Implicit Neural Multiple Description for DNA-based data storage

Trung Hieu Le, Xavier Pic, Jeremy Mateos, Marc Antonini

IMPLICIT NEURAL MULTIPLE DESCRIPTION FOR DNA-BASED DATA STORAGE

Trung Hieu Le, Xavier Pic* , Jeremy Mateos, Marc Antonini*

I3S laboratory, Côte d'Azur University and CNRS, UMR 7271, Sophia Antipolis, France

ABSTRACT

DNA exhibits remarkable potential as a data storage solution due to its impressive storage density and long-term stability, stemming from its inherent biomolecular structure. However, developing this novel medium comes with its own set of challenges, particularly in addressing errors arising from storage and biological manipulations. These challenges are further conditioned by the structural constraints of DNA sequences and cost considerations. In response to these limitations, we have pioneered a novel compression scheme and a cutting-edge Multiple Description Coding (MDC) technique utilizing neural networks for DNA data storage. Our MDC method introduces an innovative approach to encoding data into DNA, specifically designed to withstand errors effectively. Notably, our new compression scheme overperforms classic image compression methods for DNA-data storage. Furthermore, our approach exhibits superiority over conventional MDC methods reliant on auto-encoders. Its distinctive strengths lie in its ability to bypass the need for extensive model training and its enhanced adaptability for fine-tuning redundancy levels. Experimental results demonstrate that our solution competes favorably with the latest DNA data storage methods in the field, offering superior compression rates and robust noise resilience.

***Index Terms*—** DNA data storage, Multiple Description Coding (MDC), Implicit Neural Network (INR), Quaternary Shannon Fano Entropy Coder (SFC4).

1. INTRODUCTION

The memory of humanity hinges on our capacity to effectively handle ever-expanding volumes of data, spanning timeframes ranging from mere years to several centuries. As our current storage media struggle to keep pace, there is an urgent need to explore groundbreaking solutions that can be swiftly put into practical use. In the development of alternative data storage methods, synthetic molecules, particularly synthetic DNA, appear as one of the most promising options. Due to its density, durability, and its low energy consumption, synthetic DNA is an ideal storage support candidate for long-term data storage. The initial phase in the data encoding process involves constructing a sequence of nucleotides A, T, C, and G (referred to as nts). However, it is imperative that the DNA-encoded information stream follows specific biochemical constraints. These constraints include avoiding homopolymers, maintaining a balanced GC content, and preventing repetitive patterns. Additionally, it is crucial to acknowledge that the biochemical procedures involved in this process can introduce errors that may compromise the integrity of the stored data. Operations such as synthesis, sequencing, storage, and DNA manipulation can introduce errors in the form of substitutions and indels (insertions or deletions of nucleotides). During the last decade, information therorists have developed different schemes for the encoding of digital data into DNA, with some of them targeting the storage of images [1, 2]. Some compression algorithm and coders have been developed specifically for this paradigm of data storage [3, 4, 5, 6]. This work introduces a Single Description Coder (SDC) and a Multiple Description Coder (MDC) designed for DNA data storage with the SDC method exhibiting superior compression performance compared to the existing state of the art. MDC for image encoding involves encoding multiple representations of an image; if one is lost or corrupted during transmission, the remaining descriptions can still be used to reconstruct the original image with some quality degradation. Recent research [7, 8] show a potential use of neural networks to generate different descriptions, which involve Generative Networks and Compressive Autoencoders. However, the main drawback of this method is its long training process that has a high computational cost. Furthermore, the training process must be performed with very large datasets to converge towards an optimal model. This is even more challenging in the MDC context due to the redundancy adaptation mechanism, which requires retraining the model.

In recent works on image compression using neural networks, the so-called Implicit Neural Representation (INR), learns to represent an image implicitly through its weights, a coordinate map, and possibly a latent space [9, 10]. More recently, the Coordinate-based Low Complexity Hierarchical Image Codec (COOL-CHIC) framework [11] has achieved superior performance compared to traditional image compression methods. The first MD scheme using INR (INR-MDSQC) is proposed in [12] with the following advantages: generalized model training is unnecessary, high performance and flexible redundancy tuning. However, INR-MDSQC's drawback is the number of descriptions, which is limited to two. Moreover, those descriptions are not balanced. The goals of implementing Multiple Description Coding (MDC) in DNA data storage are twofold: minimizing the reading cost, and enhancing noise robustness. This is particularly crucial due to the biochemical constraints inherent in the process, which can result in the absence of certain oligos. To our knowledge, this work constitutes the first MDC application for DNA data storage. More precisely, we propose a Spatial Frequency Multiple Description based on INR (SF-MDC) generalized to N descriptions, and evaluate its performance on the Kodak Lossless True Color Image Suite dataset.[1].

2. SPATIAL FREQUENCY MDC

In this section, we introduce a SF-MDC approach that incorporating an INR. The SF-MDC architecture, as depicted in Fig. 1, comprises four main components:

- N sets of hierarchical latent spaces

*These authors contributed equally to this work

[1] http://r0k.us/graphics/kodak/

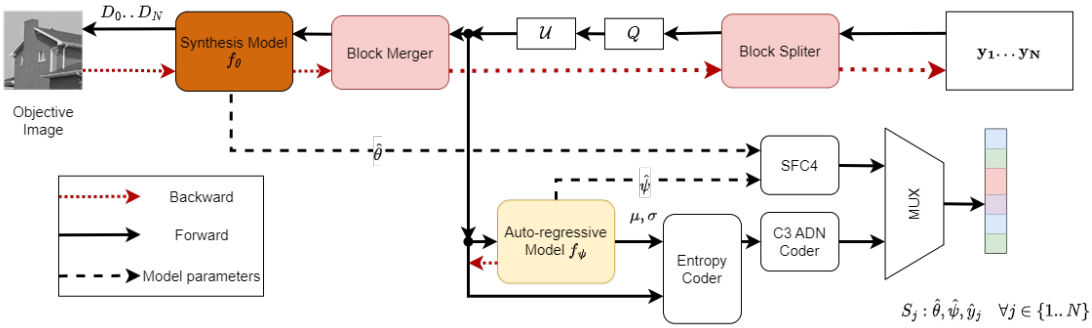


Fig. 1: **SF-MDC**: During the training process, N latent sets are initially divided into blocks of size 8x8. Each block is then quantized independently with added uniform noise. These quantized latent blocks are then fed into the Block Merger module. In this module, each block is categorized as either redundancy or principal. Principal blocks are merged to form the central description, as illustrated in Fig. 2. Both these side descriptions and the central description are then input into the synthesis module, which generates the corresponding reconstruction and computes the related distortion. The latent space is updated using the back-propagation process, which is based on the distortion measured in MSE. Simultaneously, the Auto-regressive model is refined to better estimate the distribution of the quantized latent space.

- f_θ: Synthesis Model with θ its parameters
- f_ψ: Auto-regressive Model with ψ its parameters
- Block Splitter/Merger

2.1. Synthesis model

The quantization process is defined as follows:

$$\hat{s} = Q(s, \Delta s) \tag{1}$$

where s is the element to be quantized, and Δs is its associated quantization step. The latent spaces corresponding to each description $\mathbf{y}_j \in \{\mathbf{y}_1..\mathbf{y}_N\}$ are hierarchically organized at different levels of resolution. Accordingly, we denote by $\mathbf{y}_{k|j}$ the latent space corresponding to resolution level k for description j. In our solution, each description contains a mix of redundancy (low rate, low quality) and principal (high rate, high quality) blocks. At the decoder, when all the descriptions are received correctly, the decoder will merge all the principal blocks to form the central description. Otherwise, the redundancy blocks will be used to replace any corrupted principal blocks. Therefore at the encoding phase, to distribute equal amounts of redundancy across descriptions, block splitter divides each $\mathbf{y}_{k|j}$ into M blocks, each of size 8x8. The segment of the latent delineated by block b is denoted as $y^b_{k|j}$, where $b \in \{0, 1, ..., M-1\}$ is the block index. Each $y^b_{k|j}$ is quantized with a unique quantization step $\Delta y^b_{k|j}$. A principal block uses a finer step, and a redundancy block uses a coarser step. The central description $\hat{\mathbf{y}}_0$ is merged from the principal blocks as depicted in Fig. 2. Hence, each quantized block $\hat{y}^b_{k|j}$ is expressed as:

$$\hat{y}^b_{k|j} = Q(y^b_{k|j}, \Delta y^b_{k|j}) \tag{2}$$

Therefore, the quantized latent space $\hat{\mathbf{y}}_{\mathbf{k|j}}$ is defined as:

$$\hat{\mathbf{y}}_{k|j} = \{\hat{y}^b_{k|j} \in \mathbb{Z}^{8\times 8}, b \in \{0, 1, ..., M-1\}\} \tag{3}$$

From this, we deduce description j composed by the set of different quantized latent spaces $\hat{\mathbf{y}}_{k|j}$:

$$\hat{\mathbf{y}}_j = \{\hat{\mathbf{y}}_{k|j} \in \mathbb{Z}^{H_k \times W_k}, k \in \{0, 1, ..., L-1\}\} \tag{4}$$

where $H_k = \frac{H}{2^k}$, $W_k = \frac{W}{2^k}$, and L denotes the hierarchical depth of $\hat{\mathbf{y}}_j$. As discussed in [12], $\hat{\mathbf{y}}_j$ is sequentially input into the synthesis model f_θ with shared parameters, transforming set of latent spaces into a reconstructed image. In the synthesis model f_θ, each $\hat{\mathbf{y}}_{k|j}$ is first upsampled using bi-cubic interpolation to match the target image shape before being fed into the MLP. The output of f_θ is defined as:

$$\hat{\mathbf{x}}_j = f_\theta(\hat{\mathbf{y}}_j) \quad \text{where} \quad j \in \{0, .., N\} \tag{5}$$

The distortion of each $\hat{\mathbf{x}}_\mathbf{j}$ compared to the target image is denoted by D_j and measured using Mean Squared Error (MSE). Given that the latent space is discrete and the quantization process is non-differentiable, uniform noise is introduced to enable differentiable operations, as described in [13]. Thus, the latent space quantization is defined as:

$$\hat{y}^b_{k|j} = \begin{cases} y^b_{k|j} + u, & u \sim \mathcal{U}[-0.5, 0.5] \text{ during training} \\ Q(y^b_{k|j}) & \text{otherwise} \end{cases}$$

where $\mathcal{U}$ is the uniform noise and $j \in \{1, ..., N\}$ (6)

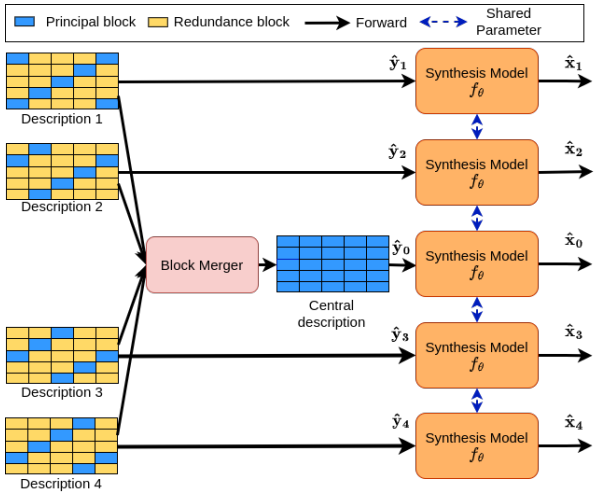


Fig. 2: **Block Merger Module**: In this example, the number of descriptions is $N = 4$. The Principal and Redundancy blocks are assigned using the principle of round-robin item attribution. The central description is derived from the principal blocks of the 4 descriptions. Each description is then sequentially fed into the Synthesis model.

2.2. Autoregressive model

The auto-regressive probability model f_ψ, implemented as MLP aims to closely estimate the quantized latent distribution p_ψ. Since the distribution of each pixel in the latent space is conditioned by their neighbors, according to [14] the probability of the pixels is determined by a factorized model:

$$p_\psi(\hat{\mathbf{y}}_j) = \prod_{i,k} p_\psi(\hat{y}_{ik|j} | c_{ik|j}) \tag{7}$$

where $\hat{y}_{ik|j}$ is the latent pixel at the position i of level k of description j and $c_{ik|j} \in \mathbb{Z}^{\mathcal{C}}$ are the set of decoded neighboring pixels $\mathcal{C}$ of $\hat{y}_{ik|j}$ representing decoding context. The auto-regressive model p_ψ uses the Laplace distribution as described in [11] to approximate the real probability of latent space and by using the factorized model equation (7), the rate for each description $\hat{\mathbf{y}}_\mathbf{j}$ can be expressed as:

$$R(\hat{\mathbf{y}}_j) = -log_2(p_{\psi_j}(\hat{\mathbf{y}}_\mathbf{j})) = -log_2 \prod_{i,k} p_{\psi_j}(\hat{y}_{ik|j}|c_{ik|j})$$

$$= -\sum_{i,k} log_2 p_{\psi_j}(\hat{y}_{ik|j}|c_{ik|j}) \quad (8)$$

2.3. Multiple description optimization

The optimization process is divided into two distinct phases: training and post-training. The objective of the training phase is to update the model parameters θ and ψ, and to adapt the various latent spaces $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ to the dynamics of the target image. Its cost function is defined as:

$$J_t = D_0(\hat{\mathbf{y}}_0) + \alpha \sum_{j=1}^{N} D_j(\hat{\mathbf{y}}_j) + \sum_{j=1}^{N} \lambda_j R(\hat{\mathbf{y}}_j) \quad (9)$$

where $\alpha \in [0, 1]$ is the redundancy factor, $R(\hat{\mathbf{y}}_\mathbf{j})$ denotes the rate as defined in equation (8), D_j is the side distortion, and D_0 is the central reconstruction distortion. The differences in distortion, represented by $D_1, \ldots, D_N$, between the side reconstructions $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N$ and the central reconstruction distortion D_0, are dependent on the redundancy factor α. The configuration of cost function (9) pushes the Synthesis model to partition the image information into N distinct descriptions and converges towards maintaining the lowest D_0 possible while accommodating different rates. After training the network, the model parameters ψ, θ are represented as 32-bit values, but such precision is not necessary for transmission. Thus, in the post-training phase the model parameters θ and ψ are quantized according to equation (1), transforming them into $\hat{\theta}$ and $\hat{\psi}$, respectively. The quantization steps for $\hat{\theta}$ and $\hat{\psi}$ are optimized as outlined in [12] by minimizing the post-training cost function:

$$J_p = D_0(\hat{\mathbf{y}}_0, \hat{\theta}, \hat{\psi}) + \alpha \sum_{j=1}^{N} D_j(\hat{\mathbf{y}}_j, \hat{\theta}, \hat{\psi})$$

$$+ \sum_{j=1}^{N} \lambda_j (R(\hat{\mathbf{y}}_j, \hat{\theta}, \hat{\psi}) + R(\hat{\theta}) + R(\hat{\psi})) \quad (10)$$

Where, $R(\hat{\theta})$ and $R(\hat{\psi})$ represent the estimated rate utilizing a Laplace model.

3. ENTROPY CODER ADAPTED TO DNA

3.1. Description coder: Range Transcoder

In the binary case, the Range coder [15] has been used to entropy code the latent space in different MDC schemes. Since the Range coder offers high performance at a very low entropy, we decided to adapt it to DNA by designing a transcoder that encodes its output into DNA. The principle of context latent coding is depicted in Fig. 3. The encoded values from Range coder are then fed to the C_3 coder described in [16]. In this paper, we introduced an arithmetic coder inspired by the JPEG 2000 MQ coder. This coder is based on a fixed-length code C_3 composed of 48 elements. Further inspired by the Run-length Limited (RLL) binary codes, it has been designed to prevent

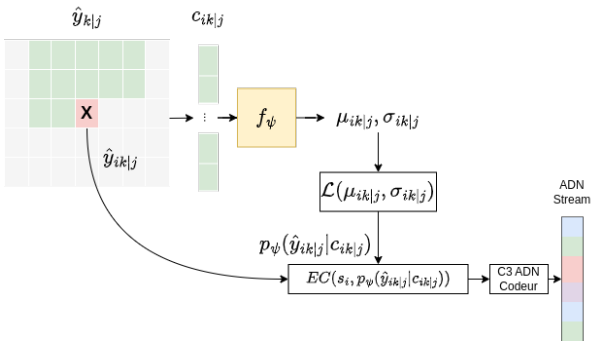


Fig. 3: **Context Entropy Coding with C3 DNA coder:** In this example, the model uses 12 pixels, $c_{ik|j}$, to yield $\mu_{ik|j}$ and $\sigma_{ik|j}$, modeling a Laplacian distribution. The symbol probability is calculated, and an entropy decoder estimates the latent pixel, $\hat{y}_{ik|j}$, from a bitstream. The bitstream is then converted to quaternary code by using the C3 DNA coder.

the occurrence of homopolymers, which are repetitions of the same nucleotide too many times consecutively. The ANS coder output will be represented in base 48. Its base 48 development will be encoded in DNA with the C_3 code.
$C_3 = \{AAT, AAC, AAG, ATA, ATC, ATG, ACA,$ ...,
$GCT, GCG, GGA, GGT, GGC\}$
$|C_3| = 48$

3.2. ARM and Synthesis Models coder: SFC4

In [17], we introduced a novel constrained quaternary entropy coder adapted to the biochemical constraints of DNA data storage, with increased performance over the state of the art Huffman/Goldman algorithm [4]. In [12], the MLP can be modeled by a Laplace distribution, so the code-book is initialized with a frequency table following this Laplace model. After initialization, the SFC4 encoder will be used to encode all the parameters of the ARM and synthesis models, since they are necessary for decoding.

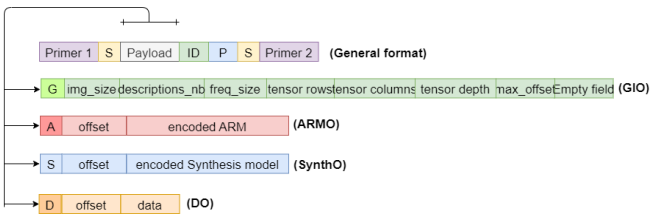
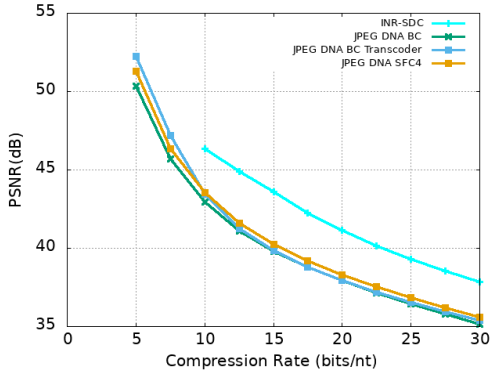


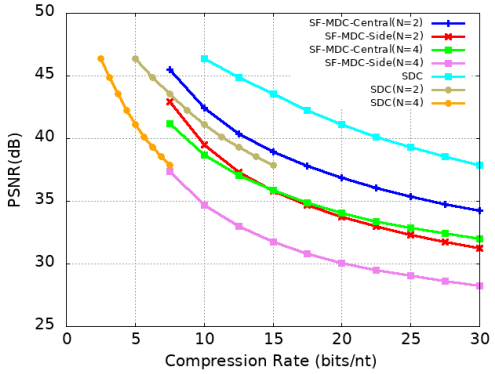
Fig. 4: **Design of the differents oligos format**. General format: The format remains consistent across all oligos, with the only variation occurring in the payload. "S" is the orientation nt, ID is the encoded file's label, P is a set of 4 parity nucleotides. GIO: General informations for the encoded image such as the image size, the number of descriptions and the coding dynamics. ARMO: Contains the weight and the bias of the ARM model. SynthO: Contains the weight and the bias of the Synthesis model. DO: Contains the latent spaces' pixels.

4. OLIGO STRUCTURE

DNA data storage requires the use of short strands called oligos, of length generally lying between 100 and 300 nts. In this work, we use oligos of length 200 nts. The decodability is ensured only if we manage to decode at least one of the descriptions, the Auto Regressive Model, and the synthesis model. In our design, we separated the different parts of the encoded data into separate oligos. Some oligos will encode the ARM model, some the synthesis, and other oligos will encode separate latent spaces, as presented in Fig. 4.



(A) SDC's Rate Distortion



(B) MDC's Rate Distortion

Fig. 5: (A) Average results over the kodak dataset. Our novel SDC coding scheme overperforms all the state of the art coders by at least 0.5 to 3 dB (JPEG DNA BC: [2], JPEG DNA BC Transcoder: [18], JPEG DNA SFC4: [17]), (B) Average result curve over kodak dataset, the MDC side curve is the mean curve across different descriptions. The benchmark is done with the following configuration: N number of descriptions with $N = \{2, 4\}$, $\alpha = 0.1$. The SDC ($N = \{2, 4\}$) is its rate $\times N$, it is equivalent to the compression rate used with MDC ($N = \{2, 4\}$), it allows us to compare SDC and MDC in terms of quality for the same rates.

(A1) Side description 1: 38.757 dB

(A2) Side description 2: 38.789 dB

(A3) Central description: 41.829 dB

(B1) Side description 1: 20.329 dB

(B2) Side description 2: 20.039 dB

(B3) Central description: 35.829 dB

Fig. 6: **Loss simulation on kodim01**: The image is encoded with two side description shown in (A1) and (A2). The central reconstruction computed from these side descriptions is shown in (A3). Noise was then introduced (oligo loss), removing entire latent spaces from those side descriptions. (B1) and (B2) are respective visual results of this noise added to the side descriptions (A1) and (A2), and (B3) is the visual result on the central reconstruction computed from (B1) and (B2).

5. EXPERIMENTS

In the following subsections, we are going to introduce comparative results from different DNA coding methods. All the coders presented here use oligos of length 200 to avoid generating side effects on one of the method's performance. The images used to conduct the test are extracted from the previously mentioned kodak dataset. The number of hierarchy levels used is six ($L = 6$).

5.1. Performance study

The new SDC shows better performance over the state of the art image coding methods adapted to DNA as shown in Fig. 5(A). With this new method, we were able to show gains between 0.5 and 3 dB in terms of quality of reconstruction in comparison to the best previous method (JPEG DNA SFC4 Transcoder). The results have been computed and averaged on the kodak dataset.

To ensure the validity of SF-MDC, its performance at central reconstruction should neither surpass the upper limit of the SDC nor fall below the SDC at an $N\times$ Rate. As shown in Fig. 5 (B), with $\alpha = 0.1$, the solution approaches the upper bound limit of the single SDC as the rate increases, and never goes under the lower bound limit for different N. Besides, we observed that the compression rate increases with the number of descriptions used. On the other hand, increasing the number of description makes the coder more robust to noise.

5.2. Noise robustness

In this section, we simulate the loss for the MDC case $N = 2$. As each latent space is entropy coded and independently decodable. Therefore, to analyze a typical case scenario, we drop three out of six latent spaces from each description, alternating between different levels of descriptions (Description 1: 77% oligo loss, Description 2: 23% oligo loss, and Central Description: 50% oligo loss). The results have been computed on the image kodim01 of the kodak dataset previously mentioned. As observed in Fig. 6, the MDC demonstrates a high resilience capacity, maintaining a loss of only 5dB when losing a big part of the information contained in the different latent spaces.

6. CONCLUSION

This work introduces an innovative DNA-based image codec that achieves substantial improvements in reconstruction quality when compared to existing DNA-based image codecs. On average, these improvements amount to 3 dB, with peak gains of up to 5 dB. These notable enhancements result from the utilization of the ARM, synthesis networks, and the DNA-adapted ANS coder, which deliver exceptional performance even at low entropy levels.

Furthermore, we present a Multiple Description Coder (MDC) capable of generating a variable number of descriptions. This MDC enhances the resilience of oligos to the noise inherent in DNA data storage channels. We also conducted experiments that involved introducing noise into the storage channel. The result shows that we only lost 5dB in the worst scenario.

In future works, we aim at building a noise model for the DNA data storage channel that could further improve the noise robustness of the MDC.

Acknowledgement

We would like to extend our gratitude to Dr. Eva Gil San Antonio for her innovative ideas, suggestions that improved the quality of the paper.

7. REFERENCES

[1] S. M. Hossein Tabatabaei Yazdi, Ryan Gabrys, and Olgica Milenkovic, "Portable and error-free dna-based data storage," *Nature*, 2017.

[2] Melpomeni Dimopoulou, Eva Gil San Antonio, and Marc Antonini, "A jpeg-based image coding solution for data storage on dna," *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 786–790, 2021.

[3] George M. Church, Yuan Gao, and Sriram Kosuri, "Next-generation digital information storage in dna," *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.

[4] N. Goldman, P. Bertone, and S. Chen, "Towards practical, high-capacity, low-maintenance information storage in synthesized dna," *Nature*, 2013.

[5] Marius Welzel, Peter Michael Schwarz, Hannah F. Löchel, Tolganay Kabdullayeva, Sandra Clemens, Anke Becker, Bernd Freisleben, and Dominik Heider, "Dna-aeon provides flexible arithmetic coding for constraint adherence and error correction in dna storage," *Nature Communications*, 2023.

[6] Melpomeni Dimopoulou, Marc Antonini, Pascal Barbry, and R. Appuswamy, "A biologically constrained encoding solution for long-term storage of images onto synthetic dna," *EUSIPCO*, pp. 1–5, 09 2019.

[7] Lijun Zhao, Huihui Bai, Anhong Wang, and Yao Zhao, "Multiple description convolutional neural networks for image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2494–2508, 2019.

[8] Lijun Zhao, Jinjing Zhang, Huihui Bai, Anhong Wang, and Yao Zhao, "LMDC: learning a multiple description codec for deep learning-based image compression," *Multim. Tools Appl.*, vol. 81, no. 10, pp. 13889–13910, 2022.

[9] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein, "Implicit neural representations with periodic activation functions," in *arXiv*, 2020.

[10] Yannick Strümpler, Janis Postels, Ren Yang, Luc Van Gool, and Federico Tombari, "Implicit neural representations for image compression," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVI*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, Eds. 2022, vol. 13686 of *Lecture Notes in Computer Science*, pp. 74–91, Springer.

[11] Théo Ladune, Pierrick Philippe, Félix Henry, Gordon Clare, and Thomas Leguay, "Cool-chic: Coordinate-based low complexity hierarchical image codec," 2023.

[12] Trung Hieu Le, Xavier Pic, and Marc Antonini, "Inr-mdsqc: Implicit neural representation multiple description scalar quantization for robust image coding," *IEEE 25th Workshop on Multimedia Signal Processing (MMSP 2023)*, 2023.

[13] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli, "End-to-end optimization of nonlinear transform codes for perceptual quality," in *2016 Picture Coding Symposium, PCS 2016, Nuremberg, Germany, December 4-7, 2016*. 2016, pp. 1–5, IEEE.

[14] David Minnen, Johannes Ballé, and George Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, Eds., 2018, pp. 10794–10803.

[15] Robert Bamler, "Understanding entropy coding with asymmetric numeral systems (ans): a statistician's perspective," *arXiv preprint arXiv:2201.01741*, 2022.

[16] Xavier Pic and Marc Antonini, "Mq-coder inspired arithmetic coder for synthetic dna data storage," *30th International Conference on Image Processing (ICIP 2023)*, 2023.

[17] Xavier Pic and Marc Antonini, "A constrained shannon-fano entropy coder for image storage in synthetic dna," *European Signal Processing Conference (EUSIPCO)*, 2022.

[18] Luka Secilmis, Michela Testolina, Davi Nachtigall Lazzarotto, and Touradj Ebrahimi, "Towards effective visual information storage on dna support," *Applications of Digital Image Processing XLV*, 2022.