An efficient and perceptiual-oriented image compression method

Jun Jiang (junjiang02@126.com)

September 16, 2025

Abstract

The goal of our model is to compress a set of images to specified bitrates while preserving high perceptual quality, particularly at low bitrates. To this end, we employ a simple yet efficient mean-scale-hyperprior model, which effectively captures spatial dependencies in the latent representation through a hyperprior structure. We finetune the model using a perceptual quality-oriented loss function to enhance visual fidelity. To accommodate a range of target bitrates, we train four distinct models corresponding to different rate-distortion trade-off parameters, and dynamically select the most appropriate model for each input image based on its complexity and target bitrate.

1 Introduction

Learned image compression has advanced rapidly in recent years, demonstrating competitive or even superior performance compared to traditional codecs such as H.264 [WSBL03], H.265 [SOHW12] and H.266 [BWY+21].

In response to the image track of the 7th Challenge on Learned Image Compression (CLIC), we adopt a simple yet highly effective architecture based on the variational image compression framework with a scale hyperprior [BMS+18]. The model consists of an analysis transform encoder g_a, a synthesis transform decoder g_s, and a hyperprior-based entropy model that utilizes a second-level latent variable z to model the spatially varying scale parameters of the main latent representation y. This structure enables the model to capture spatial dependencies in y through side information, effectively improving entropy coding efficiency.

To enhance perceptual quality—especially under low-bitrate conditions—we employ a composite loss function that combines multiple objectives: mean squared error (MSE), multi-scale structural similarity (MS-SSIM) [WSB03], learned perceptual image patch similarity (LPIPS) [ZIE+18], and bit-rate (BPP) regularization. The overall training objective is formulated as:

$$\mathcal{L} = R + \lambda \cdot (\alpha \cdot \text{MSE} + \beta \cdot \text{MS-SSIM} + \gamma \cdot \text{LPIPS})) \tag{1}$$

where λ controls the rate-distortion trade-off, and α, β, γ balance the contributions of the perceptual components and were set 1, 0.1 and 0.1 separately in our case.

2 Implementation Details

Our model is built upon CompressAI [BRFP20]. We trained the models using the training split of the Vimeo-90k [XCW+19] dataset, with a batch-size of 8. Each input frame was randomly cropped into patches of size 256×256. we finetuned the model for 50 epochs with the learning rate of $1e-5$ to optimize perceptual quality and compression efficiency. All the experiments were conducted on a sinlgle NVIDIA RTX 3090. To address diverse bitrate requirements, we trained four independent models, each configured with a distinct λ value (specifically, $\lambda \in \{128, 256, 1024, 2048\}$). This multi-model setup enables flexible adaptation to a wide range of compression targets. During inference, we select the most suitable model for each input image by considering two key factors: the desired target bitrate and the image's content characteristics. This adaptive selection strategy ensures an optimal balance between compression efficiency (bitrate) and perceptual quality for every image.

References

[BMS+18] 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.

[BRFP20] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. Compressai: a pytorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020.

[BWY+21] 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.

[SOHW12] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012.

[WSB03] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[WSBL03] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the h. 264/avc video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7):560–576, 2003.

[XCW+19] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019.

[ZIE+18] 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.