IBC2024

QUANTITATIVE ASSESSMENT OF FILM GRAIN SIMILARITY: AN OBJECTIVE MODEL

M. Jiang, Z. Li, J. Sapra, H. Yeganeh

(mjiang@imax.com, zli@imax.com, jsapra@imax.com, hyeganeh@imax.com)

IMAX Streaming and Consumer Technology (SCT), Canada

ABSTRACT

Digital cinematography has advanced, yet many artists prefer film rolls for their distinctive texture and essence, with film grain being integral to their artistic expression. However, Over-the-Top (OTT) providers and streamers face challenges with this high-entropy signal, unfriendly to compression due to limited bandwidth. Preserving film grain involves removing it at the source and resynthesizing it post-decoding, a strategy supported by codecs like AV1 and VVC, albeit potentially compromising grain fidelity. Our subjective studies, presented at IBC 2023 (1), examined existing film grain synthesis methods, revealing shortcomings in replicating the original grain appearance. We advocate for a perceptual approach to assessing film grain synthesis quality, emphasizing subjective evaluation's complexity. We propose an objective film grain similarity model using a data-driven approach, which aligns closely with human perception, demonstrating a high correlation with subjective studies. This metric optimizes auto-regression film grain synthesis parameters, resulting in faithful replication of the original film grain, as confirmed by subsequent subjective studies.

INTRODUCTION

Film grain, a distinctive texture resulting from the random distribution of silver halide crystals in traditional film photography, plays a crucial role in the visual aesthetic of films. It adds depth, texture, and authenticity, allowing directors to control brightness, contrast, and mood. Moreover, it contributes to the overall look and feel of a film, enhancing certain elements and drawing attention to specific areas of the frame. Film grain removal and synthesis are integral to video encoding systems, facilitating bandwidth savings while maintaining perceptual quality. This process involves estimating and removing grain from the source video before encoding, then reintroducing it after decoding based on estimated grain statistics (see Figure 1).

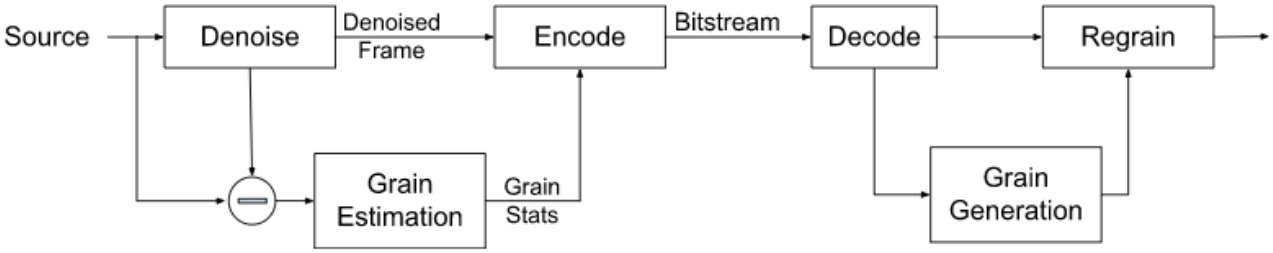

Figure 1- Framework for a grain-aware video encoding system

In modern codecs, Auto-Regressive (AR) and Frequency Filtering methods have been adopted to synthesize film grain after decoding, utilizing grain statistics estimated by denoiser filters on the encoder side. These methods were chosen primarily for their implementation efficiency, although their perceptual quality requires examination. Numerous academic studies have focused on generating film grain that closely resembles true film grain. For instance, Oh et al. (2) described an advanced method for film grain extraction and synthesis, while Dai et al. (3) proposed temporal filtering for grain removal before encoding and subsequent synthesis using an autoregressive model. Other approaches, such as those by Hwang et al. (4) and Newson et al. (5), introduced variations to enhance grain removal and rendering realism. Meanwhile, Ameur et al. (6) employed deep convolutional neural networks for realistic film grain synthesis. Despite these advancements, evaluating film grain quality remains challenging, as most methods rely on implicit metrics or visual inspections rather than typical full-reference metrics, which are often inadequate due to the nature of synthesized film grain. This limitation arises from the nature of full-reference quality metrics such as SSIM, PSNR, and VMAF, which rely on pixel-to-pixel measurements. However, in the case of synthesized film grain, there is no assurance that the instantiation of grain will occur in the same locations across frames. Consequently, there is a pressing need for an objective perceptual quality metric. Our work addresses this need by presenting the IMAX FGS metric, an AI-based approach developed to accurately assess film grain similarity. Extensive validation demonstrates its effectiveness, performing comparably to an average expert viewer in evaluating film grain similarity.

SUBJECTIVE EXPERIMENTS AND BASELINE EVALUATION

Utilizing an objective metric to evaluate film grain similarity serves multiple purposes, including optimizing various components within film grain synthesis frameworks and fine-tuning parameters and configurations in workflows to preserve creative intent. Modern codecs like AV1 or VVC exemplify such frameworks. Without a reliable objective metric assessing film grain similarity between "as graded" and "as delivered" content, the design process often relies on expert viewers whose recommendations serve as validation tools. For instance, the absence of a film grain similarity metric necessitates expert evaluation of the impact of different parameters in AR film grain synthesis methods. Therefore, the ultimate criterion for approving and employing a film grain similarity metric lies in its ability to match the expertise of an average expert viewer.

Before designing any metric, it's essential to determine if there exists a metric that can serve as an average subject for assessing film grain similarity. To achieve this, we conducted a comprehensive subjective study. This approach offers several benefits. Firstly, it allows us to evaluate the accuracy of existing metrics that could potentially serve as film grain similarity metrics. Secondly, it provides insights into how expert viewers perceive and evaluate the look and feel of film grain. Thirdly, it enables us to create a unique dataset that can be utilized

in designing a new film grain similarity metric if none of the baseline models meet the success criteria, which is to replicate the judgment of expert viewers.

In our study, we utilized ten UHD HDR sources featuring authentic film grain, previously screened in IMAX theaters. To generate multiple versions of synthesized grain, we employed the AOMedia Film Grain Synthesis 1 (AFGS1) approach (7) categorized as an Auto-Regressive method. AFGS1 is also employed in the AV1 codec (8), providing an end-to-end solution for removing grain from the source, encoding the content with suppressed grain, and subsequently decoding and adding the synthesized grain. This methodology allows us to showcase the practicality of utilizing a film grain similarity metric. Using these ten source files (referred to as the Source Reference Circuit, or SRC), we generated nine test sequences (referred to as the Hypothetical Reference Circuit, or HRC). Below are the details of each HRC.

To generate HRC 1, we initially employed the IMAX Digital Media Remastering (DMR) tool to partially suppress film grain. Subsequently, we encoded the source with reduced grain using the SVT-AV1 implementation of the AV1 codec at a bitrate of 20Mbps. We intentionally selected a relatively high bitrate to minimize compression artifacts. In HRC 1, we did not utilize any AV1 film grain capabilities and treated AV1 solely as an encoder. This was achieved by setting the FGS configuration to 0 in SVT-AV1. However, we still embedded AFGS1 film grain metadata using the grav1synth tool (9). Eventually, we utilized a customized dav1d decoder to decode the AV1 stream with the synthesized grain. The customized dav1d decoder closely resembles the standard dav1d decoder (10) but employs a 4x4 block size during grain synthesis. Smaller block size helped alleviate the repetitive pattern in the synthesized film grain. In this process, the AFGS1 metadata was manually generated. Specifically, an IMAX expert viewer manually tuned the AFGS1 parameters to achieve the most similar grain look and feel to that of the original source. Figure 2 depicts the outlined framework for generating HRC 1.

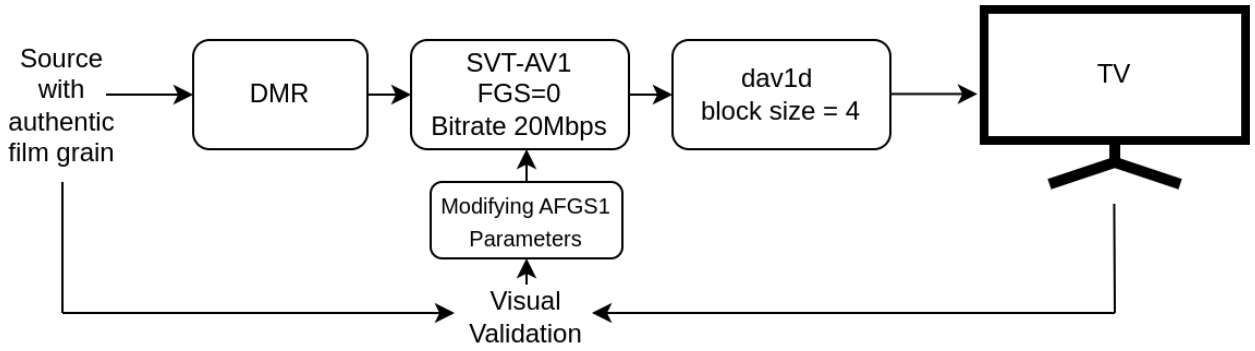

Figure 2 - Framework to generate HRC 1

For HRC 2 generation, we employed the pipeline outlined in HRC 1, with a single modification. Utilizing the optimal AFGS1 configuration derived during the HRC 1 generation process as a foundation, we generated approximately 240 distinct versions of synthesized grain by adjusting various AFGS1 parameters. Specifically, we varied parameters such as GrainScaling, GrainScaleShift, and AR_Coeff. This pool encompassed the "best" AFGS1 parameters identified during the generation of HRC 1. Ultimately, we employed an internal metric to choose the version most closely resembling the source.

HRC 3, 4, 5, and 6 follow a comparable approach to HRC 2 in their generation, albeit utilizing different baseline models. HRC 3 is generated using a texture similarity model proposed in (11). For HRC 4 and 5, universal quality metrics recommended in (12) and (13) are employed, respectively. Additionally, HRC 6 is generated by selecting the synthesized grain

closest to that of the source, employing a texture similarity measure outlined in (14). It's important to highlight that we opted not to utilize any existing full-reference quality metrics due to their inability to accurately assess the similarity between synthesized grain and the original source. As mentioned previously, the FR metrics are inadequate for evaluating film grain similarity, prompting the selection of alternative assessment approaches such as texture similarity methods.

In HRC 7, we employed the AV1 codec with film grain synthesis capability. Specifically, we utilized the SVT-AV1 implementation and configured the FGS parameter to be 35, which dictates the denoiser level in SVT-AV1. It's important to note that this setup deviates from the standard AV1 configuration, as we utilized a 4x4 block size during decoding with dav1d. HRC 8 represents the conventional end-to-end SVT-AV1 process, incorporating an FGS parameter of 35. During decoding, we utilized dav1d with the default block size of 32x32. Figures 3 and 4 depict the pipelines used in generating HRC 7 and HRC 8, respectively. Finally, HRC 9 was generated using the DMR version of the source.

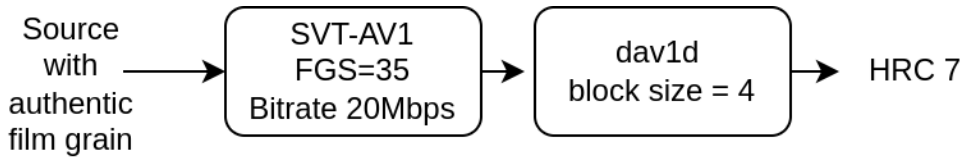

Figure 3 – Pipeline to generate HRC 7

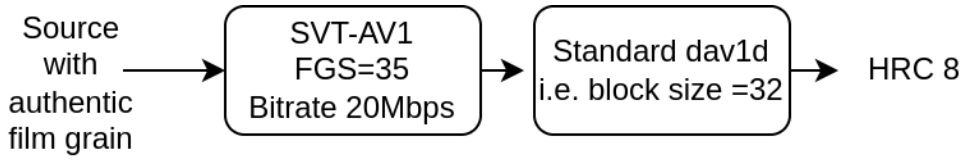

Figure 4 – Pipeline to generate HRC 8

The rationale behind utilizing different baseline models to select the closest candidate among 240 different versions of synthesized grain is to assess whether any existing texture similarity metrics can function similarly to an average expert viewer. By requesting subjects to rate the film grain similarity in test sequences and comparing the Mean Opinion Score (MOS) given to HRC 1 and other HRCs, we can gauge the confidence level in using the existing baseline models. Ultimately, this test underscores the necessity of designing a new metric specifically for film grain similarity if none of the baseline models meet the success criteria.

Ten IMAX expert viewers were invited to participate in the subjective study, which was conducted using an LG C2 65" 4K OLED evo with ThinQ AI TV (15). The IMAX experts are already highly sensitive to grain, but we provided them with specific instructions for this study, such as focusing on the grain fidelity in terms of the look and feel between the synthesized grain and the source. The TV was calibrated for HDR, with all advanced processing options disabled. The study employed the double-stimulus impairment scale (DSIS) method (16) where subjects first watched the source video followed by the synthesized grain version. They were then asked to provide a Film Grain Similarity (FGS) score. The impairment scale adhered to the Absolute Category Rating (ACR) convention, utilizing 10-level scores ranging from 1 to 10, corresponding to “very different” and “indistinguishable,” respectively. The experiment was conducted in a dark room, resembling a lab-study environment (17), where viewers had the flexibility to position themselves as

close to the TV as desired. This setup deviates from a typical video quality subjective study, necessitating a controlled environment to ensure accurate assessments.

OBSERVATIONS AND FINDINGS

The current study seeks to assess the perceived film grain similarity between test video sequences and their respective sources. Initially, subjective scores provided by participants were normalized into Z-scores per subject to account for any discrepancies in the use of the quality scale. Following normalization, an outlier removal procedure, as recommended in ITU Rec BT. 500 (16), was implemented, resulting in the absence of outliers. The resultant Z-scores were then rescaled linearly to fit within the range of 1 to 10. Mean opinion scores (MOS) for each individual video were calculated by averaging the rescaled Z-scores from all valid participants. Figure 5 illustrates the average Spearman rank-order correlation coefficient (SRCC) between scores provided by each participant and the MOS. This chart serves as a reliable indicator of the consensus among expert viewers in this specific task. Notably, the bar chart reveals that evaluating grain similarity for Src3 posed significant challenges, whereas subjects found assessing grain similarity for Src6 comparatively easier.

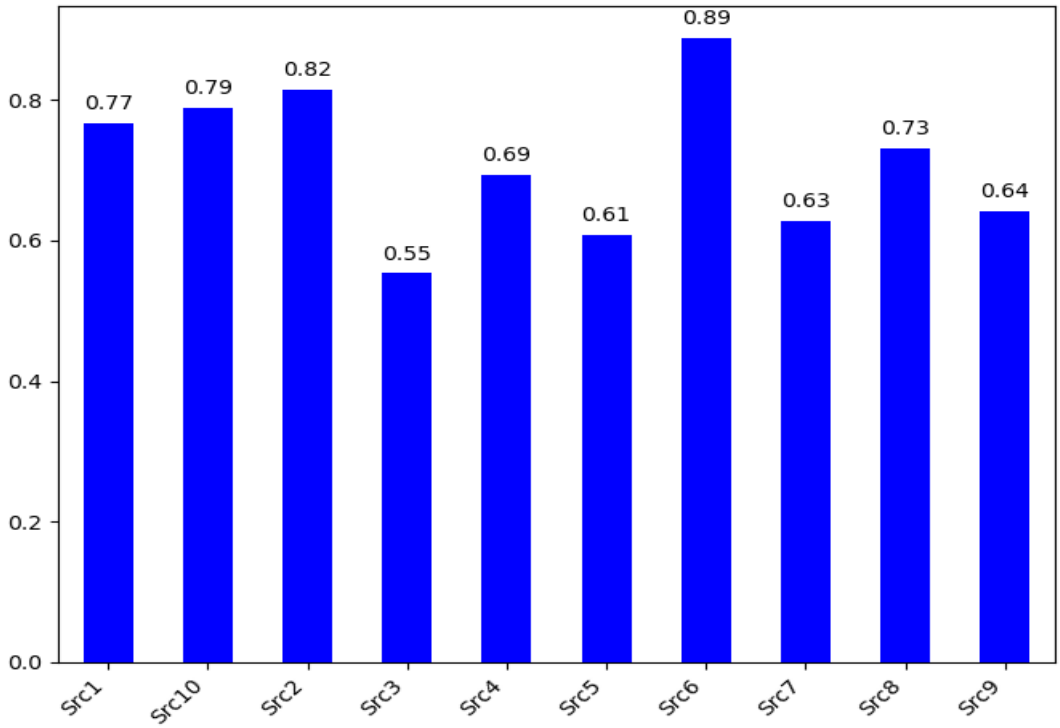

Figure 5: Average of SRCC Per Content for Each Subject with MOS(excluding itself)

Table 1 displays the Mean Opinion Score (MOS) for each SRC and HRC. The last two columns present the average and standard deviation values of MOS assigned to each HRC collectively, serving as an assessment of the existing baseline metrics. The average MOS for HRC 1 indicates a preference among expert viewers for the synthesized version produced by manually optimized AFGS1. The noticeable discrepancies in average MOS across different HRCs suggest that no baseline model matches the performance of expert viewers. The data indicates that the Gaty's model is the baseline model that performs closest to human subjects. However, the MOS values suggest that even though the Gaty's model shows promise, it is still not the ideal metric for measuring film grain similarity. These findings

serve as motivation to pursue the development of a new metric capable of assessing film grain similarity more effectively.

Furthermore, upon comparing the rows corresponding to HRC 7 and 8, it becomes apparent that expert viewers favor the use of a 4x4 block size over a 32x32 block size in the dav1d decoder. Subsequent interviews with the subjects revealed that one of the primary reasons for their preference for the 4x4 block size was the generation of repetitive or periodic patterns in the synthesized grain when using the 32x32 block size.

HRC	Src1	Src2	Src3	Src4	Src5	Src6	Src7	Src8	Src9	Src10	AVG	STD
1	8	8.5	8.33	8	5.68	8	7.67	8.5	8.17	8	**7.89**	**0.82**
2	5	4	8.8	7.5	5.83	3.83	6.33	6.83	8.6	7	6.37	1.73
3	7	4	7.5	8.17	6.83	7.17	5.83	7.5	6.5	6.83	**6.73**	**1.15**
4	7	4.17	7.67	7.67	7.67	3.5	4.5	7.17	6.17	6.83	6.24	1.59
5	5.67	3.83	7.5	8	7	3.5	4.83	7	7	7.67	6.20	1.64
6	1	1.17	1	1	1.17	1.17	1	1	1	1	1.05	0.08
7	7.5	6.17	7.67	6.5	5.4	6.67	5	6.67	6.5	5.17	6.33	0.91
8	6.83	6.33	6.67	5.83	5.67	6.33	4.33	5.83	7	5.33	6.02	0.80

Table 1: Subjective Testing Results

MODEL DESIGN

Motivated by our analysis of subjective testing results presented in Table 1 and drawing inspiration from the Gatys model (11) we have introduced the IMAX Film Grain Similarity (FGS) metric. Illustrated in Figure 6 is the workflow of IMAX FGS, which operates on two key inputs: the source video with native grain (used as a reference) and the test video containing synthesized grain (generated through film grain synthesis). The output of IMAX FGS is a score ranging from 0 to 100, with higher scores indicating greater similarity between the reference native grain and the synthesized grain in the test video. The model is structured around four core components. First is the saliency detection block. Notably, previous subjective testing has revealed that the human visual system is particularly attuned to film grain disparities within the flat regions of video frames. Hence, the saliency detection block ensures that our metric emphasizes grain similarity specifically within these flat areas. The second block is dedicated to feature extraction. Leveraging a neural-network-based machine learning model, this component adeptly extracts grain-related features from both reference and test video frames. Next, statistical measurements are performed on these extracted features. Here, we draw inspiration from the operations outlined in Gatys model, following a similar procedure using Gram matrix for statistical measurements to derive meaningful insights. Finally, the last step involves a similarity measurement, conducted on the statistical data obtained from the preceding step. This comprehensive approach, encompassing saliency detection, feature extraction, statistical analysis, and similarity measurement, collectively forms the foundation of the IMAX FGS quality assessment metric, enabling a comprehensive evaluation of film grain synthesis fidelity. The mathematical operations involved in step 3 and step 4 is shown in Equation 1.

$$\text{IMAXFGS} = \frac{2 \cdot G(x) \cdot G(y) + c1}{G^2(x) + G^2(y) + c1}$$

x

: reference features obtained from feature extraction step

y

: test features obtained from feature extraction step

$G(x)$

: Gram matrix used as statistical measurements operator

$c1$

: stability constant

Equation 1 – Mathematical operations involved in IMAXFGS

In the formula, x represents features of reference signal and y represents features of test signal. G represents the Gram matrix, which works as the statistical measurement operator. c1 is the stability constant to avoid divide by zero error. The formula also illustrates the similarity measure we employ, which is inspired by the luminance part of the SSIM (18) visual quality metric.

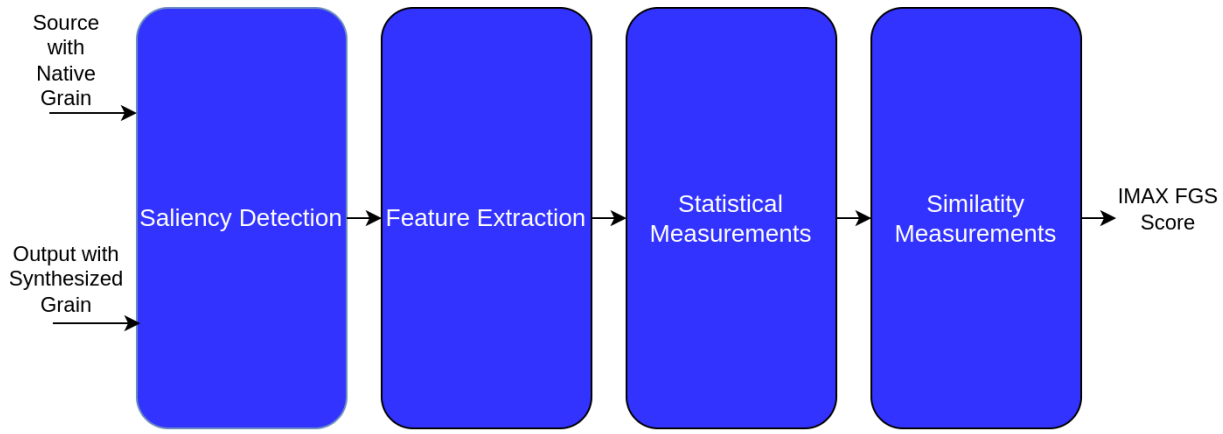

Figure 6 – IMAX FGS Model Subblocks

As for the training process, the model is trained end-to-end using previously mentioned subjective dataset. The dataset is splitted based on sources using a predefined 7:1:2 training-validation-testing split. The best model is picked based on its performance on the validation set.

VALIDATAION

To validate whether the developed model can effectively replace an average expert viewer, we computed the Spearman rank-order correlation coefficient (SRCC) between the IMAX FGS metric and the Mean Opinion Scores (MOS) obtained from our subjective study. Figure 7 illustrates the per-content SRCC between the IMAX FGS metric and MOS, as well as the

average subjective SRCC with respect to MOS. The final bar indicates that, on average and across all content, the IMAX FGS metric achieves a SRCC of 0.66, while the average SRCC for expert viewers is 0.72. This demonstrates the confidence level in using the metric for optimizing components in modern codecs with film grain synthesis support or for designing new film grain synthesis approaches. Please note that it is not unusual for IMAX FGS to show higher correlation with some assets compared to the average subjects for a particular piece of content. In fact, this behavior is consistently observed when we single out a subject from the population. It is similar to one subject having a higher correlation with respect to the Mean Opinion Score (MOS)

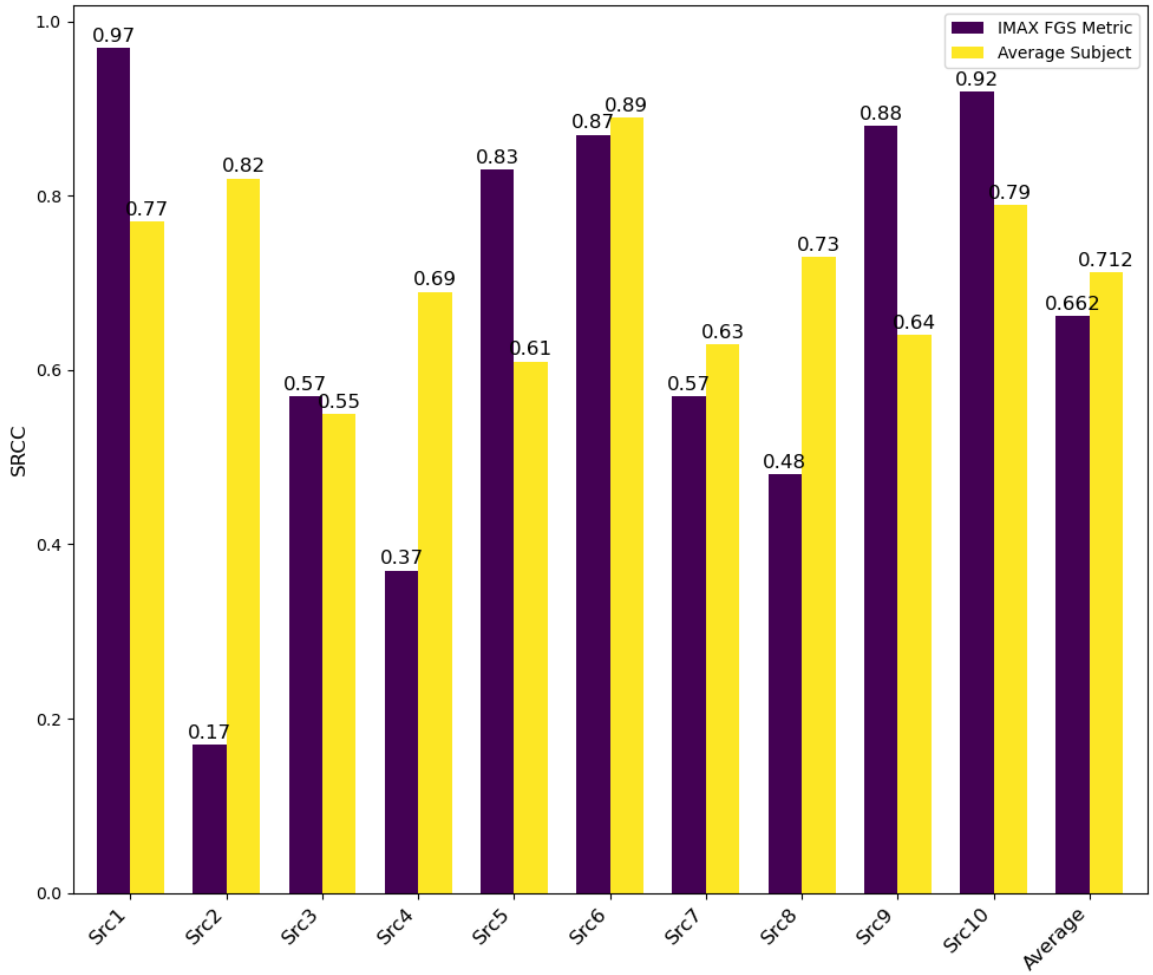

Figure 7 – SRCC per Content for subjective study presented above

The same validation methodology is applied using the subjective data presented in (14). This dataset is entirely novel to IMAX FGS, making it a robust validation test for the proposed metric. Figure 8 illustrates the per-content SRCC as well as the average SRCC across all 9 contents used in the subjective study. The final bar chart confirms that even for this previously unseen data, the IMAX FGS metric performs consistently close to expert viewers. The data indicates that while the average SRCC for expert viewers is 0.57, it reaches 0.64 for the IMAX FGS metric.

The IMAX FGS raw scores range from 0 to 1 and are not linearly scaled. Establishing an effective mapping between these raw scores and perceptual film grain similarity scores is

crucial. To achieve this, we conducted an experiment using 15 new source content, generating a diverse range of synthesized grain. IMAX expert viewers were then asked to assign scores ranging from 0 to 100 to denote the level of similarity. By correlating the IMAX FGS raw scores with the subjective scores, we were able to derive perceptually linear scores for the IMAX FGS metric. Consequently, the output of the metric now spans from 0 to 100, representing "very dissimilar" to "similar", respectively. Figure 9 provides an illustration featuring a frame with analog film grain, along with an extracted patch from the sky containing authentic grain. A corresponding patch with a spectrum of synthesized grain is presented, accompanied by the corresponding IMAX FGS scores.

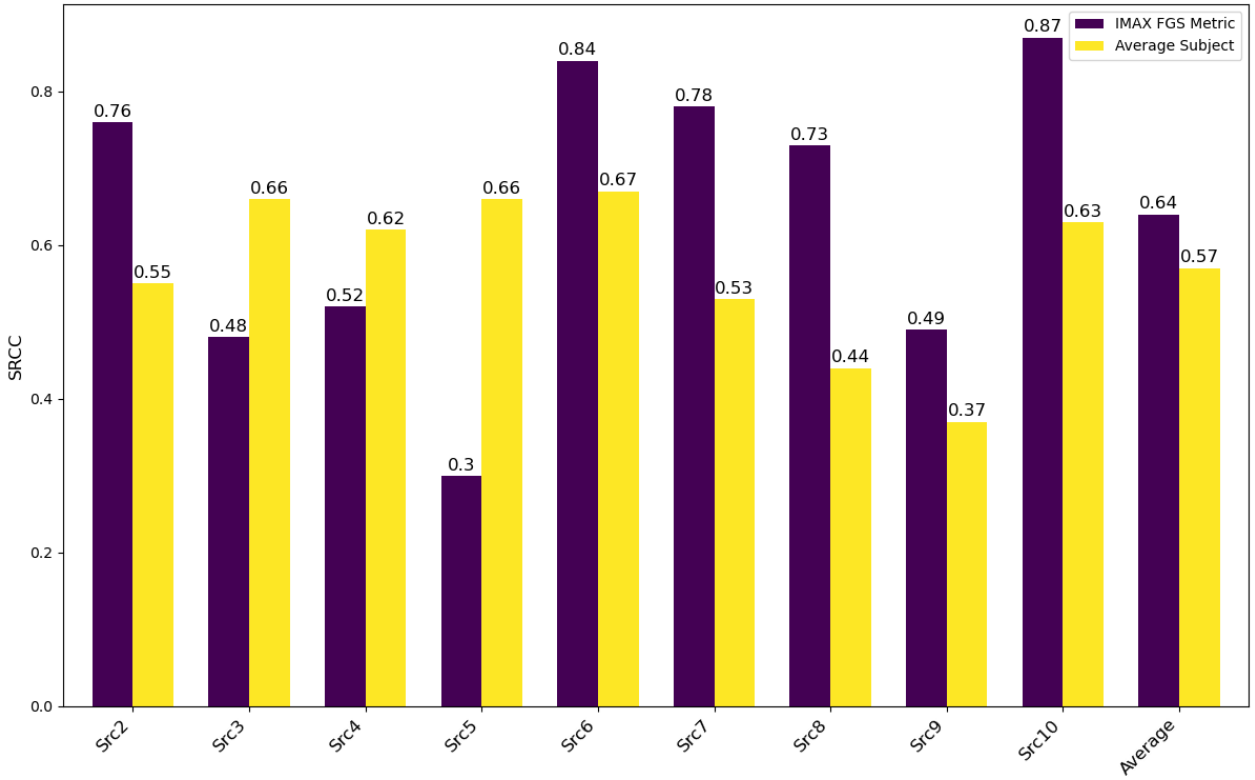

Figure 8 - SRCC per Content for subjective study discussed in (1)

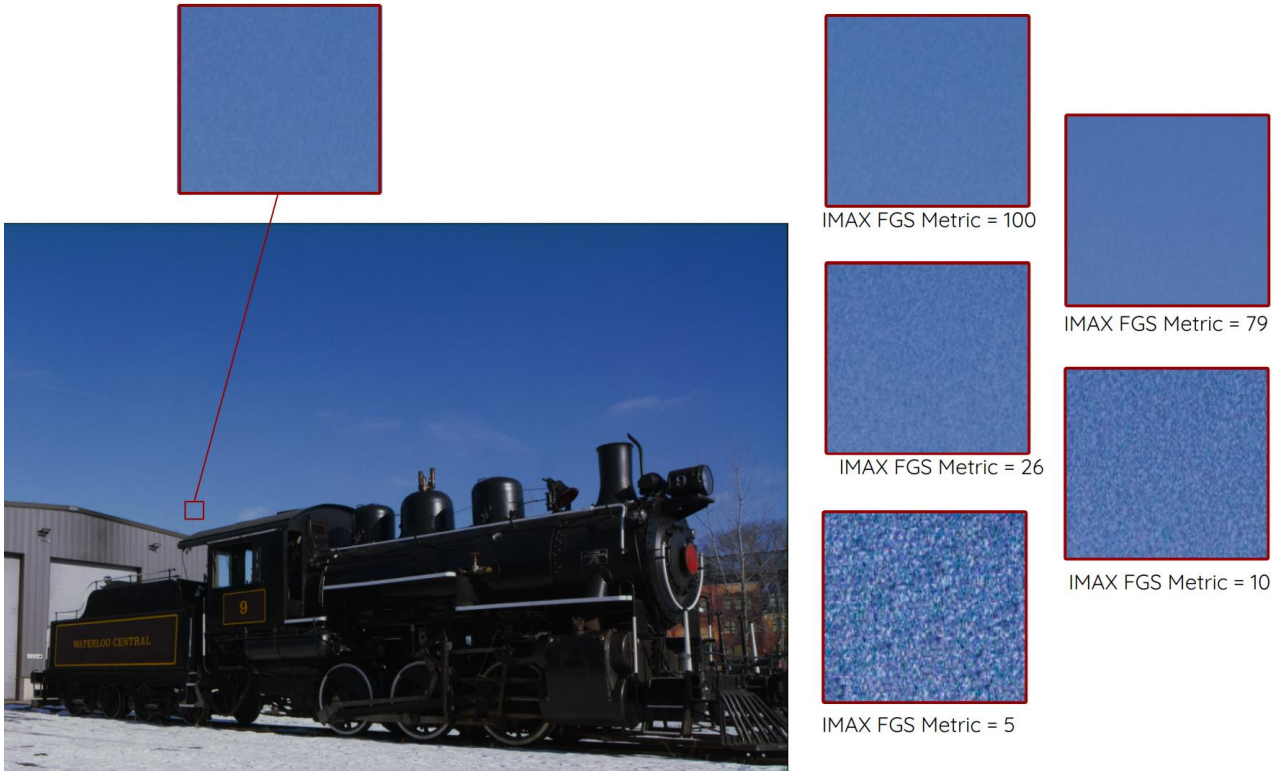

Figure 9 - Analog vs. Synthesized Film Grain with IMAX FGS Scores

CONCLUSION

In conclusion, the development of the IMAX FGS metric represents a significant step forward in the field of film grain similarity assessment. Through a comprehensive validation process, we have demonstrated its efficacy in accurately evaluating film grain similarity, performing comparably to average expert viewers. By leveraging AI approaches, we have addressed the limitations of existing metrics, offering a reliable tool for quantifying the performance of grain processing methods in modern codecs. Our metric's ability to provide perceptually linear scores further enhances its practical utility, facilitating its integration into video encoding systems. With its proven effectiveness and potential to optimize film grain synthesis frameworks, the IMAX FGS metric stands as a valuable contribution to the pursuit of preserving film grain authenticity in digital video distribution.

REFERENCES

1. J. Sapra, K. Zeng, H. Yeganeh, A Subjective Study of Film Grain Synthesis for The Preservation of Creative Intent, Technical Paper, IBC, September 2023

2. B. T. Oh, S. Lei, and C. C. J. Kuo, Advanced film grain noise extraction and synthesis for high-definition video coding. in IEEE Transactions on Circuits and Systems for Video Technology, vol. 19, no. 12, pp. 1717-1729, Dec. 2009.

3. J. Dai, O. C. Au, C. Pang, W. Yang, and F. Zou, Film grain noise removal and synthesis in video coding. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA.

4. I. Hwang, J. Jeong, S. Kim, J. Choi, and Y. Choe, Enhanced film grain noise removal and synthesis for high fidelity video coding. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2013.

5. A. Newson, N. Faraj, B. Galerne, and J. Delon, Realistic Film Grain Rendering. Image Processing On Line, 2017, pp. 165-183.

6. Z. Ameur, W. Hamidouche, E. Francois, M. Radosavljevic, D. Menard, and C. H. Demarty, Deep-based film grain removal and synthesis. Image and Video Processing eess.IV, 2022.

7. AOMedia Film Grain Synthesis 1 (AFGS1) specification https://aomediacodec.github.io/afgs1-spec/

8. A. Norkin, and N. Birkbeck, Film grain synthesis for AV1 video codec. 2018 Data Compression Conference, Snowbird, UT, USA, 2018, pp. 3-12.

9. Grain Synth analyzer and editor for AV1 files, https://github.com/rust-av/grav1synth

10. dav1d: AV1 cross-platform decoder, https://code.videolan.org/videolan/dav1d

11. L. Gatys, A. S. Ecker, and M. Bethge, “Texture synthesis using convolutional neural networks,” in Conference on Neural Information Processing Systems, 2015, pp. 262–270.

12. K. Ding, K. Ma, S. Wang and E. P. Simoncelli, "Image Quality Assessment: Unifying Structure and Texture Similarity," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 5, pp. 2567-2581, 1 May 2022

13. K. Ding, Y. Liu, X. Zou, S. Wang and K. Ma, " Locally Adaptive Structure and Texture Similarity for Image Quality Assessment" in Proceedings of the 29th ACM International Conference on Multimedia, pp. 2483-2491, October 2021

14. Portilla, J., Simoncelli, E.P. A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. International Journal of Computer Vision 40, 49–70 (2000).

15. LG C2 65" 4K OLED evo with ThinQ AI, https://www.lg.com/ca_en/tvs/lg-oled65c2pua

16. ITU, BT.500: Methodologies for the subjective assessment of the quality of television images, Tech. Rep., Intl. Telecomm. Union, 2019.

17. ITU, BT.2022: General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays, Tech. Rep., Intl. Telecomm. Union, 2017.

18. Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," in IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, April 2004