Comparison of metrics for predicting image and video quality at varying viewing distances

Dounia Hammou, *University of Cambridge, UK*, dh706@cl.cam.ac.uk
Lukáš Krasula, *Netflix Inc., USA*, lkrasula@netflix.com
Christos G. Bampis, *Netflix Inc., USA*, christosb@netflix.com
Zhi Li, *Netflix Inc., USA*, zli@netflix.com
Rafał K. Mantiuk, *University of Cambridge, UK*, rafal.mantiuk@cl.cam.ac.uk

***Abstract*—Viewing distance and display resolution have arguably a significant impact on perceived image quality; images seen on a mobile phone with high pixel density reveal fewer distortions than the same images seen on a large TV from a close distance. However, only a few image and video quality metrics account for the effect of viewing distance and resolution. Those that do, typically rely on contrast sensitivity functions (CSFs) of the visual system. Other metrics can be potentially adapted to different viewing distances by rescaling input images. In this paper, we investigate the performance of such adapted metrics together with those that natively account for viewing distance. The results for three testing datasets indicate that there is no evidence that the metrics based on the CSF outperform those that rely on rescaled images. Moreover, we found that both methods are not successful to account for the changes in quality introduced by the change in viewing distance. We conclude that accounting for viewing distances requires better models.**

***Index Terms*—Image/Video Quality Assessment, Viewing Distance, Effective Resolution, Contrast Sensitivity function (CSF)**

I. Introduction

When assessing perceptual image and video quality, it is desirable to account for the resolution, physical dimensions of the display and the viewing distance. Those can differ substantially between display devices. For example, a 6.3" smartphone display with the resolution 1080×2400 px and seen from 40 cm has an effective resolution (pixel density) of 116 pixels per visual degree (ppd). But a 47" FullHD TV set seen from the recommended distance of three display heights has about half of that resolution, at 57 ppd. VR headsets feature much lower resolution between 17 and 35 ppd. When a high-frequency image distortion is introduced, its visibility will differ between these devices [1], [2]. Yet, few metrics account for the effect of display resolution and viewing distance by design. Nevertheless, it is possible to adapt any metric to this scenario simply by rescaling the input images to maintain the effective resolution in pixels per degree.

In this work, we compare metrics that natively account for the viewing distance with those that do not but are instead provided with resized images. The comparison is performed on three publicly-available datasets, which provide subjective ratings collected at multiple viewing distances. The fourth dataset is used for training to determine what is the best effective resolution for each quality metric. The main contribution of this work is a benchmark study on the performance of the adapted quality metrics and the metrics that natively account for the viewing distance.

II. Related work

A. The effect of viewing distance on perceptual quality

Previous works have explored the effect of viewing distance on the assessment of quality degradation due to distortions such as compression, upscaling, and others. When measuring the quality degradation due to image distortions (DMOS), a larger viewing distance reduces the effect (the visibility) of distortions [3], [4]. However, when the quality of the reference image/video is measured alone, the overall quality (MOS) decreases with larger viewing distances [1], [5]. Additionally, Sugito et al. [6] studied the effect of both screen size and viewing distance on compressed and upscaled videos. They found that the distortions have a larger effect on quality at smaller viewing distances, as reported in other studies, and for larger screen sizes. Mikhailiuk et al. [7] measured the visually lossless thresholds, at which JPEG and WebP image compression distortions become invisible. The measurements were done for different viewing distances and display luminance levels. The results showed that the effect of viewing distance is content-dependent and does not follow a consistent pattern. For most content, distortions were less noticeable at longer distances, but for some, there was no change or the effect was reversed. Authors in [8], [9] explored a different but related problem: the effect of viewing distance on the perception of UHD and HD videos. Both studies concluded that there is no noticeable difference between UHD and HD videos when viewed from distances that are larger than the recommended viewing distance of 1.5 the display height.

In contrast to these works, we do not collect a new dataset. Instead, gather information from all publicly available datasets ([1], [3], [4], [7]) and use them to investigate how well the existing quality metrics account for the effect of viewing distance. We had to exclude the datasets from [6], [9] as those are not publicly available, from [8] because it contained only relative quality scores between HD and UHD, and from [5] because we lacked information on the screen size and resolution, which was essential to our study.

B. Quality assessment metrics at varying viewing distances

The metrics which consider the viewing distance do so by incorporating a contrast sensitivity function (CSF). CSF [10],

[11] models the smallest contrast detectable by an average observer on a uniform background. It can be used in a quality metric to pre-filter compared images, as done in sCIELab [12], or to weight each band of a band-pass decomposition by the corresponding sensitivity, as done in SSIMplus [13], HDR-VDP-3 [14] and FovVideoVDP [15]. A CSF alone does not model the discrimination of contrast differences above the detection threshold because of the contrast constancy of the visual system [16]. Therefore, CSF is often combined with the models of contrast masking [14], [15].

Some metrics like VMAF [17], have distinct models for different devices or resolutions, as a result, they lack the flexibility to account for any possible viewing distance. Others [2], [18] employ a neural network that takes the viewing distance as a feature. However, these models perform poorly when applied to new viewing distances that were not included in the training. Thus, these two metrics will not be included in the study. Gu et al. studied different approaches to adapt the existing metrics to varying viewing distances. In [19], they used a scaling function of the viewing distance while in [4], [20] they used the wavelet decomposition to prefilter compared images. We experimented with all these models but excluded them from our analysis as they performed worse than simple image rescaling, which we explain later in Section IV.

III. Reporting viewing distance and resolution

Different works report viewing distance in a different manner. Here, we want to clarify the notation and units used in this paper.

The viewing distance is often reported in multiplies of physical display height. This is because the recommended viewing distance for a FullHD and 4K television is often reported in such units. Some authors report the viewing distance in image heights, where the image occupies only a portion of the screen. Other authors enlarge the image to the full display resolution before displaying the images. Alternatively, the viewing distance can be reported as a distance in meters together with the screen dimensions and resolution. All those eclectic measures can be unified and simplified to a single measure that is directly related to the perceived resolution, also known as effective resolution — the resolution of the image that will be projected on the retina. The perceived resolution is best reported as the number of pixels within one visual degree, which can be computed as:

$$\rho = \frac{\pi\, r_y}{360\,\mathrm{atan}\left(\frac{0.5\,h}{d}\right)}\,, \tag{1}$$

where r_y is the display's vertical resolution in pixels, h is the display height and d is the viewing distance, both in the same units (e.g. meters). The equation assumes square pixels and a negligible effect of the viewing angle.

IV. Rescaling images to account for the viewing distance

Most metrics do not incorporate CSF and do not account for the viewing distance. However, we can use a simple observation: The size of the image projected on the retina is proportional to the viewing distance (refer to Eq. (1)). Hence, doubling the viewing distance approximately halves the size of the projected image on the retina. Using this observation, we can easily account for the viewing distance by rescaling an image accordingly.

We need to assume that each metric has been calibrated and performs the best at a certain viewing distance d_m. If we want to compute metric scores at a different viewing distance d, resulting in a different effective resolution, we need to rescale the image by a factor z_h:

$$z_\mathrm{h} = \frac{d_\mathrm{m}}{d} = \frac{\alpha h}{d} \tag{2}$$

where h represents the display height in the same units as d and α corresponds to the multiplies of viewing heights that correspond to the optimal effective resolution for a given metric. This approach is similar to the rescaling in [2], [19].

We optimized parameter α individually for each metric to yield the best Spearman correlation. This was done using the VDID2014 dataset (see Section V), which we will consider as a training dataset and exclude from testing. The value was found by performing an exhaustive search on the set $\alpha = \{1.5, 2, 2.5, 3\}$. The elements of the set correspond to the recommended viewing distances for QA datasets. The rescaling was performed using a box filter to simulate a display of lower/higher effective resolution. The default parameters of MATLAB and Python PIL functions are used for the filter.

In this work, we aim to do a cross-dataset comparison study of perceptual quality metrics for varying viewing distances. To accomplish this, we carefully select a set of well-known quality metrics that are most relevant to our study. A summary of the quality metrics we consider in addition to the α parameter chosen for each of them and the corresponding effective resolution is provided in Table I.

V. Quality Assessment datasets for varying viewing distance

Our focus in this paper is on image/video quality datasets that contain data collected at more than one viewing distance. The key information about each dataset is summarized in Table II and explained in more detail below.

a) VDID2014: The dataset [4] contains 8 pristine images with 2 different resolutions 512×512 and 768×512 px, and 160 distorted images produced by 4 distortion types (Gaussian Blur, White noise, JPEG2000 and JPEG) at 5 different levels. DMOS scores were collected using the single stimulus method. No information regarding the display size was provided, but, fortunately, it is not needed for our study, as they provided the viewing distance as a multiple of display height (refer to Eq. (2)).

b) CID:IQ: The dataset [3] contains 23 pristine images. Six different distortions have been used in the dataset: JPEG and JPEG2000 compression artefacts, Poisson Noise, Gaussian Blur, and two types of gamut mapping methods, ΔE and $SGCK$. All the distortions have been applied at five different

TABLE I: A list of the metrics used for the benchmark study. The "type" column states whether the metric is full-reference (FR) or no-reference (NR), intended for video or images. The calibrated resolution is the effective resolution at which the metric performs the best on the VDID2014 dataset, given as ppd (Eq. (1)) and α (Eq. (2)).

Metric	Metric type	Calibrated resolution: ppd (α)	Details
PSNR	FR-Image	29.30 (1.5)	A peak-signal-to-noise ratio in decibels.
SSIM [21]	FR-Image	29.30 (1.5)	A structural similarity index.
MS-SSIM [22]	FR-Image	57.09 (1.5)	A multi-scale version of SSIM.
FSIM [23]	FR-Image	38.49 (2.5)	Measures the similarity between the phase congruency and the gradient magnitude features.
GMSD [24]	FR-Image	38.49 (1.5)	Utilizes the gradients to predict the quality.
SFF [25]	FR-Image	38.49 (3)	Measures the fidelity between sparse encoding of the images.
VIF [26]	FR-Image	47.77 (1.5)	Natural scene statistic (NSS) based metric that models the images in the wavelet domain.
LPIPS [27]	FR-Image	29.30 (1.5)	Compares features extracted by deep--learning networks.
HDR-VDP3 [14]	FR-Image	-	A metric modeling low-level vision including CSF and contrast masking.
sCIELab [12]	FR-Image	-	Prefilters images with CSF and then calculates differences in the CIELab colour space.
FovVideoVDP [15]	FR-Video	-	An achromatic metric that models spatial and temporal low-level vision, the same principles as HDR-VDP-3.
VMAF [17]	FR-Video	57.09 (1.5)	Fuses VIF, ADM and motion features for quality predictions.
SpEED-QA [28]	FR-Video	57.09 (2.5)	Computes the difference between conditional block entropies of both the reference and distorted input.
BRISQUE [29]	NR-Image	38.49 (2)	An SVM-based model trained on NSS features.
BIQI [30]	NR-Image	57.09 (2.5)	A distortion-aware quality metric.
NIQE [31]	NR-Image	47.77 (2.5)	Employs a multivariate gaussian model on quality aware features.
FRIQUEE [32]	NR-Image	38.49 (2)	Captures statistics of the real-word images to predict quality.
PIQE [33]	NR-Image	38.49 (2.5)	A perception-based metric that uses features from spatially active regions.
HyperIQA [34]	NR-Image	29.30 (1.5)	Employs a self-adaptive hyper network for training the quality predictor model.

severity levels resulting in a total of 690 distorted images. DMOS scores were collected using a 9-categories ACR.

c) VCIP21: The dataset [1] contains 10 reference videos encoded with the HEVC codec to generate distorted videos at three bitrates to obtain "excellent", "good" and "fair" qualities. MOS scores were collected using the ACR method. Because no reference videos were provided with this dataset, we used the missing videos from AOM common test conditions v2.0 [35]. We could match reference videos for only half of the dataset (5 videos).

d) VLIC: The dataset [7] is different from the others as it contains only very subtle distortions — the compression level at which the distortions become just noticeable — Visually Lossless Threshold or VLT. The VLT was measured at two viewing distances and two display brightness levels, but we consider only the brighter level (220 cd/m^2) in our study. The dataset provides the probability of detecting the distortion at each compression level (for JPEG and WebP), which we converted into Just Objectionable Difference (JOD) units using Eq. (5) from [36]. We demonstrated that such JOD units are linearly related to the quality measured as mean-opinion scores [37]. Because the conversion from the probability of detection to JOD can introduce a large error close to probabilities of 0 and 1, we used only the conditions for which the probability of detection was between 0.1 and 0.9 at both viewing distances.

TABLE II: A summary of the datasets used in our study. The first column reports the number of conditions with respect to the number of references, the second column reports the display and visual field sizes, and the two last columns report the viewing distance used for each dataset as the effective resolution (Eq. (1)) and multiplies of display heights.

Dataset	Conditions (Reference)	Display size [in] / Field size [°]	Effective resolution [ppd]	Viewing distance as multiplies of display heights
VDID2014 [4]	160 (8)	unknown	38 57	2 3
CID:IQ [3]	690 (23)	24" / 56° × 33° 30° × 17°	32 64	1.7 3.3
VCIP21 [1]	15 (5)	46" / 31° × 18° 21° × 12° 15° × 8°	60 92 129	3.2 4.9 6.8
VLIC [7]	184 (20)	27" / 74° × 46° 41° × 24°	32 61	1.2 2.4

VI. Performance Analysis

In this section, we will evaluate the performance of existing state-of-the-art image and video quality metrics on the quality assessment task across different viewing distances. The metrics that do not natively account for viewing distances are adapted using Eq. (2). We do not evaluate the performance of the metrics without adaptation as it is not related to this study.

A. Evaluation Protocol

We used four metric performance indicators: the Spearman rank-order correlation coefficient (SROCC), the Pearson linear correlation coefficient (PLCC), the Kendall rank-order coefficient (KROCC) and the root mean squared error (RMSE). Due to space constraints, we include SROCC and RMSE in the main paper and the other two in the supplementary document [1]. A non-linear regression function was applied to the predictions before calculating the PLCC and RMSE coefficients. We used Eq. (16) from [4].

B. Confidence intervals

When comparing quality metrics, it is essential to account for the variance in subjective data. Such variance can cause some (if not most) of the performance differences between quality metrics to lack statistical significance. We used parametric resampling based on the reported standard errors. The standard errors were generously provided to us by the authors of [1], [3] or included in the datasets.

To compute the distribution of each performance indicator, we generated 1000 random samples. For each sample, the estimated subjective score value for each condition was drawn from $\mathcal{N}(\mu_i, \sigma_i)$, where μ_i was the reported quality value and σ_i was the standard error of the mean. Then, each metric performance indicator was computed separately for each sample. The distributions of the computed indicators can be found in Figure 1.

[1] You can find the supplementary material using this link.

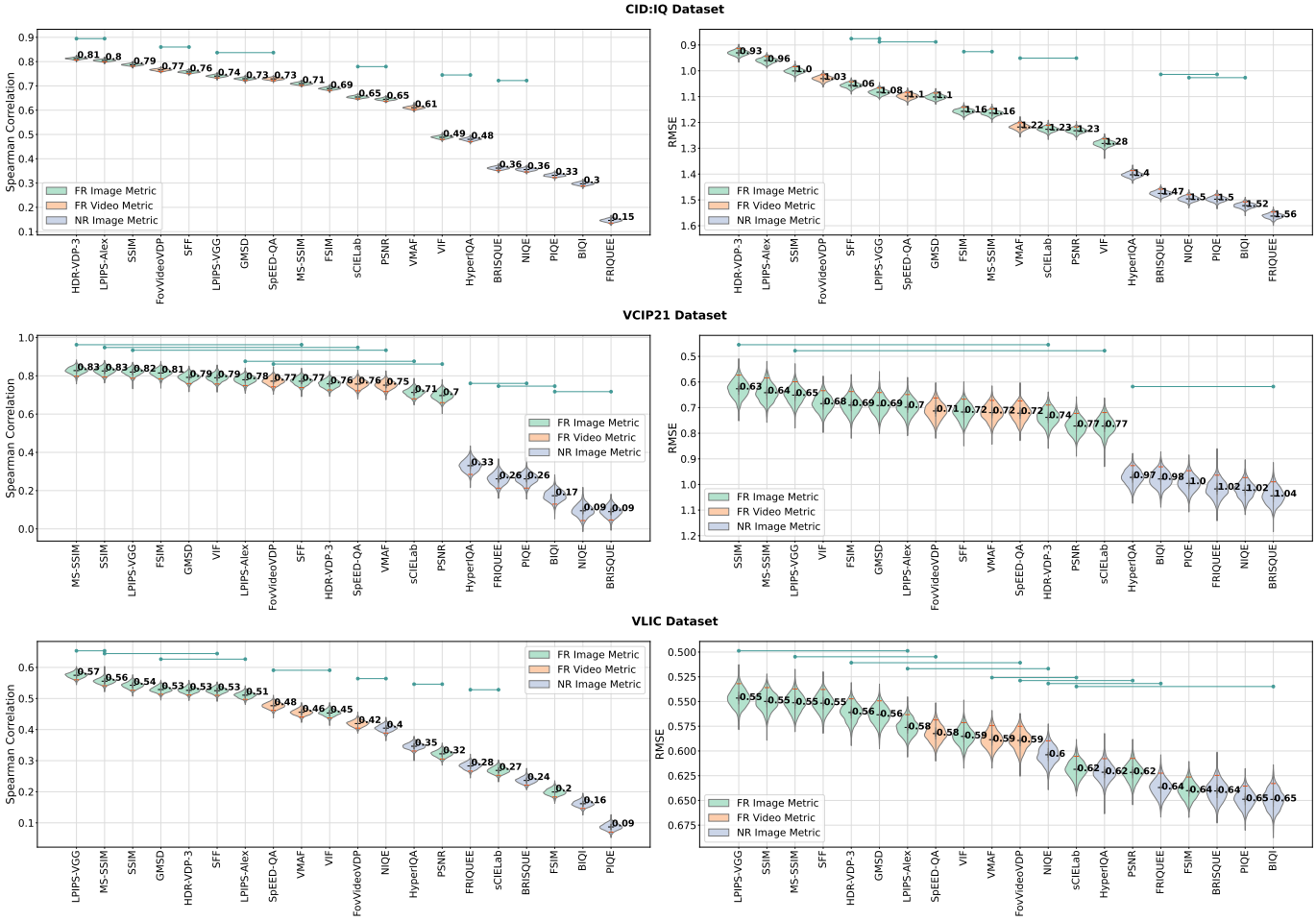

Fig. 1: Metric performance for the three tested datasets (rows), shown as SROCC and RMSE (columns). The violin plots visualize the error distribution due to the variance in subjective data. The black "-" denotes the mean score, and the red "-" denotes the 5th percentile which estimates the bad-case performance. The green solid lines group the metrics for which there is no statistical evidence that one metric is better than another. Refer to the supplementary for the PLCC and KROCC performance plots.

Furthermore, to test the statistical evidence of the results, we applied the one-tailed t-test at 0.05 significance level. The degrees of freedom were equal to $2N - 2$ where N was the number of observers. Prior to the test on the correlation coefficients, Fisher's Z-transform was applied to ensure a normal distribution of compared samples. Moreover, to account for the risk of obtaining false positive results when conducting multiple comparisons between metrics, the Benferroni-Holm correction [38] was applied to the p-values. We visualize the lack of statistical significance by plotting green lines in Figure 1 and Figure 2. When a line groups metrics, there is no statistical evidence to prove that the performance of one metric is better than any other in the same group.

C. Metrics Evaluation

The results for the CID:IQ and VCIP21 datasets, shown in Figure 1, indicate that most full-reference quality metrics perform well, except for VIF on CID:IQ. No-reference metrics perform poorly. We believe the problem could be the metrics' inability to generalize to new data. The results for the VLIC dataset (bottom row in Figure 1) show much worse performance for all the metrics. One explanation is that this dataset consists of subtle, just noticeable, distortions, which most metrics were not designed or trained for. We also observed that the results for this dataset are worse at 61 ppd than at 32 ppd (see the supplementary document).

The best-performing metrics differ from one dataset to another. However, we can single out LPIPS (both models), MS-SSIM, SSIM, SFF, and HDR-VDP-3 as the top metrics across all datasets. VCIP21 results show fewer statistically significant differences as this is a smaller dataset. Given the results for all three testing datasets, we do not see a pattern indicating whether scaling-based or CSF-based metrics perform better.

The performance results of the metrics at each fixed viewing distance (refer to the supplementary) report noteworthy

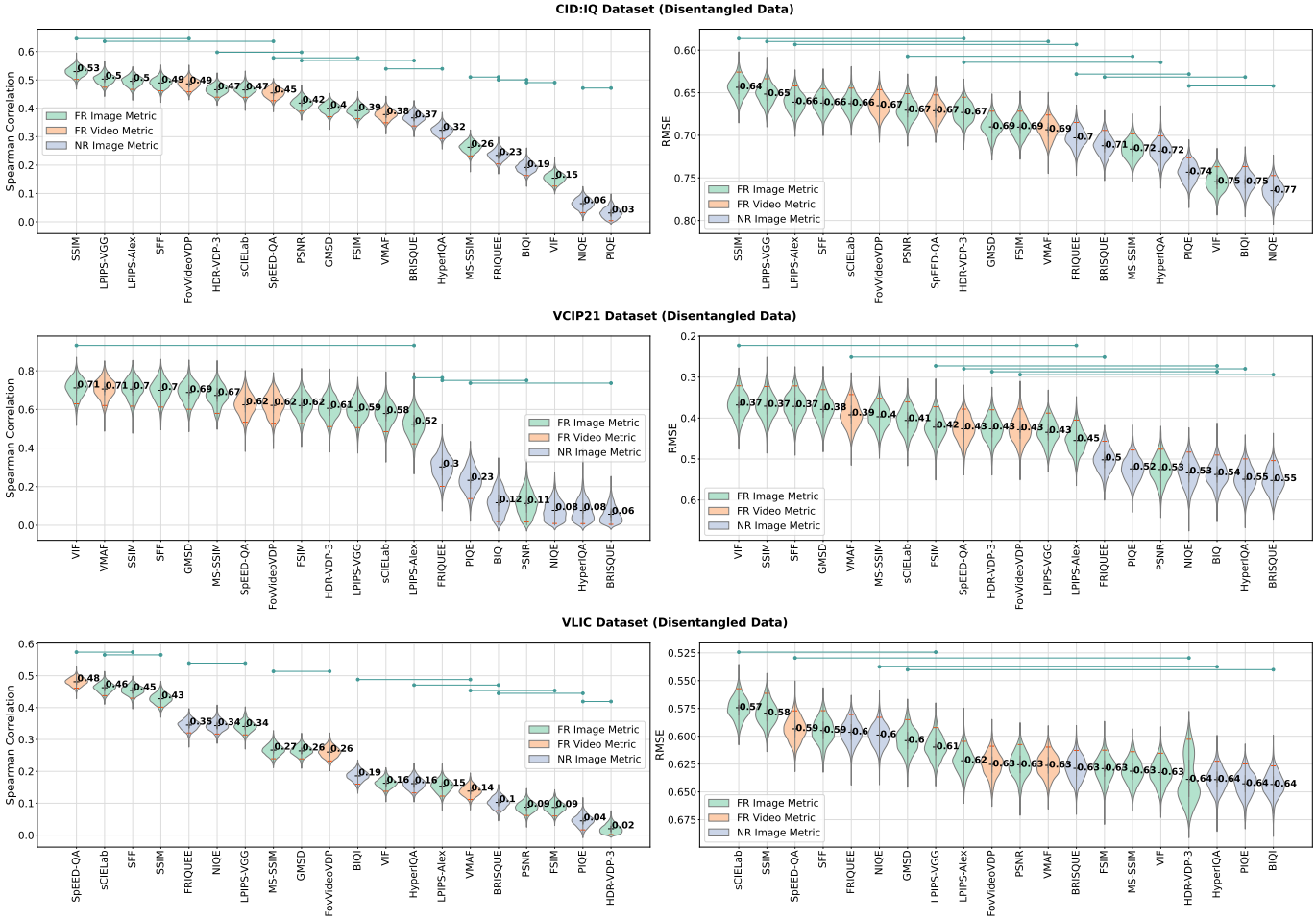

Fig. 2: Metric performance in terms of predicting quality differences between viewing distances (see Section VI-D). The notation is the same as in Figure 1. Refer to the supplementary for the PLCC and KROCC performance plots.

changes in performance between different viewing distances. For the CID:IQ dataset, the metrics' performance is comparable between viewing distances, with each metric performing slightly better at a specific viewing distance. In contrast, the VCIP21 dataset displays a distinct pattern, where most metrics perform the best at an effective resolution of 60 ppd, with a decrease in performance for higher effective resolutions. Similarly, the VLIC dataset exhibits a drastic drop in performance with the increase of effective resolution, excluding LPIPS-VGG, LPIPS-Alex and FSIM, which were able to perform better at an effective resolution of 61 ppd.

D. The effect of viewing distance disentangled

Our main research question is which metric does the best job in accounting for the viewing distance, and, to a lesser extent, which metric performs well at predicting the quality.

To disentangle the effect of viewing distance from general metric performance on the testing datasets, we calculate the differences between metric predictions at two viewing distances: $\Delta x_{i,kl} = x_{i,k} - x_{i,l}$, where $x_{i,k}$ is the metric prediction for the condition i and the viewing distance k. Then, we do the same for the subjective quality values and calculate SROCC and RMSE for the above differences. This way, we can report how well the metric compensates for the viewing distance and not whether it performs well on a given dataset.

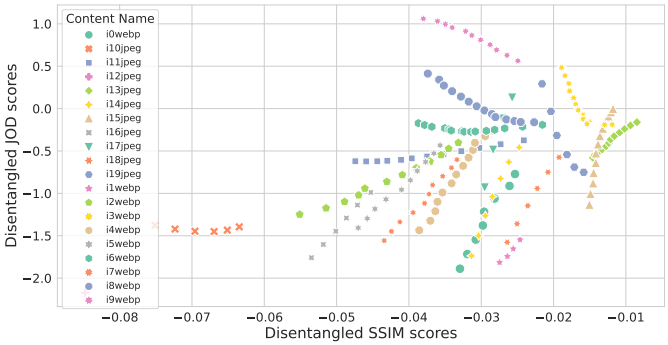

Fig. 3: Scatter plot of the disentangled quality scores for the SSIM metric and VLIC dataset. The values show the difference between the 32 ppd (near) and 61 ppd (far) conditions.

The results for the disentangled performance indices are re-

ported in Figure 2. We observe that the correlation coefficients for the disentangled data are low, in particular for CID:IQ and VLIC datasets. We investigated this further (see Figure 3 and the scatter plots in the supplementary document) and observed that the effect of the viewing distance on quality is irregular and very different for different content. While we expect the distortions to become less visible at larger viewing distances, there are examples that show otherwise. For example, the VLIC dataset contains three images in which the distortions become more noticeable at larger distances. Furthermore, the effect of distance is very different for different content — some are almost unaffected by the distance, while others show a substantial change in quality. Neither image rescaling, nor CSF used in the metrics, can predict this effect well.

VII. Conclusion

In this paper, we investigated the performance of existing metrics on the task of predicting quality at different viewing distances. In particular, we compared the metrics that natively account for the viewing distance (via CSF) with those that were provided with rescaled images. Using four publicly available datasets designed for this task, our results indicated that there is no statistical evidence suggesting that one approach is better than another. Rather, both failed to predict the changes in quality introduced by the change in viewing distance. Our results suggest that the effect of the viewing distance of quality is complex, and it requires better models than those used in the existing metrics.

References

[1] H. Amirpour, R. Schatz, C. Timmerer, and M. Ghanbari, "On the impact of viewing distance on perceived video quality," in *VCIP*. IEEE, 2021, pp. 1–5.

[2] Ye, Nanyang and Wolski, Krzysztof and Mantiuk, Rafal K., "Predicting visible image differences under varying display brightness and viewing distance," in *CVPR*. IEEE, 2019, pp. 5429–5437.

[3] X. Liu, M. Pedersen, and J. Y. Hardeberg, "CID: IQ–a new image quality database," in *ICISP*. Springer, 2014, pp. 193–202.

[4] K. Gu, M. Liu, G. Zhai, X. Yang, and W. Zhang, "Quality assessment considering viewing distance and image resolution," *IEEE TBC*, vol. 61, no. 3, pp. 520–531, 2015.

[5] R. Fang, D. Wu, and L. Shen, "Evaluation of image quality of experience in consideration of viewing distance," in *ChinaSIP*. IEEE, 2015, pp. 653–657.

[6] Y. Sugito, Y. Kondo, D. Arai, and Y. Kusakabe, "Modeling Perceived Quality on 8K VVC Video Under Various Screen Sizes and Viewing Distances," *IEEE Access*, vol. 10, pp. 97 237–97 247, 2022.

[7] A. Mikhailiuk, N. Ye, and R. K. Mantiuk, "The effect of display brightness and viewing distance: a dataset for visually lossless image compression," in *HVEI*. IS&T, 2021.

[8] J. Kufa and T. Kratochvil, "Visual quality assessment considering ultra HD, Full HD resolution and viewing distance," in *Radioelektronika*. IEEE, 2019, pp. 1–4.

[9] A. Lachat, J.-C. Gicquel, and J. Fournier, "How perception of ultra-high definition is modified by viewing distance and screen size," in *IQSP*, vol. 9396. SPIE, 2015, pp. 306–313.

[10] P. G. J. Barten, *Contrast sensitivity of the human eye and its effects on image quality*. SPIE Press, 1999.

[11] R. K. Mantiuk, M. Ashraf, and A. Chapiro, "stelaCSF - A Unified Model of Contrast Sensitivity as the Function of Spatio-Temporal Frequency, Eccentricity, Luminance and Area," *ACM TOG*, vol. 41, no. 4, pp. 1–16, 2022.

[12] X. Zhang and B. A. Wandell, "A spatial extension of CIELAB for digital color-image reproduction," *Journal of the society for information display*, vol. 5, no. 1, pp. 61–63, 1997.

[13] A. Rehman, K. Zeng, and Z. Wang, "Display device-adapted video quality-of-experience assessment," in *HVEI*, vol. 9394. SPIE, 2015, pp. 27–37.

[14] R. K. Mantiuk, D. Hammou, and P. Hanji, "HDR-VDP-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content," *arXiv preprint*, 2023.

[15] R. K. Mantiuk, G. Denes, A. Chapiro, A. Kaplanyan, G. Rufo, R. Bachy, T. Lian, and A. Patney, "Fovvideovdp: A visible difference predictor for wide field-of-view video," *ACM TOG*, vol. 40, no. 4, pp. 1–19, 2021.

[16] B. Y. M. A. Georgeson and G. D. Sullivan, "Contrast constancy: deblurring in human vision by spatial frequency channels," *Journal of Physiology*, vol. 252, no. 3, pp. 627–656, 1975.

[17] 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, p. 2, 2016.

[18] A. Chetouani and M. Pedersen, "Image Quality Assessment without Reference by Combining Deep Learning-Based Features and Viewing Distance," *Applied Sciences*, vol. 11, no. 10, p. 4661, 2021.

[19] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Self-adaptive scale transform for IQA metric," in *ISCAS*. IEEE, 2013, pp. 2365–2368.

[20] K. Gu, G. Zhai, M. Liu, Q. Xu, X. Yang, J. Zhou, and W. Zhang, "Adaptive high-frequency clipping for improved image quality assessment," in *VCIP*. IEEE, 2013, pp. 1–5.

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

[22] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *ACSSC*, vol. 2. IEEE, 2003, pp. 1398–1402.

[23] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE TIP*, vol. 20, no. 8, pp. 2378–2386, 2011.

[24] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE TIP*, vol. 23, no. 2, pp. 684–695, 2013.

[25] H.-W. Chang, H. Yang, Y. Gan, and M.-H. Wang, "Sparse feature fidelity for perceptual image quality assessment," *IEEE TIP*, vol. 22, no. 10, pp. 4007–4018, 2013.

[26] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE TIP*, vol. 15, no. 2, pp. 430–444, 2006.

[27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.

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

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

[30] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE SPL*, vol. 17, no. 5, pp. 513–516, 2010.

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

[32] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *JoV*, vol. 17, no. 1, pp. 32–32, 2017.

[33] N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *NCC*. IEEE, 2015, pp. 1–6.

[34] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *CVPR*, 2020, pp. 3667–3676.

[35] X. Zhao, Z. R. Lei, A. Norkin, T. Daede, and A. Tourapis, "AOM Common Test Conditions v2. 0," *Alliance for Open Media, Codec Working Group Output Document*, 2021.

[36] M. Perez-Ortiz and R. K. Mantiuk, "A practical guide and software for analysing pairwise comparison experiments," *arXiv preprint*, 2017.

[37] M. Perez-Ortiz, A. Mikhailiuk, E. Zerman, V. Hulusic, G. Valenzise, and R. K. Mantiuk, "From pairwise comparisons and rating to a unified quality scale," *IEEE TIP*, vol. 29, pp. 1139–1151, 2020.

[38] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," *Biom J*, vol. 50, no. 3, pp. 346–363, 2008.