Supplementary Materials

Yixuan Li[1], Bolin Chen[1], Baoliang Chen[1], Meng Wang[1], Shiqi Wang[1*]

[1*]Computer Science, City University of Hong Kong, Kowloon, Hong Kong, China.

*Corresponding author(s). E-mail(s): shiqwang@cityu.edu.hk;
Contributing authors: yixuanli423@gmail.com; bolinchen3-c@my.cityu.edu.hk;
blchen6-c@my.cityu.edu.hk; mwang98-c@my.cityu.edu.hk;

Abstract

This is the supplementary material for the paper entitled "Perceptual Quality Assessment of Face Video Compression: A Benchmark and An Effective Method". In this supplementary, we provide more details of the construction of our compressed face video quality assessment (CFVQA) dataset, including the compression parameter selection and implementation of six compression models. Subsequently, a study of the distortion types caused by different compression models is conducted. In addition, to further verify the effectiveness of the proposed video quality assessment (VQA) model, we visualize the extracted deep features and carry out case studies on video examples, providing useful evidence regarding the high correlation between the perceptual quality and predicted quality. Finally, the agreement testing results of subjective opinions for each compressed video reflect the trustworthiness of the dataset. The CFVQA benchmark is now publicly available on our project page: https://github.com/Yixuan423/Compressed-Face-Videos-Quality-Assessment.

Keywords: Face video compression, video quality assessment, subjective and objective study

1 Details of the CFVQA Dataset Construction

1.1 Compression Parameter Selection

In the dataset construction phase, a key component is selecting appropriate quantization parameters (QPs) for video compression. The philosophy of the selection lies in that: (1) For each specific codec, the video quality resulting from different compression levels should be discriminative and widespread enough. (2) For different codecs, the quality of compressed videos should be in an overlapped range. Following the two principles above, we conduct a user study of the QP selection for each codec. In detail, we first select five representative reference videos from the video pool and compress them with multiple potential QPs. As shown in the Table 1 and Fig. 1, the QPs cover a range from 22 to 47, which is practical in real applications. Then we collect the Mean Opinion Score (MOS) of each compressed video. The results are shown in Fig. 1, from which we can observe that: (1) the quality of videos compressed by two GAN-based compression models (FOMM Siarohin et al (2019) and CFTE Chen et al (2022)) cover a narrow range compared with other codecs. As such, we adopt all the potential QPs for our dataset construction. (2) The quality would be indiscriminate when the difference between two QPs is not significant. For example, the MOS difference of videos compressed by

VVC Bross et al (2021) in QP=22 and QP=27 are averagely smaller than 6, revealing the quality of those videos would not be distinguished easily by the subjects. To ensure the MOS reliability and quality diversity, several candidate QPs are finally removed for VVC and DVC Lu et al (2019) when the whole dataset is constructed. In the Table 1, we further present the selected QPs for each codec. In total, $135 \times 24 = 3,240$ compressed face videos are generated from 135 reference face videos.

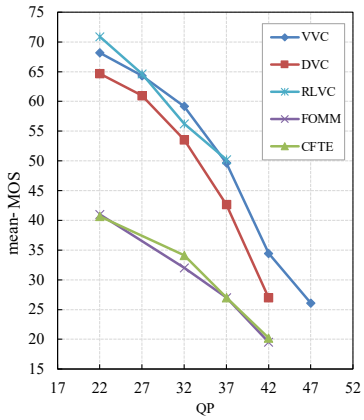


Fig. 1 MOS ranging along with the changes of QP values in terms of VVC, DVC, RLVC, FOMM, and CFTE compressed video collections.

Table 1 Candidate QP values for each compression method in the user study

Method	Candidate QPs	Selected QPs
VVC	22, 27, 32, 37, 42, 47	22, 32, 37, 42, 47
DVC	22, 27, 32, 37, 42	22, 32, 37, 42
RLVC	22, 27, 32, 37	22, 27, 32, 37
FOMM	22, 32, 37, 42	22, 32, 37, 42
CFTE	22, 32, 37, 42	22, 32, 37, 42

1.2 Implementation of Different Compression Methods

In the proposed CFVQA dataset, six video compression methods are employed, including the latest traditional hybrid video coding scheme VVC Bross et al (2021), reduced resolution compression, DVC Lu et al (2019), RLVC Yang et al (2020a), FOMM Siarohin et al (2019), and CFTE Chen et al (2022). The implementation details of the six codecs are elaborated as follows.

- **VVC.** The VVC is the latest hybrid video coding scheme which obtains significant rate-distortion gains compared with former standards. The standard VVC Test Model VTM-11.0 Bross et al (2021) is adopted to generate the VVC-compressed videos, working in the random access (RA) mode, and the group of picture (GOP) size is set to 32. By adjusting the QP value to 22, 32, 37, 42, and 47, the reference videos are compressed with five bit-rate levels.
- **Reduced Resolution Compression.** For the reduced-resolution compression method, the reference videos are first downsampled to three sizes using the bicubic linear interpolation method, while sustaining the original video aspect ratio. The downsampled videos are then compressed by the VVC method with the QP value of 32 and upsampled to the original video size.
- **DVC.** The DVC Lu et al (2019) is the first E2E video compression scheme where motion estimation, motion compression, and residual compression are jointly optimized. DVC adopts motion compensation to reduce the temporal redundancy, and the motion and residual information are compressed by two compression networks. For our dataset construction, we adopt the implementation in Yang et al (2020b), working in the PSNR-optimized mode with the GOP size of 16. By adjusting the *lambda* value, reference videos are compressed into four bit-rate levels. For the models with $lambda = 2048, 512, 256$, corresponding to $QP = 22, 32, 37$ for the I-frame respectively, the pretrained models are directly adopted. Additionally, we trained the PSNR-optimized model with $lambda = 128$ based on the Vimeo90K dataset Xue et al (2019). The parameter settings follows Yang et al (2020b).
- **RLVC.** The RLVC Yang et al (2020a) adopts the recurrent auto-encoder to capture temporal information in a long time range in the form of latent features, which are employed to reconstruct compressed frames. In our dataset, we utilize the author-released implementation Yang et al (2020a) for video compression. The RLVC works with the PSNR-optimized mode. The GOP structure is bi-IPPP and GOP

size is set to 13 (6 forward P frames and 6 backward P frames). By adjusting the *lambda* value, reference videos are compressed into 4 bit-rate levels.

- **FOMM and CFTE.** The FOMM Siarohin et al (2019) and CFTE Chen et al (2022) face video compression algorithms arise from GAN-based face image animation models. The two generative schemes follow an encoder-decoder workflow. In general, video frames are represented in a compact feature domain, such as keypoints or 3DMM parameters, and then are inter-predicted, quantized, and entropy-coded at the encoder side. Then the features are decoded to reconstruct frames at the decoder side. During the implementation of FOMM and CFTE methods, the first frame of the reference video sequence is initially compressed with VVC codec, and the consequent frames are reconstructed according to the compressed key frame and decoded motion information. The first frame is compressed with the QP levels of 22, 32, 37, and 42, which generate compressed videos in four bit-rate levels. Instead of adopting the models with the resolution of 256×256 in original papers, the adopted FOMM and CFTE codecs are newly trained in the resolution of 512×512. The models are trained for 100 epochs on NVIDIA TESLA V100 GPUs with 32GB memory capacity based on the VoxCeleb2 dataset Nagrani et al (2017) upsampled to the size of 512×512. The specific hyper-parameters for model training follow the settings in Siarohin et al (2019) and Chen et al (2022) respectively.

1.3 RD Performance

To demonstrate the rate-distortion performance of these methods in terms of the commonly-used objective quality measures PSNR Wang and Bovik (2009) and DISTS Ding et al (2020), 768 compressed video sequences generated from 32 reference videos are randomly selected to obtain the rate-distortion curves. Fig. 2 demonstrates that generative face video compression methods possess significant bit-rate savings compared with traditional and E2E methods in the low bit-rate range.

2 Distortion Type analysis for Different Compression Models

To further investigate the compression artifacts, we conduct a comprehensive study of the distortions caused by different compression models. In particular, the study is conducted in the forms of distortion type selection when the subjects are judging the quality of the compressed videos. In each trial, six video sets are formed and each set comprises two videos with different content compressed by the same video compression method. Subjects are forced to select the perceived distortion types from the given choice set. The candidate distortion types include blur, blockiness, noise, flickering artifacts, content twisting, information loss, unreality, color artifact, and edge artifacts. Note that candidate distortion types are not necessarily independent of each other, e.g., content twisting and blur can jointly or solely be the cause of unreality, while unreality can also be rooted in information loss. The information loss and content twisting are merged as geometric distortions in the main paper.

We collect subjects' votes on each video collection, and the results are shown in Fig. 3. The most commonly mentioned distortions include blur, content twisting, blockiness, edge artifact, and flickering. More specifically, the dominant, i.e., mostly voted, distortion type for each compression method diversifies. The dominant perceived distortions for VVC-compressed face videos are blur, blockiness, and color artifact. For the RL-compressed face videos, those are blur, blockiness, and edge artifacts. For the DVC- and RLVC-compressed face videos, those are flickering, unreality, and edge artifact. The dominant distortions for the FOMM- and CFTE-compressed videos include content twisting, information loss, unreality, and edge artifact.

From the results, we can observe that: (1) The HVS responding to distortion perception shows high relation to codecs. Traditional codecs, End-to-End (E2E) codecs, and generative codecs can generate diversified distortions. (2) Some distortion types signify the masking effect of other distortions in compressed face videos. Besides, blur is significantly underestimated in the E2E compressed videos, while blur usually appears in

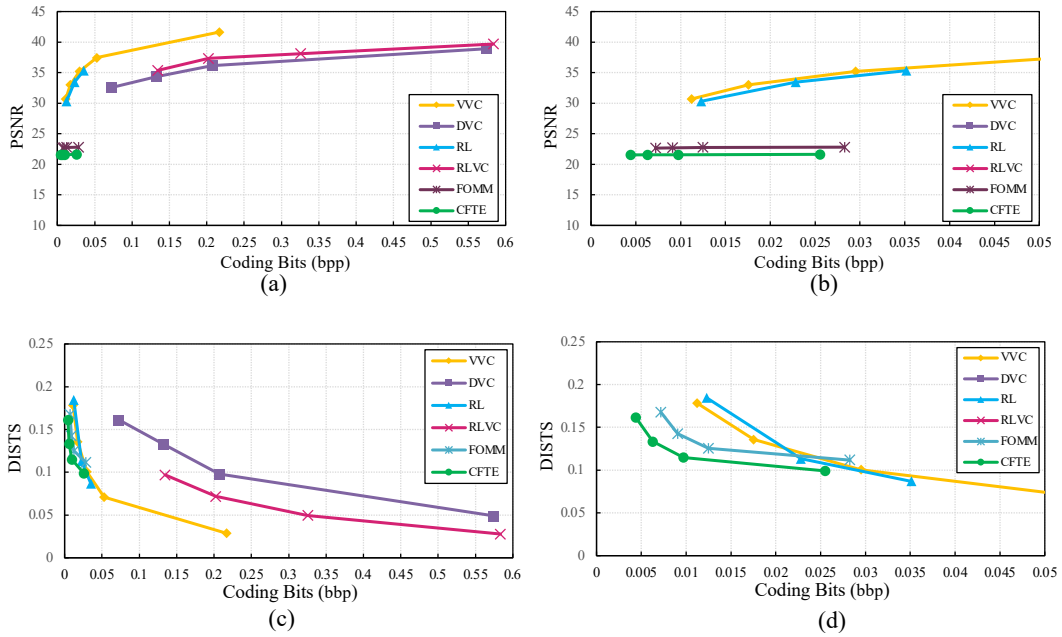


Fig. 2 RD performance comparisons of VVC, RL, DVC, RLVC, FOMM, and CFTE codecs. The corresponding QPs are as listed in the Table 1. The charts on the right side are with logarithmic horizontal axis for better comparison on lower bit-rates.

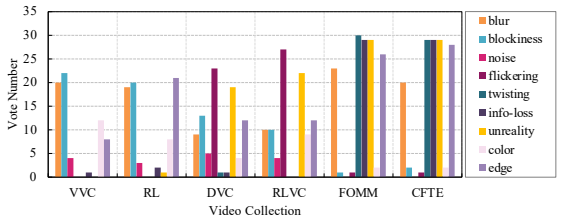


Fig. 3 Vote numbers of different distortion types for the six compression models.

such videos widely and obviously. Subjects do not reckon such blur as an artifact but a unique image style instead, resulting in the unreality artifact. This is due to the content specialty of face videos compared with natural scene videos.

3 Proposed Face VQA Framework

3.1 Feature Visualization

In Fig. 4, the feature maps of the reference frame and the corresponding compressed frames are visualized. In addition, we also show the residual maps to present the quality corruption caused by compression.

From the figure, we can observe that: (1) The quality degradation exists in different layers, revealing all layers of the feature extractor are important for the distortion measure. (2) The feature difference largely lies in the facial region, indicating that the feature extractor is highly aware of the facial content. This characteristic is also in line with human perception when they are visualizing the face video. (3) By comparing the residual maps in the same layer, we can observe the quality corruption patterns of different video codecs diversify significantly. This phenomenon reveals that it is the corruption on the low-level structure that leads to different spatial distortions.

3.2 Case Study

The proposed face video quality assessment framework is based on the human face prior and temporal memory prior. It shows superior performance compared with other VQA methods. In Fig. 5, Fig. 6, Fig. 7, Fig. 8, Fig. 9, and Fig. 10, several compressed video sequences and corresponding predicted quality scores are presented. From those samples, we can observe our VQA model shows a high consistency with human opinions.

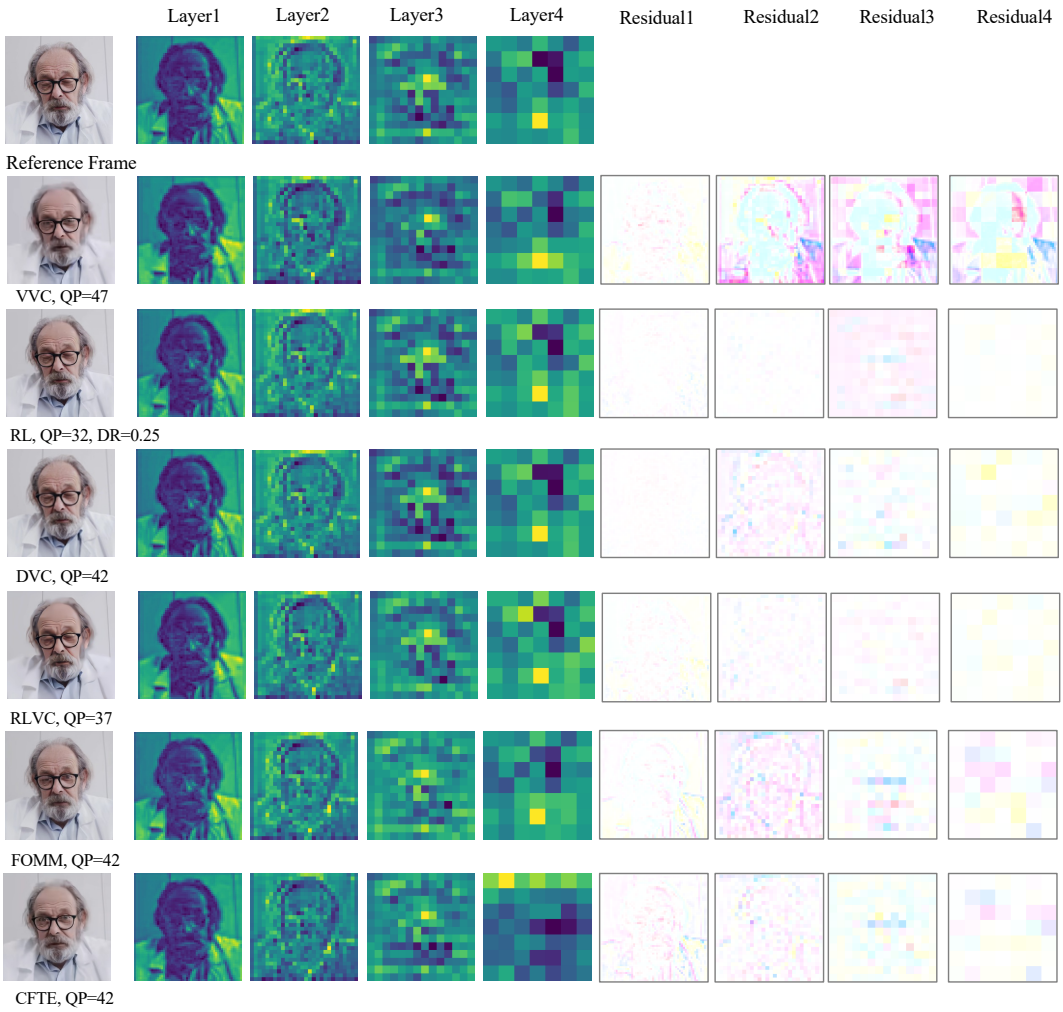


Fig. 4 Sampled feature maps of the reference frame and corresponding compressed frames. Particularly, the sixth to tenth columns are the residual maps between the features of the compressed and reference frames.

4 Subjective Rating Analyses

The raw subjective rating data need to be examined on the agreement of subjective ratings. Given one distorted video, it is generally believed that most subjects should reach agreements on the perceived quality of each video even though divergence exists. The quartiles of ratings are adopted to evaluate the subject agreement. In Fig. 11, we show the first and third quartiles of collected ratings of each compressed video. In general, a video whose interquartile range (IQR) is smaller than or equal to 1 is regarded as a video with the agreement. As listed in Table 2, we observe that 96.48% of the compressed face videos attain subjective agreement on subjective quality, implying the trustworthiness of the subjective quality in proposed database.

Table 2 Subjective agreement ratio.

IQR	Agreement		Outlier
	0	1	2
Number	547	2,579	114
Percentage	0.1688	0.7960	0.0352
	0.9648		0.0352

References

(2023) Compressed-face-videos-quality-assessment. [EB/OL], https://github.com/Yixuan423/Compressed-Face-Videos-Quality-Assessment, Accessed March 5, 2023

Bross B, Wang YK, Ye Y, et al (2021) Overview of the Versatile Video Coding (VVC) standard and its applications. IEEE Transactions on Circuits and Systems for Video Technology 31(10):3736–3764

Chen B, Wang Z, Li B, et al (2022) Beyond keypoint coding: Temporal evolution inference with compact feature representation for talking face video compression. In: Data Compression Conference, IEEE, pp 13–22

Ding K, Ma K, Wang S, et al (2020) Image quality assessment: Unifying structure and texture similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(5):2567–2581

Lu G, Ouyang W, Xu D, et al (2019) DVC: An end-to-end deep video compression framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11006–11015

Nagrani A, Chung JS, Zisserman A (2017) Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:170608612

Siarohin A, Lathuilière S, Tulyakov S, et al (2019) First order motion model for image animation. Advances in Neural Information Processing Systems 32

Wang Z, Bovik AC (2009) Mean squared error: Love it or leave it? a new look at signal fidelity measures. IEEE Signal Processing Magazine 26(1):98–117

Xue T, Chen B, Wu J, et al (2019) Video enhancement with task-oriented flow. International Journal of Computer Vision (IJCV) 127(8):1106–1125

Yang R, Mentzer F, Van Gool L, et al (2020a) Learning for video compression with recurrent auto-encoder and recurrent probability model. IEEE Journal of Selected Topics in Signal Processing 15(2):388–401

Yang R, Van Gool L, Timofte R (2020b) OpenDVC: An open source implementation of the DVC video compression method. arXiv preprint arXiv:200615862

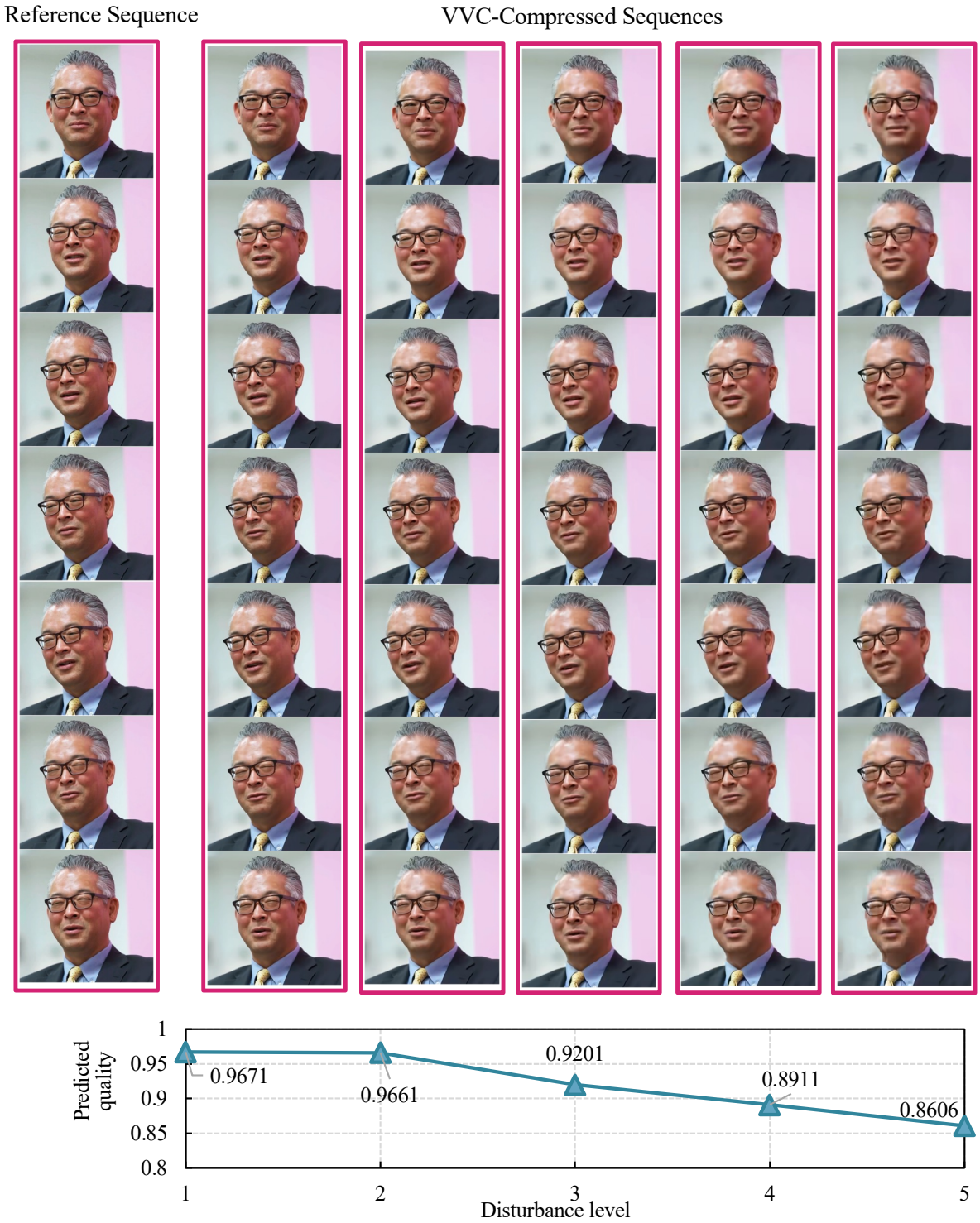


Fig. 5 Quality comparisons of VVCBross et al (2021) compressed face videos. For each row, from the left to right are the reference frame and the distorted frames generated in five quality disturbance levels (low to high). A higher MOS value indicates a lower quality disturbance level. The chart below shows the quality score predicted by our VQA model on the compressed face videos in different quality disturbance levels. The artifacts are more apparent when viewing the video samples that can be found on the project pagesup (2023).

Fig. 6 Quality comparisons of RL compressed face videos. For each row, from the left to right are the reference frame and the distorted frames generated in three quality disturbance levels (low to high). A higher MOS value indicates a lower quality disturbance level. The chart below shows the quality score predicted by our VQA model on the compressed face videos in different quality disturbance levels. The artifacts are more apparent when viewing the video samples that can be found on the project pagesup (2023).

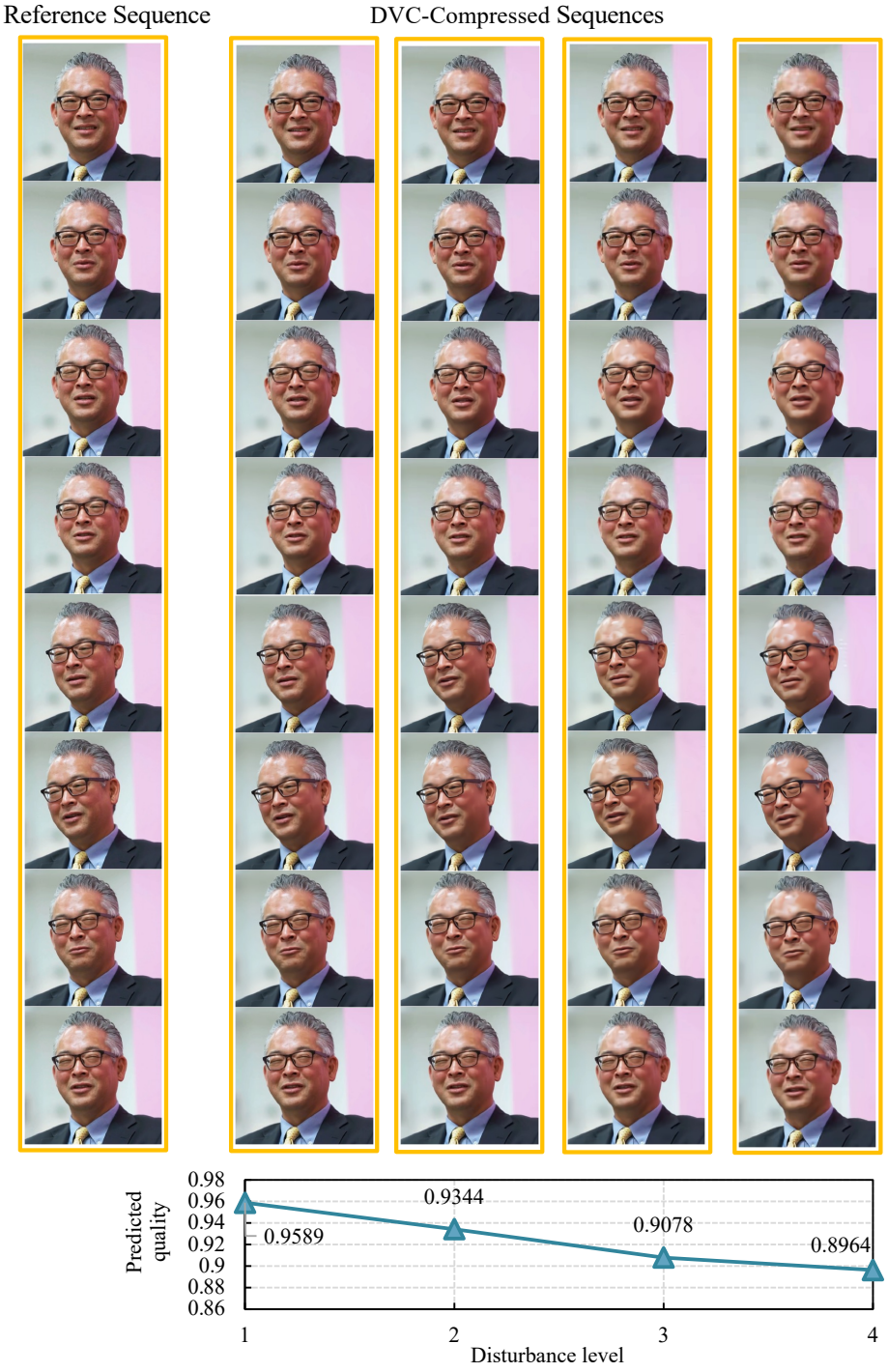


Fig. 7 Quality comparisons of DVCLu et al (2019) compressed face videos. For each row, from the left to right are the reference frame and the distorted frames generated in four quality disturbance levels (low to high). A higher MOS value indicates a lower quality disturbance level. The chart below shows the quality score predicted by our VQA model on the compressed face videos in different quality disturbance levels. The artifacts are more apparent when viewing the video samples that can be found on the project pagesup (2023).

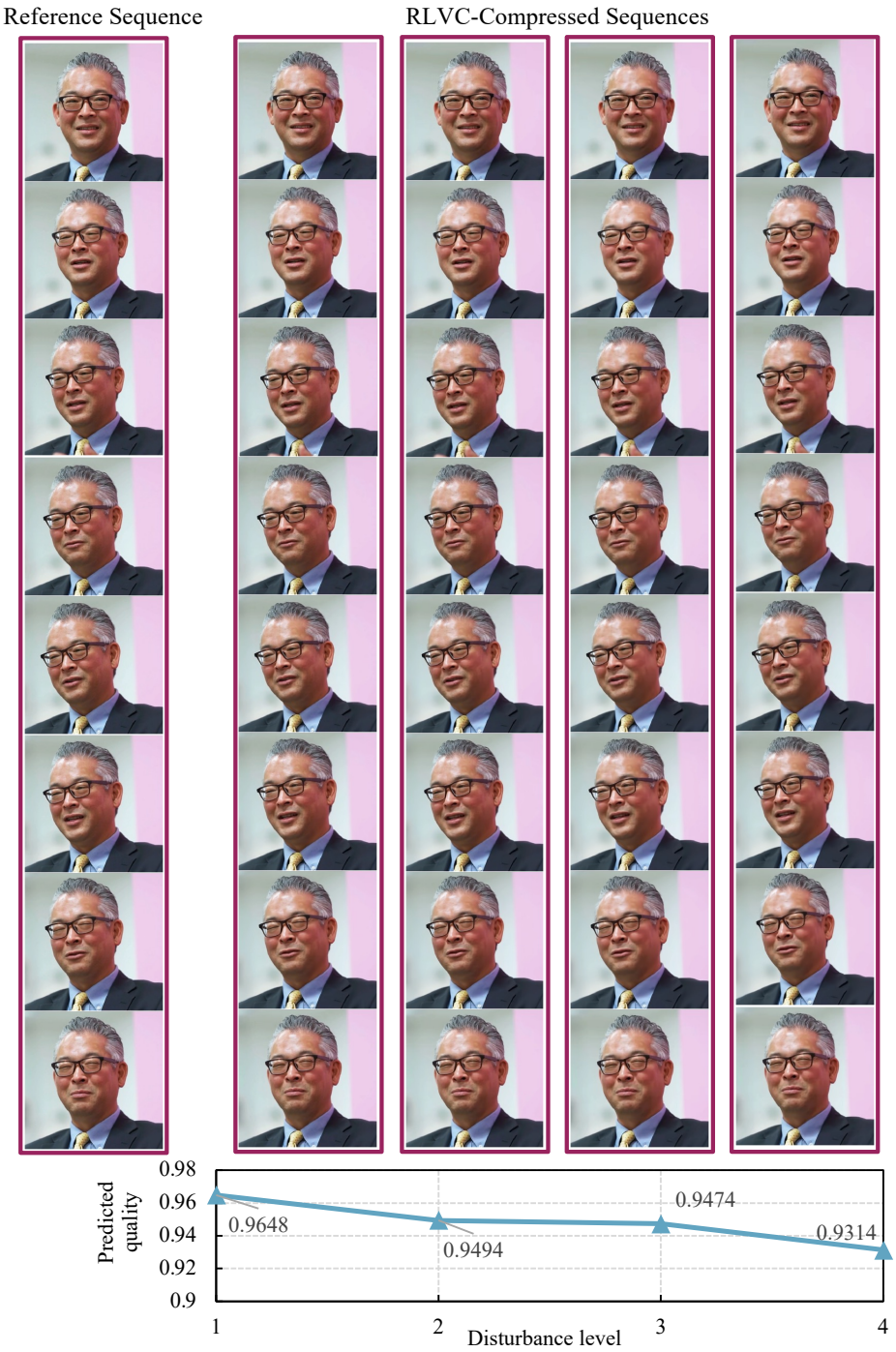


Fig. 8 Quality comparisons of RLVCYang et al (2020a) compressed face videos. For each row, from the left to right are the reference frame and the distorted frames generated in four quality disturbance levels (low to high). A higher MOS value indicates a lower quality disturbance level. The chart below shows the quality score predicted by our VQA model on the compressed face videos in different quality disturbance levels. The artifacts are more apparent when viewing the video samples that can be found on the project pagesup (2023).

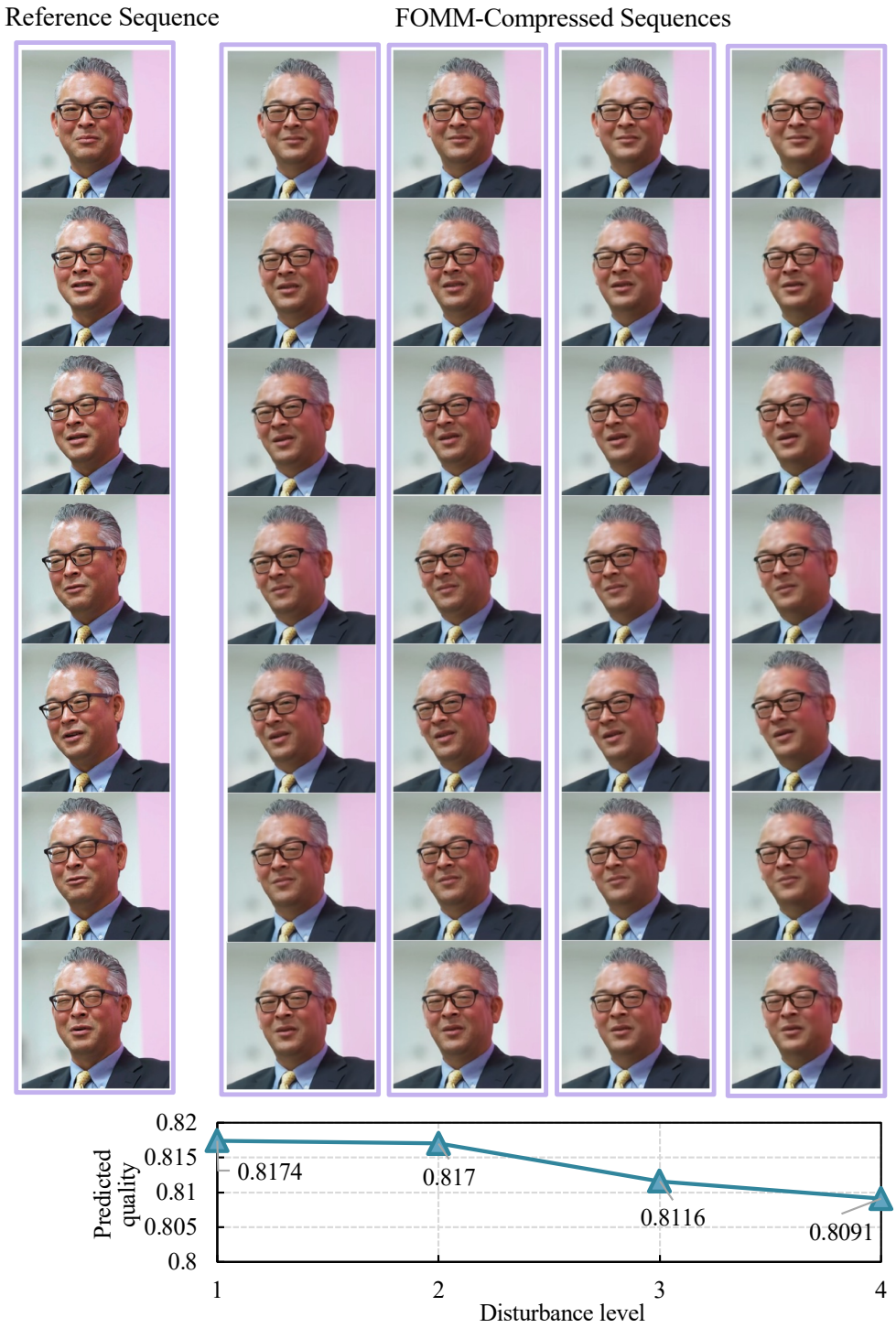


Fig. 9 Quality comparisons of FOMMSiarohin et al (2019) compressed face videos. For each row, from the left to right are the reference frame and the distorted frames generated in five quality disturbance levels (low to high). A higher MOS value indicates a lower quality disturbance level. The chart below shows the quality score predicted by our VQA model on the compressed face videos in different quality disturbance levels. The Video examples can be found on the project pagesup (2023).

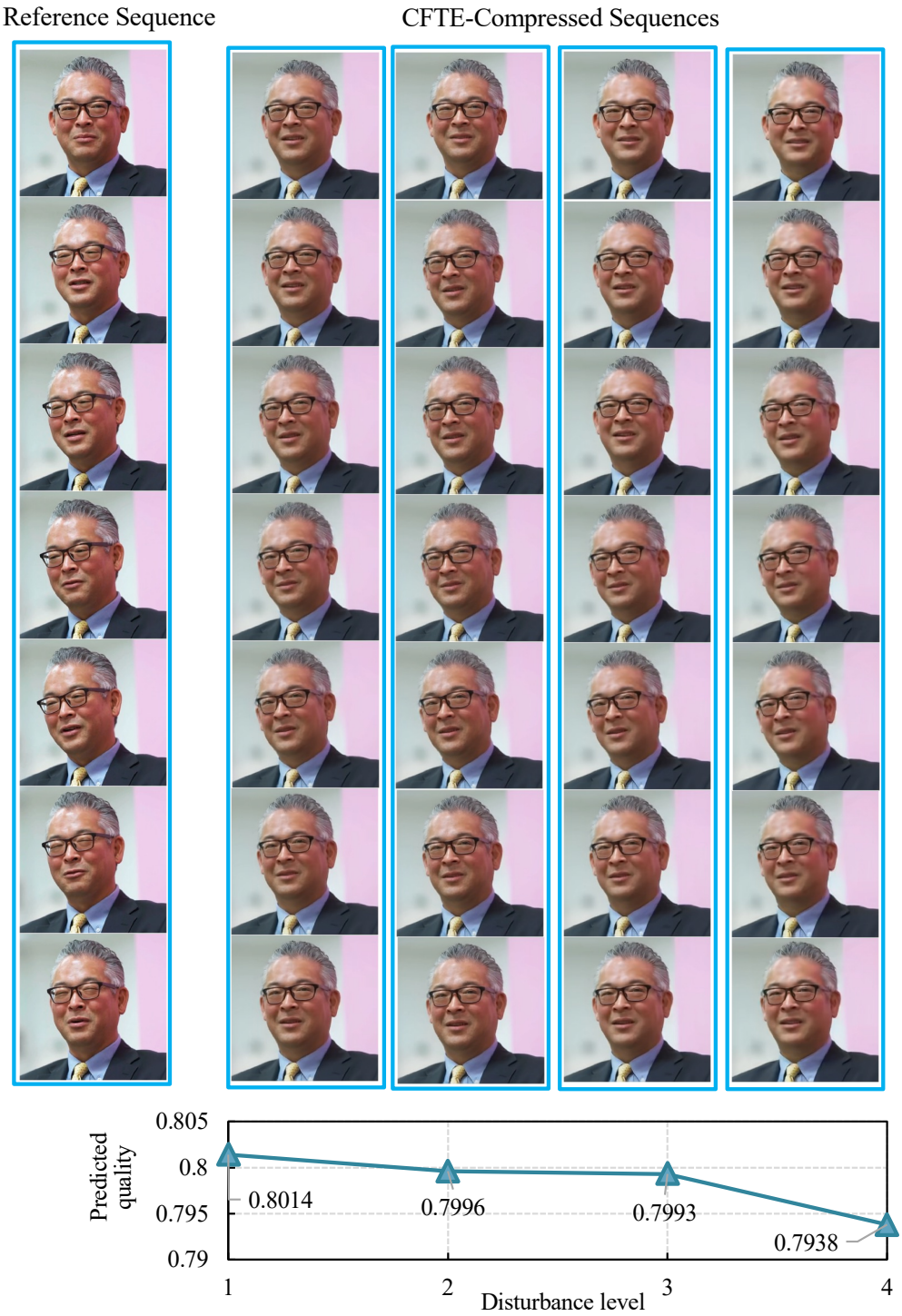
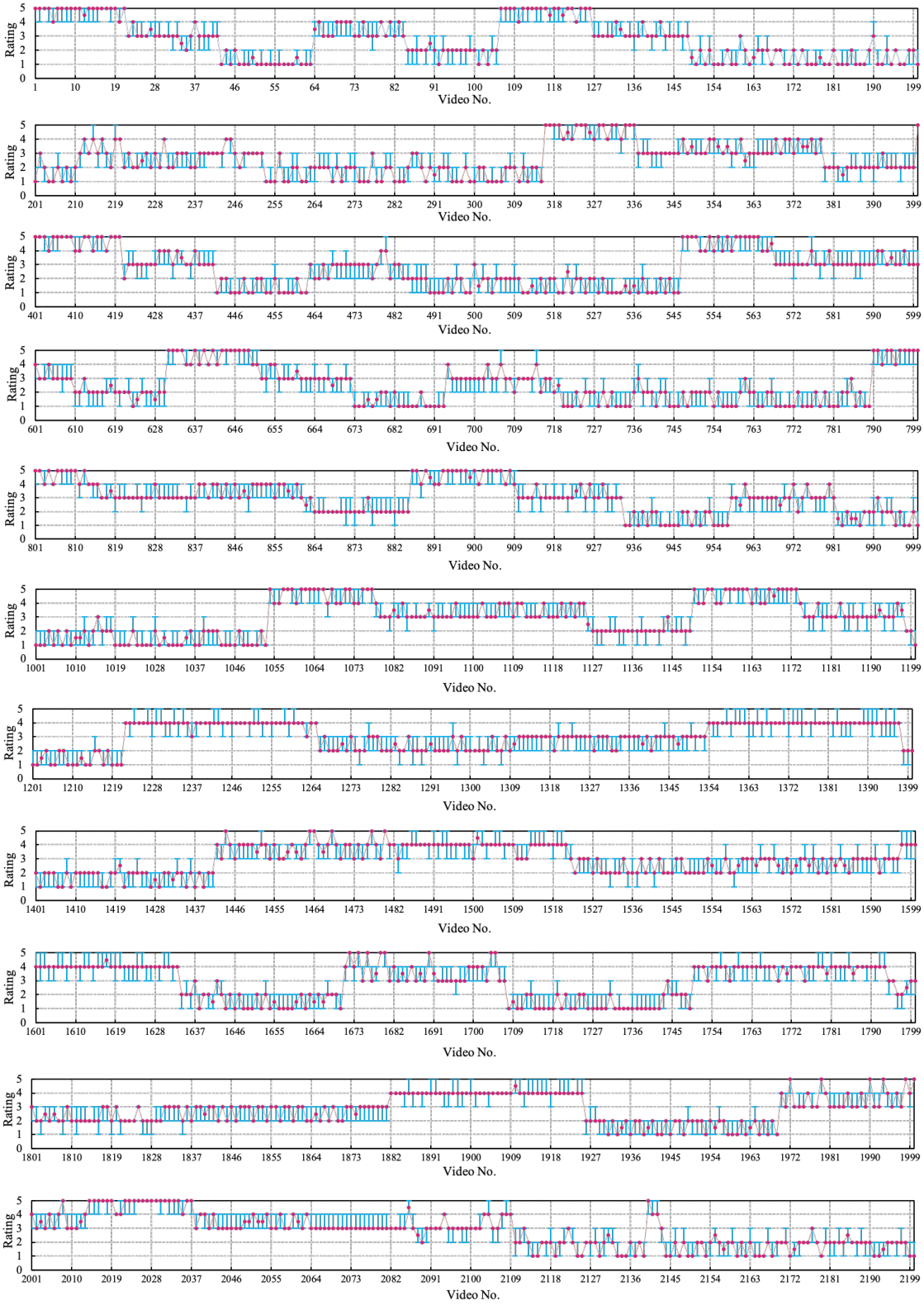


Fig. 10 Quality comparisons of CFTEChen et al (2022) compressed face videos. For each row, from the left to right are the reference frame and the distorted frames generated in four quality disturbance levels (low to high). A higher MOS value indicates a lower quality disturbance level. The chart below shows the quality score predicted by our VQA model on the compressed face videos in different quality disturbance levels. The artifacts are more apparent when viewing the video samples that can be found on the project pagesup (2023).


Rating
Video No.
0 1 2 3 4 5
1 10 19 28 37 46 55 64 73 82 91 100 109 118 127 136 145 154 163 172 181 190 199
201 210 219 228 237 246 255 264 273 282 291 300 309 318 327 336 345 354 363 372 381 390 399
401 410 419 428 437 446 455 464 473 482 491 500 509 518 527 536 545 554 563 572 581 590 599
601 610 619 628 637 646 655 664 673 682 691 700 709 718 727 736 745 754 763 772 781 790 799
801 810 819 828 837 846 855 864 873 882 891 900 909 918 927 936 945 954 963 972 981 990 999
1001 1010 1019 1028 1037 1046 1055 1064 1073 1082 1091 1100 1109 1118 1127 1136 1145 1154 1163 1172 1181 1190 1199
1201 1210 1219 1228 1237 1246 1255 1264 1273 1282 1291 1300 1309 1318 1327 1336 1345 1354 1363 1372 1381 1390 1399
1401 1410 1419 1428 1437 1446 1455 1464 1473 1482 1491 1500 1509 1518 1527 1536 1545 1554 1563 1572 1581 1590 1599
1601 1610 1619 1628 1637 1646 1655 1664 1673 1682 1691 1700 1709 1718 1727 1736 1745 1754 1763 1772 1781 1790 1799
1801 1810 1819 1828 1837 1846 1855 1864 1873 1882 1891 1900 1909 1918 1927 1936 1945 1954 1963 1972 1981 1990 1999
2001 2010 2019 2028 2037 2046 2055 2064 2073 2082 2091 2100 2109 2118 2127 2136 2145 2154 2163 2172 2181 2190 2199

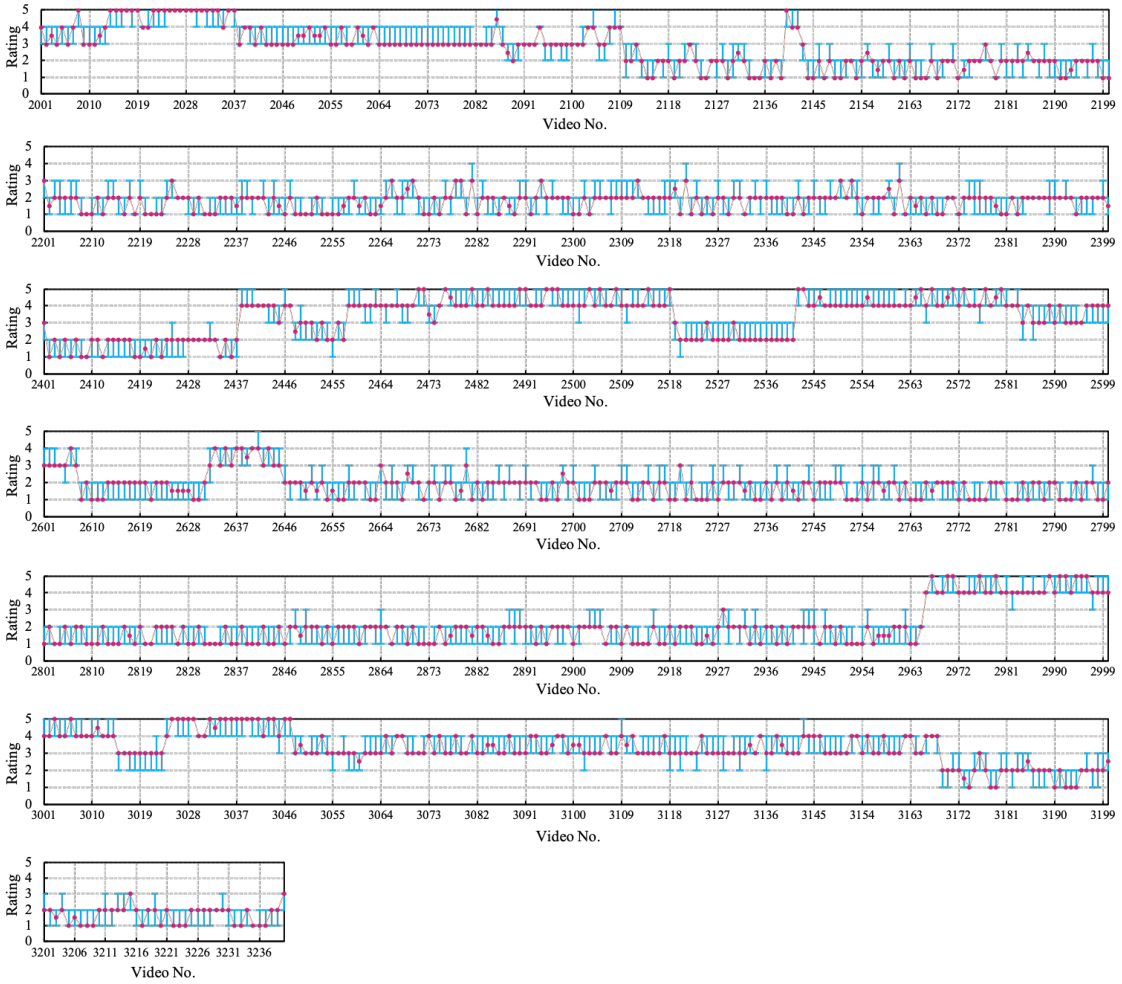


Fig. 11 The subjective ratings and MOS values for each compressed video. Each chart contains the subjective rating scores for 200 videos. The red dot denotes the median value among all subjects. The blue error bar denotes the first (25%) and third (75%) quartiles of subjective ratings.