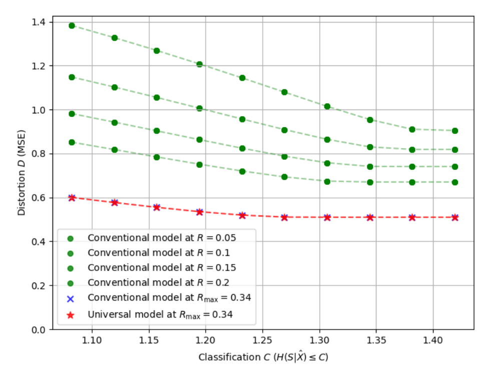

Figure 4: CDR functions for a Gaussian source.

This section verifies that no rate penalty arises when replacing multiple distortion-classification specific Gaussian encoders (i.e. conventional model) with a single universal encoder (i.e., universal model). We consider a scalar Gaussian source $X \sim \mathcal{N}(0, 1)$ and classification variable $S \sim \mathcal{N}(0, 1)$ with correlation $\rho = 0.7$, yielding a maximum rate of $R_{\max} = 0.34$.

The conventional model is evaluated at rates $[0.05, 0.1, 0.15, 0.2, 0.34]$, with each rate generating a distinct set of (C, D) tradeoffs via Theorem 2. In contrast, the universal model uses a fixed encoder at $R_{\max}$ (by Theorem 3) and varies the decoder to explore achievable (C, D) pairs.

Figure 4 illustrates that the universal model closely traces the boundary of the conventional model, thereby confirming that a single encoder is sufficient to achieve the entire classification-distortion tradeoff for Gaussian sources without incurring any rate penalty.

B. Universal Representation for Lossy Compression

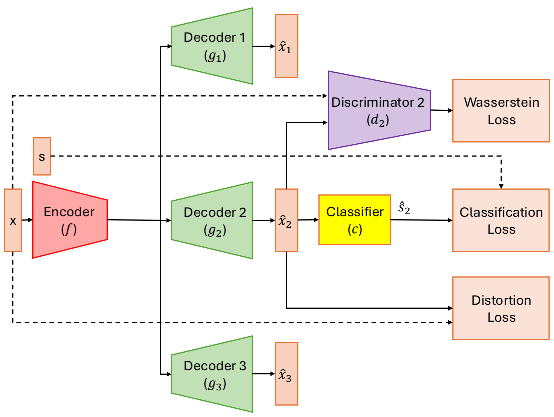

Figure 5: An illustration of the universal RDC scheme.

We empirically validate our theory using a deep learning-based image compression framework on the MNIST dataset, showing that the distortion gap between conventional and universal models aligns with our theoretical predictions.

1) Training: We use a stochastic autoencoder with a pre-trained classifier and GAN-based discriminator, consisting of an encoder f, decoder g, classifier c, and discriminator d, as shown in Figure 5. In the conventional setup, f, g, and d are trainable. The distortion loss is measured by MSE. The output $\hat{X}$ is passed through the classifier c to produce the predicted label distribution $\hat{S}$, with classification loss computed via cross-entropy $\mathrm{CE}(S, \hat{S})$, an upper bound on conditional entropy [21]. The compression rate is upper bounded by $h \times \log_2(L)$, where h is the encoder output size and L the quantization level.

To ensure that the condition $W_2^2(p_X, p_{\hat{X}_{D_3,C_3}}) = 0$ in Theorem 5 is satisfied, we augment the training loss with a squared 2-Wasserstein distance regularization term, rather than relying solely on MSE and cross-entropy losses. Following the approach in [2] and the fact that Wasserstein-2 loss can be bounded by Wasserstein-1 up to a factor (see Appendix B in [22]), we estimate the 2-Wasserstein distance using a discriminator d that takes both X and $\hat{X}$ as inputs and employs the Wasserstein-1 loss in a GAN-based framework. The complete formulation of the loss function is provided below.

$$\mathcal{L} = \lambda_d \, \mathbb{E}[\|X - \hat{X}\|^2] + \lambda_c \, \mathrm{CE}(S, \hat{S}) + \lambda_p \, W_1(p_X, p_{\hat{X}}), \quad (16)$$

where λ_d, λ_c, and λ_p controlling the trade-off.

To construct the universal model, the trained encoder f is frozen, and a new decoder g_1 and discriminator d_1 are trained using:

$$\mathcal{L}_1 = \lambda_d^1 \, \mathbb{E}[\|X - \hat{X}_1\|^2] + \lambda_c^1 \, \mathrm{CE}(S, \hat{S}) + \lambda_p^1 \, W_1(p_X, p_{\hat{X}}), \quad (17)$$

where λ_d^1, λ_c^1, and λ_p^1 adjust the task-specific trade-offs.

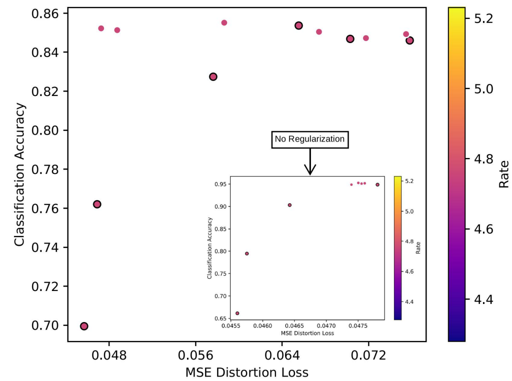

Figure 6: RDC function at $R = 4.75$ on MNIST dataset.

Figure 6 shows the RDC tradeoff on the MNIST dataset at a fixed rate $R = 4.75$, obtained by varying loss coefficients. The black outlined points represent the conventional model trained jointly for specific classification-distortion objectives. Other points correspond to the universal model with decoders trained on a fixed encoder optimized for low classification loss C. The result corresponding to the fixed encoder trained at high C can be analyzed in a similar manner. As expected, in conventional models under a fixed rate constraint, reducing the cross-entropy loss (equivalently, improving classification accuracy) results in increased distortion. This observation highlights the inherent tradeoff between classification performance and reconstruction fidelity. In addition, despite using a fixed encoder, the universal model achieves distortion levels comparable to the conventional model, confirming that an encoder trained for low C can still support diverse tradeoffs through decoder retraining. The sub-figure illustrates the benefits of incorporating the regularization term. These observations support the validity of Theorem 5.

However, a noticeable classification gap remains: universal decoders cannot recover low classification performance if the encoder is trained only for high-distortion objectives. This highlights the decoder's limited generative capacity when the encoder fails to preserve classification-task information.

VI. Conclusion

We proposed a universal RDC framework that enables a single encoder to support multiple task objectives through specialized decoders, removing the need for separate encoders per distortion-classification tradeoff. For the Gaussian source with MSE distortion, we proved that the full RDC region is achievable with zero rate penalty using a fixed encoder. For general source, we characterized the achievable region using MMSE estimation and the 2-Wasserstein distance, identifying conditions under which encoder reuse incurs negligible distortion penalty. Empirical results on the MNIST dataset support our theory, showing that universal encoders, trained with Wasserstein loss regularization, achieve distortion performance comparable to task-specific models. These findings highlight the practicality and effectiveness of universal representations for multi-task lossy compression.

References

[1] Y. Wang, Y. Wu, S. Ma, and Y.-J. A. Zhang, "Lossy compression with data, perception, and classification constraints," in *2024 IEEE Information Theory Workshop (ITW)*, 2024, pp. 366–371.

[2] G. Zhang, J. Qian, J. Chen, and A. Khisti, "Universal rate-distortion-perception representations for lossy compression," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 11 517–11 529.

[3] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 1999.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[5] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. IEEE, 2003, pp. 1398–1402.

[6] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 221–231.

[7] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6228–6237.

[8] ——, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *International Conference on Machine Learning*, 2019, pp. 675–685.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," vol. 27, 2014, pp. 2672–2680.

[10] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017, pp. 214–223.

[11] M. Tschannen, E. Agustsson, and M. Lucic, "Deep generative models for distribution-preserving lossy compression," in *Advances in Neural Information Processing Systems*, 2018, pp. 5929–5940.

[12] S. Nowozin, B. Cseke, and R. Tomioka, "f-gan: training generative neural samplers using variational divergence minimization," in *Advances in Neural Information Processing Systems*, 2016, pp. 271–279.

[13] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International Conference on Machine Learning*, 2016, pp. 1558–1566.

[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[15] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/referenceless image spatial quality evaluator," in *2011 conference record of the forty fifth asilomar conference on signals, systems and computers (ASILOMAR)*. IEEE, 2011, pp. 723–727.

[16] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[17] N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *2015 Twenty First National Conference on Communications (NCC)*. IEEE, 2015, pp. 1–6.

[18] D. Liu, H. Zhang, and Z. Xiong, "On the classification-distortion-perception tradeoff," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[19] Y. Zhang, "A rate-distortion-classification approach for lossy image compression," *Digital Signal Processing*, vol. 141, p. 104163, Sep. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.dsp.2023.104163

[20] N. Nguyen, T. Nguyen, T. Nguyen, and B. Bose, "Universal rate-distortion-classification representations for lossy compression," 2025. [Online]. Available: https://arxiv.org/abs/2504.09025

[21] M. Boudiaf, J. Rony, I. M. Ziko, E. Granger, M. Pedersoli, P. Piantanida, and I. B. Ayed, "A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses," *arXiv preprint*, 2021, arXiv 2003.08983.

[22] B. Lyu, T. Nguyen, P. Ishwar, M. Scheutz, and S. Aeron, "Barycentric-alignment and reconstruction loss minimization for domain generalization," *IEEE Access*, vol. 11, pp. 49 226–49 240, 2023.