Optimizing Video Quality Estimation Across Resolutions

Abhinau K Venkataramanan
Electrical and Computer Engineering
University of Texas at Austin
Austin, United States of America
abhinaukumar@utexas.edu

Chengyang Wu
Electrical and Computer Engineering
University of Texas at Austin
Austin, United States of America
chengyangwu@utexas.edu

Alan C Bovik
Electrical and Computer Engineering
University of Texas at Austin
Austin, United States of America
bovik@ece.utexas.edu

***Abstract*—Many algorithms have been developed to evaluate the perceptual quality of images and videos, based on models of picture statistics and visual perception. These algorithms attempt to capture user experience better than simple metrics like the peak signal-to-noise ratio (PSNR) and are widely utilized on streaming service platforms and in social networking applications to improve users' Quality of Experience. The growing demand for high-resolution streams and rapid increases in user-generated content (UGC) sharpens interest in the computation involved in carrying out perceptual quality measurements. In this direction, we propose a suite of methods to efficiently predict the structural similarity index (SSIM) of high-resolution videos distorted by scaling and compression, from computations performed at lower resolutions. We show the effectiveness of our algorithms by testing on a large corpus of videos and on subjective data.**

***Index Terms*—Image/Video Quality Assessment, Structural Similarity (SSIM), Human Vision System (HVS)**

I. Introduction

The Structural Similarity Index (SSIM) [1] is a globally-deployed picture quality model that offers significantly higher correlation with subjective quality assessment relative to the Peak Signal-to-Noise Ratio (PSNR), albeit at higher computational cost. SSIM is calculated by first computing a quality map expressed in terms of perceptually relevant local first and second order statistics. The average value of this quality map is usually reported as the SSIM score. SSIM is an $O(MN)$ algorithm, where M and N are the height and width of the image/video frame. This quadratic growth, while not a problem at lower resolutions, is consequential given the emergence of social media and streaming platforms which deliver immense volumes of high-resolution picture and video content at global scales. Monitoring picture and video quality at such large scales is proving to be expensive.

Since SSIM demonstrated significant gains over PSNR, several full and reduced-reference metrics have been proposed which attempt to model subjective visual quality of images and videos. Notable among these are Visual Information Fidelity (VIF) [2], Spatio-Temporal Reduced Reference Entropic Differences (ST-RRED) [3], SpEED-QA [4] and the fusion metric Video Multi-method Assessment Fusion (VMAF) [5]. However, due to their computational requirements, SSIM continues to be one of the most widely deployed models. As a result, we choose to optimize the prediction of SSIM. This work can be naturally extended to other models by computing similar quality features using different models.

On streaming and social media platforms, videos are commonly encoded at lower resolutions for transmission. This is done either because the source has low-complexity content and can be downsampled with relatively little additional loss, or if the available bandwidth requires it, or to decrease the decoding load at the user's end. Perceptual distortion models are becoming more common tools for determining the quality of encodes for Rate Distortion Optimization (RDO) [6]. With advances in video hardware enabling accelerated encoding and decoding of videos, the distortion estimation step has become bottleneck when carrying out RDO over a set of encoding "recipes." Due to the quadratic growth of SSIM's computation, this is an increasingly relevant issue given the prevalence of high-resolution videos.

Within this context, it is of great interest to be able to accurately predict the quality of high-resolution videos that are distorted in two steps - scaling followed by compression. For example, consider High Definition (HD) videos that are first resized to a lower resolution, which we call the compression resolution, then encoded and decoded using, for example, H.264 at this compression resolution. The videos are then upsampled to the original resolution before they are rendered for display. We will refer to this higher resolution as the rendering resolution.

To reduce the computational burden of perceptually-driven RDO, we aim to bypass the computation of SSIM at the rendering resolution between the HD source and rendered HD video. Our invention, which we call Scaled SSIM, predicts SSIM by only using SSIM values computed at the lower compression resolution during runtime. The setup of the compression pipeline that includes this SSIM prediction process is shown in Figure 1.

II. Related Work

SSIM belongs to the family of full-reference (FR) quality assessment algorithms. While the relative computational loads of these algorithms has not been systematically investigated, their performance against subjective data has been reported many times.

This research was supported by funding from Facebook.

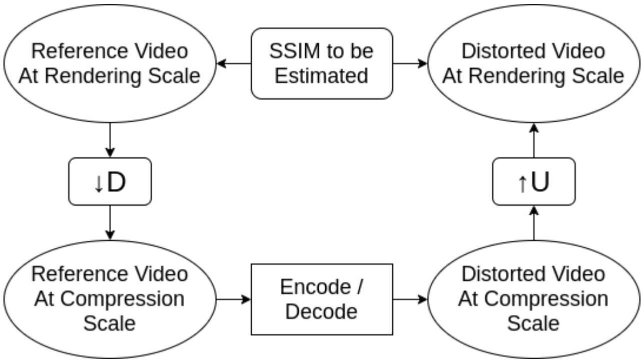
Reference Video At Rendering Scale
SSIM to be Estimated
Distorted Video At Rendering Scale
↓D
↑U
Reference Video At Compression Scale
Encode / Decode
Distorted Video At Compression Scale

Fig. 1. Video Compression Pipeline

Comprehensive studies have been conducted [7] [8] which evaluate Full Reference (FR) models on a host of databases using common performance metrics like the Spearman Rank Order Correlation Coefficient (SROCC), Kendall Tau Correlation (KCC), Pearson Correlation (PCC) and the Root Mean Squared Error (RMSE).

There has also been theoretical work comparing SSIM to other FR models under some statistical assumptions [9]. Such theoretical results provide better insight into the performance of SSIM with respect to other picture and video perceptual quality prediction algorithms.

Since SSIM provides a more meaningful distortion measurement than Mean Squared Error (MSE), or equivalently, PSNR, it has received a great deal of attention as a distortion predictor for evaluating and optimizing compression algorithms. In [10], Richter et. al. proposed an MS-SSIM optimal JPEG 2000 decoder. In [11], Yeo et. al. presented a method for RDO using SSIM as the distortion metric. SSIM has also been useful for optimization of inter-frame encoding [12], RDO of video encodes [13], and perceptual rate control [14].

III. Proposed Models

We propose two classes of models that efficiently predict Scaled SSIM, which we refer to as

- Histogram Matching
- Feature-based models

All of the proposed models operate on a per-frame basis.

A. Histogram Matching

SSIM at any scale is computed from a SSIM map, which expresses information about local quality. We observe that there is a non-linear relationship (which is to be estimated) between frame-wise SSIM values at different scales. Since the overall SSIM value is the mean of the SSIM map for each frame, we can seek to relate the SSIM values at two different scales by matching the histograms of the two SSIM maps obtained on the same frames at the two scales.

However, the goal is to *bypass* the calculation of the true SSIM, and hence SSIM map, at the rendering scale. So, we instead compute the “true” SSIM map at the rendering scale only once every k frames. Further, we assume that the shape of the true histogram of each subsequent frame over the short time period of $k-1$ frames is approximately constant.

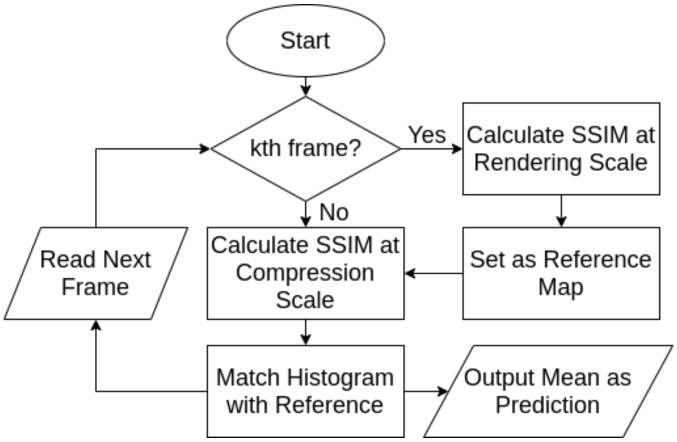
Start
kth frame?
Yes
No
Calculate SSIM at Rendering Scale
Read Next Frame
Calculate SSIM at Compression Scale
Set as Reference Map
Match Histogram with Reference
Output Mean as Prediction

Fig. 2. Histogram Matching Solution

The true SSIM map is then reused on the next $k-1$ frames as a heuristic model against which the shapes of the next $k-1$ histograms are matched. Thus, each frame’s SSIM map at the compression scale is transformed so that its histogram matches that of the reference map. The predicted frame SSIM value is then taken as the mean of the transformed SSIM map. After $k-1$ frames, the same process is repeated so that the true SSIM map at the rendered scale is computed once every k frames.

Later, we will examine the quality prediction performance of this scheme. However, assuming near-parity in performance (as we will show), we are mainly interested in the speedup obtained via this estimation process, compared to calculating the rendered SSIM on every frame. Let us denote the factor by which we downsample the source video by $\alpha \in (0,1)$. Then, the ratio of required computation using our proposed approach, to SSIM computation directly at the rendered scale is (approx.)

$$\left(1-\frac{1}{k}\right)\alpha^2\left(1+\beta+\gamma\right)+\frac{1}{k}(1+\beta) \tag{1}$$

The factors β and γ account for computing and matching the histograms respectively, which are both $O(MN)$ operations. This ratio is a decreasing function of k, and approaches $\alpha^2(1+\beta+\gamma)$ as $k \to \infty$.

By comparison, if the rendered SSIM map were not sampled, the ratio would be (approx.) α^2. In practice, we observe that the time taken to compute and match histograms is comparable to the time taken to compute the SSIM map at the compression scale. So, the computational burden of the matching step is small, albeit not negligible. The Histogram Matching solution is illustrated as a flowchart in Figure 2.

As an ablation study, we consider the algorithm obtained by simply reusing the SSIM value from the reference map, instead of matching the histograms. Note that this is identical to “skipping” $k-1$ frames after calculating the reference SSIM map. In Figure 4, we compare the SROCC with subjective

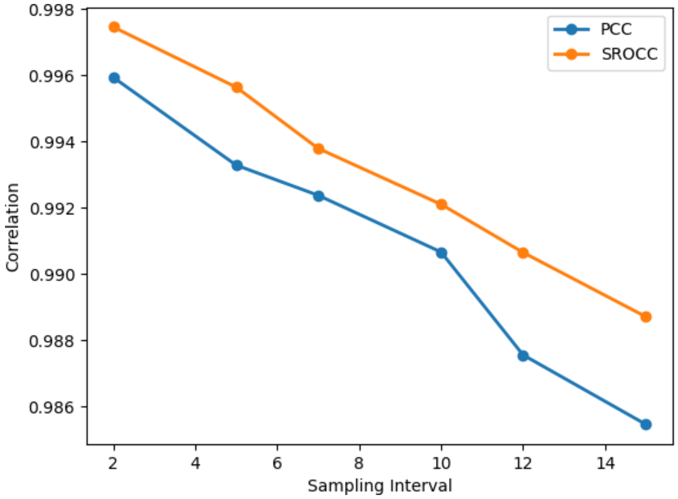

Fig. 3. Correlation vs Sampling interval for Histogram Matching

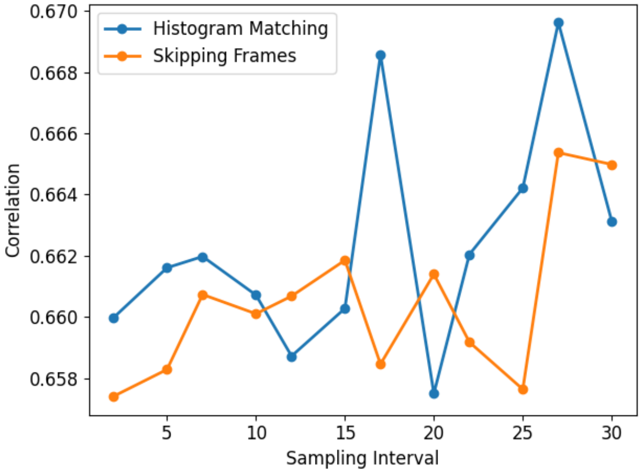

Fig. 4. Comparison of Histogram Matching with the ”Skip” Baseline

scores achieved by the two algorithms on the Netflix Public Database. We observe from this plot that the Histogram Matching generally outperforms the baseline, illustrating the positive effect of the histogram matching step.

A drawback of the Histogram Matching method is that it requires “guidance” in the form of the true SSIM map at rendering scale, and it also implicitly assumes that video quality does not change rapidly. This means that as k increases, the reduction in computational complexity is accompanied by a reduction in accuracy, since the variation in quality increases over larger sampling intervals. This loss in accuracy is illustrated in Figure 3. As a default, we choose $k = 5$ unless otherwise mentioned. Nevertheless, the following two classes of models alleviate these limitations.

B. Feature-based Models

Quality degradation at the rendering scale is a consequence of two operations - downsampling and compression. Therefore, estimating the loss in quality arising from these two operations may supply good features to predict SSIM at the rendering scale.

Let X be an original video, and denote the video scaled by a factor of α as $S_\alpha(X)$. Then, the result of upsampling the downsampled video back to the original resolution may be denoted by $S_{\frac{1}{\alpha}}(S_\alpha(X))$. The SSIM value between X and $S_{\frac{1}{\alpha}}(S_\alpha(X))$ is a measure of the loss in quality from downsampling the video. Since this SSIM is independent of the choice of codec and compression parameters, this can be pre-computed.

The second source of quality degradation is compression. Denote the result of compressing the video X using a Quantization Parameter (QP) q as $C(X; q)$. Then, the SSIM value between $S_\alpha(X)$ and $C(S_\alpha(X); q)$ measures the loss of quality from compressing the video at the compression scale.

Models that use only the above two features will henceforth be called Two-feature Models. In addition, the scaling factor α and Quantization Parameter q can also be used as features. These models are then called Four-feature models.

Three regressors were each trained to predict the SSIM value at the rendering scale on each frame. The three regressors considered are

- Linear Support Vector Regressor (Linear SVR)
- Gaussian Radial Basis Function SVR (Gaussian SVR)
- Fully Connected Neural Network (NN)

The Neural Network is a small fully connected network having a single hidden layer with twice the number of neurons as input features. These models were compared with a simple learning-free model, which is used as a baseline. The output of the baseline model is the product of the two SSIM features. This is similar to the 2stepQA metric proposed in [15] for two-stage distorted images. We call this the Product model.

IV. Dataset

In order to develop and test our algorithms, we compiled a corpus of 60 Full HD pristine videos taken from a wide range of Video Quality databases - the Netflix Public Database [16], LIVE NFLX II Data base [17], LIVE Netflix Video QOE Database [18], CSIQ Video Quality Database [19], IRCCyN IVC Database [20], IVP Subjective Quality database [21] and the VQEG HDTV Database [22].

Each reference video was compressed at 6 compression resolutions - 144p, 240p, 360p, 480p, 540p and 720p, using FFMPEG’s H.264 (libx264) encoder, and for each resolution compressed using 11 values of the QP - $1, 5, \ldots 51$, then finally scaled back up to 1080p. All scaling operations where performed using the Lanczos-3 filter. This resulted in a total of 3960 videos, consisting of almost 1.75M frames. This formed the corpus of videos used to train and test our algorithms. The Netflix Public Database also provides distorted videos with corresponding DMOS scores. These were used to evaluate the performance of our algorithms against subjective scores.

The AVT-UHD1 database [23] is a large database that investigates the effect of scaling and compression on the quality of Ultra HD (UHD) videos. However, since the videos were not rendered at their source resolution during that subjective study, it could not be used for our subjective evaluation data.

V. Experiments

All of the models were first trained and tested on the 60 reference videos. As mentioned in Section IV, this led to a total of 26,235 reference frames. To minimize content overlap between the training and testing set, all corresponding frames (same contents) of distorted videos, across compression scales and QPs, were assigned together to either the training or testing set. 21,000 reference frames were chosen for training and the remaining 5,235 for testing. This led to a total of almost 1.4M training data points and almost 350,000 testing data points.

On the corpus, we report the Pearson Correlation Coefficient (PCC) and the Spearman Correlation Coefficient (SROCC) between the predicted SSIM and the true SSIM. We also report the variation in the best model's performance against choices of the compression scale α and q to test the consistency of the model's performance across use cases. Since the videos in the corpus have not been rated by subjects as part of a study, we can only calculate the correlation between predicted and true SSIM scores.

The end goal, however, is still to predict subjective scores, so we tested our models on the Netflix Public Database. We compared the correlation values obtained by our predicted SSIM scores with DMOS against the correlation of the true SSIM with DMOS calculated at the rendering resolution. Since the videos in these subjective quality databases were compressed using a variety of methods, we restricted our tests to the Histogram Matching and two-feature models. To obtain the two SSIM features, we downsampled the reference videos to the appropriate compression resolution.

VI. Results

The testing performance of the various models on the corpus is listed in Table I. As mentioned in Section V, the PCC and SROCC values reported are SSIM - SSIM correlations. The "2" and "4" in the model names denote the number of features used. Note that the goal of all our proposed methods is to match SSIM's performance, so we evaluate our algorithms against the performance of the true SSIM scores.

From the table, we see that among the feature-based models, the 4 feature Neural Network was the best performing model. This is expected, given the great learning capacity of neural networks. Interestingly, the Product model (baseline) outperformed almost every other model on the corpus, with the added advantage of having negligible inference time in comparison.

Finally, we see that Histogram Matching led to almost perfect predictions, outperforming all feature-based models. However, unlike the product baseline, this model presents a tradeoff. The cost of this near-perfect estimation is the presence of guiding information in the form of true SSIM maps at regular intervals, which demands additional computation.

TABLE I
Correlation with True SSIM on corpus test data

Model	PCC	SROCC
NN 2	0.9461	0.9834
NN 4	**0.9845**	**0.9869**
Linear SVR 2	0.9529	0.9759
Linear SVR 4	0.9215	0.9201
Gaussian SVR 2	0.8571	0.9591
Gaussian SVR 4	0.9598	0.9628
Product (Baseline)	0.9662	0.9829
Histogram Matching	**0.9933**	**0.9956**

TABLE II
Correlation with DMOS on Netflix Public Database

Model	PCC	SROCC
True SSIM	0.6962	0.6567
NN 2	**0.6759**	**0.6425**
Linear SVR 2	0.6746	0.6196
Gaussian SVR 2	0.6756	0.6373
Product (Baseline)	0.6715	0.6215
Histogram Matching	**0.6848**	**0.6616**

The variation in performance of the Product Model, Histogram Matching and the best learning-based model (4-feature Neural Network) with choice of compression scale on the vertical axis and QP on the horizontal axis, is depicted in Figure 5.

From the figures, it may be seen that Histogram Matching performed well under almost all conditions. We observed a slight dip in performance at lower QPs. At low QPs, compression is performed at high quality. As a result, the SSIM map at the compression scales yields mostly values close to 1. So, the histogram is a narrow distribution close to 1. As a result, it is more difficult to match the reference histogram.

Table II compares the correlation of true and predicted SSIM scores against subjective opinion scores (DMOS). All SSIM scores were transformed to quality predictions by fitting to subjective scores using the five-parameter logistic function

$$Q(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp(\beta_2 (x - \beta_3))} \right) + \beta_4 x + \beta_5 \quad (2)$$

The SROCC was calculated using the predicted SSIM scores, while PCC was calculated after applying this transformation. From the table, it may be seen that the performance of SSIM estimated by Histogram Matching matched the performance of true SSIM. We also observe that the feature-based models approach true SSIM's performance, with the Product Model offering a good low complexity alternative to the learning-based models.

VII. Conclusion

In this work, we have motivated and proposed the problem of efficiently estimating Scaled SSIM. We propose two main approaches to solving this problem, one using Histogram Matching, and the other using features which can either be precomputed, or are computed at lower resolutions. We demonstrate the effectiveness of these approaches by quantifying the

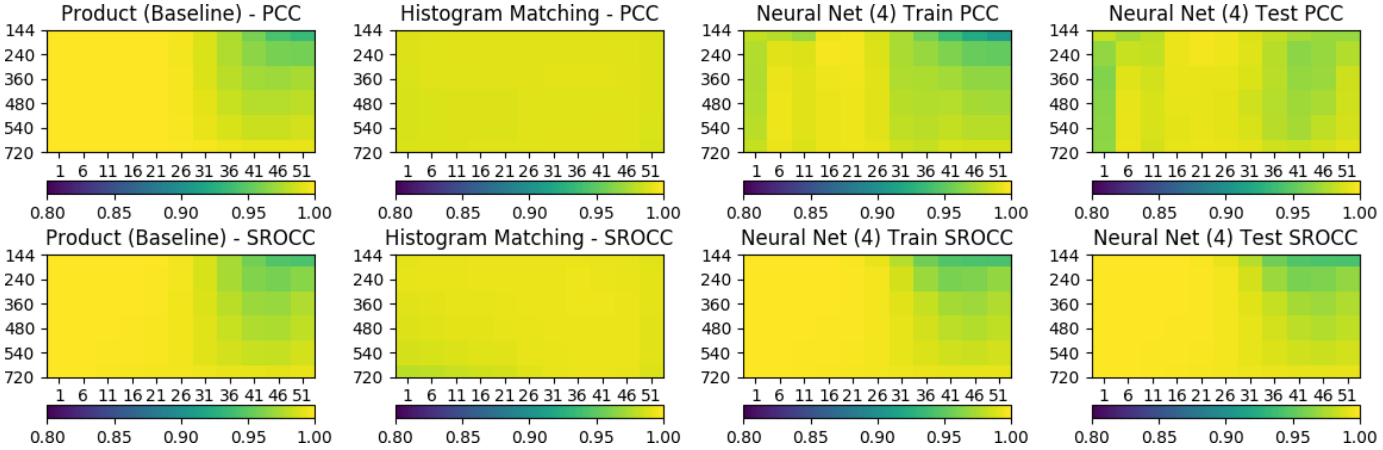

Fig. 5. Variation in performance with choice of Encoding Scale and QP

accuracy with which they predict the true SSIM, and their correlation against subjective data. In this way, we achieved the goal of reliably predicting the SSIM score between a pair of reference and test videos at a fraction of the computational cost.

In the future, we see merit in exploring better temporal aggregation strategies of frame-wise SSIM models. This work would potentially also pave the way to calculate SSIM at lower frame rates. This approach may also be improved by considering richer, potentially codec-dependent features, leading to better predictive performance of feature-based models.

References

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

[2] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[3] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal etropic differencing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 684–694, 2012.

[4] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "Speed-qa: Spatial efficient entropic differencing for image and video quality," *IEEE signal processing letters*, vol. 24, no. 9, pp. 1333–1337, 2017.

[5] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Dynamic optimizer - a perceptual video encoding optimization framework," https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652.

[6] Netflix, "Dynamic optimizer - a perceptual video encoding optimization framework," https://netflixtechblog.com/dynamic-optimizer-a-perceptual-video-encoding-optimization-framework-e19f1e3a277f.

[7] 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*. IEEE, 2012, pp. 1477–1480.

[8] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[9] K. Seshadrinathan and A. C. Bovik, "Unifying analysis of full reference image quality assessment," in *2008 15th IEEE International Conference on Image Processing*. IEEE, 2008, pp. 1200–1203.

[10] T. Richter and K. J. Kim, "A MS-SSIM optimal JPEG 2000 encoder," in *2009 Data Compression Conference*. IEEE, 2009, pp. 401–410.

[11] C. Yeo, H. L. Tan, and Y. H. Tan, "On rate distortion optimization using SSIM," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 7, pp. 1170–1181, 2013.

[12] C. L. Yang, R. K. Leung, L. M. Po, and Z. Y. Mai, "An SSIM-optimal H.264/AVC inter frame encoder," in *2009 IEEE International Conference on Intelligent Computing and Intelligent Systems*. IEEE, 2009, vol. 4, pp. 291–295.

[13] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-motivated rate-distortion optimization for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 4, pp. 516–529, 2011.

[14] T. S. Ou, Y. H. Huang, and H. H. Chen, "SSIM-based perceptual rate control for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 5, pp. 682–691, 2011.

[15] X. Yu, C. G. Bampis, P. Gupta, and A. C. Bovik, "Predicting the quality of images compressed after distortion in two steps," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5757–5770, 2019.

[16] Netflix, "VMAF - Video Multi-Method Assessment Fusion," https://github.com/Netflix/vmaf.

[17] C. G. Bampis, Z. Li, I. Katsavounidis, T. Y. Huang, C. Ekanadham, and A. C. Bovik, "Towards perceptually optimized end-to-end adaptive video streaming," *arXiv preprint arXiv:1808.03898*, 2018.

[18] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "LIVE Netflix Video Quality of Experience Database," *Online: http://live. ece. utexas. edu/research/LIVE _ NFLXStudy/index. html*, 2016.

[19] P. V. Vu and D. M. Chandler, "ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *Journal of Electronic Imaging*, vol. 23, no. 1, pp. 013016, 2014.

[20] S. Péchard, R. Pépion, and P. Le Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm," in *International Workshop on Image Media Quality and its Applications*, Kyoto, Japan, Sep 2008, p. 6, IMQA2008.

[21] Image and Video Processing Laboratory (IVP) at The Chinese University of Hong Kong, "IVP Subjective Quality Video Database," *Online: http://ivp.ee.cuhk.edu.hk/research/database/subjective/*, 2016.

[22] M. Barkowsky, M. Pinson, R. Pépion, and P. Le Callet, "Analysis of freely available dataset for HDTV including coding and transmission distortions," 2010.

[23] R. R. R. Rao, S. Göring, W. Robitza, B. Feiten, and A. Raake, "AVT-VQDB-UHD-1: A Large Scale Video Quality Database for UHD-1," in *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2019, pp. 17–177.