Soft then Hard: Rethinking the Quantization in Neural Image Compression

Zongyu Guo [1] **Zhizheng Zhang** [1] **Runsen Feng** [1] **Zhibo Chen** [1]

Abstract

Quantization is one of the core components in lossy image compression. For neural image compression, end-to-end optimization requires differentiable approximations of quantization, which can generally be grouped into three categories: additive uniform noise, straight-through estimator and soft-to-hard annealing. Training with additive uniform noise approximates the quantization error variationally but suffers from the train-test mismatch. The other two methods do not encounter this mismatch but, as shown in this paper, hurt the rate-distortion performance since the latent representation ability is weakened. We thus propose a novel *soft-then-hard* quantization strategy for neural image compression that first learns an expressive latent space softly, then closes the train-test mismatch with hard quantization. In addition, beyond the fixed integer quantization, we apply scaled additive uniform noise to adaptively control the quantization granularity by deriving a new variational upper bound on actual rate. Experiments demonstrate that our proposed methods are easy to adopt, stable to train, and highly effective especially on complex compression models.

1. Introduction

Lossy image compression is a fundamental technique for image transmission and storage. Over the past several years, deep learning methods are reshaping this field. Despite the short history, learned image compression schemes (Toderici et al., 2017; Rippel & Bourdev, 2017; Ballé et al., 2018; Li et al., 2018; Minnen et al., 2018; Lee et al., 2019; Cheng et al., 2020) have surpassed almost all classical standards in terms of rate-distortion performance. Moreover, neural image compression is promising to be more perceptual friendly (Blau & Michaeli, 2019; Mentzer et al., 2020).

[1]University of Science and Technology of China. Correspondence to: Zongyu Guo <guozy@mail.ustc.edu.cn>, Zhibo Chen <chenzhibo@ustc.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Quantization is one of the key challenges for neural image compression. Since the gradient of quantization is zero almost everywhere, it makes the standard back-propagation inapplicable. Although some recent works try to forgo the quantization process entirely (Havasi et al., 2019; Flamich et al., 2020), these methods are computationally costly and statistically inefficient (Agustsson & Theis, 2020). Therefore, quantization remains indispensable for designing an efficient neural image codec. To enable end-to-end optimization, a popular approach is to train with additive uniform noise to approximate the test-time quantization (Ballé et al., 2017). However, this method introduces stochasticity during training, leading to the train-test mismatch and thus hurting the rate-distortion performance in this way.

Other competitive alternatives for quantization include straight-through estimator (STE) with its variants (Bengio et al., 2013; Theis et al., 2017; Mentzer et al., 2018) and recently, soft-to-hard annealing (Agustsson et al., 2017; Yang et al., 2020; Agustsson & Theis, 2020), both of which avoid the mismatch issue. In this paper, we introduce a new analysis of these three quantization methods and argue that:

- Training with STE or soft-to-hard annealing is equal to optimizing a deterministic autoencoder, in which it is hard to learn a smooth latent space due to the lack of regularization term at training (Ghosh et al., 2019).
- STE-based or annealing-based quantization suffers from some training troubles such as biased gradient or unstable gradient, rendering the encoder suboptimal.
- Therefore, these two quantization methods cannot ensure the latent representation ability. Expressive latent variables are significant for compression, where transmitted symbol is expected to convey more information.
- In contrast, optimizing a compression model with additive uniform noise can be interpreted as variational optimization (Ballé et al., 2017) and does not encounter the training troubles. It is superior in learning an expressive latent space, as we demonstrate in this paper.

In short, additive uniform noise is particularly well-suited to train a compression model except for the mismatch between training and test phases. Unlike the posterior and prior mismatch in VAEs (Kingma et al., 2016; Dai & Wipf, 2019),

this mismatch originates from approximating quantization with uniform noise because the quantization error is a deterministic function regarding the signal rather than truly random noise (Gray & Neuhoff, 1998). We thus contribute to remedy this mismatch while maintaining the advantages of additive uniform noise.

Upon our analysis, we propose a novel *soft-then-hard* quantization strategy for neural image compression. Inspired by the two-stage training in recent deep generative models (Van Den Oord et al., 2017; Razavi et al., 2019; Ghosh et al., 2019), we first apply additive uniform noise as a *soft* approximation of quantization to learn a powerful encoder. To close the mismatch caused by the noise-relaxed quantization, we then conduct ex-post tuning for the decoder with *hard* quantization. It allows the decoder to be optimized for the true rate-distortion trade-off without hurting the latent representation ability. We call such proposed technique *soft-then-hard* quantization strategy.

In addition, we propose to use scaled additive uniform noise by deriving a new variational upper bound on actual rate. While previous work is inflexible to control the granularity of quantization, this scaled uniform noise enables the compression model to determine element-wise adaptive quantization step. It reinforces the *soft* noise-relaxed quantization and can be extended to the *hard* tuning stage. As we will show, the commonly used standard uniform noise is a special case of our proposed scaled uniform noise.

The *soft-then-hard* strategy along with the scaled uniform noise is *plug-and-play* to all previous noise-relaxed compression models. Experiments demonstrate that they improve the rate-distortion performance upon different base models (Minnen et al., 2018; Cheng et al., 2020; Guo et al., 2020). Specifically, our new techniques achieve 8.9% BD-rate savings when deployed in (Cheng et al., 2020).

2. Learned Lossy Image Compression

From the view of classical transform coding (Goyal, 2001), prevalent end-to-end optimized lossy image compression framework commonly follows a pipeline consisting of nonlinear transform, quantization and lossless compression. Specifically, a natural image $\boldsymbol{x}$ is first mapped to latent representations $\boldsymbol{y}$, which are then quantized, yielding discrete $\hat{\boldsymbol{y}}$. Since the gradient of quantization is zero almost everywhere, it hinders the back propagation of gradients to the encoder and thus requires differentiable approximations.

2.1. Variational Lossy Image Compression

Most neural image compression methods implement additive uniform noise during training to approximate the test-time quantization. Early works (Ballé et al., 2017; 2018) illustrate the relationship between the rate-distortion objective and variational inference in this noise-relaxed case:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}} D_{\mathrm{KL}}(q(\tilde{\boldsymbol{y}}|\boldsymbol{x})|p(\tilde{\boldsymbol{y}}|\boldsymbol{x})) = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}} \log p(\boldsymbol{x}) + \\ \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}} \mathbb{E}_{\tilde{\boldsymbol{y}}\sim q}[\log q(\tilde{\boldsymbol{y}}|\boldsymbol{x}) - \log p_{\boldsymbol{x}|\tilde{\boldsymbol{y}}}(\boldsymbol{x}|\tilde{\boldsymbol{y}}) - \log p_{\tilde{\boldsymbol{y}}}(\tilde{\boldsymbol{y}})].\end{aligned} \tag{1}$$

The first RHS term is the log likelihood of natural images, which is a constant during optimization since the image x is given in the task of compression. The second RHS term evaluates to zero in the case of additive standard uniform noise (as a stand-in for quantization during training):

$$q(\tilde{\boldsymbol{y}}|\boldsymbol{x}) = q(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \mathcal{U}(\tilde{\boldsymbol{y}}|\boldsymbol{y} - 0.5, \boldsymbol{y} + 0.5) = 1. \tag{2}$$

The rest two terms $-\log p_{\boldsymbol{x}|\tilde{\boldsymbol{y}}}(\boldsymbol{x}|\tilde{\boldsymbol{y}})$ and $-\log p_{\tilde{\boldsymbol{y}}}(\tilde{\boldsymbol{y}})$ in Eq.1 correspond to the weighted *distortion* and the estimated *rate*, respectively. The specific form of distortion is linked to the assumption on $\boldsymbol{x}$, *e.g.*, a squared error loss equals to choosing a Gaussian assumption. But we can generalize Eq.1 to other distortion metrics. Note that the actual rate is the discrete entropy of $\hat{\boldsymbol{y}}$ (at test time, $\hat{\boldsymbol{y}} = \lceil \boldsymbol{y} \rfloor$) that is non-differentiable. Following (Theis et al., 2017), such discrete entropy is upper-bounded by the differential entropy of $\tilde{\boldsymbol{y}}$ during training with Jensen's inequality as [1]:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{y}\sim q}[-\log P(\hat{\boldsymbol{y}})] &\approx \mathbb{E}_{\boldsymbol{y}\sim q}[-\log \int_{[-0.5,0.5]} p(\boldsymbol{y}+\boldsymbol{u})d\boldsymbol{u}] \\ &\leq \mathbb{E}_{\boldsymbol{y}\sim q}[-\int_{[-0.5,0.5]} \log p(\boldsymbol{y}+\boldsymbol{u})d\boldsymbol{u}] \\ &= \mathbb{E}_{\tilde{\boldsymbol{y}}\sim q}[-\log p_{\tilde{\boldsymbol{y}}}(\tilde{\boldsymbol{y}})].\end{aligned} \tag{3}$$

Therefore, minimizing the relaxed differential entropy with distortion is equivalent to minimizing the *upper bound* of the actual rate-distortion value. And the rate-distortion optimization is associated with the goal of variational inference (the LHS term in Eq.1). As shown later in Section 4.2, the standard additive uniform noise here is a special case of our derived scaled uniform noise.

Many neural image compression approaches are built upon this variational compression framework, some of which aim to improve the entropy model (Minnen et al., 2018; Lee et al., 2019; Cheng et al., 2020). All of them apply additive standard uniform noise during training as a soft approximation to hard quantization. At test time, they directly quantize the latents and transmit them with entropy coding algorithms such as arithmetic coding (Witten et al., 1987). Therefore, there is a mismatch between training and test phases, which can be theoretically attributed to the variational relaxation of actual rate, leading to the suboptimal rate-distortion performance. It is unclear how much this mismatch is hurting performance (Agustsson & Theis, 2020).

[1] We extend the derivation in (Theis et al., 2017) with additive uniform noise (although not rigorously proved here). (Ballé et al., 2017) provide a statistical explanation for this inequality.

2.2. Other Quantization Methods

As introduced in Section 2.1, training the lossy image compression model with additive uniform noise approximates the quantization error variationally. In addition, there are two other methods that tackle the non-differentiable issue of quantization in neural image compression. We briefly review them as follows. Note that we focus on the integer quantization and omit the binary quantization in some early works (Toderici et al., 2016; 2017).

Straight-Through Estimator. Straight-through estimator (STE) (Bengio et al., 2013) applies the identity gradients to pass through the hard rounding layer to enable back propagation. A few previous works adopt this method as a trivial replacement of rounding to train a generative model, such as VQ-VAE families (Van Den Oord et al., 2017; Razavi et al., 2019) and integer flow model (Hoogeboom et al., 2019). Some compression works apply hard rounding in the forward pass, but instead use modified gradient in the backward pass (Theis et al., 2017; Mentzer et al., 2018), which we regard as variants of STE. Since the backward and forward passes do not match, the coarse gradient before the quantization layer is certainly not the gradient of loss function. Therefore, taking the biased gradient to update the network brings some underlying problems such as unstable convergence near certain local minima, especially with improper choices of training strategy (Yin et al., 2019).

Soft-to-Hard Annealing. Recently, some annealing-based algorithms are proposed to approximate quantization (Agustsson et al., 2017; Yang et al., 2020; Agustsson & Theis, 2020; Williams et al., 2020). By decreasing the value of a temperature coefficient (Jang et al., 2017), the differentiable approximation function goes towards the shape of hard rounding gradually. Therefore, despite using soft assignment (Agustsson et al., 2017) or soft simulation (Yang et al., 2020; Agustsson & Theis, 2020) initially, these annealing-based quantization methods eventually solve the discrepancy between training and test phases when the temperature is close to zero. However, how to adjust the temperature from soft to hard is empirically determined. As a result, they suffer from fragile training. One latest work observes that annealing-based quantization achieves similar performance compared with STE when applied into integer discrete flow (van den Berg et al., 2021).

3. Analysis of Quantization

Unlike additive uniform noise, quantization with straight-through estimator or soft-to-hard annealing can keep training and test phases consistent because they are eventually optimized for the actual rate-distortion objective. In this section, after we investigate these three quantization methods in detail, we demonstrate that STE-based or even annealing-based quantization deteriorates the latent representation ability. An expressive latent space is extremely important in the task of compression since the transmitted symbols are always expected to convey more effective information.

3.1. Illustrative Task

We start with an illustrative example to investigate the latent representation ability of these three quantization methods: additive uniform noise (AUN), straight-through estimator (STE) and soft-to-hard annealing (here we adopt stochastic Gumbel annealing (Yang et al., 2020), abbreviated as SGA). We are concerned about the situation when the latent dimensionality is rigorously restricted, because it helps us examine the latent expressiveness clearly. A simplified model is used to *compress* the data from MNIST dataset (main task). The specific experimental settings can be found in Appendix A.

We are interested in (i) the reconstruction quality, and (ii) the distribution of latent space. In Figure 1, we visualize the reconstruction results of different methods. We can observe that some numbers are reconstructed to the wrong numbers since the latent dimensionality is restricted and the latent representation ability is limited. The model trained with AUN performs the best including the reconstruction diversity and accuracy. We then show the t-SNE visualization of the latent distribution (Van der Maaten & Hinton, 2008), in order to examine the effects of different quantization methods. Note that we are visualizing the continuous latents that have not been quantized. Compared with the latent space in Figure 1a (trained by AUN), the latent spaces in Figure 1b (STE) and 1c (SGA) are more shallow, especially Figure 1c which tends to collapse to a low-dimensional manifold. It demonstrates that the STE-trained or the SGA-trained model cannot cover enough latent distributions to express all probable contexts. In other words, the latent representation ability of STE-trained or SGA-trained compression model does not compete with AUN-trained model.

Furthermore, in Figure 1d, we visualize the latent space from a model tuned by SGA but pretrained by AUN, which follows the training strategy suggested in (Yang et al., 2020). Although the scope of latent space is almost preserved, we can observe that the latent clusters are scattered and mixed (e.g., the red and the blue clusters), which implies the inaccurate expression of the latent variables.

This illustrative task demonstrates that additive uniform noise is superior in learning an expressive latent space especially when the latent capacity is constrained. A similar observation is found in (Williams et al., 2020), where it is termed as mode-dropping behaviour of STE. However, we show that even the annealing-based quantization will hurt the latent expressiveness in the task of compression.

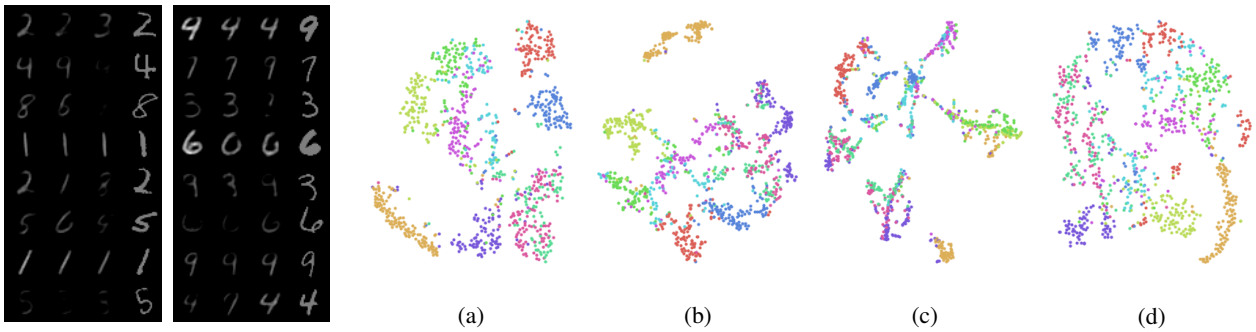


Figure 1. Left pictures: visualizations of reconstructions, where from left to right are the reconstruction results from models trained with AUN, STE, SGA and the ground truth image. (a)-(c) The learned latent distributions by AUN, STE, SGA. (d) The learned latent distribution by tuning the AUN-pretrained model with SGA. Every color represents a category of number in the MNIST dataset.

Table 1. Comparisons of three quantization methods. Improper training strategy may cause unstable convergence of STE-based model (Yin et al., 2019), thus represented as -. Our proposed soft-then-hard (STH) strategy and scaled uniform noise (SUN) are meaningful.

	AUN	STE	Annealing-Based	STH (Ours)	STH + SUN (Ours)
Train-Test Consistency	✗	✓	✓	✓	✓
Latent Expressiveness	✓	✗	✗	✓	✓
Variational Compression	✓	✗	✗	✓	✓(more flexible)
Exact Gradient	✓	✗	✓	✓	✓
Stable Training	✓	-	✗	✓	✓

3.2. Variational or Deterministic?

Theoretically, if applying additive uniform noise as an approximation of quantization, the rate-distortion objective of compression is associated with the goal of variational inference, as illustrated in Eq.1 (Ballé et al., 2017; 2018). Optimizing a compression model with additive uniform noise is thereby equal to learning a variational autoencoder (Kingma & Welling, 2014). However, in the case of STE-based or annealing-based quantization, the rate-distortion optimization is not variational any more since the second RHS term in Eq.1 (now is $\log q(\hat{\boldsymbol{y}}|\boldsymbol{x})$) is not always zero. Training a compression model with STE or annealing degrades to optimizing a deterministic autoencoder (Hinton & Salakhutdinov, 2006). From another view, the additive uniform noise works as a regularization term for variational training, which is beneficial to learn a smooth and expressive latent space (Ghosh et al., 2019).

3.3. Summary of Three Quantization Methods

As mentioned in Section 3.2, training a compression model with STE or annealing degrades to optimizing a deterministic autoencoder. Lack of regularization term during training is one of the reasons for the weak latent representation ability. In addition, STE takes the biased gradient for optimization and results in searching in the negative direction (Yin et al., 2019). And soft-to-hard annealing will suffer from unstable training caused by infinite gradient when the temperature coefficient is closed to zero. Even if recent work tries to reduce the variance of gradients by calculating the expectation of gradients (Agustsson & Theis, 2020), it requires to impose some assumptions and fails when bitrate is high. The issue of biased gradient or unstable gradient renders the encoder suboptimal, which is another reason for the inexpressive latent space.

In contrast, applying additive uniform noise (AUN) as a quantization approximation is superior in learning an expressive latent space. That is because: (i) The noise injection mechanism works as a regularization term to aid variational learning, which thus ensures the smoothness of latent space. (ii) Applying AUN makes the training process stable with exact gradient backward. Consequently, the encoder is optimized properly to be powerful enough. But AUN still encounters the mismatch between training and test phases, resulting in rate-distortion performance degradation.

We summarize the properties of these three quantization methods as shown in Table 1. In short, none of them enable the neural compression model to simultaneously achieve an expressive latent space and the train-test consistency. As a result, they cannot achieve the optimal rate-distortion performance for compression. In the following section, we will introduce our proposed *soft-then-hard* (STH) strategy that is able to solve the train-test mismatch of AUN-based quantization while preserving the latent representation ability. In addition, we derive a new variational upper bound on actual

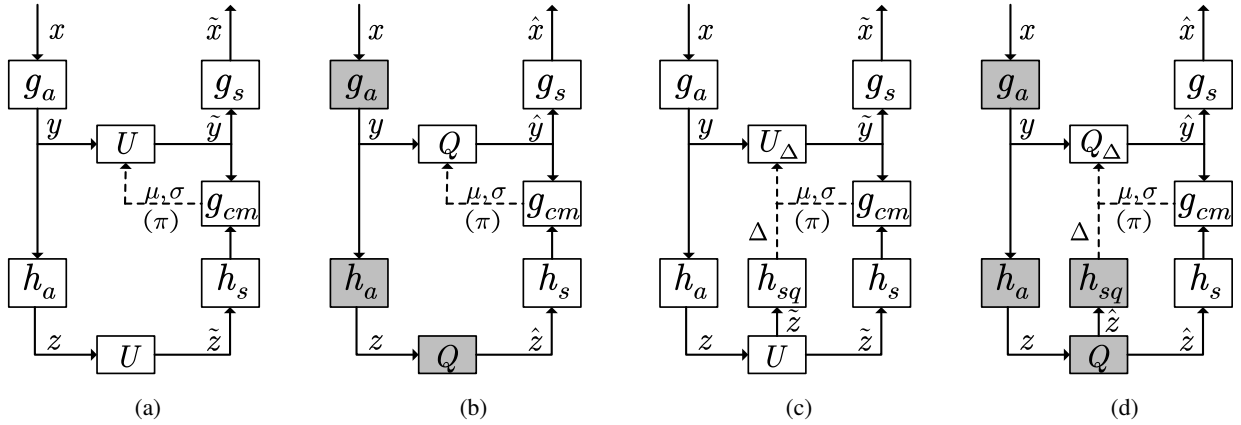


Figure 2. (a) Applying additive uniform noise at the soft training stage, same as previous work (Minnen et al., 2018). (b) Applying hard quantization at the ex-post tuning stage in our soft-then-hard strategy. (c) Flexible quantization with element-wise noise scale Δ. (d) Combining our proposed two methods together. The gray boxes represent the components that are fixed at the ex-post tuning stage.

rate that incorporates scaled uniform noise (SUN) for more flexible quantization.

4. Proposed Methods

4.1. Soft-then-Hard Strategy

We propose a novel *soft-then-hard* (STH) strategy for neural image compression that contains a two-stage training process. By using this strategy, the compression model will first learn a powerful encoder with additive uniform noise that simulates the quantization layer in a soft manner. The train-test mismatch is then solved through ex-post tuning of decoder with hard quantization.

Figure 2a presents the training process of a conventional image compression model that is trained with additive uniform noise (Minnen et al., 2018). It consists of an analysis encoder $\boldsymbol{g_a}$, a synthesis decoder $\boldsymbol{g_s}$, a hyper analysis encoder $\boldsymbol{h_a}$, a hyper synthesis decoder $\boldsymbol{h_s}$ and a context model $\boldsymbol{g_{cm}}$ that generates latent distribution parameters, *i.e.*, μ, σ in (Minnen et al., 2018) and π, μ, σ in (Cheng et al., 2020). During training, additive uniform noise is added to both $\boldsymbol{y}$ and $\boldsymbol{z}$ for end-to-end optimization. This soft noise-relaxed quantization is denoted as U in Figure 2a.

We take the abovementioned process as the soft training stage, which is able to learn a powerful encoder. After obtaining a pretrained model, the encoders $\boldsymbol{g_a}$ and $\boldsymbol{h_a}$ will not participate in the second tuning stage. As Figure 2b shown, we then directly quantize the latents $\boldsymbol{y}$, $\boldsymbol{z}$ to $\hat{\boldsymbol{y}}$, $\hat{\boldsymbol{z}}$ as the input data to tune the rest components of the compression model. The hard quantization here is denoted as Q in Figure 2b. For simplicity, the non-parametric density estimation network of $\hat{\boldsymbol{z}}$ is also fixed, which is observed to have negligible influences experimentally.

Since the encoders are fixed at this tuning stage, the learned latent variables will not be changed and thus the latent representation ability is preserved. At this stage, the mismatch issue of AUN-based quantization is solved, as now we are minimizing the exact discrete entropy along with the true distortion:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{y}\sim q}[-\log P(\hat{\boldsymbol{y}}) - \log p_{\boldsymbol{x}|\hat{\boldsymbol{y}}}(\boldsymbol{x}|\hat{\boldsymbol{y}})]. \tag{4}$$

Actually, by detaching the decoder from the encoder, the entire ex-post tuning stage can be regarded as a joint optimization of two independent tasks: treating $P(\hat{\boldsymbol{y}})$ as a new base distribution to optimize a reconstruction (generation) model, and learning a prior likelihood model to estimate the discrete latent distribution $P(\hat{\boldsymbol{y}})$.

If the encoders are not fixed, the second tuning stage formally equals to train a compression model with STE, which will fail to learn expressive latent variables as discussed in Section 3. In short, our proposed soft-then-hard strategy circumvents the trade-off between latent expressiveness and quantization mismatch. It is simple yet effective and does not require additional parameters.

4.2. Scaled Uniform Noise

In our proposed *soft-then-hard* strategy, the compression model takes additive uniform noise to replace hard rounding at the soft training stage. While adding standard uniform noise successfully approximates the integer quantization error and associates the rate-distortion optimization with variational inference, it is inflexible to control the granularity of quantization. To sidestep this issue, we propose to learn the scale of uniform noise during training by deriving a new variational upper bound on actual rate. At test time, the adaptive noise scale will determine the quantization step.

As shown in Figure 2c, a new branch h_{sq} will generate the

noise scale Δ from $\tilde{\boldsymbol{z}}$. The noise scale is *element-wise* adaptive that is encoded / decoded from hyperprior in advance. It enables the model to determine a consistent quantization granularity for arithmetic coding at both encoder and decoder. Then, we have $\tilde{\boldsymbol{y}}$ as the summation of $\boldsymbol{y}$ and a random scaled uniform noise in the interval $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$ as

$$\begin{aligned} \Delta &= h_{sq}(\tilde{\boldsymbol{z}}), \\ \tilde{\boldsymbol{y}} &= \boldsymbol{y} + \boldsymbol{u}, \boldsymbol{u} \sim \mathcal{U}(\frac{\Delta}{2}, \frac{\Delta}{2}), \\ q(\tilde{\boldsymbol{y}}|\boldsymbol{x}) &= q(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = q(\boldsymbol{u}|\boldsymbol{y}) = \frac{1}{\Delta}, \\ p_{\tilde{\boldsymbol{y}}}(\tilde{\boldsymbol{y}}) &= \int_{\tilde{\boldsymbol{y}}-\frac{\Delta}{2}}^{\tilde{\boldsymbol{y}}+\frac{\Delta}{2}} \frac{1}{\Delta} p(\boldsymbol{y}) d\boldsymbol{y}. \end{aligned} \tag{5}$$

At test time, the learnable noise scale determines quantization step, generating $\hat{\boldsymbol{y}}$ from $\boldsymbol{y}$:

$$\begin{aligned} \Delta &= h_{sq}(\hat{\boldsymbol{z}}), \\ \hat{\boldsymbol{y}} &= \Delta \cdot \lceil \frac{\boldsymbol{y}}{\Delta} \rfloor. \end{aligned} \tag{6}$$

Since the additive uniform noise here does not subject to standard uniform distribution, the derivation in Eq.3 should be modified. We derive a new variational upper bound on actual rate, which holds for scaled uniform noise $q(\boldsymbol{u}|\boldsymbol{y})$ as

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{y}\sim q}[-\log P(\hat{\boldsymbol{y}})] \\ &\approx \mathbb{E}_{\boldsymbol{y}\sim q}[-\log \int_{[-\frac{\Delta}{2}, \frac{\Delta}{2}]} p(\boldsymbol{y}+\boldsymbol{u}) d\boldsymbol{u}] \\ &= \mathbb{E}_{\boldsymbol{y}\sim q}[-\log \int_{[-\frac{\Delta}{2}, \frac{\Delta}{2}]} q(\boldsymbol{u}|\boldsymbol{y}) \frac{p(\boldsymbol{y}+\boldsymbol{u})}{q(\boldsymbol{u}|\boldsymbol{y})} d\boldsymbol{u}] \\ &\leq \mathbb{E}_{\boldsymbol{y}\sim q}[-\int_{[-\frac{\Delta}{2}, \frac{\Delta}{2}]} q(\boldsymbol{u}|\boldsymbol{y}) \log \frac{p(\boldsymbol{y}+\boldsymbol{u})}{q(\boldsymbol{u}|\boldsymbol{y})} d\boldsymbol{u}] \\ &= \mathbb{E}_{\tilde{\boldsymbol{y}}\sim q}[-\log \frac{p_{\tilde{\boldsymbol{y}}}(\tilde{\boldsymbol{y}})}{q(\boldsymbol{u}|\boldsymbol{y})}] \\ &= \mathbb{E}_{\tilde{\boldsymbol{y}}\sim q}[\log q(\tilde{\boldsymbol{y}}|\boldsymbol{x}) - \log p_{\tilde{\boldsymbol{y}}}(\tilde{\boldsymbol{y}})]. \end{aligned} \tag{7}$$

Therefore, the rate-distortion optimization still conforms the goal of variational inference as Eq.1 shown. The true training objective which relates to the rate term is:

$$\begin{aligned} \mathcal{L}_{rate} &= \mathbb{E}_{\tilde{\boldsymbol{y}}\sim q}[-\log \int_{\tilde{\boldsymbol{y}}-\frac{\Delta}{2}}^{\tilde{\boldsymbol{y}}+\frac{\Delta}{2}} p(\boldsymbol{y}) d\boldsymbol{y}] \\ &= \mathbb{E}_{\tilde{\boldsymbol{y}}\sim q}[\log \frac{1}{\Delta} - \log p_{\tilde{\boldsymbol{y}}}(\tilde{\boldsymbol{y}})] \\ &= \mathbb{E}_{\tilde{\boldsymbol{y}}\sim q}[-\log \frac{p_{\tilde{\boldsymbol{y}}}(\tilde{\boldsymbol{y}})}{q(\boldsymbol{u}|\boldsymbol{y})}] \\ &\geq \mathbb{E}_{\boldsymbol{y}\sim q}[-\log P(\hat{\boldsymbol{y}})]. \end{aligned} \tag{8}$$

Notice that the additive standard uniform noise in Eq.3 is a special case of our proposed scaled uniform noise. The noise scale allows the compression model to determine quantization granularity that is adaptive to image contexts.

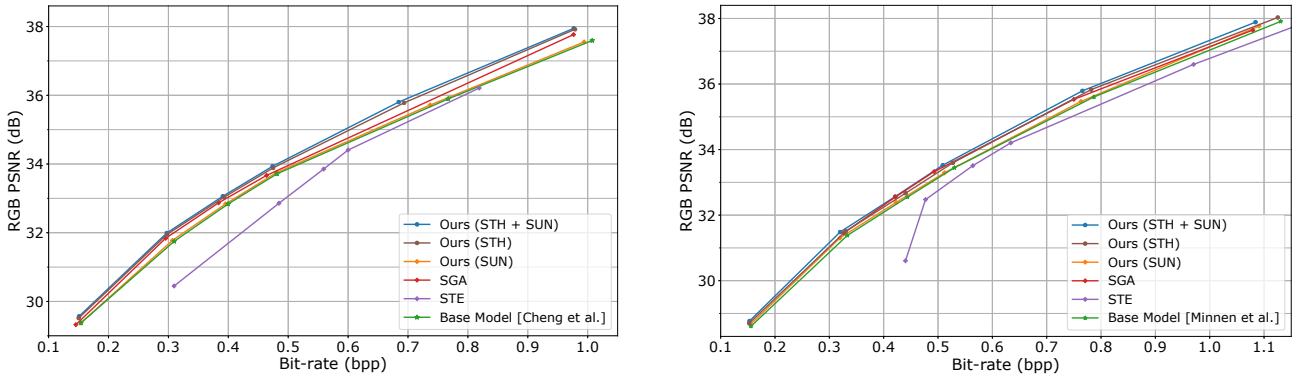
The proposed scaled uniform noise (SUN) here works at the soft quantization stage and can be extended to the hard tuning stage. As shown in Figure 2d, when conducting ex-post tuning with scaled uniform noise, the branch h_{sq} that learns noise scale is also fixed. The latent variables are then quantized according to the scale values. Experiments demonstrate that coupling our proposed two methods improves performance by a large margin.

4.3. Related Works and Discussions

In the field of neural image compression, many works apply additive uniform noise and suffer from the mismatch between training and test phases. Most of them are dedicated to improving the entropy model. After (Ballé et al., 2018) propose the hierarchical entropy model, the works of (Minnen et al., 2018; Lee et al., 2019) design an autoregressive context model (Chen et al., 2017; Van Oord et al., 2016) to capture local correlations among the latent variables. The latent distribution is also improved from zero-mean Gaussian scale model (Ballé et al., 2018) to single Gaussian in (Minnen et al., 2018; Lee et al., 2019), and recently, Gaussian mixture model (Cheng et al., 2020).

In addition to them, some works are closely related to our new ideas. VQ-VAEs are perhaps the most successful VAE architectures for image and audio generations (Van Den Oord et al., 2017; Razavi et al., 2019). In (Ghosh et al., 2019), the authors argue that despite the name, VQ-VAEs are neither stochastic, nor variational, but they are deterministic. The work of (Ghosh et al., 2019) applies a regularization term to learn a meaningful latent space with ex-post density estimation. All of these works empirically optimize the models with two-stage training. Recently, VQ-VAE is deployed with the annealing-based quantization to mitigate the mode-dropping issue caused by STE (Williams et al., 2020). However, it is observed that the annealing-based method would perform similarly as STE (van den Berg et al., 2021), which corresponds to our analysis in Section 3. Our proposed soft-then-hard strategy is inspired by them, reasonably suitable for the task of image compression.

As for our proposed scaled uniform noise, it is inspired by the variational dequantization in flow models (Ho et al., 2019). Unlike the factorized noise in (Ho et al., 2019), we derive a variational upper bound on actual rate to enable flexible quantization, where the additive noise is still uniform and the model learns the noise scale. The work of (Choi et al., 2019) employs a group of pre-defined quantization steps with universal quantization, and builds a variable rate compression network. However, the quantization step in their work is fixed at every single bitrate that does not adapt to images. In our work, we believe that flexible quantization



(a) (b)

Figure 3. Ablation results on Kodak dataset. (a) Base model is (Cheng et al., 2020). (b) Base model is (Minnen et al., 2018).

step is important to support spatial bit allocation, because even in traditional compression, different frequency bases are assigned with different quantization steps to adapt image contexts (Sullivan et al., 2012).

5. Experiments

Both our proposed two methods are *plug-and-play*, compatible with all previous noise-relaxed lossy image compression models. Only the new branch h_{sq} that learns noise scales requires minor additional parameters. Experimentally, we observe that the mismatch issue between training and test phases deteriorates the compression performance more obviously when the compression model is more complex (see Appendix B for empirical evidence). Therefore, we evaluate our proposed new techniques upon different base models: a simplified model (Minnen et al., 2018) and two powerful models (Cheng et al., 2020; Guo et al., 2020). These models are entirely reproduced by us unless otherwise stated. For fair comparison, we keep all experimental conditions as the same as possible (*e.g.*, training data is not public in some previous works). We train the compression models on the full ImageNet training set (Deng et al., 2009) and test the rate-distortion performance on Kodak dataset (Kodak, 1993), a widely used dataset for evaluating the performance of image compression model. Other experimental details are presented in Appendix C including the specific structures of the scale generation branch h_{sq}.

5.1. Ablations

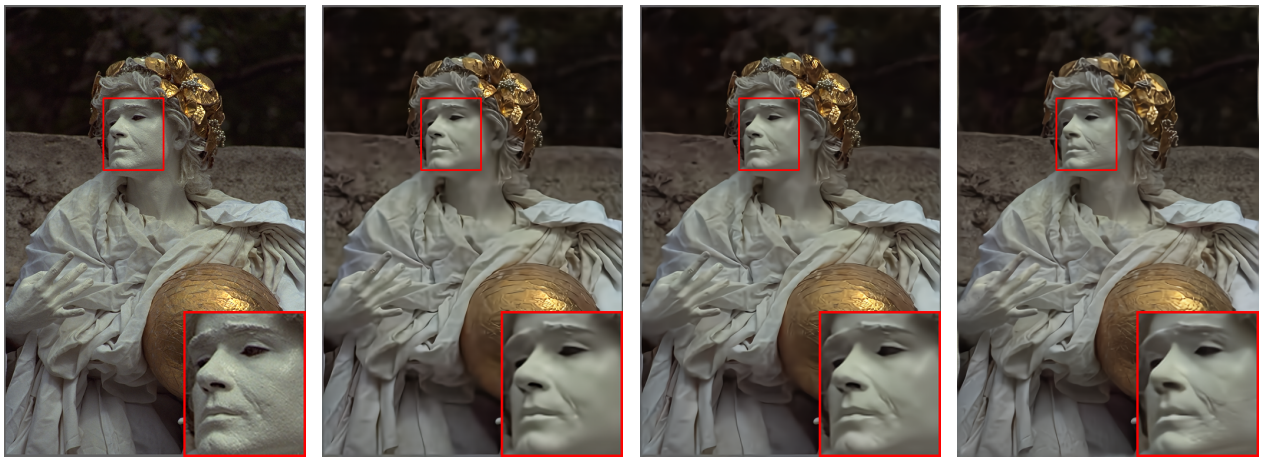
We study the effectiveness of our proposed soft-then-hard (STH) strategy and scaled uniform noise (SUN) through extensive ablation experiments. We first compute the rate-distortion curves in terms of bit per pixel (bpp) versus peak signal-to-noise ratio (PSNR). The loss function now is

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{rate} + \lambda \cdot \mathcal{L}_{distortion} \\ &= \mathcal{L}_{rate} + \lambda \cdot \mathcal{L}_{MSE}(\hat{\boldsymbol{x}}, \boldsymbol{x}).\end{aligned} \tag{9}$$

We adjust the Lagrange Multiplier λ from 192 to 4096 and obtain six models at multiple bitrates.

Baseline-1. The compression model in (Cheng et al., 2020) is a powerful codec that is trained with additive uniform noise (AUN). As shown in Figure 3a, STH brings obvious gains and SUN improves the performance marginally (see Appendix D for zoom-in RD-curves). However, we would like to emphasize that SUN provides a novel mechanism for spatial bit allocation, which is promising for variable rate compression (*e.g.*, using multiple noise generation branches in one model). Statistically, employing our proposed two techniques together achieves 8.9% BD-rate savings compared with the base model. If we replace the AUN-based quantization by STE (Bengio et al., 2013), the performance drops a lot especially at low bitrates. Tuning the AUN-pretrained model by stochastic Gumbel annealing (SGA), as suggested in (Yang et al., 2020), improves performance as well, but still has a gap to the combination of STH + SUN (even have a gap to STH alone). Note that some points of STE and SGA are missing due to unstable training.

Baseline-2. Figure 3b presents the results upon a base model with relatively weak performance (Minnen et al., 2018). We find that loading the AUN-pretrained model and tuning with SGA cannot converge here. Tuning the STE-pretrained model with SGA somehow achieves good performance (the gray line) at some bitrates. This implies that SGA struggles with fragile training. The STE-trained model encounters instability issue as well and always presents the worst rate-distortion performance. In contrast, our proposed methods, both STH and SUN, improve the performance stably. They show similar albeit weaker effects compared with the strong baseline as in Figure 3a.



Ground Truth (a) 0.210 / 32.80 / 0.9689 (b) 0.203 / **33.15** / 0.9710 (c) 0.169 / 28.86 / **0.9767**

Figure 4. Qualitative comparisons. (a) Base model (Cheng et al., 2020) optimized for PSNR. (b) Employing our methods optimized for PSNR. (c) Employing our methods optimized for MS-SSIM. The statistics are the values of bit-rate (bpp) / PSNR (dB) / MS-SSIM.

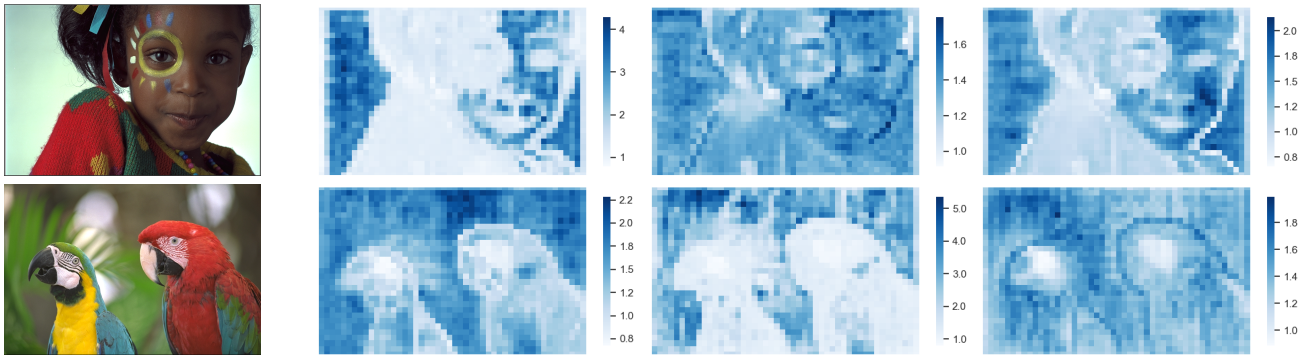


Figure 5. Visualizations of the noise scale. Left: ground truth. Right three columns: noise scale in different channels.

Other Experiments. We also evaluate our proposed two techniques by employing them in another powerful model (Guo et al., 2020), the results of which are shown in Appendix D. Deploying our methods in such more powerful compression model delivers the state-of-the-art image compression results, outperforming VVC (Sullivan & Ohm, 2018), the latest compression standard. These three groups of experiments demonstrate the robustness of our proposed methods. In addition, the soft-then-hard strategy solves the train-test mismatch and improves performance stably at all bitrates, different from (Agustsson & Theis, 2020), where the annealed universal quantization would hurt RD performance at high bitrates. When optimized for MS-SSIM (Wang et al., 2004), our methods contribute stable improvements as well, which is shown in Appendix D.

5.2. Visualizations

Reconstructions. As shown in Figure 4, we visualize the reconstruction results when the base model is (Cheng et al., 2020). Compared with the base model, employing our methods (STH + SUN) improves compression performance both quantitatively and qualitatively. When optimized for MS-SSIM, it achieves more pleasant perceptual quality. More visualizations are provided in Appendix E.

Scaled Uniform Noise. In addition, our proposed scaled uniform noise is element-wise adaptive to each latent variable. We also visualize the noise scale across different channels as presented in Figure 5. Notice that the learned noise scale is highly correlated to image textures and covers different contexts across channels. It verifies that our derived scaled uniform noise is effective to generate image-adaptive quantization step.

6. Conclusion

In this paper, we rethink the three quantization methods that are implemented as differentiable approximations for neural image compression. Among them, we demonstrate that additive uniform noise is superior to STE-based and even annealing-based quantization in terms of the latent

representation ability through our detailed analysis. We propose a novel soft-then-hard quantization strategy that achieves train-test consistency and latent expressiveness simultaneously. We also derive a new variational upper bound on actual rate that incorporates the scale of additive uniform noise into optimization and thus enable flexible quantization. Our proposed two methods are simple yet effective, achieving stable improvements at any bitrates.

Acknowledgements

Zhibo Chen is the corresponding author. This work was supported in part by NSFC under Grant U1908209, 61632001, and the National Key Research and Development Program of China 2018AAA0101400.

References

Agustsson, E. and Theis, L. Universally quantized neural compression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12367–12376, 2020.

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 *Advances in Neural Information Processing Systems 30*, pp. 1141–1151, 2017.

Ballé, J., Laparra, V., and Simoncelli, E. P. End-to-end optimized image compression. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. Variational image compression with a scale hyperprior. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

Bengio, Y., Léonard, N., and Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

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

Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., and Abbeel, P. Variational lossy autoencoder. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7939–7948, 2020.

Choi, Y., El-Khamy, M., and Lee, J. Variable rate deep image compression with a conditional autoencoder. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3146–3154, 2019.

Dai, B. and Wipf, D. P. Diagnosing and enhancing VAE models. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Flamich, G., Havasi, M., and Hernández-Lobato, J. M. Compressing images by encoding their latent representations with relative entropy coding. In *Advances in Neural Information Processing Systems 33*, 2020.

Ghosh, P., Sajjadi, M. S., Vergari, A., Black, M., and Scholkopf, B. From variational to deterministic autoencoders. In *International Conference on Learning Representations*, 2019.

Goyal, V. K. Theoretical foundations of transform coding. *IEEE Signal Processing Magazine*, 18(5):9–21, 2001.

Gray, R. M. and Neuhoff, D. L. Quantization. *IEEE transactions on information theory*, 44(6):2325–2383, 1998.

Guo, Z., Wu, Y., Feng, R., Zhang, Z., and Chen, Z. 3-d context entropy model for improved practical image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 116–117, 2020.

Havasi, M., Peharz, R., and Hernández-Lobato, J. M. Minimal random code learning: Getting bits back from compressed model parameters. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science*, 313 (5786):504–507, 2006.

Ho, J., Chen, X., Srinivas, A., Duan, Y., and Abbeel, P. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning*, pp. 2722–2730. PMLR, 2019.

Hoogeboom, E., Peters, J., van den Berg, R., and Welling, M. Integer discrete flows and lossless compression. In *Advances in Neural Information Processing Systems*, pp. 12134–12144, 2019.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29:4743–4751, 2016.

Kodak, E. Kodak Lossless True Color Image Suite (PhotoCD PCD0992). http://r0k.us/graphics/kodak/, 1993.

Lee, J., Cho, S., and Beack, S. Context-adaptive entropy model for end-to-end optimized image compression. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

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*, pp. 3214–3223, 2018.

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*, pp. 4394–4402, 2018.

Mentzer, F., Toderici, G. D., Tschannen, M., and Agustsson, E. High-fidelity generative image compression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11913–11924, 2020.

Minnen, D., Ballé, J., and Toderici, G. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems 31*, pp. 10771–10780, 2018.

Razavi, A., van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems 32*, pp. 14837–14847, 2019.

Rippel, O. and Bourdev, L. Real-time adaptive image compression. In *International Conference on Machine Learning*, pp. 2922–2930. PMLR, 2017.

Sullivan, G. J. and Ohm, J.-R. Versatile video coding towards the next generation of video compression. In *2018 Picture Coding Symposium (PCS)*, 2018.

Sullivan, G. J., Ohm, J.-R., Han, W.-J., and Wiegand, T. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012.

Theis, L., Shi, W., Cunningham, A., and Huszár, F. Lossy image compression with compressive autoencoders. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

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 *4th International Conference on Learning Representations, ICLR 2016*, 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*, pp. 5306–5314, 2017.

van den Berg, R., Gritsenko, A. A., Dehghani, M., Kaae Sønderby, C., and Salimans, T. Idf++: Analyzing and improving integer discrete flows for lossless compression. In *9th International Conference on Learning Representations, ICLR 2021*, 2021.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pp. 6306–6315, 2017.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Van Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pp. 1747–1756. PMLR, 2016.

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*, 13(4):600–612, 2004.

Williams, W., Ringer, S., Ash, T., MacLeod, D., Dougherty, J., and Hughes, J. Hierarchical quantized autoencoders. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4524–4535, 2020.

Witten, I. H., Neal, R. M., and Cleary, J. G. Arithmetic coding for data compression. *Communications of the ACM*, 30(6):520–540, 1987.

Yang, Y., Bamler, R., and Mandt, S. Improving inference for neural image compression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 573–584, 2020.

Yin, P., Lyu, J., Zhang, S., Osher, S. J., Qi, Y., and Xin, J. Understanding straight-through estimator in training activation quantized neural nets. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.