Variable Rate ROI Image Compression Optimized for Visual Quality

Yi Ma[1,2†], Yongqi Zhai[1†], Chunhui Yang[1,2], Jiayu Yang[1,2], Ruofan Wang[1,2]
Jing Zhou[3], Kai Li[3], Ying Chen[3], Ronggang Wang[1,2*]
[1]Shenzhen Graduate School, Peking University, China
[2]Peng Cheng Laboratory, Shenzhen, China [3]Alibaba Group, China
{mayi,yangchunhui,jiayuyang,wangruofan20}@pku.edu.cn {rgwang}@pkusz.edu.cn
{zhaiyongqi}@stu.pku.edu.cn {zj271504,kaishi.lk,chenying.ailab}@alibaba-inc.com

Abstract

With the development of compression technology, objective metrics (e.g. PSNR, MS_SSIM) cannot satisfy our need, especially in extreme low bit-rate compression, thus more attention is being paid on perceptual quality. Since people have different standards for objective evaluation. For this reason, we simplify the topic with the consideration that people will strict more on interested region, so a ROI(region of interest) based image compression model is proposed with team name 'Sub201'. For the ROI, we expect its reconstructed part to be more accurate, while the background, server distortion is tolerable, and fake texture can be generated. Firstly, a weighted mask from saliency map is used. Secondly, to balance the difference of ROI and background area, different losses are applied separately. What's more, GAN and LPIPS are utilized to generate more texture in background. At last, variable rate method is adopted to realize rate control, and it performs well with perceptual metric. Experiment shows that our method can achieve better performance both in visual and objective quality.

1. Introduction

Image compression as a mature technology has been developed for decades, which aims to balance the tradeoff between rate and distortion: entropy of discretized representation and error arising while constructing[17]. For traditional codec, such as JPEG and JPEG2000, rate and distortion are optimized by hand separately. Recently, neural compression made rate and distortion optimization in an end-to-end manner[4, 1, 5, 14, 12, 8].

In extrame low rate compression, objective metrics perform bad visually, perceptual quality enhancement attracts more attention, methods[15, 2, 18, 11, 3, 13] had been proposed with GAN to generate perceptual texture. Perceptual quality as a kind of high-level metric also adopted, such as VGG[16], LPIPS[21]. Thus compression transfers to the optimization of rate-distortion-perception. Since people have different standards for subjective criterion, for some content sensitive images, such as faces, documents, keeping its authenticity is more important than generating vivid but fake texture.

From this point, we propose a ROI based image compression method. Based on the framework of [13], we introduce ROI mask from salience map to guide the network firstly. Secondly, to further utilize the prior information, for areas of background and ROI, different loss functions are used separately to obtain optimal visual quality technically. Basically, more bits are allocated to the area of ROI. At last, to satisfy target bits, our model is trained with a variable rate compression method inspired from [9], and it performs better than non-variable rate model.

2. Method

Figure 1 provides a high-level overview of our proposed method. In the following chapters, we will separately introduce the network structure, ROI compression, variable-rate implementation.

2.1. Network architecture

Our network is based on a main auto-encoder with hyperprior network. The main encoder architecture is shown in Figure 2, which contains residual and attention mechanism. In order to capture both channel-wise and spatial-wise relationships, we utilize a channel-spatial attention block in our main autoencoder, as shown in Figure 3. Different from previous work [19, 20], we introduce residual blocks both in trunk and attention branch to extract more powerful features. Batch normalization layers are removed and ReLU is used in residual blocks.

*Corresponding author
†These authors contribute equally.
‡Thanks to National Natural Science Foundation of China 61672063, Shenzhen Research Projects of JCYJ 20180608092141929O, 20180503182128089 and RCJC 20200714114435057.

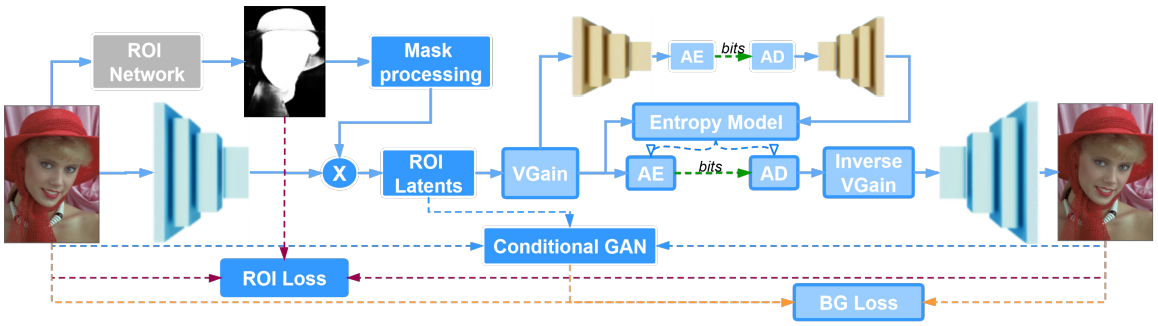

Figure 1. Overall architecture of the proposed image compression framework. The blue stacked layer represents the image compression network, and the yellow stacked layer represents the hyperprior network. The ROI Network is not trainable. VGain and Inverse VGain is used to implement variable rate. AE/AD are short for arithmetical encoder/decoder. MASK processing will be described in Section 2.2.3.

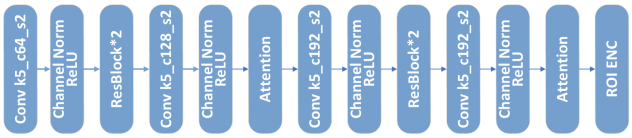

Figure 2. Network architecture of our main encoder.

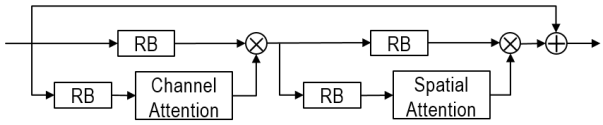

Figure 3. The structure of our channel-spatial attention module."RB" means residual block.

2.2. ROI Compression

In our model, to design corresponding optimization methods for different image contents, the image is divided into two types of regions. The first type of area includes human faces, text, etc. People require such textures to be accurately reconstructed. For the second, more attention will be paid on the perceptual quality even it deviates its original. Thus, a ROI guided optimization method is introduce.

2.2.1 ROI Mask

When considering segmentation, instead of labeled semantic segmentation, visual saliency detection can distinguish the image into the focused area and background, which is more suitable to our strategy. Different from [6], saliency regions are generated offline through a saliency detection network[7], which is fixed as a strong supervision while training.

$$Mask_{2D} = \sigma(Detection(x)) \tag{1}$$

where $Detection$ denotes the saliency detection network and σ refers to sigmoid function.

For the saliency map, there are sharp boundaries between different regions, so transition method should be used. Figure 5 shows that the decoder generates noise at such boundaries with gan loss. Therefore, we adapt a convolution layer (the filter size is 51, and weights are all set to 1) to generate a 2D ROI mask RM_{2D} to smooth the saliency map.

$$RM_{2D} = Smooth_{conv}(Mask_{2D}) \tag{2}$$

2.2.2 Distortion Loss

Under the guidance of RM_{2D}, we use differentiated loss functions to optimize the ROI and the background area, d_{ROI} and d_{BG}.

$$d_{ROI} = RM_{2D} \otimes MSE(x, \hat{x}) \tag{3}$$

$$d_{BG} = 1 - MS\_SSIM(x, \hat{x}) + \lambda_p \times d_P \tag{4}$$

x and $\hat{x}$ denote the input and reconstructed image. And $\otimes$ refers to element-wise multiplication. d_{ROI} uses MSE as a measurement, and it only takes effect in the ROI. While, d_{BG} includes MS_SSIM and a perceptual loss LPIPS as d_P, which proves to be closer to human visual evaluation standards. The default λ_p is 0.5.

2.2.3 ROI Latents

From the perspective of visual quality optimization, more bits are allocated to the ROI to enhance the accuracy of the reconstructed features. When the image is mapped into latent representations by the encoder, the spatial characteristics are still preserved even down-scaled by 16x. So for the latents, we can generate ROI mask RM_{Latent} applied on it by averaging pooling (stride is set as 16):

$$RM_{Latent} = AvgPool(RM_{2D}) \tag{5}$$

With the weighted RM_{Latent}, latents in ROI are magnified, thus the area of ROI will occupy more bits in the

Figure 4. Visual quality comparison of reconstructed images. Comparison of the visual quality of the reconstructed image. BASE represents the compression model without deploying the ROI module. The w/o ch and w/o GAN isolate the channel protection strategy and the GAN model respectively. ROI1.5 and ROI1.0 represent the complete ROI model, and the alpha is set to 1.5 and 1.0 respectively.

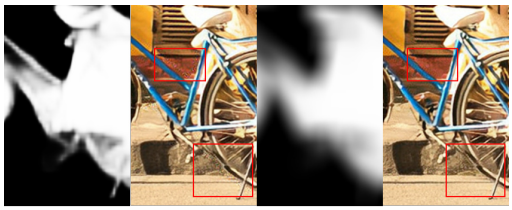

Figure 5. Comparisons with different masks.

generated code stream. In addition, we use α to control the weight of the ROI in terms of rate allocation.

$$Latent_{ROI} = \frac{RM_{Latent} + \alpha}{\alpha} \otimes Latent \tag{6}$$

Here, a smaller α means more bits are allocated to the ROI area in latents. What's more, we protect a certain number of channels to retain appropriate information for the background to avoid the fading of its reconstructed texture.

$$Latent_{ROI} = Latent_{ch0-ch47} || Latent_{ROIch48-ch191} \tag{7}$$

Assuming there are 192 channels in latents, the first 48 feature maps are protected, and the following channels are weighted with [6] for corresponding channels.

2.3. Variable Rate

To realize rate control, we adopt a variable-rate strategy as in [9]. In the encoder, a scaled matrix $M \in R^{c*n}$ is introduced to scale the encoded latent representation $y \in R^{c*h*w}$ channel by channel, where c, h, w, n represent the number of the channels, the height, width of latents, and the number of scaled vectors respectively. The scaled vector can be denoted as $v_s = \{\alpha_{s(0)}, \alpha_{s(1)}, ..., \alpha_{s(c-1)}\}, \alpha_{s(i)} \in R$, where s represents the index of the scaled vectors in the scaled matrix. The scaled matrix is trained to obtain different bit rates by scaling the channels of the latent representation as Eq.8. Here y represents $Latent_{ROI}$.

$$\bar{y}_s = G(y, s) = y \odot v_s, \tag{8}$$

where $G(\cdot)$ represents the scale process, $\odot$ represents channel-wise multiplication, $\bar{y}_s$ is the scaled latent representation.

In the decoder side, another scaled matrix $M' \in R^{c*n}$ is applied to rescale the quantized scaled latent representation $\hat{y}_s$. The inverse-scale vector is denoted as $v'_s = \{\beta_{s(0)}, \beta_{s(1)}, ..., \beta_{s(c-1)}\}, \beta_{s(i)} \in R$. The inverse-scale process works as Eq.9.

$$y'_s = IG(\hat{y}_s, s) = \hat{y}_s \odot v'_s, \tag{9}$$

Each pair of the scaled vector v_s, v'_s are corresponding to a specific Lagrange multiplier which are included in the loss function for training to acquire models with variable rate.

For purpose of accurate rate control, a continuous variable rate model is need in inference.

$$v_s \cdot v'_s = C, \tag{10}$$

where v_s, v'_s ($s \in [0, 1, ..., n-1]$) represent existing scaled vector pairs, and $C \in R^c$ is a constant vector. More vectors can be interpolated linearly through these scaled vector pairs as [9].

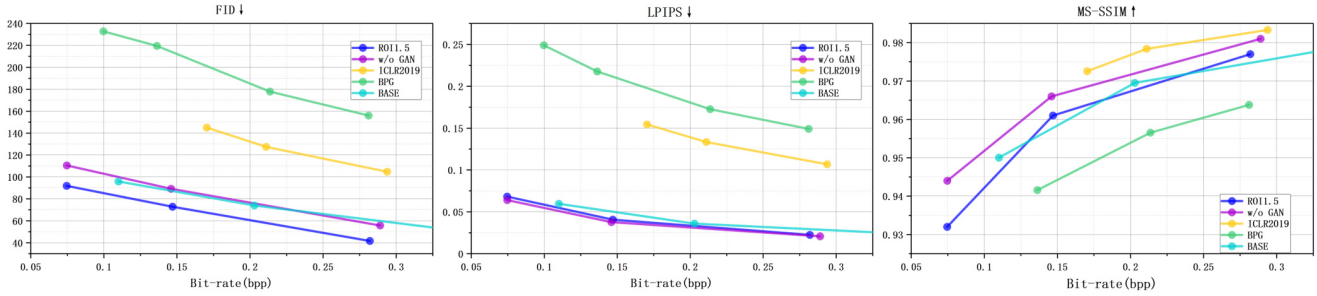

Figure 6. Comparison of rate-distortion performance of Our model with BPG and ICLR2019 [10]. ↑ and ↓ respectively represent larger and smaller values are better.

2.4. Quantization and Entropy Model

In our model, an additive i.i.d uniform noise is used to approximate quantization on latent representations to make the framework end-to-end trainable.

Following the work of Cheng *et al.* [8], we introduce Gaussian mixture model to parameterize flexible conditional distributions of $Latent_{ROI}$ representations combine with an auto-regressive context prior and hyperprior. For the latents of hyperprior $\hat{z}$, it's modeled by a non-parametric, fully factorized density model. Finally, the total bit rate cost r is defined as Eq.11.

$$r = r_{\hat{Latent}_{ROI}} + r_{\hat{z}} \tag{11}$$

2.5. Adversarial Training

With a ROI loss that protects key information of contents and reduce substantial redundancy in backgrounds, we further introduce a conditional GAN in the rate-distortion trade-off to maintain high perceptual fidelity of reconstructed images at low bit-rate, as that in [13], where the information used in conditional GAN is ROI latents, as is defined in Eq.[3,4].

3. Experiments

3.1. Training

Models are trained in two stages. Firstly, it's trained without GAN to initialize parameters stably, then the model with GAN are trained to improve subjective quality. The size of the images is cropped to 256×256, and we use Adam optimization with the initial learning rate of $1e^{-4}$. Meanwhile, batch size is set to 8, and it takes $1e^6$ iterations for the model without GAN and with GAN respectively.

While training for variable rate, three models of 0.075bpp, 0.15bpp and 0.3bpp are optimized. For each variable-rate model, we set six sets of scaled vectors and Lagrange multipliers $[v_s, v'_s, \lambda_s]$ in training. For 0.075bpp, λ is selected from [120, 220, 320, 420, 520, 720], and [30, 90, 140, 190, 240, 290] and [10, 20, 30, 50, 70, 90] for 0.15bpp and 0.3bpp separately.

3.2. Subjective Quality Evaluation

Figure 4 shows the details of the reconstruction of our model. Compared with HIFIC [13], a learning image compression method with the most advanced visual quality, our model can reconstruct more accurate textures resulting in higher visual quality. From the hat area on the left, we can see that the colors of our reconstructed image are more accurate, while in the face area, there are less noise with more vivid and accurate textures in our image. As for the text details, on the right side of the figure, due to the existence of GAN, the fake texture generated by HIFIC has seriously affected the quality and the words are out of shape. In our ROI model, this problem is solved regardless of whether GAN is included or not.

3.3. Objective Quality Evaluation

Figure 6 demonstrates that the rate-distortion curve of our model and other advanced compression models in CLIC2021 validation set. It can be seen that, compared with ICLR2019 [10] and BPG, our ROI compression model has a great advantage in perceptual metrics (LPIPS, FID), while its performance on MS_SSIM is mediocre. In the curve of MS_SSIM, $ROI\ 1.5$ and $w/o\ GAN$ perform better than the $BASE$, which indicates that the objective quality did not decrease with the deployment of the ROI. We assume such result to the protection of channel as explained in 7.

4. Conclusion

In this paper, ROI based image compression method is proposed to improve visual quality. To fully extract the information of ROI, we utilize it not only in loss but also latents, and method to obtain ROI based latents is proposed. A better balance of rate and distortion between ROI and background are discovered. At last, we also verify the effectiveness of variable rate method, that is one model can get different rates with different subjective quality in one model. Experiments results prove that our method can surpass the state-of-the-art method both in subjective and some high-level objective metrics, such as LPIPS, FID.

References

[1] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc Van Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. *arXiv preprint arXiv:1704.00648*, 2017. 1

[2] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 221–231, 2019. 1

[3] Hiroaki Akutsu, Akifumi Suzuki, Zhisheng Zhong, and Kiyoharu Aizawa. Ultra low bitrate learned image compression by selective detail decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 118–119, 2020. 1

[4] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016. 1

[5] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018. 1

[6] Chunlei Cai, Li Chen, Xiaoyun Zhang, and Zhiyong Gao. End-to-end optimized roi image compression. *IEEE Transactions on Image Processing*, 29:3442–3457, 2019. 2

[7] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. Global context-aware progressive aggregation network for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10599–10606, 2020. 2

[8] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 4

[9] Z. Cui, J. Wang, B. Bai, T. Guo, and Y. Feng. G-vae: A continuously variable rate deep image compression framework. 2020. 1, 3

[10] Jooyoung Lee, Seunghyun Cho, and Seungkwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. 2019. 4

[11] Jooyoung Lee, Donghyun Kim, Younhee Kim, Hyoungjin Kwon, Jongho Kim, and Taejin Lee. A training method for image compression networks to improve perceptual quality of reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1

[12] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Practical full resolution learned lossless image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10629–10638, 2019. 1

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

[14] David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. *arXiv preprint arXiv:1809.02736*, 2018. 1

[15] Shibani Santurkar, David Budden, and Nir Shavit. Generative compression. In *2018 Picture Coding Symposium (PCS)*, pages 258–262. IEEE, 2018. 1

[16] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[17] N. J. A. Sloane and A. D. Wyner. *Coding Theorems for a Discrete Source With a Fidelity CriterionInstitute of Radio Engineers, International Convention Record, vol. 7, 1959.*, pages 325–350. 1993. 1

[18] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[19] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. *Springer, Cham*, 2018. 1

[20] J. Yang, C. Yang, Y. Ma, S. Liu, and R. Wang. Learned low bit-rate image compression with adversarial mechanism. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. 1

[21] 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. 1