Perceptual Friendly Variable Rate Image Compression

Yixin Gao[*], Yaojun Wu[*], Zongyu Guo, Zhizheng Zhang, Zhibo Chen[†]
CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System
University of Science and Technology of China
{gaoyixin, yaojunwu, guozy, zhizheng}@mail.ustc.edu.cn, chenzhibo@ustc.edu.cn

Abstract

In this paper, we study high fidelity variable rate compression framework. Both conventional and learned codecs in prior works are optimized for objective quality commonly measured by PSNR or SSIM, leaving perceptual quality optimization underexplored. Besides, to circumvent the need of training separate models under different rate conditions, we design a novel coding framework to support variable rate compression. Aside from the variable rate functionality, we propose an adaptive bit allocation unit to strengthen rate-distortion optimization across different rates. Extensive experimental results demonstrate that our proposed approach achieves better subjective quality than methods optimized by the objective metrics such as MSE, and MS-SSIM on CLIC 2021 validation dataset.

1. Introduction

Image compression is an important technology to reduce the resources of storage and transmission. In recent years, the rapid development of deep learning has spawned abundant end-to-end learning-based compression frameworks [1, 11, 7, 6, 2].Among them, VAE-based frameworks are of favorable rate-distortion performance thus more and more prevalent, which adapt joint optimization of rate and distortion as guidelines.

Employing some objective distortion metrics, most of the existing VAE-based works are optimized on objective indicators like mean square error (MSE) and MS-SSIM, which are not very consistent with human perception. Rippel *et al.* [12] first employ adversarial training to pursue realistic reconstructions. And Mentzer *et al.* [10] strive to a further step, by weighing ”rate-distortion-perception” and introducing perceptual loss to improve reconstruction quality effectively. Even though they provide a combined metric for pleasing visual reconstructions, there is no subjective explanation for the choice of metric in their method. Therefore, optimization for high perceptual quality is still underexplored. Besides, most VAE based methods train several discrete models for rate adaptation. This brings heavy cost for the deployment in the real industry scenarios. To address this problem, Choi et al. [3] provide two adjustable parameters: Lagrange multiplier and quantization bin size. The former corresponds to discrete rates with large intervals, while the latter changes rate in a small range. By inserting a pair of channel-wise units into the network, G-VAE [4] realizes the continuous rate changing and the increase of parameters is almost negligible. However, when performing rate adjustment, they lack explicit guidance for allocating bits according to different spatial contents, which is essential to the visual quality of the reconstruction.

To alleviate the above issues, we propose a variable rate compression framework with pleasing perceptual quality optimization in this paper. First, we choose suitable distortion metrics as the training supervisions to improve the perceptual quality. Considering distortion in the compression process lies on different granularities, we group MSSSIM, LPIPS and adversarial loss to instruct the optimization of the network. Second, we design an adaptive spatial bit allocation unit, which achieves quantization with different scales according to their contents. Then referring to G-VAE, we expanded the bit allocation unit as the role of ”gain unit” to realize variable rate. It cleverly changes the rate by adjusting the distortion of space content, which has obvious advantages in the perception quality. With the proposed methods, our team (IMCL_PQ) produces visually pleasing reconstructions during the validation phase, and detailed experimental results are shown in Section 3.

2. Proposed Methods

2.1. The overall framework

The framework of the proposed method can be seen in Figure 1. It can be divided into four parts: encoder, quantizer, entropy model, and decoder. The backbone network is an improved version based on [7]. Specifically, an input

[*] Yixin Gao and Yaojun Wu contribute equally to this work.
[†] Zhibo Chen is the corresponding author.

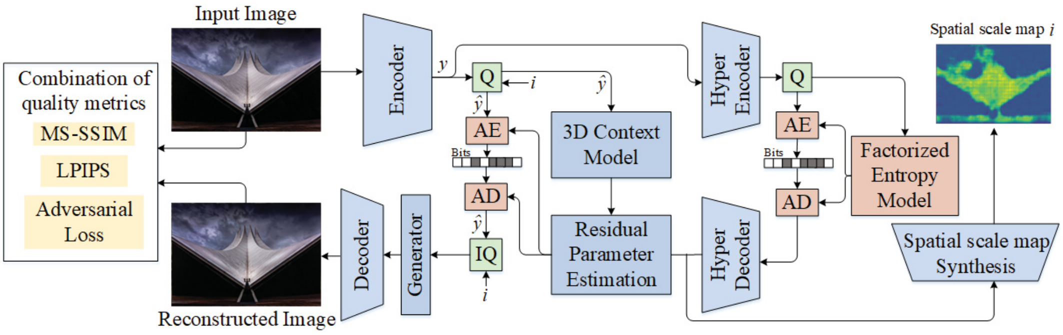

Figure 1. The overall framework of our image compression model.

image will first be transformed into latent representation y through the non-linear encoder. Then the latent feature y will be quantized with the aid of the spatial scale map i, the detailed operation of the quantization will be discussed in Section 2.3. After quantization, the quantized feature $\hat{y}$ will be coded to bit-stream by arithmetic coding. To formulate the distribution of $\hat{y}$ for entropy coding, Gaussian Mixture Model (GMM) is utilized in our model, where the parameters of the GMM are estimated through hyper-prior network (including hyper encoder and hyper decoder), 3D context model, and Residual Parameter Estimation (RPE) module. Compared with the former work [7], we improve the decoder by adding a generator [10] in front of it, which is utilized to enhance the quality of reconstruction images. In addition, we combine different quality metrics to jointly optimize the model for better visual quality, which will be discussed in Section 2.2.

2.2. Improvement of Perceptual Quality

The aim of the compression is to reduce the distortion D between reconstructed and original image as lower as possible under the constraints of the limited transmission bits R, which can be formulated by the Lagrange multiplier method:

$$L = R + \lambda \cdot D. \qquad (1)$$

Eq. 1 is usually employed as the loss function in learning based compression frameworks. Among various distortion metrics, MSE is commonly utilized, which can measure the fidelity from the perspective of pixel level. However, in the low rate setting, the guidance of MSE will result in blur artifacts, which severely influences the experience of human subjective quality. Although the model optimized by MS-SSIM can produce clear reconstructions compared with MSE, it still fails to reconstruct texture details. For example, it is hard to reconstruct the text information in most cases. To obtain decoded images with more texture details, neural network-based quality assessment metrics (such as LPIPS [13] and DISTS [5]) are feasible in compression network optimization. Nevertheless, it will also introduce artifacts such as checkboard effects in images. On the other hand, generative Adversarial Networks (GAN) can also bring more details on reconstruction by measuring the difference in data distribution, while some fake patterns would influence the visual quality. To utilize the advantages of different distortion metrics and instruct the compression network to reconstruct more pleasant results, we explore merging different distortion metrics and try to find the best combination. The detailed subjective results can be seen in Section 3.2. Specifically, we use MS-SSIM, LPIPS and adversarial loss as our final distortion metrics, which can be formulated as:

$$L = \lambda_R \cdot R + \lambda_{MS} \cdot (1 - D_{MS}) + \lambda_L \cdot D_L + \lambda_G \cdot D_G, \quad (2)$$

where D_{MS}, D_L, and D_G are respectively the distortion metrics calculated by MS-SSIM, LPIPS and discriminator. λ_{MS}, λ_L, and λ_G are the corresponding weights. The detailed structure of the discriminator is based on [10]. We empirically set λ_{MS} to 3×2^{-6}, λ_L to 5×10^{-3}, and λ_G to 7.5×10^{-4}. In training procedure, we only need to adjust the setting of the λ_R to obtain models on different rates.

2.3. Variable Rate with Bit Allocation

Since human are sensitive to different image contexts, we can adjust the spatial-wise quantization step to allocate bits and control local distortion. Furthermore, to achieve variable rate, the spatial quantization intervals under different rates can be used as a rate adaptation unit.

We design a spatial scale map synthesis module for adaptive bit allocation. Its structure is shown in Figure 2. Concretely, the model will generate spatial scale map i from the output of hyper decoder, which is calculated before the arithmetic coding of y. This spatial scale map has the same height and width of $\hat{y}$ but the channel number is 1. It represents the spatial quantization scale of y, which is formulated as

$$Q(y) = round(y * i). \qquad (3)$$

$Q(y)$ is the quantization operation and $*$ means spatial-wise multiplication. Since the $round()$ operation is not

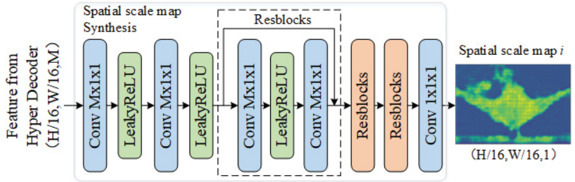

Figure 2. Detailed structure of the spatial scale map synthesis network.

differentiable, we simulate it with modified uniform noise, which is learned to adapt to image contexts. It can also be seen as a simplified noise scale to realize better spatial bit allocation [8]. To recover the information after the scaled quantization, we perform the corresponding inverse quantization at the input of the generator:

$$IQ(y) = \hat{y}/i, \tag{4}$$

where $IQ(y)$ represents the inverse quantization and $/$ means spatial-wise division. We use quantized hyper prior $\hat{z}$ to generate the spatial scale map i for two reasons. First, as the prior information that can effectively predict the probability distribution of quantized latent $\hat{y}$, $\hat{z}$ actually contains the spatial information of y. Second, there is no extra transmission overhead compared with discrete bit-rate model.

Based on the operation of gain units in G-VAE [4], we group spatial maps $\{i_0, i_1, \cdots, i_{K-1}\}$ at different rates for continuous variable rate, where K is a hyperparameter corresponding to the number of discrete rate models. So the quantization in variable rate involves rate selection

$$Q(y,k) = round(y * [(i_k)^l (i_{k+1})^{1-l}]), \tag{5}$$

where $k \in [0, 1, \cdots, K-1]$ is the index of rates and l represents an interpolation coefficient as in G-VAE. Then the inverse quantization is

$$IQ(y,k) = \hat{y}/[(i_k)^l (i_{k+1})^{1-l}]. \tag{6}$$

Eq. 2 is the loss function at one rate. Here, we optimize the network by accumulating the loss under K rates:

$$\begin{aligned} L_{VBR} = \sum_{k=0}^{K-1} [\lambda_{Rk} \cdot R_k + \lambda_{MS} \cdot (1 - D_{MSk}) \\ + \lambda_L \cdot D_{Lk} + \lambda_G \cdot G_k]. \end{aligned} \tag{7}$$

3. Experiments

3.1. Implementation Details

We follow the setting in [7] to build our network. We utilize CLIC training set as our training dataset. During training, the input image is randomly cropped to 256×256 patches with minibatches of 8. To meet the requirements at three different bit-rates (0.075 bpp, 0.15 bpp, and 0.3 bpp) in CLIC competition, we train three varaiable rate models, and each model can adapt to a certain range of rate changes. Specifically, for 0.075 bpp, we set $\lambda_{Rk} \in \{2^{-4.5}, 2^{-5.3}, 2^{-6}\}$ to realize the variable rate, it can finally achieve the rate range of 0.058 bpp to 0.085 bpp in CLIC validation dataset. In 0.15 bpp, the setting of the λ_R is $2^{-6}, 2^{-7}, 2^{-8}$, which realize the range between 0.126 bpp and 0.198 bpp. In 0.3 bpp, we set the λ_R to $\{2^{-7.5}, 2^{-9.33}, 2^{-11}\}$, and the range of the rate can be 0.242∼0.412 bpp. The training procedure of our method can be divided into three steps. First, we train network with high rate $\lambda_R = 2^{-11}$, and only optimize it for two distortion metrics including MS-SSIM and LPIPS. The aim of the first stage training is to obtain a high rate model as pre-train weights, which can accelerate the convergence of each rate points in the next stage. In the second stage, we start to train single rate-distortion points by utilizing the loss in Eq. 2, where the λ_R is respectively set to $2^{-6.25}$, 2^{-8}, $2^{-9.75}$. The final stage training will introduce the proposed spatial scale map synthesis network to realize variable rate for each rate point. In different training process, we set the initial learning rate lr as $5e-5$, and decay it after 300000 iterations. We utilize Adam as the optimizer and train our network on NVIDIA GTX 1080Ti GPU. It takes about 500000 iterations to finish the training of the first stage, while the training of second stage needs 300,000 iterations. And the training of the final stage only needs 100,000 iterations.

Table 1. Rate-Distortion Results of Different Metrics Combinations on Kodak at 0.30 bpp.

Metric	PSNR	MS-SSIM	LPIPS	DISTS
MSE	31.14	0.9611	0.2010	0.1467
MSE+ LPIPS+GAN	29.85	0.9523	0.0567	0.0871
MSE+ DISTS+GAN	30.44	0.9560	0.1564	0.1286
MS-SSIM	27.63	0.9749	0.2055	0.1543
MS-SSIM+ LPIPS	26.20	0.9498	0.0734	0.1312
MS-SSIM+ LPIPS+GAN	27.00	0.9675	0.0653	0.1068

3.2. Metrics Combinations for Perceptual Quality

To explore better visual quality for image reconstruction, we experiment different combinations on MSE, MS-SSIM, LPIPS, DISTS, and adversarial loss. We utilize Kodak as the testset. The reconstructed results are shown in Figure 3 and the rate-distortion results are in Table 1.

In our experiments, since the codec optimized with MS-SSIM metric has demonstrated better visual quality compared with the codec optimized with MSE, we continue to verify hybrid metrics based on MS-SSIM. We further

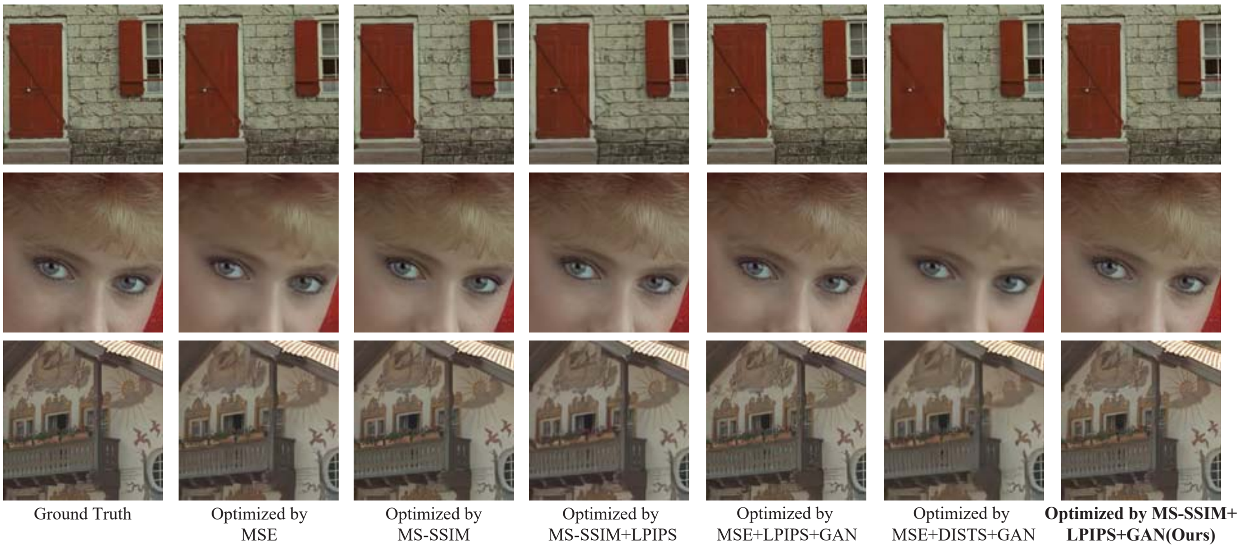
Ground Truth
Optimized by MSE
Optimized by MS-SSIM
Optimized by MS-SSIM+LPIPS
Optimized by MSE+LPIPS+GAN
Optimized by MSE+DISTS+GAN
Optimized by MS-SSIM+ LPIPS+GAN(Ours)

Figure 3. Visual Comparison between Different Metrics Combinations at 0.30 bpp.

Table 2. Comparison on adopting our variable rate (VBR) method.

Methods	Our method without VBR	**Our method with VBR**
BPP	0.07416	0.07402
PSNR	26.890	26.837
MS-SSIM	0.93353	0.93017
Average Preference	27.8%	**72.2%**

Table 3. Results on CLIC 2021 validation dataset.

Task	FID	PSNR	MS-SSIM
Image 075(valid)	177.240	26.311	0.94040
Image 150(valid)	160.317	28.918	0.97109
Image 300(valid)	147.937	30.823	0.98388

use LPIPS with MS-SSIM to enhance performance in details, but the experimental results show that it will instead generate worse visual quality due to additional texture artifacts (checkboard effects). Inspired by the success of GAN in [10], we combine MS-SSIM, LPIPS, and GAN, and compared it with the model optimzed by MS-SSIM. With the aid of the generator, the decoded images can obtain realistic texture details compared with the reconstructions from MS-SSIM optimized model. Besides, as reported in [9], compared with LPIPS, DISTS is proven to have better consistency with human subjective quality scoring. Therefore we try to replace LPIPS with DISTS and compare their performances. However, the model optimized by DISTS will generate more blurry images, which reduces the subjective quality. Finally, we compare the proposed hybrid metric(MS-SSIM, LPIPS and GAN) with the model that optimized by the loss utilized in [10], and better perceptual quality can be achieved from our experiments.

3.3. Variable Rate

Table 2 shows compression performance influences of VBR in terms of both subjective and objective quality, which is tested on the CLIC validation dataset. The performance drop of VBR is minor, which is difficult to compare quality differences. Therefore, we employ 5 experts to conduct subjective experiments and test the objective metrics (PSNR and MS-SSIM) to evaluate the performance of VBR. Obviously, our variable rate approach has overwhelming subjective advantages while nearly no drop on MSE and MS-SSIM. Based on the proposed framework, we obtain objective results on CLIC competition, and performance results on validation phase can be seen in Table 3.

4. Conclusion

In this paper, we introduce our work utilized in CLIC 2021 competition. We first focus on the improvement of perceptual quality. In detail, we analyze the influence of different metrics in this part and find a better metric combination for pleasing visual reconstruction. Then we introduce spatial scale map to realize variable rate in our work. It utilizes different scales in spatial contents for better rate adaption. As shown in experiments, our approach achieves best subjective quality compared with models optimized by other metrics, and realizes variable rate with even better visual quality.

Acknowledgment

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

References

[1] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations (ICLR)*, 2018. 1

[2] Zhibo Chen, Tianyu He, Xin Jin, and Feng Wu. Learning for video compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2):566–576, 2019. 1

[3] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Variable rate deep image compression with a conditional autoencoder. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3146–3154, 2019. 1

[4] Ze Cui, Jing Wang, Bo Bai, Tiansheng Guo, and Yihui Feng. G-vae: A continuously variable rate deep image compression framework, 2020. 1, 3

[5] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[6] Runsen Feng, Yaojun Wu, Zongyu Guo, Zhizheng Zhang, and Zhibo Chen. Learned video compression with feature-level residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 120–121, 2020. 1

[7] Zongyu Guo, Yaojun Wu, Runsen Feng, Zhizheng Zhang, and Zhibo Chen. 3-d context entropy model for improved practical image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 116–117, 2020. 1, 2, 3

[8] Zongyu Guo, Zhizheng Zhang, Runsen Feng, and Zhibo Chen. Soft then hard: Rethinking the quantization in neural image compression, 2021. 3

[9] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pages 633–651. Springer, 2020. 4

[10] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 2, 4

[11] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, pages 10794–10803, 2018. 1

[12] Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2922–2930, 2017. 1

[13] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2