ROI Image codec Optimized for Visual Quality

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. 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. For the ROI, we expect its reconstructed part to be more accurate, while the background, 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. 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.

1.1. 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 [8, 9], 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.

1.2. 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.

1.2.1 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 [1], saliency regions are generated offline through a saliency detection network[2], 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.

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

1.2.2 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}.

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

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

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. We also use L1 as d_{ROI}. Although L1 makes the texture slightly clearer than MSE, the RD performance is reduced, and subjectively it may not be better than using L2. While, d_{BG} includes SSIM and a perceptual loss LPIPS as d_P, which proves to be closer to human visual evaluation standards.

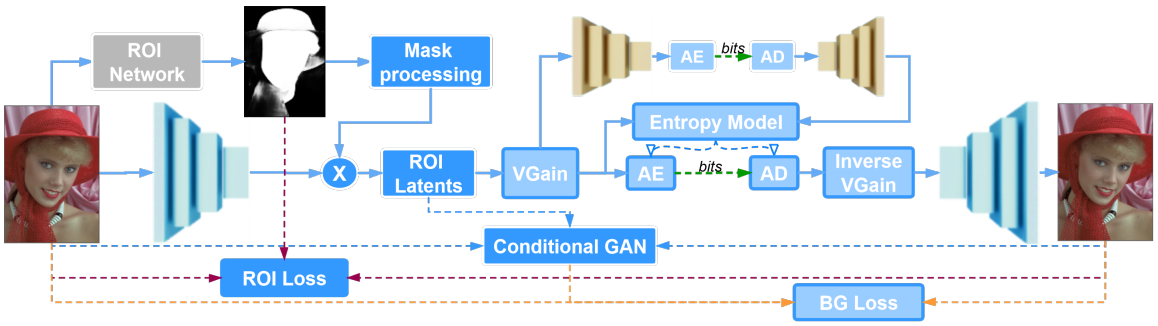

Figure 1. Overall architecture of the proposed image compression framework. The blue stacked layer represents the image compression network, and the yellow stacked layer represents the hyperprior network. The ROI Network is not trainable. VGain and Inverse VGain is used to implement variable rate. AE/AD are short for arithmetical encoder/decoder. MASK processing will be described in Section 1.2.3.

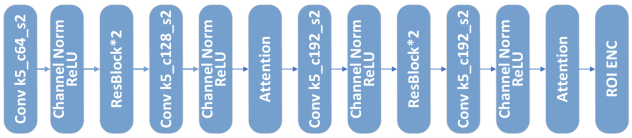

Figure 2. Network architecture of our main encoder.

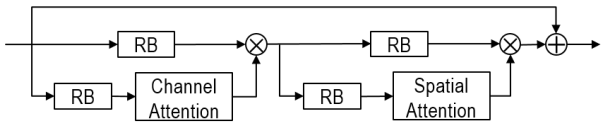

Figure 3. The structure of our channel-spatial attention module."RB" means residual block.

We observed that using MSSSIM alone sometimes resulted in a color temperature shift in the reconstructed map. The default λ_p is 0.5.

1.2.3 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 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.

1.3. Variable Rate

To realize rate control, we adopt a variable-rate strategy as in [4]. 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.

Figure 4. Visual quality comparison of reconstructed images.

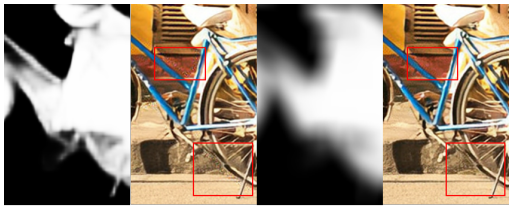

Figure 5. Comparisons with different masks.

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 [4].

1.4. 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.* [3], 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.

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

1.5. 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 [7], where the information used in conditional GAN is ROI latents, as is defined in Eq.[3,4]. In addition to the conditional GAN, we also tried to use LSGAN [6], which did not bring subjective performance gains.

2. Experiments

2.1. 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.

2.2. Objective Quality Evaluation

Figure 6 demonstrates that the rate-distortion curve of our model and other advanced compression models in CLIC2022 validation set. It can be seen that, compared with ICLR2019 [5] 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

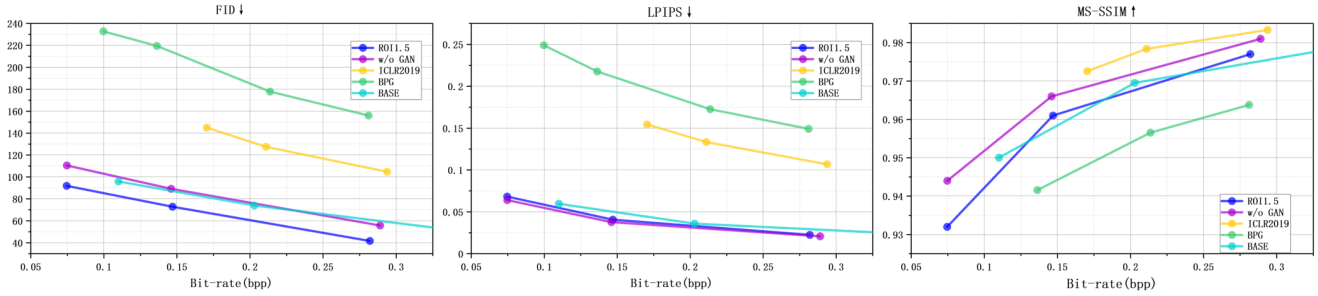

Figure 6. Comparison of rate-distortion performance of Our model with BPG and ICLR2019 [5]. ↑ and ↓ respectively represent larger and smaller values are better.

Table 1. Quantitative results on different target bpp of CLIC2022 validation image.

Bpp	PSNR↑	MS-SSIM↑	LPIPS↓	ROI-PSNR↑
0.106	30.06	0.952	0.055	29.29
0.195	31.76	0.971	0.032	31.19
0.293	32.89	0.979	0.021	33.16

result to the protection of channel as explained in 7. Table 1 shows our evaluation on the validation set.

3. 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] 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. 1

[2] 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. 1

[3] 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. 3

[4] Z. Cui, J. Wang, B. Bai, T. Guo, and Y. Feng. G-vae: A continuously variable rate deep image compression framework. 2020. 2, 3

[5] Jooyoung Lee, Seunghyun Cho, and Seungkwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. 2019. 3, 4

[6] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 3

[7] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33, 2020. 3

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

[9] 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