MCM: Multi-channel Context Model for Entropy in Generative Image Compression

Yeda Chen, Qingzhu Yuan, Ziwen Zhang, Yi Feng, XiangJi Wu
AttrSense Inc
wuxiangji@attrsense.com

Abstract

In this paper, we present an extended end-to-end image compression network for CLIC2021 image compression track. Perception loss and GAN loss are taken into account for better human perception of image quality, a simple attention module is deployed to enhance the network to capture structures and edges of objects in images which human may pay more attention to. Besides, in order to fully utilize the decoded information, we split features into multiple splits and recursively decode features. Also, we recognize that the image compression tasks have to have some rigor. Pure perceptually optimized model might introduces eye pleasing but yet fake details, this issue could be enlarged for low rate compression tasks. Therefore, our model is optimized both perceptually and objectively. Thus, the model is able to generate much more visually pleasing reconstructions compared to traditional compression methods, while maintaining the authenticity of small details. We demonstrate our methods with human-judged experiments.

1. Introduction

Recently, subjective evaluation methods of image compression has derived great interests, many perception critics are proposed [5, 1, 4, 2, 7]. Compared with traditional image quality evaluation critics, well designed perception loss can analyze the inner feature of images generated by network, which is more precious to assess structural information. As to GAN loss, this is an excellent method to supervise the reconstruction to follow the distribution of original images, resulting in more realistic visual effects compared to traditional evaluation critics [2, 7].

Instead of considering each pixel equally, human eyes are usually more sensitive to image color contrast and structure information. To address this problem, Kim *et al*. [8] and Woo *et al*. [12] developed channel/spatial attention module for feature selection and calibration during image reconstruction process.

One more thing to consider: The practical significance of pure perceptual evaluation. How we value the importance of different metrics in terms of the original intention of the image compression task. The human evaluation process described by the task is Full-Reference Image Quality Assessment, this might introduce one potential issue, i.e. different individual has different region of interest. Thus the difference of texture pattern, brightness, saturation, hue, sharpness, contrast, and the general image content stability between the reconstruction and the original image, and other perceptual evaluation methods such as MS-SSIM, FID, LPIPS, NIQE, MMD, PIM, DISTS, etc. According to our human evaluation experiments, we can hardly define a unified evaluation equation to meet different individual's tastes, details will be given in Section 4. In addition, [3] proposed a reasonable explanation that perceptual-distortion tradeoff is an inevitable matter of fact. Thus we decide to design our model to cover as much aspects as possible, yet remaining our core value, following the original data and reduce the fake content generation effects.

2. Method

2.1. Overview of the Proposed Model

Our variable rate image compression framework, which is illustrated in Figure 1, has three main parts: encoder, decoder and entropy prediction module. The encoder module consists of space2depth, residual block, attention module and gain unit [6], while the decoder module consists of inversed gain unit [6], stacked residual block, residual block, attention module and depth2space. For encoder and decoder, we adapt GDN and iGDN as activations respectively.

Our stack residual block consists of five normal residual blocks with 3*3 kernel. Note that the inputs of the stack residual block are concatenation of decoded feature and

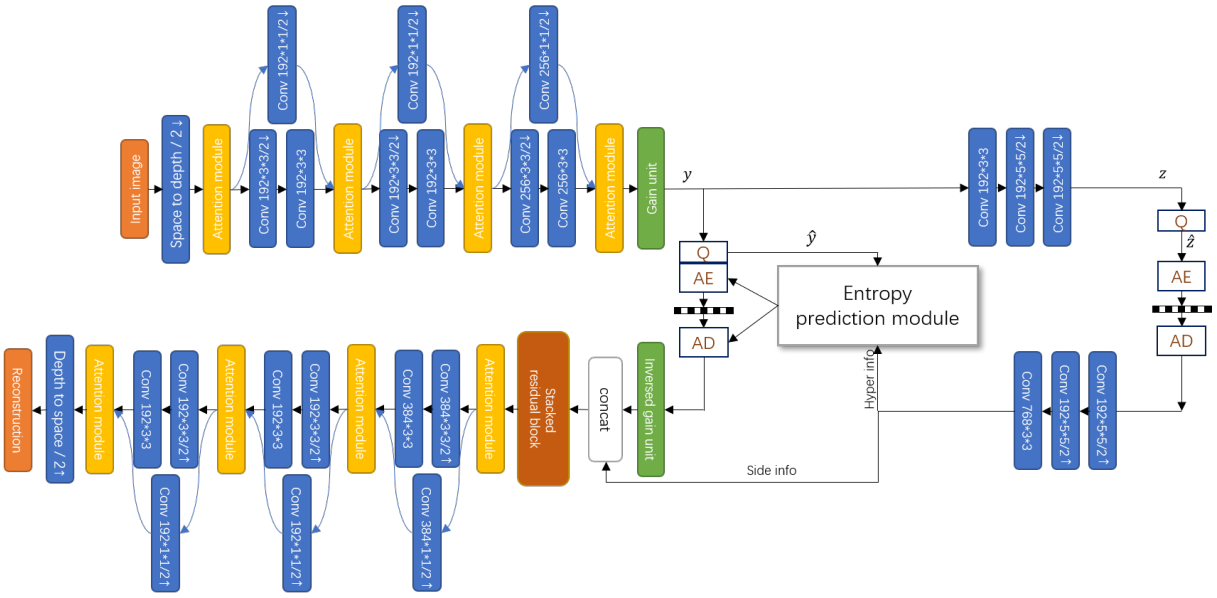


Figure 1. Overview of the proposed architecture.

side information from the hyper part. This stack residual block has the benefit of generating better textures, since wider convolution kernels can help reconstruct more details of images [11].

2.2. Attention module

The proposed attention module, shown in Figure 3, is implemented after each downsampling and before each upsampling. During the practice of attention modules tuning, we realize that heavy attention modules would incur network hard to train, especially for generative adversarial training process. Two convolutions are replaced by space2depth and depth2space layers respectively to alleviate unstable training situation of too deep networks.

2.3. Entropy prediction module

The entropy prediction module is inspired by channel split [10] and multi-scale mask convolutions [6]. In our proposed framework, quantized feature y is split into 8 parts, and for each part, we do multi-scale mask convolutions to traverse each pixel. Each split part will utilize the decoded feature slice as prior knowledge to predict mu and sigma of current feature, except the first split part, as shown in red dash line in Figure 2.

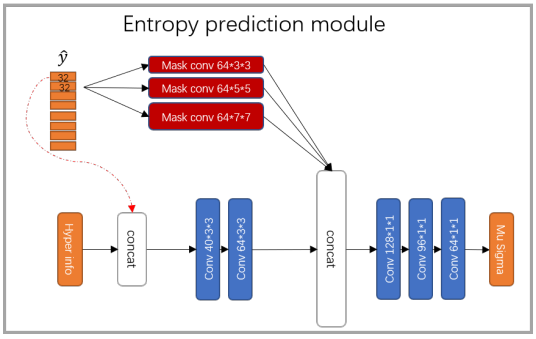


Figure 2. Illustration of the entropy prediction model

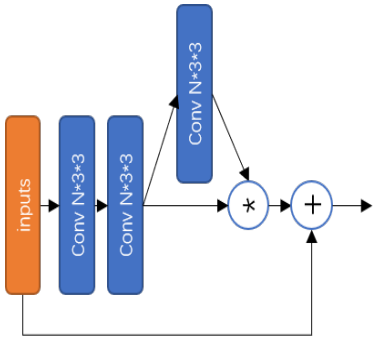


Figure 3. Illustration of attention module

image name	model	Bpp	PSNR	MSSSIM	LPIPS	Pim	DISTS
felix-russell-saw-140699.png	hific	0.177	33.01	0.9665	0.0447	9.998	0.0533
	anf	0.170	34.74	0.9700	0.0600	9.843	0.0556
todd-quackenbush-27493.png	hific	0.142	34.91	0.9868	0.0211	6.623	0.0394
	anf	0.139	36.38	0.9884	0.0261	6.817	0.0498

Table 1. Metrics tested on two images sampled from CLIC valid.

2.4. Loss

We optimized the model in a two-step-training fashion:

Step1. In the first step, the following loss function was trained for 1,000,000 steps.

$$L = \lambda \cdot (D_{mse} + \alpha \cdot D_{LPIPS} + \beta \cdot D_{msssim}) + R, \quad (1)$$

$\mathcal{D}_{mse}$ represents mean squared error loss, $\mathcal{D}_{LPIPS}$ represents perceptual loss [9], $\mathcal{D}_{msssim}$ represents multi-scale SSIM loss, R represents the total rate loss, while λ is factor to balance the quality of reconstruction and total bits. α, among 0.01$\sim$0.006, is a factor to control LPIPS loss and β, among 0.0$\sim$0.02, controls multi-scale MS-SSIM loss. From our experience, loss only with LPIPS can reconstruct bright-coloured images better, but not the small region structural accuracy. Therefore, using β to mix some part of multi-scale SSIM has benefits to recover the structure of details, especially for high-rate compression situations.

Step2. For the second step, the following loss function was used. $\mathcal{D}_{gan}$ represents generative adversarial loss. Base on the pretrained model generated in step 1, we add generative adversarial loss to train the model again.

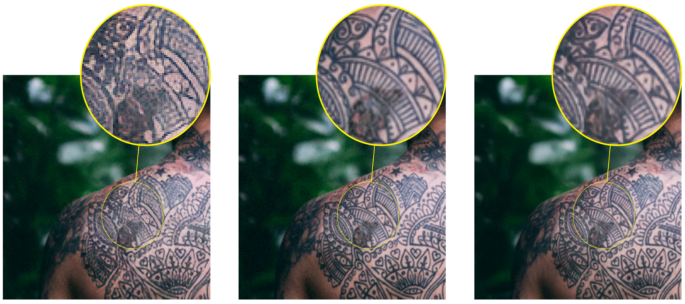
$$L = \lambda \cdot (D_{mse} + \alpha \cdot D_{LPIPS} + \beta \cdot D_{msssim} + 0.001 \cdot D_{gan}) + R, \quad (2)$$

3. Experimental setup

Training data Our model is first trained on a union of different data sets, including OpenImage, CLIC 2020, COCO 2014, COCO 2017, and YFCC 100m. We also download 121226 high resolution images (averaged in 4k*4k) from Unsplash with highest quality. The experiments show that the model trained on Unsplash can provide the best visual results without bending the details too much.

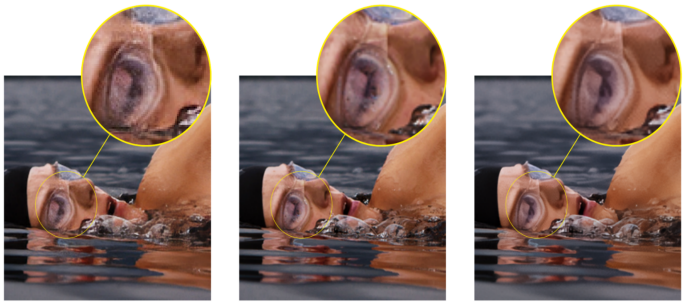
4. Results

Qualitative results In Figure 4, we compare two images sampled from CLIC Valid set between hific [9] and our model names anf on Table 1. We can conclude that anf has more authentic textures of reconstructed images.



(a) hific (b) original (c) ours

Figure 4. comparison between hific[9] and our model tested on felix-russell-saw-140699.png



(a) hific (b) original (c) ours

Figure 5. comparison between hific [9] and our model tested on todd-quackenbush-27493.png

Quantitative results on CLIC2021 valid set We evaluate our proposed model on all 41 CLIC 2021 valid images. For PSNR and MS-SSIM our model has higher score than hific [9], while hific performs slightly better on LPIPS as in Figure 6. We also carry out a human evaluation process where 20 guests are invited to do a blind selection between hific and our reconstructions. For a total 840 votes, our model has 570 votes, hific has 202 votes, and 48 votes are discarded. figure 4 and figure 5 show different reconstruction details of hific and our model. Although hific has lower lpips score, our model still generates better details compared to the hific.

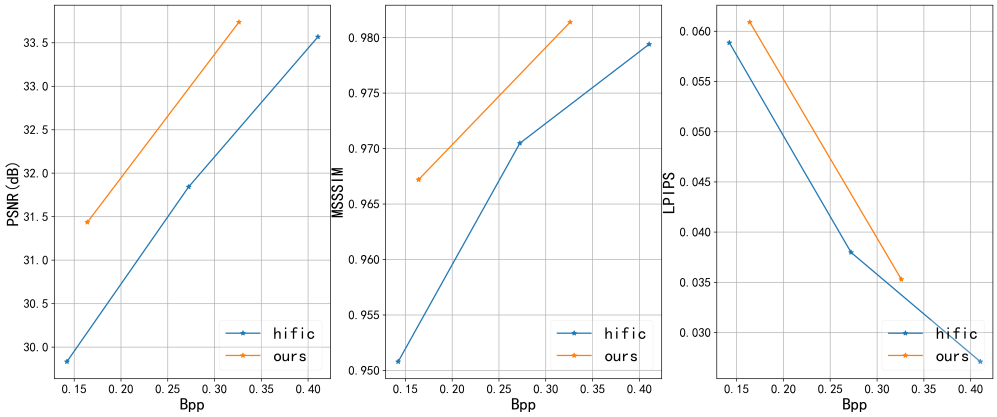


Figure 6. PSNR, MSSSIM, LPIPS comparison between hific [9] and our model

5. Conclusion

In this paper, we propose a novel image compression model for human perceptual evaluation. Multi-channel context model is adopted to capture both channel and spatial prior information. Attention mechanism and space2depth modules are applied to balance more semantic information and less calculation. Besides, we design an effective loss function to prevent eye pleasing but yet fake details, through manual evaluation, such real features are more convincing. As shown in the results of the validation set, our model "anf" yeilds outstanding performance in both objective and perception metric.

References

[1] Sangnie Bhardwaj, Johannes Ballé, Ian Fischer, and Troy Chinen. An Unsupervised Information-Theoretic Perceptual Quality Metric. *arXiv*, (NeurIPS):1–19, 2020. 1

[2] Mikołaj Binkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD gans. *arXiv*, (2017):1–36, 2018. 1

[3] Y. Blau and T. Michaeli. The perception-distortion tradeoff. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 1

[4] 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*, pages 1–14, 2020. 1

[5] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Comparison of Full-Reference Image Quality Models for Optimization of Image Processing Systems. *International Journal of Computer Vision*, pages 1–25, 2021. 1

[6] Tiansheng Guo, Jing Wang, Ze Cui, Yihui Feng, Yunying Ge, and Bo Bai. Variable rate image compression with content adaptive optimization. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2020-June:533–537, 2020. 1, 2

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 2017-December(Nips):6627–6638, 2017. 1

[8] Jun Hyuk Kim, Jun Ho Choi, Manri Cheon, and Jong Seok Lee. MAMNet: Multi-path adaptive modulation network for image super-resolution. *Neurocomputing*, 402:38–49, 2020. 1

[9] Fabian Mentzer, George Toderici, Michael Tschannen, and Eirikur Agustsson. High-Fidelity Generative Image Compression. *arXiv*, (NeurIPS), 2020. 3, 4

[10] D. Minnen and S. Singh. Channel-wise autoregressive entropy models for learned image compression. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3339–3343, 2020. 2

[11] Mingxing Tan and Quoc V Le. EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks. 2019. 2

[12] Sanghyun Woo, Jongchan Park, Joon Young Lee, and In So Kweon. CBAM: Convolutional block attention module. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11211 LNCS:3–19, 2018. 1