REDEFINING VISUAL QUALITY:
THE IMPACT OF LOSS FUNCTIONS ON INR-BASED IMAGE COMPRESSION

Lorenzo Catania, Dario Allegra

University of Catania
Department of Mathematics and Computer Science
lorenzo.catania@phd.unict.it, dario.allegra@unict.it

ABSTRACT

Implicit Neural Representations (INR) are a novel data representation technique which is gaining ground in the image compression field due to its simplicity and interesting results in terms of rate/distortion ratio. Although a variety of methods based on this paradigm were proposed, limited interest has been given to the analysis of the loss function and the impact of compression artifacts on the visual quality of the reconstructed images, which are mainly due to the adoption of the simple Mean Squared Error (MSE) loss function and to the evaluation done merely in terms of Peak Signal-to-Noise Ratio (PSNR), which do not often correlate with the human perception. In this paper, we evaluate a set of five loss functions in the context of training INRs for image compression, applied to three state-of-the-art architectures, and evaluate their effect on a broader collection of quantitative metrics and the visual fidelity of the decoded images to the originals. The presented outcomes show that the reconstructions obtained by training with some loss functions as MSE suffer from over-smoothing and aliasing artifacts. Our findings reveal that through the employing of a suitable loss function, state-of-the-art architectures quantitatively and qualitatively outperform the results reported in their original papers.

***Index Terms*—** Image compression, Implicit neural representations

1. INTRODUCTION

Lossy image compression has always been widely explored by the academy and the industry. Traditional codecs, such as JPEG [1] and AVIF [2], are the most common way to encode and transmit images due to their high computational and compression efficiency. However, during the last decade, new paradigms based on autoencoders have emerged [3]. These learned methods do not need hand-crafted heuristics, making them easier to design and implement, achieve state-of-the-art results and have also proven to be able to beat traditional codecs in terms of reconstruction quality [4, 5] and have attracted the interest of the historical JPEG committee, that is developing an AI-based standard [6]. Nevertheless, these outstanding compression capabilities require high computational demands. Training these networks may require various days, a consistent amount of computational power and large datasets. On the receiver's end, these complex neural networks must be stored and inference is required to decode the compressed image. Although recent works [7] attempt to mitigate this issue, these limitations make these methods still far from being practical. An emerging paradigm for data representation is Implicit Neural Representations (INR), in which a signal is interpreted as a mapping from coordinates to samples and a neural network is overfitted to this mapping. If the purpose is to compress data, then the weights of the neural networks are compressed and transmitted, then decoded by the receiver. The signal is therefore reconstructed by inference through the neural network. This approach effectively transforms a data compression problem into a model compression one, leading to an alternative viewpoint to the task of encoding signals. This kind of network has proven to efficiently represent both videos [8] and images [9, 10, 11]. When this approach is applied to images, it is often referred to as *Implicit Image Compression* or *Coordinate-based overfitted codec*. A fundamental step when defining an INR pipeline is to choose a proper loss function that represents the distortion between the original signal and the one reconstructed by the network. In the case of images, the most common loss is the L2 mean, also known as *Mean Square Error* (MSE), and compression distortion is commonly evaluated by using the traditional *Peak Signal-to-Noise Ratio* (PSNR). However, these simple metrics may not match the perceived quality of decoded images, therefore most complex metrics that take into account the spatial structure and the characteristics of the human vision system were developed, for instance, *Structure Similarity* (SSIM) [12], *Multi-Scale Structure Similarity* (MS-SSIM) [13] and *Learned Perceptual Image Patch Similarity* (LPIPS) [14]. A strength of learned methods is that it is possible to explicitly optimize perceived quality during training, but most works [9, 11, 15] rely on PSNR for evaluation and overlook perceptual quality, although recent

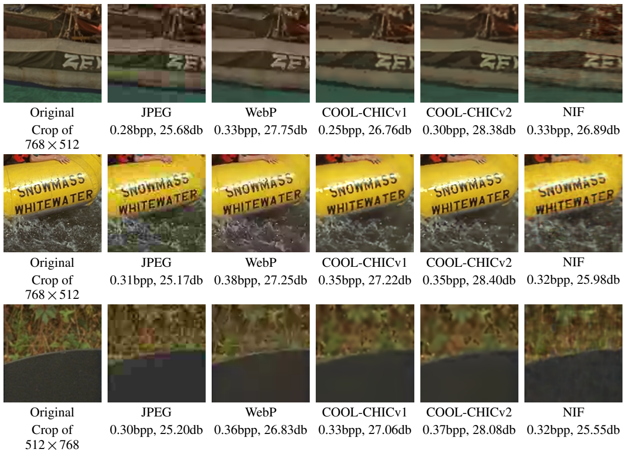


Fig. 1: Samples of distortion artifacts introduced by traditional codecs compared with INR-based methods, the latter do not suffer from blocky and colour aliasing artifacts. Bits-per-pixel and PSNR values are reported below each picture in the same order.

research [10] has shown that INR-based methods can obtain visually smooth and accurate results. These methods do not suffer from well-known noise and block artifacts which are instead common in traditional codecs, as shown in Figure 1.

The purpose of this paper is to evaluate the impact of various loss functions on state-of-the-art INR-based image codecs. The contributions of this research are the following:

- The evaluation of five functions as losses in three SotA INR-based image compression, a paradigm in which the choice of the loss function is fundamental yet nearly unexplored. Results are presented in terms of averaged quantitative results, visual fidelity of specific samples and appearance of artifacts.
- We examine in depth the potential of adding structural factors in loss functions when training INRs for image compression, proposing recommendations that consistently improve the state-of-the-art in terms of perceptive metrics while maintaining a high PSNR. Also, decoded images benefit reduced artifacts and better visual fidelity.
- The code used for the experiments and the full results are publicly released to the community on GitHub [1].

The remainder of this paper is structured as follows. Section 2 presents an overview of the state-of-the-art of the topic. Section 3 describes the formalisms used in the paper, the network architecture, the analysed loss functions and the evaluation metrics used. Section 4 presents the experimental results and quantitatively and visually compares the effects of the five loss functions on the considered methods. Section 5 concludes the paper with some considerations about the impact of the proposals and possible further developments.

2. RELATED WORK

The recent interest in INRs was born by the seminal work of Mildenhall et al. [16] which used a multi-layer perception (MLP) to approximate radiance fields, defining *Neural Radiance Fields* (NeRFs). To solve the well-known spectral bias [17], which is the inclination of neural networks to focus on low-frequency details, they map input coordinates to a positional encoding [18]. Sitzmann et al. [19] proposed an MLP architecture named SIREN, which uses sinusoidal activations with specific initialization schemes and enables learning high-frequency details without any input encoding. The first INR-based codec which concerns the images domain is COIN [9], which uses a simple 16-bit quantization on weights and a naive SIREN architecture and has demonstrated how implicit compression was able to match the ratio of an established codec such as JPEG. Strumpler et al. [20] added positional encoding and a better 8-bit quantization scheme to the COIN pipeline, improving the overall results at the cost of increasing the computational complexity of the method. Catania et al. [10] propose *Neural Imaging Format* (NIF), an INR-based image codec which matches or improves state-of-the-art results of other INR methods by consistently reducing the encoding complexity, bringing execution times from hours to minutes on a single GPU, by using greedy optimization steps during training. In [10], the visual quality of the reconstruction is given special attention, with a structural factor added to the loss function. Recent works such as *(Coordinate-based Low Complexity Hierarchical Image Codec)* (COOL-CHIC) [11] and its extensions [15, 21] replace positional features with learned latent grids [22] as inputs to the MLP, which then require fewer neurons. These methods achieve results comparable to state-of-the-art traditional codecs with a much lower per-pixel decoding complexity concerning other learned methods, but they still exhibit long encoding times, their decoding process is not trivially parallelizable, plus their results in terms of perceptive metrics have not been explored yet. Adding structural information to loss functions has been demonstrated to be effective in works regarding INR-based video compression as well, such as NeRV [8] and further evolutions [23]. In this work, we evaluate the results in terms of quantitative metrics and qualitative results obtained by state-of-the-art architectures such as NIF [10], COOL-CHICv1 [11] and COOL-CHICv2 [15] when using various loss functions, with and without structural factors added.

3. PROPOSED METHODOLOGY

In this section, we outline the formalisms adopted in the paper, then present the chosen network architectures and the

[1] https://github.com/INRAnalysis-ICIP24

loss functions we evaluate on each of those architectures. In the end, we define the metrics used to quantitatively evaluate the results and their correlation with the accuracy and with the perceived quality of the reconstructed signal.

3.1. Functional representation of an image

The idea behind designing a functional image representation that a neural network may fit is to define a mapping from some input features to the image colour for each pixel.

In the case where the network expects positional features, the input features $i_{x,y}$ of a pixel p are calculated by normalizing in the range $[-1, 1]$ its position (p_x, p_y), where $p_x \in \{0, 1, ..., W-1\}$ and $p_y \in \{0, 1, ..., H-1\}$:

$$i_{x,y} = \left(\frac{2p_x}{W-1} - 1,\ \frac{2p_y}{H-1} - 1 \right) \tag{1}$$

Where (W, H) are the width and the height of the image, respectively.

If the method is instead designed to learn L two-dimensional latent grids $g^0, g^1, ..., g^{L-1}$, the input features are given by the following concatenation:

$$(i_{x,y}) = (\hat{g}^0_{x,y}, \hat{g}^1_{x,y}, ..., \hat{g}^{L-1}_{x,y}) \tag{2}$$

Where $\hat{g}^i_{x,y}$ is the value at indices x, y of the i-th upsampled latent grid $\hat{g}^i$, which is calculating by upsampling g^i at resolution (W, H).

The pixel-wise functional representation of an image data I is given by a mapping of the form:

$$I(i_{x,y}) = (R_c, G_c, B_c) \tag{3}$$

Where (i_x, i_y) are the input features associated to the pixel p with RGB values (R_c, G_c, B_c). The purpose of an INR-based method is, then, to propose a network architecture and a training strategy which accurately fits the function I.

3.2. Network architectures

We consider three network architectures, adopting variegated approaches to evaluate the loss functions in a wide scenario.

NIF [10]: A SIREN architecture which takes positional features as input. A modulation module alters the period of each activation based on the coordinates of the pixel. Also, the number of features on each layer is reduced proportionally to its depth. This technique has been empirically proved to enhance the bitrate/distortion ratio [10].

COOL-CHIC v1 [11]: A multi-layer perception with ReLU activations which takes latent grid features as input. The purpose of this architecture is to reduce the decoding complexity of the method limiting the amount of the operations needed to decode each pixel. An auto-regressive probability model is added to estimate the parameters' distribution and an entropy factor is added to the loss function to minimize the parameters' entropy.

COOL-CHIC v2 [15]: An evolution of [11] which adds convolutional layers to the original architecture and adaptive upsampling instead of fixed one to upsample grid features.

3.3. Loss functions

We propose the following five loss functions to be used during training. In the formula, y is the original sample, $\hat{y}$ is the reconstructed sample and N is the number of pixels.

L1: A function calculated as the absolute difference between two signals, also known as *Mean Average Error*.

$$L1(y, \hat{y}) = \frac{\sum |y - \hat{y}|}{N} \tag{4}$$

MSE: The most common loss in INR-based compression.

$$MSE(y, \hat{y}) = \sqrt{\frac{\sum (y - \hat{y})^2}{N}} \tag{5}$$

Compared to L1, this loss function penalizes large errors and is less sensitive to small differences.

LogCosh: It exhibits a shape similar to L1 for large values and MSE for small values, obtaining the best of both worlds. In practice, the following approximation of $log(cosh(x))$ is used to avoid infinite growth for large differences:

$$Lc(y, \hat{y}) = \frac{\sum ((y - \hat{y}) + sp(2 * (y - \hat{y})) - ln(2))}{N} \tag{6}$$

Where sp is the SoftPlus function and is calculated as:

$$sp(x) = ln(1 + e^x) \tag{7}$$

LcSSIM (LogCosh + SSIM): A combination of LogCosh and SSIM [12], first proposed in [10], which aids the training process to consider structural information on the image instead of optimizing each point independently:

$$LcSSIM(y, \hat{y}) = Lc(y, \hat{y}) + \alpha * (1 - SSIM(y, \hat{y})) \tag{8}$$

Where α is a factor which scales the SSIM, as it is usually much bigger than LogCosh and may dominate the loss value.

L1SSIM (L1 + SSIM): A combination of L1 and SSIM similar to the one formulated above:

$$L1SSIM(y, \hat{y}) = (1-\alpha) * L1(y, \hat{y}) + \alpha * (1 - SSIM(y, \hat{y})) \tag{9}$$

In this case, the α factor increases the influence of SSIM on the values and decreases the influence of L1.

3.4. Evaluation metrics

The following evaluation functions are used to quantitatively compare the results for the various combinations of architecture and loss. PSNR (Peak Signal-to-Noise Ratio): This standard metric for compression scales logarithmically compared to MSE. MS-SSIM (Multi-Scale Structural Similarity) [13]: An evolution of SSIM [12] which takes into account the hierarchical structure of the human vision system. It is calculated as a combination of SSIM at various resolutions. It was adopted in previous works about INRs [10] as well. LPIPS (Learned Perceptual Image

Patch Similarity) [14]: A learned metric based on hidden features of classification architectures. Even if no previous works on INR compression adopt this metric, it has been demonstrated to capture distortions which may not affect PSNR or MS-SSIM [24, 5].

4. EXPERIMENTAL RESULTS

4.1. Dataset and settings

The following experiments were run on the Kodak [25] dataset, a historical standard for image processing. Although more recent datasets are available, this is the most commonly used for INR-based compression assessment as those methods are characterised by long encoding times, which make it difficult to widely test them on high-resolution images [10]. NIF and COOL-CHICv1 were trained with the faster configurations available for every bitrate, performing between 1000 and 1500 optimization steps, while COOL-CHICv2 was trained using the *faster* preset provided by the authors. Although better results can be obtained by using slower presets, these fast training methods achieve a sufficiently good rate/distortion ratio for our purposes. The α factor is set to 0.01 for LogCosh+SSIM and to 0.2 for L1+SSIM. In line with previous works [3, 10], MS-SSIM values are normalized as $-10log_{10}(1-MS\text{-}SSIM)$ to better visualize small differences in the plots. All comparisons are made in the RGB colorspace.

4.2. Quantitative results

The following paragraphs discuss the various results obtained when replacing the original loss function of each architecture with our proposed ones. The original loss function is reported in green, while losses with and without a structural factor are reported in red and blue respectively. The marker shape distinguishes between the point-wise distortion function adopted (L1, MSE and LogCosh). The arrow beside the metric's name indicates if the metric should be maximized (up-arrow) or minimized (down-arrow). Two traditional codecs, AVIF [2] and JPEG [1] are reported as baselines using dashed lines. Note that, especially on COOL-CHIC architectures, different loss functions may obtain different ranges of bits-per-pixel. This is because in those methods the rate/distortion ratio is controlled by a λ factor which balances distortion and latent parameters entropy. As the various losses have different magnitudes on average, the same λ parameter obtains different average bits-per-pixel values. In our experiments, we stick to the default MSE-tuned λ values.

4.2.1. COOL-CHICv1

Results for COOL-CHICv1 are reported in Figure 2a. In terms of PSNR, the original MSE loss obtains the best results, although LogCosh performs similarly. That's the awaited outcome as not only is PSNR calculated based on the MSE value, but this is also the metric used for evaluations in the original paper, therefore it is straightforward that the adopted loss function is the one which maximizes the results in that setting. Regarding MS-SSIM, every loss achieves similar results, although these with a structural factor achieve slightly better values. Finally, by using the original loss function this method is outperformed by JPEG in terms of LPIPS. However, losses with an SSIM factor obtain better results and compare or improve over JPEG at every bitrate. In summary, COOL-CHICv1 is always outperformed by AVIF.

4.2.2. COOL-CHICv2

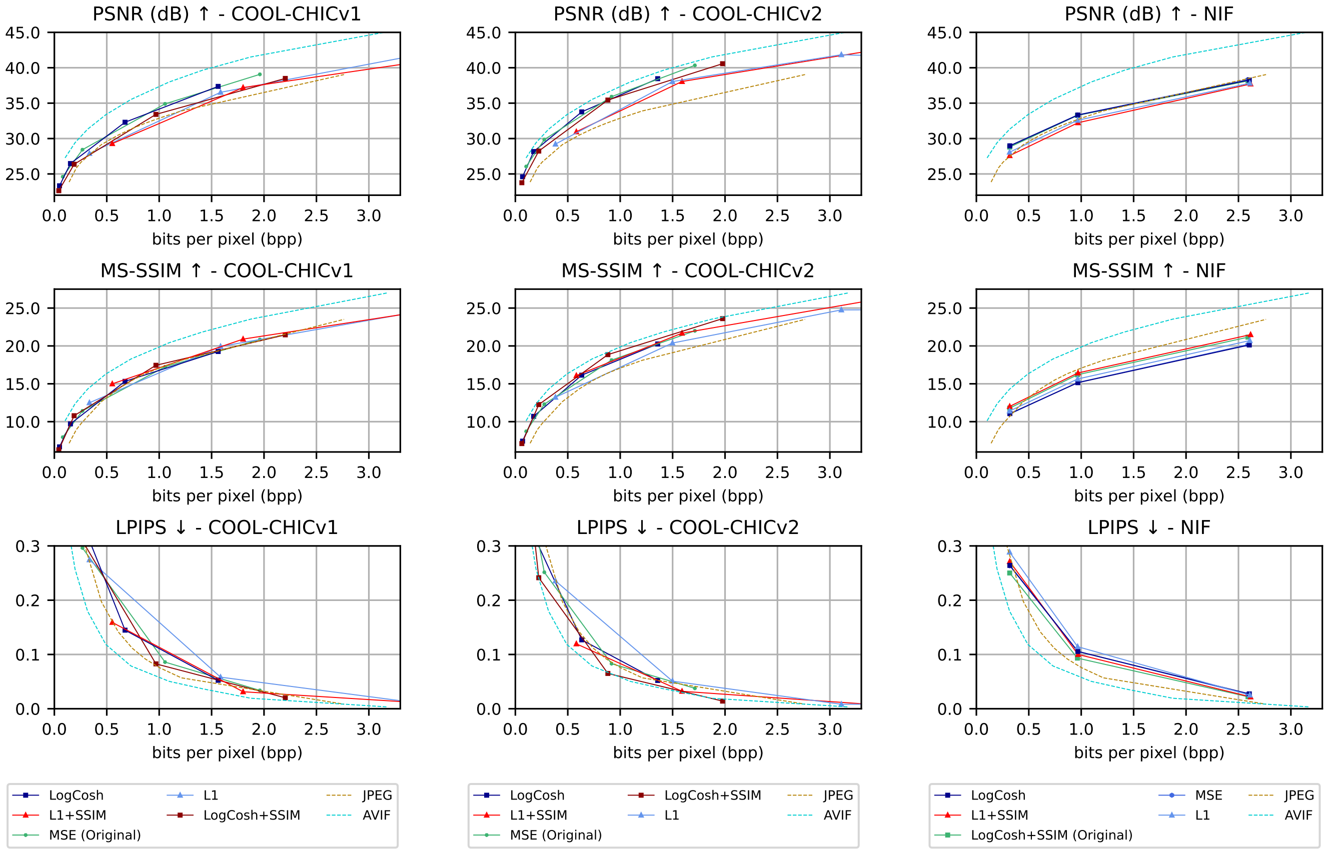
Analogously to the COOL-CHICv1, the originally adopted MSE loss is the best performing in terms of PSNR for COOL-CHICv2 in figure 2b, and the observations done in the previous paragraph hold here as well. In terms of MS-SSIM, losses with an SSIM factor consistently improve over the original MSE loss. In particular, training with LogCosh+SSIM obtains comparable results with AVIF between 0.9 and 2.0 bits-per-pixel, while MSE loss is substantially outperformed. Similarly, structural losses obtain similar LPIPS values compared to AVIF while the original loss function is not able to. Considering that the reconstruction metrics may be overall increased by using slower presets, these results demonstrate that INR-based methods can outperform well-established codecs traditional codecs when appropriately tuned.

4.2.3. NIF

Figure 2c reports the results for NIF, which adopts LogCosh+SSIM as the default loss. In terms of PSNR, the original loss function seems to be the best-performing, although removing the structural factor slightly increases the results. Results are similar for MS-SSIM, although the combination L1+SSIM obtains slightly better results. Regarding LPIPS, most losses obtain better results than JPEG at low bits-per-pixel but are outperformed at higher bitrates, and the original LogCosh+SSIM loss is the best performer. Overall, NIF is not able to reach the same performance as AVIF, but it is fair to point out that its purpose is to provide faster encoding times compared to other INR-based methods.

4.3. Qualitative comparisons

Figure 3 reports some samples to visually compare the results of INR-based models trained with the proposed set of loss functions, along with standard hand-crafted codecs such as JPEG and WebP and the modern JPEG XL and AVIF formats. JPEG and WebP reconstructions both suffer from block artifacts and blurry effects. JPEG XL reconstructions are blurry overall, while AVIF ones are smoother but the codec tends to miss finer details, such as the building decoration



(a) COOL-CHICv1 results

(b) COOL-CHICv2 results

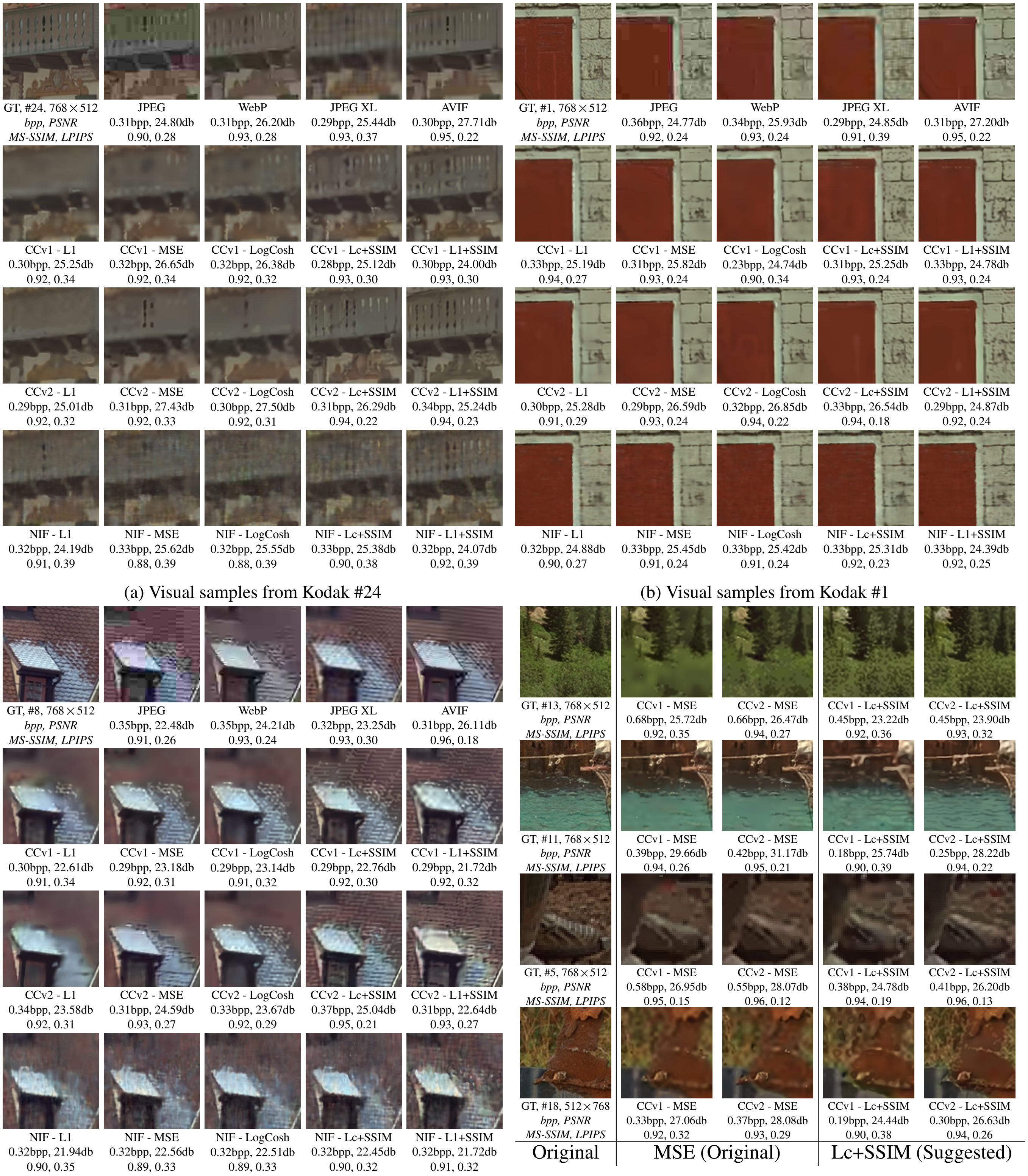
(c) NIF results

Fig. 2: Quantitative results of the evaluated loss functions applied to the considered architectures. The original loss is reported in green, while red and blue blends are used for point-wise losses and losses with an SSIM factor respectively. Dashed lines are used for traditional codecs reported as baselines.

on the bottom in sample #24 (Figure 3a). COOL-CHIC reconstructions exhibit distorted edges when trained without our proposed structural loss factor, which instead mitigates this issue. Still, some artifacts represented by bright lines may appear on edges and color aliasing may be present. These are especially present on COOL-CHICv1, such as in the bottom part of Kodak #24, and on COOL-CHICv2 in the same image when encoding with L1+SSIM. NIF reconstructions do not suffer from such artifacts but are characterised by the presence of high-frequency noise, which is presumably due to the positional encoding applied to the input coordinates to increase the frequency spectrum to which the network is sensible. However, NIF occasionally better represent fine lines such as the door details on Kodak #1 (Figure 3b), which are sometimes smoothed by COOL-CHIC methods. On Kodak #8 (Figure 3a), COOL-CHIC methods are not able to reconstruct the roof details when trained with MSE, but our proposal of adding an SSIM factor to the loss solves this issue. This is evident in Figure 3d, where in some patches COOL-CHIC methods trained with their default loss, which is MSE, smooth away important image segments such as vegetation, sea waves and ground details. Our proposed LogCosh+SSIM loss, instead, produces more fidel reconstructions and obtains similar or better results in terms of MS-SSIM and LPIPS.

5. CONCLUSION

In this paper, we have evaluated a set of five loss functions for INR-based image compression. These losses were applied to the training of three different state-of-the-art methods and their impact on the encoding process has been evaluated. We have given special attention to the differences between the standard MSE loss and our proposals with an SSIM factor. The presented results show that, in a dataset which is small but still a reference for INR-based compression, our approach achieves better results in terms of both visual fidelity and perceptive metrics, while not considerably affecting the PSNR. Further explorations of more complex structural losses should be done, as these novel image compression methods must produce visually pleasing results to be adopted in practice.



(a) Visual samples from Kodak #24

(b) Visual samples from Kodak #1

(c) Visual samples from Kodak #8

(d) Visual samples of artifacts introduced by training COOL-CHIC models with MSE, with alternatives obtained by using the LogCosh+SSIM loss function.

Fig. 3: Visual comparisons on various Kodak image details. COOL-CHIC models are reported as "CC" and LogCosh+SSIM loss is reported as ”Lc+SSIM" for conciseness. The uncompressed crop is reported in the top-left corner of each sequence, along with the resolution of the full image, while bits-per-pixel, PSNR, MS-SSIM and LPIPS values are reported below each crop in this same order.

6. REFERENCES

[1] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, 1992.

[2] N. Barman and M. G. Martini, "An evaluation of the next-generation image coding standard AVIF," in *International Conference on Quality of Multimedia Experience*, 2020.

[3] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[4] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-cnn architectures," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[5] E. Agustsson, D. Minnen, G. Toderici, and F. Mentzer, "Multi-realism image compression with a conditional generator," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[6] J. Ascenso, E. Alshina, and T. Ebrahimi, "The jpeg ai standard: Providing efficient human and machine visual data consumption," *IEEE MultiMedia*, 2023.

[7] Y. Yang and S. Mandt, "Computationally-efficient neural image compression with shallow decoders," in *IEEE International Conference on Computer Vision*, 2023.

[8] H. Chen, B. He, H. Wang, Y. Ren, S. N. Lim, and A. Shrivastava, "NeRV: Neural representations for videos," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[9] E. Dupont, A. Goliński, M. Alizadeh, Y. W. Teh, and A. Doucet, "COIN: Compression with implicit neural representations," 2021. [Online]. Available: https://arxiv.org/abs/2103.03123v2

[10] L. Catania and D. Allegra, "NIF: a fast implicit image compression with bottleneck layers and modulated sinusoidal activations," in *ACM International Conference on Multimedia*, 2023.

[11] T. Ladune, P. Philippe, F. Henry, G. Clare, and T. Leguay, "COOL-CHIC: Coordinate-based low complexity hierarchical image codec," in *IEEE International Conference on Computer Vision*, 2023.

[12] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, 2004.

[13] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems, and Computers*, 2003.

[14] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[15] T. Leguay, T. Ladune, P. Philippe, G. Clare, F. Henry, and O. Déforges, "Low-complexity overfitted neural image codec," in *IEEE International Workshop on Multimedia Signal Processing*, 2023.

[16] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision*, 2020.

[17] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in *International Conference on Machine Learning*, 2019.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017.

[19] V. Sitzmann, J. N. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Neural Information Processing Systems*, 2020.

[20] Y. Strümpler, J. Postels, R. Yang, L. V. Gool, and F. Tombari, "Implicit neural representations for image compression," in *European Conference on Computer Vision*, 2022.

[21] H. Kim, M. Bauer, L. Theis, J. R. Schwarz, and E. Dupont, "C3: High-performance and low-complexity neural compression from a single image or video," 2023. [Online]. Available: https://arxiv.org/abs/2312.02753

[22] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics*, 2022.

[23] Y. Bai, C. Dong, C. Wang, and C. Yuan, "PS-NeRV: Patch-wise stylized neural representations for videos," in *IEEE International Conference on Image Processing*, 2023.

[24] R. Yang and S. Mandt, "Lossy image compression with conditional diffusion models," *Advances in Neural Information Processing Systems*, 2024.

[25] Kodak. (1999) Lossless true color image suite. [Online]. Available: https://r0k.us/graphics/kodak/