PO-ELIC: Perception-Oriented Efficient Learned Image Coding

Dailan He*, Ziming Yang*, Hongjiu Yu*, Tongda Xu, Jixiang Luo,
Yuan Chen, Chenjian Gao, Xinjie Shi, Hongwei Qin
SenseTime Research
{hedailan, yangziming, yuhongjiu}@sensetime.com

Yan Wang†
SenseTime Research
Tsinghua University
wangyan1@sensetime.com
wangyan@air.tsinghua.edu.cn

Abstract

In the past years, learned image compression (LIC) has achieved remarkable performance. The recent LIC methods outperform VVC in both PSNR and MS-SSIM. However, the low bit-rate reconstructions of LIC suffer from artifacts such as blurring, color drifting and texture missing. Moreover, those varied artifacts make image quality metrics correlate badly with human perceptual quality. In this paper, we propose PO-ELIC, i.e.*, Perception-Oriented Efficient Learned Image Coding. To be specific, we adapt ELIC, one of the state-of-the-art LIC models, with adversarial training techniques. We apply a mixture of losses including hinge-form adversarial loss, Charbonnier loss, and style loss, to finetune the model towards better perceptual quality. Experimental results demonstrate that our method achieves comparable perceptual quality with HiFiC with much lower bitrate.*

1. Introduction

Learned image compression (LIC) has outperformed traditional methods like JPEG [29] and BPG [4] in terms of PSNR and MS-SSIM. In 2018, the classical hyperprior framework [3, 24] dramatically improves the rate distortion performance of LIC. More recently, various context models [16, 21] have been proposed to accurately predict the distribution of latents, so as to further reduce bitrate. Although these models perform well on full-reference metrics, the reconstructed images show various artifacts when bpp is low (*e.g.* $\leq$ 0.3). For example, it is well-known that MSE-optimized models produce blurry reconstruction images. The similar phenomenon occurs when optimizing MS-SSIM and other metrics. Those artifacts become increasingly intolerable as bpp grows even lower (*e.g.* 0.075). In fact, no full-reference metric is fully consistent with perceptual quality, and optimizing towards any of the metrics brings visual artifacts. This is known as perception-distortion trade-off [6].

To address this issue, previous works introduce generative adversarial network (GAN) [14] to enhance perceptual quality. [1] efficiently compress images at low bit-rate and maintain image details by introducing adversarial training. HiFiC [23] exploits generator and conditional discriminator architectures for perceptual quality. However, to some extent they all face common GAN problems, such as unnatural texture and drifted color. To tackle these challenges, we follow these existing approaches and further investigate the perceptual optimized LIC. Our target is to encode images in lower bitrates with higher perceptual quality.

In this paper, we contribute in two aspects:

- We propose PO-ELIC, which can utilize lower bit-rate to achieve comparable visual quality against previous approaches. The reconstructions below 0.15bpp still retain clear and realistic details (See Fig. 3 and Fig. 4).
- We exploit the advantage of GAN at low bit-rate and context model at medium bit-rate to balance distortion and rate in Sec. 3. And Fig. 2 shows we have the fastest decoder among other learning based methods on CLIC 2022 leaderboard.

2. Background

2.1. LIC with context model

Lossy image compression aims to optimize the rate distortion function $\mathcal{R} + \lambda\mathcal{D}$. Denoting the image as x, encoder as g_a and decoder as g_s, the neural network has the following objective:

$$\mathcal{L} = \mathbb{E}[-\log p(g_a(x)) + \lambda d(x, g_s(g_a(x)))] \quad (1)$$

where $\mathbb{E}$ is the expectation over $p(x)$, g_a extracts the input image x as latent variable $\hat{y} = g_a(x)$ and g_s transforms it

*Equal contribution.
†Corresponding author.

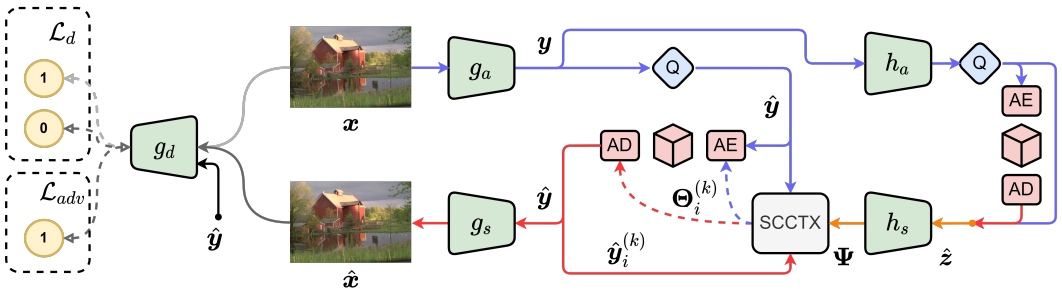

Figure 1. Diagram of the adopted networks. The right part is ELIC [15]. We use the same architecture of g_a, g_s, h_a and h_s as the original paper. SCCTX denotes the spatial-channel context model. We use the uneven 5-group scheme with parallel context models [16]. The left part shows the adversarial training. We use the same discriminator (g_d) structure as HiFiC [23].

into reconstruction $\hat{x}$. $\mathcal{D}, \mathcal{R}$ are the MSE reconstruction loss and bit-rate computed via learned prior.

Auto-regressive context model is the key factor to promote compression performance by more accurately modeling symbol probability. To be specific, the estimation of current symbol y_i can leverage previous symbols $y_{<i}$:

$$p(y_i|y_{<i}) = p(y_i|\Psi(y_{<i})) \tag{2}$$

where Ψ is context model of various form. Minnen *et al*. [24] utilizes spatial masked convolution as context model. Then channel-wise context model is proposed [25]. ELIC [15] adopts a spatial-channel context modelling.

2.2. LIC with generative adversarial networks

GAN has been successful in improving perceptual quality of end-to-end image compression [9, 12, 23]. Usually, a conditional GAN (cGAN) is adopted to constrain the consistency between the decoded image and the original input. The most common adversarial loss of GAN is the non-saturated binary cross-entropy (BCE). Given a discriminator g_d, the BCE adversarial loss is:

$$\mathcal{L}_{adv} = -\mathbb{E}\left[\log g_d(\hat{x}, \hat{y})\right] \tag{3}$$

where the condition $\hat{y} = g_a(x)$ is the coding-symbols, according to Eq. 2.1. By optimizing g_a, g_s guided by g_d we constrain the reconstruction image $\hat{x}$ to be closer to the original one. To train the discriminator g_d, an auxiliary discriminator loss is introduced:

$$\mathcal{L}_d = -\mathbb{E}\left[\log g_d(x, \hat{y})\right] - \mathbb{E}\left[\log\left(1 - g_d(\hat{x}, \hat{y})\right)\right] \tag{4}$$

Introducing the $\mathcal{L}_{adv}$ term to $\mathcal{D}$ extends the rate-distortion optimization to rate-distortion-perception optmization, as GAN demonstrates better correlation with human perception.

3. Architecture

We use ELIC [15] as our coding architecture. Fig. 1 shows its diagram. When optimizing MSE, it achieves better RD performance than VVC [8] w.r.t. both PSNR and MS-SSIM. The model adopts a multi-dimension context model SCCTX, recognizing redundancy in latents from both channel and spatial dimensions. Because of the usage of parallel context model [16], it gets rid of slow serial decoding and can decompress a 720P image within 100ms.

4. Objective

We take the rate-constrained RD optimization from HiFiC:

$$\mathcal{L} = \mathcal{D} + \lambda(\mathcal{R}, \mathcal{R}^*) \cdot \mathcal{R} \tag{5}$$

where $\mathcal{D}$ and $\mathcal{R}$ are (perceptual) distortion and rate terms. The multiplexer $\lambda(\mathcal{R}, \mathcal{R}^*)$ is conditioned on the given target bitrate $\mathcal{R}^*$:

$$\lambda(\mathcal{R}, \mathcal{R}^*) = \begin{cases} \lambda_\alpha, & \mathcal{R} \geq \mathcal{R}^* \\ \lambda_\beta, & \mathcal{R} < \mathcal{R}^* \end{cases} \tag{6}$$

Our summarized perceptual $\mathcal{D}$ loss function is:

$$\mathcal{D} = \lambda_1 \mathcal{L}_{perc} + \lambda_2 \mathcal{L}_{recon} + \lambda_3 \mathcal{L}_{adv} + \lambda_4 \mathcal{L}_{sty} \tag{7}$$

where the perceptual loss $\mathcal{L}_{perc}$ is LPIPS-VGG [30]. $\mathcal{L}_{recon}$ is a pixel-wise reconstruction loss (L_2, L_1, Charbonnier loss [20], *etc*.). $\mathcal{L}_{adv}$ is the adversarial loss, and $\mathcal{L}_{sty}$ is the style loss constraining the texture consistency. Similar loss functions have been successfully used in low-level tasks like image translation [7] and super-resolution [28]. We will discuss these loss terms in detail in this section.

4.1. Perceptual optimization with SNGAN

The BCE adversarial loss (eq. 3 and eq. 4) function suffers from the modal collapse issue [2]. Inspired by [26] and [7], we instead apply the hinge loss to train a synthesizer with a spectral normalization constrained discriminator:

$$\begin{aligned} \mathcal{L}_{adv} &= -\mathbb{E}[g_d(\hat{x}, \hat{y})] \\ \mathcal{L}_d &= -\mathbb{E}[\mathrm{ReLU}(-1 + g_d(x, \hat{y}))] \\ &\quad -\mathbb{E}[\mathrm{ReLU}(-1 - g_d(\hat{x}, \hat{y}))] \end{aligned} \tag{8}$$

Table 1. Objective results at $0.075, 0.15$ and 0.30bpp with validation dataset. ↑ means higher is better and ↓ vice versa.

BPP	PSNR↑	MSSSIM↑	LPIPS↓	FID↓	KID↓	PieAPP↓	DISTS↓	IQT↑
0.075	27.5324	0.9179	0.1982	33.8917	-0.0286	0.7560	0.0480	0.6783
0.15	30.2501	0.9424	0.1604	23.5175	-0.0292	0.4905	0.0325	0.7136
0.3	32.6412	0.9720	0.1083	13.9438	-0.0298	0.3788	0.0207	0.7377

note that when this hinge loss is used, the output of g_d is the non-activated logits. In our experiments, it outperforms the non-saturated BCE loss.

Other alternatives of BCE adversarial loss include least-square form [22] and relativistic form [19], which are also adopted by recent perceptual LIC approaches [12, 18].

4.2. Learning smoother pixel-wise reconstruction using Charbonnier loss

L_1 loss is frequently used in low-level vision tasks to provide a gentler pixel-wise supervision than L_2 (MSE) loss. However, it has an ill-defined gradient when the input is zero. We instead apply a smoother variant of L_1 loss called Charbonnier loss [20]:

$$\mathcal{L}_{recon}^{(\text{Charb})}(x) = \sqrt{x^2 + \epsilon^2}, \tag{9}$$

where we set $\epsilon = 10^{-6}$.

4.3. Improving texture generation with patched style loss

Borrowed from style-transfer [13], the style loss is widely adopted in low-level tasks to match the texture pattern (or, the so-called *style*) of source and generated images:

$$\mathcal{L}_{sty}(x, \hat{x}) = \sum_{\ell} \left\| G\left(\Phi^{(\ell)}(x)\right) - G\left(\Phi^{(\ell)}(\hat{x})\right) \right\| \tag{10}$$

where the operator $G(\cdot)$ denotes the Gram matrix of the given vector. Φ is the pretrained feature extraction network (*e.g.*, VGG) and $\Phi^{(\ell)}(x)$ is the feature map output by its ℓ-th selected layer when feed x to the network. The loss matches the global statistics of each feature map, yet the texture usually has locality. As [28], we split the feature maps to 16×16 patches and calculate this loss per patch.

This loss is connected to the LPIPS perceptual loss. In fact, an 1×1 patch style loss is the same as LPIPS without finetuning stage. The LPIPS pays more attention to constraining the global image content and style loss supervises the local texture statistics.

5. Experiments

5.1. Training settings

We use the ELIC models optimized for MSE as our pretrained models. Following previous works, we use a 8000-image ImageNet subset as training set. To optimize for the objective losses, we train each model for 500 epochs with a batch size of 128. We use Adam optimizer and cosine annealing learning rate scheduler with a base learning rate set to 8e-4.

We finetune the pretrained ELIC model with the above mentioned objective (*i.e.* perceptual loss, reconstruction loss, adversarial loss, and style loss, as summarized in eq. 7) to finally obtain the perception-oriented model.

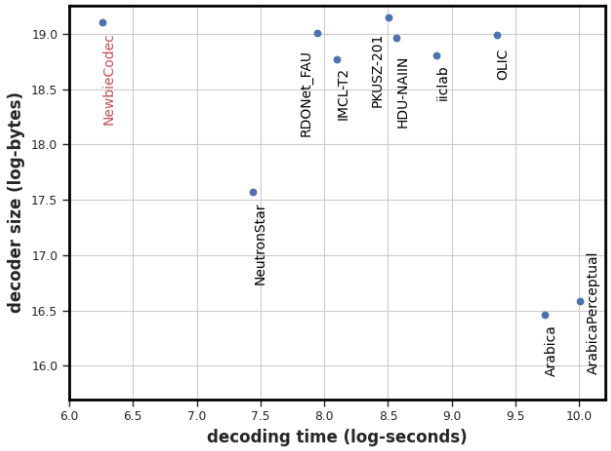

Figure 2. Logarithm decoder size and decoding time of CLIC2022 *Image 075*. Conventional coders (JPEG, BPG and AVIF) are omitted. Our method (red) is the fastest among learning based approaches even with relative large decoder. *Image 150/300* have the similar trend.

5.2. Quantitative results

To verify the effectiveness of our method, we utilize *LPIPS* [30], *FID* [17], *KID* [5], *PieAPP* [27], *DISTS* [11] and *IQT* [10] to guide the evalution of reconstructions. The combination of these scores is consistent with MOS to some degree. And our major scores are shown in Tab. 1.

5.3. Qualitative results

We compare PO-ELIC with HiFiC, and experiments demonstrate that our method has higher fidelity at even lower bit-rate. Fig. 3 shows our method has more details for dark area at right column with yellow rectangles, and more structures on the butterfly at bottom row with red rectangles. Fig. 4 gives another example.

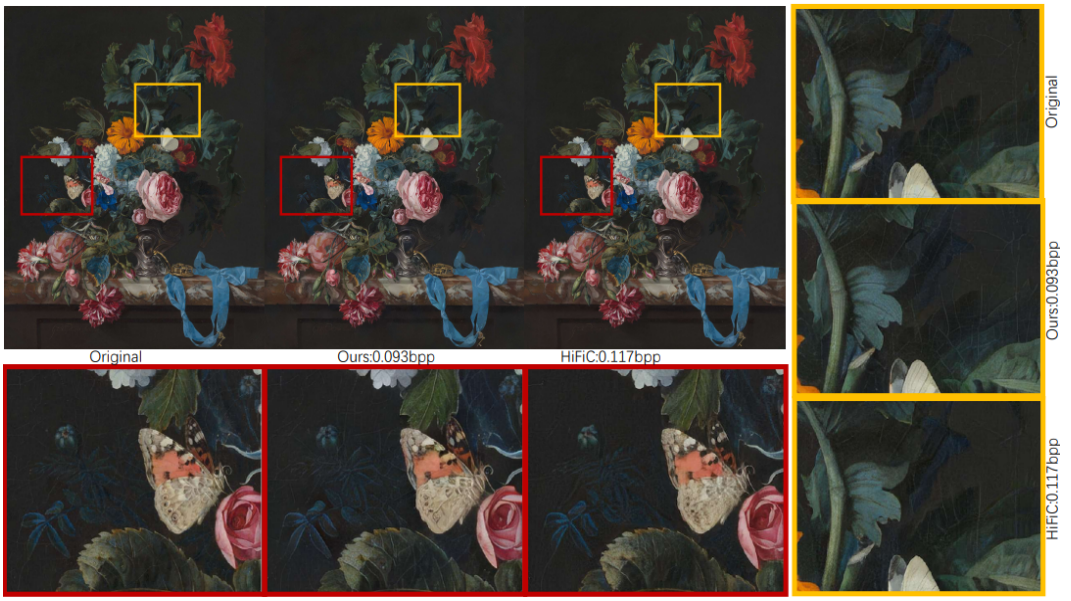

Figure 3. The visualization of our method and HiFiC at low bit-rate. Our method has more details for dark area at right column with yellow rectangles, and more structures on the butterfly at bottom row with red rectangles.

Figure 4. The visualization of our method and HiFiC at low bit-rate. Our method has more details for cable at medium row with red rectangles, and more structures on the train at bottom row with yellow rectangles.

6. Conclusion

In this paper we propose PO-ELIC, which introduces the hybrid context and generative model. It utilizes less bits and achieves more pleasant reconstructions compared to HiFiC. Moreover, it further improves the visual quality for LIC at even lower bit-rate (0.075bpp). Perceptual metrics such as *LPIPS* and *IQT* indicate that PO-ELIC obtains high-fidelity images with more texture.

References

[1] 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

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 2

[3] 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

[4] Fabrice Bellard. Bpg image format. *URL https://bellard.org/bpg*, 1:2, 2015. 1

[5] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 3

[6] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff, 2018. 1

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 2

[8] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021. 2

[9] Y. Chen, Q. Yuan, X. wu, Z. Zhang, and Y. Feng. Mcm: Multi-channel context model for entropy in generative image compression. In *4th Challenge on Learned Image Compression*, Jun 2021. 2

[10] Manri Cheon, Sung-Jun Yoon, Byungyeon Kang, and Junwoo Lee. Perceptual image quality assessment with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–442, 2021. 3

[11] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*, 2020. 3

[12] S. Gao, Y. Shi, T. Guo, Z. Qiu, Y. Ge, Z. Cui, Y. Feng, J. Wang, and B. Bai. Perceptual learned image compression with continuous rate adaptation. In *4th Challenge on Learned Image Compression*, Jun 2021. 2, 3

[13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 3

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[15] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. *arXiv preprint arXiv:2203.10886*, 2022. 2

[16] Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14771–14780, 2021. 1, 2

[17] 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*, 30, 2017. 3

[18] Shoma Iwai, Tomo Miyazaki, Yoshihiro Sugaya, and Shinichiro Omachi. Fidelity-controllable extreme image compression with generative adversarial networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8235–8242. IEEE, 2021. 3

[19] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. In *International Conference on Learning Representations*, 2018. 3

[20] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 3

[21] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. *arXiv preprint arXiv:1809.10452*, 2019. 1

[22] 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

[23] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33:11913–11924, 2020. 1, 2

[24] David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. *arXiv preprint arXiv:1809.02736*, 2018. 1, 2

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

[26] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 2

[27] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018. 3

[28] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE*

international conference on computer vision, pages 4491–4500, 2017. 2, 3

[29] Gregory K Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992. 1

[30] 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, 3