Stable Training for Perception-Oriented Learned Image Compression

Renjie Zou
Malanshan audio & video Laboratory
Changsha, China
zourenjie@mlslabs.com.cn

Defa Wang
Malanshan audio & video Laboratory
Changsha, China
wangdefa@mlslabs.com.cn

Zhiwei Huang
Malanshan audio & video Laboratory
Changsha, China
huangzhiwei@mlslabs.com.cn

***Abstract*—Despite achieving state-of-the-art rate-distortion performance exceeding VVC in PSNR and MS-SSIM, recently learned image compression (LIC) methods still exhibit significant perceptual limitations at low bitrates. Reconstructed images often suffer from blurring, inaccurate colors, and lack of textural detail, highlighting the well-known divergence between conventional metrics and human visual perception. Although several perceptual LIC approaches have been proposed to bridge this gap, many are hampered by unstable training that hinders their practical applicability. To bridge this gap, we propose ST-LIC (Stable Training for Perception-Oriented Learned Image Compression). Our approach introduces two key innovations for stable and effective perceptual optimization. First, during the initial training phase, we analyze the gradient contribution of each loss component to identify a balance point, preventing any single loss from dominating or becoming negligible during updates. Second, we integrate a UNet-based refiner module after the decoder. This module applies distortion and perceptual losses to distinct outputs, enabling a more precise and balanced optimization of the Rate-Distortion-Perception trade-off. Experimental results demonstrate that ST-LIC achieves significantly more stable training when incorporating adversarial loss while simultaneously delivering reconstructions with superior subjective visual quality. And we compete under the team name Evolve.**

I. Introduction

Despite significant advances in traditional evaluation metrics such as PSNR and MS-SSIM—where many Learned Image Compression (LIC) methods [2]–[6] now outperform even traditional standards like VVC [1]. Reconstructed images, particularly at low bit rates, continue to exhibit blurring, color shifts, and texture degradation that are poorly captured by conventional distortion measures. Although several perception-oriented methods [7], [8] have improved visual alignment with human perception, many however suffer from unstable training and limited robustness owing to the adversarial loss [9].

To address this issue, we propose an UNet-based refiner module that operates on the output of the decoder. Our method significantly improves the quality of subjective perception with only a small impact on rate-distortion performance. Experimental results demonstrate that the proposed approach achieves pleasing perceptual quality while maintaining stable and efficient training across diverse image contents.

In this paper, we contribute in two aspects:

- During the initial training phase, we analyze the gradient contribution of each loss component to identify a balance point, preventing any single loss from dominating or becoming negligible during updates.
- We introduce a UNet-based refiner module that processes the decoder's output. By applying distortion and perceptual losses to separate outputs, our model enables a more precise and balanced optimization of the Rate-Distortion-Perception trade-off [10].

II. Proposed Methods

A. Framework Overview

The proposed ST-LIC is built upon LALIC [6]. We retain the original encoder, decoder, and entropy model unchanged, and introduce a UNet-based refiner module to post-process the decoder's output. An overview of the framework is illustrated in Fig. 1.

B. Objective Optimization

We decompose the overall objective into two distinct components: distortion loss and perceptual loss. The distortion term comprises ℓ_1 and ℓ_2 losses, which focus on pixel-level fidelity and are applied only to the main network, i.e., the encoder, decoder, and entropy model. In contrast, the perceptual term, including LPIPS [11], Wasserstein loss [12], and adversarial loss [9] for visual realism, is applied globally, encompassing both the main network and the UNet-based refiner module.

Leveraging the properties of second-order optimizers, we assign relatively small coefficients λ to all perceptual losses. This design encourages them to exert their primary influence on the refiner while imposing only mild gradient updates on the main network. As a result, we achieve a semi-decoupled optimization between distortion and perceptual objectives — effectively stabilizing training while preserving the distinct roles of each module.

Our overall objective function is decomposed into:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{distortion}} + \mathcal{L}_{\text{perceptual}} + \lambda_{rate} \cdot \mathcal{L}_{\text{rate}} \tag{1}$$

The distortion loss comprises pixel-level fidelity terms applied exclusively to the main compression network $\Theta_{\text{main}} = \{\text{encoder, decoder, entropy model}\}$:

$$\mathcal{L}_{\text{distortion}} = \lambda_{\ell_1} \cdot \mathcal{L}_{\ell_1} + \lambda_{\ell_2} \cdot \mathcal{L}_{\ell_2} \tag{2}$$

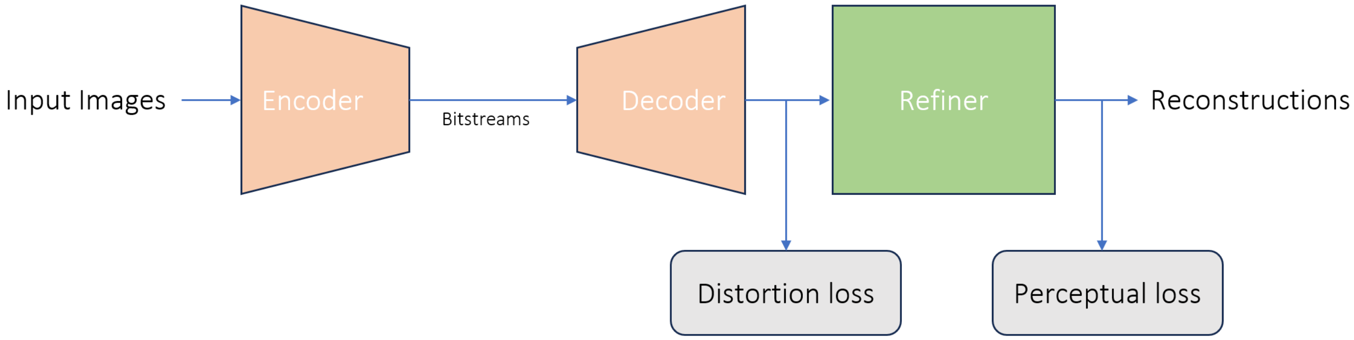

Fig. 1. The framework overview of proposed ST-LIC.

where x is the original image and $\hat{x}$ is the reconstructed image from the main network.

The perceptual loss encompasses multiple criteria applied globally to both the main network and the refiner module Θ_{refiner}:

$$\begin{aligned}\mathcal{L}_{\text{perceptual}} =&\lambda_{lpips} \cdot \mathcal{L}_{\text{LPIPS}} + \lambda_{wasserstein} \cdot \mathcal{L}_{\text{Wasserstein}} \\ &+ \lambda_{adversarial} \cdot \mathcal{L}_{\text{ adversarial}}\end{aligned} \tag{3}$$

We leverage second-order optimization properties by assigning relatively small coefficients to all perceptual losses. This coefficient scheme encourages the perceptual losses to primarily influence the refiner module while imposing only mild gradient updates on the main network parameters.

$$\nabla_{\Theta_{\text{refiner}}} \mathcal{L}_{\text{total}} = \nabla_{\Theta_{\text{refiner}}} \mathcal{L}_{\text{perceptual}} \tag{4}$$

$$\nabla_{\Theta_{\text{main}}} \mathcal{L}_{\text{total}} = \nabla_{\Theta_{\text{main}}} \mathcal{L}_{\text{distortion}} + \nabla_{\Theta_{\text{main}}} \mathcal{L}_{\text{perceptual}} \tag{5}$$

$$+ \lambda_{rate} \cdot \nabla_{\Theta_{\text{main}}} \mathcal{L}_{\text{rate}} \tag{6}$$

C. Gradient-Guided coefficient Initialization

During the initial training phase, we analyze the gradient contribution of each loss component to identify a balance coefficient strategy. Table I shows the gradient contribution of each loss component to different modules in the early stage.

TABLE I
Gradient Contribution Percentages by Module

Loss Component	Main Network (%)	Refiner Module (%)
rate	18.62	-
ℓ_2	31.88	-
ℓ_1	32.80	-
lpips	11.29	56.14
wasserstein	1.56	11.62
adversarial	3.84	32.24

D. Absence of Explicit GAN Regularization

To stabilize adversarial training, previous methods often incorporate explicit regularization techniques into the loss function or the network architecture. Common approaches include Spectral Normalization [13] and Wasserstein GAN with Gradient Penalty (WGAN-GP) [14].

Spectral Normalization stabilizes training by constraining the Lipschitz constant of the discriminator. It achieves this by normalizing the weight matrices in the discriminator using their largest singular value, effectively controlling the gradient flow and preventing the explosive growth that leads to mode collapse.

WGAN-GP explicitly enforces the Lipschitz constraint required by the Wasserstein distance by adding a gradient penalty term to the loss function.

Although these techniques are widely adopted, explicitly constraining the discriminator's Lipschitz constant may limit its learning capacity and expressive power. An over-constrained discriminator can fail to provide sufficiently informative gradients to the generator, potentially leading to suboptimal convergence and a loss of fine-grained texture details.

In our framework, we deliberately forgo these explicit regularization techniques. We posit that the proposed semi-decoupled optimization strategy inherently ensures training stability. The refiner module, dedicated to perceptual enhancement, receives strong and clear gradients from the perceptual loss, while the main network is shielded from their potentially disruptive effects by the small coefficient. This architectural separation of concerns, combined with the properties of second-order optimizers, creates a stable training environment without the need for additional, costly constraints on the discriminator. Consequently, our adversarial component is free to learn with greater flexibility, ultimately contributing to higher visual realism without the common pitfalls of adversarial training instability.

E. Adversarial Loss

Fig. 2 presents the training dynamics of both the generator and discriminator adversarial losses, smoothed using a sliding window for clarity. The curves reveal that the adversarial training process remains highly stable throughout optimization, with both losses converging to values consistent with theoretical equilibrium.

The adversarial objective is formulated as a two-player minimax problem. Let G denote the generator and D the

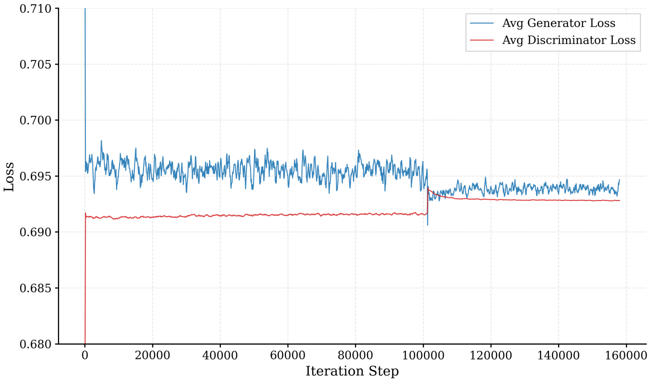

Fig. 2. Training dynamics of adversarial losses. The sudden loss shift due to training dataset switching.

discriminator. Given real images $\mathbf{x} \sim p_{\text{data}}$ and generated images $\hat{\mathbf{x}} \sim p_g$, the adversarial losses are defined as:

$$\mathcal{L}_D = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}\big[-\log D(\mathbf{x})\big] + \frac{1}{2}\mathbb{E}_{\hat{\mathbf{x}}\sim p_g}\big[-\log(1-D(\hat{\mathbf{x}}))\big] \tag{7}$$

$$\mathcal{L}_G = \mathbb{E}_{\hat{\mathbf{x}}\sim p_g}\big[-\log D(\hat{\mathbf{x}})\big] \tag{8}$$

At the Nash equilibrium, the optimal discriminator satisfies $D(\mathbf{x}) = D(\hat{\mathbf{x}}) = 0.5$ for all inputs. Substituting into the loss functions yields:

$$\mathcal{L}_D^* = \frac{1}{2}(-\log 0.5) + \frac{1}{2}(-\log 0.5) \approx 0.693 \tag{9}$$

$$\mathcal{L}_G^* = -\log 0.5 \approx 0.693 \tag{10}$$

As shown in Fig. 2, both the discriminator and generator losses stabilize near 0.693, with only minor fluctuations. This behavior confirms that neither player dominates the other, and the training has settled into a stable adversarial equilibrium within the optimal regime.

F. Variable Rate Framework

To align with the target bitrates of the CLIC2025 challenge, we implement the variable rate framework proposed in [15].

III. Experiments

A. Training settings

We train our models using the OpenImages dataset [16] and images sourced from Pexels [19]. To align with the target bitrates specified by the challenge, we train three separate models, each optimized using the AdamW optimizer. All models are trained with an initial learning rate of 1×10^{-4}, which is reduced to 1×10^{-5} for final finetuning. We employ a progressive training strategy where the patch size starts at 256×256 and increases to 512×512 for final finetuning. Training converged after approximately 600,000 to 800,000 iterations, depending on the target bitrate. Table II summarizes the detailed training configurations for all models.

TABLE II
Training configurations

Parameter	**0.075 bpp**	**0.15 bpp**	**0.30 bpp**
λ_{ℓ_1}	3.0	3.6	4.0
λ_{ℓ_2}	40	50	50
λ_{lpips}	0.55	0.24	0.24
$\lambda_{wasserstein}$	0.03	0.012	0.008
$\lambda_{adversarial}$	0.5	0.5	0.32
λ_{rate}	[0.55,6.40]	[0.135,1.70]	[0.055,0.63]

B. Quantitative Results

We evaluate the proposed method at the target bitrates using four quality metrics: PSNR, MS-SSIM [18], LPIPS [11], and FID [17]. It should be noted that our PSNR is computed by averaging the per-image PSNR values over the CLIC2025 test set (30 images). This differs from the official CLIC leaderboard practice, which computes PSNR by first averaging the pixel-level MSE across the entire dataset.

TABLE III
Quantitative evaluation across different bitrates

Bitrate (bpp)	**PSNR ↑**	**MS-SSIM ↑**	**LPIPS ↓**	**FID ↓**
0.075	27.07	0.914	0.206	39.88
0.150	29.47	0.952	0.163	27.56
0.300	32.78	0.977	0.114	23.17

C. Visualization

This section presents visual comparisons between our method and HiFiC [7] at low bitrates. Results are shown in Fig. 3 and Fig. 4.

IV. Conclusion

In this whitepaper, we presented a semi-decoupled optimization framework for learned image compression that effectively balances distortion and perceptual quality. Our key innovation lies in the architectural separation of roles: a main network dedicated to pixel-level fidelity and a refiner module focused on perceptual enhancement. Crucially, by leveraging second-order optimizers and strategically weighting loss components, we achieve stable training without relying on explicit GAN regularization techniques.

References

[1] Bross B, Wang Y K, Ye Y, et al. Overview of the versatile video coding (VVC) standard and its applications[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(10): 3736-3764.

[2] Jiang W, Yang J, Zhai Y, et al. MLIC++: Linear Complexity Multi-Reference Entropy Modeling for Learned Image Compression[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2025, 21(5): 1-25.

[3] He D, Yang Z, Peng W, et al. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 5718-5727.

[4] Zou R, Song C, Zhang Z. The devil is in the details: Window-based attention for image compression[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 17492-17501.

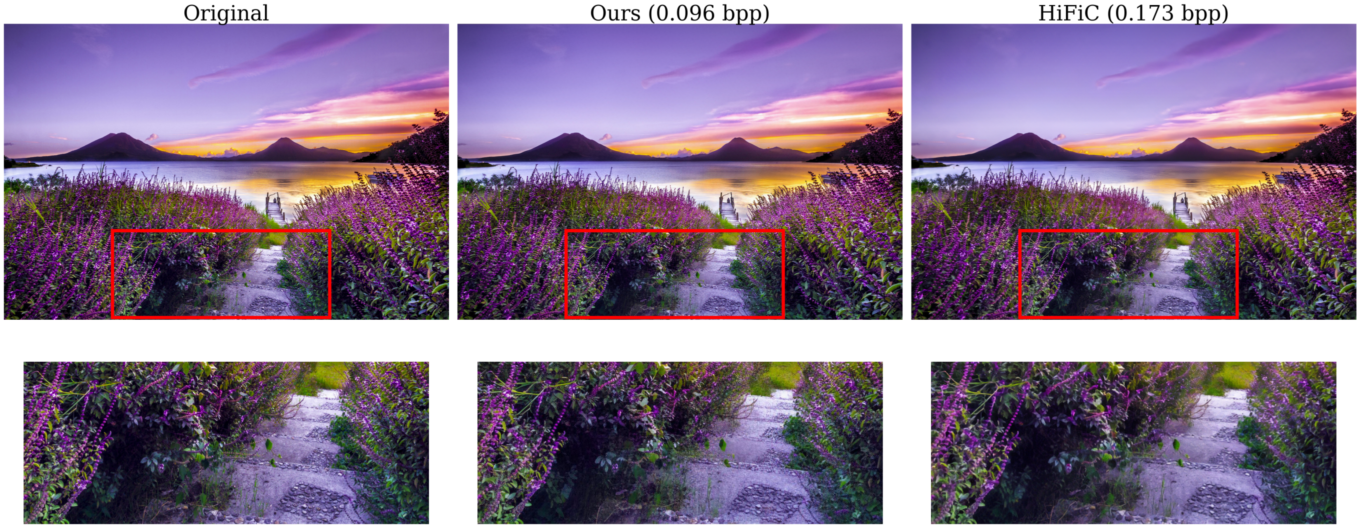

Fig. 3. Visual comparison between our method and HiFiC at low bitrate.

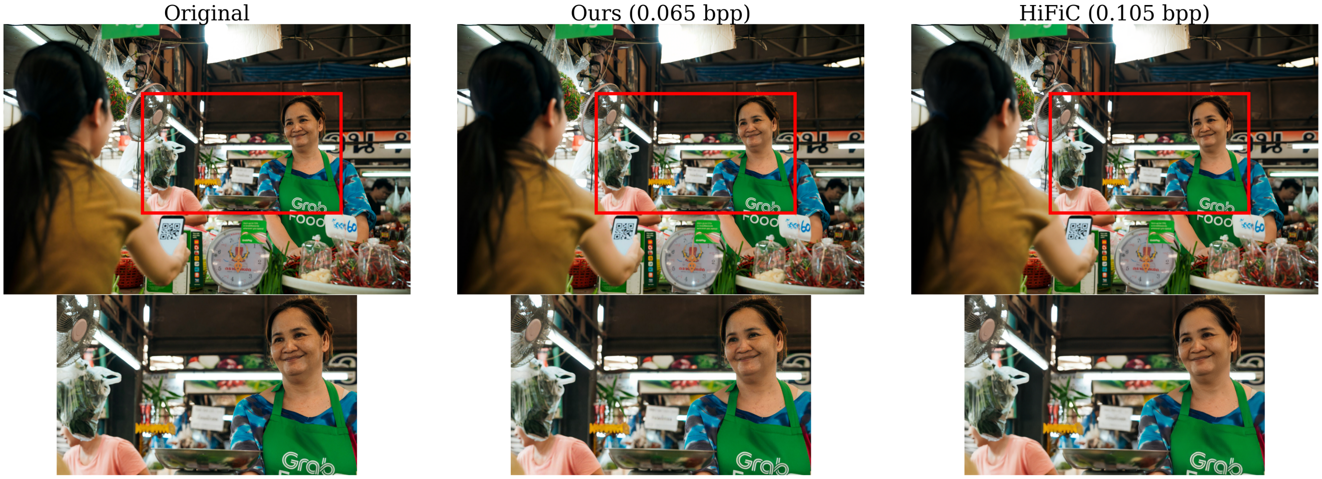

Fig. 4. Another visual comparison at low bitrate.

[5] Liu J, Sun H, Katto J. Learned image compression with mixed transformer-cnn architectures[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 14388-14397.

[6] Feng D, Cheng Z, Wang S, et al. Linear Attention Modeling for Learned Image Compression[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 7623-7632.

[7] Mentzer F, Toderici G D, Tschannen M, et al. High-fidelity generative image compression[J]. Advances in neural information processing systems, 2020, 33: 11913-11924.

[8] He D, Yang Z, Yu H, et al. Po-elic: Perception-oriented efficient learned image coding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 1764-1769.

[9] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[10] Blau Y, Michaeli T. Rethinking lossy compression: The rate-distortion-perception tradeoff[C]//International Conference on Machine Learning. PMLR, 2019: 675-685.

[11] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 586-595.

[12] Ballé J, Versari L, Dupont E, et al. Good, cheap, and fast: Overfitted image compression with Wasserstein distortion[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 23259-23268.

[13] Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks[J]. arXiv preprint arXiv:1802.05957, 2018.

[14] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans[J]. Advances in neural information processing systems, 2017, 30.

[15] Tong K, Wu Y, Li Y, et al. Qvrf: A quantization-error-aware variable rate framework for learned image compression[C]//2023 IEEE International Conference on Image Processing (ICIP). IEEE, 2023: 1310-1314.

[16] Krasin I, Duerig T, Alldrin N, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification[J]. Dataset available from https://github. com/openimages, 2017, 2(3): 18.

[17] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems, 2017, 30.

[18] Wang Z, Simoncelli E P, Bovik A C. Multiscale structural similarity for image quality assessment[C]//The thrity-seventh asilomar conference on signals, systems computers, 2003. Ieee, 2003, 2: 1398-1402.

[19] https://www.pexels.com/