User-Generated Video Quality Assessment: A Subjective and Objective Study

Yang Li, Shengbin Meng, Xinfeng Zhang, *Senior Member, IEEE*, Meng Wang, Shiqi Wang, *Senior Member, IEEE*, Yue Wang, and Siwei Ma, *Senior Member, IEEE*

***Abstract*—Recently, we have observed an exponential increase of user-generated content (UGC) videos. The distinguished characteristic of UGC videos originates from the video production and delivery chain, as they are usually acquired and processed by non-professional users before uploading to the hosting platforms for sharing. As such, these videos usually undergo multiple distortion stages that may affect visual quality before ultimately being viewed. Inspired by the increasing consensus that the optimization of the video coding and processing shall be fully driven by the perceptual quality, in this paper, we propose to study the quality of the UGC videos from both objective and subjective perspectives. We first construct a UGC video quality assessment (VQA) database, aiming to provide useful guidance for the UGC video coding and processing in the hosting platform. The database contains source UGC videos uploaded to the platform and their transcoded versions that are ultimately enjoyed by end-users, along with their subjective scores. Furthermore, we develop an objective quality assessment algorithm that automatically evaluates the quality of the transcoded videos based on the corrupted reference, which is in accordance with the application scenarios of UGC video sharing in the hosting platforms. The information from the corrupted reference is well leveraged and the quality is predicted based on the inferred quality maps with deep neural networks (DNN). Experimental results show that the proposed method yields superior performance. Both subjective and objective evaluations of the UGC videos also shed lights on the design of perceptual UGC video coding.**

***Index Terms*—Deep neural network, user-generated content, video quality assessment.**

I. Introduction

VIDEO content is historically created by professional content producers. Recently, with the development of multimedia and network technologies, as well as the advances of acquisition devices, there has been an explosion of user-generated content (UGC) videos and related sharing services. Enormous videos generated without professional routines and practices are uploaded to sharing platforms such as Facebook, YouTube and TikTok. Comparing to professionally-generated content (PGC) videos, the low barriers in video production and sharing make the UGC content extremely diverse. In particular, the lack of proper shooting skills and professional video capture equipment make the perceptual quality of UGC videos even worse. Besides, special effects are sometimes incorporated to enhance the user experience, thereby increasing the difficulty of quality assessment and compression. UGC videos generally undergo multi-distortion stages before ultimately viewed by end-users. As captured by non-professional users, in-capture distortions such as noise, shaking and under/over-exposed may be induced, then these videos are compressed and uploaded to hosting platforms, and finally the uploaded videos are transcoded according the requirements of the hosting platform hence another round of compression distortion is induced. Exponential increase in the demand for high-quality videos poses great challenges in practice. As such, effective UGC video quality assessment (VQA) algorithms become critical to guide the optimization of the hosting platform, in an effort to deliver videos with better visual quality under limited bandwidth.

Traditional reference-based quality assessment methods, including full-reference (FR) and reduced-reference (RR) models, are generally designed based on pristine sources, such that the quality of the distorted video can be predicted by signal or feature level comparisons. However, straightforwardly applying this strategy to UGC videos is problematic, as the source videos in the hosting platform have already been corrupted due to acquisition and compression distortions introduced before uploading to the hosting platform. As such, the reference-based algorithms may be mislead by the distorted reference and fail to predict the quality of the ultimately viewed UGC videos. One extreme example is that an excessively high bit rate is applied to transcode a video with extremely poor quality. In this scenario, the objective reference quality is not consistent with the subjective quality due

Manuscript received 27 February 2021; revised 15 August 2021; accepted 7 October 2021. Date of publication 29 October 2021; date of current version 13 January 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62025101, 61961130392, and 61931014, in part by the National Key Research and Development Project under Grant 2019YFF0302703, in part by Hong Kong ITF UICP under Grant 9440203, in part by Fundamental Research Funds for the Central Universities, and PKU-Baidu Fund under Grant 2019BD003, and in part by High performance Computing Platform of Peking University. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Manoranjan Paul. *(Corresponding author: Prof. Siwei Ma.)*

Yang Li and Siwei Ma are with the Institute of Digital Media, Peking University, Haidian District, Beijing 100871, China (e-mail: liyang.00@pku.edu.cn; swma@pku.edu.cn).

Xinfeng Zhang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: xfzhang@ucas.ac.cn).

Meng Wang and Shiqi Wang are with the Department of Computer Science, City University of Hong Kong, Kwoloon, China (e-mail: mwang98-c@my.cityu.edu.hk; shiqwang@cityu.edu.hk).

Shengbin Meng and Yue Wang are with the VideoArch Department, Bytedance Inc., Haidian District, Beijing 100871, China (e-mail: mengshengbin@bytedance.com; wangyue.v@bytedance.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2021.3122347.

Digital Object Identifier 10.1109/TMM.2021.3122347

to the high similarity with the corrupted reference. Besides, relying on no-reference (NR) algorithms only may omit the useful reference information, and may not be able to ensure the accurate prediction with high robustness on such diverse content.

In our previous work [1], we create a database to facilitate the study of quality assessment of UGC videos where subjective ratings of source UGC videos and associated transcoded versions are involved. Based on the UGC database, in this paper, we propose a corrupted-reference-based quality assessment framework which delivers accurate predictions of the perceptual quality for the transcoded UGC videos. The proposed algorithm measures the perceptual quality by combining the local distortions of the source and transcoded videos relying on the prediction of the quality maps. In particular, the quality maps are predicted in a data-driven manner, and fused through a learned network such that the overall quality score is estimated by gradually pooled features. The three main contributions of our work are as follows.

1) We construct a dedicated exploration database for UGC videos, including the source and transcoded videos in the hosting platforms, as well as their subjective scores. We further demonstrate that innovative quality assessment approaches should be developed based on careful investigations.
2) We propose a novel corrupted-reference VQA method for UGC videos based on deep neural networks (DNNs). In contrast with traditional FR quality models, the intrinsic quality of the corrupted-reference is incorporated to accurately infer the quality.
3) We show that the performance of the proposed framework outperforms the state-of-the-art methods in the application domain of UGC video processing. While the field of UGC video coding and processing is still quickly evolving, we also envision the future perceptual UGC video compression scheme based on the proposed quality measure.

II. Related Works

A. Objective VQA Measures

1) Reference-Based VQA: FR and RR VQA algorithms deliver predictions based on the accessible reference information. In [2], hysteresis effect in the subjective testings is observed, and a hysteresis based temporal pooling strategy is applied to extend image quality assessment (IQA) metrics such as PSNR and SSIM [3] to VQA, which has been proved to be better than average pooling. In [4] and [5], video quality measures have been designed based on structural features. Lu *et al.* [6] describes the degradation of video quality via spatiotemporal 3D gradient differencing. A VQA model based on statistical characteristics of optical flows is proposed in [7]. In [8], a multi-scale framework for evaluating video fidelity by motion quality along computed motion trajectories is presented. In [9], ViS3 estimates video quality by combining perceived degradation due to spatial distortion and joint spatial-temporal distortion. ST-RRED [10] is a RR VQA model in which a Gaussian scale mixture model is used to measure the spatial and temporal information differences, these measured differences between the reference and distorted videos are further combined to predict video quality. SpEED-QA [11] is a natural scene statistics (NSS) based model that applies local entropic differencing between reference and distorted videos. Machine learning has also played a critical role in the development of modern VQA models. In [12], several perceptual-relevant features and methods have been combined by random forest regression algorithm to boost the performance. In [13], video multi-method assessment fusion (VMAF) produces remarkably improved quality prediction performance by mapping multiple features to human-quality opinions using support vector regression (SVR). Motivated by the great success of convolutional neural network (CNN) on numerous visual analysis tasks, a DNN based approach has been developed by joint learning of local quality and local weights in [14], and a pairwise-learning framework is proposed in [15] to train a perceptual image-error metric. Kim *et al.* [16] propose to quantify the video quality via a CNN and a convolutional neural aggregation network, which are used for spatio-temporal sensitivity learning and temporal pooling, respectively. Zhang *et al.* [17] propose a FR VQA metric for compressed videos by integrating transfer learning with a CNN, samples are enriched by transferring the distorted images as the related domain. C3DVQA [18] exploits 3D convolution in FR VQA task, specifically, 2D convolutional layers are used to extract spatial features and 3D convolutional layers are used for learning spatio-temporal features.

2) No-Reference VQA: NR VQA is a more natural and preferable way to assess the perceived video quality as the pristine videos are unavailable in many practical video applications. Many methods focus on estimating the perceived quality of videos with specific distortions, such as compression distortion [19], transmission error [20] and scaling artifacts [21]. For distortion-unaware NR VQA methods, NSS is usually used as it is sensitive to diverse distortions. Saad *et al.* [22] propose a NR VQA algorithm, known as VBLIINDS, which contains a NSS model and a motion model that quantifies motion coherency. Mittal *et al.* [23] propose a VQA model termed as the video intrinsic integrity and distortion evaluation oracle (VIIDEO), which quantifies disturbances introduced due to distortions according to the NSS model. In [24], the video content is disassembled into the predicted part and the uncertain part, their quality degradations are separately evaluated by NSS model and further yield the overall quality. Li *et al.* [25] propose a NR VQA metric based on NSS in the 3D discrete cosine transform (DCT) domain. Recently, CNN based NR-VQA methods have also been developed. Li *et al.* [26] propose a shearlet- and CNN-based NR VQA (SACONVA), where spatiotemporal features extracted by 3D shearlet transform are fed to a CNN to predict a perceptual quality score. Liu *et al.* [27] propose a 3D CNN model for codec classification and quality assessment of compressed videos. In [28], a NR VQA framework based on weakly supervised learning with a CNN and a resampling strategy is presented. Li *et al.* [29] propose a NR VQA framework for in-the-wild videos by incorporating content dependency and temporal-memory effects. Specifically, content-aware features are extracted using a pre-trained CNN, and temporal effects are modeled by a gated recurrent unit (GRU) network and a subjectively-inspired temporal pooling layer. TLVQM [30] is a learning-based VQA model which jointly employs low-complexity features from full video

sequences and high-complexity features from representative video frames, and it is further improved to CNN-TLVQM [31] by replacing spatial high-complexity features to deep features. Moreover, generative networks have also been used to predict reference information given the distorted images to help blind IQA task [32], [33].

B. VQA Databases

There are several publicly available video databases for VQA. LIVE [34] collects 10 uncompressed high-quality videos as reference videos, and correspondingly 150 distorted videos are created using four different distortion types and strengths. LIVE Mobile [35] consists of 200 distorted videos created from 10 RAW HD reference videos, and dynamically varying distortions are also considered. VMAF+ video quality database [36] contains 29 video contents from TV shows and movies. These clips are further downsampled to six different resolutions and encoded using three constant rate factors (CRFs) for the study of the combined effects of compression and scaling artifacts on the perceived subjective quality. BVI-HD [37] is a video quality database for HEVC compressed and texture synthesized content, in which 12 distorted versions are generated by HEVC [38] and its synthesis integrated version HEVC-SYNTH. BVI-SynTex [39] contains 196 videos clustered in three texture types and HEVC codec is adopted to study the impact of content to compression efficiency and perceptual quality. TJU-SVQA [40] focus on the VQA of symmetrically/asymmetrically distorted stereoscopic videos. In MCL-JCV [41], a VQA database for compressed videos is created based on the just noticeable difference (JND) model. CVD2014 [42] contains a total of 234 videos that are recorded using 78 different cameras, along with open-ended quality descriptions such as sharpness, graininess and color balance provided by the observers. LIVE-Qualcomm [43] consists of 208 videos captured using 8 different mobile devices which model six common in-capture distortion categories. LIVE-VQC [44] contains 585 videos captured using 101 different devices with a wide range of distortion levels. KoNViD-1k [45] contains 1200 public-domain videos that are fairly sampled from a large public video database YFCC100 M. Youtube UGC [46] collects around 1500 UGC video clips across 15 categories sampling from millions of videos according to spatial, color, temporal and chunk variation features. The issues regarding how to evaluate quality degradation caused by compression for these already distorted UGC videos are discussed. For LIVE Wild Compressed Video Quality Database [47], 55 videos are randomly selected from LIVE-VQC Database. These videos are further down-scaled to different resolutions, compressed with multiple compression levels and sampled. Finally 275 videos along with their subjective scores are finally collected.

Apparently, databases with high quality source videos such as LIVE and LIVE Mobile may not align with the UGC application scenarios, where databases with diverse authentic acquisition distortion such as CVD2014, LIVE-Qualcomm are more realistic. KoNViD-1 k is a UGC video quality assessment database. However, only the UGC videos uploaded by users are contained and the transcoded versions of these UGC videos are absent. Hence, database dedicated to UGC videos by considering the UGC video compression still remains absent and there is a strong desire for an adequate database sufficing to simulate the UGC production chain from acquisition to processing on the hosting platform.

III. UGC Video Database

A. Video Collection

To cover typical content and characteristics representing UGC videos, 400 videos are randomly selected from the videos uploaded to TikTok that meet the following criteria:

- With a resolution of 1280×720 (height × width);
- Belonging to the category of selfie, indoor, outdoor or screen content.
- Last longer than 10 seconds;
- Played at 30 frames per second (FPS);

Since 720p is one of the most widely adopted UGC video formats, we ensure that all selected videos share this resolution. Selfie, indoor, outdoor and screen content videos are common types of UGC videos. In particular, most areas of selfie videos are occupied by human face, and screen content videos are mainly game screen recording. Moreover, indoor videos are life scenes shot in close-up, and outdoor videos are outdoor scenery acquired with a distant view. A few videos with special content have been filtered out. Since we will crop all these videos to 10 seconds, videos shorter than 10 seconds are not considered.

B. Video Sampling

Subsequently, we sample the videos selected in the previous step according to the statistical characteristics of videos to obtain the final source videos. Specifically, three attributes including spatial perceptual information (SI), temporal perceptual information (TI) and blur index have been employed. Among these indicators, SI and TI are highly correlated with the levels of distortion when the video is lossy transmitted, as suggested in [48]. Since UGC videos uploaded by users are usually accompanied by varying degrees of blurry artifacts which significantly affect the perceptual quality, the blur metric is also included.

SI: SI quantifies spatial complexity and variety of a video, and it is defined as the maximum standard deviation over all Sobel-filtered frames,

$$SI = max_{time}\{std_{space}[Sobel(F_n)]\}, \tag{1}$$

where F_n represents frame n, $Sobel(\cdot)$ is Sobel filter and std_{space} represents the standard deviation over space.

TI: TI quantities the temporal changes of a video, and it is given by the maximum standard deviation of the frame difference derived from adjacent frames. As such, it can be formulated as follows,

$$TI = max_{time}\{std_{space}[M_n(i,j)]\}, \tag{2}$$

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j), \tag{3}$$

where $F_n(i,j)$ is pixel value at (i,j) of frame n.

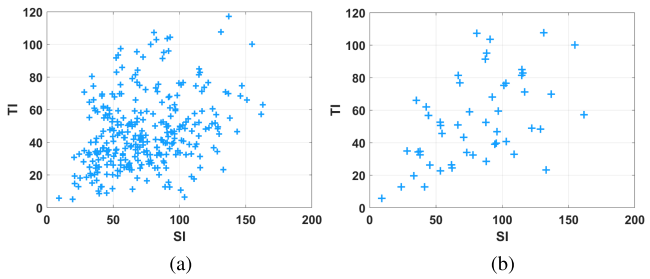


Fig. 1. Distribution of SI and TI indices for source UGC videos. (a) 400 videos before sampling, (b) 50 videos after sampling.

Blur: The cumulative probability of blur detection (CPBD) indicator [49] is adopted here to evaluate the levels of blur. The average CPBD value of the all frames is used to indicate the blurriness of the video.

Before sampling from these videos, we crop these videos to 10 seconds and remove the audio parts. Fair sampling mechanism should produce a broader diversity of video properties than a random sampling mechanism, hence we adopt the sampling strategy introduced in [50] to enable the characteristics of sampled videos uniformly distributed in terms of these attributes. In particular, the original videos are characterized with a set S,

$$S = \left\{ \boldsymbol{q}_i | \boldsymbol{q}_i \in \mathbb{R}^M, \boldsymbol{q}_i \sim D_S^M \right\}_{i=1}^K, \tag{4}$$

where M and K represent the number of features and videos, respectively (here $M = 3, K = 400$). The main objective is to select a subset of N videos,

$$s = \left\{ \hat{\boldsymbol{q}}_i | \hat{\boldsymbol{q}}_i \in S, \hat{\boldsymbol{q}}_i \sim D_s^M \right\}_{i=1}^N, \tag{5}$$

with the uniform distribution $\boldsymbol{D} \in \mathbb{R}^{H \times M}$ (each of its columns $\boldsymbol{D}_{*j}$ denoting the probability mass function (PMF) across the j^{th} dimension which is quantized into H bins). As such, we introduce a set of M binary matrices $B = \{\boldsymbol{B}^m\}_{m=1}^M$, in which b_{ij}^m denotes whether or not the j^{th} item of S belongs to i^{th} interval of the target PMF for the dimension m, and binary vector $\boldsymbol{x} \in \mathbb{Z}_2^K$, where x_i is decision variable determining whether i^{th} item of S belongs to subset s. As such, this problem can be formulated as follows,

$$\min_{\boldsymbol{x}} \sum_{m=1}^{M} \|\boldsymbol{B}^m \boldsymbol{x} - N \boldsymbol{D}_{*m}\|_1 \ \text{ s.t. } \|\boldsymbol{x}\|_1 = N. \tag{6}$$

By finding the best solution with the optimization objective, a subset that is closest to the uniform distribution on all features can be sampled from the original database. Finally, 12, 13, 13, 12 videos were chosen from selfie, indoor, outdoor and screen content videos, respectively, using this sampling strategy with $H = 5$. Fig. 1 shows the plots of SI against TI for 400 videos before sampling and 50 videos after sampling, which is apparent that sampled videos span a wide range of SI-TI spaces. Moreover, the snapshots of some example sampled videos from each content category are shown in Fig. 2.

C. Video Transcoding

Considering that our primary goal of building this database is to simulate the UGC production chain from acquisition to processing on the hosting platform, based on which quality assessment algorithm can be developed in an effort to further improve the transcoding performance, we further transcode these sampled source videos using different codecs to compression levels. H.264/AVC [51] and H.265/HEVC are commonly used video compression standards in practice and they are adopted to simulate the transcoding process in the hosting platforms. Specifically, FFmpeg software [52] and its internal libraries are used to perform the video transcoding. Constant quantization parameter (QP) mode is used for each codec and different QP values are set to control the quality levels. QP values of 22, 27, 32, 37 are commonly used for these two compression standards, besides, as some low quality UGC videos are less sensitive to compression, that is, there are no significant quality differences between source and transcoded versions of these QP values, so a larger QP value of 42 is also applied. As such, each source video can be transcoded to 5 quality levels by setting QPs to 22, 27, 32, 37 and 42 for each codec. The source videos, as well as the corresponding transcoded videos, are of YUV 4:2:0 format. Finally, there are 550 videos in our built UGC-VIDEO database including source videos.

D. Subjective Testing and Analyses

After collecting the videos, subjective testing is further conducted to obtain the subjective scores using absolute category rating with hidden reference (ACR-HR) [53] method. Specifically, the videos are displayed one by one in their native resolution without scaling, and the subjects are asked to provide an opinion score according to the five-grade rating scales. A liquid crystal display (LCD) monitor with a display resolution of 1920×1440 is used for display, and all subjects conduct experiments using the same monitor under stable office environment with moderate ambient brightness to ensure the fairness and effectiveness of our subjective experiment. The full database is divided into three sessions, each containing 16 or 17 source videos along with their respective transcoded versions. Hence, each session lasts about half an hour to minimize viewer fatigue. In particular, at the beginning of each session, "dummy presentations" with various levels of perceptual quality have been introduced to stabilize the opinion of subjects and the opinion data of these presentations are not taken into account in the final result of the experiment. The subjects are required to click the corresponding button within a few seconds to choose from "Excellent,""Good," "Fair," "Poor" and "Bad," corresponding to 5$\sim$1 points. A total of 28 subjects participated in this test. Since this is a hidden-reference study, the source videos have also been included in the subjective testings. As such, the mean opinion scores (MOSs) of all 550 videos in our database can be obtained.

The screening of subjects is further conducted as specified in ITU-R BT 500.13 [54]. The kurtosis of scores are calculated to determine if the scores for each test presentation are normally distributed. Score range of each video is then computed as 2

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2. Examples of source videos in our database. (a)–(b): selfie videos. (c)–(d): indoor videos; and (e)–(f): outdoor videos. (g)–(h): screen content videos.

TABLE I
PERFORMANCE COMPARISONS OF QUALITY ASSESSMENT ALGORITHMS IN TERMS OF SROCC

Methods	Selfie	Indoor	Ourdoor	Screen	Full set
BRISQUE	0.436	0.327	0.580	0.346	0.354
NIQE	0.511	0.480	0.453	0.128	0.314
VIIDEO	0.113	0.348	0.218	0.026	0.085
VBLIINDS	0.382	0.386	0.051	0.462	0.175
PSNR	0.715	0.700	0.664	0.489	0.612
VIF	0.837	0.803	0.807	0.629	0.736
SSIM	0.842	0.798	0.857	0.464	0.714
MS-SSIM	0.821	0.783	0.842	0.507	0.722
SpEED-QA	0.839	0.747	0.838	0.746	0.786
ViS3	0.762	0.706	0.823	0.699	0.746
VMAF	0.823	0.821	0.856	0.825	0.814

TABLE II
PERFORMANCE COMPARISONS OF QUALITY ASSESSMENT ALGORITHMS IN TERMS OF PLCC

Methods	Selfie	Indoor	Ourdoor	Screen	Full set
BRISQUE	0.416	0.346	0.611	0.328	0.315
NIQE	0.509	0.511	0.520	0.056	0.176
VIIDEO	0.251	0.178	0.326	0.032	0.157
VBLIINDS	0.415	0.421	0.001	0.464	0.216
PSNR	0.717	0.733	0.639	0.452	0.579
VIF	0.862	0.820	0.850	0.633	0.626
SSIM	0.866	0.847	0.857	0.590	0.769
MS-SSIM	0.845	0.841	0.865	0.626	0.773
SpEED-QA	0.748	0.671	0.730	0.724	0.673
ViS3	0.787	0.744	0.872	0.754	0.783
VMAF	0.884	0.886	0.907	0.830	0.863

or $\sqrt{20}$ standard deviations from the mean scores according to whether the scores are normally distributed. For each subject i, we count the number of scores above and blow this range, denoted as P_i and Q_i. As such, the subject i will be rejected when

$$\frac{P_i + Q_i}{T} > 0.05 \text{ and } |\frac{P_i - Q_i}{P_i + Q_i}| < 0.3, \tag{7}$$

where T is the number of test videos. Based on the our analysis, no subject has been rejected at this stage.

E. Performance of Existing Models

We evaluate the performance of several objective quality assessment algorithms on the established database using Spearman's rank ordered correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC). In particular, the larger the values of SROCC and PLCC, the better the performance. Besides, before computing PLCC, the predicted scores are passed through a logistic non-linearity regression as suggested in [55]:

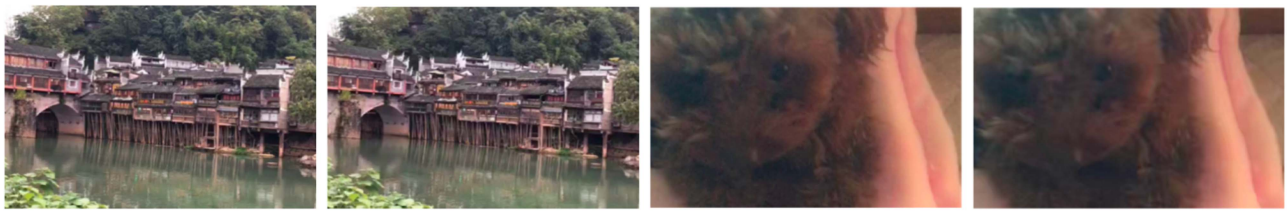
$$f(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (x - \beta_3)}} \right) + \beta_4 x + \beta_5. \tag{8}$$

IQA algorithms are extended to VQA methods by averaging frame-level quality scores. The tested quality measures include reference-based models PSNR, SSIM, MS-SSIM [56], VIF [57], SpEED-QA, ViS3 and VMAF, as well as NR models BRISQUE [58], NIQE [59], VBLIINDS and VIIDEO. Table I and Table II tabulate the SROCC and PLCC between the algorithm scores and MOS for each content category, as well as across the full database. It is disappointing to find that the existing algorithms may not be able to provide adequate predictions on the UGC videos. However, these results still provide some useful insights that could benefit the design of the UGC VQA models. First, as illustrated in Fig. 3, the quality based on comparisons against the reference could be problematic due to the corrupted reference. This suggests the importance of including the intrinsic quality of the reference. Second, most of the tested algorithms perform the worst on screen content videos, and this may be attributed to the particularity of this type of videos compared to natural videos. These observations motivate a specifically designed VQA framework that equips the intrinsic quality of the corrupted reference as well as the data-driven model for learning the statistics of the video content.

IV. OBJECTIVE QUALITY ASSESSMENT

A. Framework

In the UGC video product chain, users capture videos and upload the videos to the host platform. Before distributing to end users, transcoding is conducted, yielding transcoded videos with different levels of compression distortions. We target at evaluating the quality of transcoded videos through the proposed UGC VQA framework. As illustrated in Fig. 4, the proposed framework leverages the intrinsic quality of the source videos as well as the comparisons between the source and transcoded videos. We propose to learn and fuse the intermediate quality maps, which meaningfully indicate the spatially variant quality of different regions. The inferred quality maps are fed to a pooling network, such that the local distortions are aggregated in a data-driven manner for final quality prediction. As such,



(a) MOS: 4.43 (b) MOS: 4.11 (c) MOS: 3.21 (d) MOS: 2.96

Fig. 3. Performance of FR metrics on the proposed database. (a) One sample frame of high quality reference video, (b) Corresponding frame from HEVC transcoded video of (a), PSNR: 41.64 dB, SSIM: 0.996, (c) One sample frame of low quality reference video, (d) Corresponding frame from HEVC transcoded video of, and (c) PSNR: 41.73 dB, SSIM: 0.977.

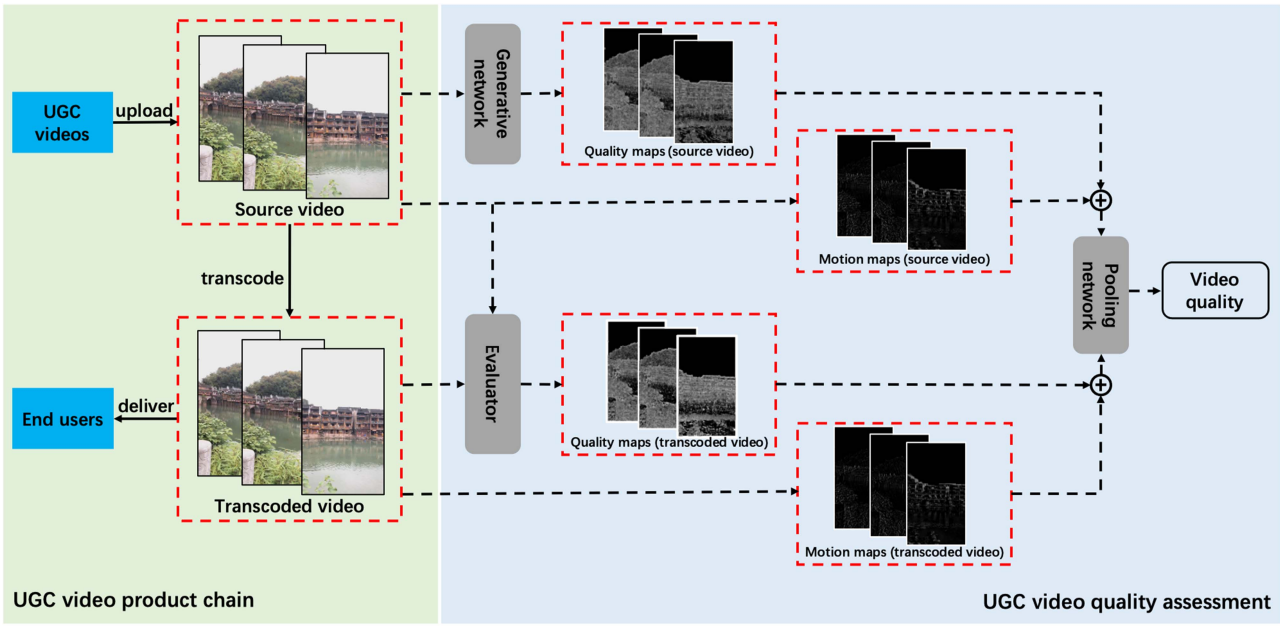


Fig. 4. Framework of the proposed objective quality assessment method. The quality maps from the source videos, the comparisons between source and transcoded videos, as well as their motion maps are fused with a pooling network to obtain the final predicted quality.

the proposed quality evaluation framework mainly consists of three modules, including a generative network G_ϕ that generates the quality maps of the source videos, an evaluator E_ω which produces the relative quality maps between the source and transcoded videos, and a pooling network f_θ that fuses the quality maps to obtain the final quality score. These modules are parameterized by ϕ, ω and θ, respectively.

Given a source video $\boldsymbol{V}_s$ and its transcoded version $\boldsymbol{V}_t$, we first predict the quality maps $\boldsymbol{M}_s$ of $\boldsymbol{V}_s$ using G_ϕ:

$$M_s^i = G_\phi(I_s^i), \tag{9}$$

where I_s^i is the i-th frame of $\boldsymbol{V}_s$ and M_s^i represents its corresponding quality map. Meanwhile the relative perceptual degradation between $\boldsymbol{V}_s$ and $\boldsymbol{V}_t$ can also be measured,

$$M_t^i = E_\omega(I_s^i, I_t^i), \tag{10}$$

where I_t^i and M_t^i are the i-th frame of $\boldsymbol{V}_t$ and its corresponding quality map, respectively. The motion information of the source video and transcoded video is also employed, where the frame differences are calculated. The motion information of the i-th frames in $\boldsymbol{V}_s$ and $\boldsymbol{V}_t$ are given by

$$d_s^i = |I_s^{i+1} - I_s^i|, \tag{11}$$

$$d_t^i = |I_t^{i+1} - I_t^i|. \tag{12}$$

Finally, a quality pooling network f_θ combines the maps from source video and transcoded video, which delivers the intrinsic quality of source video as well as the relative quality between source and transcoded video,

$$\hat{S} = f_\theta(G_\phi(\boldsymbol{V}_s) \oplus \boldsymbol{D}_s, E_\omega(\boldsymbol{V}_s, \boldsymbol{V}_t) \oplus \boldsymbol{D}_t) \tag{13}$$

where $\boldsymbol{D}_s = \{d_s^1, d_s^2, \ldots, d_s^n\}$, $\boldsymbol{D}_t = \{d_t^1, d_t^2, \ldots, d_t^n\}$, $\oplus$ is the concatenation operator and $\hat{S}$ is the predicted score of the transcoded video.

B. Quality Maps Generation Based on $\boldsymbol{V}_s$ and $\boldsymbol{V}_t$

In the hosting platform, $\boldsymbol{V}_s$ is further transcoded into $\boldsymbol{V}_t$, such that the difference lying between them originates from compression artifacts. As such, given $\boldsymbol{V}_s$ and $\boldsymbol{V}_t$, to evaluate the relative distortion between them, we can leverage existing quality metrics such as SSIM, MDSI [60] and VIF to reflect the local

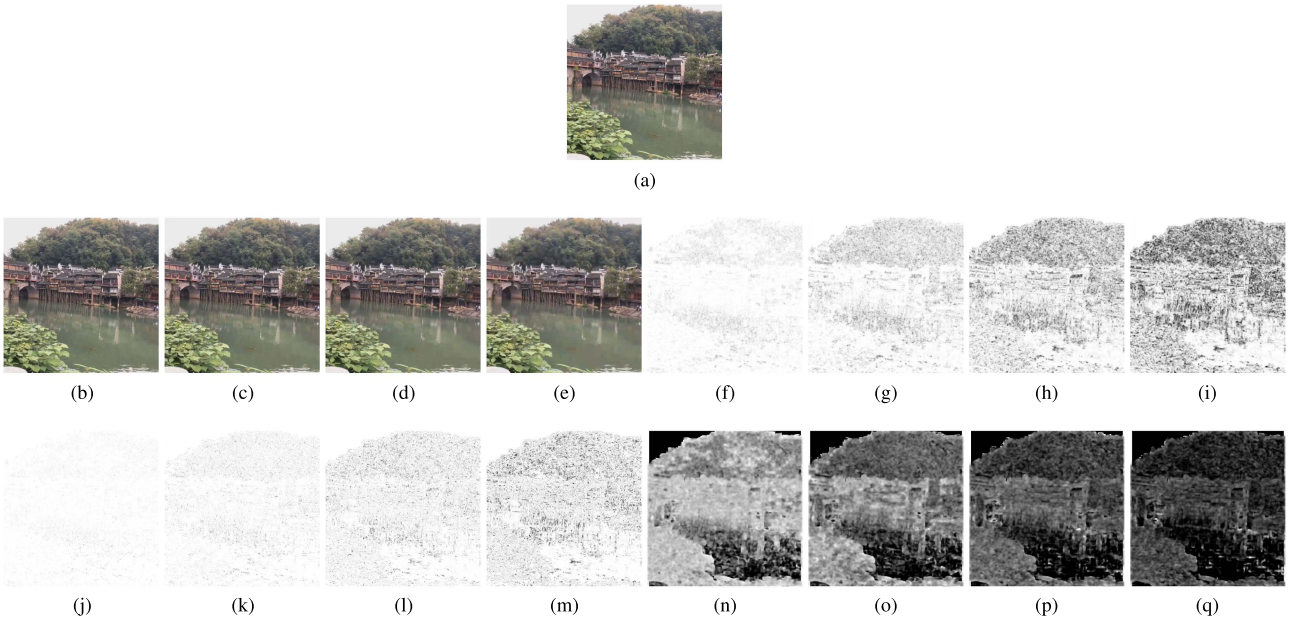


Fig. 5. Quality maps generated based on $\boldsymbol{V}_s$ and $\boldsymbol{V}_t$. (a) One frame from $\boldsymbol{V}_s$, (b)-(e) Frames from HEVC transcoded versions of $\boldsymbol{V}_s$ with QP 27, 32, 37 and 42, (f)–(i) corresponding SSIM maps, (j)–(m) corresponding MSDI maps, and (n)–(q) corresponding VIF maps.

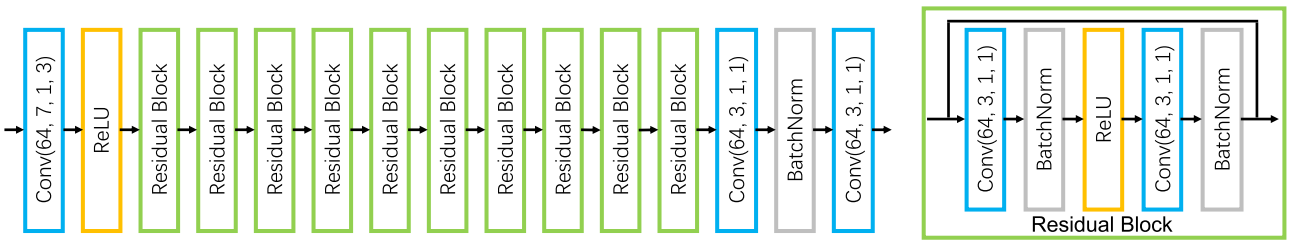


Fig. 6. The architecture of the generative network that produces the quality maps from $\boldsymbol{V}_s$. Blue box: a convolutional layer $Conv(d, f, s, p)$ with d filters of size $f \times f$, a stride of s and a padding of p; yellow box: ReLU layer; gray box: batch normalization layer.

distortion from I_s^i to I_t^i from the perspectives of structure, gradient and visual information, respectively. Regarding SSIM, only luminance component is considered and the derived single channel luminance similarity map of each frame pair is used as SSIM quality map. With respect to MDSI, the combination of gradient similarity map and chromaticity similarity map is used as MDSI map. Since VIF is one number that quantifies the information fidelity for the entire image, we employ a sliding-window to generate the VIF quality map, visually illustrating the perceptual quality of the test frame varies over space. These quality maps are shown in Fig. 5, which imply that the adopted quality maps well predict the visual quality. The values in the quality maps are normalized to the range of $[0, 1]$ to facilitate subsequent training in DNN. Y channel of source video frames and transcoded video frames are utilized to estimate relative quality maps of transcoded video in this step.

C. Quality Maps Generation From Source Video $\boldsymbol{V}_s$

Given the source video $\boldsymbol{V}_s$, we aim at blindly estimating the quality map of each frame since the pristine reference is not available. We adopt the deep neural networks ensuring the robust and accurate quality map prediction. In particular, ResNet [61] is employed with the consideration that residual connections make the training of identical function easier, which gradually facilitate the adding of distortions from low level to high level. The detailed architecture of the generative network is shown in Fig. 6. More specifically, quality maps of the input frame are predicted after 10 identical residual blocks, each of which contains two 3×3 conventional layers with 64 feature maps, and all convolution layers are with stride 1×1 and zero-padding. As such, the size of the final output feature map is consistent with the original input frame. Besides, batch normalization [62] and rectified linear unit (ReLU) are used after convolution. RGB channels of source video frames are utilized to estimate quality maps of source video in this step.

Regarding the training of the quality map generative network, Waterloo database [63] is adopted. In particular, it includes 4744 pristine images, as well as distorted versions with Gaussian noise, Gaussian blur, JPEG compression and JPEG2000 compression. It is worth mentioning that we use the pristine images in the database to regenerate the distorted images. To model the

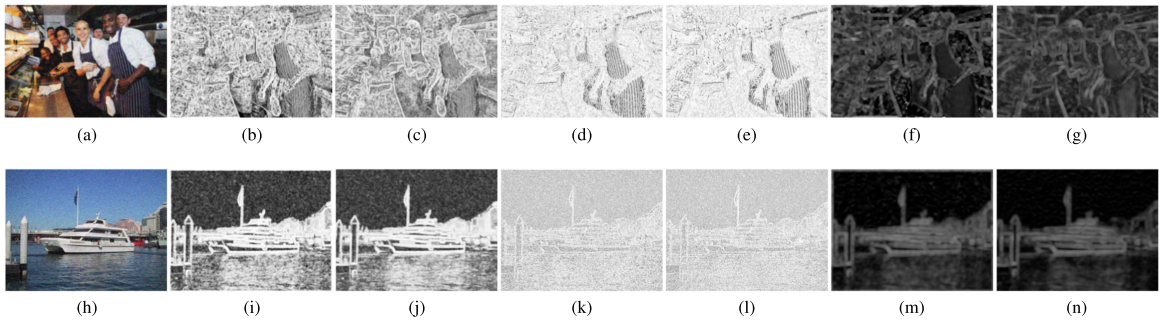


Fig. 7. Illustration of the predicted quality maps and the corresponding ground-truth maps. (a)(h) distorted images with multiple distortions, (b)(i) ground-truth SSIM maps, (c)(j) predicted SSIM maps, (d)(k) ground-truth MDSI maps, (e)(l) predicted MDSI maps, (f)(m) ground-truth VIF maps, and (g)(n) predicted VIF maps.

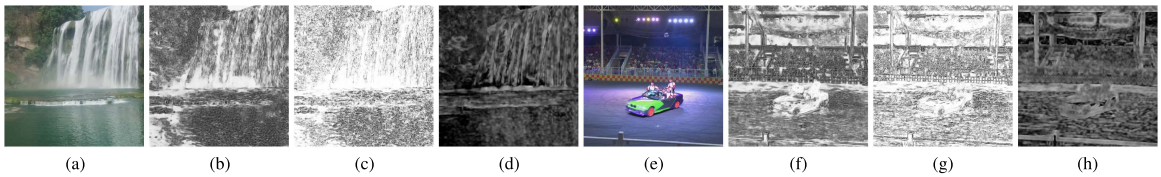


Fig. 8. Illustration of the predicted quality maps in our database. (a)(e) frames of $\boldsymbol{V}_s$, (b)(f) predicted SSIM maps of $\boldsymbol{V}_s$, (c)(g) predicted MDSI maps of $\boldsymbol{V}_s$, and (d)(h) predicted VIF maps of $\boldsymbol{V}_s$.

distortions contained in $\boldsymbol{V}_s$, multiple distortion stages are applied on these pristine images. Gaussian blur or Gaussian noise of different levels is injected to the pristine image, and subsequently these distorted images are compressed with different compression levels by JPEG or JPEG2000 compression. The images after compression are used as training inputs and their quality maps are used as groundtruth labels. As described in Section IV-B, different quality maps derived from existing FR models can be adopted as training labels. Different quality maps predicted by the generative network and their corresponding ground-truth labels are shown in Fig. 7. The generative networks trained on Waterloo database are then applied on our database to generate quality maps of $\boldsymbol{V}_s$, as shown in Fig. 8.

The loss function is designed with full consideration of the quality maps' characteristics. In particular, the pixel values of the quality map reflect the quality degradation of specific location, and the texture of quality map reveals the quality variation over space. The designed loss function for the generative network consists of a structural loss characterized by SSIM and pixel-wise loss. SSIM-based loss well preserves the contrast for high-frequency regions but reveals less sensitivity to uniform biases. This may cause deviations of pixel intensity. L1 loss is expert in preserving colors and luminance but shows less efficiency in maintaining the contrast level. We try to capture the best characteristics of both error functions by combining them. The effectiveness of this combination of loss function has been corroborated in [64]. The loss function is given by,

$$L_G(P_k^0, P_k) = \alpha \cdot L^{SSIM}(P_k^0, P_k) + (1-\alpha) \cdot L^{L1}(P_k^0, P_k), \tag{14}$$

where P_k^0 is the ground-truth quality map patch, P_k is the corresponding generated patch, and α here is an empirically set weighting factor. In this work we set $\alpha = 0.75$ so that the contribution of the two losses would be roughly balanced The structural loss based on SSIM is formulated as,

$$L^{SSIM}(P_k^0, P_k) = 1 - SSIM(P_k^0, P_k). \tag{15}$$

D. Quality Map Pooling

After the generation of quality maps from the source video and transcoded video, a pooling network is trained to fuse these quality maps and generate a final quality score. Moreover, as temporal masking effects of the video may influence the perceptual quality, we utilize the motion information as temporal cues for quality assessment. Specifically, the frame difference is extracted from consecutive neighboring frames and scaled as motion maps. The motion maps of source videos and transcoded videos are concatenated with quality maps, serving as the input of the pooling network. In general, convolutional networks have been widely used to progressively reduce the resolution of feature maps, while such loss of spatial acuity may limit the performance. In our framework, a dilated residual network (DRN) [65] is employed, in which dilated convolutions are used to increase the resolution of output feature maps without reducing the receptive field of individual neurons. As shown in Fig. 9, source video maps and transcoded video maps flow through independent convolutional layers, and feature maps are concatenated after the first convolutional layer. The four dilated residual structures with 3×3 convolution kernel are used to extract feature representations, and global average pooling as well as fully connected layers are used to regress to the final score. The loss function of pooling network is mean square error (MSE) loss,

$$L_{REG} = \left\| f_\theta(G_\phi(\boldsymbol{V}_s) \oplus \boldsymbol{D}_s, E_\omega(\boldsymbol{V}_s, \boldsymbol{V}_t) \oplus \boldsymbol{D}_t) - S \right\|^2, \tag{16}$$

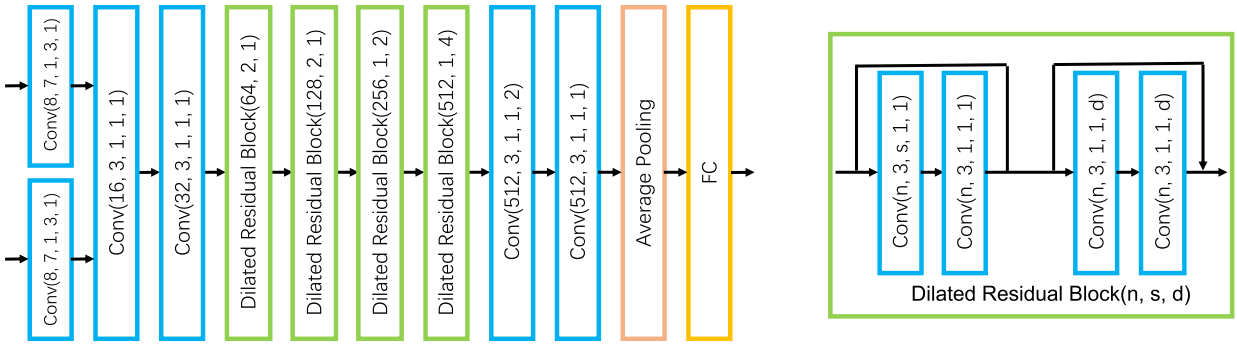


Fig. 9. Detailed architecture of the pooling network. Blue box: a convolutional layer $Conv(n, f, s, p, d)$ with n filters of size $f \times f$, a stride of s, a padding of p and a dilation of d; red box: average pooling layer; yellow box: full connection layer. It is worth mentioning that batch normalization and ReLU layers after convolutional layers are omitted here for simplification.

where S denotes video score of human evaluation for $\boldsymbol{V}_t$, serving as training label for each input pair.

V. Experimental Results

A. Experimental Settings

1) Database: Due to the lack of databases that align with the UGC application scenario, in particular from acquisition to processing on the hosting platform, the UGC-VIDEO database newly introduced in Section III is used for evaluating our proposed method.

2) Compared Methods: Both reference-based and NR quality assessment algorithms are applicable for quality assessment of UGC videos. In particular, for reference-based methods, corrupted source videos with various quality levels are used as references to help evaluate the quality of transcoded videos. NR methods are directly applied on the transcoded videos for evaluating the video quality. Various traditional reference-based and NR methods are used for comparison, including PSNR, SSIM, MS-SSIM, VIF, SpEED-QA, ViS3, VMAF, NIQE, BRISQUE, VBLIINDS, VMAF. In addition, 2stepQA [66] method, which serves as a flexible framework based on different combinations of FR and NR methods, is also considered. Furthermore, DNN-based NR VQA mothods VSFA, CNN-TLVQM and FR method C3DVQA are retrained and compared.

B. Training Details

The training process consists of two steps: (1) training generative network on the modified Waterloo Exploration Database; (2) training the pooling network on the UGC-VIDEO database.

In the original Waterloo Exploration Database, 94880 distorted images are created from 4744 pristine natural images by introducing four types of distortion (blur, noise, JPEG and JPEG2K), each with five levels. To enable the generation network to capture mixture distortions similar to that in the source UGC videos, we have designed a new way to generate the distorted images. More specifically, noise or blur distortions of random levels are first induced to these pristine images, and subsequently compression distortion is injected by JPEG or JPEG2000 with random compression levels. As such, 4744 distorted images with multiple distortions are created. SSIM and VIF quality maps are calculated according the distorted images and the corresponding pristine images. Both distorted images and their quality maps are cropped into 64×64 non-overlapping patches. Generative network is trained based on the inputs (patches from distorted image) and labels (corresponding patches from quality map) using Adam optimizer [67] at the learning rate of 10^{-3} for 50 epochs. The loss curves of the generative network during training are shown in Fig. 10, as we have seen, there appears the network convergence after about 40 epochs.

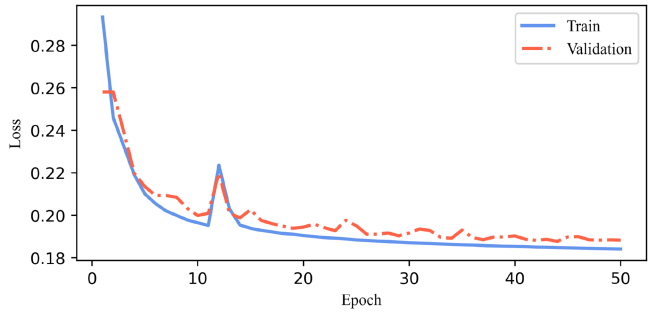


Fig. 10. Loss curves of the generative network during training.

Subsequently, the pooling network is trained using the pre-trained generative network model and the score is regressed using quality maps. 20 frames are uniformly sampled from videos in this step. Once each quality map is derived from the previous generative network, we freeze the weights of generative network, and train the pooling network using MSE loss and Adam optimizer for 100 epochs, cooperated with L2 regularization factor 5×10^{-4}. The quality maps and motion maps are downsampled by a factor of 2 before being fed to the pooling network. We set the initial learning rate to 10^{-3} and drop it every 30 epochs using a multiplicative factor of 0.1, and batch size is set to 16. The model with the highest SROCC performance on the validation set is selected for test.

C. Performance Comparisons

In this work, we focus on the VQA of transcoded UGC videos, to ensure fair comparisons with existing conventional

TABLE III
MEAN AND STANDARD DEVIATION OF PERFORMANCE VALUES OF CONVENTIONAL FR AND NR METHODS IN 10 RUNS ON UGC-VIDEO DATABASE, I.E., MEAN (±STD). THE BOLDFACED ENTRIES INDICATE THE BEST PERFORMANCE

Method	SROCC	PLCC	RMSE
PSNR	0.655 (±0.113)	0.667 (±0.092)	0.651 (±0.040)
SSIM	0.714 (±0.118)	0.766 (±0.090)	0.557 (±0.089)
MS-SSIM	0.723 (±0.107)	0.769 (±0.089)	0.556 (±0.093)
VIF	0.780 (±0.070)	0.752 (±0.117)	0.592 (±0.141)
SpEED-QA	0.794 (±0.060)	0.802 (±0.070)	0.522 (±0.089)
ViS3	0.759 (±0.077)	0.779 (±0.079)	0.549 (±0.070)
VMAF	**0.819 (±0.044)**	**0.860 (±0.044)**	**0.444 (±0.072)**
NIQE	0.346 (±0.107)	0.314 (±0.092)	0.810 (±0.042)
BRISQUE	0.364 (±0.110)	0.314 (±0.094)	0.810 (±0.049)
VIIDEO	0.086 (±0.070)	0.126 (±0.065)	0.845 (±0.041)
VBLINDS	0.274 (±0.092)	0.285 (±0.083)	1.384 (±1.534)

TABLE IV
MEAN AND STANDARD DEVIATION OF PERFORMANCE VALUES OF 2STEPQA MODEL USING DIFFERENT COMBINATIONS OF FR AND NR METHODS IN 10 RUNS ON UGC-VIDEO DATABASE, I.E., MEAN (±STD). THE BOLDFACED ENTRIES INDICATE THE BEST PERFORMANCE

Method	SROCC	PLCC	RMSE
PSNR+NIQE	0.693 (±0.117)	0.714 (±0.115)	0.602 (±0.082)
PSNR+BRISQUE	0.768 (±0.037)	0.764 (±0.040)	0.576 (±0.043)
PSNR+VIIDEO	0.641 (±0.114)	0.661 (±0.099)	0.679 (±0.105)
PSNR+VBLIINDS	0.679 (±0.105)	0.672 (±0.092)	0.648 (±0.042)
SSIM+NIQE	0.727 (±0.100)	0.775 (±0.092)	0.542 (±0.097)
SSIM+BRISQUE	0.795 (±0.041)	0.812 (±0.048)	0.516 (±0.057)
SSIM+VIIDEO	0.712 (±0.120)	0.764 (±0.090)	0.559 (±0.090)
SSIM+VBLIINDS	0.760 (±0.113)	0.783 (±0.088)	0.536 (±0.093)
VIF+NIQE	0.780 (±0.070)	0.750 (±0.121)	0.596 (±0.148)
VIF+BRISQUE	0.778 (±0.070)	0.749 (±0.121)	0.597 (±0.149)
VIF+VIIDEO	0.780 (±0.070)	0.750 (±0.121)	0.595 (±0.147)
VIF+VBLIINDS	0.775 (±0.067)	0.743 (±0.124)	0.612 (±0.182)
VMAF+NIQE	0.812 (±0.044)	0.861 (±0.045)	0.442 (±0.074)
VMAF+BRISQUE	**0.831 (±0.027)**	**0.871 (±0.031)**	**0.429 (±0.063)**
VMAF+VIIDEO	0.807 (±0.049)	0.857 (±0.047)	0.451 (±0.077)
VMAF+VBLIINDS	0.820 (±0.038)	0.863 (±0.042)	0.439 (±0.071)

and learning-based methods, the 500 transcoded videos in our database are randomly divided into non-overlapping 60% training set, 20% validation set and 20% test set, according to the content of source videos. Conventional quality measures which are not learning-based are directly evaluated on the 20% testing data after the parameters in (8) are optimized with the training and validation data. For the 2stepQA method, the training and validation sets are merged together to fit the relevant parameters before evaluating on the test set. For our method and other deep learning-base methods, the models with the highest SROCC value on the validation set during the training are chosen for testing. This procedure has been repeated for 10 times and all above methods are tested on the same 20% test set for each repetition. The SROCC, PLCC and RMSE are calculated according to the predicted quality scores and MOS of transcoded videos, in particular, the mean and standard deviation of performance values in 10 runs are reported.

Table III shows the performance of conventional methods. reference-based algorithms tend to perform better than NR algorithms, and VMAF performs the best by fusing different metrics. This may due to the fact that the NSS-based NR metrics developed on synthetically distortions fail to handle such complicated authentic distortions. Moreover, the 2stepQA performances with different combinations of FR and NR models are shown in Table IV, we can see that the performances of reference algorithms have been improved in most cases. For example, the SROCC performance of SSIM method is greatly improved from 0.714 to 0.795 by introducing the BRISQUE score of source video using 2stepQA model. However, due to the simplicity of the 2stepQA model and the lack of efficient NR models with high generalization capability, 2stepQA method may degrade the performance of FR algorithms. For example, the performances of VIF combined VBLIINDS, and VMAF combined VIIDEO are inferior to VIF and VMAF, respectively.

The performance comparisons of deep learning-based models including VSFA, CNN-TLVQM, C3DVQA and the proposed method are shown in Table V. For our models, SSIM and VIF quality maps are employed for source videos and transcoded videos Besides, to demonstrate the effectiveness of motion maps, we show the performances with and without the introduction of motion map, where motion maps are concatenated with spatial quality maps before feeding into the pooling network

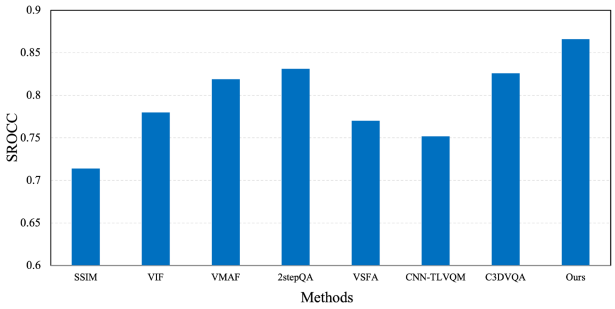


Fig. 11. Comparison of SROCC performances of different algorithms over 10 trials on the UGC-VIDEO database.

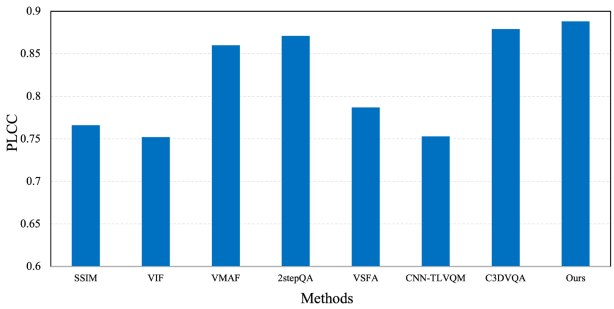


Fig. 12. Comparison of PLCC performances of different algorithms over 10 trials on the UGC-VIDEO database.

and the number of input channels of pooling network should be adjusted accordingly. We can find that, for our method, performance varies according to the selected quality map combination. All combinations show high correlation with the ground truth and it outperforms the other deep learning-based methods. Besides, motion maps are helpful for improving the performance.

To demonstrate the effectiveness of our framework, the average SROCC, PLCC and RMSE performances of conventional FR methods, 2stepQA method and deep learning-based methods are compared in Fig. 11, Fig. 12 and Fig. 13, respectively. It can be seen from the results that our method achieves the best overall performance in terms of predicting correlation (i.e.,SROCC) and accuracy (i.e.,PLCC and RMSE), and the proposed method significantly surpasses the second-best method in terms of SROCC.

TABLE V
MEAN AND STANDARD DEVIATION OF PERFORMANCE VALUES OF DEEP LEARNING-BASED MODELS IN 10 RUNS ON UGC-VIDEO DATABASE, I.E., MEAN (±STD). FOR OUR PROPOSED MODEL, DIFFERENT COMBINATIONS OF QUALITY MAPS ARE TESTED, I.E., TYPE OF QUALITY MAPS FOR SOURCE VIDEO + TYPE OF QUALITY MAPS FOR TRANSCODED VIDEO. THE BOLDFACED ENTRIES INDICATE THE BEST PERFORMANCE

Method	SROCC	PLCC	RMSE
VSFA	0.770 (±0.072)	0.787 (±0.073)	0.612 (±0.095)
CNN-TLVQM	0.752 (±0.048)	0.753 (±0.054)	0.656 (±0.084)
C3DVQA	0.826 (±0.039)	0.879 (±0.024)	0.407 (±0.046)
Ours:VIF+SSIM	0.825 (±0.056)	0.853 (±0.043)	0.461 (±0.061)
Ours:VIF+VIF	0.864 (±0.043)	0.880 (±0.046)	0.413 (±0.083)
Ours:(VIF, Motion)+(VIF, Motion)	**0.875 (±0.050)**	**0.881 (±0.050)**	**0.406 (±0.081)**
Ours:SSIM+SSIM	0.820 (±0.075)	0.857 (±0.057)	0.450 (±0.071)
Ours:SSIM+VIF	0.860 (±0.039)	0.892 (±0.039)	0.390 (±0.066)
Ours:(SSIM, Motion)+(VIF, Motion)	**0.866 (±0.035)**	**0.888 (±0.038)**	**0.399 (±0.052)**

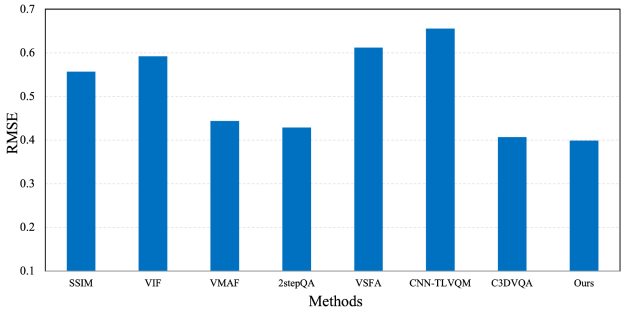


Fig. 13. Comparison of RMSE performances of different algorithms over 10 trials on the UGC-VIDEO database.

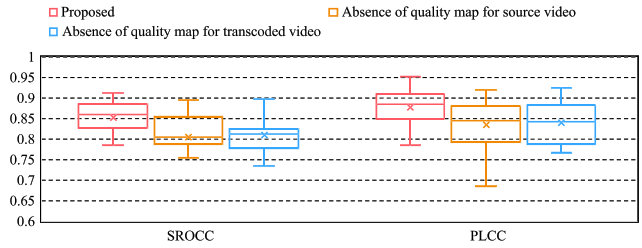


Fig. 14. Box plot in the ablation studies. The mark × in the middle represents the average. The bottom, middle and top bounds of the box represent the 25%, 50% and 75% percentage points, respectively.

D. Ablation Studies

To further provide evidence regarding the effectiveness of the proposed framework, we have conducted ablation studies by removing the quality maps of source and transcoded videos.

1) Absence of Quality Map for Source Video: We first show the performance variation due to the removal of the quality maps of source videos. In particular, these quality maps are replaced by source video frames, such that frames of source videos and quality maps of transcoded videos are fed to the pooling network.

2) Absence of Quality Map for Transcoded Video: The performance variations due to the removal of the quality maps of transcoded videos are further studied. In this manner, the quality maps of source videos and frames of transcoded video are fed to the pooling network.

We compare the full version of our proposed method (red) with source video quality map removed configuration (yellow) and transcoded video quality map removed configuration (blue), as shown in Fig. 14. Replacing quality maps with frames causes significant performance drop. Specifically, the absence of source video quality maps causes 5.51% and 4.90% decrease in terms of SROCC and PLCC, respectively. The absence of transcoded video quality maps leads to 5.04% and 4.44% performance drop in SROCC and PLCC, respectively. These results further verify the effectiveness of quality maps of source videos or transcoded videos in our framework.

VI. CONCLUSIONS

In this paper, we have systematically studied the video quality of UGC content. To facilitate the development of VQA for transcoded UGC videos, we have constructed a new subjective quality database. This database contains diverse UGC video sources along with their transcoded versions under different compression standards and levels. The subjective ratings of these videos are also provided as the ground truth. Based on the interesting observations from the developed database, we propose a new objective video quality model with the design philosophy that the quality prediction does not only rely on the divergence of source video and transcoded video, but also the intrinsic quality of the source videos. The experimental results show that our method outperforms the state-of-the-art quality assessment methods. The proposed VQA method is also envisioned to be further adopted to regularize the quality of the output UGC videos of sharing platforms, in an effort to provide a new paradigm of quality driven UGC video coding.

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and anonymous reviewers for their valuable comments that significantly helped them in improving the quality of the article.

REFERENCES

[1] Y. Li *et al.*, "UGC-VIDEO: Perceptual quality assessment of user-generated videos," in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2020, pp. 35–38.
[2] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 1153–1156.
[3] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[4] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process.: Image Commun.*, vol. 19, no. 2, pp. 121–132, 2004.

[5] Y. Wang, T. Jiang, S. Ma, and W. Gao, "Novel spatio-temporal structural information based video quality metric," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 7, pp. 989–998, Jul. 2012.

[6] W. Lu, R. He, J. Yang, C. Jia, and X. Gao, "A spatiotemporal model of video quality assessment via 3D gradient differencing," *Inf. Sci.*, vol. 478, pp. 141–151, 2019.

[7] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2480–2492, Jun. 2016.

[8] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[9] P. V. Vu and D. M. Chandler, "ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imag.*, vol. 23, no. 1, 2014, Art. no. 013016.

[10] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2013.

[11] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "SpEED-QA: Spatial efficient entropic differencing for image and video quality," *IEEE signal Process. Lett.*, vol. 24, no. 9, pp. 1333–1337, Sep. 2017.

[12] P. G. Freitas, W. Y. Akamine, and M. C. Farias, "Using multiple spatio-temporal features to estimate video quality," *Signal Process.: Image Commun.*, vol. 64, pp. 1–10, 2018.

[13] J. Y. Lin, T.-J. Liu, E. C.-H. Wu, and C.-C. J. Kuo, "A fusion-based video quality assessment (FVQA) index," in *Proc. Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2014, pp. 1–5.

[14] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[15] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual image-error assessment through pairwise preference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1808–1817.

[16] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 219–234.

[17] Y. Zhang, X. Gao, L. He, W. Lu, and R. He, "Objective video quality assessment combining transfer learning with CNN," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 8, pp. 2716–2730, Aug. 2020.

[18] M. Xu *et al.*, "C3DVQA: Full-reference video quality assessment with 3D convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 4447–4451.

[19] K. Zhu, C. Li, V. Asari, and D. Saupe, "No-reference video quality assessment based on artifact measurement and statistical analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 4, pp. 533–546, Apr. 2014.

[20] F. Zhang, W. Lin, Z. Chen, and K. N. Ngan, "Additive log-logistic model for networked video quality assessment," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1536–1547, Apr. 2013.

[21] 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.

[22] 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.

[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] Y. Zhu, Y. Wang, and Y. Shuai, "Blind video quality assessment based on spatio-temporal internal generative mechanism," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 305–309.

[25] 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.

[26] Y. Li *et al.*, "No-reference video quality assessment with 3D shearlet transform and convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 6, pp. 1044–1057, Jun. 2016.

[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*, 2018, pp. 546–554.

[28] Y. Zhang, X. Gao, L. He, W. Lu, and R. He, "Blind video quality assessment with weakly supervised learning and resampling strategy," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2244–2255, Aug. 2019.

[29] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 2351–2359.

[30] 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.

[31] J. Korhonen, Y. Su, and J. You, "Blind natural video quality prediction via statistical temporal features and deep spatial features," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3311–3319.

[32] H. Ren, D. Chen, and Y. Wang, "RAN4IQA: Restorative adversarial nets for no-reference image quality assessment," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7308–7314.

[33] D. Pan *et al.*, "Blind predicting similar quality map for image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6373–6382.

[34] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[35] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. De Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 652–671, Oct. 2012.

[36] 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.

[37] F. Zhang, F. M. Moss, R. Baddeley, and D. R. Bull, "BVI-HD: A video quality database for HEVC compressed and texture synthesized content," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2620–2630, Oct. 2018.

[38] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[39] A. V. Katsenou, G. Dimitrov, D. Ma, and D. R. Bull, "BVI-SynTex: A synthetic video texture dataset for video compression and quality assessment," *IEEE Trans. Multimedia*, vol. 23, pp. 26–38, 2021.

[40] J. Yang *et al.*, "No-reference quality assessment of stereoscopic videos with inter-frame cross on a content-rich database," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3608–3623, Oct. 2019.

[41] H. Wang *et al.*, "MCL-JCV: A JND-based H.264/AVC video quality assessment dataset," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 1509–1513.

[42] M. Nuutinen *et al.*, "CVD2014—A database for evaluating no-reference video quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3073–3086, Jul. 2016.

[43] D. Ghadiyaram *et al.*, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2061–2077, Sep. 2017.

[44] 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. 2018.

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

[46] 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.

[47] X. Yu *et al.*, "Predicting the quality of compressed videos with pre-existing distortions," *IEEE Trans. Image Process.*, vol. 30, pp. 7511–7526, 2021.

[48] ITU-T, "P. 910: Subjective video quality assessment methods for multimedia applications," Tech. Rep., Apr. 2008.

[49] N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2678–2683, Sep. 2011.

[50] V. Vonikakis, R. Subramanian, and S. Winkler, "Shaping datasets: Optimal data selection for specific target distributions across dimensions," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 3753–3757.

[51] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Systems Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[52] FFmpeg. Accessed: Oct. 2021. [Online]. Available: https://www/ffmpeg.org

[53] ITU-T, "Subjective video quality assessment methods for multimedia applications," Tech. Rep., Feb. 2009.

[54] ITU-R, "Recommendation ITU-R BT.500-13: Methodology for the subjective assessment of the quality of television pictures," Tech. Rep., Jan. 2012.

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

[56] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, pp. 1398–1402.

[57] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[58] 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.
[59] 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.
[60] H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet, "Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator," *IEEE Access*, vol. 4, no. 1, pp. 5579–5590, 2016.
[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[62] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Int. Conf. Mach. Learn. PMLR*, 2015, pp. 448–456.
[63] K. Ma *et al.*, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 1004–1016, Feb. 2017.
[64] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.
[65] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 472–480.
[66] X. Yu, C. G. Bampis, P. Gupta, and A. C. Bovik, "Predicting the quality of images compressed after distortion in two steps," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5757–5770, Dec. 2019.
[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd Int. Conf. Learn. Representations, ICLR*, 2015, pp. 1–15.

Yang Li received the B.S. degree in electronic information engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2017. He is currently working toward the Ph.D. degree with Peking University, Beijing, China. Since 2018, he has been a Research Intern with Bytedance Inc. From 2019 to 2020, he was a Research Assistant with the Department of Computer Science, City University of Hong Kong. His research interests include data compression and image/video quality assessment.

Shengbin Meng received the B.S. degree in automation engineering from Tsinghua University, Beijing, China, in 2011, and the Ph.D. degree in computer science from Peking University, Beijing, China, in 2016. He is currently working with Video Architecture Team of ByteDance, Beijing, China. His research interests focus on how to optimize video streaming system and provide best quality-of-experience.

Xinfeng Zhang (Senior Member, IEEE) received the B.S. degree in computer science from the Hebei University of Technology, Tianjin, China, in 2007, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2014. From 2014 to 2017, he was a Research Fellow with the Rapid-Rich Object Search Lab, Nanyang Technological University, Singapore. From October 2017 to October 2018, he was a Postdoctoral Fellow with the School of Electrical Engineering System, University of Southern California, Los Angeles, CA, USA. From December 2018 to August 2019, he was a Research Fellow with the Department of Computer Science, City University of Hong Kong. He currently is an Assistant Professor with the School of Computer Science and Technology, University of Chinese Academy of Sciences. He authored more than 150 refereed journal/conference papers and was the recipient of the Best Paper Award of IEEE Multimedia 2018, the Best Paper Award at the 2017 Pacific-Rim Conference on Multimedia (PCM) and the Best Student Paper Award in IEEE International Conference on Image Processing 2018. His research interests include video compression and processing, image/video quality assessment, and 3D point cloud processing.

Meng Wang received the B.S. degree in electronic information engineering of Honors Program from China Agricultural University, Beijing, China, in 2015, the M.S. degree in computer application technology from Peking University, Beijing, China, in 2018, and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2021. She is currently a Postdoc with the Department of Computer Science, City University of Hong Kong. She has been a research intern with Bytedance Inc., since 2017. Her research interests include data compression and image/video coding.

Shiqi Wang (Senior Member, IEEE) received the B.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2008 and the Ph.D. degree in computer application technology from Peking University, Beijing, China, in 2014. From 2014 to 2016, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. From 2016 to 2017, he was a Research Fellow with the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore. He is currently an Assistant Professor with the Department of Computer Science, City University of Hong Kong. He has proposed more than 50 technical proposals to ISO/MPEG, ITU-T, and AVS standards, and authored or coauthored more than 200 refereed journal articles/conference papers. His research interests include video compression, image/video quality assessment, and image/video search and analysis. He was the recipient of the 2021 IEEE multimedia rising star award, the Best Paper Award from IEEE VCIP 2019, ICME 2019, IEEE Multimedia 2018, and PCM 2017 and is the coauthor of an article that received the Best Student Paper Award in the IEEE ICIP 2018.

Yue Wang received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2006, and the Ph.D. degree from the Graduate University of the Chinese Academy of Sciences, Beijing, China, in 2012. He did the Postdoctoral Research with the Department of Computer Science, Peking University, Beijing, China, from 2015 to 2017. He is currently the Director of VideoArch Department. His current research interests include image and video coding, processing, and transmission technology.

Siwei Ma (Senior Member, IEEE) received the B.S. degree from Shandong Normal University, Jinan, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005. He held a Postdoctoral position with the University of Southern California, Los Angeles, CA, USA, from 2005 to 2007. He joined the School of Electronics Engineering and Computer Science, Institute of Digital Media, Peking University, Beijing, where he is currently a Professor. He has authored more than 300 technical articles in refereed journals and proceedings in image and video coding, video processing, video streaming, and transmission. He was/is as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and the *Journal of Visual Communication and Image Representation*.