Perceptual Evaluation of Pre-processing for Video Transcoding

Shiyu Huang, Ziyuan Luo, Jiahua Xu,Wei Zhou, Zhibo Chen*
CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System
University of Science and Technology of China
Hefei, China
chenzhibo@ustc.edu.cn

2021 International Conference on Visual Communications and Image Processing (VCIP) | 978-1-7281-8551-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/VCIP53242.2021.9675438

***Abstract*—Recently, the pre-processed video transcoding has attracted wide attention and has been increasingly used in practical applications for improving the perceptual experience and saving transmission resources. However, very few works have been conducted to evaluate the performance of pre-processing methods. In this paper, we select the source (SRC) videos and various pre-processing approaches to construct the first Pre-processed and Transcoded Video Database (PTVD). Then, we conduct the subjective experiment, showing that compared with the video sent to the codec directly at the same bitrate, the appropriate pre-processing methods indeed improve the perceptual quality. Finally, existing image/video quality metrics are evaluated on our database. The results indicate that the performance of the existing image/video quality assessment (IQA/VQA) approaches remain to be improved. We will make our database publicly available soon.**

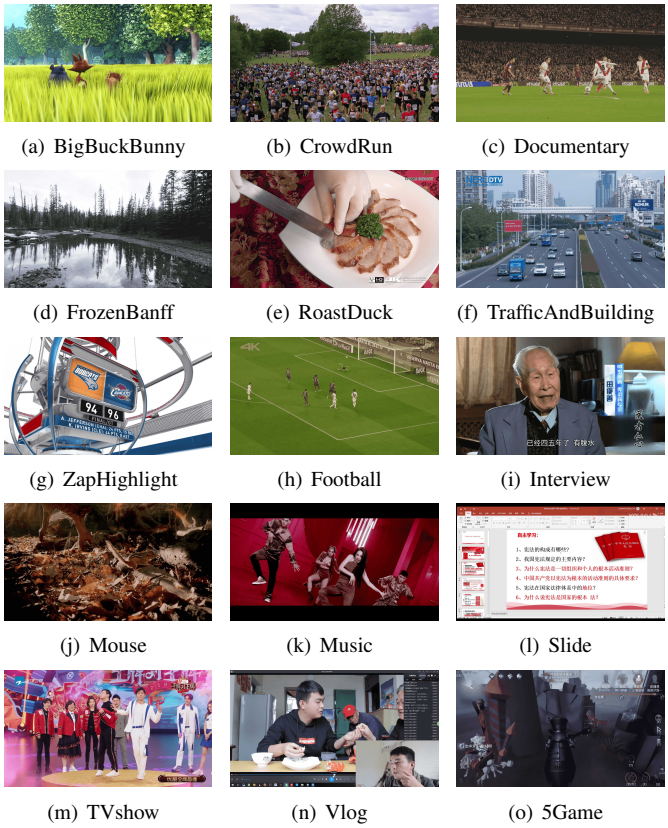
***Index Terms*—Video quality assessment, Subjective evaluation, Database, Pre-processing for video transcoding**

I. Introduction

With the rapid development of media technology and the popularity of mobile display devices, professionally-generated content (PGC) videos and user-generated content (UGC) videos become very popular. However, due to limited transmission bandwidth and coding efficiency, the video quality is often severely compromised. To provide users with better experience when watching videos with same or even fewer transmission resources, pre-process have been applied in online streaming services for video coding [1]. Therefore, an evaluation model that reflects the quality of user experience is needed [2], [3]. But there are few works on quality assessment (QA) that explore how to measure the subjective quality of pre-processed based video transcoding. And thus, the QA method for the pre-processed and transcoded video is attracted and in high demand.

The image/video quality assessment (IQA/VQA) can be divided into two categories: subjective and objective, according to whether human is involved or not. The subjective QA requires humans as the observers to evaluate the quality [4]–[6]. The subjective QA is precise, reliable, and indispensable since the human perception scores are usually identified as the ground truth to evaluate objective QA methods [7]. However, it is too labor-intensive and time-consuming to be applied to practical applications [8]. On the contrary, the objective QA model is an alternative QA solution for practical applications,

Zhibo Chen is the corresponding author. (chenzhibo@ustc.edu.cn)



(a) BigBuckBunny (b) CrowdRun (c) Documentary
(d) FrozenBanff (e) RoastDuck (f) TrafficAndBuilding
(g) ZapHighlight (h) Football (i) Interview
(j) Mouse (k) Music (l) Slide
(m) TVshow (n) Vlog (o) 5Game

Fig. 1. The representative frames of all the 15 source (SRC) videos in our established PTVD covering various types

usually taking the characteristics of spatial-temporal information and human visual system (HVS) into account [9], [10]. The quality scores computed for the input images/videos by a well designed objective QA model are expected to concur with subjective perception.

In general, objective IQA/VQA methods can be classified into three categories: full-reference (FR), reduced-reference (RR), and no-reference (NR). For FR methods, the full information of original contents is needed, e.g. mean square error (MSE) and peak signal-to-noise ratio (PSNR). Considering the HVS, structure similarity between original and distorted images is measured in structural similarity index (SSIM) [11], with several variants, e.g. MS-SSIM [12] and FSIM [13]. The visual information fidelity (VIF) [14] measures the degree of information loss of the distorted images compared to the

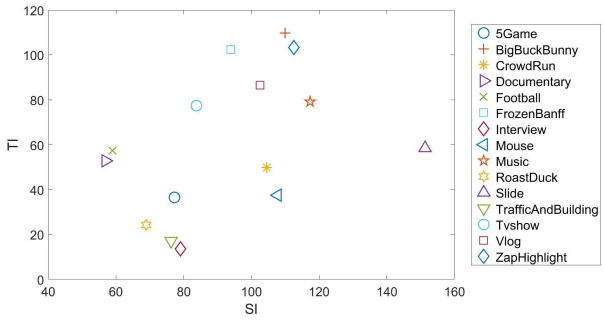


Fig. 2. Distribution of SI and TI.

reference images. Recently, video multi-method assessment fusion (VMAF) [15] proposed by Netfix has attracted much attention. The support vector machine (SVM) [16] is utilized in VMAF to fuse several metrics, proved to be more consistent with subjective feelings. RR methods only require the part of the information from original contents [3]. The NR methods estimate the quality without the original multimedia information, which are more applicable in most real-world scenarios, e.g. Mittal *et al*. [17] used natural scene statistics in the spatial domain, and binocular vision mechanisms were utilized in BSVQE [18]. However, they are not skilled in distinguishing the quality with similar bit rates and evaluating the performance of pre-processing methods for video transcoding. There exists a few works focusing on enhanced images/videos quality assessment [19]–[21]. Nonetheless, the impacts of transcoding distortion are not taken into account.

In this paper, we make the first attempt to carry out the specific subjective database for the evaluation of pre-process based video transcoding, which contains both static frames and dynamic videos. The video database, called Pre-processed and Transcoded Video Database (PTVD), is built for the purpose of QA for the PTVs. The proposed PTVD has two subsets: PTVD-I for static frames of PTVs and PTVD-II for original dynamic videos. Each of the subsets contains 15 reference videos, 570 processed PTVs, and the corresponding subjective scores from 30 observers.

The rest of the paper is organized as follows: Section II introduces the establishment of our datasets PTVD. The analyses of the PTVD are shown in Section III. Section IV presents the experimental results of existing metrics and we conclude the paper in Section V.

II. Pre-processed and Transcoded Video Database

To investigate the perceptual quality of the PTVs, the first pre-processed and transcoded video database is established, including two subsets: PTVD-I for static frames and PTVD-II for dynamic video sequences. We apply several typical pre-processing methods (including sharpening, smoothing, a mixture of them, etc.) to selected original video contents.

A. Selection of PTV Contents

Our constructed PTVD contains 15 source (SRC) videos, covering various contents (e.g. cartoon, sports, games, screen content, natural scenes). All SRC videos are cut into 7 seconds sequences with the same resolution of 1920×1080. For demonstration, Fig. 1 shows a typical frame of 15 SRC videos. Fig. 1 (a)-(g) are from public domain while Fig. 1 (h)-(o) are provided by Youku Inc [22].

In order to accurately describe the characteristics of these video contents, the spatial information (SI) and temporal information (TI) [23] of the SRC videos are illustrated in Fig. 2. Here, SI represents scene details while TI describes temporal variations. One can see that the distributions of the SI and TI are expansive, reflecting that our SRC videos cover a large range of content features.

B. Generation of PTV sequences

Pre-processing helps to turn images/videos into a form that can be easily processed by codecs [24]. To be specific, after appropriate pre-processing methods, codecs may save the bit rates with similar subjective quality, or visibly improve the subjective experience using equivalent bandwidth resources. Generally, the pre-processing methods for a SRC video include sharpening, smoothing, a mixture of the processing methods mentioned above, and network-based end-to-end frameworks [25], [26]. Specifically, sharpening methods aim to enhance the structure so that objects can still keep their shape when transcoding distortion is introduced, leading to better visual experience. Suitable smoothing approaches help codecs to utilize fewer bit rates for approaching identical subjective quality. In common applications, sharpening and smoothing approaches can be leveraged together in order to approach the best performance. Moreover, an end-to-end network is able to predict proper pre-processing method, the representative one among which is Narrowband HD 2.0 [27].

We use the sharpening and smoothing algorithm introduced in Narrowband HD 1.0 [28], which provides one parameter to control sharpening amplitude and the other one to control smoothing amplitude. Following the principle of the variable-controlling approach, we choose 10 sets of the pre-processing methods mentioned above with various groups of parameters

TABLE I
Hypothetic Reference Circuits (HRCs) of each source (SRC) video.

HRC-ID	Sharpening	Smoothing	Bit rates
1-5*	0	0	$VL,\ L,\ M,\ H,\ VH$ **
6-8	1	1	$L,\ M,\ H$
9-11	1	10	$L,\ M,\ H$
12-14	1	100	$L,\ M,\ H$
15-17	2	1	$L,\ M,\ H$
18-20	2	10	$L,\ M,\ H$
21-23	2	100	$L,\ M,\ H$
24-26	5	1	$L,\ M,\ H$
27-29	5	10	$L,\ M,\ H$
30-32	10	1	$L,\ M,\ H$
33-35	10	10	$L,\ M,\ H$
36-38	nbhd2		$L,\ M,\ H$

* original PTVs without pre-processing
** VL: Very Low; L: Low; M: Medium; H: High; VH: Very High

covering usual application scenarios from tweaking to massive alters, and one end-to-end pre-processing network: Narrow-band HD 2.0 (nbhd2), as is shown in Table I.

Each SRC video is encoded into 5 bitrate levels as {Very Low (VL), Low (L), Medium (M), High (H), Very High (VH)}, while other Hypothetic Reference Circuits (HRCs) are encoded into 3 bitrate levels as {L, M, H}. The HRC-1 (original PTV at VL bit rates) and the HRC-5 (original PTV at VH bit rates) are applied to provide the benchmark for other 36 types of HRCs. The codec for all HRCs is x264 [29].

C. Subjective Testing

To guarantee the accuracy of the mean opinion score (MOS) values for PTVD, 30 ratings are collected for each video. To demonstrate the stability and consistency in the subjective evaluation, 30 subjects are randomly divided into two groups with 15 individuals.

1) **PTVD-I**: As mentioned above, 570 original and processed PTVs have been generated to establish the PTVD. For each HRC video, a typical frame is selected randomly from all frames. PTVD-I is made up of 570 PTV frames mentioned above. Note that the frames of the same content are from the same original frame of the SRC video.

The subjective test is conducted with the single stimulus method under a normal lighting condition following the guidance of the ITU-R BT.500 [30]. We adopt the absolute category rating (ACR) scale. The scores are measured on 5 discrete scales with 1 for bad and 5 for excellent. Since our subjective experiment aims to compare different pre-processing methods, the subjects are encouraged to pay attention to details such as the edge, texture, and unnatural artifacts.

The subjective test system is displayed on the laptop with a 14-inch LED monitor with 1920×1080 resolution, 8GB RAM, and 64-bit Windows operating system. The PTV frames will be played on full screen for 5 seconds and then the subjects give their subjective scores for the current image by pressing a button from 1 to 5. The 30 subjects include 16 males and 14 females. Each of them is required to evaluate all 570 images while taking a break every half an hour watching. Before the formal subjective test, there exists an instruction provided to each subject for clearly explaining how to give the subjective scores for the PTVs, helping them to get familiar with the test and establishing stable assessment criteria.

2) **PTVD-II**: Different from the PTVD-I composed of PTV static frames, which shows the impact of the pre-processing methods in the spatial domain, the PTVD-II requires the subjects to evaluate the entire PTVs. The setting of PTVD-II subjective test is consistent with that of PTVD-I subjective test. It is clear that PTVD-II is closer to the real application scenarios and more effective to measure the performance of pre-processing methods in the spatial-temporal domain. The subjective setting of PTVD-II follows the regulation of the ITU-T P.910 [31].

III. Analyses of The Subjective Test

The subjects with the correlation coefficient to the average quality lower than 0.75 are considered as outliers, and their ratings are removed from our database [3]. At the same time, a new subject is required to take part in the subjective test. There remain 30 valid subjects for each PTVD subset after outlier removal. For each PTV, any single score S_i that does not fit the 3σ criterion will be regarded as exceptional data and will be eliminated until every S_i fits the criterion [32].

$$\left|S_i - \overline{S}\right| < 3\sigma, \tag{1}$$

where $\overline{S}$ means the average value, and σ means the standard deviation of valid scores in one PTV content. MOS values are computed for each PTV in the database by averaging scores of valid scores.

30 subjects are randomly divided into two bisected groups, each of which consists of 15 individuals. The correlation between the two groups is shown in Fig. 3. With the participants in each group increasing, the correlation between the two groups of MOS increases obviously. In PTVD-I, the correlation is 0.9627, and in PTVD-II, the correlation is 0.9739. It is sufficient to show the robustness and feasibility of our subjective experiments.

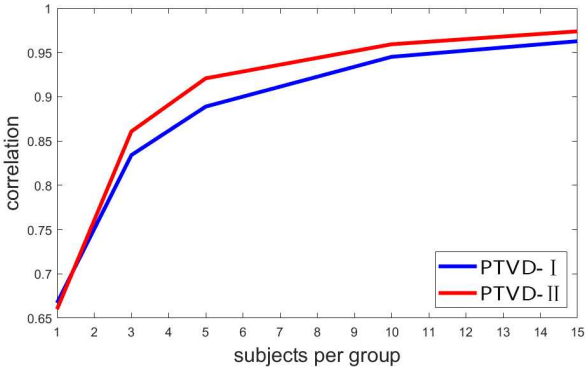


Fig. 3. The correlation between two groups of subjective scores. The X-axis represents the number of subjects per group, while the Y-axis represents the correlation.

Fig. 4 shows the relation of perceptual experience to pre-processing parameters for all contents. Appropriate pre-processing methods indeed effectively improve the subjective quality, while improper algorithm parameters will impair the perceived quality. When the sharpening parameter is at a relatively low level (1 or 2), the subjective quality increases at first and then decreases as the smoothing parameter increases. The turning points of the MOS curve may dissimilate at different bit rates. When the smoothing parameter is fixed (1 or 10), the subjective quality performs much similarly as the sharpening parameter increases. The nbhd2 method performs positively in these scenarios at various bit rates.

However, not all videos meet these rules, since the pre-processing algorithm has a different impact on videos with different content. For example, the images of animated content (BigBuckBunny) show monotonous increase in subjective quality with the increase of smoothing parameter in Fig. 5, when the sharpening parameter is small (1 or 2). And we can

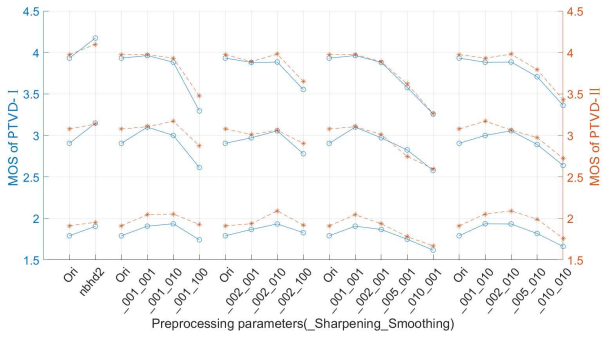


Fig. 4. The relation of average MOS to pre-processing parameters and bitrates for all contents. The X-axis represents pre-processing parameters, while the Y-axis represents MOS values. The blue solid line refers to PTVD-I and the orange dotted one refers to PTVD-II. The three bunches of lines from top to bottom mean bitrate H, M, L, respectively.

notice that this video does not meet the pattern in Fig. 4. To explain this phenomenon, we visualize the representative frames of BigBuckBunny with different pre-processing parameters, respectively in Fig. 6 (a) and Fig. 6 (b). Fig. 6 (a) and (b) take the same sharpening parameter 2 and the same coding mode of low bit-rate, while adapting different smoothing parameters respectively as 1 and 100. We can observe that compared with smoothing parameter as 1 in Fig. 6 (a), the details of tree trunk and leaf lines in the background are almost completely erased in Fig. 6 (b), but observers are generally more concerned about the quality of the foreground, so that the visual information loss is hardly noticed by the observers. Meanwhile, we can notice that high smoothing parameters can also reduce the block effect at low bit-rate point and improve the quality of coded video sequence to some extent. Therefore, a larger smoothing parameter can give the video a better subjective score within a certain range.

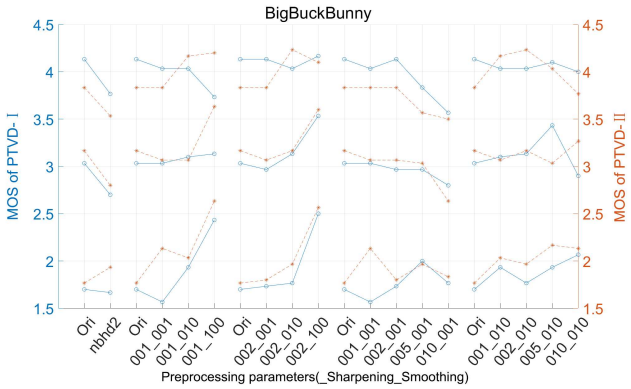


Fig. 5. The relation of average MOS to pre-processing parameters and bitrates for BigBuckBunny content.

IV. Experimental Result

Three evaluation criteria are selected to evaluate the performance of different metrics, consisting of Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC), and Root Mean Squared Error (RMSE). The PLCC evaluates the linear relationship between predicted score and MOS value, while the SROCC measures the monotonicity. The RMSE measures the accuracy of prediction. The better consistency with human perception is reflected in PLCC and SROCC closing to 1 as well as RMSE closing to 0.

(a) Sharpening parameter = 2, smoothing parameter = 1, low bit rate (b) Sharpening parameter = 2, smoothing parameter = 100, low bit rate

Fig. 6. Comparison of animated content BigBuckBunny with the same sharpening parameters

The performance of the objective metrics is shown in TABLE II. FSIM and VIF have outstanding performance compared with other objective metrics. However, the correlation between subjective and objective scores is still relatively low. Thus, there is plenty of room for performance improvement. Since few specific algorithms are proposed for assessment of pre-processing for video transcoding, a specifically designed model for PTV quality assessment is in demand.

TABLE II
PERFORMANCE OF THE OBJECTIVE METRICS IN OUR DATABASE.

	PTVD-I			PTVD-II		
Metric	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
PSNR	0.4848	0.4306	0.7460	0.5504	0.4179	0.7053
SSIM [11]	0.4543	0.4232	0.7598	0.5394	0.3907	0.7114
MS-SSIM [12]	0.7739	0.7646	0.5402	0.7390	0.5736	0.5691
FSIM [13]	**0.8615**	**0.8555**	**0.4331**	**0.8333**	0.7043	**0.4671**
VIF [14]	0.8590	0.8485	0.4366	0.8291	**0.8284**	0.4723
VMAF [15]	0.8168	0.8056	0.4921	0.8105	0.8101	0.4948
IFC [33]	0.7487	0.7440	0.5654	0.7135	0.7022	0.5919
BRISQUE [17]	0.4279	0.4011	0.7708	0.3929	0.3781	0.7769
NIQE [34]	0.5117	0.4993	0.7327	0.4088	0.3979	0.7710

The best performance results are highlighted in bold

V. Conclusion

In this work, we create the first PTVD for the increasingly popular pre-processed video transcoding application scenarios, covering both static frames and dynamic videos of PTVs. The database considers the perceived quality of pre-processing for video transcoding and demonstrates the impact of pre-processing methods. We also evaluate the performance of existing objective metrics in this database. The results show that a specific model for these scenarios is urgently required. Meanwhile, we will make our database publicly available. In the future, we will introduce more pre-processing methods and consider building a new model.

VI. Acknowledge

This work was supported in part by NSFC under Grant U1908209, 61632001, 62021001 and the National Key Research and Development Program of China 2018AAA0101400.

REFERENCES

[1] G.-M. Su, X. Su, Y. Bai, M. Wang, A. V. Vasilakos, and H. Wang, "Qoe in video streaming over wireless networks: perspectives and research challenges," *Wireless networks*, vol. 22, no. 5, pp. 1571–1593, 2016.

[2] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. IEEE, 2002, pp. IV–3313.

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

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

[5] J. Xu, C. Lin, W. Zhou, and Z. Chen, "Subjective quality assessment of stereoscopic omnidirectional image," in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 589–599.

[6] L. Shi, S. Zhao, W. Zhou, and Z. Chen, "Perceptual evaluation of light field image," in *2018 IEEE International Conference on Image Processing (ICIP)*, 05 2018.

[7] W. Zhou, N. Liao, Z. Chen, and W. Li, "3d-hevc visual quality assessment: Database and bitstream model," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2016, pp. 1–6.

[8] S. Cheng, H. Zeng, J. Chen, J. Hou, J. Zhu, and K.-K. Ma, "Screen content video quality assessment: Subjective and objective study," *IEEE Transactions on Image Processing*, vol. 29, pp. 8636–8651, 2020.

[9] W. Zhou, Z. Chen, and W. Li, "Dual-stream interactive networks for no-reference stereoscopic image quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3946–3958, 2019.

[10] J. Xu, Z. Luo, W. Zhou, W. Zhang, and Z. Chen, "Quality assessment of stereoscopic 360-degree images from multi-viewports," in *2019 Picture Coding Symposium (PCS)*. IEEE, 2019, pp. 1–5.

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

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

[13] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[14] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3. IEEE, 2004, pp. iii–709.

[15] "Vmaf:toward a practical perceptual video quality metric," https://github.com/Netflix/vmaf.

[16] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

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

[18] Z. Chen, W. Zhou, and W. Li, "Blind stereoscopic video quality assessment: From depth perception to overall experience," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 721–734, 2017.

[19] K. Gu, D. Tao, J.-F. Qiao, and W. Lin, "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 4, pp. 1301–1313, 2017.

[20] C. T. Vu, T. D. Phan, P. S. Banga, and D. M. Chandler, "On the quality assessment of enhanced images: A database, analysis, and strategies for augmenting existing methods," in *2012 IEEE Southwest Symposium on Image Analysis and Interpretation*, 2012, pp. 181–184.

[21] M. BARKOWSKY, E. MASALA, G. V. WALLENDAEL, K. BRUNNSTRÖM, N. STAELENS, and P. L. CALLET, "Objective video quality assessment — towards large scale video database enhanced model development," *IEICE Transactions on Communications*, vol. E98.B, no. 1, pp. 2–11, 2015.

[22] https://www.youku.com.

[23] I.-T. R. BT, "Methodology for the subjective assessment of video quality in multimedia applications (2007)," 1788.

[24] J. Deng, L. Wang, S. Pu, and C. Zhuo, "Spatio-temporal deformable convolution for compressed video quality enhancement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 696–10 703.

[25] Z. Chen, W. Lin, and K. N. Ngan, "Perceptual video coding: Challenges and approaches," in *2010 IEEE International Conference on Multimedia and Expo*. IEEE, 2010, pp. 784–789.

[26] P. Panda, K. H. El-Maleh, and H.-t. Li, "Adaptive filtering to enhance video encoder performance," Mar. 8 2011, uS Patent 7,903,733.

[27] "Apsaravideo for media processing: Transcoding multimedia data," 2017, https://developer.aliyun.com/article/225411.

[28] "Narrowband hd 1.0 - apsaravideo for media processing," 2018, https://www.alibabacloud.com/help/doc-detail/57691.html.

[29] L. Merritt and R. Vanam, "x264: A high performance h. 264/avc encoder," *online] http://neuron2. net/library/avc/overview_x264_v8_5. pdf*, 2006.

[30] B. Series, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation ITU-R BT*, pp. 500–13, 2012.

[31] P. ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," *International telecommunication union*, 1999.

[32] E. W. Grafarend, *Linear and nonlinear models: fixed effects, random effects, and mixed models*. de Gruyter, 2006.

[33] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on image processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[34] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.