Barriers Towards No-reference Metrics Application to Compressed Video Quality Analysis: on the Example of No-reference Metric NIQE

A. Zvezdakova[1], D. Kulikov[1,2], D. Kondranin[3], D. Vatolin[4]
azvezdakova@graphics.cs.msu.ru|dkulikov@graphics.cs.msu.ru|denis.kondranin@graphics.cs.msu.ru|
dmitriy@graphics.cs.msu.ru
[1]Lomonosov Moscow State University, Moscow, Russia;
[2]Dubna State University, Dubna, Russia

This paper analyses the application of no-reference metric NIQE to the task of video-codec comparison. A number of issues in the metric behavior on videos was detected and described. The metric has outlying scores on black and solid-colored frames. The proposed averaging technique for metric quality scores helped to improve the results in some cases. Also, NIQE has low-quality scores for videos with detailed textures and higher scores for videos of lower bit rates due to the blurring of these textures after compression. Although NIQE showed natural results for many tested videos, it is not universal and currently can't be used for video-codec comparisons.

Keywords: *video quality, no-reference metric, quality measuring, video-codec comparison.*

1. Introduction

Today video content takes the biggest part of world Internet traffic (more than 70%). According to the forecasts [1], its rate will grow up to 82% in 2022. This trend leads to the creation of new encoding standards and improvements in existing encoders. There is a number of video-codec comparisons which are conducted to find the best codecs for different tasks and use cases and to help users and customers to find appropriate encoders for their needs. The tar-get for video encoding is to deliver high visual qual-ity with reduced file size, so the only reliable way to compare encoded videos quality is to perform a subjective evaluation. It requires a proofed methodology and a high number of observers to achieve reasonable results. In general, subjective comparisons are still very expensive to perform, however, there are some services which help researchers to perform qualitative subjective comparison [2]. This obstacle increases the importance of objective metrics for video quality comparison.

Objective quality metrics can be divided into three general categories: full-reference metrics, no-reference metrics and reduced-reference metrics. Full-reference metrics are easy to interpret and useful in application to video compression quality estimation. Unlike full-reference metrics which require source video to compare with compressed, no-reference metrics are useful when you don't have a source and want to estimate the quality of the compressed video. This case is usual for example for cloud encoding when videos are uploaded compressed by a built-in encoder in smartphones or non-professional cameras. Reduced-reference metrics require just some part of information about source video and can also be used in some of the listed cases.

2. Related work

There is a number of no-reference metrics which were created using databases with subjective quality scores. Such quality assessment models were trained to estimate subjective quality, and so their scores depend on training and testing sets. For example, DIIVINE (2011) [10], LBIQ (2011) [12], BRISQUE (2012) [9] and V-Bliinds (2012) [11] were trained on LIVE data set. In 2015, a metric called IL-NIQE [15] was proposed. It was based on NIQE [8] metric, which is studied in this paper, but used multivariate Gaus-sian (MVG) model to predict the quality of image patches instead of using a single global MVG model for an image.

Another group contains metrics which weren't trained on any data sets and use only data from a source image to estimate its quality. For example, CORNIA (2012) [14] combined feature and regression training. Recently several approaches which use neural networks architectures have been developed. The authors of COME (2018) [13] proposed an approach based on convolution neural network AlexNet and multi-regression which outperformed V-Bliinds on a number of video sets.

No-reference metrics are created to approximate users perception of video quality, but in case of estimating the quality of encoding and compression, they can be used only as an addition to reference metrics. No-reference metrics can't become the main criteria for encoders comparison because in the opposite way encoders could win the comparison producing a visually ideal result which has little common with the input video. The authors of this paper organize worldwide video-codec comparisons for 16 years. Currently, full-reference metric SSIM is used in these comparisons as the main metric supplemented with a number of additional metrics (PSNR, VMAF). At the same time, several researchers and industry experts con-sider measuring and taking into account no-reference metrics in video-codec comparisons. This paper de-scribes the authors' experience of using no-reference metric NIQE (Natural Image Quality Evaluator) [8] created by Anish Mittal, Rajiv Soundararajan and Alan C. Bovik in video-codec comparison. This met-ric is one of the most popular nowadays and shows good results for image quality assessment.

We used NIQE to access the quality of encoded video sequences during the video-codec comparison.

The main idea of NIQE metric is based on construct-ing a collection of quality-aware features and fitting them to a multivariate Gaussian (MVG) mode. NIQE score represents the degree of distortions in the frame, and the lower score is, the higher quality is the frame. Accordingly, rate-distortion graphs for encoded videos look unusually inverted, so on the plots in this paper NIQE scores are presented inverted to make the results more familiar and interpreting.

There is an open implementation on MATLAB provided by the authors [5]. In order to increase computational speed, we used an implementation from MSU Video Quality Measurement Tool (VQMT) which is currently faster. The tool has a free version (it includes NIQE) and can be downloaded [6]. Speed was important in this case because the metric was used for video quality assessment.

3. Experimental setup

For the evaluation, 28 different FullHD video sequences were used with number of frames per second from 24 to 60 and which were generated by real users. The videos were chosen from MSU video collection which consists of 15,833 videos. The collection was divided into 28 clusters by spatiotemporal complex-ity [7] and one sequence from each cluster, which was close to the cluster center, was chosen for the final testing set. Each video was encoded by x264 and x265 encoders. There were three encoding use cases (“fast”, “universal” and “ripping”) based on different encod-ing speed/quality ratios and 7 different bit rates from 1 Mbps to 12 Mbps. An overall number of encoded streams which were evaluated by NIQE is 1176.

The final video set was used in 2018 Moscow State University (MSU) video-codec comparison [3]. The comparison results are available on the link, but the results of NIQE were not published on-line because of several issues found in NIQE application to video quality measurement. Some of them were noted it the original article, the others were resolved with our proposed averaging technique which will be described in the article. Unfortunately, some issues can’t be fixed without the metric improvement (completing the training set or other fixes). In this article, we suggest the method of metric results processing to solve the detected problems on metric application to videos.

4. Metric behavior on videos

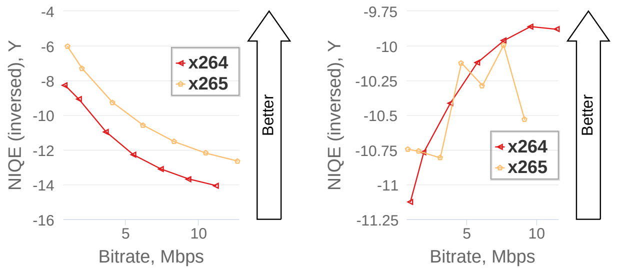
For most of the encoded videos, NIQE showed the results which reflected the usual perceptual video quality on different bit rates. But there were some cases in which NIQE showed the results with some issues; the following sections describe the detected issues and their reasons.

4.1 Cases with relevant results

According to the authors, NIQE is not applica-ble to unnatural distortions in scenes and scenes from unnatural source (e.g. computer graphics), as such scenes were not used during the training. However, we checked metric scores on cartoons from our video set.

At *Sita* (part from the cartoon movie), rate-distortion curve looks inverted (Fig. 1a), NIQE shows worse quality scores for high bit rates that for low bit rates. This means that the metric is really not applicable to this type of content. At *Sintel* (part from CGI movie trailer), NIQE showed non-monotonic scores for x265 encoder on fast use case bit rate map, but acceptable results for universal and ripping use cases (Fig. 1b). Thus, the metric is said to be not applica-ble to cartoons, but we revealed that it works for some types of realistic animation, such as for video gaming (sequences *Witcher3*, *Rust*).

There were some examples, where the rate-distortion curve looked unnatural, but the metric correctly ranked worse visual quality to higher bit rates. For example, on *Hera* video sequence (a part of a mu-sic clip with grain effects) NIQE showed worse score for x264 encoding on 4000 kbps than on 2000 kbps in fast use case (Fig. 2). The metric had better scores for almost all frames of the lower bit rate. It is shown in the example frame on Fig. 3, where x264 encoding of the video on 4000 kbps produced worse visual quality and more compression artifacts than on 2000 kbps.



(a) *Sita* video sequence

(b) *Sintel* video sequence

Fig. 1. Rate-distortion graph for animation.

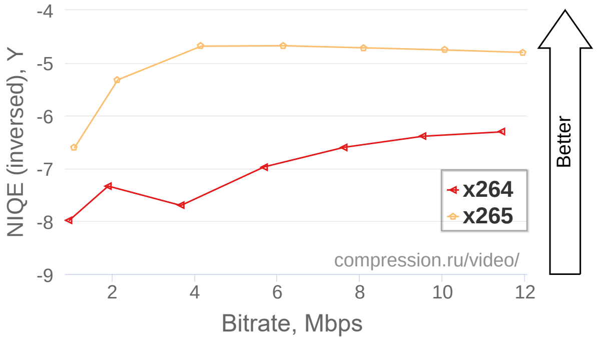


Fig. 2. Rate-distorion graph for *Hera*

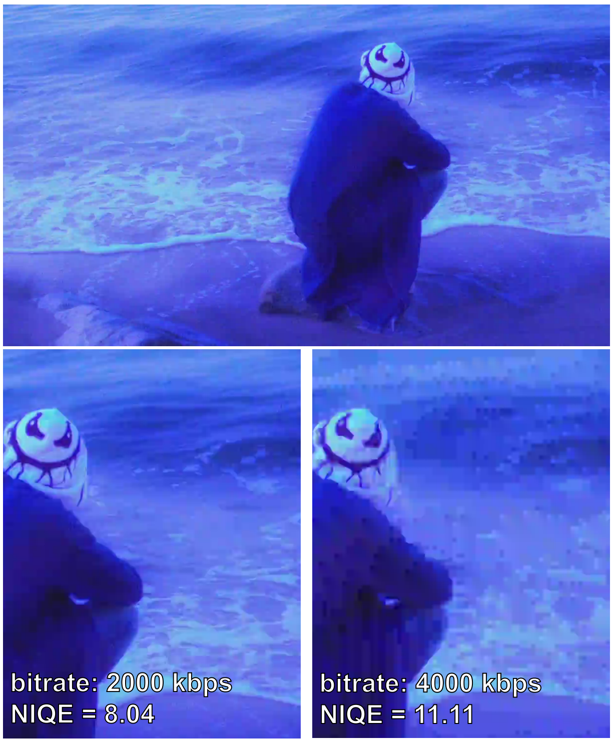


Fig. 3. Frame 208 from *Hera* video sequence, codec: x264, fast use case. According to NIQE, left image is visually better.

4.2 Cases with irrelevant results

4.2.1 Dark scenes

The metric was said to be not applicable to the cartoons, but some other types of video content also had inaccurate NIQE scores. One of the most frequent cases in video sequences with completely black frames (for example, in the beginning). These frames, according to NIQE, are perceptually worse than the other frames and has an extremely high metric score. This might happen because of the absence of such kind of content in training data used for NIQE creation.

For example, for x264 encoding NIQE showed worse score on 2000 kbps than on 1000 kbps at *Fire* video sequence (Fig. 4). It contains close shooting of a fire in a dark. In this sequence, the metric showed better scores on a group of frames where the camera started a slow movement.

Another example which demonstrates this issue is presented in Fig. 5. *Music Clip* video sequence was quite complicated for many encoders in MSU comparison. It consists of short scenes which quickly switch and a lot of special effects, such as red sparkles and grain. NIQE shows unnatural results on this sequence for all use cases: the rate-distortion curve is not mono-tonic because of an anomaly big values on dark frames.

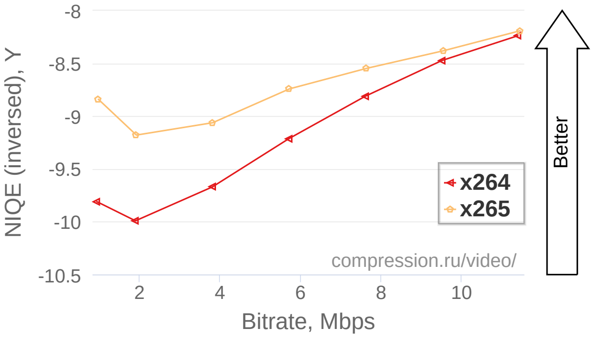
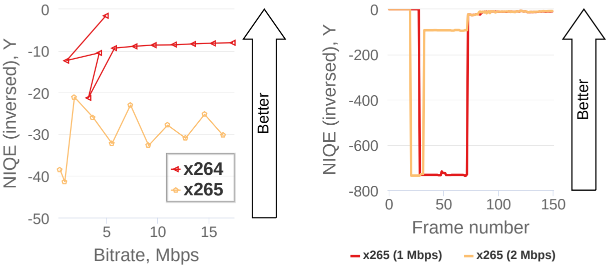


Fig. 4. Rate-distortion graph for *Fire*.



(a) Rate-distortion graph **(b)** Per-frame NIQE scores

Fig. 5. *Music clip* video sequence.

The videos described above contained completely black or dark frames. In these videos, NIQE had large values mostly on these frames, which was the main reason for the wrong overall quality score for the en-tire video. The following examples demonstrate an-other case in which NIQE was not applicable to video quality estimation.

4.2.2 Noisy scenes/scenes with lots of details

A number of cases where the metric took wrong values appear in videos with noise or a lot of small and textured details, like sand, water waves and grass. For x265-encoded *Bay time-lapse* sequence, NIQE showed worse score on 2000 kbps than on 1000 kbps in universal use case (Fig. 6). This video contained a scene with water and grass, and the grass and waves on the water are smoother in a lower-bit rate video stream.

In another example, NIQE showed worse score on 4000 kbps than on 2000 kbps in ripping use case on *Playground* video sequence for both encoders. This video contains a lot of bright frames with highly structural and detailed grass and sand. Such texture is quite complicated for compression, and on low bit rates, there were visible compression artifacts, but NIQE had a worse score on high bit rates (Fig. 7). This happened due to NIQE perception of finely textured grass as noise, while blurred compressed grass was expected to be visually better by NIQE. This is why the rate-distortion curve looks inverted on bit rates higher than 2000 kbps.

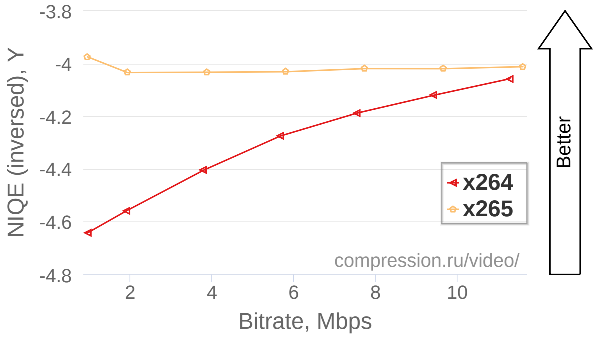


Fig. 6. Rate-distortion graph for *Bay time lapse*.

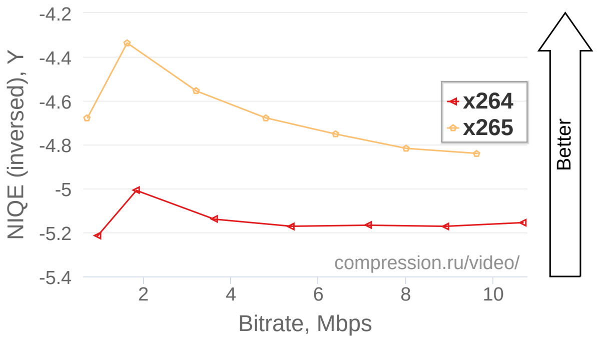


Fig. 7. Rate-distortion graph for *Playground*.

4.3 Proposed processing technique

During the analysis of per-frame NIQE results, it was revealed, that values greater than 40 don't usually appear in most of the video frames. Extreme values often occur in solid-colored or dark frames. We pro-posed and applied a special averaging technique to eliminate these cases. Our NIQE score for the video V was computed in the following way:

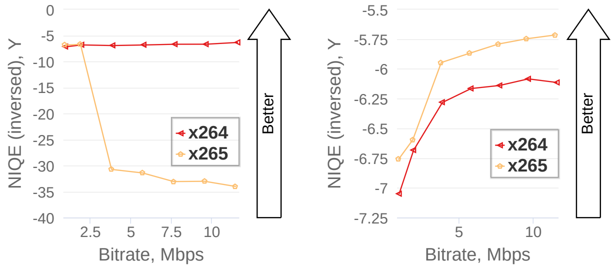
$$Score_V = \frac{\sum_i m_i * k_i}{\sum_i k_i}, i \in [0, N],$$

$$k_i = \begin{cases} 1, m_i \in [0, 15), \\ -0.04 * m_i + 1.6, m_i \in [15, 40), \\ 0, m_i \in [40, +\infty), where \end{cases} \qquad (1)$$

m_i – NIQE score for frame i,
k_i – weighting coefficient for m_i score,
N – number of frames.

The proposed averaging formula helped to improve NIQE scores for some of the video sequences. The following results demonstrate the corrected rate-distortion curves, which can be compared to the original results presented above.

With the proposed averaging technique rate-distortion curve for *Forest dog* doesn't contain outlying points (Fig. 8b). Another example, where the results were corrected by the proposed averag-ing for both encoders, is *Music clip* video sequence (Fig. 9). The non-monotonic curve of x264 encoding was caused by high spatial complexity of this video.



(a) Original rate-distortion graph.

(b) Rate-distortion graph after smart averaging.

Fig. 8. *Forest dog* video sequence.

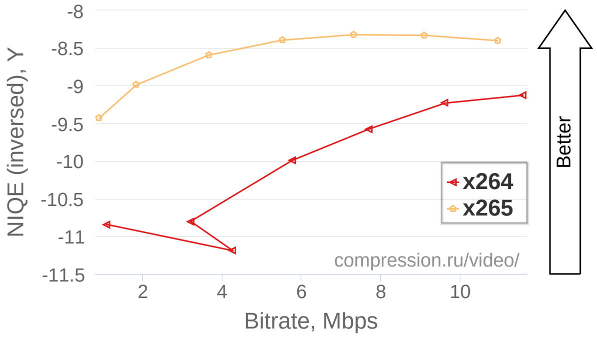


Fig. 9. Rate-distortion graph for *Music clip* after smart averaging.

5. Correlation with subjective scores

The obtained NIQE quality scores were compared to the subjective scores on part of test videos. A pairwise subjective comparison was conducted as one of the parts of 2018 MSU Video-Codec Compari-son, where a total of 22542 valid answers were re-ceived from 473 subjects. The detailed description and methodology can be found in the report [4]. Five videos were used in this comparison, and none of them contained animated scenes or black frames for which NIQE could show inaccurate results. In addition, several full-reference quality metrics were measured (SSIM, PSNR, VMAF and their variations). The Pearson correlation coefficient was calculated for the results on each video separately (Fig. 10). The averaged correlation scores across all videos reveal that NIQE has the lowest correlation with subjective scores (0.85) while VMAF v.0.6.1 for phones has the highest correlation (0.99). It should also be noted that at the moment NIQE has even lower correlation to subjective quality than PSNR (0.98), which is long considered to have low similarity to subjective quality for compression algorithms comparison.

The lowest correlation of NIQE with subjective scores was obtained for *Playground* video sequence. As it was described above for this video sequence, NIQE showed worse scores for detailed textures (grass and sand) in this video sequence, which is illustrated in Fig. 11.

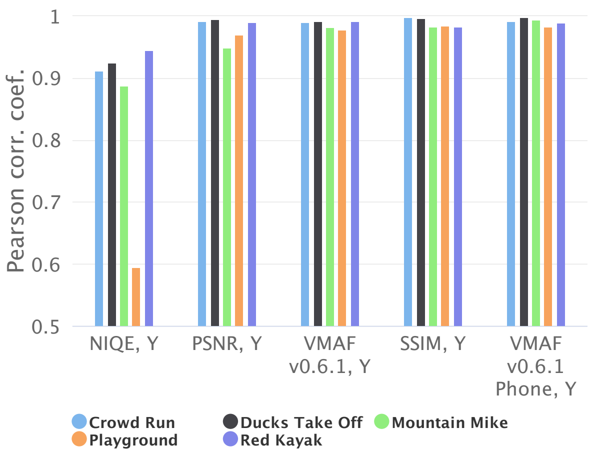


Fig. 10. Correlation between objective quality metrics and subjective scores.

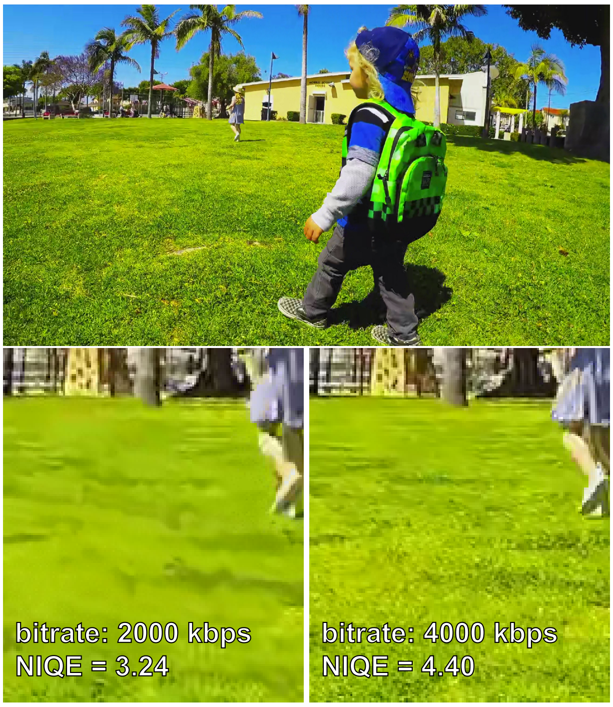


Fig. 11. Frame 58 from *Playground* video sequence, codec: x265, ripping use case. According to NIQE, left image is visually better.

6. Conclusion

During the experiments, NIQE showed good results for most of the videos. But still, there are many cases for which the metric is not applicable. This is why NIQE is not universal and can not be used in video-codec comparisons at the moment. The results of this comparison show NIQE deficiencies that need to be corrected, such as an application to animated cartoons, videos with completely black and solid-colored frames, noise and highly detailed/textured frames. For example, the abundance of fine details (grass, sand, grain effects) increases the values of NIQE despite the high bit rate of the encoded video, which leads to incorrect results. At the same time, in the original paper, NIQE was said to be not appli-cable to computer graphics, but in our investigation, it was found that the metric works for some types of animation (particularly for a screen capture of video gaming).

7. Acknowledgments

Special thanks to Georgiy Osipov who helped to analyze all detected issues and improved NIQE implementation in MSU VQMT. This work was partially supported by the Russian Foundation for Basic Research under Grant 19-01-00785a.

8. References

[1] Cisco Report VNI 2017-2022, 2018 update https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html

[2] Crowd-sourced subjective quality evaluation platform subjectify.us

[3] HEVC Video Codec Comparison 2018 (Thirteen MSU Video Codec Comparison) http://compression.ru/video/codec_comparison/hevc_2018/

[4] HEVC Video Codec Comparison 2018 (Thirteen MSU Video Codec Comparison), Part II: FullHD Content, Subjective Evaluation http://compression.ru/video/codec_comparison/hevc_2018/#subjective_report

[5] MathWorks Documentation: Naturalness Image Quality Evaluator (NIQE) no-reference image quality score https://www.mathworks.com/help/images/ref/niqe.html

[6] MSU Quality Measurement Tool: Download Page http://compression.ru/video/quality_measure/vqmt_download.html

[7] C. Chen, S. Inguva, A. Rankin, and A. Kokaram, “A subjective study for the design of multi-resolution ABR video streams with the VP9 codec,” in *Electronic Imaging*, 2016(2), pp. 1-5.

[8] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a «completely blind» image quality analyzer,” in *IEEE Signal Processing Letters*, 2012, 20(3) pp. 209-212.

[9] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” in *IEEE Transactions on Image Processing*, 2012, 21(12), pp. 4695-4708.

[10] A. K. Moorthy and A. C. Bovik, “Blind image quality assessment: From natural scene statistics to perceptual quality,” in *IEEE Transactions on Image Processing*, 2011, 20(12), pp. 3350–3364.

[11] M. Saad, A. C. Bovik, and C. Charrier, “Blind image quality assessment: A natural scene statis-tics approach in the DCT domain,” in *IEEE Transactions on Image Processing*, 2012, 21(8), pp. 3339–3352.

[12] H. Tang, N. Joshi, and A. Kapoor, “Learning a blind measure of perceptual image quality,” in *IEEE CVPR*, 2011, pp. 305-312.

[13] C. Wang, S. Li, and W. Zhang, “COME for No-Reference Video Quality Assessment,” in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2018.

[14] P. Ye, J. Kumar, L. Kang, and D. Doermann, “Unsupervised feature learning framework for no-reference image quality assessment,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 1098–1105.

[15] L. Zhang, L. Zhang, and A. C. Bovik, “A feature-enriched completely blind image quality evaluator,” in *IEEE Transactions on Image Processing*, 2015, 24(8), pp. 2579-2591.