SwinCoder: A Swin Transformer-based Image Compression Model with Perceptual Optimization For CLIC2025

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

***Abstract*—In this paper, we present a low-complexity image compression model based on the Swin Transformer, which only requires 100kMACs/pixel. We introduce a perceptual optimization strategy by incorporating adversarial training. Experimental results demonstrate that our model improves the perceptual quality of compressed images. This paper is a solution of CLIC2025 challenge and our team name are *TestC* and *RunRun* for GPU and CPU tracks, respectively.**

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

I. Introduction

Recent advances in neural image compression have achieved superior performance compared to traditional codecs in terms of PSNR and MS-SSIM. However, these models often suffer from perceptual artifacts like blurring and texture loss, especially at low bitrates, which degrade the visual quality of reconstructed images. In this work, we propose a low-complexity image compression model based on the Swin Transformer [10], which only requires 100kMACs/pixel. We introduce a perceptual optimization strategy by incorporating adversarial training. Experimental results demonstrate that our model improves the perceptual quality of compressed images. This paper is a solution of CLIC2025 challenge and our team name are *TestC* and *RunRun* for GPU and CPU tracks, respectively. *Please note that we adopt the same variable model for GPU and CPU tracks.*

II. Method

A. Architecture

The Swin Transformer [10] has recently demonstrated state-of-the-art (SOTA) performance in various vision tasks, such as image classification and object detection, due to its hierarchical architecture and efficient self-attention mechanism. Inspired by its success, several studies [8], [9], [15], [16] have explored the application of Swin Transformer in the field of learned image compression. In our work, we build upon the Swin Transformer-based architecture as described in [16], and further tailor it for efficient and effective image compression.

Our model consists of an encoder-decoder framework, both of which are constructed using Swin Transformer blocks. The encoder progressively downsamples the input image to a compact latent representation. Specifically, the input image is reduced to $1/16$ of its original spatial resolution through four consecutive downsampling stages, each implemented by the analysis transform network. Each stage contains a certain number of Swin Transformer blocks, with the encoder comprising $\{1, 2, 2\}$ blocks at each respective stage. The number of hidden channels in the encoder is set to $\{128, 256, 384\}$, allowing the network to capture increasingly abstract features as the resolution decreases.

The decoder mirrors the encoder structure, reconstructing the image from the latent representation. It upsamples the latent features back to the original resolution through four upsampling stages, each realized by the synthesis transform network. The decoder employs $\{1, 1, 2\}$ Swin Transformer blocks at each stage, with hidden channel dimensions of $\{64, 112, 256\}$, which are carefully chosen to balance reconstruction quality and computational efficiency. Additionally, we increase the number of latent channels to 256 and the number of hyper latent channels to 192, which enhances the model's capacity to represent complex image content and improves the accuracy of entropy modeling.

For entropy modeling, we also leverage the Swin Transformer to capture dependencies within the latent representation. The latent features are divided into four groups along both spatial and channel dimensions to facilitate parallel processing and efficient context modeling. Initially, the latent representation is split into two groups using a checkerboard pattern [5], ensuring that each group contains spatially interleaved information. Each of these groups is then further divided into two sub-groups along the channel dimension, resulting in a total of four groups. To model the dependencies

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

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 AMD EPYC 7R13 CPU.

BPP	PSNR↑	MSSSIM↑	LPIPS↓	DISTS↓	MACs/pixel	Decoding Time (s)
0.075	27.25	0.9169	0.2532	0.0763	100k	26
0.15	29.56	0.9525	0.1847	0.0434	100k	26
0.30	32.28	0.9751	0.1260	0.0233	100k	26

among these groups in a causal manner, we employ a Swin Transformer block with masked self-attention, which prevents information leakage from future (yet-to-be-encoded) groups. The output of this Swin Transformer block is used to predict the probability distribution parameters for each group. We model the probability distribution of each group using a single Gaussian distribution. The mean and variance of the Gaussian distribution are predicted by a Swin Transformer block, which follows a group-wise causal modeling [8].

For variable rate compression, we insert into the encoder and decoder with AdaLN [4] layers to modulate the latent representation. The AdaLN layers are inserted after the last layer of the encoder and the first layer of the decoder. The AdaLN layers are used to modulate the latent representation, which is the same as the one in [4].

B. Perceptual Optimization

Inspired by previous works [1], [7], [11], we introduce a perceptual optimization strategy by incorporating adversarial training.

To enhance the perceptual quality of the compressed images, we draw inspiration from the multi-stage training process of RealESRGAN [13]. This state-of-the-art super-resolution framework leverages a combination of pixel-wise, perceptual, and adversarial losses to achieve high-fidelity results. We adapt this paradigm to our compression model, progressively refining the output from pixel accuracy to perceptual realism.

Initially, the model is trained exclusively with a Mean Squared Error (MSE) loss. This stage focuses on minimizing the raw distortion between the original and compressed images, ensuring accurate reconstruction of low-level details and stabilizing the training process, which provides a solid foundation for subsequent perceptual enhancement. For variable rate compression, we randomly sample λ from the range of $[16, 1024]$.

Subsequently, we introduce perceptual $\mathcal{L}_{\text{per}}$ and style $\mathcal{L}_{\text{style}}$ losses, while retaining the MSE loss $\mathcal{L}_{\text{MSE}}$. This combined objective encourages the model to preserve high-level semantic content and texture information. The loss function for this stage is:

$$\mathcal{L}_{\text{stage2}} = 150 \cdot \mathcal{L}_{\text{MSE}} + 1.0 \cdot \mathcal{L}_{\text{per}} + 0.1 \cdot \mathcal{L}_{\text{style}} \quad (1)$$

The perceptual loss measures the L_2 similarity between feature representations extracted by a pre-trained VGG network [12]. The style loss captures the correlation between VGG feature maps using Gram matrices, promoting the retention of fine-grained textures and natural image statistics.

Finally, we incorporate adversarial training to further refine the perceptual quality and eliminate artifacts such as over-smoothing. We introduce a discriminator network, adopting the architecture and non-saturating GAN loss from RealESRGAN. The generator (our compression model) is trained to produce images that are indistinguishable from real ones, encouraging more realistic and natural-looking outputs. The discriminator architecture is the same as the one in RealESRGAN. The complete loss function for this final stage is:

$$\mathcal{L}_{\text{final}} = 150 \cdot \mathcal{L}_{\text{MSE}} + 1.0 \cdot \mathcal{L}_{\text{per}} + 0.1 \cdot \mathcal{L}_{\text{style}} + 0.01 \cdot \mathcal{L}_{\text{adv}} \quad (2)$$

III. Experiments

A. Experimental Setup

We train our model on randomly cropped 256×256 image patches sampled from the test split of the OpenImages V7 dataset [6]. The patches are augmented with random horizontal flips and random rotations. The AdamW optimizer is used with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Each training stage runs for 2 million iterations.

For evaluation, we utilize the CLIC2025 Test set, which contains 30 high-resolution 2K images. Model performance is measured using a range of metrics, including MSE, MS-SSIM, LPIPS [14] and DISTS [3].

B. Quantitative Results

We assess the effectiveness of our approach on the CLIC2025 Test set. The quantitative results are summarized in Table I.

C. Qualitative Analysis

As illustrated in Fig. 1, Fig. 2, and Fig. 3, our method achieves superior visual quality at comparable bitrates. Compared to MS-ILLM, our approach better preserves fine details, such as leaf textures, plant structures, and eye features, while maintaining high fidelity to the original images.

Fig. 1: Visual comparison of Ground Truth, our method, and VTM [2] on *07b9f93f170a0381836bdf301280a5b80b2c4be6e66f793a3c335dc200fb4e5b.png* from the CLIC2025 Test set. The reconstructed image is generated at 0.075bpp.

IV. Conclusion

In this paper, we present a low-complexity image compression model based on the Swin Transformer, which only requires 100kMACs/pixel. We introduce a perceptual optimization strategy by incorporating adversarial training. Experimental results demonstrate that our model improves the perceptual quality of compressed images. This paper is a solution of CLIC2025 challenge and our team name are *TestC* and *RunRun* for GPU and CPU tracks, respectively.

References

[1] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. V. Gool, "Generative adversarial networks for extreme learned image compression," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 221–231.

[2] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.

[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. Duan, M. Lu, J. Ma, Y. Huang, Z. Ma, and F. Zhu, "Qarv: Quantization-aware resnet vae for lossy image compression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[5] D. He, Y. Zheng, B. Sun, Y. Wang, and H. Qin, "Checkerboard context model for efficient learned image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 771–14 780.

[6] 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.

[7] 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.

[8] 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.

[9] H. Li, S. Li, W. Dai, C. Li, J. Zou, and H. Xiong, "Frequency-aware transformer for learned image compression," in *The Twelfth International Conference on Learning Representations*, 2024.

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[11] F. Mentzer, G. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," *ArXiv*, vol. abs/2006.09965, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219721015

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *International Conference on Computer Vision Workshops (ICCVW)*.

[14] 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.

[15] Y. Zhu, Y. Yang, and T. Cohen, "Transformer-based transform coding," in *International Conference on Learning Representations*,

Fig. 2: Visual comparison of Ground Truth, our method, and VTM [2] on *0369d229ba4c9965d5caeb38c359a027a810968eee930b81520b604e76b4df14.png* from the CLIC2025 Test set. The reconstructed image is generated at 0.075bpp.

2022.

[16] R. Zou, C. Song, and Z. Zhang, "The devil is in the details: Window-based attention for image compression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 492–17 501.

Fig. 3: Visual comparison of Ground Truth, our method, and VTM [2] on *608cb09e6ffc66d4bc838d4088bf4c0ab889d7e83d4f5d78805cbc4497e432a1.png* from the CLIC2025 Test set. The reconstructed image is generated at 0.075bpp.