Efficient Variable Bit-Rate Neural Image Compression with Perceptual-Enhanced Optimization For CLIC2025

Daxin Li [1,2], Yuanchao Bai[1] *, Yiwei Zhang [2], Zhe Zhang [2], Sixin Lin [2], Xianguo Zhang [2], Kai Wang [1], Xianming Liu [1], Deming Zhai [1]
[1] Faculty of Computing, Harbin Institute of Technology
[2] Shannon Lab, Tencent Inc.

***Abstract*—In this paper, we propose a novel and enhanced image compression framework that builds upon the state-of-the-art (SOTA) DCVC-RT intra model, with a particular emphasis on advancing perceptual quality in compressed images. Although DCVC-RT demonstrates outstanding rate-distortion performance and real-time processing capabilities, it is still susceptible to generating perceptual artifacts, such as blurring and loss of fine textures, especially at lower bitrates. To effectively mitigate these issues, we introduce a comprehensive perceptual optimization strategy that leverages a semantic ensemble loss. This loss function is meticulously designed by integrating multiple complementary components, including Charbonnier loss for robust pixel-wise fidelity, perceptual loss to preserve high-level semantic features, style loss to maintain texture and style consistency, and a non-binary adversarial loss to further enhance the realism of reconstructed images. Our approach is developed as a solution for the CLIC2025 challenge, and we participate under the team name *Vcoder*. Through experiments, we demonstrate that our method significantly improves the perceptual quality of compressed images.**

***Index Terms*—Generative Image Compression, Learned Image Compression**

I. Introduction

In recent years, neural image compression has emerged as a transformative approach in the field of image coding, leveraging deep learning techniques to surpass the performance of traditional codecs such as JPEG, JPEG2000, and BPG [1], [11], [12]. These neural methods [5], [10], particularly those based on end-to-end optimized autoencoders, have demonstrated remarkable improvements in both rate-distortion efficiency and perceptual quality metrics. Among these, the DCVC-RT intra model [4] stands out for its state-of-the-art performance in terms of objective measures like Peak Signal-to-Noise Ratio (PSNR) and Multi-Scale Structural Similarity (MS-SSIM), as well as its real-time processing capabilities.

Despite these advances, a persistent challenge in neural image compression is the presence of perceptual artifacts, such as blurring, loss of fine textures, and unnatural color shifts, especially at lower bitrates. These artifacts can significantly degrade the subjective visual quality of reconstructed images, limiting the practical adoption of neural codecs in applications where human perception is paramount [2], [15]. Addressing this issue requires not only optimizing for traditional distortion metrics but also incorporating perceptually motivated loss functions and training strategies.

In this work, we present an enhanced version of the DCVC-RT intra model with a strong emphasis on perceptual quality improvement. Drawing inspiration from recent advances in perceptual optimization [9], [14], we introduce a novel semantic ensemble loss that integrates multiple complementary components: Charbonnier loss for robust pixel-wise fidelity [8], perceptual loss to preserve high-level semantic features [6], [15], style loss to maintain texture and style consistency [14], and a non-binary adversarial loss to further enhance the realism of reconstructed images [13]. This comprehensive loss formulation is designed to guide the training process towards generating images that are not only quantitatively accurate but also visually pleasing and semantically faithful.

Experimental results on benchmark datasets demonstrate that our enhanced DCVC-RT intra model achieves significant improvements in perceptual fidelity compared to baseline methods, delivering superior visual quality at equivalent bitrates.

II. Method

A. Architecture

Recently, DCVC-RT [4] has set a new benchmark in real-time video compression, demonstrating state-of-the-art (SOTA) performance in both efficiency and quality. The core innovation of DCVC-RT lies in its highly parallelizable intra-frame architecture, which leverages efficient depth-wise separable convolution blocks (DCBs) and a quadtree-based partitioning strategy for latent representations. This design not only accelerates encoding

Daxin Li is an intern in Shannon Lab, Tencent Inc.
* Corresponding Author: Yuanchao Bai

and decoding but also enhances the model's ability to capture diverse spatial contexts, which is crucial for high-fidelity image reconstruction.

Building upon the strengths of DCVC-RT, we adapt its intra-frame architecture for the task of learned image compression, introducing several targeted enhancements to further boost performance and perceptual quality. Specifically, we retain the original encoder and decoder structures from DCVC-RT, which are composed of stacks of DCBs in both the analysis (encoder) and synthesis (decoder) transforms. The use of DCBs, as highlighted in the DCVC-RT paper, significantly reduces computational complexity while maintaining strong representational power, making the architecture well-suited for real-time applications.

A key feature inherited from DCVC-RT is the quadtree-based partitioning of the latent space. This mechanism adaptively divides the latent representation into spatially and channel-wise varying blocks. By doing so, the model can efficiently model dependency between latent elements, improving both rate-distortion and perceptual performance. The quadtree partitioning also enables parallel entropy coding of independent blocks, further accelerating the overall compression pipeline.

To push the limits of compression quality and model expressiveness, we scale up the network significantly. Our enhanced model increases the total parameter count to 149M by expanding both the depth and width of the network. Concretely, the analysis transform is deepened to 9 layers of DCBs, while the synthesis transform is extended to 15 layers, allowing for more complex feature extraction and reconstruction. The number of channels in each layer is increased to 384, providing greater capacity for feature representation. Additionally, the dimensionality of the latent and hyper-latent spaces is enlarged to 320 and 192 channels, respectively, which facilitates richer modeling of both the primary and side information required for effective entropy coding.

These architectural modifications, inspired by the design principles and empirical findings of the DCVC-RT paper, enable our model to achieve superior compression performance, particularly in terms of perceptual quality at low bitrates. By combining the efficient and scalable backbone of DCVC-RT with our enhancements, we lay a strong foundation for the subsequent integration of advanced perceptual optimization strategies.

B. Variable Bit-Rate Training with Semantic Ensemble Loss

DCVC-RT inherently supports variable bit-rate compression within a single model by leveraging learnable quantization vectors and hyperprior models to control the bitrate. Building on this foundation, we propose a three-stage training strategy to develop a perceptually optimized variable bit-rate compression model.

In the first stage, we train the model using a standard rate-distortion loss, as in the original DCVC-RT framework. The objective function is defined as:

$$\mathcal{L} = \lambda \mathcal{R} + \mathcal{D} \tag{1}$$

where $\mathcal{R}$ denotes the rate, estimated from the noised latent representations to ensure differentiability, and $\mathcal{D}$ represents the distortion, initially measured by mean squared error (MSE) between the original and reconstructed images. To enable training across a range of bitrates, we sample 64 different Lagrange multipliers λ per batch, which are linearly spaced in the log-domain within the range $[0.004, 0.1]$.

In the second stage, we fine-tune the model using a combined distortion metric $\mathcal{D} = 150 \cdot \mathcal{L}_{\text{MSE}} + 1.0 \cdot \mathcal{L}_{\text{LPIPS}}$, incorporating LPIPS [15]. The inclusion of LPIPS encourages the model to retain more perceptually relevant and high-frequency information in the latent representations, thereby enhancing the perceptual quality of the reconstructions.

In the final stage, we freeze the encoder and exclusively fine-tune the decoder using a semantic ensemble loss. This loss, which redefines the distortion term $\mathcal{D}$, integrates a Charbonnier loss [8], a perceptual loss [15], a style loss [14], and a non-binary adversarial loss [13]. This comprehensive formulation guides the model to produce reconstructions that are both visually realistic and perceptually appealing. The semantic ensemble loss is defined as:

$$\mathcal{D} = \alpha \mathcal{L}_{\text{rec}} + \beta \mathcal{L}_{\text{per}} + \gamma \mathcal{L}_{\text{style}} + \delta \mathcal{L}_{\text{adv}} \tag{2}$$

where α, β, γ, and δ are the weighting coefficients, which we empirically set to 64, 1.0, 0.1, and 0.01, respectively. Specifically, the reconstruction loss $\mathcal{L}_{\text{rec}}$ employs the Charbonnier loss [8] to assess pixel-level similarity between the original and compressed images. The perceptual loss $\mathcal{L}_{\text{per}}$ [15] is calculated as the L_2 distance between VGG-extracted features of the original and reconstructed images. The style loss $\mathcal{L}_{\text{style}}$ [14] is computed as the L_2 distance between Gram matrices of 16×16 feature patches, promoting the preservation of local texture information. The adversarial loss $\mathcal{L}_{\text{adv}}$ is derived from the output of a discriminator network.

To further enhance the realism of the reconstructed images, we adopt the non-binary adversarial discriminator from [13]. This discriminator, as detailed therein, is designed to more effectively distinguish between real and generated images at a semantic level. The adversarial training objectives are as follows:

$$\begin{aligned} \mathcal{L}_{\text{disc}}(\boldsymbol{\phi}) &= \mathbb{E}_{\boldsymbol{x} \sim P_{\boldsymbol{X}}} \left[- \langle u(\boldsymbol{x}), \log D_{\boldsymbol{\phi}}(\boldsymbol{x}) \rangle \right] \\ &\quad + \mathbb{E}_{\hat{\boldsymbol{x}} \sim P_{\hat{\boldsymbol{X}}}} \left[- \langle \boldsymbol{b}_0, \log D_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}) \rangle \right] \\ \mathcal{L}_{\text{adv}}(\boldsymbol{\varphi}, \boldsymbol{\omega}, \boldsymbol{v}) &= \mathbb{E}_{\hat{\boldsymbol{x}} \sim P_{\hat{\boldsymbol{X}}}} \left[- \langle u(\boldsymbol{x}), \log D_{\boldsymbol{\phi}}(\hat{\boldsymbol{x}}) \rangle \right] \end{aligned} \tag{3}$$

where $D_{\boldsymbol{\phi}}$ denotes the discriminator parameterized by $\boldsymbol{\phi}$. The vector $u(\boldsymbol{x})$ is a one-hot encoding indicating the

TABLE I: Performance of our model on the CLIC2025 Test set. Objective results at 0.075, 0.15 and 0.30bpp. ↑ means higher is better and ↓ vice versa. The decoding time is measured on whole CLIC2025 Test set using a single NVIDIA L4 GPU and a AMD EPYC 7R13 CPU. Note that the PSNR and MS-SSIM are measured by averaging each image in the test set, which is **different** from the Leaderboard metrics.

Method	BPP	PSNR↑	MSSSIM↑	LPIPS↓	DISTS↓	Decoding Time (s)
DCVC-RT	0.075	28.39	0.9249	0.3965	0.2144	18
	0.15	30.85	0.9572	0.3332	0.1641	18
	0.30	33.68	0.9771	0.2655	0.1160	18
Ours	0.075	26.17	0.9014	0.1968	0.0495	24
	0.15	28.29	0.9437	0.1458	0.0305	24
	0.30	30.80	0.9716	0.1026	0.0183	24

closest codebook entry for $\boldsymbol{x}$, and $\boldsymbol{b}_0$ is the label assigned to generated (fake) samples. The encoder, entropy model, and decoder are parameterized by $\boldsymbol{\varphi}$, $\boldsymbol{\omega}$, and $\boldsymbol{v}$, respectively.

III. Experiments

A. Experimental Settings

To train our model, we utilize image patches of size 256×256 randomly sampled from the test split of the OpenImages V7 dataset [7]. The patches are augmented with random horizontal flips and rotations. The training process is carried out using the AdamW optimizer, with hyperparameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Each stage of training is run for a total of 2 million iterations, ensuring thorough convergence and robust learning of the model parameters. This extensive training regimen allows the model to effectively capture both low-level and high-level image features, which are crucial for high-fidelity compression and reconstruction.

For evaluation, we adopt the CLIC2025 Test set, which consists of 30 high-resolution images at 2K resolution. This challenging benchmark provides a comprehensive assessment of our model's performance in real-world scenarios. We evaluate our method using a suite of both distortion-based and perceptual quality metrics. Specifically, we report mean squared error (MSE), multi-scale structural similarity (MS-SSIM), and learned perceptual image patch similarity (LPIPS) [15]. In addition, we include the DISTS metric [3] as a reference-based perceptual measure. This comprehensive evaluation protocol ensures a holistic understanding of both the objective and subjective quality of the compressed images.

B. Quantitative Results

To rigorously validate the effectiveness of our proposed approach, we conduct extensive quantitative experiments on the CLIC2025 Test set. We also provide the test results of DCVC-RT for reference. The detailed results are summarized in Table I.

C. Qualitative Analysis

In addition to quantitative evaluation, we provide qualitative comparisons to further illustrate the advantages of our method. We use the perceptual optimized MS-ILLM [13] as the baseline for comparison. As depicted in Fig. 1, Fig. 2, and Fig. 3, our approach produces reconstructions that are visually closer to the original images compared to competing methods, especially at equivalent bitrates. Notably, our model excels at preserving intricate details and textures, such as the fine structure of leaves, plant surfaces, and subtle features like eye details. These results demonstrate that our method not only achieves high compression ratios but also maintains a high degree of visual fidelity, making it particularly suitable for applications where perceptual quality is paramount.

IV. Conclusion

In this work, we propose an advanced image compression framework that builds upon the state-of-the-art (SOTA) DCVC-RT intra model, with a particular emphasis on enhancing perceptual quality. By introducing a novel perceptual optimization strategy—centered around a semantic ensemble loss that integrates Charbonnier loss, perceptual loss, style loss, and a non-binary adversarial loss—we enable the model to generate reconstructions that are both visually realistic and semantically meaningful. Extensive experiments on the challenging CLIC2025 Test set demonstrate that our enhanced DCVC-RT intra model achieves significant improvements in perceptual fidelity, delivering superior visual quality at comparable bitrates. These results underscore the effectiveness of our approach and its potential for advancing the field of learned image compression.

References

[1] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[2] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:215763824

Fig. 1: Visual comparison of Ground Truth, our method, and perceptual optimized MS-ILLM [13] on *d0208a1e9b4fde642b9752da2907d82d5b530a6e2e6e18ef71d5524a92d1cf9a.png* from the CLIC2025 Test set. The reconstructed images are generated at 0.075bpp.

[3] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 5, pp. 2567–2581, 2020.

[4] Z. Jia, B. Li, J. Li, W. Xie, L. Qi, H. Li, and Y. Lu, "Towards practical real-time neural video compression," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 543–12 552.

[5] W. Jiang, J. Yang, Y. Zhai, P. Ning, F. Gao, and R. Wang, "Mlic: Multi-reference entropy model for learned image compression," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7618–7627.

[6] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016*, 2016.

[7] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *IJCV*, 2020.

[8] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 624–632.

[9] D. Li, Y. Bai, K. Wang, J. Jiang, and X. Liu, "Semantic ensemble loss and latent refinement for high-fidelity neural image compression," in *2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2024, pp. 1–5.

[10] D. Li, Y. Bai, K. Wang, J. Jiang, X. Liu, and W. Gao, "Grouped-mixer: An entropy model with group-wise token-mixers for learned image compression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 10, pp. 9606–9619, 2024.

[11] F. Mentzer, L. V. Gool, and M. Tschannen, "Learning better lossless compression using lossy compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6638–6647.

[12] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in neural information processing systems*, vol. 31, 2018.

[13] M. J. Muckley, A. El-Nouby, K. Ullrich, H. Jégou, and J. Verbeek, "Improving statistical fidelity for neural image compression with implicit local likelihood models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 25 426–25 443.

[14] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4491–4500.

[15] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

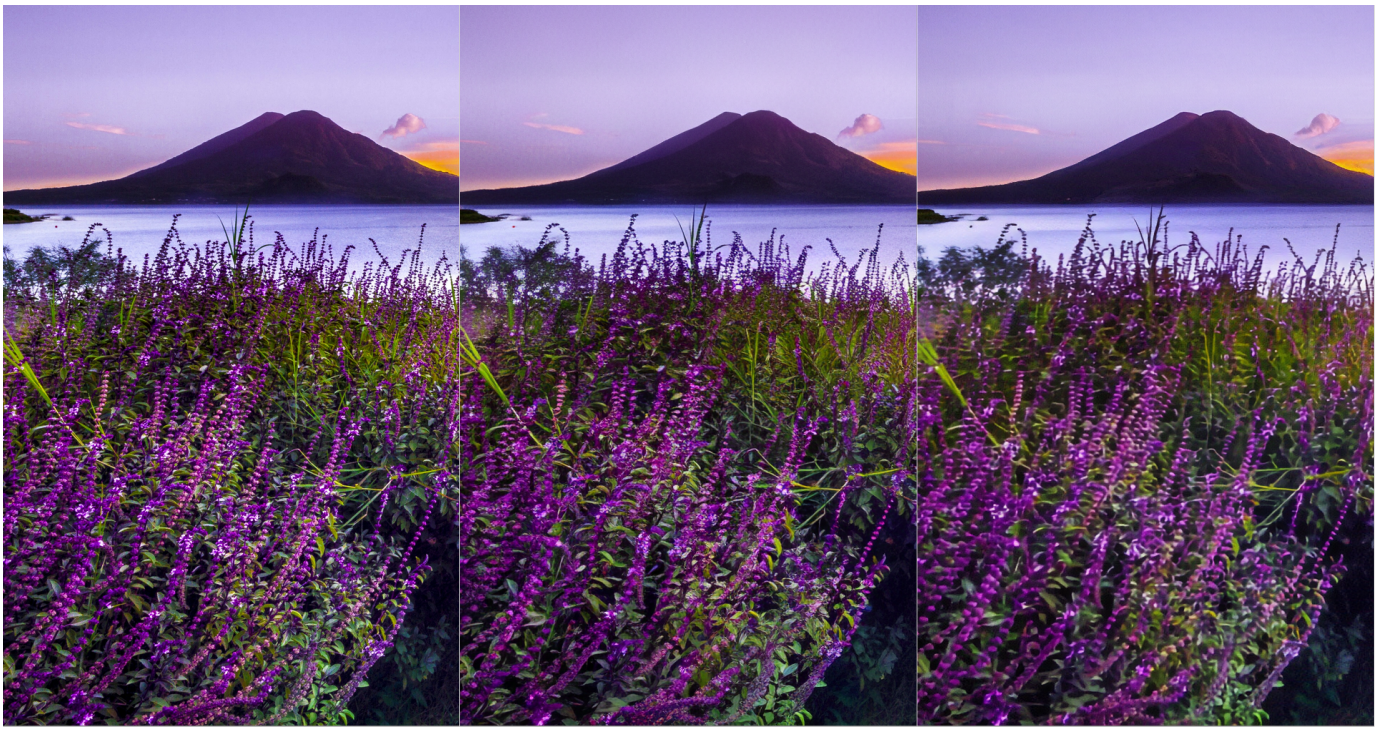

Fig. 2: Visual comparison of Ground Truth, our method, and perceptual optimized MS-ILLM [13] on *b0b3744cc6b5ad6426b1ac02909b4389ea7fc4140d72a888ccdfb21f573a6db4.png* from the CLIC2025 Test set. The reconstructed images are generated at 0.075bpp.

Fig. 3: Visual comparison of Ground Truth, our method, and MS-ILLM [13] on *2684452db505ddbbb53f42a3f3bcfe86fdd0d6d8d98c029db4b4c6fc1f55b750.png* from the CLIC2025 Test set. The reconstructed images are generated at 0.075bpp.