Perceptually optimised Cool-chic for CLIC 2025

Pierrick Philippe, Théo Ladune, Guillaume Lorand, Gordon Clare and Félix Henry
Orange Innovation, France
pierrick.philippe@orange.com

***Abstract*—Neural image compression has proven to be highly effective compared to conventional approaches such as JPEG, HEVC or the latest standard VVC.**

Recently low complexity neural image coding using overfitting has proven to reach the level of performance of VVC. Indeed, at the CLIC 2024 image challenge, Cool-chic demonstrated to be perceptually equivalent to the VVC performance at the 3 target bitrates. The decoding complexity was limited, with less than 2200 operations per pixel, permitting decoding on any legacy CPU.

For this CLIC 2025 candidate, it is proposed to improve Cool-chic quality using a perceptually driven distortion metric and the addition of a random noise.

This paper describes the approach and presents a preliminary subjective evaluation that demonstrates the effectiveness of the solution: 50% rate reduction is demonstrated with the proposed distortion metric. The decoder complexity is reduced to approximately 1700 operations per pixel, it is written in C language and it is operated on CPU.

All contributions are made open-source at https://github.com/Orange-OpenSource/Cool-Chic/tree/clic2025 [1].

***Index Terms*—Image coding, perceptual distortion, overfitted neural coding**

I. Introduction

In the past few years, the CLIC challenges have demonstrated that neural image coding significantly surpasses conventional coding in terms of perceptual quality. The best performing neural coders rely on the auto-encoder approach, optimised using a mixture of distortion metrics (often the Mean Square Error, the MS-SSIM [2] and the LPIPS [3]). Adversarial training techniques (such as GAN) and style losses are also largely used [4], [5].

Although the perceptual performance of theses auto-encoders is significantly higher than the conventional codecs, they exhibit an important decoding complexity that might prevent their adoption in the market: indeed, they require a number of Multiplication Accumulation (MAC) operations per pixel often in the order of 100,000 or more. As such, they require dedicated and powerful devices that incorporate neural units (GPUs, NPUs, etc.).

Another paradigm involves overfitted neural codecs. The image representation is made of a latent representation as for auto-encoders but also includes the decoder neural weights. These weights, specifically adapted to the image, are conveyed in the compressed representation. As the decoder is specific to the image, it does not need to generalise to other types of images and can consequently be much lighter than usual neural decoders: Cool-chic [6], [7] typically uses less than 2000 neural weights and typically requires in the order of 300-2500 MAC per pixel.

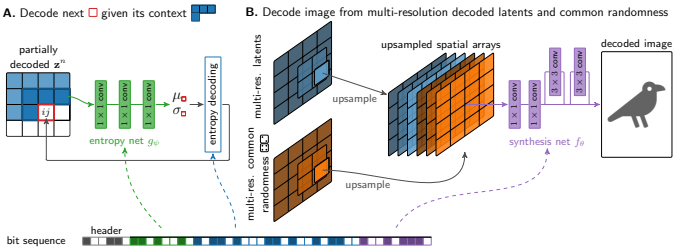

Fig. 1: Functional description of a Cool chic decoder with common randomness, illustration extracted from [10]

Although perceptual tuning of overfitted codecs has been proposed in the past [8], [9], it was limited to the usage of MS-SSIM optimisation. This paper proposes to apply the approach proposed in [10] which motivates the Wasserstein distance applied to a VGG latent representation [11] to drive the Cool-chic perceptual optimisation. As proposed by Ballé et al. [10], the addition of a random noise with a constant seed (called *common randomness*) is also considered.

The objectives of this paper are to:

1) Optimise Cool-chic perceptually using the approach of [10];
2) Assess the effect of the Wasserstein distance and further adapt the distortion metric to improve visual quality;
3) Provide the community with open software and example bitstreams to reproduce and continue the approach [1].

The paper is organised as follows: in a first section, we briefly present the Cool-chic architecture and the encoding strategy. This encoding strategy is examined using the Wasserstein distance in the subsequent section and motivating an adaptation of the initial method toward a better visual quality. The encoding process and the results, provided using visual testing, conclude this paper.

II. Overview and adaptation of Cool-chic

A Cool-chic decoder is made of four elements as shown in figure 1.:

1) a series of $L = 7$ latents, organised in a pyramidal fashion: their resolution ranges from 1 to $\frac{1}{64}$ of that of the image;
2) an entropy decoder, based on an autoregressive MLP, which provides a statistical description of the latents. These statistics drive a binary arithmetic decoder;
3) a linear upsampling block, built with a series of $\times 2$ upsamplers to convert the pyramidal latents to a dense representation for the latents, at the image resolution;

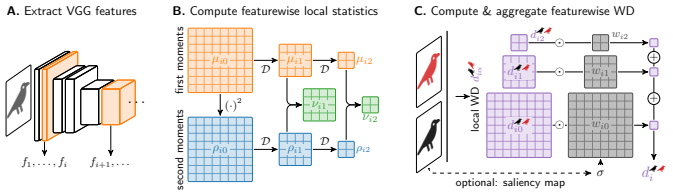

Fig. 2: Wasserstein distance computation illustration extracted from [10]

4) a synthesis neural network which takes the resulting dense latents and outputs the RGB image color channels.

Depending on the desired complexity, the autoregressive and the synthesis networks complexity is adjusted by adding more neurons and hidden layers.

The key objective of Cool-chic is to provide a low complexity decoder whose weights are learned in an end-to-end fashion. The quantised latents are trained alongside the decoder weights in a rate distortion manner, such that the quantity $J(\lambda) = D + \lambda \cdot R$ is minimised. Any derivable distortion D can be used in that context as shown in [8], R is the number of bits per pixel and λ balances the two quantities.

The standard Mean Square Error (*mse*) metric is often used, although perceptual tuning using distortions more aligned with the human visual system have been successfully tested [8], [9].

In [10] it is suggested to complement the latent representation with a pyramid of noise latents. Those latents are obtained by drawing i.i.d. elements from a gaussian pseudo-random number generator with fixed seed. The training process uses this conditioned random noise which is identical at the decoding side using the same seed and generator.

This noise component helps at generating a decoded image whose distribution is similar to the original image. A simple pixel-wise distance is not relevant for that purpose and a distance that considers distribution distance is more appropriate: the Wasserstein is suggested in the same paper.

III. Distortion metric for training

The paper [11] suggests the use of a Wasserstein distance in a feature domain to evaluate the perceptual distance between images. The principle of the computation is displayed in figure 2. A VGG network is applied to both the original image and the decoded one. The obtained features are used to derive the latent's first and second order moments at different scales. The Wasserstein distance is evaluated through a pooling kernel whose spreading ($\sigma_{x,y}$) is adjustable and can vary inside the image. The spreading function can be made made narrow or wide to account for the distribution difference on small or large areas. If the area is small, this is equivalent to a pointwise distance similar to the the LPIPS computation.

Ballé et al. [10] also suggest to adjust the spreading following a saliency map where smaller spreading is chosen in image regions that are likely of interest, and larger spreading otherwise. In this paper, we prefer focusing on the use of the Wasserstein solely with a constant spreading value ($\sigma = 8$) for all image regions.

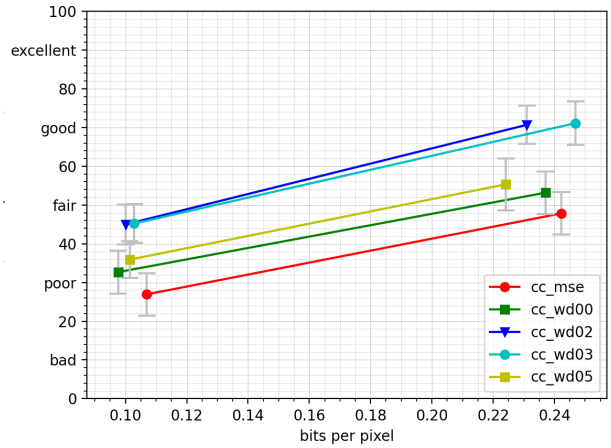

Fig. 3: Subjective test results for different balance factor α

To appreciate the effect of the Wasserstein distance approach in Cool-chic , a preliminary experiment is conducted where the Cool-chic training distortion is chosen as:

$$D = \alpha \|x - \hat{x}\|^2 + \frac{1-\alpha}{200} \mathrm{WD}(x, \hat{x}) \tag{1}$$

Where α balances the mean square error and the Wasserstein distance computed between the target and reconstructed images. When $\alpha = 1$ the *mse* is used, while when $\alpha = 0$ the Wasserstein is solely considered.

The distortion with different α is used to encode the validation set of this edition of CLIC and we target two bit rates for each value: around 0.10 and 0.25 bit per pixel (bpp). We run the experiment with $\alpha = \{0.0, 0.2, 0.3, 0.5, 1.0\}$ and evaluate through a subjective test the average quality of the decoded images.

The raters are asked to rate the quality of the coded items on a scale *bad, poor, fair good and excellent*. The raters can interactively and instantly switch on demand between any coded item and the explicit original image. The scale is continuous where the intervals are 0-20 for bad, 0-40 for poor, 40-60 for fair, 60-80 for good and excellent above. The scores among raters and image clips are averaged and the confidence intervals derived. Figure 3 reports the results.

As it can be noticed that the worse performing solution is the one labeled *mse* where $\alpha = 1$. The version using solely the Wasserstein distance, when $\alpha = 0$, is significantly rated above this pure *mse* version. The best performing balances are for $\alpha = 0.2$ and $\alpha = 0.3$ which appear statistically indistinguishable.

Based on this experiment, we retain $\alpha = 0.2$ for the distortion term used in the encoding process.

IV. Training procedure and results

For the challenge, we select the high operating point (*hop* configuration) of Cool-chic as a baseline for the experiment. For this level of complexity, Cool-chic uses an autoregressive

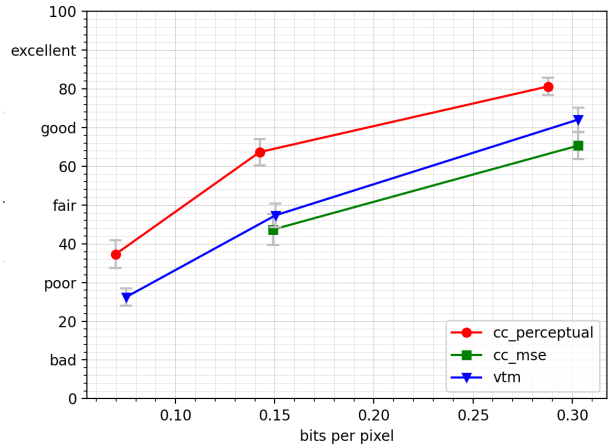

Fig. 4: Subjective test results on the CLIC 2025 test set.

modeling with 2 layers of 16 neurons and it uses 16 context pixels. The synthesis uses a first hidden layer with 48 outputs and feeds the second one that outputs 3 latents. Those 3 latents are subsequently filtered using 2 cascaded convolutional layers giving the reconstructed RGB image.

Compared to the usual Cool-chic *hop* configuration, 7 additional common randomness latents are added, thus the pyramidal latent domain is doubled from 7 to 14. Overall the decoding complexity for this candidate is consequently increased from 1536 to 1728 MAC per pixel.

The training is performed with 30,000 iterations (called the *medium* preset) which gives a good compromise in terms of encoding time versus quality. A set of 14 different rate-distortion balance λ is selected to provide for each clip a range of bit rate largely exceeding the 0.075-0.30 bpp challenge range.

After encoding, the number of bit per pixel for each image is selected to get the best overall distortion (in the sense of the distortion used during training) given the overall bitstream size constraint.

The decoder is identical for the 3 image challenge targets. It is written in C language and is run on CPU to demonstrate the portability of the Cool-chic approach on any device.

In order to verify the submission, a subjective evaluation was organized to compare the proposal with two anchors: the VTM encoder, the reference implementation of VVC (labeled: vtm) and Cool-chic optimised using with the mean square error as distortion (labeled cc_mse).

The 3 testing points are the ones selected for the challenge: 0.075 bpp, 0.15 bpp and 0.3 bpp. The subjective testing methodology follows the one described above. The results are reported in figure 4.

Two main conclusions can be derived from this subjective assessment:

1) Cool-chic optimised using mainly the Wasserstein distance (labelled cc_perceptual) surpasses significantly the 2 other codecs.
2) The proposed approach provides approximately 50% bit saving over the legacy Cool-chic when optimised using the mean square error.

V. Conclusion and Further Work

The purpose of this challenge submission is to demonstrate that a lightweight neural decoder can offer a significant improvement over the latest compression standard such as VVC. This neural decoder can be operated on any device and does not require any particular neural accelerator.

The gap in quality compared to the legacy Cool-chic approach is obtained through the approach presented in [10] where the usage of the Wasserstein distance is motivated along with the addition of a random noise in the latent space. Indeed, the gains obtained are impressive according to the viewers involved in the subjective assessment. It is evaluated that the bit savings is in the order of 50%.

Instead of using a saliency map in the distortion metric we preferred to study solely the effect of this new distortion metric. The usage of a saliency map, as suggested in the cited paper, constitutes an obvious perspective for this work. Also, the incorporation of the VGG neural network used to derive the Wasserstein in the training process leads to a significant increase in the encoding time, being 3 times slower than the usual *mse* optimisation. Accelerating the encoding time is also a dimension of exploration for future developments.

References

[1] Orange Research, "The Cool-chic image and video codec," https://github.com/Orange-OpenSource/Cool-Chic.

[2] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2003*, 2003, vol. 2, pp. 1398–1402 Vol.2.

[3] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," 2018, https://arxiv.org/abs/1801.03924.

[4] D. He, Z. Yang, H. Yu, T. Xu, J. Luo, Y. Chen, C. Gao, X. Shi, H. Qin, and Y.Wang, "PO-ELIC: Perception-oriented efficient learned image coding," in *5th Challenge on Learned Image Compression*, 2022.

[5] D. Li, K. Wang, Y. Bai, X. Liu, and W. Gao, "Local semantic loss and latent refinement for perception-oriented neural compression," Tech. Rep., 2024, https://arxiv.org/abs/2401.14007.

[6] Théo Ladune, Pierrick Philippe, Félix Henry, Gordon Clare, and Thomas Leguay, "Cool-chic: Coordinate-based low complexity hierarchical image codec," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 13515–13522.

[7] Théophile Blard, Théo Ladune, Pierrick Philippe, Gordon Clare, Xiaoran Jiang, and Olivier Déforges, "Overfitted image coding at reduced complexity," in *2024 European Signal Processing Conference (EUSIPCO)*, 2024, https://eurasip.org/Proceedings/Eusipco/Eusipco2024/pdfs/0000927.pdf.

[8] T. Ladune, P. Philippe, G. Clare, F. Henry, and T. Leguay, "Cool-chic: Perceptually tuned low complexity overfitted image coder," Tech. Rep., 2024, https://arxiv.org/abs/2401.02156.

[9] M. Testolina P. Philippe G. Lorand M. Antonini, C. Couvreur and T. Ladune, "Comparative analysis of JPEG AI and Cool-Chic in the binary domain," in *ISO/IEC/JTC1 M108052-ICQ Report*, 2025.

[10] Jona Ballé, Luca Versari, Emilien Dupont, Hyunjik Kim, and Matthias Bauer, "Good, cheap, and fast: Overfitted image compression with wasserstein distortion," 2025, https://arxiv.org/abs/2412.00505.

[11] Yang Qiu, Aaron B. Wagner, Johannes Ballé, and Lucas Theis, "Wasserstein distortion: Unifying fidelity and realism," 2024, https://arxiv.org/abs/2310.03629.