A Completely Blind Video Quality Evaluator

Qi Zheng, Zhengzhong Tu, *Graduate Student Member, IEEE*, Xiaoyang Zeng, Alan C. Bovik, *Fellow, IEEE*, and Yibo Fan

***Abstract*—Automatic video quality assessment of user-generated content (UGC) has gained increased interest recently, due to the ubiquity of shared video clips uploaded and circulated on social media platforms across the globe. Most existing video quality models developed for this vast content are trained on large numbers of samples labeled during large-scale subjective studies, which are often fail to exhibit adequate generalization abilities on unseen data. Thus, it is also desirable to develop opinion-unaware, "completely blind" video quality models, that are free of training, yet can compete with existing learning-based models. Here we propose such a model called VIQE (VIdeo Quality Evaluator), which we designed based on a comprehensive analysis of patch- and frame-wise video statistics, as well as of space-time statistical regularities of videos. The statistical features desired from the analysis capture complementary predictive aspects of perceptual quality, which are aggregated to obtain final video quality scores. Extensive experiments on recent large-scale video quality databases demonstrate that VIQE is even competitive with state-of-the-art opinion-aware models. The source code is being made available at https://github.com/uniqzheng/VIQE.**

***Index Terms*—Completely blind, video quality assessment, user-generated content, natural scene statistics, linear model.**

I. Introduction

GIVEN the pervasiveness of social media platforms like YouTube, Facebook, Instagram, and TikTok, it has become quite important that these providers be able to monitor, analyze, and control the perceptual quality of the massive amount of user-generated content (UGC) videos that are now shared and streamed across the globe. Moreover, UGC videos often suffer from wide ranges of types and severities of mixtures of in-capture and/or post-capture distortions, which if unaddressed, can greatly detract from viewers' quality of experience (QoE). Objective video quality assessment tools are already widely used to automatically monitor and evaluate video codecs, communication systems, and quality enhancement algorithms. Reference video quality assessment (R-VQA) models measure the *perceptual* difference between high-quality original reference videos and their distorted counterparts. In many practical scenarios, however, 'pristine' reference signals are unavailable. In fact, this is generally the case when designing VQA models for UGC videos [1]. In such circumstances, only blind (no-reference or NR) VQA models can be used to predict the perceptual quality of videos, since there is no possible access to the presumed 'pristine' videos.

Recently, data-driving methods such as deep neural networks have been shown to deliver superior performance on numerous computer vision tasks [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. Likewise, the majority of existing blind video quality assessment (BVQA) methods belong to the "opinion aware" (learning-based) category, wherein a learned regression model, either deep or shallow, is trained on databases of distorted videos that have been human-labeled in the form of mean opinion scores (MOS) [1], [12], [13], [14], [15], [16], [17], [18], [19], [20]. However, real-world user-generated videos often suffer from multiple commingled, unpredictable distortions that can interact to create new distortions, and that are impossible to model and difficult to populate in video quality databases. This makes these kinds of opinion-aware (OA) BIQA/BVQA models often suffer from unsatisfactory generalization capability in practical commercial scenarios. Therefore, it is important to also study "opinion-unaware" (OU) or "completely blind" models that do not rely on training on human-labeled videos. High-performing OU models instead predict perceptual video quality by measuring statistical shifts between naturalistic and distorted versions of them. In this regard, OU BVQA models are distortion agnostic, and hence are potentially more generalizable to new, unseen distortions, and mixtures of systems often encountered in UGC video streaming and sharing.

There has been previous work done on opinion-unaware (OU) BIQA/BVQA models [21], [22], [23], [24], [25], [26], [27]. Among them, NIQE [21] and IL-NIQE [22] were designed using perception-inspired statistical features that have been empirically observed to reliably follow natural scene statistics (NSS) models. Quality predictions are computed by measuring statistical distances between the distributions of distorted images and those of high-quality, natural images. Following the design framework of NIQE [21], NPQI [26] explores NSS features from a local binary map and the locally normalized coefficients of images, while SNP-NIQE [27] measures structural variations as well as naturalness deviations. VIIDEO [23] models the temporal regularities of natural videos, using them to assess video quality. SLEEQ [24] is another OU video quality predictor based on 'self-referenced' features, and was specifically designed for compression and scaling artifacts. A more recent completely blind BVQA models targeting UGC videos, called STEM [25], quantifies losses of "perceptual straightness" [28] to measure temporal quality.

Manuscript received 6 September 2022; accepted 6 October 2022. Date of publication 17 October 2022; date of current version 7 November 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62031009, in part by Alibaba Innovative Research (AIR) Program, in part by Fudan University-CIOMP Joint Fund under Grant FC2019-001, in part by Fudan-ZTE Joint Lab, in part by the Pioneering Project of Academy for Engineering and Technology Fudan University under Grant gyy2021-001, and in part by the CCF-Alibaba Innovative Research Fund For Young Scholars. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Lu Gan. *(Corresponding author: Yibo Fan.)*

Qi Zheng, Xiaoyang Zeng, and Yibo Fan are with the State Key Laboratory of ASIC & System, College of Microelectronics, Fudan University, Shanghai 200000, China (e-mail: qzheng21@m.fudan.edu.cn; xyzeng@fudan.edu.cn; fanyibo@fudan.edu.cn).

Zhengzhong Tu and Alan C. Bovik are with the Laboratory for Image and Video Engineering (LIVE), Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: zhengzhong.tu@utexas.edu; bovik@ece.utexas.edu).

Digital Object Identifier 10.1109/LSP.2022.3215311

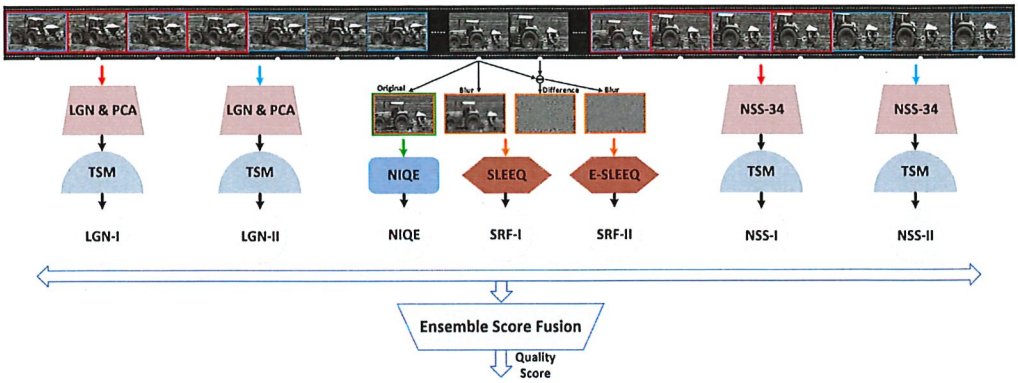

Fig. 1. Illustration of the overall processing flow of the VIQE model.

These previous completely blind OU algorithms are designed either targeting quality perception of limited distortion types, or by modeling only a few specific statistical regularities. Therefore, none of them have been able to deliver accurate predictions of human judgments (MOS) of video quality on the most recent large-scale UGC video datasets [29], [30], [31]. Towards advancing progress on this problem, we have developed an effective and efficient, opinion-unaware, 'completely blind' BVQA model that we call VIQE, that integrates complementary video quality factors expressive of multiple of perception, including multi-scale perceptual straightening of statistical regularities, enhanced self-referencing features, and NSS features. Our experiments show that VIQE significantly outperforms previous state-of-the-art (SOTA) OU BVQA models on recent UGC datasets.

The rest of the paper is organized as follows. Section II details the proposed VIQE model. Experimental results are presented in Section III, and Section IV concludes the paper.

II. Proposed Method

Fig. 1 illustrates the overall modular processing flow of VIQE. It comprises four sub-modules. One model measures the temporal quality of the video using two features derived from a multi-scale temporal straightening model (TSM). Another measures spatial quality using NIQE [21] in a computationally efficient way. Inspired by the self-referencing strategy in [24], we compute two sets of robust statistical features, which we refer to as SRF-I and SRF-II, which we use to compute self-referencing scores (SLEEQ [24] and Enhanced SLEEQ, or E-SLEEQ). Finally, we aggregate these perceptually-inspired quality indicators by fusing them into a single predictor of the overall quality of the video.

A. Natural Scene Statistics Feature Extractor

The study of natural scene statistics (NSS) [32], and how they are altered by distortion, has inspired a number of popular BIQA/BVQA models [13], [14], [15], [20], [23], [33]. NSS-derived statistical features have been defined that deliver strong quality prediction performance on diverse image spaces, such as bandpass luminance [23], bandpass chroma [15], spatial gradients [13] and laplacians [13], [15], and temporal bandpass responses [34]. Similar to [20], we devise a module that computes strong, well-grounded NSS features on multiple visual domains. A summary of the 34 features that are extracted by the module, which we will refer to as NSS-34, is presented in Table I. We also identify three submodules: NSS-2 ($f_1 - f_2$), NSS-18 ($f_1 - f_2, f_5 - f_{20}$) and NSS-34 ($f_1 - f_{34}$), which are used to capture different and complementary aspects of video quality, as explained in the following.

TABLE I
Summary of the 34-Dim NSS Feature Extractor

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

B. Multi-Scale Perceptual Straightening

Henaff et al. [28] hypothesized that the visual system transforms incoming streams of visual input, making them more predictable by a process of "perceptual straightening". In this model, a cascade of retinal [35] and cortical representations are computed and used to straighten the temporal trajectories of the video, enabling prediction via linear extrapolation. In other words, for a high-quality video, the perceptual representation of a present frame can be linearly extrapolated from those of previous frames.

We map each input video onto two types of perceptual representations to achieve better predictability: 1) a bandpass "LGN" model (model of lateral geniculate nucleus) [25], [28], and 2) NSS-inspired statistical models [20]. Fig. 2 depicts the processing of the LGN model, which performs spatial bandpass filtering (gray boxes) followed by nonlinear luminance and contrast gain control [28]. The input image is decomposed into six scales of bandpass difference of Gaussian (DoG) filters. The features from the sixth scale are used, as in [25], [35].

To understand the straightening process, consider a video having N frames denoted by $\mathbf{X} = [X_1, X_2, X_3, \ldots X_N]$. Both the LGN model and NSS-34 are computed on all the video frames, yielding sets of frame-wise features that we denote as f_{LGN} and $F_{\text{NSS-34}}$, respectively. Then project f_{LGN} into F_{LGN},

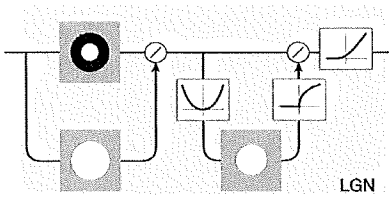

Fig. 2. Processing flow of the LGN model [28].

a lower dimensional space, using principal component analysis (PCA). Then train a linear extrapolation model to predict the straightened trajectories at every fifth frame, to conserve computation. A third-order extrapolator [28] is fitted to F_{LGN} and $F_{\text{NSS-34}}$ over two temporal scales (one original, the other downsampled by two) to model multi-scale straightening:

$$\begin{aligned}
\hat{F}^{t+3}_{\text{LGN-I}} &= \beta_0 + \beta_1 F^{t}_{\text{LGN}} + \beta_2 F^{t+1}_{\text{LGN}} + \beta_3 F^{t+2}_{\text{LGN}},\\
\hat{F}^{t+6}_{\text{LGN-II}} &= \beta_0 + \beta_1 F^{t}_{\text{LGN}} + \beta_2 F^{t+2}_{\text{LGN}} + \beta_3 F^{t+4}_{\text{LGN}},\\
\hat{F}^{t+3}_{\text{NSS-I}} &= \beta_0 + \beta_1 F^{t}_{\text{NSS-34}} + \beta_2 F^{t+1}_{\text{NSS-34}} + \beta_3 F^{t+2}_{\text{NSS-34}},\\
\hat{F}^{t+6}_{\text{NSS-II}} &= \beta_0 + \beta_1 F^{t}_{\text{NSS-34}} + \beta_2 F^{t+2}_{\text{NSS-34}} + \beta_3 F^{t+4}_{\text{NSS-34}}, \qquad (1)
\end{aligned}$$

where the perceptual features at previous timestamps F^t, F^{t+1}, F^{t+2} are fed into the model, which uses them to predict the current feature $\hat{F}^{t+3}$. The two scales, indexed I and II, are arrived at by first gaussian low-pass filtering, then downsampling by factor two, exactly as in BRISQUE [12] but in the 1D time direction. The weights $\beta_i|_{i=0}^{i=3}$ are estimated using multiple least-squares regression. The distances between the ground truth and predicted features are expressed by the root mean squared error (RMSE), averaged over the entire video:

$$\begin{aligned}
Q_{\text{LGN-I}} &= \log\left(\frac{1}{n}\sum_{j=0}^{n} D^{j}_{\text{LGN-I}}\right),\\
Q_{\text{LGN-II}} &= \log\left(\frac{1}{n}\sum_{j=0}^{n} D^{j}_{\text{LGN-II}}\right),\\
Q_{\text{NSS-I}} &= \log\left(\frac{1}{n}\sum_{j=0}^{n} D^{j}_{\text{NSS34-I}}\right),\\
Q_{\text{NSS-II}} &= \log\left(\frac{1}{n}\sum_{j=0}^{n} D^{j}_{\text{NSS34-II}}\right), \qquad (2)
\end{aligned}$$

where D^j is the prediction error at timestamp j, representing deviations from the temporal straightening hypothesis, which are powerful temporal quality indicators.

C. Enhanced Self-Reference Quality

Self-referenced video quality models have been shown to effectively capture inter-dependencies between the features of original videos and their blurry variants [24], but without reliance on any reference signals. This idea is highly relevant to the UGC-VQA problem, where the underlying content may be distorted by combinations of any number of unknown distortions. Hence, another module computes feature-enriched self-referenced quality scores, which we will refer to as enhanced SLEEQ, or E-SLEEQ, since it builds on the model in [24].

Given an input frame f with its frame-difference d (relative to the previous frame), apply a 2-D Gaussian smoothing filter to obtain blur versions f' and d'. Following SLEEQ [24], then divide the four frames (including before and after blur) into patches, then apply NSS-2 and NSS-18 on these patches, respectively. Thus, two sets of NSS features are then extracted from each patch: $(S_{\text{NSS-2}}, S_{\text{NSS-18}})$ and $(S'_{\text{NSS-2}}, S'_{\text{NSS-18}})$ are spatial features extracted from f and f', while $(T_{\text{NSS-2}}, T_{\text{NSS-18}})$ and $(T'_{\text{NSS-2}}, T'_{\text{NSS-18}})$ are temporal features computed from d and d'. Then, the absolute differences of the spatial and temporal features are computed:

$$\begin{aligned}
\Delta S_{\text{NSS-2}} &= \left|S_{\text{NSS-2}} - S'_{\text{NSS-2}}\right|,\\
\Delta S_{\text{NSS-18}} &= \left|S_{\text{NSS-18}} - S'_{\text{NSS-18}}\right|,\\
\Delta T_{\text{NSS-2}} &= \left|T_{\text{NSS-2}} - T'_{\text{NSS-2}}\right|,\\
\Delta T_{\text{NSS-18}} &= \left|T_{\text{NSS-18}} - T'_{\text{NSS-18}}\right|. \qquad (3)
\end{aligned}$$

To account for temporal masking effects arising from large motions [36], the self-reference-based spatial and temporal scores of each patch location P are further weighted:

$$\begin{aligned}
Q^{(P)}_{\text{SRF-I}} &= \left(1 - m^{(P)}\right)\cdot \Delta S_{\text{NSS-2}} + m^{(P)}\cdot \Delta T_{\text{NSS-2}},\\
Q^{(P)}_{\text{SRF-II}} &= \left(1 - m^{(P)}\right)\cdot \Delta S_{\text{NSS-18}} + m^{(P)}\cdot \Delta T_{\text{NSS-18}}, \qquad (4)
\end{aligned}$$

where $m^{(p)}$ is the normalized average frame difference of patch location P. Given that the version system is highly sensitive to spatial change, only those patches where the standard deviation $\Delta\overline{\sigma}_p$ lies within the q^{th} percentile overall patches in the video are chosen, where $q = 90 - 5\left\lceil\frac{W}{768}\right\rceil\left\lceil\frac{H}{432}\right\rceil$ (H, W: width, height). Then, the final quality scores $Q_{\text{SRF-I}}$ and $Q_{\text{SRF-II}}$ are found by average-pooling [37] the $Q^P_{\text{SRF-I}}$ and $Q^P_{\text{SRF-II}}$ scores over all patches of the video.

D. Spatial Naturalness Quality

To enhance the spatial quality estimation, we employ the powerful and popular completely blind IQA algorithm NIQE [21], using its responses as spatial quality features. We have observed that applying NIQE on only a single frame each second performance delivers the same performance, but with greatly reduced cost. A spatial quality feature Q_{NIQE} is attained by averaging NIQE across all the sampled frames.

E. Score Fusion

All the models described above are capable of capturing certain aspect of perceptual quality, either mostly spatial or temporal; but we have observed that none of these models is able to generalize well on unseen databases, especially challenging UGC video datasets [1]. UGC videos often contain complex combinations of spatial-time distortions, which are beyond the prediction capabilities of existing completely-blind quality models. Here, we deploy a simple model fusion approach [38] to combine multiple aspects of video quality. Similar approaches have been shown to deliver promising results on other video quality tasks [39], [40], [41], [42]. Specifically, we employ a simple weighted sum of all seven quality indices to define the

TABLE II
PERFORMANCE COMPARISON OF EVALUATED MODELS ON THREE PUBLICLY AVAILABLE DATASETS

DATASET	KoNViD-1k [20]			LIVE-VQC [19]			YouTube-UGC [21]		
MODEL \ METRIC	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE
VBLIINDS [6]	0.5720	0.5650	0.5260	0.6916	0.7150	11.8693	0.5327	0.5290	0.5451
TLVQM [9]	0.7687	0.7652	0.4125	0.7870	0.7913	10.3543	0.6738	0.6648	0.4799
PaQ-2-PiQ [26]	0.6130	0.6014	0.5148	0.6436	0.6683	12.6190	0.2658	0.2935	0.6153
NIQE [11]	0.5456	0.5625	0.5291	0.5912	0.6327	13.0796	0.2217	0.2745	0.6174
IL-NIQE [13]	0.5120	0.5303	0.5423	0.4842	0.5316	14.2884	0.2910	0.3234	0.6071
NPQI [16]	0.4519	0.4755	0.5628	0.5425	0.5787	13.7750	0.3567	0.3887	0.5916
SNP-NIQE [17]	0.5472	0.5648	0.5281	0.6254	0.6647	12.6230	0.2677	0.3210	0.6077
VIIDEO [13]	0.2988	0.3002	0.6101	0.0332	0.2146	16.6540	0.0580	0.1534	0.6339
STEM [15]	0.6193	0.6266	0.4985	0.5938	0.6292	13.1397	0.2840	0.3180	0.6359
VIQE	**0.6284**	**0.6380**	**0.4924**	**0.6598**	**0.6943**	**12.1517**	**0.5130**	**0.4769**	**0.5616**

The boldfaced entries indicate the top performer among the opinion-unaware (completely blind) models. The italicized entries indicate supervised models.

final VIQE model:

$$Q_{\text{VIQE}} = \sum_{i=1}^{N} w_i \cdot Q_i, \tag{5}$$

where i indexes the quality models in [LGN-I, LGN-II, NSS-I, NSS-II, SRF-I, SRF-II, NIQE], and w represents the learned weight for each aspect of quality followed by SLEEQ [24].

III. EXPERIMENTAL RESULTS

We conducted extensive experiments to compare the performances of VIQE against other opinion-unaware (OU) BVQA models on the three existing UGC video datasets: LIVE-VQC [29], KoNViD-1k [30] and YouTube-UGC [31]. We evaluated six OU BIQA/BVQA algorithms: NIQE [21], IL-NIQE [23], NPQI [21], SNP-NIQE [23], VIIDEO [23], and STEM [25]. We also included three leading opinion-aware (OA) blind IQA/VQA models, two handcrafted models TLVQM [19] and VBLIINDS [16], and a deep learning model PaQ-2-PiQ [16].

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). Following convention, we randomly divided the database into training and test sets comprising approximately 80% and 20% of the data 50 times, and report the median evaluation results on the test partitions. Since the data distributions of the three evaluated UGC datasets vary significantly, we conducted a grid-search on the ensemble weights to select the best performing model on each dataset.

The performances of all the compared models are shown in Table II. As may be seen, VIQE significantly improves SROCC upon the best performances of state-of-the-art OU BIQA/BVQA (completely blind) models by 76.3%, 5.5%, and 1.5% on the YouTube-UGC, LIVE-VQC, and KoNViD-1 k datasets, respectively. It is also worth mentioning that VIQE delivered performance comparable to that of OA BIQA/BVQA models (trained on MOS labels) on the LIVE-VQC [29] and YouTube-UGC [31], and even surpassed some leading learning-based models on KoNViD-1k [30]. Since VIQE does not require training (hence is not subject to dataset bias), VIQE may be able to better generalize to future, unseen data.

We also studied the computational complexity of VIQE in terms of CPU processing time on an AMD Ryzen 7 4800 U equipped with a Radeon Graphics@1.80 GHz processor and 16 G RAM. As may be observed in Table III, VIQE is much faster than the OU STEM model and the state-of-the-art OA deep learning based PaQ-2-PiQ model, while delivering better prediction accuracy. To further understand the contributions of the individual components of VIQE, we also conducted an ablation study of VIQE feature importance on LIVE-VQC [29]. Table IV tabulates the SROCC performance of each of the VIQE feature subsets against processing time when sequentially adding features starting from the empty set. It may be observed that all of the features contribute to the overall performance of VIQE, while the overall fused VIQE yields the best results at only a moderate computational cost.

TABLE III
AVERAGE RUNTIME COMPARISON OF 1080p VIDEOS

MODEL	TIME IN SEC
TLVQM [9]	256.5
VBLIINDS [6]	1968.8
PaQ-2-PiQ [6]	279.32
NIQE [11]	214.9
IL-NIQE [13]	906.2
NPQI [11]	2712.6
SNP-NIQE [13]	5634.1
VIIDEO [13]	674.8
STEM [15]	231.3
VIQE	200.6

TABLE IV
CONTRIBUTION ANALYSIS ON LIVE-VQC [29]

SCORE	SROCC	RUNTIME
SRF	0.3863	33.9
SRF+NIQE	0.5892	41.9
SRF+NIQE+LGN	0.6307	84.0
SRF+NIQE+LGN+NSS	0.6598	200.6

IV. CONCLUSION

We have described an opinion-unaware, "completely blind" video quality model called VIQE that significantly outperforms all previous such models on the challenging UGC video quality prediction task. VIQE employs several well-defined, perceptually-inspired quality-aware features that analyze the patch-wise, frame-wise, and space-time statistics of potentially distorted videos. Extensive experiments have shown that VIQE delivers superior performance on large-scale UGC datasets, and is comparable to learning-based models but with better compute efficiency.

REFERENCES

[1] 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.

[2] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3575–3585.

[3] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-VQ:'patching up'the video quality problem," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14019–14029.

[4] 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, 2021.

[5] Z. Tu et al., "MAXIM: Multi-axis MLP for image processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5769–5780.

[6] L. Yang et al., "Diffusion models: A comprehensive survey of methods and applications," 2022, *arXiv:2209.00796*.

[7] R. Xu et al., "Pik-fix: Restoring and colorizing old photo," 2022, *arXiv:2205.01902*.

[8] Z. Fan, Y. Zhu, Y. He, Q. Sun, H. Liu, and J. He, "Deep learning on monocular object pose detection and tracking: A comprehensive overview," *ACM Comput. Surv.*, 2021.

[9] Z. Meng, R. Xu, and C. M. Ho, "Gia-Net: Global information aware network for low-light imaging," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 327–342.

[10] X. Dong, J. Guo, A. Li, W.-T. Ting, C. Liu, and H. Kung, "Neural mean discrepancy for efficient out-of-distribution detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19217–19227.

[11] Z. Tu et al., "MaxViT: Multi-axis vision transformer," 2022, *arXiv:2204.01697*.

[12] 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, Dec. 2012.

[13] 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, Nov. 2014.

[14] 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, Jun. 2017.

[15] D. Ghadiyaram, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, 2017, Art. no. 32.

[16] 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, Mar. 2014.

[17] Z. Tu et al., "Regression or classification? new methods to evaluate no-reference picture and video quality models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 2085–2089.

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

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

[20] 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.

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

[22] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[23] 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, Jan. 2016.

[24] D. Ghadiyaram, C. Chen, S. Inguva, and A. Kokaram, "A no-reference video quality predictor for compression and scaling artifacts," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 3445–3449.

[25] P. Kancharla and S. S. Channappayya, "Completely blind quality assessment of user generated video content," *IEEE Trans. Image Process.*, vol. 31, pp. 263–274, 2021.

[26] Y. Liu, K. Gu, X. Li, and Y. Zhang, "Blind image quality assessment by natural scene statistics and perceptual characteristics," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 3, pp. 1–91, 2020.

[27] Y. Liu et al., "Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 929–943, Apr. 2020.

[28] O. J. Hénaff, R. L. Goris, and E. P. Simoncelli, "Perceptual straightening of natural videos," *Nature Neurosci.*, vol. 22, no. 6, pp. 984–991, 2019.

[29] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 612–627, Feb. 2019.

[30] V. Hosu et al., "The konstanz natural video database (KoNViD-1 k)," in *Proc. 9th Int. Conf. Qual. Multimedia Experience*, 2017, pp. 1–6.

[31] Y. Wang, S. Inguva, and B. Adsumilli, "YouTube UGC dataset for video compression research," in *Proc. IEEE 21st Int. Workshop Multimedia Signal Process.*, 2019, pp. 1–5.

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

[33] Q. Zheng, Z. Tu, P. C. Madhusudana, X. Zeng, A. C. Bovik, and Y. Fan, "Faver: Blind quality prediction of variable frame rate videos," 2022, *arXiv:2201.01492*.

[34] Q. Zheng, Z. Tu, Y. Fan, X. Zeng, and A. C. Bovik, "No-reference quality assessment of variable frame-rate videos using temporal bandpass statistics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 1795–1799.

[35] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelli, "Perceptual image quality assessment using a normalized Laplacian pyramid," *Electron. Imag.*, vol. 2016, no. 16, pp. 1–6, 2016.

[36] L. K. Choi and A. C. Bovik, "Video quality assessment accounting for temporal visual masking of local flicker," *Signal Process.: image Commun.*, vol. 67, pp. 182–198, 2018.

[37] 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.

[38] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits Syst. Mag.*, vol. 6, no. 3, pp. 21–45, Jul.–Sep. 2006.

[39] 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, Aug. 2019.

[40] 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, Nov. 2015.

[41] 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, 2021.

[42] A. K. Venkataramanan, C. Stejerean, and A. C. Bovik, "Funque: Fusion of unified quality evaluators," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 2147–2151.