# Towards effective visual information storage on DNA support

Luka Secilmis, Michela Testolina, Davi Lazzarotto, and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
luka.secilmis@epfl.ch, michela.testolina@epfl.ch, davi.nachtigalllazzarotto@epfl.ch,
touradj.ebrahimi@epfl.ch

## ABSTRACT

DNA is an excellent medium for efficient storage of information. Not only it offers a long-term and robust mechanism but also it is environmental friendly and has an unparalleled storage capacity, However, the basic elements in DNA are quaternary, and therefore there is a need for efficient coding of information in quaternary representation while taking into account various biochemical constraints involved. Such constraints create additional complexity on how information should be represented in quaternary code. In this paper, an efficient solution for the storage of JPEG compressed images is proposed. The focus on JPEG file format is motivated by the fact that it is a popular representation of digital pictures. The proposed approach converts an already coded image in JPEG format to a counterpart represented in quaternary representation while taking into account the intrinsic structure of the former. The superiority of the proposed approach is demonstrated by comparing its rate distortion performance to two alternative approaches, namely, a direct transcoding of the binary JPEG compressed file into a quaternary codestream without taking into account its underlying structure, and a complete JPEG decoding followed by an image encoding for DNA storage.

**Keywords:** DNA compression, transcoding, JPEG

## 1. INTRODUCTION

Although the volume of digital information produced by humans is growing at an exponential pace, conventional storage technologies for archival are rapidly approaching their physical limits, resulting in a deceleration in the exponential growth of storage capacities as witnessed during the past several decades. In addition, archival infrastructures, in particular those used in data centers, continue to consume significant amounts of energy for their maintenance, because despite progress in more energy-efficient storage technologies, the amount of information that needs to be archived, grows even faster, offsetting gains in more environmental-friendly data storage solutions, in a world that is becoming more conscious of the impact of human activities on climate change. Furthermore, a large majority of storage technologies have a life-span in the order of decades and require replacement, further increasing the total cost of ownership in long-term data storage. A potentially attractive solution to respond to these challenges resides in archival of information on a DNA support mimicking the way nature stores the genetic code of life in living organisms. Deoxyribonucleic acid (DNA) is a macromolecule composed in quaternary units which encode the genetic information of living organisms. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Similar to encoding digital information in binary digits (01) on electromagnetic and optical devices, DNA can be used as a medium to store information by creating artificial molecules that code the latter based on a quaternary representation (AGCT). DNA data storage provides several advantages with respect to the state of the art in storage technologies. For instance, DNA storage does not require much energy to maintain and thus would significantly reduce the carbon footprint of data storage based on such an approach. It has a much longer lifetime - spanning over hundreds and even thousands of year. Furthermore, DNA is a medium with extremely high density of storage. But currently, it also faces a number of challenges, such as the cost associated with sequencing and synthesis of molecules, and the difficulties in efficient random access as well as larger latency. Furthermore, a number of biochemical constraints must be taken into account when storing and retrieving information coded in DNA oligos (single strands of synthetic DNA) which are error-prone processes.

Recent progress in DNA storage point to a near future where challenges mentioned above are resolved and DNA storage becomes a competitive technology for a large number of applications requiring archival such as long-term data preservation as well as cold storage. In this paper, we propose a DNA based coding solution that stores information represented as JPEG formatted digital pictures in a efficient way and compare it to two alternatives that can be used for this purpose. The first alternative consists in directly transcode the binary bitstream of compressed pictures into quaternary code without taking advantage of the underlying structure and syntax of the latter. The second alternative first decodes the JPEG coded picture and then re-encodes the uncompressed picture represented in its canonical representation in the pixel domain, using a state-of-the art image coding approach.[1] The paper is organized as follows. After this Introduction, Section 2 discusses some of the state of the art in image coding for storage in DNA. Section 3 then presents the proposed solution in details and Section 4 provides some results and discusses them. Conclusions and future work are then discussed in Section 5.

## 2. RELATED WORK

The potential advantages of DNA storage have recently attracted the attention of the research community from academia as well as industry. The large majority of approaches under investigation in data storage on DNA rely on the following workflow . First, the binary information to be stored is converted to a quaternary representation in AGCT codestream which is then synthesized into actual molecules in a physical medium. In long-term storage applications, these molecules are then encapsulated for the purpose of their preservation. In order to read the stored information and to obtain the corresponding quaternary codestream, the DNA strands must be released and subjected to a certain process which is often error-prone and produces various types of errors that must be dealt with. In this paper, we assume that the mechanism to cope with such errors is capable of extracting the original quaternary codestream without any loss and that the length of DNA strains can be as long as needed. Finally, the information represented in quaternary codestream is converted back to its original binary representation.

The design of a system for DNA storage is, in practice, subjected to biological constraints that must be respected in order to result in stable synthesis, storage and sequencing processes. Past studies have identified[2] three main factors that introduce errors into the coding process. The first is the presence of homopolymers, i.e. sections where the same nucleotide is repeated. Moreover, the percentage of G and C nucleotides should be lower or equal to those for A and T. Finally, the repetition of patterns in the oligos can also increase the probability of errors.

It is worth mentioning that the current technologies for storage based on synthetic DNA are error-prone even when the above-mentioned constraints are respected. For that reason, in recent years different simulators have been proposed to ease the implementation of new coding algorithms that can be robust to such residual errors. These simulators model the effect of physical processes on the DNA strands and add errors by simulating nucleotide insertion, deletion and substitution. Although different methods have been proposed[3,4] based on the Nanopore sequencer, MESA[5] which is an open source simulator modelling the full DNA coding channel is a popular tool.

The first attempt of encoding digital data into DNA was made by Church et al.,[6] with an encoding which simply mapped zeros to A or C and ones to T and G, allowing to study biological error of DNA storage. Goldman et al.[7] proposed a new encoding scheme which first converted a sequence of bytes into base-3 using a ternary Huffman tree, and then encoded these trits into nucleotides using a scheme that avoids the previously written nucleotide. Through rotating the set of nucleotides used to encode DNA at each step as presented above, the Goldman encoding was the first DNA encoding which respected the biological constraint of avoiding Homopolymers. Grass et al.[8] proposed the first DNA encoding scheme with error correction using Reed Solomon codes. Since then numerous works have been published on DNA storage, notably the work from Erlich et al.[9] on Fountain codes which proposed a robust and efficient encoding in terms of storage density, but at the expense of important computational complexity. Recently, Dimopoulou et al.[10] proposed a low complexity solution to construct a biologically constrained DNA encoding called PAIRCODE. In this algorithm, codewords are constructed using two distinct dictionaries $C_1$ and $C_2$:

$$C1 = \{AT, AC, AG, TA, TC, TG, CA, CT, GA, GT\} \tag{1}$$

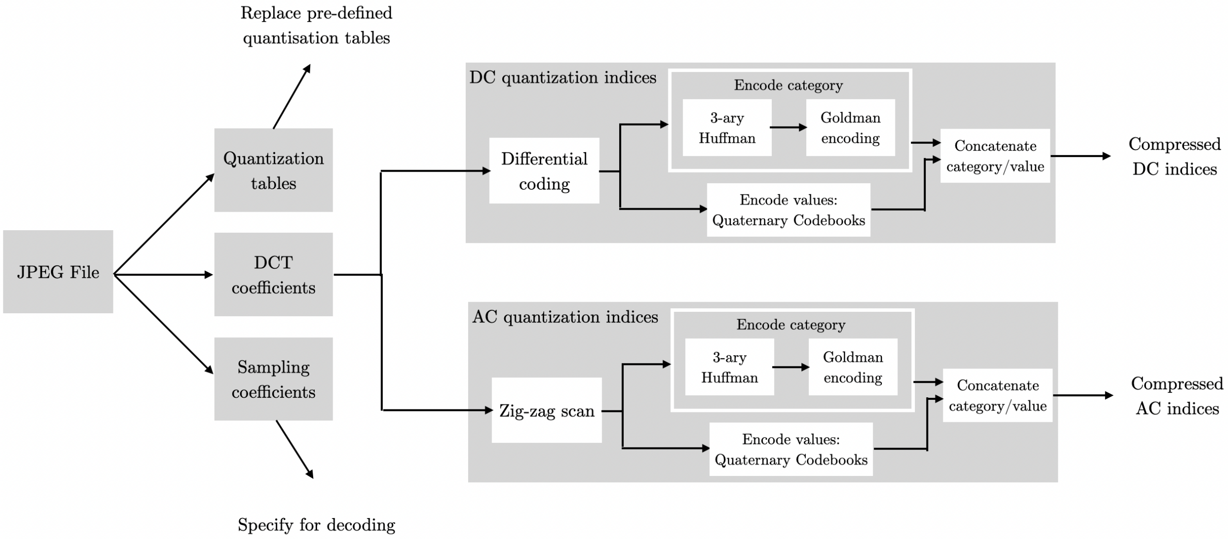


Figure 1: Proposed JPEG DNA Transcoder blockdiagram as a modificaton of JPEG DNA BC[1]

$$C2 = \{A, T, C, G\} \tag{2}$$

Even-length codewords are constructed by selecting doublets from the first dictionary, and odd-length codewords by appending one element from the second dictionary. Even if PAIRCODE is not optimized for coding efficiency, it is robust to sequencing errors and meets all previously mentioned biological constraints.

Both Goldman encoding and PAIRCODE were leveraged in the image coding solution based on JPEG for DNA storage proposed by Dimopoulou et al.[1] This work was selected by JPEG standardization committee as its first benchmark codec.[2] Recent works from Pic et al. have also proposed solutions to combine DNA storage with learning-based image coding methods,[11] as well as a hybrid architecture integrating both JPEG and a convolutional autoencoder.[12]

## 3. JPEG TO DNA TRANSCODING

The coding solution proposed in this paper to compress an already coded JPEG image into a DNA sequence (referred to as JPEG DNA Transcoder) is designed by modifying the JPEG DNA Benchmark Codec[1] (referred to as JPEG DNA BC) which encodes uncompressed images into a quaternary DNA sequence. The rationale behind proposing such a solution is motivated by the fact that a large majority of images are already coded in JPEG format. In particular the proposed JPEG DNA Transcoder is lossless and does not require estimation of a quality factor prior to generating a DNA sequence. In addition, it is more advantageous in terms of computational complexity as it avoids a JPEG decoding in order to produce a decompressed image before performing JPEG DNA BC, which would create undesired distortions because of superfluous full decoding/encoding operations. The workflow of the proposed solution is presented in Figure 1.

Essentially, the transcoding module consists of parsing a JPEG file and extracting its quantized DCT coefficients, quantization tables and sampling factors. The quantized DCT coefficients are then directly plugged into the category/value encoding of JPEG DNA BC, while using the pre-defined quantization tables of the codec as extracted from the JPEG image.

Categories correspond to pre-defined intervals on which the value of a coefficient falls. The category of a DCT coefficient is computed from its value, passed through a ternary Huffman tree and the resulting trit is then encoded using a Goldman Coder. For each category there is one pre-defined fixed-length codebook generated using PAIRCODE which is used to encode the value of the coefficient itself. The resulting nucleotides of category and value are then concatenated to form the final DNA sequence for that coefficient.

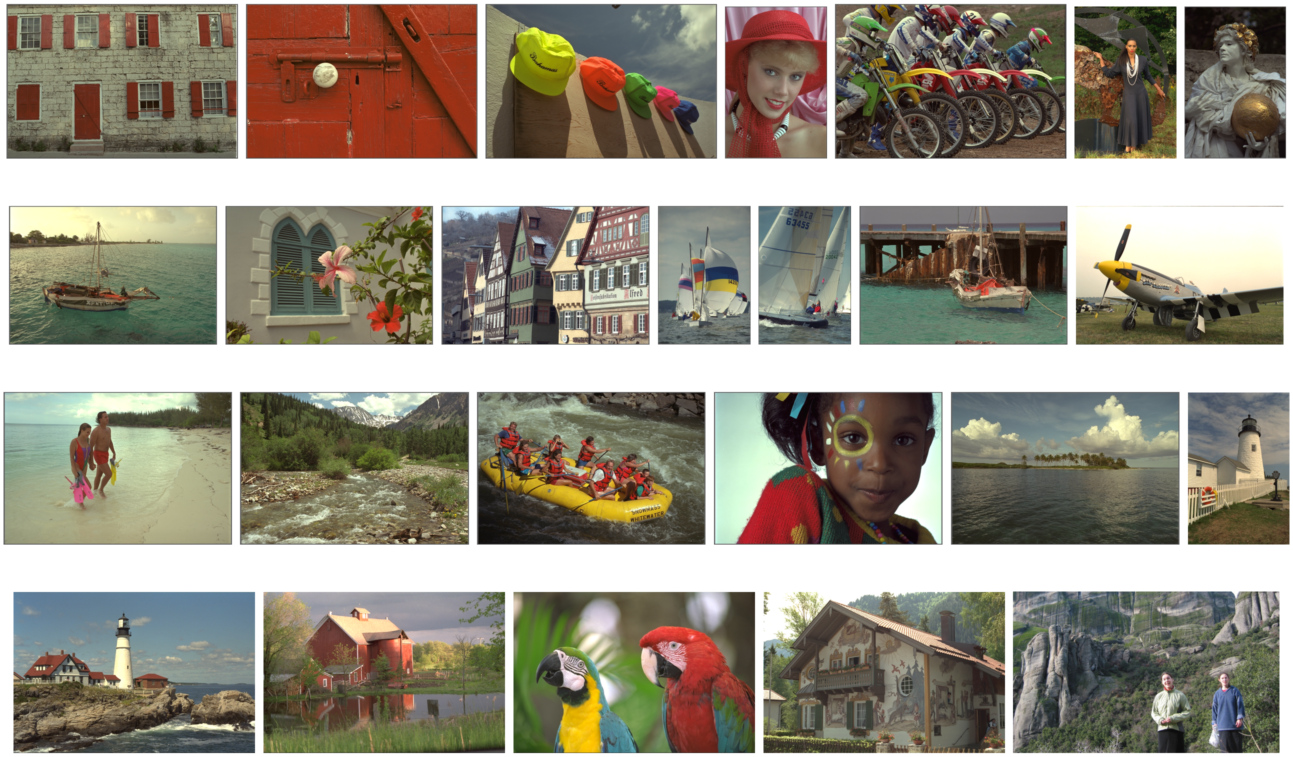


Figure 2: Kodak dataset images used in performance assessment

The encoding of the extracted DC and AC quantized indices differ in that DC coefficients are first differentially coded before computing the above discussed category/value encoding. AC coefficients, on the other hand, first undergo a Zig-zag scan, and the values are Run-length encoded (RLE) as in the legacy JPEG compression algorithm. The category encoding is replaced by a run/category encoding which is a two unit hexadecimal code, where the first unit corresponds to the number of zeros before the non-zero AC coefficient, and the second unit to the category of that non-zero coefficient. This is then again passed in the Goldman encoder, and the value of the non-zero AC coefficient itself is encoded again using the PAIRCODE codebook corresponding to its category.

## 4. RESULTS AND ANALYSIS

The performance of the proposed JPEG DNA Transcoder is assessed on the Kodak PhotoCD PCD0992 dataset,[13] which includes 25 uncompressed images with size 768x512 both in landscape and portrait orientation, and a variety of contents as presented in Figure 2. The entire dataset was first compressed using JPEG with quality levels corresponding to 60, 65, 70, 75, 80, 85, 90, 95, and 100, using the PIL library in Python. Quality levels lower than 60 are not considered in this paper, as the main target application of DNA storage is long-term preservation, where low quality factors are not desirable. After compressing with JPEG, the resulting bitstream is transcoded to DNA using the proposed solution. Results are presented in the form of rate-distortion plots, where the rate is computed in nucleotides per pixel and the PSNR is used as an objective quality metric, where the original uncompressed image is taken as a reference.

Two anchor methods are used as a baseline in this paper, notably:

- **Anchor 1**: the encoded JPEG file is transformed directly into DNA using the Goldman Codec. Notably, the JPEG bitstream is read byte by byte, and each byte is converted to base 3. Consecutively, the resulting ternary stream is Goldman coded. We refer to this anchor as "*direct transcoding*".

- **Anchor 2**: the encoded JPEG file is first decoded using a standard JPEG decoder, and subsequently re-compressed using the JPEG DNA BC. As in JPEG DNA BC the quality is determined by specifying an alpha parameter that multiplies a pre-defined quantization tables, this parameter is selected during coding such that the resulting rate, measured in nucleotides/pixel, is as close as possible to the rate of Anchor 1 or the Proposed JPEG DNA Transcoder. We refer to this anchor as "*full decoding/re-encoding*".

The rate-distortion results, averaged over the Kodak dataset, are presented in Figure 3, where the PSNR is employed as an objective quality assessment metric. From the results, it is possible to observe that the proposed transcoder is able to decrease the rate of the compressed JPEG streams without decreasing their visual quality. Moreover, the 95% confidence intervals (CI), computed assuming a t-Student distribution, reveal rather small confidence intervals. This shows that the performance of the proposed transcoder, as well as of the *direct transcoding* anchor, are only marginally dependent on the content. As presented in Figure 4, image *kodim13* presents the best performance of the presented transcoder among contents in the dataset, having the largest difference between the rate of the proposed transcoder and the rate of the *direct transcoding* anchor. Likewise, *kodim03* presents the lowest performance, where the rate of the proposed transcoder and *direct transcoding* anchor are the most comparable.

Additional observations from Figure 3 regard the *full decoding/re-encoding* Anchor 2. In particular, it can be observed that the proposed transcoder is able to achieve, for the same rate, higher objective performance in terms of PSNR compared to Anchor 2, which suffers by the additional operations performed in the *full decoding/re-encoding*. This highlights that the proposed transcoder is not only able to reduce the computational complexity of the pipeline, but also can improve its fidelity to the original uncompressed image.

Visual results in Figure 5 reveal that the three different approaches, when the highest bitrate is considered, do not present any perceptible differences. However, Figure 6 shows that subtle additional artifacts might appear in the *full decoding/re-encoding* at the lowest quality levels considered in this paper. As an example, Figure 6 shows an additional color-distortion artifact in the area where the orange and green cap overlap.

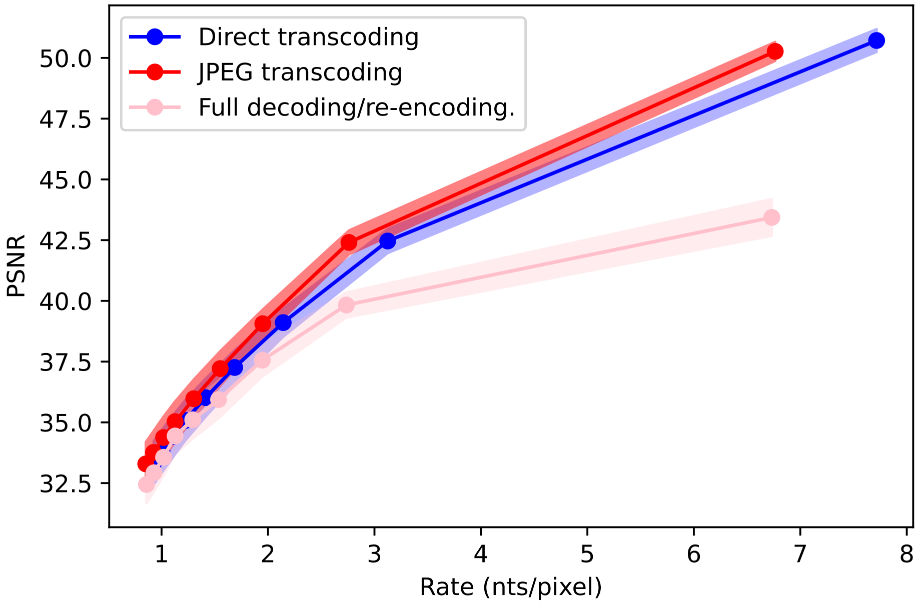
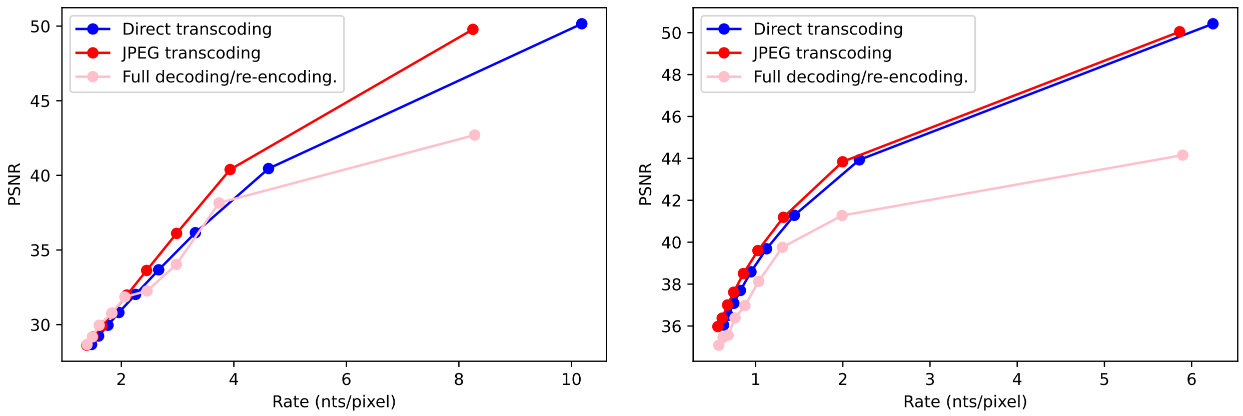


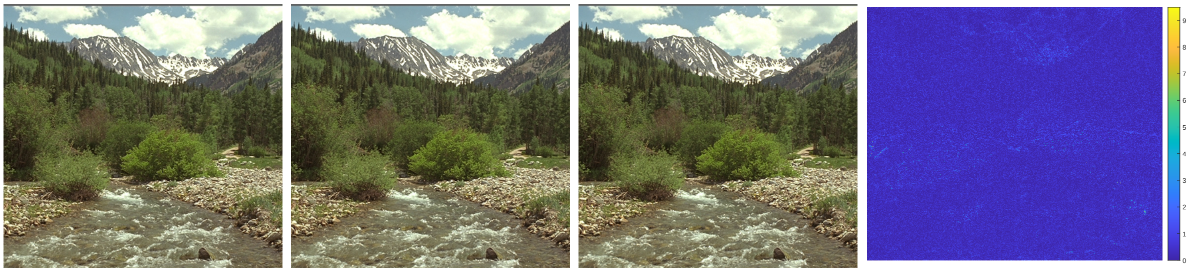
Figure 3: Average rate-distortion results over the Kodak dataset. The employed distortion metric, in this case, is the PSNR. The 95% confidence intervals (CI) were subsequently computed and depicted.



(a) Results for image *kodim13*

(b) Results for image *kodim03*

Figure 4: Rate-distortion results for images *kodim13* and *kodim03*, showing the largest and smallest gain in the proposed transcoding solution when compared to Anchor1.



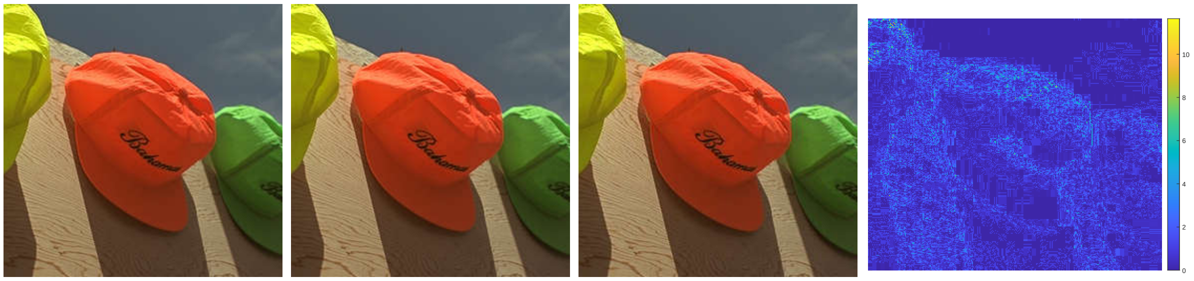
(a) Proposed transcoder

(b) Anchor 1

(c) Anchor 2

(d) Difference (a)-(c)

Figure 5: Visual results for image *kodim13*, coded at the highest quality level. The difference in luminance between the image obtained with the proposed encoder and with the *full decoding/re-encoding* anchor is presented in (d). While it reveals slight differences in the area of the sky and trees, the visual inspection of the images reveal that the difference between the different approaches, for the considered bitrates, is not perceivable by the human eye.



(a) Proposed transcoder

(b) Anchor 1

(c) Anchor 2

(d) Difference (a)-(c)

Figure 6: Visual results for image *kodim03*, coded at the lowest quality level considered in this paper (i.e. quality 60). The difference in luminance between the image obtained with the proposed encoder and with the *full decoding/re-encoding* Anchor 2 is presented in (d). The images reveal that the *full decoding/re-encoding* Anchor 2 introduces additional subtle artifacts that, in this case, are more visible for instance in the orange and green cap overlap.

## 5. CONCLUSIONS AND FUTURE WORK

In this paper, a transcoder able to convert a compressed JPEG image directly into a quaternary DNA stream without the need of full decoding is presented. The proposed method has the advantage of being able to achieve lower rates when compared to a JPEG stream directly coded with a Goldman encoder, and to present higher visual quality when compared to the results obtained by decoding using a standard JPEG decoder followed by re-encoding with JPEG DNA BC. Experimental results have been presented on the Kodak dataset, which although a popular reference, somewhat lacks the high-resolution characteristics of images obtained from recent high-end cameras, and have a reduced variety of types of content. Future work consists in validating these results for a wider variety of higher resolution images and additional types of contents and to include the impact of DNA synthesis, storage and sequencing using the MESA simulator for evaluation in more realistic conditions.

## ACKNOWLEDGMENTS

The authors would like to acknowledge support from the Swiss National Scientific Research project entitled ”Advanced Visual Representation and Coding in Augmented and Virtual Reality” under grant number 200021_178854.

## REFERENCES

[1] Dimopoulou, M., San Antonio, E. G., and Antonini, M., “A JPEG-based image coding solution for data storage on DNA,” in [*2021 29th European Signal Processing Conference (EUSIPCO)*], 786–790, IEEE (2021).

[2] Antonini, M., Cruz, L., da Silva, E., Dimopoulou, M., Ebrahimi, T., Foessel, S., San Antonio, E. G., Menegaz, G., Pereira, F., Pic, X., et al., “Dna-based media storage: State-of-the-art, challenges, use cases and requirements version 8.0.” ISO/IEC JTC1/SC29/WG1 Doc. N100154 (April 2022).

[3] Yang, C., Chu, J., Warren, R. L., and Birol, I., “NanoSim: nanopore sequence read simulator based on statistical characterization,” *GigaScience* **6**(4), gix010 (2017).

[4] Li, Y., Han, R., Bi, C., Li, M., Wang, S., and Gao, X., “DeepSimulator: a deep simulator for Nanopore sequencing,” *Bioinformatics* **34**(17), 2899–2908 (2018).

[5] Schwarz, M., Welzel, M., Kabdullayeva, T., Becker, A., Freisleben, B., and Heider, D., “MESA: automated assessment of synthetic DNA fragments and simulation of DNA synthesis, storage, sequencing and PCR errors,” *Bioinformatics* **36**(11), 3322–3326 (2020).

[6] Church, G. M., Gao, Y., and Kosuri, S., “Next-generation digital information storage in DNA,” *Science* **337**(6102), 1628–1628 (2012).

[7] Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., and Birney, E., “Towards practical, high-capacity, low-maintenance information storage in synthesized DNA,” *nature* **494**(7435), 77–80 (2013).

[8] Grass, R. N., Heckel, R., Puddu, M., Paunescu, D., and Stark, W. J., “Robust chemical preservation of digital information on DNA in silica with error-correcting codes,” *Angewandte Chemie International Edition* **54**(8), 2552–2555 (2015).

[9] Erlich, Y. and Zielinski, D., “DNA Fountain enables a robust and efficient storage architecture,” *science* **355**(6328), 950–954 (2017).

[10] Dimopoulou, M., *Techniques de codage pour le stockage à long terme d’images numériques dans l’ADN synthétique*, PhD thesis, Université Côte d’Azur (2020).

[11] Pic, X. and Antonini, M., “Image Storage on Synthetic DNA Using Autoencoders,” *arXiv preprint arXiv:2203.09981* (2022).

[12] Pic, X. and Antonini, M., “Image coding algorithm for DNA data storage combining JPEG and autoencoders,” in [*MWCC 2022 (Munich Workshop on Coding and Cryptography)*], (2022).

[13] “Kodak Lossless True Color Image Suite (PhotoCD PCD0992),” (accessed: 17.08.2022). "http://r0k.us/graphics/kodak/".