Comparative Analysis of Subjective Evaluations for Traditional and Neural-Based Video Enhancement Techniques

Darren Ramsook[1], Vibhoothi[2], Anil Kokaram[3]
Sigmedia Group
Trinity College Dublin
Dublin, Ireland
ramsookd@tcd.ie[1], vibhootv@tcd.ie[2], anil.kokaram@tcd.ie[3]

Angeliki Katsenou[4], David Bull[5]
Visual Information Laboratory
University of Bristol
Bristol, England
angeliki.katsenou@bristol.ac.uk[4], dave.bull@bristol.ac.uk[5]

***Abstract*—This work evaluates the effectiveness of modern video restoration methods, contrasting neural network-based techniques with traditional statistical algorithms to improve perceived video quality. Our analysis focused on three distinct methods: VBM4D, CVEGAN, and Ramsook, assessing their performance using pairwise subjective assessments with a compressed baseline. Results indicate a significant disparity between objective and subjective evaluations, with traditional methods like VBM4D showing limited improvements in perceptual quality, as demonstrated by a statistically non-significant increase in Mean-Opinion-Score (MOS). In contrast, the neural-based methods, CVEGAN and Ramsook, showed statistically significant improvements in subjective video quality. The findings highlight the superior capability of neural approaches to enhance perceptual quality, suggesting that current objective metrics may not fully capture quality as perceived by human observers. This study also contributes the results of the comparative analysis and the dataset to the research community.**

***Index Terms*—Subjective analysis, video restoration, perceptual criteria**

I. Introduction

The evolution of video compression artifact suppression algorithms has been a key to enhancing existing compression and transmission pipelines for visual media. Restoration algorithms can be classified as either hand-crafted statistical approaches or learning-based. Hand-crafted algorithms for compression artifact reduction [1], [2] refer to methodologies that have been meticulously designed through analytical processes, leveraging insights into the characteristics of images and videos as well as the human visual system.

In contrast, learning-based approaches, particularly those employing Convolutional Neural Networks (CNNs), present a paradigm shift in compression artifact reduction [3]–[5]. Unlike traditional algorithms that typically address one type of artifact at a time, neural networks are capable of learning to handle multiple artifact types simultaneously. This holistic understanding allows them to more accurately model the complex interplay of artifacts and their impact on perceived video quality. Learning-based methods have already outperformed traditional ones in general restoration tasks [6]–[8].

Traditional methodologies exhibit a notable limitation related to their foundational design principles [9]. Particularly, most methods employ simple loss functions, such as variants of the L_p norm (e.g., MSE). Although these techniques lead to increased PSNR, they inadvertently induce an oversmoothing effect, filtering out the natural texture and details of the original video. Thus, they do not correlate well with human scores [10].

However, some research efforts [4], [11] have focused on designing loss functions that are more aligned with human visual perception. This shift from pixel-wise error minimization to perceptual quality enhancement enables CNNs to generate content that not only score higher on quantitative metrics but also offer improved visual satisfaction. Taking all the above into consideration, this paper explores the complex relationship between traditional objective metrics and perceived human quality scores, highlighting the inconsistency and proposing avenues for more effective measures. A visual example of the outcome of the difference in optimization criteria is shown in Figure 1, where the restored version with most visible detail ranks the lowest with respect to PSNR. Furthermore, this paper contributes to the field by conducting a subjective study that compares the latest neural-based video restoration methods with traditional statistical hand-crafted methods. Our aim is to evaluate the impact of these methods on perceived video quality, thereby offering insights that are beyond the limitations of conventional metrics. Our specific contributions made are as follows:

- A dataset of compressed and restored videos and associated subjective and objective metrics for three different methods: VBM4D [12], CVEGAN [5] and Ramsook et al. [11].
- A detailed statistical analysis of the subjective scores obtained for each video restoration algorithm that is the basis for a comparative analysis of the performance, in terms of subjective scores, of these algorithms.

(a) VBM4D (31.1, 60.2) (b) CVEGAN (31.2, 66.7) (c) RAMSOOK (29.6, 67.2) (d) Compressed (30.9, 62.6) (e) Reference

Fig. 1: *Compressed (d) and restored (a-c) patches of (e). The values reported are in the format of (PSNR/db, VMAF). These results show inconsistencies between a traditional objective metric such as PSNR and VMAF. Based on VMAF, Ramsook (c) ranks the highest, while based on PSNR, CVEGAN (b) ranks the highest.*

II. Background

While it would be easier to individually model compression artifacts, and then propose solutions for each type, in reality compression artifacts are a combination of different artifacts. This means that an artifact-specific solution would not always be the best one. One of the most robust hand-crafted algorithms for video restoration is VBM4D [12]. VBM4D is an advanced video denoising technique designed to remove noise from video sequences while preserving important details and structures. It builds on concepts introduced by its predecessor, the BM3D (Block-Matching and 3D Filtering) [13] algorithm, which was highly successful in image denoising. The VBM4D algorithm extends these principles into the temporal domain, making it possible to effectively reduce noise across both spatial and temporal dimensions of video data. The VBM4D algorithm uses Wiener filtering [14] and, therefore, has an optimization criterion based on pixel-wise mean squared error.

A promising learning-based approach is CVEGAN [5], which incorporates a mixture of multiple objective metrics including SSIM and MSSIM in its loss function. This approach includes the direct comparison of complete deep features between a degraded-reference pair, similar to [15]. This model shows substantial improvement in PSNR and VMAF. CVEGAN uses a Relativistic-GAN [16] as the basis of their adversarial training setup.

Another interesting approach is presented by Ramsook et al. [11]. This approach uses a Wasserstein-GAN [17], which consists of a neural perceptual critic network. This critic network has a pre-processing step in which the restored content is passed through a pre-trained classification network to generate deep features. Selected deep features that correlate well with perceptual criteria are used by the critic in calculating a score of the restored content. This score is used within the optimization process of the restoration network.

Recent research in measurement of the subjective quality of neural based approaches trained on different optimization criteria [18], demonstrated that participants preferred content that was restored using either DISTS [19] or MS-SSIM [20] as the optimization criteria. This indicates that structural based metrics have a higher correlation with the human visual system. However, this study was performed on still images and did not consider traditional statistical restoration algorithms.

III. Evaluation

A. Dataset

A corpus of 10 clips with a spatial resolution of 1920×1080 was selected from different datasets as described in [21]. Each clip has a duration of 5 seconds with a playback framerate of 24 fps. The clips were randomly selected from spatial and temporal information (SI/TI) histograms to cover varying combinations of SI/TI. The formula for the computations of SI/TI is detailed in ITU-T P.910 [22]. This ensured that the clips used for subjective testing represented a diverse span of content complexity. Each clip was compressed using the libx265 encoder, which is an open-source implementation of the HEVC standard [23]. The clips were compressed at two quantization parameter (QP) values, 32 and 37, which are QP values within the range of the common testing conditions by JVT-VC [24]. Each clip was then restored using three restoration algorithms: VBM4D [12], CVEGAN [5] and Ramsook et al. [11]. The sequences used can be accessed via our GitHub repo [21].

B. Environmental Setup

The study was conducted in the subjective quality testing room in the Visual Information Laboratory at the University of Bristol. The environment was set-up to comply with ITU-BT.500-14 [25] simulating a living room. The television used was 60 inch 4k/UHD, with the content being shown on screen at original resolution with no upsampling applied. The ratio of the illuminance of the background behind the television and the inactive television was kept at 0.02. The ratio of the illuminance of the background behind the television and the peak luminance of the television was kept at 0.14. The distance of the viewer to the content was set to $1.6H$, where H is the height of the 1080p content being viewed on the 4K television without upsampling.

C. Experimental Setup

Given our dataset of 10 clips with 2 QP points per clip and restored with either three methods, this results in a

TABLE I: *Table of results from statistical tests and collected metrics. Both tests compare the null hypothesis: scores between compressed version and restored version are equal, and the alternative hypothesis: scores from the restored version are higher than the compressed version. The Thurstone MLE and BT MLE are estimated from results obtained from Experts.*

		Paired t-test		Wilcoxon Ranked Sum test		Thurstone MLE	BT MLE	PSNR	VMAF
	Average Δ MOS	p-value	Reject Null?	p-value	Reject Null?				
VBM4D [12]	0.394	0.201	No	0.243	No	1.4e-6	3.73	32.77	68.53
CVEGAN [5]	0.816	0.013	Yes	0.011	Yes	5.1e-7	3.84	32.85	76.21
RAMSOOK [11]	6.17	1.45e-31	Yes	6.8e-32	Yes	2.6e-7	3.75	30.89	78.41
COMPRESSED								31.79	72.16

total of 60 trials per person. For our experiment we follow the Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) method as described by ITU-BT.500-14 [25]. The experiment consists of showing the subject a restored version and a compressed version which is played three times. The subjects are then presented with continuous scales ranging from 0-100 where they are asked to rank the quality of each video. Before each experiment, the subjects undergo a training exercise where they are given a brief explanation of the system. The first two rounds are for training purposes with no recordings of the scoring. Overall, 62 pairs of videos are presented per person. Each clip is played for 15 sec (3 times), and the user inputs the score on average within 12 seconds. In addition to that, there is a three second gray screen that is displayed between trials for eye calibration. This results in a total time of ≈30 min per subject in the experiment.

D. Participants & Data Processing

In total 25 subjects participated in the experiment (with a median age of 28), with 15 being classified as "Experts", as they have some prior experience with compression artifacts in image or video, and 10 as "Non-experts". Consent forms were signed by the subjects and the results were anonymized. The collected scores are available for general access here [21].

IV. Results & Discussion

The presented results from the comparison of three video processing methods—VBM4D [12], CVEGAN [5] and Ramsook [11] are shown in Table I. These results are summarized using paired t-tests, Wilcoxon ranked-sum tests, and Mean Opinion Score (MOS) differences, alongside the objective quality metrics Peak Signal-to-Noise Ratio (PSNR) and Video Multimethod Assessment Fusion (VMAF). The Thurstone and Bradley-Terry MLE estimates were also calculated for results obtained by experts through the use of Sureal [26].

VBM4D displayed the smallest improvement with a Δ MOS of 0.394. These MOS improvements were not statistically significant, as indicated by both paired t-tests (p-value = 0.201) and Wilcoxon tests (p-value = 0.243). In stark contrast, CVEGAN demonstrated a more substantial increase in Δ MOS of 0.816, with significant p-values in both the paired t-test (0.013) and Wilcoxon test (0.011), decisively rejecting the null hypothesis. Finally, Ramsook showed the most significant enhancements, achieving an average Δ MOS of 6.17. This improvement was supported by strikingly significant statistical results, with p-values approaching zero in both tests, underscoring the method's effectiveness.

These results reveal that while VBM4D did not significantly enhance video quality over baseline compression, both CVEGAN and Ramsook markedly improved the perceptual quality of the same videos. The disparity and inconsistency between objective metrics and subjective scores (see Table I) is highlighted by these statistical tests, emphasizing the lack of a consistent relationship between Peak Signal-to-Noise Ratio (PSNR) and subjective assessments. Notably, the variance in subjective scores for Ramsook is much higher than that observed for CVEGAN and VBM4D. However, both the mean and median improvements are significantly greater with Ramsook, as illustrated in Figure 2.

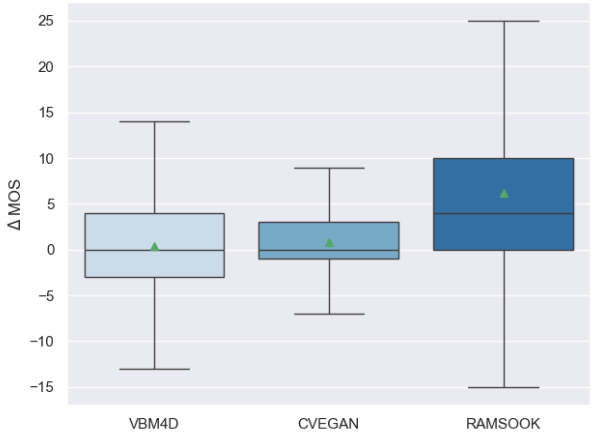

Fig. 2: *Box-plots of Δ MOS between compressed and restored content. Triangles indicate the mean change in MOS.*

Further analysis using Thurstone and Bradley-Terry Maximum Likelihood Estimation (BT MLE) models suggests that CVEGAN has the higher perceived change in quality, likely due to the larger variance in the subjective scores associated with the Ramsook method.

Overall, this analysis conclusively demonstrates that advanced neural processing techniques like CVEGAN [5] and Ramsook [11] can significantly enhance the visual quality of compressed video, offering compelling evidence of their ability to optimize for perceptual criteria that is not captured by mean squared error optimized algorithms.

REFERENCES

[1] H. C. Reeve and J. S. Lim, "Reduction of blocking effect in image coding," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1983.

[2] K. Fukuda and A. Kawanaka, "Reduction of blocking artifacts by adaptive dct coefficient estimation in block-based video coding," *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, vol. 3, pp. 969–972 vol.3, 2000.

[3] C. Dong, Y. Deng, C. C. Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *Proceedings of the IEEE international conference on computer vision*, pp. 576–584, 2015.

[4] F. Zhang, C. Feng, and D. R. Bull, "Enhancing vvc through cnn-based post-processing," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2020.

[5] D. Ma, F. Zhang, and D. R. Bull, "Cvegan: a perceptually-inspired gan for compressed video enhancement," *arXiv preprint arXiv:2011.09190*, 2020.

[6] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, "Deep learning on image denoising: An overview," *Neural Networks*, vol. 131, pp. 251–275, 2020.

[7] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8160–8168, 2019.

[8] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

[9] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[10] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," in *2012 19th IEEE International Conference on Image Processing*, pp. 1477–1480, 2012.

[11] D. Ramsook and A. Kokaram, "Learnt deep hyperparameter selection in adversarial training for compressed video enhancement with a perceptual critic," in *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 2420–2424, 2023.

[12] M. Maggioni, G. Boracchi, A. Foi, and K. Egiazarian, "Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms," *IEEE Transactions on image processing*, vol. 21, no. 9, pp. 3952–3966, 2012.

[13] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[14] A. C. Kokaram, *Noise Reduction for Image Sequences*, pp. 241–260. London: Springer London, 1998.

[15] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1905–1914, 2021.

[16] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," 2018.

[17] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017.

[18] S. Mohammadi and J. Ascenso, "Perceptual impact of the loss function on deep-learning image coding performance," in *2022 Picture Coding Symposium (PCS)*, pp. 37–41, IEEE, 2022.

[19] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 5, pp. 2567–2581, 2020.

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

[21] "Project github repo." https://github.com/DarrenR96/subjectivetestDataset, 2024.

[22] ITU-T, "Subjective video quality assessment methods for multimedia applications." Recommendation of ITU-T, 07 2022.

[23] "Hevc." https://www.itu.int/rec/T-REC-H.265. Accessed: 2024.

[24] "Hevc common testing conditions." http://phenix.it-sudparis.eu/jct/doc_end_user/current_document.php?id=7281. Accessed: 2024.

[25] ITU-R, "Bt.500-14: Methodologies for the subjective assessment of the quality of television images," tech. rep., International Telecommunication Union, 2019.

[26] Netflix, "Sureal." https://github.com/Netflix/sureal, 2024. Accessed: 2024.