Low-rate Image Compression with Super-resolution Learning

Wei Gao[1,2,*], Lvfang Tao[1,2], Linjie Zhou[1,2], Dinghao Yang[2], Xiaoyu Zhang[1,2] , Zixuan Guo[1,2]

[1]School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University,
Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
{gaowei262, ltao, ljzhou, zxy2019, gzx2019}@pku.edu.cn dinghowyang@gmail.com

Abstract

In this paper, we propose an end-to-end learned image compression framework for low-rate scenarios. Based on variational autoencoder, our method features a pair of compact-resolution and super-resolution networks, a set of hyper and main codec networks, and a conditional context model. The learning process of this framework is facilitated with integrated non-local attention modules and phase congruency priors. Multiple models are obtained from training with different hyper-parameters, and are jointly used in the image-level model selection process for rate control, which ensures that the bit rate constraint of the CLIC challenge is satisfied. Experimental results demonstrate that the proposed method can achieve an averaged multi-scale structural similarity (MS-SSIM) score of 0.9648 with bit rate consumption of 0.1499 bits per pixel, which outperforms the BPG image coding method significantly.

1. Introduction

As the popularization of image and video applications, the volume of visual data becomes increasingly huge. Therefore, lossy image compression, especially with low bit rate, becomes a challenging task. By consuming low bit rate, image compression algorithm should provide much smaller perceived distortions.

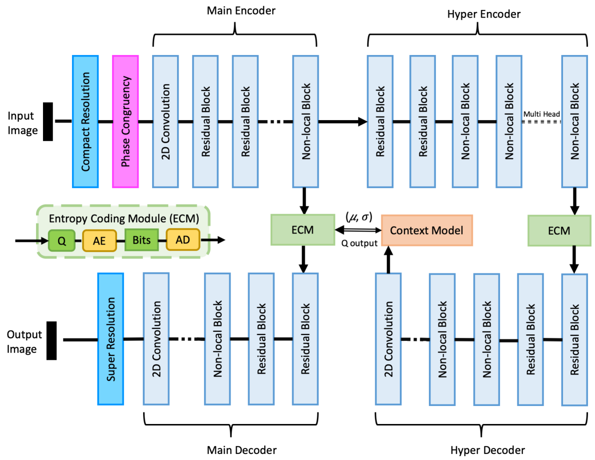

Figure 1. The overall structure of the proposed image compression method, which is mainly consist of variational autoencoder, compact resolution, super resolution, non-local and phase congruency modules. The symbols of Q, AE, AD represent quantizer, arithmetic encoder and arithmetic decoder, respectively.

Recently, with the development of neural networks, deep learning based image compression techniques have been proposed and achieve superior rate-distortion performance than traditional image codecs like JPEG [12], JPEG2000 [9], and HEVC-intra [10]. Autoencoders [1, 11] are widely used in end-to-end image compression, which include two major components, i.e., encoders and decoders. Encoders extract features from raw image to reduce the data redundancy, thereby expressing the image as a more compact feature representation. Decoders can utilize the feature expressions for image reconstruction in an inverse process. Ballé et al. [1] propose variational autoencoder (VAE) framework, where a hyper encoder is studied for better entropy modeling. Li et al. [7] add compact-resolution (CR) and super-resolution (SR) modules on traditional coding meth-

[*]Corresponding author. This work was supported by Natural Science Foundation of China under Grant 61801303, Guangdong Basic and Applied Basic Research Foundation under Grant 2019A1515012031, Shenzhen Science and Technology Plan Basic Research Project under Grant JCYJ20190808161805519, the Open Projects Program of National Laboratory of Pattern Recognition (NLPR) under Grant 202000045, the Open Project Program of the State Key Lab of CAD&CG (Grant No. A2009), Zhejiang University, the start-up fund of Shenzhen Graduate School of Peking University under Grant 2390101081, and CCF-Tencent Open Fund under Grant IAGR20190101.

ods as an multi-branch framework, including block-level adaptive scheme and frame-level scheme, to achieve bits saving. Jiang et al. [4] develop an end-to-end learning-based compression algorithm with compact Convolutional Neural Network (CNN) and reconstruction CNN, which shows superior results over traditional codecs. Since the perceptual quality for low-rate scenarios becomes much more important, the learning-based image compression has not been fully investigated.

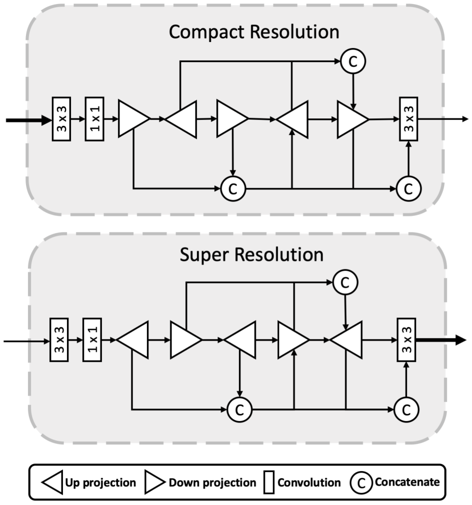

Figure 2. The structure of the compact-resolution and super-resolution modules.

In this work, we propose to improve the VAE architecture in [1] by introducing a paired CR and SR network. CR network is capable of acquiring the dense representation of images, while SR network can be jointly trained to restore the information loss during the down-scaling process in CR. Because CR can greatly reduce the amount of pixels, the use of CR and SR is quite suitable for low-rate image compression. In addition, by using phase congruency [6] for texture evaluation, the structures of images can be better exploited for feature representation and learning in the network. The adopted non-local [13] module can capture long-range dependencies by attention mechanism to obtain global information. Finally, we also use an iterative degradation-based rate control algorithm to balance rate and distortion trade-off on a per-image basis.

2. Proposed Method

2.1. Paired Compact-Resolution (CR) and Super-Resolution (SR) Networks

For further effective bit rate reduction, the compact-resolution and super-resolution are employed at the beginning and end of the overall process, respectively. We would like to construct a CNN to get a better representation of an image which preserves more informative content after the down-sampling process. The output down-sampled image is then encoded and decoded through an image compression network. Then, super-resolution network can generate the full resolution image.

The structure of CR and SR network is depicted as Figure 2. We implement the SR network based on deep back-projection network (DBPN) [3]. For light-weight model parameters and memory usage, we have reduced the number of network layers to 3 up and down sampling units rather than 7 in the original DBPN. Additionally, the number of feature maps is also greatly cut down. To boost the SR performance to restore the degraded information by CR, inspired by VAE [1], we intuitively construct a CR network which is symmetrical to the SR network.

The CR network includes initial feature extraction, up and down sampling projection, and final reconstruction. The first stage contains two convolutional layers to obtain the primary features. A convolutional layer with 3×3 filter size firstly generates $H_h \times W_h \times C_1$ feature map from $H_h \times W_h \times C$ image, and then 1×1 filter is used to reduce the feature dimension to $H_h \times W_h \times C_2$ $(C_2 < C1)$. The second stage includes 3 down-sampling units and 2 up-sampling units. Similar to [3], each unit is constructed as back-projection form based on residual learning. All previous outputs are concatenated as input to the next unit. The down-sampling unit outputs $H_l \times W_l \times C_2$ feature from $H_h \times W_h \times (C_2 \times n)$ feature, and the up-sampling unit outputs $H_h \times W_h \times C_2$ feature from $H_l \times W_l \times (C_2 \times n)$ feature, where n represents the number of concatenated features. A convolutional layer with 3×3 filter size is used for reconstruction from $H_l \times W_l \times C_2$ to $H_l \times W_l \times C$ image. The SR network finally super-resolves the decoded image from $H_l \times W_l \times C$ to $H_h \times W_h \times C$.

The CR and SR networks are pre-trained before the training of the overall compression network. SR network is trained by minimizing loss function L_{sr},

$$L_{sr} = \|f_{sr}(g(x)) - x\|_2^2 \tag{1}$$

where x is the input image, f_{sr} represents the SR network, g is the bicubic interpolation.

Since the ground truth for CR is difficult to achieve, we use the SR network to assist the training for CR network. The training method in [7] is adopted, where CR network is followed by the trained SR network, and the weights of SR network are fixed. The loss function L_{cr} is defined as

$$L_{cr} = \|f_{sr}(f_{cr}(x)) - x\|_2^2 + \lambda\|f_{cr}(x) - g(x)\|_2^2 \tag{2}$$

where f_{cr} represents the CR network, and λ is the parameter balancing between visual quality and the amount of contained information.

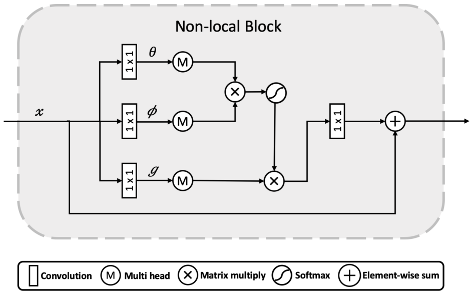

Figure 3. The structure of the adopted non-local mechanism.

2.2. Non-Local Module

Moreover, we would like to take advantage of the non-local mechanism as an enhanced attention method to better perceive image features adaptively. In Figure 3, the structure of the proposed non-local module is illustrated, the input feature map x is processed into three flow $\theta(x)$, $\phi(x)$, $g(x)$, where θ, ϕ, g are implemented by 1×1 convolution. Additionally, we apply multi-head mechanism to learn different representation from different subspaces jointly, which are created by channel splitting. The matrix multiplication outputs of different subspaces are concatenated to aggregate information, then activated by softmax to obtain an attention mask. Furthermore, we add the global attention output with input x to get more abundant feature.

2.3. Phase Congruency

In this paper, phase congruency (PC) [6] is employed to represent sharp transitions of image, which can evaluate the textures effectively. The PC of 2D image p with scale s and orientation r can be calculated as

$$PC = \frac{\sum_r \sum_s M_r(p) \lfloor L_{sr}(p)\Delta\Theta_{sr}(p) - N_r \rfloor}{\sum_r \sum_s L_{sr}(p) + \xi} \tag{3}$$

where $M_r(p)$ is a metric for frequency spread, and $L_{sr}(p)$ and $\Delta\Theta_{sr}(p)$ are amplitude and phase deviation of p, respectively. N_r is a quantity used to reduce noise effect, while the symbol of $\lfloor\rfloor$ means that the enclosed quantity equals itself if the value is positive, otherwise equals zero. ξ is used for avoiding zero-division.

The PC image is down-sampled to the same size of main encoder layers via convolutions. Then multi-scale PC features are concatenated to the corresponding main encoder layers, which provides edge texture information.

2.4. Models and Learning Details

For context and entropy modeling, the quantized outputs of main encoder and hyper encoder are denoted as $\hat{u}$ and $\hat{v}$, respectively. The context model with 3D masked convolution network [8] can predict the mean and standard deviation of $\hat{u}$ with lower computational cost. Similar to [1] and [8], $R_{\hat{u}}$ and $R_{\hat{v}}$ are obtained as rate estimations of $\hat{u}$ and $\hat{v}$, respectively. Then, similar to [15], the loss function is designed as Eq. (4) so that the joint training of our learned model can be considered as a process of rate-distortion optimization. D_w is a weighted mixed distortion criterion, which is devised as Eq. (5) by combining mean squared error (MSE) and multi-scale structural similarity (MS-SSIM) [14] score,

$$L = D_w + R_{\hat{u}} + R_{\hat{v}} \tag{4}$$

$$D_w = \lambda_1 \times \|x - \hat{x}\|_2^2 + \lambda_2 \times (1 - \text{MS-SSIM}) \tag{5}$$

where λ_1 and λ_2 denote weighting coefficients, x and $\hat{x}$ denote original and compressed images, respectively. By varying λ_1, λ_2, models with different average bit rates can be obtained through multiple training.

In this work, three models will be trained to have different compression ratios. Therefore, we need to implement a rate control algorithm, which is responsible for choosing the most proper model for each individual image to be compressed. The algorithm should also ensure that the average bit rate of all compressed images is below but enough close to the target bit rate [2, 8, 15].

Initially, we assume all the images are compressed with the highest available quality. Then, similar to [8], we employ greedy algorithm to select images to be degraded based on the slope of MS-SSIM [14] and generated bitstream size. The degradation process is iteratively executed, one image at a time, until global bit constraint is satisfied.

3. Experimental Results

We use all images in CLIC 2020 dataset for training, including both professional and mobile sub-sets. After randomly resizing, the input image is cropped into multiple 192×192 patches to train the networks. $H_h = 192, W_h = 192$ is used for the input image of CR and the output image of SR, while $H_l = 96, W_l = 96$ is used for the output image of CR and the input image of SR. We choose the scale of 2 for CR and SR, and set the parameter $C_1 = 32$, $C_2 = 16$. We set the parameter $\lambda = 0.7$, which shows relatively better performance in [7]. Two NVIDIA Tesla V100 GPUs are used during training and validation phase. The Adam [5] optimizer is adopted in the experiments. The initial learning rate is set to 10^{-4}, and the batch size is 64.

In Figure 4, we compare rate-distortion results of our model on CLIC 2020 validation dataset with BPG (Better

Table 1. Evaluation results on CLIC 2020 validation dataset.

Method	bit rate	MS-SSIM			
	(bpp)	Mean	Max	Min	SD
Ours	0.1499	**0.9648**	**0.9852**	**0.9193**	**0.0142**
BPG	0.1498	0.9519	0.9796	0.8968	0.0164

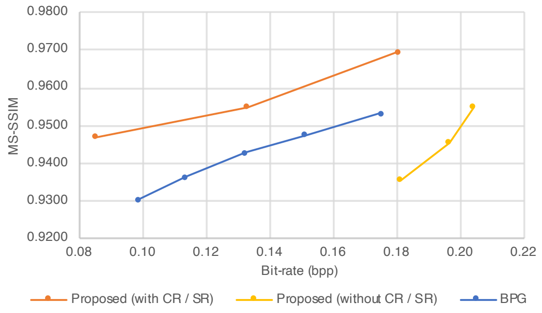

Figure 4. Rate-distortion performance of different models on CLIC 2020 validation dataset.

Portable Graphics) which is a state-of-the-art engineered image codec. Obviously, the proposed method can outperform BPG 4:2:0 within the low-rate range. For BPG codec, a QP range from 37 to 41 are used to generate bitstream and decoded images. For fairness, we apply the same image-level rate control algorithm discussed above to these two methods, and then we can evaluate their overall performance under the same bit rate constraint, which should be lower than 0.15 bits per pixel. The results on CLIC 2020 validation dataset are listed in Table 1, from which we can see that our proposed method outperforms BPG by higher average, maximum and minimum MS-SSIM scores. Meanwhile, our method maintains a less severe MS-SSIM fluctuation (SD: standard deviation) across 102 validation images.

To validate the effectiveness of the proposed super-resolution based method, we conduct more experiments by training and testing the proposed method with original 192 × 192 image patches with removal of CR/SR models. Results of this ablation experiment are also depicted in Figure 4, from which we can find that the adoption of the paired compact-resolution and super-resolution networks shall account for the performance gain. Additionally, the performance of proposed method can be further improved if a larger image dataset could be trained and tested.

4. Conclusion

In this paper, we propose a novel learned image compression framework based on super-resolution learning. The use of paired compact-resolution (CR) and super-resolution (SR) networks in proposed framework shall be highlighted. Besides, efforts such as designing efficient non-local attention modules and providing phase congruency are also made to facilitate training convergence. From the ablation experiment, it can be seen that the adoption of proposed paired CR and SR networks can be beneficial for learning-based low-rate image compression tasks.

References

[1] 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, 2, 3

[2] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Deep residual learning for image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 3

[3] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673, 2018. 2

[4] Feng Jiang, Wen Tao, Shaohui Liu, Jie Ren, Xun Guo, and Debin Zhao. An end-to-end compression framework based on convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):3007–3018, 2017. 2

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[6] Peter Kovesi. Phase congruency: A low-level image invariant. *Psychological research*, 64(2):136–148, 2000. 2, 3

[7] Yue Li, Dong Liu, Houqiang Li, Li Li, Zhu Li, and Feng Wu. Learning a convolutional neural network for image compact-resolution. *IEEE Transactions on Image Processing*, 28(3):1092–1107, 2019. 1, 2, 3

[8] Haojie Liu, Tong Chen, Qiu Shen, and Zhan Ma. Practical stacked non-local attention modules for image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2019. 3

[9] Majid Rabbani and Rajan Joshi. An overview of the jpeg 2000 still image compression standard. *Signal processing: Image communication*, 17(1):3–48, 2002. 1

[10] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012. 1

[11] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*, 2017. 1

[12] Gregory K. Wallace. The jpeg still picture compression standard. *Communications of the Acm*, 38(1):xviii–xxxiv, 1992. 1

[13] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2

[14] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 3

[15] Lei Zhou, Zhenhong Sun, Xiangji Wu, and Junmin Wu. End-to-end optimized image compression with attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 3