Full-reference Video Quality Assessment for User Generated Content Transcoding

Zihao Qi[†], Chen Feng[†], Duolikun Danier[†], Fan Zhang[†], Xiaozhong Xu[§], Shan Liu[§] and David Bull[†]
[†]*Visual Information Laboratory, University of Bristol, Bristol, UK, BS1 5DD*
{zihao.qi, chen.feng, duolikun.danier, fan.zhang, dave.bull}@bristol.ac.uk
[§]*Tencent Media Lab, Palo Alto, CA 94306, USA*
{xiaozhongxu, shanl}@tencent.com

***Abstract*—Unlike video coding for professional content, the delivery pipeline of User Generated Content (UGC) involves transcoding where unpristine reference content needs to be compressed repeatedly. In this work, we observe that existing full-/no-reference quality metrics fail to accurately predict the perceptual quality difference between transcoded UGC content and the corresponding unpristine references. Therefore, they are unsuited for guiding the rate-distortion optimisation process in the transcoding process. In this context, we propose a bespoke full-reference deep video quality metric for UGC transcoding. The proposed method features a transcoding-specific weakly supervised training strategy employing a quality ranking-based Siamese structure. The proposed method is evaluated on the YouTube-UGC VP9 subset and the LIVE-Wild database, demonstrating state-of-the-art performance compared to existing VQA methods. The source code of the developed quality metric and the associated training data are available from https://zihaoq1.github.io/FRUGC/.**

***Index Terms*—Video quality assessment, UGC, video transcoding, deep learning**

I. Introduction

Recent advances in video technology alongside rapidly increasing user numbers on various social media platforms have resulted in an explosion of User Generated Content (UGC) on the internet. Before being uploaded to video platform servers, many user generated source videos will exhibit distortions caused by inadequate photography skills, unprofessional equipment, and/or video compression on capture devices [1]. Once uploaded onto a steaming platform, the distorted source content is further transcoded and distributed to viewers, as shown in Fig. 1. When the UGC content is transcoded, existing video coding systems perform rate-distortion optimisation (RDO) to achieve maximum quality fidelity for a given bit rate. Crucial to the success of such RDO process is a full reference video quality assessment (VQA) method that can accurately measure the quality degradation of the transcoded version compared to its corresponding unpristine original video.

Existing video quality metrics can be divided into two major categories according to the availability of reference content:

The authors thank the funding from Tencent (US), University of Bristol, and the UKRI MyWorld Strength in Places Programme (SIPF00006/1).

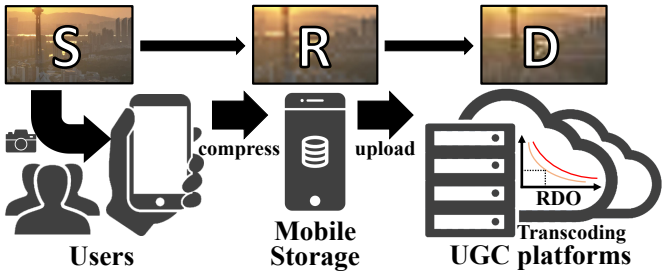


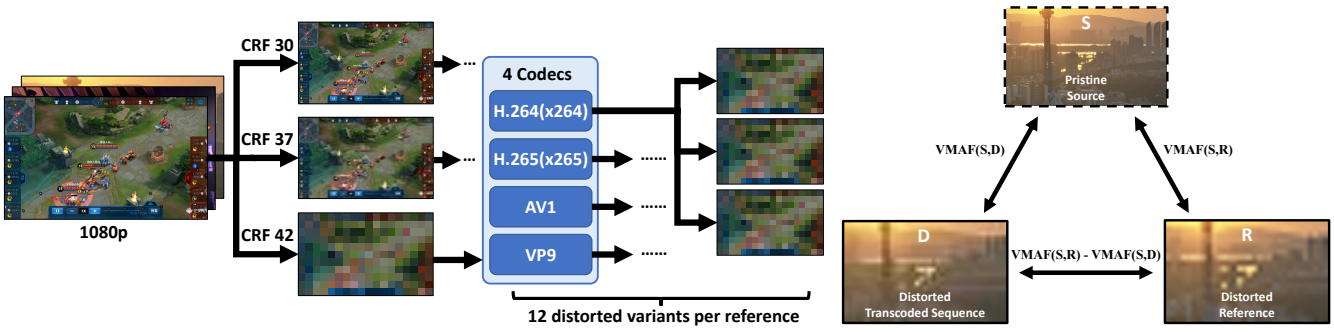
Fig. 1. Illustration of the UGC video delivery pipeline. The source content ($\mathbf{S}$) captured by a user is directly compressed on the user device for storage. These videos, denoted by distorted references ($\mathbf{R}$), are then uploaded onto UGC streaming platforms and further compressed into transcoded videos ($\mathbf{D}$) before being transmitted to the viewer.

full-reference (FR) and no-reference (NR)[1]. FR methods aim to capture the quality difference between the reference and distorted videos, which is inherently suitable for guiding the RDO in video coding. However, existing methods [2–9], either conventional or learning-based, assume pristine reference content, i.e. a lack of distortion in the reference. This assumption is often reflected in model design which relies on natural scene statistics [6, 10] and in training datasets which only include pristine references. For UGC coding, however, reference videos are often not pristine and can exhibit significant levels of distortion. Hence the use of conventional FR metrics may not be appropriate and could potentially lead to suboptimum performance (as shown in TABLE I).

NR VQA methods [11–15], on the other hand, aim to provide an absolute quality index for the distorted videos without considering the reference. In this case the artefacts inherited from the unpristine reference are considered alongside those arising from transcoding. However, due to the absence of a reference, the precision of NR methods can be inherently low in terms of measuring the perceptual degradation relative to the reference sequence. Additionally, taking the NR quality index difference between the unpristine reference and the reconstructed sequence can also be ineffective (as shown in TABLE I) due to the accumulation of quality estimation errors.

[1]Although reduced reference (RR) VQA methods do exist in the literature, they are not commonly deployed in practical applications, in particular in the context of UGC.

2024 Picture Coding Symposium (PCS) | 979-8-3503-5848-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/PCS60826.2024.10566416



(A) (B)

Fig. 2. (A) The generation of the training database for optimising the PQANet. (B) The illustration on how VMAF was used in the quality label annotation.

In this paper we investigate the unique characteristics of UGC full-reference VQA methods for transcoding applications, and expose the need for further research through quantitative experiments. To address this, we propose a FR VQA model, which is trained in a weakly-supervised learning manner based on a Siamese structure. To enable this training methodology, we have developed a new large-scale database with associated quality labels annotated using a proxy quality metric, VMAF. The resulting model was evaluated through a quantitative experiment and compared with both conventional and state-of-the-art FR/NR metrics. Our results confirm the unsatisfactory performance of existing VQA metrics in the targeted scenario of UGC transcoding, and demonstrate the evident improvement achieved by the proposed method.

II. Proposed Algorithm

This section first describes the unique characteristics of full reference video quality assessment for UGC transcoding, then presents a new training strategy for optimising a deep VQA network which predicts the quality difference between a transcoded video and its corresponding unpristine reference. This training strategy has inspired the development of a new large training database with reliable quality annotations for FR VQA in the context of UGC transcoding.

A. The UGC Transcoding Pipeline

In the UGC delivery pipeline of Fig. 1, the video is processed in three stages. When a video is captured, typically using a mobile device, the "pristine source", $\mathbf{S}$, is not stored losslessly. Instead, it is directly compressed by a video codec integrated within the device, e.g, by H.264/AVC [16] or H.265/HEVC [17], to generate a version with much lower bit rate, denoted as the "distorted reference", $\mathbf{R}$. This signal tends to contain visible compression artefacts compared to the original uncompressed UGC video. These distorted references are then uploaded by users to a UGC video platform, where they are further transcoded into distorted versions, denoted as the "distorted transcoded sequence", $\mathbf{D}$ (another layer of compression), before distribution to viewers across variable bandwidth networks. While conventional FR VQA focuses on the single-layer compression scenario, where the pristine $\mathbf{S}$ is encoded into $\mathbf{R}$, here we target the transcoding from $\mathbf{R}$ to $\mathbf{D}$.

B. The Employed Network Architecture

We use the network architecture proposed in [9]. As illustrated by Fig. 3. (A), this network first takes a pair of transcoded and reference patches, $\mathbf{P}_{\mathbf{D}_1}$ and $\mathbf{P}_{\mathbf{R}_1}$, as input, which correspond to the co-located 256×256×12 (H×W×T) video patches extracted from the transcoded video $\mathbf{D}_1$ and its distorted reference $\mathbf{R}_1$ respectively. The input patches are then processed in two stages. In Stage 1, a transformer-based patch-wise quality assessment network (PQANet) is employed to output a quality index, $Q_{(\mathbf{P}_{\mathbf{R}_1},\mathbf{P}_{\mathbf{D}_1})}$. In the second stage, the quality indices generated by PQANet for all the video patches from the same transcoded and reference videos are then passed to the Spatio-Temporal Aggregation Network (STANet) to obtain the sequence level quality index for $\mathbf{D}_1$ (against $\mathbf{R}_1$). Details on network architecture designs can be found in [9].

C. The Training Strategy

It is noted that [9] uses a ranking-inspired training methodology which allows the use of a proxy quality metric (VMAF [6] in this case) to generate reliable quality labels for the training content. This enables the creation of a large and diverse database to improve the generalisation of the deep VQA models without performing expensive and extensive subjective experiments. However, this training methodology cannot be directly applied here, because existing quality metrics such as VMAF cannot provide a robust quality prediction if the reference content contains various UGC artefacts (this has been confirmed by [18] and by our results in TABLE I).

To address this issue in our training process, we additionally employ the pristine source, $\mathbf{P}_{\mathbf{S}_1}$ and $\mathbf{P}_{\mathbf{S}_2}$, for quality labelling to improve the annotation reliability, as shown in Fig. 3. (B). It is noted that these pristine source versions are not input to the network in either training or inference stages.

Specifically, during the training process, based on the ranking-inspired training method used in [9, 19], two pairs of video patches, $(\mathbf{P}_{\mathbf{R}_1},\mathbf{P}_{\mathbf{D}_1})$ and $(\mathbf{P}_{\mathbf{R}_2},\mathbf{P}_{\mathbf{D}_2})$ are input into PQANet and, based on their outputs $Q_{(\mathbf{P}_{\mathbf{R}_1},\mathbf{P}_{\mathbf{D}_1})}$ and $Q_{(\mathbf{P}_{\mathbf{R}_2},\mathbf{P}_{\mathbf{D}_2})}$, the probability p of patch $\mathbf{P}_{\mathbf{D}_1}$ being of higher quality (with respect to its reference counterpart) than $\mathbf{P}_{\mathbf{D}_2}$ is obtained using a sigmoid function:

$$p = \text{sigmoid}(Q_{(\mathbf{P}_{\mathbf{R}_1},\mathbf{P}_{\mathbf{D}_1})} - Q_{(\mathbf{P}_{\mathbf{R}_2},\mathbf{P}_{\mathbf{D}_2})}). \quad (1)$$

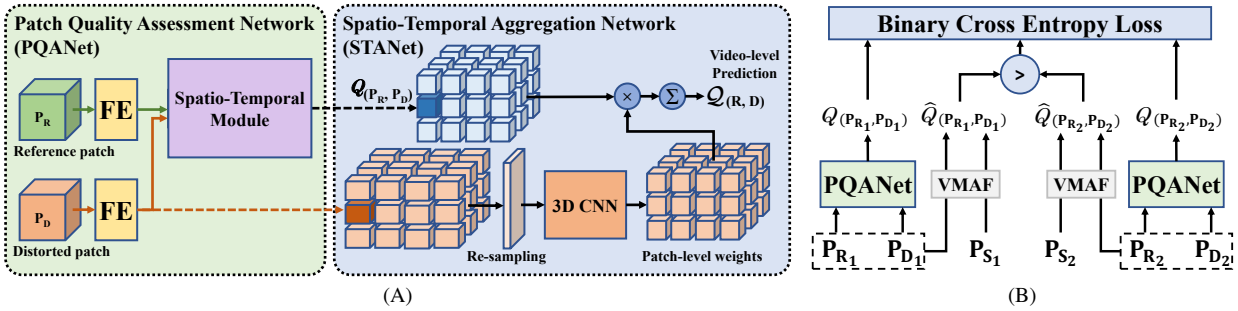


Fig. 3. (A) The overall picture of the inference workflow. (B) The illustration of the proposed training strategy.

To obtain the training targets, we also employ VMAF [6] as a proxy metric here, but based on a different approach compared to [9]. Specifically, we calculate the VMAF values, $\mathrm{VMAF}(\mathbf{P_{S_1}}, \mathbf{P_{D_1}})$, $\mathrm{VMAF}(\mathbf{P_{S_2}}, \mathbf{P_{D_2}})$, $\mathrm{VMAF}(\mathbf{P_{S_1}}, \mathbf{P_{R_1}})$ and $\mathrm{VMAF}(\mathbf{P_{S_2}}, \mathbf{P_{R_2}})$ separately. Based on these, we then obtain the quality differences, $\hat{Q}_{(\mathbf{P_{R_1}}, \mathbf{P_{D_1}})}$ and $\hat{Q}_{(\mathbf{P_{R_2}}, \mathbf{P_{D_2}})}$ between the unpristine reference and the distorted content for both patch pairs:

$$\begin{cases} \hat{Q}_{(\mathbf{P_{R_1}}, \mathbf{P_{D_1}})} = \mathrm{VMAF}(\mathbf{P_{S_1}}, \mathbf{P_{R_1}}) - \mathrm{VMAF}(\mathbf{P_{S_1}}, \mathbf{P_{D_1}}) \\ \hat{Q}_{(\mathbf{P_{R_2}}, \mathbf{P_{D_2}})} = \mathrm{VMAF}(\mathbf{P_{S_2}}, \mathbf{P_{R_2}}) - \mathrm{VMAF}(\mathbf{P_{S_2}}, \mathbf{P_{D_2}}) \end{cases} \tag{2}$$

It is noted that in the VMAF calculations above, as we always use pristine content as references, the accuracy of the quality prediction has been maintained. This process has also been illustrated by Fig. 2.(B).

As in [9], rather than minimising the difference between the quality difference $\hat{Q}$ and the network output Q, we further obtain the quality ranking information r between two quality difference values $\hat{Q}_{(\mathbf{P_{R_1}}, \mathbf{P_{D_1}})}$ and $\hat{Q}_{(\mathbf{P_{R_2}}, \mathbf{P_{D_2}})}$:

$$r = \begin{cases} 1, \text{if } \hat{Q}_{(\mathbf{P_{R_1}}, \mathbf{P_{D_1}})} - \hat{Q}_{(\mathbf{P_{R_2}}, \mathbf{P_{D_2}})} > \sigma \\ 0, \text{if } \hat{Q}_{(\mathbf{P_{R_1}}, \mathbf{P_{D_1}})} - \hat{Q}_{(\mathbf{P_{R_2}}, \mathbf{P_{D_2}})} < -\sigma \end{cases} \tag{3}$$

Here σ is a threshold, which is based on the ranking ability of $\hat{Q}_{(\cdot,\cdot)}$. For cases when $-\sigma \leq r \leq \sigma$, we have excluded them in the database generation process. More detailed definition of σ and its values are described in Section II-D. Based on this, we can train the network with the binary cross entropy loss:

$$\mathcal{L} = -(r \cdot \log p + (1 - r) \cdot \log(1 - p)). \tag{4}$$

This training strategy is illustrated by Fig. 3.(B).

The training methodology for STANet remains the same as that in the original RankDVQA paper [9]. The only change is the training database, as described below.

D. Training Databases

Training database for Stage 1. For the training of PQANet, to closely simulate the UGC transcoding pipeline described in Section II-A, we firstly collated 252 pristine source sequences from the BVI-DVC dataset [20], the CVPR 2022 CLIC challenge training dataset [21] and the YouTube-UGC 2K database [18] (non-overlapping with VP9 subset) alongside 6 self-captured videos. These 258 original sequences ($\mathbf{S}$) were then compressed with x264 [22] (CRF = 30, 37, 42, medium preset), a commonly used video codec on mobile devices, to generate distorted reference $\mathbf{R}$. This is to simulate the compression process on user-capture devices. Each distorted reference video was further compressed into 12 transcoded sequences ($\mathbf{D}$) at three quantisation levels, by 4 different codecs (x264, x265 [22], AV1 [23], VP9 [24]). This emulates the transcoding process on UGC platforms. The workflow is illustrated in Fig. 2 (A). This results in 9,288 distorted sequences.

Based on the generated transcoded sequences described above and their associated distorted references ($\mathbf{R}$), we adopted the same patch generation method as in [9], which randomly crops 256×256×12 spatio-temporal patches. We further combined two different patch groups as a training instance $\{(\mathbf{P_{R_1}}, \mathbf{P_{D_1}}), (\mathbf{P_{R_2}}, \mathbf{P_{D_2}}), r\}$ to enable ranking-based optimisation. In each training instance, two $(\mathbf{P_R}, \mathbf{P_D})$ pairs can either correspond to the same distorted reference sequence, denoted by single source (SS), or to a different distorted reference sequence, labelled as dual source (DS). The label r is obtained by Equation (2) and (3).

To ensure the reliability of the VMAF-based training labels here, we followed the approach described in [9, 19] to evaluate the ranking ability of $\hat{Q}_{(\cdot,\cdot)}$ in Equation (3) and determine the value of the threshold σ, in the context of the UGC transcoding scenario based on the ICME Challenge database [25]. It contains 900 unpristine references alongside 6,300 transcoded sequences (excluding the testing set). Specifically, we calculate the accuracy of $\hat{Q}_{(\cdot,\cdot)}$ when it is used to differentiate quality differences between every two pairs of transcoded sequences and their unpristine references (the same or different) based on the subjective ground truth. The results are shown in Fig. 4. It can be observed that, when the $\hat{Q}_{(\cdot,\cdot)}$ predicted quality difference, $|\hat{Q}_{(\mathbf{P_{R_i}}, \mathbf{P_{D_i}})} - \hat{Q}_{(\mathbf{P_{R_j}}, \mathbf{P_{D_j}})}|$ (following Equation (2)), is larger than 0 (for the single source scenario) or 6 (for the dual sources), the ranking accuracy according to the actual subjective score is above 96% (which provides a good trade off between the number of training instances and reliability). These two thresholds are hence used to examine the labelled patches - for each training instance; if the absolute difference in $\hat{Q}_{(\cdot,\cdot)}$ is smaller than these thresholds, we exclude it. Otherwise, we obtain the label according to Equation (3). This results in 315,059 training instances (80% SS, 20% DS).

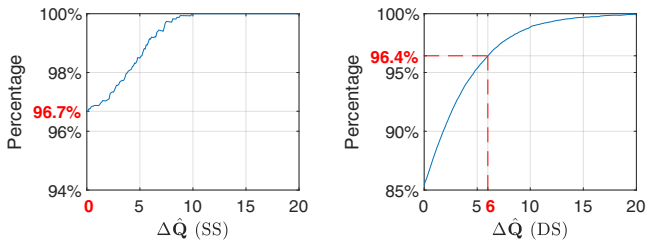


Fig. 4. $\hat{Q}$ difference (see Equation (3)) stands for the difference between two $\hat{Q}$ values (on video-level). Percentage: the accuracy ratio of $\hat{Q}$ difference-based ranking results that are consistent with human judgement. Results produced on ICME database [25]. (Left) Single sources (SS). (Right) Dual sources (DS).

Training database for Stage 2. For STANet, we followed the same training strategy as in [9], but used the ICME Challenge 2021 [25] database to optimise the spatio-temporal pooling network. Here we expect STANet to take all the patch-level quality indices and output the final sequence-level quality score. Detailed training procedures can be found in [9].

E. Implementation Details

Pytorch 1.10 was used to implement both PQANet and STANet. The training parameters used were: Adam optimisation with hyper-parameters of $\beta 1 = 0.9$ and $\beta 2 = 0.999$; 60 training epochs; batch size of 4; the initial learning rate is 0.0001 with a weight decay of 0.1 after every 20 epochs. Both training and evaluation were executed on a computer with a 2.4GHz Intel CPU and an NVIDIA P100 graphic card.

III. Results and Discussion

Most existing UGC databases, such as YouTube-UGC [18], KoNVid-1k [26], and LIVE-VQC [27], are designed for no-reference video quality assessment, and hence cannot be employed here for evaluating full reference quality metrics. Therefore we conducted a benchmark experiment based on two UGC video quality databases with unpristine references, YouTube-UGC VP9 subset [18] and LIVE-WILD [28], which can simulate the UGC transcoding scenario. The YouTube-UGC VP9 subset is part of YouTube-UGC [18] database, containing 507 transcoded sequences compressed by VP9 based on 169 reference sequences with visual artefacts. LIVE-WILD consists of 220 distorted sequences which are derived from 55 unpristine reference sequences. It is noted that the overall perceptual quality of the reference content in the LIVE WILD database is higher than that of the YouTube-UGC VP9 subset. This implies that the latter may be more challenging when used for evaluating FR VQA methods.

Our proposed method was tested against 7 existing FR VQA methods (PSNR, SSIM, MS-SSIM [3], LPIPS [7], VMAF [6], C3DVQA [8] and RankDVQA [9]), and 6 NR quality metrics (NIQE [11], BRISQUE [13], VIIDEO [29], VBLIINDS [12], VIDEVAL [14] and SimpleVQA [15]). Here PSNR, SSIM, MS-SSIM, NIQE, BRISQUE, VIIDEO and VBLIINDS are conventional VQA methods, while VMAF is a regression-based quality metric. C3DVQA, RankDVQA, VIDEVAL and SimpleVQA are deep learning-based VQA methods, and their pre-trained models were used for benchmarking. For all learning-based methods including ours, we did not perform intra database cross-validation, and there is no overlap between the training material and all the sequences in these two benchmark databases.

TABLE I

FR and NR results on available FR UGC datasets. In each cell, the values x(y) corresponds to the SROCC or KRCC value (x) and the F-test result (y) at 95% confidence interval. Here y= 1 indicates that the metric is superior to our proposed method (y= −1 if the opposite is true), and y= 0 means that there is no significant performance difference between them.

Metrics	YT-UGC VP9		LIVE-WILD	
	SROCC	**KRCC**	**SROCC**	**KRCC**
Full-reference				
PSNR	0.3946 (-1)	0.2737 (-1)	0.7613 (-1)	0.5687 (-1)
SSIM	0.5384 (-1)	0.3752 (-1)	0.8711 (0)	0.6755 (0)
MS-SSIM [3]	0.5290 (-1)	0.3677 (-1)	0.8753 (0)	0.6788 (0)
LPIPS [7]	0.4491 (-1)	0.3087 (-1)	0.8587 (0)	0.6618 (0)
VMAF 0.6.1 [6]	0.4378 (-1)	0.3025 (-1)	0.8957 (0)	0.7127 (0)
C3DVQA [8]	0.4732 (-1)	0.3220 (-1)	0.6980 (-1)	0.5157 (-1)
RankDVQA [9]	0.4725 (-1)	0.3272 (-1)	0.8803 (0)	0.6958 (0)
No-reference				
NIQE [11]	0.1718 (-1)	0.1136 (-1)	0.3443 (-1)	0.2352 (-1)
BRISQUE [13]	0.0934 (-1)	0.0632 (-1)	0.7297 (-1)	0.5371 (-1)
VIIDEO [29]	0.2121 (-1)	0.1425 (-1)	0.0201 (-1)	0.0148 (-1)
VBLIINDS [12]	0.2042 (-1)	0.1365 (-1)	0.3660 (-1)	0.2529 (-1)
VIDEVAL [14]	0.1868 (-1)	0.1089 (-1)	0.3402 (-1)	0.2291 (-1)
SimpleVQA [15]	0.4741 (-1)	0.3318 (-1)	0.8226 (-1)	0.6277 (-1)
Ours	**0.6974**	**0.4951**	**0.8975**	**0.7149**

All FR quality metrics tested are used to predict the quality differences between transcoded videos and their unpristine references. The Spearman Ranking Order Correlation Coefficients (SROCC), and Kendall Ranking Correlation Coefficients (KRCC) between their predicted quality difference scores and the true subjective different mean opinion scores (DMOS) on each database are employed to measure their performance. For each NR metric, we adapt to the FR transcoding scenario to calculate the quality index of a transcoded video and its unpristine reference separately, and then obtain their quality differences to correlate with the DMOS[2].

Table I summarises the performance results for all the tested full reference and no reference VQA methods on both databases. It can be observed that the proposed FR quality metric outperforms all other tested FR and NR VQA methods on both databases. Furthermore, the improvement of our method over the second performer on the YT-UGC VP9 dataset is much higher than that based on the LIVE-WILD database. This is most likely because the reference sequences in the former contain more visual artefacts compared to the latter. This makes the YT-UGC VP9 dataset much more challenging, as confirmed by the relatively low SROCC values of all the tested quality metrics (below 0.7). It is also noted that on both

[2]We did not evaluate the quality predictions of NR methods on the distorted videos against their MOS values (their corresponding SROCC values are relatively low anyway - up to 0.3882 on YT-UGC VP9 and 0.7113 on LIVE-WILD), because our focus is on the quality differences between transcoded content and unpristine references.

databases, the proposed method performs well compared to RankDVQA - this verifies the effectiveness of the new training methodology, which is one of our primary contributions.

Table I also provides statistical test results to differentiate the performance between the proposed quality metric and each of the benchmark results. Here we employed an F-test following [30], between the prediction residuals (after a non-linear fitting) of our method and other benchmark approaches. The results show that the proposed method is significantly better than all the other tested quality metrics on the YT-UGC VP9 database. On the LIVE-WILD database, there is no significant difference between SSIM, MS-SSIM, VMAF, RankDVQA, and our approach, but the latter statistically performs better than all the other FR and NR metrics.

IV. Conclusion

In this paper, we first justified the need for tailored metrics for assessing transcoded UGC videos, because existing VQA methods cannot accurately predict the quality difference between them and their unpristine references. To address this issue, a new FR VQA method is proposed based on a weakly-supervised Siamese training methodology and a large training dataset with reliable quality annotations. Our results demonstrate the improvement of the proposed method over existing FR and NR quality metrics when tested on two full reference UGC transcoding databases. Future work should focus on further improving the correlation performance by enhancing the architecture for the focused application scenario.

References

[1] Z. Zhang, W. Wu, W. Sun, D. Tu, W. Lu, X. Min, Y. Chen, and G. Zhai, "MD-VQA: Multi-dimensional quality assessment for ugc live videos," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023, pp. 1746–1755.

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

[3] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conf. on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[4] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *2011 18th IEEE International Conf. on Image Processing*. IEEE, 2011, pp. 2505–2508.

[5] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. on image processing*, vol. 19, no. 2, pp. 335–350, 2009.

[6] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, no. 2, 2016.

[7] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. of the IEEE Conf. on computer vision and pattern recognition*, 2018, pp. 586–595.

[8] M. Xu, J. Chen, H. Wang, S. Liu, G. Li, and Z. Bai, "C3DVQA: Full-reference video quality assessment with 3d convolutional neural network," in *ICASSP 2020-2020 IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4447–4451.

[9] C. Feng, D. Danier, F. Zhang, and D. R. Bull, "RankDVQA: Deep vqa based on ranking-inspired hybrid training," *arXiv preprint arXiv:2202.08595*, 2022.

[10] 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. of the European Conf. on Computer Vision*, 2018, pp. 219–234.

[11] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

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

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

[14] 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. on Image Processing*, vol. 30, pp. 4449–4464, 2021.

[15] W. Sun, X. Min, W. Lu, and G. Zhai, "A deep learning based no-reference quality assessment model for ugc videos," in *Proc. of the 30th ACM International Conf. on Multimedia*, 2022, pp. 856–865.

[16] I. Telecom, "Advanced video coding for generic audiovisual services," *ITU-T Recommendation H. 264*, 2003.

[17] V. Sze, M. Budagavi, and G. J. Sullivan, "High efficiency video coding (hevc)," in *Integrated circuit and systems, algorithms and architectures*. Springer, 2014, vol. 39, p. 40.

[18] Y. Wang, S. Inguva, and B. Adsumilli, "Youtube ugc dataset for video compression research," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–5.

[19] Q. Hou, A. Ghildyal, and F. Liu, "A perceptual quality metric for video frame interpolation," in *European Conf. on Computer Vision*. Springer, 2022, pp. 234–253.

[20] D. Ma, F. Zhang, and D. Bull, "BVI-DVC: A training database for deep video compression," *IEEE Trans. on Multimedia*, 2021.

[21] "Workshop and challenge on learned image compression (clic2022)," in *CVPR*, 2022. [Online]. Available: https://clic.compression.cc/2022/

[22] "ffmpeg." [Online]. Available: http://ffmpeg.org/

[23] "Alliance for open media." [Online]. Available: https://aomedia.org

[24] D. Mukherjee, J. Bankoski, A. Grange, J. Han, J. Koleszar, P. Wilkins, Y. Xu, and R. Bultje, "The latest open-source video codec vp9-an overview and preliminary results," in *2013 Picture Coding Symposium (PCS)*. IEEE, 2013, pp. 390–393.

[25] H. Wang, G. Li, S. Liu, and C.-C. J. Kuo, "Challenge on quality assessment of compressed ugc videos," 2021.

[26] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The konstanz natural video database (konvid-1k)," in *2017 Ninth International Conf. on Quality of Multimedia Experience (QoMEX)*. IEEE, 2017, pp. 1–6.

[27] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. on Image Processing*, vol. 28, no. 2, pp. 612–627, 2018.

[28] X. Yu, N. Birkbeck, Y. Wang, C. G. Bampis, B. Adsumilli, and A. C. Bovik, "Predicting the quality of compressed videos with pre-existing distortions," *IEEE Trans. on Image Processing*, vol. 30, pp. 7511–7526, 2021.

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

[30] F. Zhang and D. R. Bull, "A perception-based hybrid model for video quality assessment," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 26, no. 6, pp. 1017–1028, 2015.