Towards Efficient Image Compression Without Autoregressive Models

Muhammad Salman Ali [1], **Yeongwoong Kim**[1], **Maryam Qamar** [1],
Sung-Chang Lim[2], **Donghyun Kim**[2], **Chaoning Zhang**[1], **Sung-Ho Bae**[*1], **Hui Yong Kim**[*1]
[1] Kyung Hee University, Republic of Korea
[2] Electronics and Telecommunications Research Institute (ETRI), Republic of Korea
{salmanali, duddnd7575, maryamqamar}@khu.ac.kr,
{sclim, kimddng}@etri.re.kr,
chaoningzhang1990@gmail.com
{shbae, hykim.v}@khu.ac.kr

Abstract

Recently, learned image compression (LIC) has garnered increasing interest with its rapidly improving performance surpassing conventional codecs. A key ingredient of LIC is a hyperprior-based entropy model, where the underlying joint probability of the latent image features is modeled as a product of Gaussian distributions from each latent element. Since latents from the actual images are not spatially independent, autoregressive (AR) context based entropy models were proposed to handle the discrepancy between the assumed distribution and the actual distribution. Though the AR-based models have proven effective, the computational complexity is significantly increased due to the inherent sequential nature of the algorithm. In this paper, we present a novel alternative to the AR-based approach that can provide a significantly better trade-off between performance and complexity. To minimize the discrepancy, we introduce a correlation loss that forces the latents to be spatially decorrelated and better fitted to the independent probability model. Our correlation loss is proved to act as a general plug-in for the hyperprior (HP) based learned image compression methods. The performance gain from our correlation loss is 'free' in terms of computation complexity for both inference time and decoding time. To our knowledge, our method gives the best trade-off between the complexity and performance: combined with the Checkerboard-CM, it attains **90%** and when combined with ChARM-CM, it attains **98%** of the AR-based BD-Rate gains yet is around **50 times** and **30 times faster** than AR-based methods respectively.

1 Introduction

In the digital world, visual media has proven to be a mixed blessing. On the one hand, it provides a mechanism to disseminate a plethora of data, information, and knowledge in a form easily consumable by humans Silk et al. [2021], while on the other, improved quality and increased content is very quickly pushing the boundaries of data storage, communication, and processing technologies Kim et al. [2021]. A software-driven solution to skew the balance of this state has been extensively presented in the form of lossy compression techniques Wallace [1991], Rabbani and Joshi [2002], Ballé et al. [2017]. In the case of images, these techniques aim to reduce the data size while maintaining its context. Traditional image codecs like JPEG Wallace [1991] and JPEG2000 Rabbani and Joshi [2002] use mathematically well-defined transformations such as discrete cosine transform (DCT) and discrete wavelet transform (DWT). These context-less transformations lead to reduced efficiency in terms of maintaining an optimal balance between the image quality and bitrate of the encoded image. In complex and varied images, this problem is especially pronounced, leading to the inclusion of various artifacts such as ringing and blocking Ballé et al. [2017]. In recent years,

[*] Corresponding Authors

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

deep learning has significantly advanced several computer vision tasks Hu et al. [2022], especially image compression, by incorporating structure and learned behavior. In particular, learned image compression (LIC) has gained popularity as a field of research that has the potential to go beyond traditional compression methods by exploring the inherent nature of images Toderici et al. [2016], Lee et al. [2022], Johnston et al. [2018], Qian et al. [2022], Cheng et al. [2019].

State-of-the-art learned image compression techniques employ transform coding strategies for lossy image compression, which maps image pixels into a quantized latent space Cheng et al. [2020], Zhong et al. [2020]. This map is then losslessly compressed using entropy coding methods such as Huffman or arithmetic coding to create a bitstream for transmission over a channel Shannon [1948]. The entropy coding algorithm requires an entropy model, which is a prior probability model on the quantized latent representation. In this framework, deep neural networks are used for both transform coding and entropy modeling. By learning both modules end-to-end, the ability of deep neural networks is fully utilized to improve the compression performance Yang et al. [2023]. The bitstream size depends on the accuracy of the entropy model, making it essential to design an accurate model for efficient compression Kim et al. [2022]. The objective of the entropy model is to estimate a joint probability distribution over the elements of the quantized latent representation. A simplistic approach is to assume complete independence among the elements, but this approach results in subpar compression performance Ballé et al. [2017, 2018]. Thus, modeling the quantized latent representation interdependencies in learned image compression is a significant challenge Lee et al. [2019], Minnen et al. [2018], Qian et al. [2021], Theis et al. [2017].

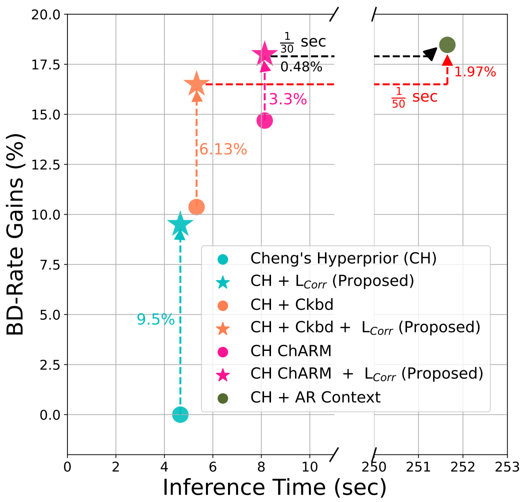


Figure 1: Performance-complexity tradeoff using various entropy models on the top of Cheng's architecture. Incorporating correlation loss to Cheng's hyperprior yields a BD rate gain of 9.5%, comparable to the gains achieved by the Checkerboard (CKBD) method. Applying correlation loss to the CKBD model results in a BD rate gain of 6.13% in comparison to the CKBD baseline, signifying approximately 90% of the gain seen in AR-based models. Notably, this improvement is achieved with a computational expense merely 1/50th that of the full AR-based model. Similarly, incorporating correlation loss into ChARM leads to a BD rate gain of 3.3% over the ChARM baseline, akin to 98% of the gain achieved by a full AR model, while utilizing only 1/30th of the computational resources.

A common approach is to use convolutional neural networks (CNNs) to extract additional features, known as "hyperprior," which captures the context from the quantized latent representation to improve the efficiency of the entropy model Ballé et al. [2018]. The underlying assumption of hyperprior-based entropy models is the probabilistic independence of the latent features, which contradicts the highly correlated nature of latent space, resulting in a discrepancy between the actual distribution and the assumed probability model Kim et al. [2022]. To minimize this discrepancy, Autoregressive context model (AR-CM) based LIC methods have been proposed Minnen et al. [2018], where the probability of each latent is updated based on the previously coded neighboring latents. While this increases the model's overall performance, it introduces sequential dependencies, making it roughly 50× slower than simple hyperprior-based models He et al. [2021] and hinders its practical implementation of neural image compression in real-world scenarios

To mitigate the above issue, a line of works have attempted to discard AR method for improving efficiency. A notable advancement in this domain is the channel-wise autoregressive entropy model (ChARM) introduced by Minnen et al. Minnen and Singh [2020]. ChARM focuses on reducing the element-level serial processing within the context model presented in Minnen et al. [2018], and it demonstrates a reasonably good RD performance compared to a full autoregressive context model. Li et al. introduced CCN Li et al. [2020], which enables faster context calculation with moderate parallelizability. However, the effectiveness of CCN is still constrained by the image size, limiting its overall efficiency. He et al. He et al. [2021] proposed a parallelizable checkerboard context model along with a two-pass decoding method, aiming to achieve a better trade-off between rate distortion (RD) performance and computational efficiency. The efficiency of the above methods is increased at a cost of performance drop compared to AR-based models.

To this end, we introduce a correlation loss to minimize the discrepancy without incurring any sequential dependency. The correlation loss forces latents to be decorrelated spatially, resulting in a better fit with the spatially independent probability model. Our findings indicate that decorrelating the latent aids in enhancing the general effectiveness of the hyperprior entropy model. Figure 1 illustrates the effectiveness of the proposed approach. Applying correlation loss to Cheng's hyperprior model yields a comparable performance improvement to that achieved by integrating efficient context modeling approaches. When our approach is integrated with the Checkerboard method, it leads to substantial BD rate improvements of approximately 90%, and when coupled with ChARM, it exhibits BD rate gains of about 98%, akin to the gains observed in a full AR-CM. Notably, our method achieves these remarkable performance gains while requiring only 1/50th and 1/30th of the inference time required by the AR model, respectively. Our key contributions can be summarized as follows:

- Our work takes the path of decorrelating latents that have never been tried in the field. The previous approaches primarily focus on CMs to enhance their performance, whereas, in our work, we address the fundamental need for employing computationally heavy CMs. With the introduction of our correlation loss, the correlation among spatially-neighbored elements in the latent space is decreased. This would minimize the discrepancy between the hyperprior entropy model's assumed probability distribution and the actual distribution of the latents, which eventually leads to an overall improvement of the rate-distortion (RD) performance.
- Since our method modifies only the loss function, i.e. adding a correlation loss to the original one, it does not require any additional memory or computational complexity beyond the base model. In contrast to the AR-based method which has severe sequential dependency among neighboring latent elements and thus could increase the inference time or decoding time up to 100× the base model, our method does not create any sequential dependency and thus can maintain the inference or decoding time of the base model.
- Our proposed correlation loss acts as a plug-in for the existing learned image compression methods and can achieve BD rate gains of up to 9.5% by combining with the base hyperprior-based models. Remarkably, the gains are comparable to the performance improvements achieved by combining the hyperprior module with efficient context-based LIC models, like the SOTA efficiency-oriented Checkerboard's context model (CM). When combined with Checkerboard method, our proposed approach acheives 90% and when combined with ChARM our proposed approach acheives 98% of the AR-CM's BD-Rate gains while incurring only 1/50th and 1/30th of the inference time of the AR-CM. Our method achieves the best trade-off between the complexity and performance.

2 Related Work

The autoregressive context based entropy model, inspired by the concept of context used in traditional codecs, is used to predict the probability of unknown codes based on latents that have already been decoded Bellard [2015]. These models use hyper latent and context to predict both the location (mean value) and scale parameters (standard deviation) of the entropy model Minnen et al. [2018]. By combining elements of differentiable entropy modeling, hyper latent, and context models, it is possible to outperform BPG Bellard [2015] in terms of PSNR and MS-SSIM Minnen et al. [2018], He et al. [2021], Qian et al. [2022]. These context models have also been extended to become more powerful but also more computationally expensive Cheng et al. [2020]. In particular, AR context-based models, essential for achieving the state of the art performance, have low computational efficiency due to the sequential dependency of decoding the latents. As AR context models are sequentially dependent, they cannot be deployed on edge devices or take advantage of parallel hardware processing. As a result, models focusing on the real-world deployment of learned image compression often exclude AR context models due to their inefficiency and high inference costs He et al. [2021].

Parallel decoding approaches like Checkerboard He et al. [2021] and ChARM Minnen and Singh [2020] offer a better performance tradeoff over computation complexity. Checkerboard introduces a parallel decoding approach by encoding and decoding only half of the latent variables using a checkerboard-shaped context and hyperprior. The remaining latents, referred to as anchors, are coded solely based on the hyperprior. On the other hand, ChARM utilizes channel conditioning instead of

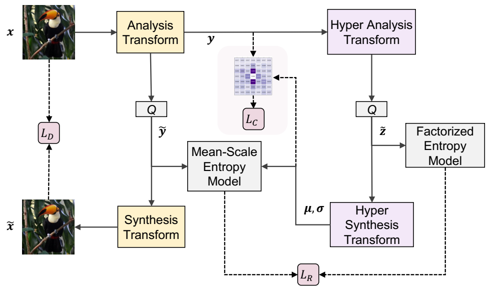


Figure 2: Operational diagram combining Hyperprior Ballé et al. [2018] with our proposed correlation loss

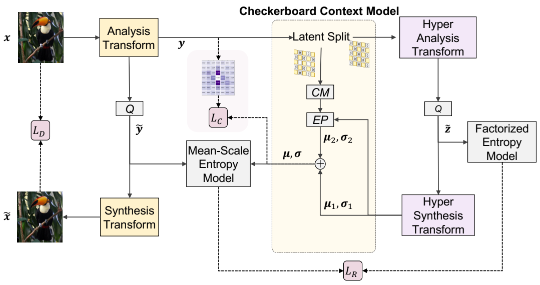


Figure 3: Operational diagram combing Checkerboard method He et al. [2021] with our proposed correlation loss

element-wise conditioning, along with latent residual prediction, to enhance the rate-distortion (RD) performance. The recent efficient SOTA methods are all built on top of the ChARM and checkerboard context model as an alternative to a fully AR-based context model Zhu et al. [2022], Qian et al. [2022], Kim et al. [2022].

Recent research in the field of entropy coding has primarily concentrated on improving the efficiency of context-based methods. These methods aim to exploit the interdependencies within the latent space more effectively, utilizing the inherent dependencies present Lei et al. [2022], Lee et al. [2019]. However, there has been a lack of attention towards enhancing the effectiveness of the hyperprior module, an integral component of prominent learned image compression (LIC) techniques. Notably, recent studies utilizing Transformer Vaswani et al. [2017] models have demonstrated that achieving a higher level of decorrelation in the latent space leads to improved rate-distortion (RD) performance Zhu et al. [2022]. This improvement can be attributed to the reduction in the discrepancy between the actual distribution of the latent space and the assumed probability model Kim et al. [2022]. Our work aims to address this research gap by introducing a novel approach that builds upon the existing findings. Specifically, we incorporate correlation loss during the training phase of our model, which promotes better decorrelation among the latent variables. By doing so, we enhance the performance of the hyperprior module, thereby improving the overall efficacy of LIC models.

3 Methodology

3.1 Formulation of Learned Image Compression Models

Image compression can be formulated through the transform coding approach by (as Figure 2 and 3)

$$\begin{aligned} y &= g_a(x; \phi) \\ \hat{y} &= Q(y) \\ \hat{x} &= g_s(\hat{y}; \theta), \end{aligned} \tag{1}$$

where x and $\hat{x}$ are input and reconstructed image, y is a latent representation, and $\hat{y}$ is the quantized latent representation. ϕ and θ are the parameters of analysis and synthesis transforms g_a and g_s, respectively, Q represents the quantization. During training, the quantization is approximated using uniform noise. While for inference, a round function is applied for quantization to generate $\hat{y}$, and then entropy encoded to generate the bitstream. Given a probability model for the quantized representation, entropy encoding techniques such as arithmetic coding Rissanen and Jr. [1981] can losslessly compress the quantized codes. Arithmetic coders are optimal entropy coders which can be used to estimate the entropy of $\hat{y}$ during training.

A prior work Ballé et al. [2017] has employed a non-adaptive density model which was shared between the encoder and decoder also called factorized prior. It proposed a hyperprior in the follow-up work as a mechanism to introduce side information z to capture the spatial dependencies in the latent space y Ballé et al. [2018]. The side information predicted the mean and scale information for the latent space as:

$$\begin{aligned} z &= h_a\left(y; \phi_h\right) \\ \hat{z} &= Q(z) \\ p_{\hat{y}|\hat{z}}(\hat{y} \mid \hat{z}) &\leftarrow h_s\left(\hat{z}; \theta_h\right), \end{aligned} \tag{2}$$

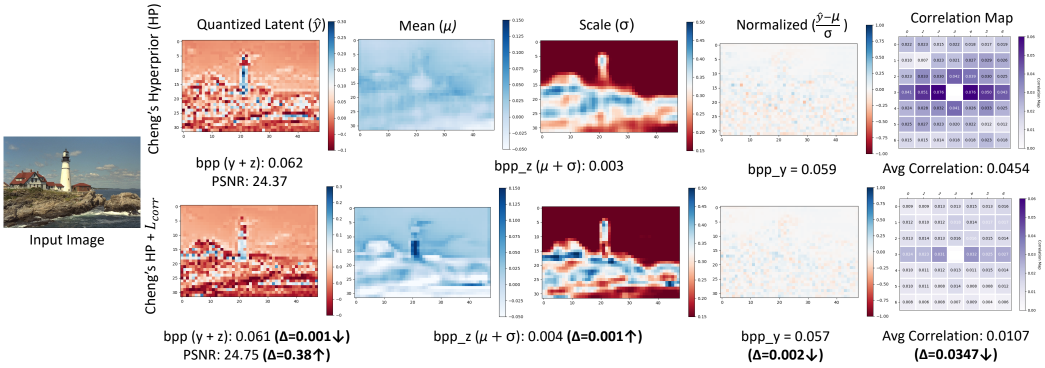


Figure 4: Visualization of quantized latent $\hat{y}$, hyperprior (μ and σ), normalized latent, and correlation map using Cheng's Hyperprior and Cheng's Hyperprior with correlation loss using an image from the Kodak dataset. The application of correlation loss enhances the details of the hyperprior and significantly reduces the correlation of latent variable $\hat{y}$, leading to improved compression efficiency and reconstruction quality.

where h_a and h_s denote the analysis and synthesis transform for the hyperprior auto-encoder. $p_{\hat{y}|\hat{z}}(\hat{y} \mid \hat{z})$ is the estimated probability distribution for the quantized latent space $\hat{y}$, given by:

$$p_{\hat{y}|\hat{z}}\left(\hat{y}_i \mid \hat{z}\right) = \left[\mathcal{N}\left(\mu_i, \sigma_i^2\right) * \mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right)\right]\left(\hat{y}_i\right) \tag{3}$$

$$p_{\hat{y}|\hat{z}}(\hat{y} \mid \hat{z}) = \prod_i p_{\hat{y}|\hat{z}}\left(\hat{y}_i \mid \hat{z}\right), \tag{4}$$

where the mean μ_i and scale σ_i parameter is the i^{th} element of $h_s(\hat{z})$ for each spatial location i in the latent space $\hat{y}$. With the premise that individual latent components are mutually independent, each spatial element in latent $\hat{y}$ is modeled as a Gaussian with its own mean and standard deviation. We may therefore estimate entropy using eq. 4. We modeled the probability of the hyper-latent $\hat{z}$ using a non-parametric fully factorized density model as:

$$p_{\hat{z}|\psi}(\hat{z} \mid \psi) = \prod_i \left(p_{z_i|\psi}(\psi) * \mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right)\right)\left(\hat{z}_i\right), \tag{5}$$

where z_i denotes the i^{th} element of z, and i specifies the position of each element, and the vectors $\psi(i)$ encapsulate the parameters of each univariate distribution $(p_{z_i|\psi}(\psi))$.

A context model can be added to boost the RD performance in the mean-scale hyperprior framework. Additionally, a parameter inference network (g_{ep}) is employed to estimate the mean and scale parameters $\Phi = (\mu, \sigma)$ by leveraging the outputs of the hyperprior module $h_s(\hat{z})$ and the context model $g_{cm}(\hat{y}_{<i})$:

$$\Phi_i = (\mu_i, \sigma_i) = g_{ep}\left(h_s(\hat{z}), g_{cm}\left(\hat{y}_{<i}\right)\right), \tag{6}$$

where $\hat{y}_{<i}$ means the causal context of the nearby visible latents to the latent $\hat{y}_i$. The context model can be calculated as:

$$g_{cm}(x) = (M \odot W)x + b, \tag{7}$$

where $\odot$ refers to the Hadamard product, b is a bias term. The variable M represents a mask commonly used to establish a top-left reference, necessitating strict Z-ordered serial decoding.

The Checkerboard context model He et al. [2021] introduces a parallel decoding approach where only half of the latent variables are encoded/decoded using a checkerboard-shaped context and hyperprior as shown in the Figure 3. The coding of the remaining half of the latents, referred to as anchors, relies solely on the hyperprior. To implement this approach, the eq. 6 above is modified as:

$$\Phi_i = \begin{cases} g_{ep}\left(h_s(\hat{z}), 0\right)_i, & \hat{y}_i \in \hat{y}_{\text{anchor}} \\ g_{ep}\left(h_s(\hat{z}), g_{cm}\left(\hat{y}_{\text{anchor}}; M \odot W\right)\right)_i, & \text{otherwise} \end{cases} \tag{8}$$

The masked convolution, denoted as g_{cm}, is applied conditioned on a checkerboard-shaped mask M, as described in eq. 7. The input for g_{cm}, denoted as $\hat{y}_{anchor}$, corresponds to the set of anchors, and i represents the index for the i-th element $\hat{y}_i$ in the latent variable $\hat{y}$. For methods that utilize

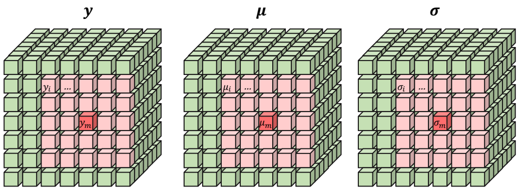


Figure 5: After normalizing the latent $\frac{y-\mu}{\sigma}$, the correlation will be calculated for all latents in the active $k \times k$ window, shown in pink, with the central point m.

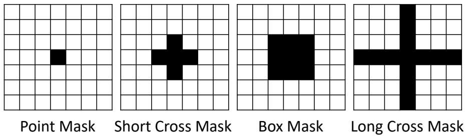


Figure 6: Our proposed mask designs for the correlation loss. We have used Point Mask as a standard for all our experiments. Please refer to the supplementary material for the detailed ablation studies.

mean-scale Gaussian entropy models, the entropy parameter Φ is represented as (μ, σ), where μ and σ correspond to the mean and standard deviation, respectively. In the presence of visible anchors, the context features of all non-anchors can be calculated simultaneously using masked convolution in parallel. The decoding of anchors also occurs in parallel. As a result, the calculation of entropy parameters in eq. 8 for decoding can be performed in two passes, which significantly improves efficiency compared to a serial context model.

LIC is a langrangian multiplier based rate-distortion optmization problem. The loss function is then defined as

$$\begin{aligned}\mathcal{L} =&\mathcal{R}(\hat{y}) + \mathcal{R}(\hat{z}) + \lambda \cdot \mathcal{D}(x, \hat{x}) \\ =&\mathbb{E}\left[-\log_2\left(p_{\hat{y}|\hat{z}}(\hat{y} \mid \hat{z})\right)\right] + \mathbb{E}\left[-\log_2\left(p_{\hat{z}|\psi}(\hat{z} \mid \psi)\right)\right] + \lambda \cdot \mathcal{D}(x, \hat{x}),\end{aligned} \tag{9}$$

where λ controls the rate-distortion tradeoff and $\mathcal{D}(x, \hat{x})$ is the distortion term usually using mean square error.

3.2 Our Proposed Correlation Loss

As shown in the eq. 4, the joint probability of $\hat{y}$ in the hyperprior-based entropy is calculated based on the assumption that the latent features are mutually independent. However, as shown in the upper portion of last column of Figure 4, the latent space is highly correlated, resulting in a discrepancy between the actual and the calculated probability distribution. To address this discrepancy, we introduce correlation loss to force the latent to be decorrelated spatially. We can therefore model the correlation loss by calculating the correlation in the latent space. We start by computing the correlation map as:

$$Corr\_Map_{k\times k}[i] = \mathbb{E}_{x\sim p(x)}\left[\left(\frac{y_i-\mu_i}{\sigma_i}\right)\left(\frac{y_m-\mu_m}{\sigma_m}\right)\right], 0 \le i < k^2 \tag{10}$$

where $k \times k$ is the window size of the correlation map, μ and σ refer to the hyper-latent's corresponding i^{th} and m^{th} elements, m refers to the central point of the window of size $k \times k$ as shown in Figure 5. We slide the window with the stride of 1 across the entire latent space to calculate the correlation of the central point m with all the other points in the window of size $k \times k$. We then take the mean of all the individual correlation maps to get the resultant correlation map of the latent y of size $k \times k$.

We then apply the $Mask$ on the correlation map. As shown in Figure 6, the point mask only masks the central location of latent space as it corresponds to self-correlation of 1 (when $i = m$). We propose several mask designs (discussed in supplementary material) to investigate the effect of the spatially relaxed decorrelation constraint. $Mask$ can then be applied as:

$$Masked\_Map_{k\times k}[i] = Corr\_Map_{k\times k}[i] \odot Mask. \tag{11}$$

The correlation loss is formulated by applying the L_2 norm on the $Masked\_Map_{k\times k}$.

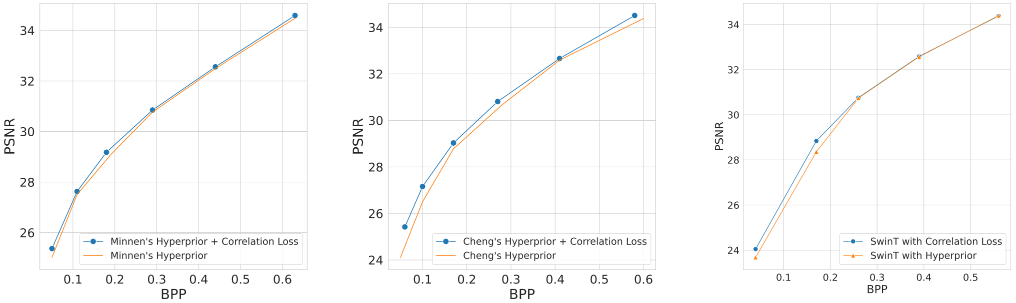
$$L_{corr} = ||Masked\_Map_{k\times k}[i]||^2. \tag{12}$$

By incorporating the correlation loss L_{corr}, the RD loss can now be given as:

$$RD_{loss} = E_{x\sim p_{(x)}}\left[-\log_2 p_{\hat{y}|\hat{z}}(\hat{y} \mid \hat{z}) - \log_2 p_{\hat{z}}(\hat{z})\right] + \lambda \cdot E_{x\sim p_{(x)}}[d(x, \hat{x})] + \alpha \cdot [L_{corr}], \tag{13}$$

where α is the scaling factor for the correlation loss.

Given that the entropy estimate presumes independent latent elements, our proposed loss function increases compression efficiency by reducing the correlation between latent features, hence reducing the discrepancy between the assumed probability distribution of the entropy model and the actual latent probability distribution.



(a) Minnen's mean scale hyperprior (b) Cheng's mean scale hyperprior (c) SwinT mean scale hyperprior

Figure 7: RD rate comparison of (a) Minnen's (b) Cheng's and (c) SwinT mean scale hyperprior models with the inclusion of correlation loss.

4 Experiments and Results

4.1 Experimental Setup

We train all our models on the Vimeo-90k dataset Xue et al. [2019], with the images cropped to the resolution of 256×256 for training. The models were optimized using the Adam optimizer Kingma and Ba [2015] with a batch size of 16 and trained for 1.5 million iterations with a learning rate of 1×10^{-4} for the first million iterations and then halved every 50,000 iterations till 1.25 million iterations.

RD loss given in eq. 9 is used to train the baseline models, whereas the modified loss in eq. 13 is used to train models with correlation loss. The rate-distortion tradeoff is guided by λ, whose value is contained in the set [0.0009, 0.0018, 0.0035, 0.0067, 0.0130, 0.0250].

We tested our model on a commonly used Kodak lossless images dataset Kodak [1993], with 24 uncompressed images of 768×512 or 512×768 resolution. Bits per pixel (bpp) provide the rate, while PSNR indicates the quality of the reconstructed image when evaluating rate-distortion performance. In order to demonstrate the coding efficiency, RD curves are drawn.

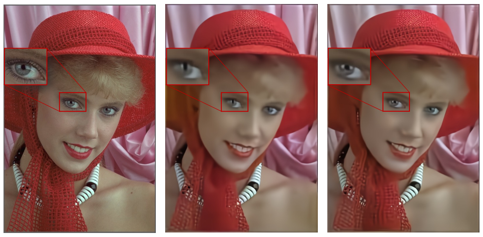
4.2 Rate-Distortion Performance

We employ Minnen's Minnen et al. [2018], Cheng's Cheng et al. [2020], and SwinT Liu et al. [2021] based mean scale hyperprior models as well as Cheng's with Checkerboard He et al. [2021] for our experiments. We perform all our experiments on the Pytorch framework Paszke et al. [2017] and use the CompressAI library Bégaint et al. [2020]. Minnen's and Cheng's models were trained using an NVIDIA 2080Ti, whereas the SwinT model was trained on an NVIDIA 3070Ti due to the transformers' high memory requirement. Note that including correlation loss has a negligible effect on the hardware requirement of these base models.

Minnens' Hyperior: Figure 7a depicts the RD curve for Minnen's model Minnen et al. [2018] with and without correlation loss; for Minnen's model, correlation loss achieves BD rate gains of 3.20%.

Cheng's Hyperprior: Figure 7b illustrates the RD curve for Cheng's model Cheng et al. [2020] with and without correlation loss. Cheng's model yields a BD rate gain of 9.5% with correlation loss. Cheng's model consists of attention modules along with CNN layers that significantly increase the overall efficiency of the correlation loss. Figure 9 depicts the spatial correlation of latent y averaged across all latent channels and compares the baseline Cheng's model with the one trained on correlation loss. Figure 9 demonstrates that the model trained with our proposed loss function exhibits significantly less correlation than the baseline model. This observation holds for the models trained with other λ values. The qualitative comparison in Figure 8 between Cheng's model with and without correlation loss demonstrates the superior performance achieved by incorporating correlation loss. It implies that models trained on correlation loss incur less redundancy across different spatial latent locations, resulting in a better rate-distortion trade-off overall.

Transformer Hyperprior: We also compared the effectiveness of correlation loss on the Transformer-based LIC model. We implemented our SwinT with a hyperprior model inspired by Zhu's work and employed a similar architectural setting Zhu et al. [2022]. Transformers tend to have a lesser correlation in their latent space than the convolutional-based LIC models Zhu et al. [2022]. The

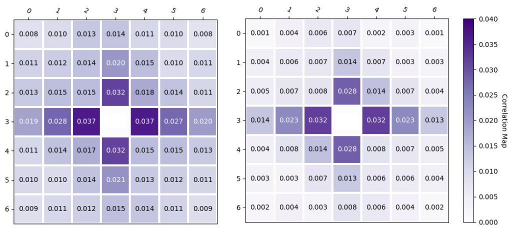


Original Image

Cheng's with Hyperprior
BPP: 0.06, **PSNR: 27.74dB**

Cheng's with Correlation loss
BPP: 0.06, **PSNR: 28.36dB**

Figure 8: Comparison of reconstruction quality of the original image from Kodak dataset using Cheng's hyperprior and Cheng's hyperprior with correlation loss.



Cheng's Hyperprior
(**λ**=0.0067)
Kodak: PSNR=30.58dB, BPP=0.27

Cheng's with Correlation Loss
(**λ** =0.0067)
Kodak: PSNR=30.84dB, BPP=0.26

Figure 9: Spatial correlation map comparison of Cheng's hyperprior (left) with Cheng's hyperprior with correlation loss (right)

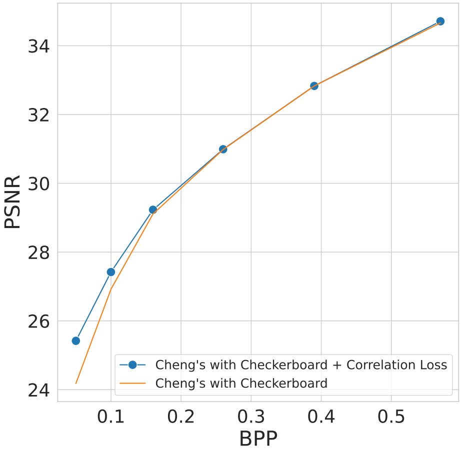


Figure 10: RD rate comparison of Cheng's + Checkerboard with the inclusion of correlation loss.

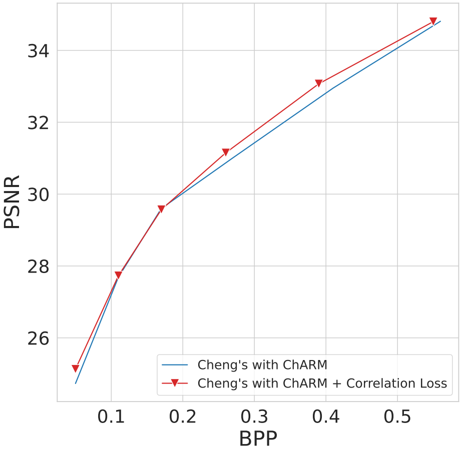


Figure 11: RD rate comparison of Cheng's + ChARM with the inclusion of correlation loss.

Figure 7c illustrates the performance of the SwinT model with and without correlation loss. We found that with the correlation loss, the SwinT model achieves a BD-rate gain of about 4.8%. The effectiveness of correlation loss for the SwinT model further indicates the efficacy of the correlation loss, which in turn improves the mean scale based entropy model.

Cheng's with Checkerboard Context Model: Figure 10 illustrates the RD curve for Cheng's model with Checkerboard He et al. [2021] with and without correlation loss. Cheng's with Checkerboard yields a BD rate gain of 6.13% with correlation loss. Checkerboard uses a parallel decoding approach by encoding and decoding only half of the latent variables using a checkerboard-shaped context. The remaining latents, referred to as anchors, are coded solely based on the hyperprior. Since half of the latents are encoded by the hyperprior module. Incorporating correlation loss into the Checkerboard context model significantly improves the performance of the hyperprior module, resulting in an overall enhancement of the model's performance.

Cheng's with ChARM: Figure 11 illustrates the RD curve for Cheng's model with ChARM Minnen and Singh [2020], both with and without the integration of correlation loss. Notably, when correlation loss is incorporated into Cheng's with ChARM, there is a substantial BD rate gain of 3.3%. ChARM effectively employs channel context and latent residual prediction for the encoding and decoding of the latents. The combination of correlation loss with ChARM leads to a significant enhancement in its overall performance, marking a performance difference of approximately 0.5% when compared to a full AR model.

4.3 Complexity Tradeoff

The primary advantage of the correlation loss lies in its capacity to minimize the discrepancy between the actual and presumed probability distribution within the entropy model. Figure 1 illustrates that for

Table 1: Average encoding and decoding time on the Kodak dataset for different entropy models applied to Cheng's Hyperprior. BD-Rate gains are evaluated using Cheng's Hyperprior as the baseline.

Architecture	BD Rate Gains (%)	Inference Time (sec)
Cheng's Hyperprior (CH)	—	4.66
CH + Correlation Loss (Proposed)	**9.5**	4.66
CH + Checkerboard	10.37	5.33
CH + Checkerboard + Correlation Loss (Proposed)	**16.50**	5.33
CH + ChARM	14.69	8.14
CH + ChARM + Correlation Loss (Proposed)	**17.99**	8.14
CH + AR Context	18.47	251.65

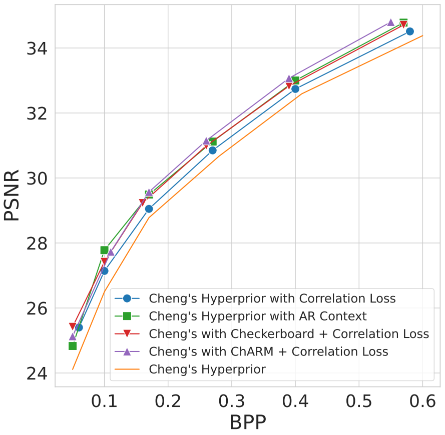


Figure 12: Comparison between Cheng's AR context, Cheng's hyperprior, Cheng's hyperprior with correlation loss, Cheng's Checkerboard with correlation loss, and Cheng's ChARM with correlation loss

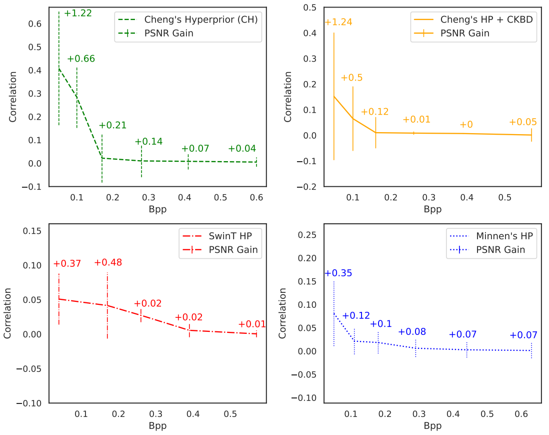


Figure 13: Visual representation depicting the relationship between bits per pixel (bpp) and correlation, and its influence on peak signal-to-noise ratio (PSNR) gains. For all the four models, both the average correlation and the gains decreases as the bpp value increases.

a given encoder and decoder architecture, which is Cheng's model in this case, the choice of entropy model determines the performance of the resulting model. Note that Cheng's hyperprior signifies the lower bound with the given encoder-decoder architecture, whereas Cheng's AR defines the upper limit with the same encoder-decoder architecture. Using Cheng's hyperprior model as the baseline, we compared the effectiveness of including an AR module vs. the correlation loss. The Figure 12 shows the RD curve comparison, while the Table 1 compares the inference time for processing the entire Kodak test dataset. Incorporating correlation loss into Cheng's hyperprior model results in a performance improvement roughly half that of the AR-context-based model, all while reducing inference costs to just 1/55th of the original. Similarly, when utilizing Cheng's Checkerboard with correlation loss, a substantial performance enhancement of approximately 90% is achieved compared to the AR-context-based model, with an inference cost as low as 1/50th of the latter. Additionally, Cheng's ChARM with correlation loss attains an impressive performance gain of about 98% when compared to the AR-context-based model, while incurring an inference cost of merely 1/30th of the AR-context-based model as shown in Figure 1.

It should be noted that when correlation loss is applied to Cheng's hyperprior, significant improvements can be achieved due to the large performance gap between Cheng's hyperprior baseline and the upper limit. On the other hand, since the gaps are diminished in models like Cheng's with Checkerboard and Cheng's with ChARM, the potential for improvement by applying correlation loss is also limited when compared to Cheng's hyperprior baseline. In summary, the introduction of correlation loss to these models would result in a closer approximation to the full AR performance, albeit with gains that are comparatively smaller.

5 Analysis

5.1 Visual Illustration of Effectiveness of the Correlation Loss

Figure 4 visualizes some of the internal workings of our models. We compare Cheng's hyperprior model with and without correlation loss. By observing the first column of Figure 4, it is clear that our correlation loss makes the latent feature preserve more detailed texture information yielding higher PSNR compared with the base model. The mean μ can also extract more distinctive and meaningful information from the latent space. Since the hyperpriors (mean μ and scale σ) contain more information, they require slightly more bits than the base model. Nevertheless, the rise in hyperprior rate is adequately offset by the relatively greater decrease in latent bitrate, as demonstrated in the final column of Figure 4, where the normalized latents exhibit improved decorrelation compared to the base model. Overall, our proposed method surpasses the base model by enhancing PSNR through more decorrelated latents while either maintaining or reducing bpp with better decorrelated latents.

5.2 Effect of Correlation score on the Model's performance

Our proposed loss function consistently demonstrates significant BD-rate gains at lower bit rates while maintaining comparable performance at high bit rates. This behavior is primarily attributed to learned image compression (LIC) models exhibiting higher correlation among latent variables at lower bpp values compared to higher bpp values as reported by Zhu et al. [2022] and also evident from the correlation maps in Figure 6 of the supplementary material. We conducted a comprehensive analysis and presented the findings in Figure 13. Figure 13 showcases the relationship between PSNR gains, bpp, and correlation for different models, including Cheng's hyperprior , Cheng's with Checkerboard, SwinT hyperprior, and Minnen's hyperprior. From the graphs in the Figure 13, a clear pattern can be observed: as the bpp decreases, the correlation of the latents increases, resulting in a higher gain in PSNR. However, as we move towards higher bpps, the correlation becomes notably reduced, resulting in decreased PSNR gain. This trend explains the reason why the efficacy of the correlation loss is larger in low bpp range and diminish at high bpp range.

6 Ablation Studies

Comprehensive ablation studies regarding various mask types, mask sizes, and α values are presented in the supplementary material.

7 Conclusion

The hyperprior-based entropy models assume probabilistic independence, which leads to a discrepancy between the actual and the assumed probability distribution. We proposed correlation loss, which reduces the correlation among spatially neighbored elements in the latent space resulting in a better fit with the spatially independent probability model. Our proposed loss function does not require any model structure or capacity changes and acts as a plug-in method for the existing neural compression methods. The proposed approach, without modifying the entropy models and increasing the inference time, significantly improves the RD performance with BD rate gains up to 17.99%.

Acknowledgments and Disclosure of Funding

Authors are thankful to Muhammad Awais for his constructive and helpful feedback. Authors are also thankful to Taegoo Kang for managing and providing the required compute resources.
This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government [Ministry of Science and Information Communication Technology (MSIT)] (Development of Audio/Video Coding and Light Field Media Fundamental Technologies for Ultra Realistic Tera-media) under Grant 2017-0-00072; in part by the IITP grant funded by the Korea Government (MSIT) (Artificial Intelligence Convergence Innovation Human Resources Development, Kyung Hee University) under Grant RS-2022-00155911; in part by the IITP grant funded by the Korea Government (MSIT) (Artificial Intelligence Innovation Hub) under Grant 2021-0-02068.

References

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=rJxdQ3jeg`.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL `https://openreview.net/forum?id=rkcQFMZRb`.

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.

Fabrice Bellard. Bpg image format. *URL https://bellard. org/bpg*, 2015.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Deep residual learning for image compression. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, page 0. Computer Vision Foundation / IEEE, 2019. URL `http://openaccess.thecvf.com/content_CVPRW_2019/html/CLIC_2019/Cheng_Deep_Residual_Learning_for_Image_Compression_CVPRW_2019_paper.html`.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 7936–7945. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00796. URL `https://openaccess.thecvf.com/content_CVPR_2020/html/Cheng_Learned_Image_Compression_With_Discretized_Gaussian_Mixture_Likelihoods_and_Attention_CVPR_2020_paper.html`.

Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14771–14780. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01453. URL `https://openaccess.thecvf.com/content/CVPR2021/html/He_Checkerboard_Context_Model_for_Efficient_Learned_Image_Compression_CVPR_2021_paper.html`.

Yueyu Hu, Wenhan Yang, Zhan Ma, and Jiaying Liu. Learning end-to-end lossy image compression: A benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8):4194–4211, 2022. doi: 10.1109/TPAMI.2021.3065339. URL `https://doi.org/10.1109/TPAMI.2021.3065339`.

Nick Johnston, Damien Vincent, David Minnen, Michele Covell, Saurabh Singh, Troy Chinen, Sung Jin Hwang, Joel Shor, and George Toderici. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4385–4393, 2018.

Jun-Hyuk Kim, Byeongho Heo, and Jong-Seok Lee. Joint global and local hierarchical priors for learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5992–6001, 2022.

Youmin Kim, Jinbae Park, Younho Jang, Muhammad Salman Ali, Tae-Hyun Oh, and Sung-Ho Bae. Distilling global and local logits with densely connected relations. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6270–6280. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00623. URL `https://doi.org/10.1109/ICCV48922.2021.00623`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

Eastman Kodak. Kodak lossless true color image suite (photocd pcd0992), 1993. URL `http://r0k.us/graphics/kodak/`.

Jae-Han Lee, Seungmin Jeon, Kwang Pyo Choi, Youngo Park, and Chang-Su Kim. Dpict: Deep progressive image compression using trit-planes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16113–16122, 2022.

Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=HyxKIiAqYQ`.

Jianjun Lei, Xiangrui Liu, Bo Peng, Dengchao Jin, Wanqing Li, and Jingxiao Gu. Deep stereo image compression via bi-directional coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19669–19678, 2022.

Mu Li, Kede Ma, Jane You, David Zhang, and Wangmeng Zuo. Efficient and effective context-based convolutional entropy modeling for image compression. *IEEE Trans. Image Process.*, 29: 5900–5911, 2020. doi: 10.1109/TIP.2020.2985225. URL `https://doi.org/10.1109/TIP.2020.2985225`.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00986. URL `https://doi.org/10.1109/ICCV48922.2021.00986`.

David Minnen and Saurabh Singh. Channel-wise autoregressive entropy models for learned image compression. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3339–3343. IEEE, 2020.

David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10794–10803, 2018. URL `https://proceedings.neurips.cc/paper/2018/hash/53edebc543333dfbf7c5933af792c9c4-Abstract.html`.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Yichen Qian, Zhiyu Tan, Xiuyu Sun, Ming Lin, Dongyang Li, Zhenhong Sun, Hao Li, and Rong Jin. Learning accurate entropy model with global reference for image compression. *International Conference on Learning Representations (ICLR)*, 2021.

Yichen Qian, Ming Lin, Xiuyu Sun, Zhiyu Tan, and Rong Jin. Entroformer: A transformer-based entropy model for learned image compression. *International Conference on Learning Representations (ICLR)*, 2022.

Majid Rabbani and Rajan Joshi. An overview of the jpeg 2000 still image compression standard. *Signal processing: Image communication*, 17(1):3–48, 2002.

Jorma Rissanen and Glen G. Langdon Jr. Universal modeling and coding. *IEEE Trans. Inf. Theory*, 27(1):12–22, 1981. doi: 10.1109/TIT.1981.1056282. URL `https://doi.org/10.1109/TIT.1981.1056282`.

Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948. doi: 10.1002/J.1538-7305.1948.TB01338.X. URL `https://doi.org/10.1002/j.1538-7305.1948.tb01338.x`.

Matthew Silk, Ricardo Correia, Diogo Veríssimo, Audrey Verma, and Sarah L Crowley. The implications of digital visual media for human-nature relationships, 2021.

Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=rJiNwv9gg`.

George Toderici, Sean M O'Malley, Sung Jin Hwang, Damien Vincent, David Minnen, Shumeet Baluja, Michele Covell, and Rahul Sukthankar. Variable rate image compression with recurrent neural networks. *International Conference on Learning Representations (ICLR)*, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4): 30–44, 1991.

Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T. Freeman. Video enhancement with task-oriented flow. *Int. J. Comput. Vis.*, 127(8):1106–1125, 2019. doi: 10.1007/S11263-018-01144-2. URL `https://doi.org/10.1007/s11263-018-01144-2`.

Yibo Yang, Stephan Mandt, Lucas Theis, et al. An introduction to neural data compression. *Foundations and Trends® in Computer Graphics and Vision*, 15(2):113–200, 2023.

Zhisheng Zhong, Hiroaki Akutsu, and Kiyoharu Aizawa. Channel-level variable quantization network for deep image compression. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 467–473. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/65. URL `https://doi.org/10.24963/ijcai.2020/65`. Main track.

Yinhao Zhu, Yang Yang, and Taco Cohen. Transformer-based transform coding. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=IDwN6xjHnK8`.