NO-REFERENCE QUALITY ASSESSMENT OF VARIABLE FRAME-RATE VIDEOS USING TEMPORAL BANDPASS STATISTICS

Qi Zheng[⋆]*, Zhengzhong Tu*[†]*, Yibo Fan*[⋆]*, Xiaoyang Zeng*[⋆]*, and Alan C. Bovik*[†]

[⋆] State Key Laboratory of ASIC and System, Fudan University [†] The University of Texas at Austin

ABSTRACT

Recent advances in mobile devices and cloud computing techniques have made it possible to capture, process, and share high resolution, high frame rate (HFR) videos across the Internet nearly instantaneously. Being able to monitor and control the quality of these streamed videos can enable the delivery of many enjoyable content and perceptually optimized rate control. However, the development of no-reference (NR) VQA algorithms targeting frame rate variations has been little studied. Here, we propose a first-of-a-kind blind VQA model for evaluating HFR videos, which we dub the Framerate-Aware Videos Evaluator w/o Reference (FAVER). FAVER uses extended models of spatial natural scene statistics that encompass space-time wavelet-decomposed video signals, to conduct efficient frame rate sensitive quality prediction. Our extensive experiments on several HFR video quality datasets show that FAVER outperforms other blind VQA algorithms at a reasonable computational cost. The code will be released on `https://github.com/uniqzheng/HFR-BVQA`.

Index Terms— Video Quality Assessment, high frame rate, no reference/blind, temporal band-pass filter, natural scene statistics, generalized Gaussian distribution

1. INTRODUCTION

A topic of very recent interest is the possibility of augmenting adaptive bitrate (ABR) streaming by all allowing control over variable frame rates (VFR). This concept has increased relevance given the availability of high-frame rate (HFR) videos which can improve viewing experiences on high-motion and rapidly changing videos. For example, just as streaming providers now routinely alter spatial resolution along with standard compression in the construction of bitrate ladders, they could also modify the streamed frame rate before compression depending on the content characteristics. Creating ABR bitrate ladders that allow for VFRs will require new VQA models and algorithms to perform perceptual optimization. In this direction, new databases have been built containing VFR contents [1, 2], and efficient predictors of compressed VFR video quality have been created [3, 4].

Just as there is now a heightened need for FR VQA models for VFR/HFR scenarios, so also is there for blind/NR models that can address the effects of frame rate changes throughout video workflows without the availability of any reference. Yet, existing BVQA models have been focused on fixed frame-rate videos, and there are no blind models specifically designed to predict the quality of VFR/HFR videos. Although a variety of temporal quality-aware features have been explored [3, 5–7], these are generally between-frame measurements with no demonstrated efficacy on VFR videos. Since the demand for HFR videos is increasing and many soon explode, so also will VFR protocals and the need for video quality prediction models capable of guiding them, including those that lead to BVQA algorithms.

Here, we propose, to best of our knowledge, *the first NR-VQA* model that targets visual distortions arising from framerate variations, which we dub the Framerate Aware Video Evaluator w/o Reference, or FAVER. The design of FAVER relies on a general, effective and efficient temporal statistics model that is based on the well-established bandpass regularities of natural videos. Our experiments show that FAVER achieves state-of-the-art performance on predicting frame-rate-variant video quality on VFR/HFR databases.

2. RELATED WORK

The earliest BVQA / BIQA algorithms were designed to analyze and quantify a single distortion type [8–12]. These have been largely replaced by much more powerful general-purpose BIQA/BVQA models that are based on measurements of distortion-induced deviations of bandpass processed images/videos from perceptually relevant natural scene statistics (NSS) [5, 13–23]. Deep convolutional neural networks have been shown to deliver superior performance on numerous image analysis tasks [24,25]. Recently, several successful deep learning-based BVQA models have emerged [7, 26–28]. Among hybrid models, RAPIQUE [6] efficiently combines low-level NSS features with high-level deep learning features

*Yibo Fan is the corresponding author.

This work was supported in part by the National Natural Science Foundation of China under Grant 62031009, in part by the Shanghai Science and Technology Committee (STCSM) under Grant 19511104300, in part by Alibaba Innovative Research (AIR) Program, in part by the Innovation Program of Shanghai Municipal Education Commission, in part by the Fudan University-CIOMP Joint Fund(FC2019-001), in part by the Fudan-ZTE Joint Lab, in part by Pioneering Project of Academy for Engineering and Technology Fudan University(gyy2021-001).

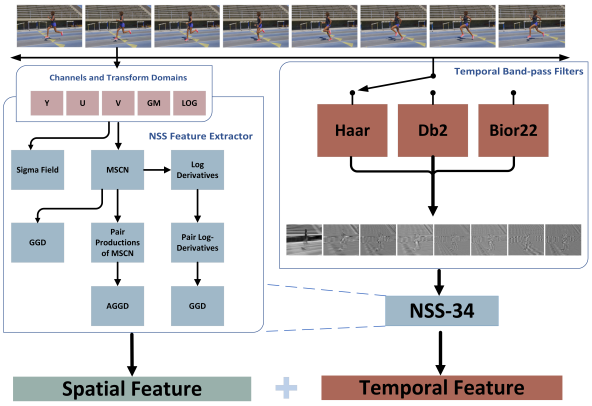


Fig. 1. The overall framework of the proposed FAVER model.

which are used to train a single regressor head that predicts video quality scores.

While there are now available a variety of BVQA models that are able to capture various temporal aspects of perceptual video quality, none have been designed to handle VFR/HFR video quality measurement. Moreover, in worst of these prior models, the composition of temporal features largely extends spatial features to the time domain, e.g., by applying them on frame differences. There is evidence that simple between-frame measurements are inadequate to capture frame rate induced visual distortions [4, 29, 30]. Therefore, it is of great interest to consider the design of no-reference VQA models applicable to videos capable of capturing temporal aspects of perceptual quality arising from frame rate variations.

3. PROPOSED METHOD

Towards advancing progress in modeling VFR video quality, we have designed a VFR-sensitive BVQA model (FAVER) whose processing is depicted in Fig. 1. FAVER is defined by two branches that respectively contain spatial feature extraction and temporal feature calculation modules. The latter module comprises temporal bandpass filters the responses of which are expressive of the expected statistical regularities of videos, which can be affected by frame rate. Next, we describe the spatial and temporal branches of FAVER.

3.1. Spatial Feature Design

Inspired by the efficacy and modularity of RAPIQUE [6], we leverage the basic 34-dim NSS feature extraction module used in RAPIQUE, applying it on several spatial feature maps to extract rich quality-aware statistical features. The processing flow of the basic feature extraction module is presented in Table 1. Feature subsets can be extracted on versions of the image/frame that have been processing spatially or temporally to arrive at a set of statistical features that account for various aspects of quality perception.

Table 1. Summary of the basic 34-dimensional NSS feature extraction module.

Index	Description	Computation Procedure
$f_1 - f_2$	(α, σ)	Fit GGD to MSCN coefficients
$f_3 - f_4$	$(\phi_\sigma, \rho_\sigma)$	Compute statistics on 'sigma' map
$f_5 - f_8$	$(\nu, \eta, \sigma_l, \sigma_r)$	Fit AGGD to H pairwise products
$f_9 - f_{12}$	$(\nu, \eta, \sigma_l, \sigma_r)$	Fit AGGD to V pairwise products
$f_{13} - f_{16}$	$(\nu, \eta, \sigma_l, \sigma_r)$	Fit AGGD to D1 pairwise products
$f_{17} - f_{20}$	$(\nu, \eta, \sigma_l, \sigma_r)$	Fit AGGD to D2 pairwise products
$f_{21} - f_{22}$	(α, σ)	Fit GGD to D1 pairwise log-derivative
$f_{23} - f_{24}$	(α, σ)	Fit GGD to D2 pairwise log-derivative
$f_{25} - f_{26}$	(α, σ)	Fit GGD to D3 pairwise log-derivative
$f_{27} - f_{28}$	(α, σ)	Fit GGD to D4 pairwise log-derivative
$f_{29} - f_{30}$	(α, σ)	Fit GGD to D5 pairwise log-derivative
$f_{31} - f_{32}$	(α, σ)	Fit GGD to D6 pairwise log-derivative
$f_{33} - f_{34}$	(α, σ)	Fit GGD to D7 pairwise log-derivative

Using the unsmoothed gradient makes it possible to capture very fine details that may affect perceived quality [17, 18]. We estimate the gradient magnitude of each video frame using the Sobel kernel with horizontal and vertical operators: $H_x = [1, 0, -1; 2, 0, -2; 1, 0, -1]$, $H_y = [1, 2, 1; 0, 0, 0; -1, -2, -1]$

The gradient magnitude (GM) of each video frame is then calculated as

$$GM = \sqrt{(I * h_x)^2 + (I * h_y)^2}, \tag{1}$$

where $*$ denotes discrete convolution.

Coarser details are captured using the smoothed twice-derivative Laplacian-of-Gaussian (LOG) operator:

$$LoG = I * h_{LoG}, \tag{2}$$

where the LOG kernel is defined as:

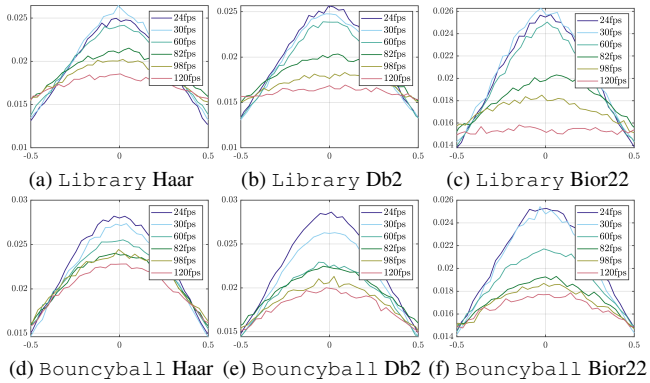
$$\begin{aligned} h_{LOG}(x, y|\sigma) &= \frac{\partial^2}{\partial x^2} g(x, y|\sigma) + \frac{\partial^2}{\partial y^2} g(x, y|\sigma) \\ &= \frac{1}{2\pi\sigma^2} \frac{x^2 + y^2 - 2\sigma^2}{\sigma^4} exp(-\frac{x^2 + y^2}{2\sigma^2}), \end{aligned} \tag{3}$$

where $g_\sigma(x, y)$ is an isotropic Gaussian function and σ is a scale parameter. The window size is 9×9.

The overall flow spatial feature extraction is as follows. As chroma features are also essential to perceptual quality [6, 18, 19, 31, 32], The YUV components of each frame in two scales (the original- and half-scale) are fed into the basic 34-dim NSS feature extractor (Table 1) to obtain low-frequency statistical measurements. The GM and LOG are applied on the Y channel to capture the high-frequency and mid-frequency information descriptive of each frame only at the half scale. The five image maps (Y, U, V, GM, LOG) are then each processed by the NSS-34 feature extraction module, yielding a total of 272 spatial quality-aware features.

3.2. Temporal Feature Design

As we shall see, these prior models perform poorly on VFR distortions, which generally develop over multiple frames.



(a) Library Haar (b) Library Db2 (c) Library Bior22

(d) Bouncyball Haar (e) Bouncyball Db2 (f) Bouncyball Bior22

Fig. 2. Empirical distributions of MSCN normalized coefficients of the fourth subband of the temporal filtered responses on sequences Library and Bouncyball from the LIVE-YT-HFR database.

Table 2. Metadata of the evaluated VQA databases.

Database	#Source videos	Framerate (fps)	CRF	Distortion	#Total videos
BVI-HFR [33]	22	15, 30, 60, 120	None	Framerate	88
LIVE-YT-HFR [2]	16	24, 30, 60, 82, 98, 120	0-63	Framerate, Compression	480

Therefore, inspired by the effective use of temporal bandpass filters in the FR model ST-GREED [4], we take a similar approach, but without access to a reference signal, to build NR-VQA models sensitive to VFR distortions.

Consider a bank of K 1D purely temporal bandpass filters denoted as $b_k, k \in \{0, ..., K-1\}$, where k is the subband index. The temporal bandpass responses of these filters to a video $F(\mathbf{x}, t)$ (where $\mathbf{x} = (x, y)$ and t are spatial and temporal coordinates, respectively) is

$$B_k(\mathbf{x}, t) = V(\mathbf{x}, t) * b_k(t), \quad k \in 1, ..., K, \tag{4}$$

where * is the discrete convolution operator, and B_k is the response of b_k. Of course, frame-differencing is a special case of (4). We compared three types of wavelets in this paper: Haar, Daubechies-2 (DB2), and Biorthogonal-2.2 (bior22).

To illustrate the temporal bandpass statistics of videos having different frame rates, we used the sequences Library and Bouncyball from LIVE-YT-HFR [2] with six variants of framerate (24, 30, 60, 82, 98, 120 fps). Fig. 2 plots the empirical distributions of the spatial MSCN-normalized responses of these videos after temporal Haar, Db2 and Bior22 filtering. As may be seen, the framerate affects the shape and spread of each video's histograms in a systematic way.

Then the 34-dim feature extractor is applied to the temporal subband responses using each of Haar/Db2/Bior22 filter sets (as three different realizations), to extract temporal features at two scales, yielding a total of 476 temporal quality-aware features. Finally, 748 FAVER features are obtained via combination of spatial and temporal feature sets.

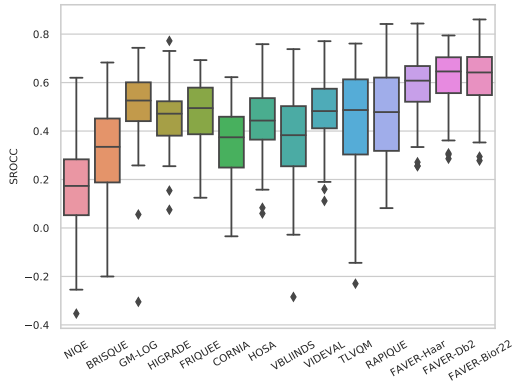


Fig. 3. Boxplots of SROCC distributions over 100 iterations of train/test splits on the LIVE-YT-HFR dataset.

3.3. Feature Learning

Most feature-based BVQA models train a single support vector machine (SVM) to perform feature-to-score mapping. However, we observed that the spatial and temporal features in FAVER have different distributions, hence we adopted ensemble learning [34, 35] to better exploit the bifurcated space/time statistical structure. Specifically, we trained two separate SVM models, one to learn spatial features, and the other to learn temporal features, respectively. FAVER then simply averages [34] the spatial and temporal predictions to produce a video quality final prediction.

4. EXPERIMENTS

4.1. Benchmarks and Evaluation Protocols

We conducted comparisons of the performances of FAVER and other IQA/VQA models on the only available HFR video quality dataset, LIVE-YT-HFR [2], that contains both VFR and compression distortions. We also conducted experiments using BVI-HFR [33], which contains VFR content but without compression variations. The main characteristics of LIVE-YT-HFR and BVI-HFR are summarized in Table 2.

Model Baselines. We studied the performance of FAVER against those of several popular and widely used BVQA models: BRISQUE [15], GM-LOG [17], HIGRADE [18], FRIQUEE [19], CORNIA [36], HOSA [37], NIQE [16]. As in [5,6], features or scores computed using each of these IQA models were computed at a rate of one frame per second, then average pooled [38] across all frames to obtain final video features or scores. We also studied several leading BVQA models: V-BLIINDS [20], VIDEVAL [5], RAPIQUE [6] and TLVQM [23].

Evaluation Protocols. We randomly subdivided the database into training and test sets by comprising approximatedly 80% and 20% of the data following the convention [15, 17, 19, 23]. We repeated the train-test process over

Table 3. Performance comparison of evaluated models on the LIVE-YT-HFR database. The boldfaced entries indicate the top three performers for each performance metric.

Model	SROCC↑	PLCC↑	RMSE↓
NIQE [16]	0.1371	0.4184	11.0853
BRISQUE [15]	0.3190	0.4196	11.0615
GM-LOG [17]	0.4950	0.6049	9.6585
HIGRADE [18]	0.4640	0.5557	10.2196
FRIQUEE [19]	0.4801	0.5723	10.0235
VBLIINDS [20]	0.3917	0.4675	10.8168
VIDEVAL [5]	0.4748	0.5665	10.0547
TLVQM [23]	0.4295	0.5047	10.5108
RAPIQUE [6]	0.4566	0.5665	9.8969
FAVER-Haar	**0.5864**	**0.6546**	**9.1789**
FAVER-Db2	**0.6195**	**0.6769**	**8.9525**
FAVER-Bior22	**0.6350**	**0.6923**	**8.7964**

Table 4. Performance Comparison of all compared models on BVI-HFR.

Model	SROCC↑	PLCC↑	RMSE↓
NIQE [16]	0.2247	0.4194	16.8898
BRISQUE [15]	0.2600	0.4448	16.6404
GM-LOG [17]	0.2778	0.4647	16.3202
HIGRADE [18]	0.2546	0.4180	16.8538
FRIQUEE [19]	0.2387	0.3733	16.8940
VIDEVAL [5]	0.3449	0.4742	16.2954
TLVQM [23]	0.3734	0.4908	16.1876
RAPIQUE [6]	0.3037	0.4631	16.3567
FAVER-Haar	**0.4117**	**0.5212**	**15.7612**
FAVER-Db2	**0.5008**	**0.5872**	**14.8300**
FAVER-Bior22	**0.5560**	**0.6390**	**14.1079**

100 random divisions, over which the exact ratios of train/test data were 13:3 and 17:5 on the LIVE-YT-HFR and BVI-HFR datasets, respectively, while guaranteeing that the training and test sets shared no versions (distorted or otherwise) of any original content. The performance metrics we employed are the Spearman's rank-order correlation coefficient (SROCC), Pearson's linear correlation coefficient (PLCC), and the root mean squared error (RMSE).

4.2. Main Results

We report the performance of all of compared models on the LIVE-YT-HFR dataset, in Table 3. The entire FAVER model (using all features) using Haar, Db2, and Bior22 filters is respectively denoted as FAVER-Haar, FAVER-Db2 and FAVER-Bior22. It may be observed from the Table that the family of FAVER models significantly outperformed the other compared BVQA models by a large margin, with FAVER-Db2 and FAVER-Bior22 delivering the best performance. Fig. 3 shows the spreads of the SROCC values for each evaluated NR-VQA model over the 100 iterations, showing that FAVER-bior22 had a much tighter confidence interval than the other models, highlighting its robustness.

Table 5. Complexity analysis of BVQA models on LIVE-YT-HFR. The time cost (seconds) is measured by averaging the seconds used for feature extraction on ten 1080p videos.

Model	Feat Dim	Time (second/video@1080p) 30fps	60fps	120fps
TLVQM	75	278.43	396.95	684.10
VIDEVAL	60	518.80	984.34	2045.80
RAPIQUE	3884	38.34	37.42	38.09
FAVER-Haar	748	29.86	33.13	32.13
FAVER-Db2	748	48.01	47.79	47.60
FAVER-Bior22	748	56.51	56.63	56.23

Since the BVI-HFR [33] database contains a limited number of videos having largely insignificant variation across framerates, we found it difficult for no-reference models to be fully trained, yielding relatively low correlations as compared to full-reference models [2]. Yet, as summarized in Table 4, FAVER outperformed all the competing BVQA models.

4.3. Runtime Comparison

We also carried out a computational complexity analysis on the evaluated models (measured in CPU for a fair comparison). We accounted for the video framerates, and recorded the time costs at 30fps, 60 fps, and 120 fps. Table 5 tabulates the feature dimension and actual time cost against framerate of the compared BVQA models. As it may be seen, the three FAVER variants are efficient as compared to other top-performing BVQA models like TLVQM and VIDEVAL. FAVER is **6x-9x** faster than TLVQM, and **9x-17x** faster than VIDEVAL on 1080p@30fps videos. It is also worth mentioning that FAVER maintains a similar computational cost as the video framerate is increased, while those of TLVQM and VIDEVAL scales linearly with respect to framerate. Moreover, FAVER obtained comparable complexity as the most efficient model, RAPIQUE, but it utilizes far fewer features while delivering significantly better performance.

5. CONCLUSION

We proposed the first robust, effective and efficient BVQA model focused on perceptual quality assessment of VFR videos, dubbed FAVER. FAVER utilizes the temporal natural video statistics of bandpass filtered videos to capture and represent aspects of temporal video quality. FAVER is trained using an ensemble of learners to temporal and spatial quality scores which are then combined into a final quality score. Our extensive experimental results show that FAVER delivers accurate and stable predictions of video quality on HFR/VFR database against human judgments, exceeding the performances of previous models. We believe this work will facilitate and inspire future research efforts on the quality assessment and intelligent compression of VFR videos.

6. REFERENCES

[1] A. Mackin, F. Zhang, and D. R. Bull, "A study of high frame rate video formats," *IEEE Trans. Multimedia.*, vol. 21, no. 6, pp. 1499–1512, 2018.

[2] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," *IEEE Access*, vol. 9, pp. 108 069–108 082, 2021.

[3] D. Y. Lee, H. Ko, J. Kim, and A. C. Bovik, "Space-time video regularity and visual fidelity: Compression, resolution and frame rate adaptation," *arXiv preprint arXiv:2103.16771*, 2021.

[4] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "ST-GREED: Space-time generalized entropic differences for frame rate dependent video quality prediction," *IEEE Transactions on Image Processing*, vol. 30, pp. 7446–7457, 2021.

[5] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. Image Process.*, vol. 30, pp. 4449–4464, 2021.

[6] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "RAPIQUE: Rapid and accurate video quality prediction of user generated content," *IEEE Open J. Signal Process.*, vol. 2, pp. 425–440, 2021.

[7] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-VQ: 'patching up' the video quality problem," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2021, pp. 14 019–14 029.

[8] N. D. Narvekar and L. J. Karam, "A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection," in *2009 International Workshop on Quality of Multimedia Experience*, 2009, pp. 87–91.

[9] R. Hassen, Z. Wang, and M. M. A. Salama, "Image sharpness assessment based on local phase coherence," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2798–2810, 2013.

[10] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. IEEE Int. Conf. Image Process.*, vol. 1, 2002, pp. I–I.

[11] Z. Wang, A. C. Bovik, and B. L. Evan, "Blind measurement of blocking artifacts in images," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3, 2000, pp. 981–984.

[12] Z. Tu, J. Lin, Y. Wang, B. Adsumilli, and A. C. Bovik, "Bband index: a no-reference banding artifact predictor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 2712–2716.

[13] D. L. Ruderman, "The statistics of natural images," *Netw.: Comput. Neural Syst.*, vol. 5, no. 4, pp. 517–548, 1994.

[14] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, 2015.

[15] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012.

[16] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[17] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, 2014.

[18] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference quality assessment of tone-mapped HDR pictures," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2957–2971, 2017.

[19] D. Ghadiyaram, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17(1), no. 32, pp. 1–25, 2017.

[20] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, 2014.

[21] Z. Tu, C.-J. Chen, L.-H. Chen, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Regression or classification? new methods to evaluate no-reference picture and video quality models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2021, pp. 2085–2089.

[22] X. Li, Q. Guo, and X. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3329–3342, 2016.

[23] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5923–5938, 2019.

[24] L.-H. Chen, C. G. Bampis, Z. Li, A. Norkin, and A. C. Bovik, "Proxiqa: A proxy approach to perceptual optimization of learned image compression," *IEEE Trans. Image Process.*, vol. 30, pp. 360–373, 2020.

[25] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.

[26] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. ACM Multimedia Conf.*, 2019, pp. 2351–2359.

[27] W. Liu, Z. Duanmu, and Z. Wang, "End-to-end blind quality assessment of compressed videos using deep neural networks," in *Proc. ACM Multimedia Conf. (MM)*, 2018, pp. 546–554.

[28] Y. Wang, J. Ke, H. Talebi, J. G. Yim, N. Birkbeck, B. Adsumilli, P. Milanfar, and F. Yang, "Rich features for perceptual quality assessment of UGC videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2021, pp. 13 435–13 444.

[29] R. M. Nasiri and Z. Wang, "Perceptual aliasing factors and the impact of frame rate on video quality," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3475–3479.

[30] R. M. Nasiri, Z. Duanmu, and Z. Wang, "Temporal motion smoothness and the impact of frame rate variation on video quality," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1418–1422.

[31] L.-H. Chen, C. G. Bampis, Z. Li, J. Sole, and A. C. Bovik, "Perceptual video quality prediction emphasizing chroma distortions," *IEEE Trans. Image Process.*, vol. 30, pp. 1408–1422, 2020.

[32] S.-C. Pei and L.-H. Chen, "Image quality assessment using human visual dog model fused with random forest," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3282–3292, 2015.

[33] A. Mackin, F. Zhang, and D. R. Bull, "A study of high frame rate video formats," *IEEE Trans. Multimedia.*, vol. 21, no. 6, pp. 1499–1512, 2019.

[34] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Trans. Circuits Syst. Video Technol*, vol. 29, no. 8, pp. 2256–2270, 2018.

[35] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[36] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.

[37] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, 2016.

[38] Z. Tu, C. J. Chen, L. H. Chen, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "A comparative evaluation of temporal pooling methods for blind video quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 141–145.