

Learned Image Compression with Channel-wise Autoregressive Entropy and Context Modelling

Sofia Iliopoulou

Department of Electrical and Computer
Engineering, University of Patras
Patras, Greece
sofia_iliopoulou@ac.upatras.gr

Dimitris Ampeliotis

Department of Digital Media and
Communication, Ionian University
Argostoli, Greece
ampeliotis@ionio.gr

Athanasios Skodras

Department of Electrical and Computer
Engineering, University of Patras,
Patras, Greece
skodras@upatras.gr

Abstract—In learning-based image compression methods, entropy modelling remains crucial for the achievement of higher compression performance with low computational cost. In this study, we introduce a framework for learned image compression that combines channel-wise autoregressive entropy modelling, LSTM-based context adaptation, and latent residual prediction (LRP) to improve entropy estimation and reduce reconstruction errors. The model employs channel-wise conditioning to progressively refine latent distributions, while an autoregressive prior captures spatial dependencies, allowing for more accurate probability predictions. Furthermore, latent residual prediction reduces quantization errors, resulting in more accurate reconstructions. Experimental results indicate that the proposed method surpasses other deep learning-based compression techniques, achieving a 12.7% BD-rate (Bjontegaard Delta-Rate) reduction compared to a baseline model that utilises channel-conditioning and LRP. These findings highlight the effectiveness of combining hierarchical entropy modelling with spatial and channel-wise context adaptation to achieve state-of-the-art compression efficiency.

Keywords—*image compression, autoregressive modelling, channel conditioning, entropy modelling, context modelling, latent residual prediction*

I. INTRODUCTION

Image compression is a fundamental area of signal and image processing that has been extensively studied due to its significant role in efficient image storage and transmission. Over the past decades, several compression standards have been developed, including JPEG [1], JPEG2000 [2,3], JPEG XL [4], and, more recently, JPEG AI [5], which leverages deep learning techniques for enhanced compression performance. Learned image compression methods have increasingly surpassed traditional approaches in both objective and subjective evaluation metrics. This shift reflects a broader trend toward end-to-end optimization.

During the initial stages of deep learning-based image compression development, most approaches adhered to the principles of traditional codecs, which involved a sequence of transformation, quantisation, and entropy coding. These methods typically substituted one of these stages with a deep learning-based element. In recent years, learning-based techniques have primarily evolved into two main categories.

One focuses on entropy modelling to minimise redundancies found within images, while the other utilises transformers and attention mechanisms to improve the efficiency of feature representation. Both approaches have demonstrated promising compression performance, although their effectiveness is optimised for different aspects of the compression process.

Estimating the code rate through entropy modelling is essential in image compression methods that rely on learning techniques. According to Shannon’s source coding theorem [6], given a discrete memoryless source that generates symbols from the set $y = \{y_0, \dots, y_N\}$, the optimal codelength for the representation of this source is given by

$$C = E_y [-\log_2 P(y_i)] = - \sum_{i=0}^N [P(y_i) \cdot \log_2 P(y_i)] \quad (1)$$

where E_y denotes the expected value over the discrete random variable y and $P(y_i)$ is the probability of symbol y_i [7]. Therefore, in the case of autoencoders it is essential to accurately estimate the probability density functions (PDFs) of the bottlenecks in order to compute the compression rate [8].

State-of-the-art models employ both forward and backward-adaptive components to enhance the predictive accuracy of the entropy model, thereby achieving higher compression rates without increasing distortion. Forward adaptation [9] typically utilises side information, such as a learned hyperprior [10]. The hyperprior approach is widely adopted, as it can be easily integrated into an end-to-end optimized network while enabling efficient encoding and decoding. In contrast, backward adaptation generally leverages predictions derived from the causal context of each symbol, including neighbouring symbols located above and to the left of the current symbol, as well as those in previously decoded channels [11,12].

In this study, an autoregressive approach with context modelling and latent residual prediction (LRP) is used to enhance compression performance. The key components of the proposed model include: (a) autoregressive (AR) modelling for sequential latent representation prediction, (b) LSTM-based context adaptation [13] to improve entropy estimation, (c) channel conditioning (CC) for better latent distribution estimation, (d) hyperprior-based entropy modelling to capture global dependencies and (e) latent residual prediction to minimise reconstruction errors.

The remainder of this paper is structured as follows: Section II provides a review of deep learning-based image compression methods. Section III presents the details of the proposed technique and the network architecture employed in this study. The experimental setup and evaluation results are discussed in Section IV. Finally, Section V summarizes the findings and outlines directions for future research.

II. RELATED WORK

Numerous end-to-end techniques for image compression utilise a variational autoencoder (VAE), a widely used probabilistic generative model that is often combined with an approximate inference framework [14]. Ballé et al. included a hyperprior model [10] to communicate the distribution of latent representations and assumed a zero-mean Gaussian distribution for each code [7]. Minnen et al. have built on this architecture and incorporated a context-based autoregressive prior into the hierarchical hyperprior introduced by Ballé [15]. Lee et al. presented an autoregressive entropy model that improved coding efficiency through masked convolutions, which helped in capturing spatial dependencies in the latent space [16].

Minnen et al. introduced a channel-wise autoregressive entropy model that processed latent representations on a per-channel basis rather than on pixel-level [11]. Instead of relying on full spatial autoregression, the model exploited causal dependencies between channels. Although autoregressive models improve compression performance, they introduce sequential dependencies that slow down the decoding process. To overcome these limitations, Guo et al. proposed a three-dimensional context entropy model that, instead of relying exclusively on two-dimensional spatial neighbors, also incorporated information from previously processed channels in the latent space [17]. On the other hand, Fu et al. introduced a framework that integrated a wavelet-domain convolution module into a wavelet-domain channel-wise autoregressive entropy model [18].

Recent advancements in learned image compression have increasingly utilised transformers and attention mechanisms. Liu et al. combined a transformer-CNN mixture block with a channel-wise entropy model incorporating parameter-efficient Swin-Transformer-based attention modules [19]. Zou et al. proposed a window-based local attention block and a symmetrical transformer framework that incorporated absolute transformer blocks in both the downsampling encoder and upsampling decoder [20]. Finally, Yang et al. combined a neighborhood window attention mechanism that improved global modelling with an enhanced SwinT transformer and a CNN block that improved the transformation capabilities of both the main autoencoder and the hyper autoencoder [21].

III. PROPOSED IMAGE COMPRESSION METHOD

A. Channel-Conditional Entropy Model

Our model builds on the architecture introduced in [11]. This framework leverages a hyperprior, a channel-wise autoregressive model, a long short-term memory (LSTM)-based spatial context model and latent residual prediction to predict latent distributions. Fig. 1 provides a high-level

overview of this architecture. Given an image x , the objective is to find a compressed representation y that minimises the rate-distortion cost:

$$L=R+\lambda\cdot D \quad (2)$$

where R is the expected bit rate, D is the distortion (e.g., the mean squared error - MSE), and λ controls the trade-off [22]. The proposed entropy model is hierarchical, first using a global prior via the hyperprior and then progressively refining the probability estimates using local dependencies via the channel-conditional and autoregressive modelling. The hyperprior architecture is similar to the one presented in [10].

Instead of modelling the entire latent space y directly, the channel-conditional entropy model decomposes y into M non-overlapping slices along the channel dimension. Each slice is coded sequentially and is conditioned on the mean and scale parameters derived by the hyperprior, as well as all the previously decoded slices. For each slice, channel-wise conditional mean and scale estimates are refined using convolutional transforms. However, images exhibit strong correlations across neighbouring pixels too, so the model is also autoregressive along the spatial dimension to minimise these spatial dependencies. The LSTM-based context model is designed to capture such dependencies. It processes previously encoded slices sequentially and updates a hidden state h_m , that is used to correct the scale parameter:

$$\sigma'_m=\sigma_m+\alpha\cdot h_m \quad (3)$$

where σ_m and σ'_m are the original and LSTM-enhanced scales respectively and α is a learnable weight.

B. Latent Residual Prediction and Quantisation

Neural image compression models rely on transforming images into a latent representation that can be entropy-coded efficiently. Like many other studies, in the proposed model a soft-to-hard quantisation approach is taken. During training, uniform noise $U(-\frac{1}{2}, \frac{1}{2})$ is added to the latents to simulate quantisation and avoid vanishing gradients [23]. On the other hand, during inference, rounding is applied.

Due to the quantisation process that takes place before the lossless entropy coding, some information loss occurs, which may lead to increased distortion in the reconstructed image. Latent residual prediction aims to reduce the quantisation error by predicting and correcting the residual error that occurs between the original latent representation and its quantised version. Previously decoded latents are used to improve the reconstruction process [24]. For each slice, information from both the hyperprior and the autoregressive conditioning is used for a more accurate estimation of the residual error.

The predicted residual is scaled and added to the quantised slice:

$$\hat{y}'_m = \hat{y}_m + \lambda_{LRP} \cdot \tanh(r_m) \quad (4)$$

where \hat{y}'_m is the slice after LRP, \hat{y}_m is the quantised slice, λ_{LRP} is a learned scaling factor to control the strength of the correction and r_m is the residual error. The hyperbolic tangent $\tanh(\cdot)$ is used to ensure that the predicted residual remains within a reasonable range. The layers of the compression network are presented in Table I.

TABLE I. NEURAL NETWORK LAYERS

Encoder	Decoder	Hyper Encoder	Hyper Decoder	Slice Transform	Context Model
Conv: 5×5 c192 s2 GDN Conv: 5×5 c192 s2 GDN Conv: 5×5 c192 s2 GDN Conv: 5×5 c1_d s2	Deconv: 5×5 c192 s2 IGDN Deconv: 5×5 c192 s2 IGDN Deconv: 5×5 c192 s2 IGDN Deconv: 5×5 c3 s2	Conv: 3×3 c320 s1 ReLU Conv: 5×5 c256 s2 ReLU Conv: 5×5 c h_d s2	Deconv: 5×5 c192 s2 ReLU Maskconv: 3×3 c192 s1 Deconv: 5×5 c256 s2 ReLU Maskconv: 3×3 c256 s1 Deconv: 3×3 c320 s1	Conv: 5×5 c224 s1 ReLU Conv: 5×5 c128 s1 ReLU Conv: 3×3 c_s d s1	CLSTM: 3×3 c 128 s1 Conv: 3×3 c 1_s s1
Conv: convolutional layer, Deconv: deconvolutional layer, c:channels, s:stride, l_d:latent depth (depends on the slice), h_d:hyperprior depth (depends on the slice), Maskconv: masked convolution, CLSTM: convolutional LSTM, l_s: latent depth/number of slices					

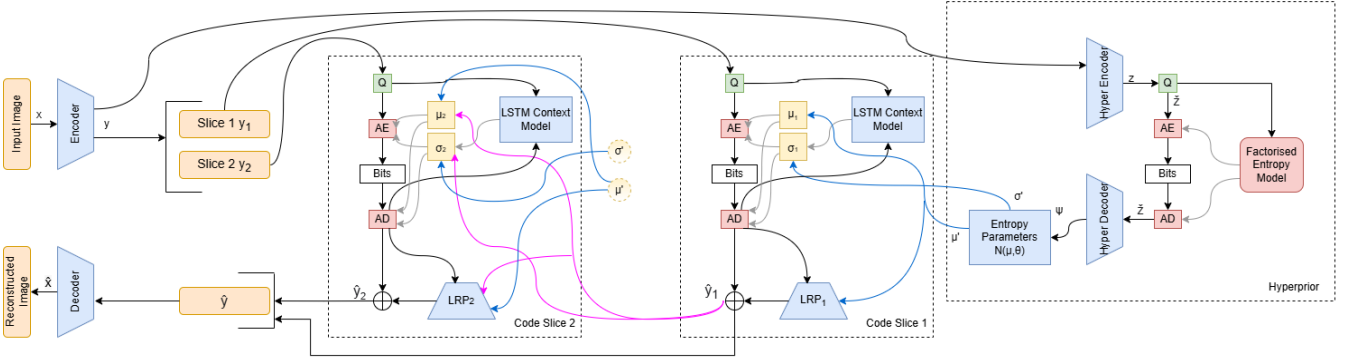


Fig. 1. The architecture of our compression model with a hyperprior, an LSTM context model, latent residual prediction and two slices for channel conditioning. Q denotes quantization and AE/AD is the lossless arithmetic encoder/decoder, μ and σ symbolise the mean and scale parameters.

IV. EXPERIMENTAL SETUP AND RESULTS

In this section, the performance of the proposed model is compared to various learned image compression techniques. The models are trained on the CLIC dataset [25] and evaluated over the Kodak dataset [26]. The number of slices that has been chosen for our model is 10. The models are trained with 256×256 patches that have been randomly cropped from the input images and a batch size of 8. For the experimental results, the model is trained for 400 epochs using the Adam optimizer with a learning rate of 10^{-4} .

The distortion metric that is used for the optimisation is the mean squared error (MSE) and experiments are conducted with $\lambda \in \{0.001, 0.005, 0.01, 0.03, 0.05\}$. The scaling factor for the latent residual prediction is set to 0.6 and the LSTM-based context model is applied with a 0.5

weight to reduce the risk of overcorrection. Fig.2 presents the rate-distortion curves for the evaluation of the proposed model with some popular image compression methods. The distortion metrics that are used in this paper are the PSNR and the MS-SSIM.

To evaluate the contribution of each component of the proposed framework, we conducted an ablation study. Specifically, we compared the baseline hyperprior model [10] against variants incorporating channel conditioning (CC), autoregressive modelling (AR), LSTM-based context modelling (LSTM), latent residual prediction (LRP), as well as their combination (CC+AR+LSTM+LRP). The models were trained and evaluated under identical conditions to ensure a fair comparison. The results, summarised in Fig. 3a, show that each component individually improves the

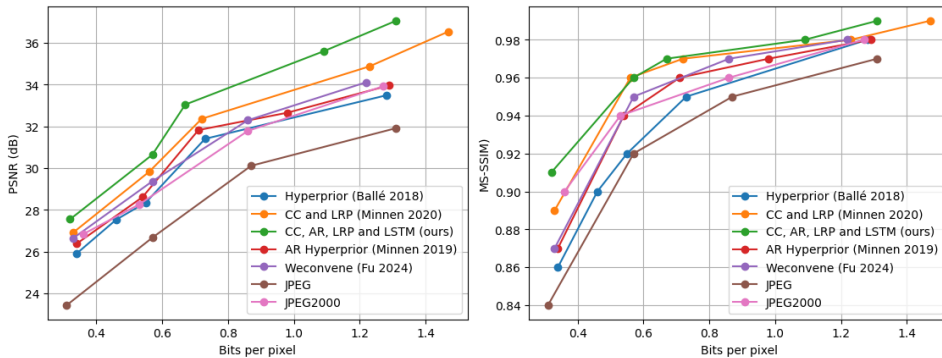


Fig. 2. Model evaluation over the Kodak dataset

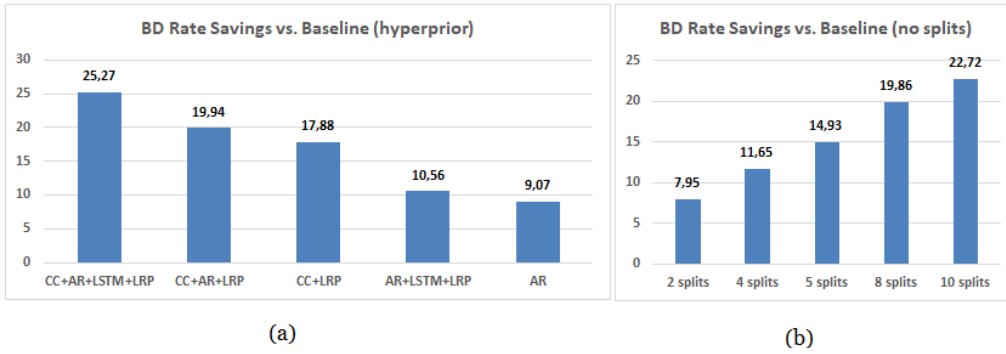


Fig. 3. The BD Rate savings of (a) different components of our framework compared to a scaled hyperprior baseline [10] and (b) models that are identical except for the number of channel splits.

compression performance compared to the hyperprior baseline [10], and also highlight how complementary these modules are, with their combined use having a significantly better performance than any component in isolation..

We further investigated the effect of the number of channel slices on the performance of the proposed model. Since channel conditioning operates on sequential slices, the level of slicing affects both the efficiency of the entropy model and the computational complexity. We evaluated models trained with different numbers of slices (2, 4, 5, 8 and 10) while keeping all other parameters unchanged. As shown in Fig. 3b, increasing the number of slices generally improves the rate-distortion performance, since it allows for more accurate inter-slice conditioning. However, beyond a certain point – in this case 10 slices –, the gain significantly decreases while the computational complexity and inference time increase, suggesting a trade-off between compression performance and computational efficiency. For this reason, a slice count of 10 was selected for the final architecture to balance coding efficiency with inference cost.

V. CONCLUSIONS

This work presents a new method of learning-based image compression that is based on entropy modelling. The incorporation of autoregression combined with LSTM-based context modelling and latent residual prediction enhances the model’s performance compared to similar learned compression approaches. The channel conditioning captures inter-channel dependencies by sequentially encoding latent slices conditioned on previously decoded slices and the autoregressive prior captures spatial correlations in the latent representation by conditioning the probability distribution of each latent element on previously decoded elements. Thus, both the entropy estimation and the bit allocation are improved. Furthermore, latent residual prediction mitigates quantization errors, resulting in superior rate-distortion performance without increasing computational complexity. Fig. 4 presents visual comparisons between the proposed method and other compression techniques, highlighting the superior reconstruction quality achieved by our approach.

The next step in our research is the refinement of this model. Studies have shown that adaptive slice sizes and number of slices can improve the overall performance of such architectures. While LSTMs are used in the proposed architecture, the inference time does not differ from that of Minnen et al. [9] which only applies channel conditioning and latent residual prediction. Despite that, we aim to replace the LSTM context model with a more lightweight alternative, making this architecture more energy compact and computationally efficient. Moreover, we plan to extend the evaluation of our method by benchmarking it against state-of-the-art traditional codecs such as the Versatile Video Coding (VVC) standard to provide a broader comparison with both traditional and learned compression techniques. Finally, since the models that include attention mechanisms display superior rate-distortion results, our goal is to reduce the computational cost of such a technique, while retaining excellent performance. Future iterations may also explore hybrid transformer - CNN modules, which have recently demonstrated strong performance in both entropy modelling and transformation stages.

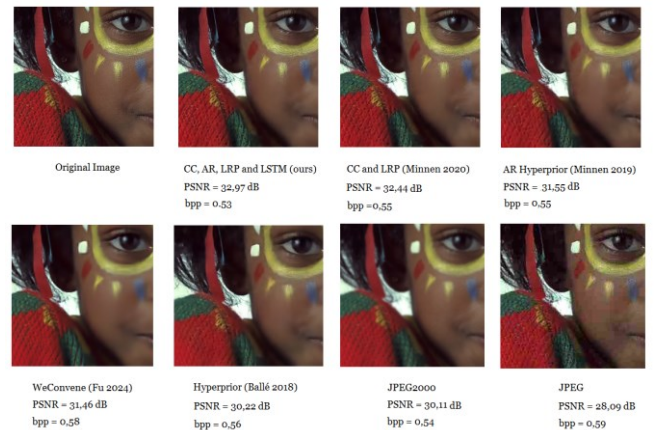


Fig. 4. At similar bit rates, our method provides the highest visual quality on the Kodak dataset.

REFERENCES

- [1] G. K. Wallace, "The jpeg still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, 1992.
- [2] M. Rabbani and R. Joshi, "An overview of the jpeg 2000 still image compression standard," *Signal processing: Image communication*, vol. 17, no. 1, pp. 3–48, 2002.
- [3] A. Skodras, C. Christopoulos and T. Ebrahimi, "The JPEG 2000 still image compression standard," in *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 36–58, Sept. 2001, doi: 10.1109/79.952804.
- [4] J. Alakuijala, R. van Asseldonk, S. Boukourt, M. Bruse, I. M. Comşa, M. Firsching, T. Fischbacher, E. Kliuchnikov, S. Gomez, R. Obryk, K. Potempa, A. Rhatushnyak, J. Sneyers, Z. Szabadka, L. Vandevenne, L. Versari and J. Wassenberg, "JPEG XL next-generation image compression architecture and coding tools," in *Proc. SPIE 11137, Applications of Digital Image Processing XLII*, 111370K, Sept. 2019, doi:10.1117/12.2529237.
- [5] E. Alshina, J. Ascenso and T. Ebrahimi, "JPEG AI: The First International Standard for Image Coding Based on an End-to-End Learning-Based Approach," in *IEEE MultiMedia*, vol. 31, no. 4, pp. 60–69, Oct.–Dec. 2024, doi: 10.1109/MMUL.2024.3485255.
- [6] T.M Cover and J.A. Thomas, "Data Compression" in "Elements of Information Theory," 2nd ed., John Wiley & Sons, pp. 103–142, 2021.
- [7] S. Iliopoulou, P. Tsinganos, D. Ampeliotis, and A. Skodras, "Synthetic Face Discrimination via Learned Image Compression," *Algorithms*, vol. 17, no. 9, 2024, doi: 10.3390/a17090375.
- [8] M. Li, K. Zhang, J. Li, W. Zuo, R. Timofte and D. Zhang, "Learning Context-Based Nonlocal Entropy Modeling for Image Compression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 3, pp. 1132–1145, March 2023, doi: 10.1109/TNNLS.2021.3104974.
- [9] Z. Cui, J. Wang, S. Gao, T. Guo, Y. Feng and B. Bai, "Asymmetric Gained Deep Image Compression With Continuous Rate Adaptation," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 10527–10536, doi: 10.1109/CVPR46437.2021.01039.
- [10] J. Balle, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in 6th Int. Conf. Learn. Rep., Vancouver, BC, Canada, Apr. 2018.
- [11] D. Minnen and S. Singh, "Channel-Wise Autoregressive Entropy Models for Learned Image Compression," 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 2020, pp. 3339–3343, doi: 10.1109/ICIP40778.2020.9190935.
- [12] D. He, Z. Yang, W. Peng, R. Ma, H. Qin and Y. Wang, "ELIC: Efficient Learned Image Compression with Unevenly Grouped Space-Channel Contextual Adaptive Coding," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 5708–5717, doi: 10.1109/CVPR52688.2022.00563.
- [13] N. Johnston, D. Vincent, D. Minnen, M. Covell, S. Singh, T. Chinen, S. J. Hwang, J. Shor and G. Toderici, "Improved Lossy Image Compression with Priming and Spatially Adaptive Bit Rates for Recurrent Networks," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 4385–4393, doi: 10.1109/CVPR.2018.00461.
- [14] G. Toderici, D. Vincent, N. Johnston, S. Hwang, D. Minnen, J. Shor and M. Covell, "Full resolution image compression with recurrent neural networks," 2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, July 2017, pp. 5435–5443.
- [15] D. Minnen, J. Balle, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 10794–10803.
- [16] J. Lee, S.-H. Cho, and S. Beack, "Context-adaptive Entropy Model for End-to-end Optimized Image Compression," *arXiv: Image and Video Processing*, Sep. 2018, [Online]. Available: <https://arxiv.org/abs/1809.10452>.
- [17] Z. Guo, Y. Wu, R. Feng, Z. Zhang and Z. Chen, "3-D Context Entropy Model for Improved Practical Image Compression," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 520–523, doi: 10.1109/CVPRW50498.2020.00066.
- [18] H. Fu, J. Liang, Z. Fang, J. Han, F. Liang, and G. Zhang, "WeConvene: Learned Image Compression with Wavelet-Domain Convolution and Entropy Model," in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part L*, 2024, pp. 37–53. doi: 10.1007/978-3-031-72973-7_3.
- [19] J. Liu, H. Sun and J. Katto, "Learned Image Compression with Mixed Transformer-CNN Architectures," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 14388–14397, doi: 10.1109/CVPR52729.2023.01383.
- [20] R. Zou, C. Song and Z. Zhang, "The Devil Is in the Details: Window-based Attention for Image Compression," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 17471–17480, doi: 10.1109/CVPR52688.2022.01697.
- [21] D. Yang, X. Fan, X. Meng, and D. Zhao, "Accurate entropy modeling in learned image compression with joint enhanced SwinT and CNN," *Multimedia Syst.*, vol. 30, no. 4, Jul. 2024, doi: 10.1007/s00530-024-01405-w.
- [22] S. Iliopoulou, P. Tsinganos, D. Ampeliotis and A. Skodras, "Learned Image Compression with Wavelet Preprocessing for Low Bit Rates," 2023 24th International Conference on Digital Signal Processing (DSP), Rhodes (Rodos), Greece, pp. 1–5, 2023, doi: 10.1109/DSP58604.2023.10167974.
- [23] J. Ballé, V. Laparra and E. P. Simoncelli, "End-to-end optimization of nonlinear transform codes for perceptual quality," 2016 Picture Coding Symposium (PCS), Nuremberg, Germany, pp. 1–5, 2016, doi: 10.1109/PCS.2016.7906310.
- [24] M. Han et al., "Causal Context Adjustment Loss for Learned Image Compression," *Advances in Neural Information Processing Systems*, vol. 37, pp. 133231–133253, 2025.
- [25] "7th Challenge on Learned Image Compression", <https://compression.cc/>
- [26] Eastman Kodak, "Kodak Lossless True Color Image Suite (PhotoCD PCDO992)", URL: <http://r0k.us/graphics/kodak/>, 1993.