Noise-to-Compression Variational Autoencoder for Efficient End-to-End Optimized Image Coding

Jixiang Luo*, Shaohui Li*, Wenrui Dai*§, Yuhui Xu*, De Cheng†, Gang Li†, and Hongkai Xiong*

*School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (§Corresponding author)
{ljx123456,lishaohui,daiwenrui,yuhuixu,xionghongkai}@sjtu.edu.cn
†Algorithm Innovation Lab, Huawei Cloud, Xi'an, Shanxi 710075, China
{chengde,ligang123}@huawei.com

Abstract

Generative model has emerged as a disruptive alternative for lossy compression of natural images, but suffers from the low-fidelity reconstruction. In this paper, we propose a noise-to-compression variational antoencoder (NC-VAE) to achieve efficient rate-distortion optimization (RDO) for end-to-end optimized image compression with a guarantee of fidelity. The proposed NC-VAE improves rate-distortion performance by adaptively adjusting the distribution of latent variables with trainable noise perturbation. Consequently, high-efficiency RDO is developed based on the distribution of latent variables for simplified decoder. Furthermore, robust end-to-end learning is developed over the corrupted inputs to suppress the deformation and color drift in standard VAE based generative models. Experimental results show that NC-VAE outperforms the state-of-the-art lossy image coders and recent end-to-end optimized compression methods in low bit-rate region, i.e., below 0.2 bits per pixel (bpp).

1. Introduction

It has been a lasting challenge to compress rapidly increasing image and video contents to accommodate limited storage space and network bandwidth. Conventional transform coding methods achieve state-of-the-art rate-distortion performance with well-designed modules of prediction, transform, quantization and entropy coding in a subsequence. JPEG [1] adopts discrete cosine transform (DCT) to transform image residues into frequency domain where cascading scalar quantization is implemented. JPEG 2000 [2] leverages discrete wavelet transform (DWT) to obtain compact and progressive representation for image compression. Better Portable Graphics (BPG) and WebP consider mode-based spatial prediction based on intra-frame coding tools in High-Efficiency Video Coding (HEVC) standard [3] and VP8 format [4]. However, these handcrafted coders cannot sufficiently exploit varying statistics within images.

With the rise of deep learning, end-to-end learning frameworks, including autoencoders, recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have been popular for accurate modeling of correlations within natural images. These models simultaneously optimize transform, quantization and coding via deep neural networks. In comparison to handcrafted transform coding schemes, end-to-end learning frameworks achieve gains in compression performance, but suffer from high

computational complexity and storage cost led by the huge amount of model parameters. Recently, extreme image compression is achieve using generative adversarial networks (GAN) [5], but its reconstruction fidelity cannot be guaranteed under the metrics of PSNR and MS-SSIM. To incorporate generative models in the end-to-end learning framework, however, computation intensive prediction network is required at the encoder and decoder side to address deformation and color drift [6].

In this paper, we propose an end-to-end optimized image compression scheme based on variational autoencoder perturbed with noise. The proposed noise-to-compression variational autoencoder (NC-VAE) leverages an efficient rate-distortion optimization (RDO) to improve decoding efficiency and compression performance. To be concrete, it minimizes the reconstruction distortion with a trainable perturbation noise and leverages the prior knowledge for latent variables to constrain the bit-rate. NC-VAE adopts an efficient variational inference based to realize RDO, rather than complicated deep prediction networks for latent variables. Furthermore, deformation and color drift in reconstruction can be suppressed using inputs with noise for NC-VAE. Experimental results demonstrate that NC-VAE yields an improved compression performance at the low bit-rate region (below 0.2 bpp).

The rest of this paper is organized as follows. Section 2 overviews the end-to-end learning frameworks and generative models for image compression. Section 3 proposes the NC-VAE model for lossy image compression, including analysis of effect of noise and theoretic relationship between prior knowledge and bit-rate. Section 4 evaluates NC-VAE on *Kodak24* dataset. Finally, Section 5 draws the conclusions.

2. Related Works

2.1 End-to-End Optimized Image Compression

Deep learning has been facilitating image compression with the end-to-end optimization. Toderici *et al.* [7] realized variable rate compression with recurrent neural networks (RNNs) to avoid the restriction led by latent variables (bottlenecks) for coding. Ballé *et al.* [8] leveraged a single network with generalized divisive normalization (GDN) to achieve the end-to-end optimization for effective image compression. However, its quantization with uniform noise is resource demanding. Theis *et al.* [9] incorporated the autoencoder for lossy compression of bottlenecks. Li *et al.* [10] adaptively generated a series of masks according to the image contents to improve compression peformance. Rippel and Bourdev [11] introduced hierarchical features in convolutional layers to extract features and regularize the optimizaiton of codec for an improved MS-SSIM performance. Ballé *et al.* [6] developed a convolutional neural network to analyze the relation between latent variables and their scale hyperpriors for better compression. Minnen *et al.* [12] considered a sophisticated prior to improve rate-distortion performance on *Kodak24* and *Tecnick* datasets, but limited by computational resources required for training on massive data. In recent Challenge on Learned Image Compression (CLIC) [1], Tucodec [13]achieved superior performance

[1] `http://www.compression.cc/challenge`.

using end-to-end learning framework, but still suffered from high computational complexity and large decoder size.

2.2 Generative Models

Generative models have emerged as a disruptive alternative for image compression that generate similar images to the inputs with a small amount of information. PixelCNN [14] and PixelRNN [15] implemented direct arithmetic coding of image pixels without requiring transform. In [16], PixelCNN was adopted to form two kinds of 3D context model for probabilistic estimation. Generative adversarial networks were incorporated to achieve extremely low bitrate coding by generating images from least retained information [5].

Variational autoencoder (VAE) achieves image generation by assuming the input x is extracted from the latent variables z. It approaches the prior distribution $p(x|z)$ with the learned posterior distribution $p(z|x)$ by maximizing the Evidence Lower BOund (ELBO) $\mathbb{E}_{z\sim p(z|x)}[-\ln q(x|z)]$. Thus, the loss function for training VAE is formulated in Eq. (1).

$$\mathcal{L} = \mathbb{E}_{x\sim\tilde{p}(x)}\left[\mathbb{E}_{z\sim p(z|x)}[-\ln q(x|z)] + \mathbb{E}_{z\sim p(z|x)}\left[\ln\frac{p(z|x)}{q(z)}\right]\right] \tag{1}$$

Here, z is commonly initialized with a normal distribution $q(z)$ for training. According to Equation (1), the encoder infers z from x based on $p(z|x)$ and the decoder generates the output $\hat{x}$ with the prior distribution $q(x|z)$. Inspired by denosing autoencoder [17], this paper proposes a noise-to-compression variational autoencoder (NC-VAE) to eliminate the deformation and guarantee the fidelity in PSNR in reconstruction of compressed images.

3. Noise-to-Compression Variational Autoencoder

Figure 1 depicts the proposed architecture for NC-VAE. We develop an end-to-end optimized image compression with noise for latent variables and achieve efficient rate-distortion optimization based on KL divergence between the encoder and decoder. Different from existing end-to-end learning framework, NC-VAE is lightweight to achieve efficient image compression with a guaranteed rate-distortion performance, especially in the low bit-rate region.

3.1 Compression with Noise

Figure 2 shows that standard VAE would lead to deformation and color drift in image reconstruction, especially for low bit-rate region. This problem would degrade the compression performance due to the distortion obscures the fidelity of reconstructed images. In this paper, we propose to eliminate the deformation and color drift by training over the images with perturbation noise.

Let us define the corrupted input $\tilde{x} = x + \epsilon$ generated by adding Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma)$ to the original image x. We can obtain the distribution $p(\tilde{x})$ of $\tilde{x}$ by

$$p\left(\tilde{x}\right) = p\left(x\right) * \mathcal{N}\left(0, \sigma^2\right), \tag{2}$$

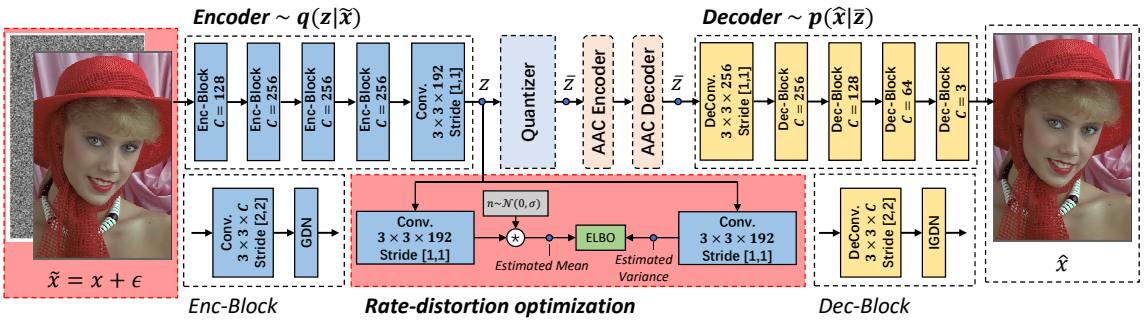


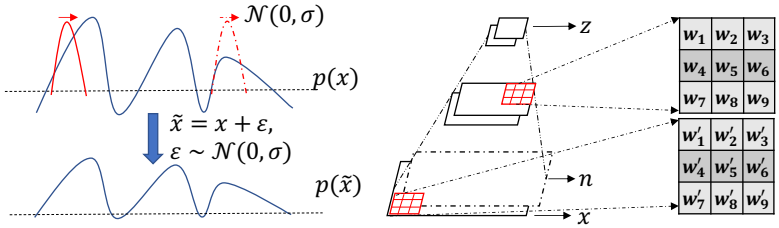
Figure 1: The proposed architecture for NC-VAE. The input $\tilde{x}$ mixes the original image x and Gaussian noise n with certain variance. The encoder transforms the input into the features z to obtain the quantized latent variables $\bar{z}$ with the quantzation layer u. For rate-distortion optimization, the distribution of $\bar{z}$ is controlled by two convolutional layer (for its mean and variance) and ELBO. Thus, the bitstream b is generated from $barz$ using adaptive arithmetic coding (AAC). At the decoder side, $\bar{z}$ is decoded into $\hat{x}$ for reconstruction that is symmetric to the encoder. Here, Enc-Block and Dec-Block consist of convolutional layer with 3×3 kernel, C channels and stride of 2, and GDN/IGDN layer. NC-VAE is trained offline for lossy image coding in an unsupervised manner.

Figure 2: Deformation and color drift led by the standard VAE, e.g. in the cloth region in the left picture and the eye of parrot in the right picture.

where $p(x)$ is the underlying distribution of x and $*$ indicates the convolution operation. Figure 3(a) illustrates the effect of noise, where a Gaussian pyramid is established according to the layer-by-layer convolutions with the noise distribution to smooth the input images . Thus, area adaptive Gaussian convolutional kernel (with weights $w'_1, \cdots, w'_9$) is developed to eliminate the deformation and color drift by considering the context of each pixel. In rate-distortion optimization, similar effect is achieved on $\tilde{x}$ for w and Gaussian noise n.

$$[w']^T = (1 + n * x^{-1})w^T \tag{3}$$

Eq. (3) implies that the energy of noise can be adjusted for various effects of smoothness. To determine the effect of the noise, we select different variances for the noise and discuss their influence on the reconstructed image, features and the statistics of symbols for AAC. Figure 4 visualizes the discrete features to evaluate the impact of noise on feature extraction. We can find that, with the growth of noise variance, the reconstructed image would be more smooth as using a white mask and the discrete features have less details. This fact suggests that the Gaussian function and kernel showed in Figure 3 can filter the images with a Gaussian pyramid. Furthermore, the noise does not affect the distribution of discrete feature, when we



(a) Gaussian function (b) Gaussian kernel

Figure 3: Two perspectives for the influence of the noise to convolution. (a) Perturbed distribution $p(\tilde{x})$ for input $\tilde{x}$ with noise ε. (b) Area adaptive Gaussian convolution with kernel $w'_1, \cdots, w'_9$ adapting to the noise n.

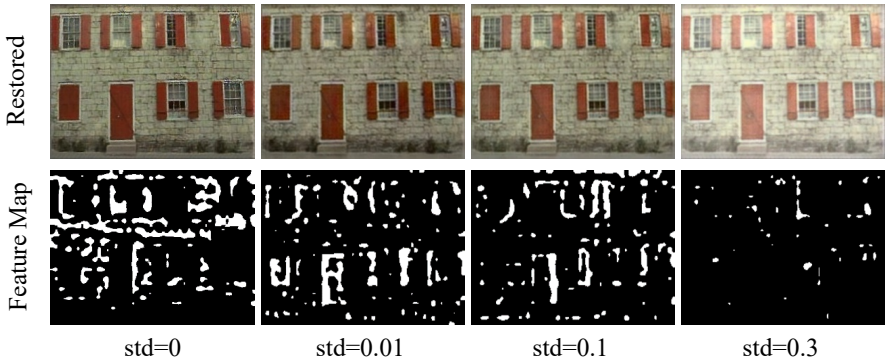


Figure 4: Effect of the energy of input noise on the latent variables at about 0.5 bpp. The first row is reconstructed images and the second row is the visualization of feature. The standard deviation (std) is set to 0, 0.01, 0.1 and 0.3 for the first to fourth column.

pose the prior knowledge on latent variables z. It is worth mentioning the approximate expectation of posterior distribution would not heavily deviate from the one of underlying distribution. Thus, the regions of deformation and color drift can be compensated using the global context with the smooth reconstruction for increasing noise variance.

3.2 Rate-Distortion Optimization

We further develop an efficient scheme for rate-distortion optimization (RDO) in the end-to-end optimized image compression. The distribution $p(z)$ of latent variables z in the bottleneck layer is adjusted with the proposed RDO module to minimize distortion under provided bit-rates. Here, we consider the variational bound [18].

$$-\log p\left(x\right) \leq \mathbb{E}_q\left[-\log p\left(x|z\right)\right] + KL\left[q\left(z|x\right) \| p\left(z\right)\right], \tag{4}$$

where $q(z|x)$ and $p(x|z)$ denote the encoder and decoder/generator in VAE. Thus, rate-distortion optimization can be formulated by minimizing the variational bound, as the first term $\mathbb{E}_q\left[-\log p\left(x|z\right)\right]$ can be interpreted as the reconstruction loss and the second term $KL\left[q\left(z|x\right) \| p\left(z\right)\right]$ the expected bit length in Eq. (4) [19, 20]. For $KL\left[q\left(z|x\right) \| p\left(z\right)\right]$, we further have

$$KL\left[q\left(z|x\right) \| p\left(z\right)\right] = \mathbb{E}_{q(z|x)}\left[-\log p\left(z\right)\right] - H\left(q\left(z|x\right)\right), \tag{5}$$

where $H(q(z|x))$ is the non-negative loss of encoder. Since introduction of prior distribution for z is equivalent to construct a hierarchical Gaussian distribution [6, 12], we propose a lighter model to reduce the complexity of decoder by eliminating its redundancy. In this paper, the proposed RDO module consists of only two linear convolutional layers to extract mean and variance of $p(z)$ and approximate its entropy. From $H(z) = -p(z)\log p(z)$, we obtain Eq. (6) as our training criterion by averaging over the encoder $q(z|x)$.

$$\begin{aligned}\min_{\theta,\theta'} KL(q\|p) &= \mathbb{E}_{q(z|\tilde{x})}\left[\log\frac{q\left(z|\tilde{x}\right)}{p\left(\hat{x}|\bar{z}\right)}\right] = \mathbb{E}_{q(z|\tilde{x})}\left[-\log p\left(\hat{x}|\bar{z}\right)+\log q\left(z|\tilde{x}\right)\right] \\ &= \mathbb{E}_{q(z|\tilde{x})}\left[-\log p\left(\hat{x}|\left(z*\mathcal{U}\left(-\frac{1}{2},\frac{1}{2}\right)\right)\right)+\log q\left(z|\left(x*\mathcal{N}\left(0,\sigma^2\right)\right)\right)\right] \\ &\approx \mathbb{E}_{q(z|\tilde{x})}\left[-\log p\left(\hat{x}|\left(z*\mathcal{U}\left(-\frac{1}{2},\frac{1}{2}\right)\right)\right)+\log q(z)\right] = D+R \qquad (6)\end{aligned}$$

Here, θ and θ' are the parameters of encoder and decoder and $\mathcal{U}(-1/2, 1/2)$ is the uniform noise for quantization. To achieve rate-distortion optimization, we introduce the hyper-parameter λ in Eq. (6). We can find the noise of input has been eliminated due to the introduction of prior distribution. This fact implies that the perturbation noise on input would not affect the output bit-rate. According to the entropy of discrete Gaussian source within a quantization width of t, the cost $b(z)$ of coding z is

$$b(z) = -\log t + \frac{\log\left(2\pi\sigma^2\right)}{2} + \frac{z^2}{2\sigma^2}. \qquad (7)$$

In Eq. (7), we set $t = 1$ for uniform quantization. Thus, the bit-rate is up to the value of z to a large extent. Note that we can adjust z with the hyper-parameter λ.

4. Experimental Results

4.1 Implementation Details

We trained the proposed codec using the CLIC training and validation dataset, which contain thousands of images with several million pixels. During the training stage, the images were cropped into 512×512 patches and fed into the codec at a batch size of 75. Adam optimizer was adopted to train parameters of the encoder and decoder, which has a learning rate initialized as 10^{-4} and decayed at a ratio of 0.9 each 2 epochs for the whole 64 epochs. As for the evaluation, Peak signal-to-noise ratio (PSNR) is utilized to measure the overall compression performance.

4.2 Rate-Distortion Performance

We evaluated the rate-distortion (R-D) performance for our method, state-of-the-art lossy image coders JPEG (jpeg-9c), JPEG 2000 (OpenJPEG-v2.3.0), Webp (libwebp-1.0.0), and BPG (bpg-0.9.8) and recent end-to-end learned compression methods Ballé *et al.* [8], and Minnen *et al.* [12] on Kodak24 dataset. Note that results for Minnen *et al.* [12] were directly copied from their paper. Figure 5(a) provides the R-D curves, where our model outperforms the benchmark methods in the region of 0.07-0.25 bpp

Table 1: Average PSNR (dB) for Kodak24 dataset at different bit-rates.

Methods	Bits Per Pixel (bpp)						BD-PSNR	BD-rate
	0.07	0.10	0.20	0.25	0.30	0.40	(dB)	(%)
JPEG	20.20	22.31	24.52	25.12	26.00	28.00	5.82	-85.97
WebP	23.90	26.00	27.30	28.04	28.90	30.00	2.98	-69.97
JPEG2000	24.95	26.01	28.00	28.68	29.58	30.81	2.52	-60.74
BPG444	25.06	27.20	29.48	30.25	31.03	32.10	1.21	-35.37
Ballé *et al.* [8]	25.49	26.81	28.40	29.78	29.96	30.96	1.88	-52.46
Minnen *et al.* [12]	-	27.00	29.98	30.75	31.35	32.02	0.71	-21.95
Ours	27.48	29.23	30.62	30.88	31.12	31.45	-	-

and yields a noticeable gain below 0.15 bpp. Table 1 presents the average PSNR for Kodak24 dataset under different bit-rates. Our method is shown to achieve an average 0.7-5.8 dB in BD-PSNR gain and 20%-85% BD-rate reduction in the low bit-rate region in comparison to the benchmark methods. These facts imply that our method leverages the generative model to reduce the number of layers for end-to-end learned image compression. NC-VAE does not outperform end-to-end learning framework in the high bit-rate region. According to the Mutual Information Neural Estimator (MINE) [21], the encoder cannot learn more useful representation from input due to the limited mutual information between input and bottleneck layers. It should be noted that the mutual information can be adjusted by λ in the form of bit-rates. However, we do not focus on optimizing λ here and will explore estimation of λ to improve the quality at high bit-rate in future.

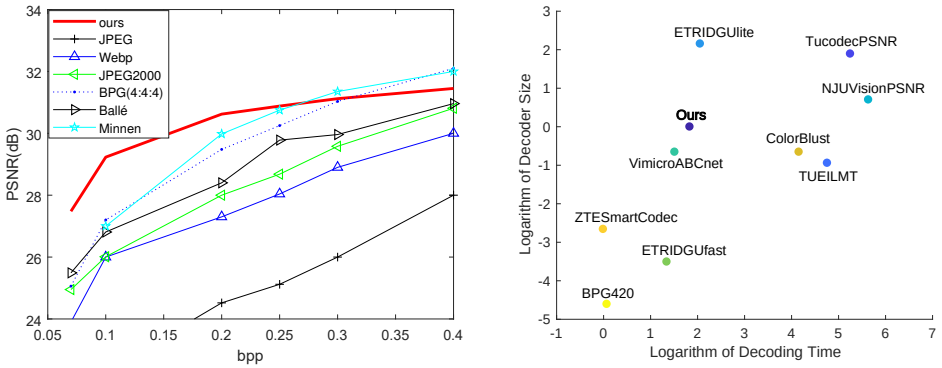
4.3 Visual Quality

Figure 6 illustrates the reconstructed *Kodak04* and *Kodak17* images obtained by our method, JPEG, JPEG 2000, WebP and Ballé *et al.* [8] at 0.07 bpp. Here, we adopt the lowest available bit-rates for JPEG and WebP, as they cannot reach the ultra-low 0.07 bpp. We do not include Minnen *et al.* [12] in Figure 6, as we have no access to their codes. Figure 6 shows that NC-VAE outperforms the benchmark methods, especially in the texture regions. JPEG 2000 cannot recover the high-frequency details under ultra low bit-rates. Regarding BPG and Ballé *et al.* [8], our method exceeds them in terms of PSNR and visual quality in the texture regions, i.e. the region of hairs in *Kodak04* image.

Table 2 compares our method and benchmarks in CLIC 2019 in terms of PSNR, decoding time and decoder size at 0.15 bpp. We present our time with one Intel(R) Xeon(R) CPU E5-2630 v4@2.20GHZ) in Table 2 and illustrate it in Figure 5(b). The original data about these models can be found on CLIC leaderboard.

4.4 Computational Complexity and Decoder Size

Compared with end-to-end learned TucodecPSNR, our method reduces the decoder time within extreme rate. We also evaluate NC-VAE with the remaining hybrid coding based methods. The hybrid compression methods in Table 2 is an improvement



(a) Rate-distortion Curves (b) Decoding Time and Decoder Size

Figure 5: Rate-distortion (R-D) performance and computational complexity and decoder size for our methods and state-of-the-art methods. (a) R-D curves at low bit-rates, i.e. 0.07-0.4 bpp. (b) Decoding time and decoder size for NC-VAE and CLIC 2019 benchmarks at 0.15 bpp. Logarithms of ratios of CLIC 2019 benchmarks against NC-VAE are measured.

Figure 6: Visual quality of reconstructed *Kodak04* and *Kodak17* images obtained by our method, JPEG, JPEG 2000, BPG, WebP and Ballé *et al.* [8], resp different methods, which contain both hand-crafted algorithms and end-to-end ones.

on current Versatile Video Coding model, which offers a powerful image compression method by detaching the intra mode. Specifically, ETRIDGUlite, NJUVisionPSNR, VimicroABCnet, ETRIDGUlite, and ColorBlust all adopt a learned post-processing network to improve the reconstruction quality over the primary reconstructed images. ZTESmartCodec adds a secondary prediction scheme to the original Versatile Video Coding (VVC) model to offer an alternative for block prediction. All these methods require only small learned models, as they are based sophisticated compression standard. On the contrary, our method improves the end-to-end learning framework with sharply reduced decoding time and decoder size. Thus, it provides a promising alternative for image coding in practical scenarios.

Table 2: The comparison among different codecs from CLIC at 0.15 bpp

Methods	PSNR (dB)	File size (bytes)	Decoding time		Decoder Size	
			(ms)	ratio	(bytes)	ratio
Ours	30.08	15742954	**1618347**	-	**37958451**	-
TucodecPSNR[13]	31.22	15748677	46174994	28.53×	252475146	6.65×
ETRIDGUlite	31.16	15748960	1891828	1.17×	327706926	8.63×
NJUVisionPSNR	31.10	15745551	67876555	41.94×	76950201	2.02×
VimicroABCnet	31.09	15748903	1100888	0.68×	19893480	0.52×
ETRIDGUfast	30.82	15748912	930953	0.58×	1117014	0.03×
ColorBlust	30.77	15739967	15464896	9.56×	19594357	0.52×
ZTESmartCodec	30.59	15749037	240315	0.15×	2563696	0.07×
BPG420	29.60	15745493	260315	0.16×	377858	0.01×

5. Conclusion

In this paper, we propose noise-to-compression VAE to improve the efficiency of end-to-end optimized image compression with an improved R-D performance in the low bit-rate regions. We first introduce the Gaussian noise to input for generative model to reduce deformation and improve fidelity and robustness for image compression. Furthermore, we optimize the parameters with an efficient rate-distortion optimization module for efficient image compression in practical scenarios.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 61971285, 61720106001, 61932022, 61425011, 61529101, 61622112 and 91838303, in part by the Program of Shanghai Academic Research Leader under Grant 17XD1401900, and in part by Huawei Cloud.

References

[1] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, February 1992.

[2] D. Taubman and M. Marcellin, *JPEG2000 Image Compression Fundamentals, Standards and Practice: Image Compression Fundamentals, Standards and Practice*, vol. 642, Springer Science & Business Media, 2012.

[3] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[4] J. Bankoski, P. Wilkins, and Y. Xu, "Technical overview of vp8, an open source video codec for the web," in *2011 IEEE International Conference on Multimedia and Expo*, July 2011, pp. 1–6.

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

[6] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada, April 2018.

[7] G. Toderici et al., "Variable rate image compression with recurrent neural networks," in *Proceedings of the 4th International Conference on Learning Representations*, San Juan, Puerto Rico, May 2016.

[8] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, April 2017.

[9] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," in *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, April 2017.

[10] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang, "Learning convolutional networks for content-weighted image compression," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018, pp. 3214–3223.

[11] O. Rippel and L. Bourdev, "Real-time adaptive image compression," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, NSW, Australia, August 2017, pp. 2922–2930.

[12] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in Neural Information Processing Systems 31*, Montreal, QC, Canada, December 2018, pp. 10771–10780.

[13] L. Zhou, Z. Sun, X. Wu, and J. Wu, "End-to-end optimized image compression with attention mechanism," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[14] A. van den Oord et al., "Conditional image generation with PixelCNN decoders," in *Advances in Neural Information Processing Systems 29*, Barcelona, Spain, December 2016, pp. 4790–4798.

[15] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proceedings of the 33rd International Conference on Machine Learning*, New York, NY, USA, June 2016, pp. 1747–1756.

[16] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, "Conditional probability models for deep image compression," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018, pp. 4394–4402.

[17] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, July 2008, pp. 1096–1103.

[18] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[19] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and Helmholtz free energy," in *Advances in Neural Information Processing Systems 7*, Denver, CO, USA, November 1993, pp. 3–10.

[20] A. Honkela and H. Valpola, "Variational learning and bits-back coding: An information-theoretic view to Bayesian learning," *IEEE Transactions on Neural Networks*, vol. 15, no. 4, pp. 800–810, April 2004.

[21] M. I. Belghazi et al., "MINE: Mutual information neural estimation," *arXiv preprint arXiv:1801.04062*, 2018.