High-Fidelity Generative Image Compression Using Conditional Decoder

Haeyoon Yang, Nam Ik Cho (Team ISPL_IC)

Dept. of Electrical & Computer Eng., INMC,
Seoul National University

1 Abstract

This paper describes the method submitted by team ISPL_IC for the Challenge on Learned Image Compression 2025 (CLIC 2025). We propose a generative image compression approach that combines a conditional discriminator and high-frequency-aware objective to improve both perceptual quality and fidelity. The method is implemented using an ELIC generator and a conditional discriminator, which are trained with rate-distortion, adversarial, high-frequency, and perceptual loss terms. The architecture, training procedure, and datasets are illustrated to facilitate reproducibility.

2 Introduction

In the field of lossy image compression, deep learning–based approaches have achieved remarkable performance, surpassing traditional hand-crafted codecs such as JPEG and VVC. The goal of lossy compression is to encode images into compact bitstreams while reconstructing them into images with minimal distortion. Although there exists a trade-off between bitrates and reconstruction quality, the field has evolved to improve both efficiency and fidelity.

One of the research directions is to use generative models to generate information lost in the compression process [5, 6, 10, 7]. Generative image compression methods, which use GANs or diffusion models, are known for producing perceptually realistic reconstruction images. However, GAN-based methods often generate images that deviate from the actual content of the original images, resulting in poor fidelity and semantic mismatches.

To address this limitation, we introduce a conditional discriminator that uses high-frequency features of the original images as conditioning information. Additionally, we employ a high-frequency–aware loss to preserve fine details. This combination ensures that reconstructed images are both perceptually convincing and faithful to the input content.

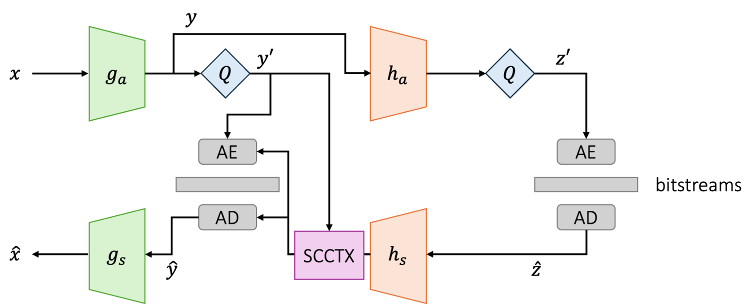

Figure 1: Structure of the generator (ELIC [1])

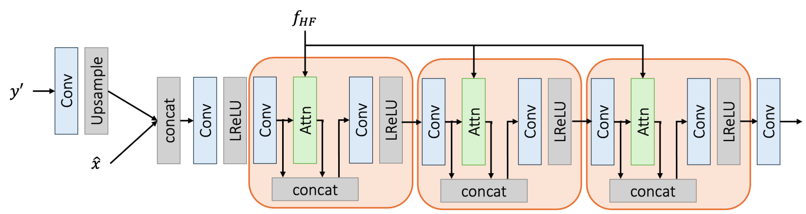

Figure 2: Structure of the discriminator

3 Proposed Method

3.1 Overview

The compression framework consists of a generator and a discriminator. The generator follows the ELIC architecture [1], which includes a space-channel context model (SCCTX) for entropy modeling. Fig 1 illustrates the generator structure.

3.2 Conditional Discriminator

Fig 2 shows the structure of the conditional discriminator. The discriminator is based on the HiFiC architecture [5], which takes the reconstructed image $\hat{x}$ and latent representation y' as input. To improve fidelity, we introduce cross-attention between the reconstructed image and the features of the original image, following ideas from SeD [3]. Unlike SeD, which uses CLIP embeddings, we extract high-frequency features, f_{HF}, using a high-frequency network (HFNet), illustrated in Fig 3. HFNet is trained separately to predict high-frequency components of the images.

3.3 High-frequency Objective

In addition to the use of high-frequency features in the proposed conditional discriminator, we incorporate a high-frequency loss to further generate image-

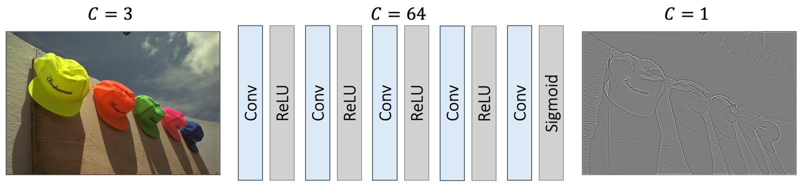

Figure 3: Structure of the HF network

related fine details. The high-frequency loss is defined as

$$L_{hf} = \|\text{HFNet}(x), \text{HFNet}(\hat{x})\|_1, \quad (1)$$

where HFNet$(\cdot)$ denotes the high-frequency network output. Since HFNet features are used in the conditional discriminator, we also employ HFNet for the pixel-level high-frequency loss to ensure consistency. Other methods capable of extracting high-frequency components could be used in place of HFNet.

3.4 Training Objective

The network is trained with four loss terms:

- Rate-distortion loss: $L_{rd} = R + \lambda \cdot D$, where R is the bitrates of y' and z' and D is the distortion between x and $\hat{x}$.
- Adversarial loss: L_{gan} from the conditional discriminator.
- High-frequency loss: L_{hf} as defined above.
- Perceptual loss: L_{lpips}, where LPIPS [11] is computed using VGG features [8] of x and $\hat{x}$.

The total loss is:

$$L = L_{rd} + \alpha \cdot L_{gan} + \beta \cdot L_{hf} + \gamma \cdot L_{lpips}, \quad (2)$$

with α, β, γ controlling the contribution of each term.

4 Experiments

4.1 High-frequency Network Training

HFNet consists of 5 convolutional layers with 64 channels, ReLU activations between the convolutional layers, and a final sigmoid layer. The ground truth images for HFNet are computed as:

$$I_{hf} = I'_{gt} - LPF(I'_{gt}), \quad (3)$$

Bitrate (bpp)	PSNR (dB)	MS-SSIM	decoding time (s)
0.075	24.745	0.910	70
0.150	25.251	0.942	69
0.300	28.593	0.972	69

Table 1: Testset results

where $LPF(\cdot)$ is a low-pass filter implemented with Gaussian blur function (kernel size 11, $\sigma = 5.0$). Here, I'_{gt} refers to one-channel gray scale ground truth image. HFNet is trained on the DIV2K dataset [9] using random crops of size 256×256 and data augmentation (flips and rotations) for 500 epochs with Adam optimizer [2] with learning rate 10^{-4}.

4.2 Generative Compression Network Training

We use 320 channels for y and 192 channels for z features. The channels of y are divided into 16, 16, 32, 64, and 192 in SCCTX (space-channel context model) for entropy modeling and checkerboard context model is used to reduce the decoding time. For the loss function, α, β, and γ are set to 0.01, 1.0, and 1.0, respectively. We used two splits of Open Images dataset [4] for training and the images are randomly flipped, resized, and cropped to size 256×256. The high-frequency feature used in the discriminator f_{HF} is extracted from the output of the second-to-last convolutional layer followed by the ReLU activation in the HF-Net. Both generator and discriminator networks are optimized with Adam [2] optimizers with learning rate 10^{-4}. We used $\lambda = 8\times10^{-5}, 10\times10^{-5}, 80\times10^{-5}$ for target bitrates 0.075bpp, 0.15bpp, and 0.3bpp, respectively. LPIPS loss L_{lpips} is omitted for the lowest bitrate model (0.075bpp) to meet the target bitrate.

4.3 Testset Refinement and Results

For the challenge testset, decoders trained on the validation set generated bit-streams exceeding the target bitrates. To correct this, we applied a low-pass filter (Gaussian blur filter) to the testset images before encoding. Table 1 lists the resulting PSNR and MS-SSIM metrics on the testset.

5 Conclusion

This paper provides a detailed description of the ISPL_IC method for CLIC 2025. By combining conditional discriminators and high-frequency-aware loss, our approach achieves perceptually realistic and content-faithful reconstructions. All network architectures, training procedures, and dataset usage have been described to allow reproducibility by third parties.

References

[1] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5718–5727, 2022.

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

[3] Bingchen Li, Xin Li, Hanxin Zhu, Yeying Jin, Ruoyu Feng, Zhizheng Zhang, and Zhibo Chen. Sed: Semantic-aware discriminator for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25784–25795, 2024.

[4] Google LLC. Open images dataset. https://storage.googleapis.com/openimages/web/index.html, 2022.

[5] 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.

[6] Matthew J Muckley, Alaaeldin El-Nouby, Karen Ullrich, Hervé Jégou, and Jakob Verbeek. Improving statistical fidelity for neural image compression with implicit local likelihood models. In *International Conference on Machine Learning*, pages 25426–25443. PMLR, 2023.

[7] Lucas Relic, Roberto Azevedo, Markus Gross, and Christopher Schroers. Lossy image compression with foundation diffusion models. In *European Conference on Computer Vision*, pages 303–319. Springer, 2024.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[9] Radu Timofte, Shuhang Gu, Jiqing Wu, Luc Van Gool, Lei Zhang, Ming-Hsuan Yang, Muhammad Haris, et al. Ntire 2018 challenge on single image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[10] Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. *Advances in Neural Information Processing Systems*, 36:64971–64995, 2023.

[11] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.