On the improvement of 2D quality assessment metrics for omnidirectional images

Abderrezzaq Sendjasni, Mohamed-Chaker Larabi, Faouzi Alaya Cheikh

On the Improvement of 2D Quality Assessment Metrics for Omnidirectional Images

Abderrezzaq Sendjasni [1,2]***, Mohamed-Chaker Larabi*** [1] ***and Faouzi Alaya Cheikh*** [2]
[1] ***CNRS, Xlim UMR 7252, Université de Poitiers, France***
[2] ***Faculty of Computer Science and Media Technology, NTNU, Norway***

Abstract

Subjective quality assessment remains the most reliable way to evaluate image quality while being tedious and money consuming. Therefore, objective quality evaluation ensures a trade-off by providing a computational approach for predicting image quality. Even though a large literature exists for 2D image and video quality evaluation, 360-degree images quality is still under-explored. One can question the efficiency of 2D quality metrics on such a new type of content. To this end, we propose to study the possible improvement of well-known 2D quality metrics using important features related to 360-degree content, i.e. equator bias and visual saliency. The performance evaluation is conducted on two databases containing various distortion types. The obtained results show a slight improvement of the performance highlighting some problems inherently related to both the database content and the subjective evaluation approach used to obtain the observers' quality scores.

Keywords: *Omnidirectional images, quality metrics, performance evaluation.*

1. Introduction

In recent years, virtual reality (VR) has experienced an impressive growth. 360-degree images represent an important part of the VR content, in which the users are provided with real world scenes to live an immersive experience. With commercial head mount displays (HMDs), the viewer is allowed to freely focus on the desired content thanks to his head movements (HM) making the interactive and the immersive experience more interesting. Accordingly, to achieve good quality of experience (QoE), immersive contents with high visual quality should be provided. Thus, 4K or even higher resolutions are required. Based on this, quality assessment of omnidirectional images becomes crucial to control QoE.

No doubt that images Quality Assessment (IQA) is an important aspect for numerous applications. There are two types of IQA, objective and subjective assessment. In the former, the aim is to find an algorithm that evaluates the quality of image/video as a human observer would do. Hence, the computed quality scores have to be well correlated with those given by human observers, since the user is the ultimate receiver of the image/video. As for the latter, it is considered as the most reliable method to obtain the quality score where subjects are asked to rate the viewed image/video [20]. Unfortunately, subjective experiments are time-consuming, expensive and require a lot of effort to provide reliable results [6]. Hence, developing objective quality metrics is perceived as a good tradeoff. Depending on the availability of the reference image, objective IQA can be classified into : 1)Full-Reference (FR) which are used when the reference image is available, 2)Reduced-Reference (RR) using partial information from the reference image, and 3)No-Reference (NR) where the reference image is not needed [1].

IQA has been extensively studied for 2D images in the past few decades resulting in a large number of quality metrics. Therefore, it seems typical to use these metrics directly on omnidirectional images, as they are generally processed, encoded and transmitted using in a 2D plane representation. Meanwhile, it has been highlighted in [13] that, sampling density from sphere to plane projection is not uniform at every pixel location, leading thus to a modification of the impact of the different regions on the final quality. Therefore, an objective metric for omnidirectional images that mitigates the aforementioned limitation is needed.

Motivated by the large number of quality metrics dedicated to 2D images, we present in this paper an extensive evaluation of the performance of these metrics when applied to on omnidirectional images. This evaluation is made using two databases containing multiple degradation. In this work, our aim is not to review these metrics. Rather, we aim to understand and analyze their behavior when applied directly on omnidirectional images and possibly propose some tuning using: 1) equator bias weighting which correspond to the way the human gaze is biased towards the equatorial line when watching 360-degree images, 2) perceptual weighting such as the one obtained by a saliency model dedicated to such image type. By adding these features to existing 2D IQA metrics, the performance could be improved when assessing omnidirectional image quality. Finally, we provide a statistical evaluation of the used IQA metrics. The provided results aim to bring insight for readers for a better understanding of the current issues regarding QA of omnidirectional images. Furthermore, the provided discussions based on the evaluation results are expected to motivate and inspire new thoughts for designing new IQA algorithms or rethinking the way the QA is performed for this emerging type of media.

2. IQA ALGORITHMS

In the past few decades, several IQA metrics have been proposed for 2D images,where the perceived quality measurement has been and still the main focus of researchers. These metrics are widely used in image quality evaluation, benchmark, image/video databases creation and evaluation. Hence, we select for this study fifteen metrics from the literature, for which source code is made available by their original authors. This way, one can avoid any doubts with regards to the original algorithm. The selected metrics include signal-based and perception-based algorithms. In the following section, the selected metrics are described. For the rest

this paper, I and $\hat{I}$ denote the reference and distorted image respectively, M and N represent the width and the height of images respectively.

Peak Signal-to-Noise Ratio (PSNR) is a widely used distortion measurement based on pixel-to-pixel error to estimate the quality of an image. The PSNR is expressed as :

$$PSNR(I,\hat{I}) = 10\log_{10}(\frac{MAX^2}{MSE(I,\hat{I})}), \quad (1)$$

where MSE denotes the mean squared error and can be obtained as follows:

$$MSE = \frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(I(i,j)-\hat{I}(i,j))^2}{MN}. \quad (2)$$

The Mean Absolute error (MAE) is a simple mathematical measurement of the error between two images. In [11], a Perceptual-fidelity aware MSE (PAMSE) is proposed based on a linear structure extraction resulting to the structural MSE. Structural similarity (SSIM) proposed in [22] is based on the assumption that the HVS is substantially adapted to extract the structural information from a scene. In SSIM, the similarity measurement in performed in three steps: luminance (3), contrast (4) and structure comparison (5). These steps are denoted as l, c, s respectively and formulated as:

$$l(I,\hat{I}) = \frac{2\mu_I\mu_{\hat{I}}+C_1}{\mu_I^2+\mu_{\hat{I}}^2+C_1}, \quad (3)$$

$$c(I,\hat{I}) = \frac{2\sigma_I\sigma_{\hat{I}}+C_2}{\sigma_I^2+\sigma_{\hat{I}}^2+C_2}, \quad (4)$$

$$s(I,\hat{I}) = \frac{\sigma_{I\hat{I}}+C_3}{\sigma_I\sigma_{\hat{I}}+C_3}, \quad (5)$$

where $C1$, $C3$ and $C2$ are constants. Finally, the SSIM index is defined as:

$$SSIM(I,\hat{I}) = \left[l(I,\hat{I})\right]^{\alpha}\left[c(I,\hat{I})\right]^{\beta}\left[s(I,\hat{I})\right]^{\gamma}, \quad (6)$$

α, β and γ are positive weighting factors used to tune the importance of each component. The Universal Quality Index (UQI) [21] is the predecessor of SSIM and is considered as a special case of this latter where $\alpha=\beta=\gamma=1$ and $C1=C2=C3=0$ [10].

Later, a Multi-scale version of the structural similarity metric namely MS-SSIM is proposed in [10], where the image details at different resolutions and viewing conditions are incorporated. First, a low-pass filter is applied. Then, the contrast comparison (eq. 4) and the structure comparison (eq. 5) are calculated at different scales $(1,\ldots,M)$. Finally, luminance comparison (eq. 3) is computed, but only at the last scale M. The overall SSIM evaluation is obtained by combining the measurement at the different scales as follows:

$$MS-SSIM(I,\hat{I}) = \left[l(I,\hat{I})\right]^{\alpha M}.\prod_{i=1}^{M}\left[c(I,\hat{I})\right]^{\beta i}\left[s(I,\hat{I})\right]^{\gamma i}. \quad (7)$$

Visual information fidelity (VIF) proposed in [7] measures the information fidelity between two images. The fidelity is quantified based on the amount of mutual information between I and $\hat{I}$ by modeling the images in the wavelet domain using Gaussian scale mixture. In [19], the feature similarity (FSIM) index is proposed based on the assumption that the HVS is attracted by low-level characteristics. FSIM is based on two features to measure local similarity,*i.e.* phase congruency (PC) and gradient magnitude (GM). The former is also used at the pooling stage as a weighting factor. An extended version of FSIM to color channels (FSIMc) through pixel-wise fidelity over chroma channels was also proposed. Riesz Transform based Feature Similarity (RFSIM) [18] is based on the fact that the HVS is attracted to image edges. Therefore, the RFSIM is computed using the 1^st and 2^nd-order Riesz transforms to extract the image's local structure and compare the feature maps at key locations marked by a feature mask generated using an edge detection algorithm.

Based on the important role played by the image gradient for the HVS, reflecting the contrast and the structure information of an image, the Gradient Magnitude Similarity Deviation (GMSD) is proposed in [12]. GMSD computes the horizontal and vertical gradients for both I and $\hat{I}$ by convolving Prewitt filter along the two directions. The image gradient maps at pixel i are obtained as follows:

$$G(i) = \sqrt{G_x(i)^2+G_y(i)^2} \quad (8)$$

where G_x and G_y represents the horizontal and vertical gradients respectively. With c a positive constant for numerical stability, N the number of pixels, the final score of GMSD is computed using the standard deviation of the Gradient Magnitude Similarity map, as follows:

$$GMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(GMS(i)-GMSM)^2} \quad (9)$$

where

$$GMS(i) = \frac{2G_I(i)G_{\hat{I}}(i)+c}{G_I(i)^2+G_{\hat{I}}(i)^2+c} \quad (10)$$

The Mean Deviation Similarity Index (MDSI), proposed in [23], benefits from the important role of the image gradient for the HVS and uses two similarity maps. The gradient map and the color distortion map are pooled by a deviation pooling strategy based on the measure of central tendency (MCT).

Visual Saliency-based Index (VSI) [17] and Spectral Residual based Similarity (SR-SIM) [16] use visual saliency and gradient magnitude. VSI exploits visual saliency as a feature when computing the local quality map of the distorted image as well as a weighting factor at the pooling stage so as to reflect the importance of a local region. SR-SIM extracts the residuals of the image in the spectral domain and constructs its corresponding spatial domain saliency map using a Fourier transform-based approach. In [5], the Haar Wavelet-based Perceptual Similarity Index is proposed based on local features to construct similarity map using the coefficients of a discrete Haar wavelet transform.

2.1 Omnidirectional IQA

Regarding omnidirectional IQA, a very few objective IQA metrics have been proposed. Most of them are simply derived from traditional PSNR, SSIM or MSE [2]. For instance, Weighted

Spherical PSNR (WS-PSNR) uses the scaling factor $w(i,j)$ from 2D plane to the sphere as a weighting factor for PSNR computation. For each pixel (i,j) in I and $\hat{I}$ the $WS-PSNR$ is obtained by $Eq.11$.

$$WS-PSNR = 10\log_{10}(\frac{MAX^2}{WMSE}), \tag{11}$$

where,

$$WMSE = \frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(I(i,j)-\hat{I}(i,j))^2.w(i,j)}{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}w(i,j)}, \tag{12}$$

and,

$$w(i,j) = \cos(\frac{\pi}{N}(j+\frac{1}{2}-\frac{N}{2})). \tag{13}$$

Other PSNR-based metrics were also proposed, particularly in [14], *Zakharchenko et al.* proposed to compute PSNR on the projection plane of Craster parabolic projection (CPP). It is named as CPP-PSNR, where PSNR is computed after re-mapping pixels of the original and distorted images from the spherical domain to the CPP projection. *Yu et al.* [13] proposed Spherical PSNR (S-PSNR) in which, the PSNR is computed in the spherical domain.

3. Improvement of 2D IQA for omnidirectional images

3.1 Equator Bias

Unlike 2D images, omnidirectional images can be viewed in a 360-degree range. However, when using Head Mounted Display (HMD), viewers tend to focus on the region near the equator as reported in [8]. In this study, eye fixations data was collected using an eye tracker combined with an HMD during the conducted experiment. Therefore, the regions around the equator are considered as regions of interest (RoI) in omnidirectional contents and may have a higher impact on the global perceived quality. Consequently, we believe that adding more importance to the equator region and incorporate the equatorial bias as a weighting factor can perceptually improve existing 2D IQA for omnidirectional images. We will refer to it as $E-Metric$. To this end, we modeled the equator bias as a Gaussian function depending on the image latitude (α). Hence, for an $M \times N$ image at latitude i the equator bias is denoted as $E(i)$.An Illustration is given in Fig. $1(b)$.

$$E(i) = \exp(\frac{-(i-\frac{N}{2})^2}{\alpha N}) \tag{14}$$

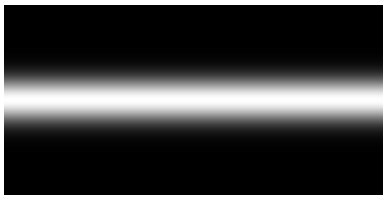
Similarly, we use the scaling factor from 2D plane to the spherical domain used in [2] as a weighting factor and, as a special case the weights in the Equirectangular projection are given in Eq.13. Therefore, we incorporated this weighting factor to 2D metrics and refer to it as $W-Metric$.

3.2 Visual Attention Weighting

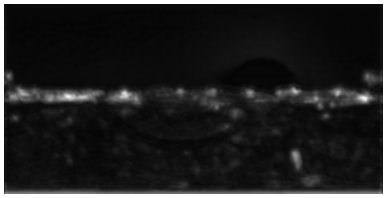
In this part, we propose to incorporate visual attention information into existing 2D metrics so as to give more importance to RoIs and add perceptual properties. As for human attention modeling, we chose to use saliency features for omnidirectional contents. For this matter, we select the model described in [15] in order to produce the saliency maps because of its efficiency and availability. Hence, every element of the obtained saliency map is used as weighting factor $w_{i,j}$ for the corresponding pixel of the image. This will be refered to as $Sal-Metric$. An illustration of the obtained weighting map is given in Fig. 1.

In addition, we investigate the combination of the equator bias and scaling factors with visual saliency in order to study the possible additivity of improvements. It will refered to as $ESal-Metric$ and $WSal-Metric$ respectively.

(a) Original image



(b) Equator Bias



(c) Saliency map

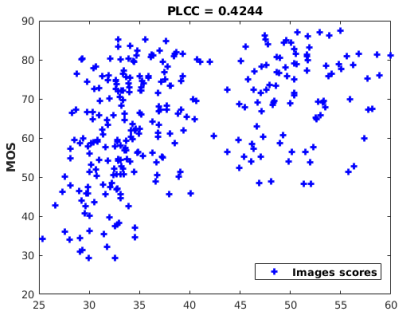
Figure 1: Illustration of the equator bias and the saliency prediction.

3.3 Omnidirectional Image Datasets

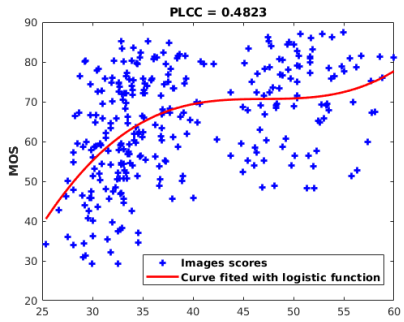
To date, IQA of omnidirectional content is suffering from the unavailability of large and reliable datasets. Very few work have been done to this end due to its complexity and difficulty. Indeed, building a dataset requires subjective experiments to gather human opinions represented as the Mean Opinion Score (MOS), in addition to an appropriate environment and test conditions. As a special case, for omnidirectional subjective tests, the observers view the images through commercial HMDs. The latter presents some shortcomings that may affect the effectiveness of the observer's quality rating as the screen door effect. Neglecting such a phenomena may result in an unreliable evaluation. Another common issue with subjective scores in general is the non-linear nature of the obtained scores requiring a non-linear regression using a five parameter logistic function, as recommended in the ITU-R recommendations [4], prior to the performance evaluation. However, it is to be recalled that the use of the logistic function cannot be done if the native correlation is below 0.7. Otherwise the correlation value cannot be considered as reliable because regression quality is very low.

For instance, Fig. 2 depicts the scatter plot of a given objec-

tive scores against the subjective scores MOS. One can notice that scores are widely spread and do not show a consistent correlation confirmed by the Pearson correlation coefficient of $PLCC = 0.42$. By applying the aforementioned regression, the correlation score improves artificially ($PLCC = 0.48$). In some cases, this may provide a significant improvement that is totally inconsistent.



(a) Without regression



(b) With regression

Figure 2: Illustration of inconsistent regression.

For this study, we chose two publicly available datasets. and . Huang et al. [3] proposed a dataset of 12 omnidirectional source images down-sampled to four different spatial resolutions and compressed using three JPEG quality factors of 20, 60, 100, resulting in a total of 144 omnidiretional images. The HTC VIVE was used as the test device. As for the test method, the Absolute Category Rating (ACR) Single Stimulus was adopted. 98 subjects participated to the subjective experiment, 53 males and 45 females with ages of 18 to 25. The subjects were standing and had to give their scores verbally. The range of this latter was divided to 100 levels, from 0 (Bad) to 100 (Excellent). Each subject observed three different image contents at four different spatial resolutions and three quality factors. A transition of 5 seconds between every two image samples was implemented.

Sun et al. [9] proposed the CVIQD2018 datatset containing 16 original omnidirectional, compressed using JPEG compression with quality factors ranging from 50 to 0 with an interval of -5, H.264/AVC and H.265/HEVC with factors from 30 to 50 with an interval of 2, leading to 528 compressed images for a single source image. As for the subjective test, the recommendations in ITU-RB500-11 were followed to validate the proposed dataset with the participation of 20 subjects including 15 males and 6 females. The adopted method is a Single Stimulus one and the HTC VIVE was used as the test device where an interaction system developed using Unity3D software was designed to facilitate the display and the collection of the subjective scores. The rating was scaled into 10 levels from the lowest to the highest quality. The experiment was conducted in an empty room and subjects were sitting on a swivel chair so they can only turn around.

Table 1 summarize both datasets' characteristics in terms of quality distortion types, number of reference/distorted images and the number of subjects. All the images in the datasets are provided in equirectangular projection.

4. Results and Discussion

In order to assess the performance of the pre-selected metrics and the improved versions, we used four common metrics including Spearman Rank Order Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), Kendal Rank Order Correlation Coefficient (KRCC) and the Root Mean Squared Error (RMSE). A metric is considered having a perfect prediction when $LCC = SRCC = KROOC = 1$ and $RMSE = 0$. We computed the predicted scores using a five parameters non-linear logistic function as given below :

$$f(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp^{\beta_2 (x - \beta_3)}} \right) + \beta_4 x + \beta_5 \tag{15}$$

where x denotes the objective score and, $f(x)$ represents the corresponding mapped score. $\beta_i (i = 1,2,3,4,5)$ correspond to the logistic function parameters to be fitted. Next, the obtained mapped scores are compared to the subjective scores (DMOS) in order to measure the performance of the selected IQA metrics using the aforementioned metrics. Furthermore, we adopted the regression criterion discussed previously; so Eq. 15 is applied only if the native PLCC ≥ 0.7.

4.1 Performance of 2D IQ metrics

With the intent of giving a comprehensive evaluation of state-of-the art 2D IQA metrics on omnidirectional images, we first evaluated fifteen FR IQ metric designed for 2D content on two public databases (see Table 1), and we compared them with two PSNR-based IQA metrics designed for omnidirectional content. Tables 2 and 3 provide the LCC, SRCC, KRCC and RMSE results without using the non-linear regression on Huang et al. and CVIQD2018 datasets, respectively. The best three performance are highlighted in each column. This evaluation is performed with the aim to inspect 2D metrics' behavior when applied directly to omnidirectional images. As we know, QA metrics designed for 2D images do not take into account omnidirectional content characteristics, but since the original and the impaired images are both with geometric distortion, 2D metrics may assess the similarity effectively. Also, an overview of the scatter plots for Huang et al. and CVIQD2018 is given in Fig. 3 and Fig. 4 respectively in order to visualize the relation between predicted and subjective scores.

From Tables 2 and 3, one can observe that globally, the predicted scores exhibit better consistency with CVIQD2018's subjective score unlike Huang et al's. Fig. 4 supports this observation, where one can see that the scatter plots depict a concentration of the dots around the diagonal. This is mainly related to the subjective scores nature, since both datasets used the same

Table 1: Description of the used omnidirectional image datasets.

Datasets	Ref Images	Dis Images	Dis Types	Subjects
Huang et al. [3]	25	300	JPEG with different QFs/Different resolution	98
CVIQD2018 [9]	16	528	JPEG with different QFs/H.264 and H.265 with different QPs	20

methodology and same equipment(HMD). An other interesting observation is that, some metrics outperform WS-PSNR and S-PSNR, PSNR-based quality metrics that are designed for omnidirectional content. This is due to the fact that PSNR does not always agree with human judgment. Although, the behavior of PSNR compared to S-PSNR and WS-PSNR is quite interesting, we can clearly see that PSNR outperformed S-PSNR and WS-PSNR in both datasets. An explanation to this behavior may be the way we computed PSNR. Specifically for this latter, we measured PSNR as the log of the sum of the three channels' MSE.

From Table 2, we can observe that VSI, SR-SIM and RF-SIM achieved the best three performance. Specifically, VSI and SR-SIM, visual attention based metrics, outperformed the other metrics. As for CVIQD2018, Table 3 reveals that structure similarity based metrics such as MDSI, SSIM and HaarPSI have the best performance. This may be related to the nature of the images and the impairment types. The use of multiple datasets bring more accuracy to the evaluation of QA models, in the same time, its makes the comparison among datasets difficult, since they contain different scenes and may use different tools. From Figs. 3 and 4, one can see that the distribution is consistent with the performance score given in Tables 2 and 3 where each of VSI, SR-SIM and RF-SIM scatter are more concentrated and correlated. This justifies the values of PLCC *i.e.* 0.755, 0.756 and 0.747 respectively. On the other hand, from the scatter plot of UQI, we can observe almost no relationship between the objective and the subjective scores which justify the PLCC value of 0.328.

Table 2: Performance evaluation (LCC, SRCC, KRCC and RMSE) of the selected 2D-IQA metrics on Huang et al. [3].

Metric	LCC	SRCC	KRCC	RMSE
WS-PSNR	0.419	0.465	0.325	**28.973**
S-PSNR	0.419	0.466	0.325	**27.798**
PSNR	0.449	0.481	0.338	**30.324**
MAE	0.39	0.453	0.317	64.162
SSIM	0.634	0.671	0.496	65.17
MS-SSIM	0.455	0.42	0.284	65.174
PAMSE	0.445	0.479	0.332	64.081
HaarPSI	0.477	0.451	0.305	65.221
RF-SIM	**0.747**	**0.707**	**0.526**	65.199
GMSD	0.493	0.474	0.326	66.133
FSIM	0.696	0.668	0.49	65.168
FSIMc	0.72	0.684	0.503	65.169
MDSI	0.672	0.683	0.504	66.022
UQI	0.328	0.323	0.216	65.398
VIF	0.411	0.422	0.286	65.456
VSI	**0.755**	**0.693**	**0.513**	65.167
SR-SIM	**0.756**	**0.705**	**0.528**	65.167

Table 3: Performance evaluation (LCC, SRCC, KRCC and RMSE) of the selected 2D-IQA metrics on CVIQD2018 [9].

Metric	LCC	SRCC	KRCC	RMSE
WS-PSNR	0.741	0.724	0.52	**18.025**
S-PSNR	0.741	0.724	0.52	**19.073**
PSNR	0.783	0.751	0.551	**18.343**
MAE	0.763	0.735	0.535	48.554
SSIM	**0.854**	0.887	0.698	50.28
MS-SSIM	0.853	0.876	0.683	50.295
PAMSE	0.718	0.789	0.583	46.931
HaarPSI	**0.923**	**0.905**	**0.728**	50.444
RF-SIM	0.84	0.846	0.648	50.362
GMSD	0.854	0.845	0.643	51.17
FSIM	0.845	**0.911**	**0.735**	50.274
FSIMc	0.844	**0.913**	**0.738**	50.274
MDSI	**0.914**	0.903	0.721	51.036
UQI	0.809	0.836	0.633	50.837
VIF	0.841	0.851	0.655	50.814
VSI	0.77	0.893	0.709	50.263
SR-SIM	0.814	0.885	0.696	50.264

4.2 Performance on Individual Distortions

Performance evaluation over different types of impairment gives a hazy idea about quality metrics behavior. Therefore, it is important to analyze their performance on individual impairments. For example, structural similarity-based metrics detect structural changes which is the main distortion existing in the used dataset. Hence, in this section, we provide an in-depth analysis of the selected 2D metrics on individual impairment types (spatial resolution for Huang et al. and AVC, HEVC, JPEG for CVIQD2018). The performance is evaluated using LCC and SRCC and the results are given in Tables 4 and 5 for huang et al. and CVIQD2018 respectively. From these tables, several observations could be made based on the obtained performance results. For instance, from Table 4 we can observe that all the metrics achieved better performance on single distortions than on the overall dataset. With the increase of spatial resolution, the performance drops, which may means that, 2D QA metrics are not suitable for high resolution content in this case. Additionally, PSNR performed poorly compared to S-PSNR and WS-PSNR unlike when evaluated globally. As for the best performance, SR-SIM still outperforms the rest. GMSD and MS-SSIM achieved quite impressive performance on content-independent distortions compared to their performance on the overall datasets. This confirms that heterogeneous resolutions may affect the performance of quality metrics. From Table 5, we can observe that structural distortions measurement-based metrics still outperform the other metrics; HaarPSI and MDSI performed the best on content-independent distortion. Meanwhile, SSIM performs worst than VIF, but its performance improved compared to the overall one, this conclusion holds for the other metrics.

Table 4: Performance evaluation (LCC, SRCC) of the selected 2D-IQA metrics on individual impairments for Huang et al. [3].

Metric	LCC				SRCC			
	4k	2k	1k	720p	4k	2k	1k	720p
WS-PSNR	0.52	0.625	0.683	0.616	0.513	0.671	0.707	0.636
S-PSNR	0.52	0.625	0.684	0.616	0.513	0.672	0.71	0.636
PSNR	0.426	0.553	0.601	0.53	0.476	0.631	0.655	0.562
MAE	0.401	0.483	0.577	0.556	0.52	0.662	0.693	0.614
SSIM	0.534	0.682	0.705	0.773	**0.701**	**0.8**	**0.815**	**0.758**
MS-SSIM	**0.748**	0.812	0.836	0.817	**0.704**	**0.797**	**0.812**	0.749
PAMSE	0.568	0.649	0.741	0.735	0.625	0.762	0.794	0.719
HaarPSI	0.735	0.808	0.844	0.823	0.628	0.759	0.78	0.705
RF-SIM	0.696	0.76	0.805	0.817	0.603	0.762	0.78	**0.761**
GMSD	**0.742**	**0.832**	**0.871**	**0.848**	0.656	0.765	0.795	0.694
FSIM	0.657	0.759	0.793	0.833	**0.692**	0.786	0.8	**0.745**
FSIMc	0.677	0.78	0.818	**0.846**	0.664	0.772	0.778	0.721
MDSI	0.667	0.762	0.789	0.72	0.613	0.764	0.779	0.687
UQI	0.646	0.686	0.684	0.68	0.603	0.734	0.716	0.651
VIF	0.629	0.694	0.732	0.684	0.663	0.752	0.769	0.725
VSI	0.736	**0.839**	**0.86**	0.795	0.62	0.767	0.783	0.692
SR-SIM	**0.757**	**0.836**	**0.862**	**0.853**	**0.688**	**0.799**	0.81	0.73

Table 5: Performance evaluation (LCC, SRCC) of the selected 2D-IQA metrics on individual impairment for CVIQD2018 [9].

Metric	LCC			SRCC		
	AVC	HEVC	JPEG	AVC	HEVC	JPEG
WS-PSNR	0.738	0.712	0.828	0.748	0.717	0.722
S-PSNR	0.738	0.712	0.828	0.747	0.717	0.722
PSNR	0.754	0.678	0.871	0.765	0.682	0.779
MAE	0.738	0.641	0.839	0.762	0.677	0.765
SSIM	0.896	0.884	0.853	**0.945**	0.922	**0.933**
MS-SSIM	0.836	0.818	0.889	0.879	0.86	0.914
PAMSE	0.71	0.669	0.757	0.843	0.803	0.873
HaarPSI	**0.924**	**0.916**	**0.947**	0.924	0.912	0.93
RF-SIM	0.89	0.874	0.844	0.916	0.895	0.912
GMSD	0.84	0.847	0.908	0.866	0.869	0.913
FSIM	0.893	0.89	0.837	**0.947**	**0.934**	0.928
FSIMc	0.896	0.89	0.839	**0.948**	**0.934**	**0.933**
MDSI	**0.942**	**0.915**	**0.96**	0.941	0.911	**0.937**
UQI	0.848	0.804	0.89	0.878	0.84	0.867
VIF	**0.913**	**0.913**	**0.945**	0.924	**0.924**	0.887
VSI	0.86	0.835	0.799	0.919	0.882	0.931
SR-SIM	0.861	0.863	0.8	0.929	0.911	0.923

Table 6: Performance evaluation (LCC, SRCC) of the selected 2D-IQA metrics on individual impairment for Huang et al."With Regression"

Metric	LCC				SRCC			
	4k	2k	1k	720p	4k	2k	1k	720p
WS-PSNR	-	-	-	-	-	-	-	-
S-PSNR	-	-	-	-	-	-	-	-
PSNR	-	-	-	-	-	-	-	-
MAE	-	-	-	-	-	-	-	-
SSIM	-	-	0.849	**0.865**	**-**	-	**0.819**	**0.758**
MS-SSIM	**0.758**	0.825	**0.862**	0.864	**0.704**	**0.797**	**0.812**	**0.749**
PAMSE	-	-	0.772	0.86	-	-	0.794	0.604
HaarPSI	**0.746**	0.819	0.86	0.86	**0.628**	0.759	0.78	0.705
RF-SIM	-	0.76	0.846	0.817	-	0.762	0.777	**0.761**
GMSD	0.742	0.832	**0.871**	0.861	0.656	0.765	0.795	0.738
FSIM	-	**0.839**	0.857	**0.866**	-	0.794	0.805	0.745
FSIMc	-	0.835	**0.863**	**0.868**	-	**0.775**	0.777	0.721
MDSI	-	0.762	0.789	0.72	-	0.764	0.779	0.687
UQI	-	-	-	-	-	-	-	
VIF	-	-	0.753	-	-	-	0.769	-
VSI	0.736	**0.839**	0.86	0.859	0.62	0.767	0.783	0.692
SR-SIM	**0.757**	**0.836**	**0.862**	0.864	**0.688**	**0.799**	**0.81**	0.73

Table 7: Performance evaluation (LCC, SRCC) of the selected 2D-IQA metrics on individual impairment for CVIQD2018."With Regression"

Metric	PLCC			SRCC		
	AVC	HEVC	JPEG	AVC	HEVC	JPEG
WS-PSNR	0.748	0.725	0.853	0.747	0.717	0.722
S-PSNR	0.751	0.73	0.86	0.748	0.719	0.732
PSNR	0.767	0.687	0.89	0.766	0.682	0.779
MAE	0.769	0.677	0.888	0.763	0.678	0.765
SSIM	**0.946**	0.884	0.853	**0.945**	0.922	**0.933**
MS-SSIM	0.881	0.861	0.963	0.879	0.86	0.914
PAMSE	0.78	0.808	0.938	0.843	0.804	0.873
HaarPSI	0.926	0.918	**0.971**	0.924	0.912	0.93
RF-SIM	0.916	0.898	0.96	0.916	0.895	0.912
GMSD	0.84	0.847	0.908	0.866	0.869	0.913
FSIM	**0.95**	**0.936**	**0.972**	**0.947**	**0.934**	0.928
FSIMc	**0.951**	**0.935**	**0.973**	**0.948**	**0.934**	**0.933**
MDSI	0.942	0.915	0.96	0.941	0.911	**0.937**
UQI	0.888	0.845	0.922	0.877	0.839	0.863
VIF	0.927	**0.927**	0.952	0.924	**0.924**	0.887
VSI	0.922	0.89	0.969	0.919	0.882	0.931
SR-SIM	0.931	0.914	0.8	0.929	0.911	0.923

4.3 Performance of improved 2D IQ metrics

In this section we discuss the performance evaluation of improved IQ metrics based on the improvement mentioned in Section 3. To understand existing 2D IQ metrics limits for omnidirectional IQA, we applied the mentioned adjustments to eight 2D metrics namely PSNR, MAE, SSIM, MS-SSIM, PAMSE, UQI, SR-SIM and VSI. Tables 8 and 9 list the performance evaluation results (LCC and SRCC) on individual distortions for Huang et al. and CVIQD2018, respectively, after applying regression. Cells containing dashes represent the performance that did not fit the condition of $PLCC \geq 0.7$. We first observe the improvements over standard versions, for Huang et al. Improvement could be noticed for SSIM, MS-SSIM, PAMSE and UQI with the five different tunings. In the mean time, the performance dropped for VSI and SR-SIM. From this latter observation, one can conclude that the incorporation of omnidirectional properties with visual attention-based 2D metrics may drop the performance. But, the opposite can be observed in Table 9, where VSI and SR-SIM outperformed the their native version, specially for JPEG compression. Although, SSIM and MS-SSIM, performed better than their native version for Huang et al. and their performance dropped for CVIQD2018.

The observation mentioned before regarding the relationship between spatial resolution and quality metrics performance observed for Huang et al. still hold even with the proposed tunings. For instance, MS-SSIM performed the best among the eight metrics and, we can observe from Table 8 that its performance is increasing for lower resolutions. The same observation applies to the rest of the metrics.

5. Conclusion

In this work, we evaluated fifteen state-of-the art IQA models designed for 2D images on two 360-degree image databases

with the aim to understand their behavior for this emerging media. The main contributions of this work consist of an exploration of the existing 2D QA metrics' limit when assessing the quality of 360-degree images and, the possibility to tune them according to the characteristics and the nature of this type of content. To this end, we proposed to incorporate in selected metrics perceptual properties such as equator bias and visual attention weightings in addition to the scaling factor to account for the spherical nature of the image. Results showed some improvement in comparison to standard 2D metrics. However, this improvement is relatively limited depending on the metric and the content. These results question about the reliability of both available databases and existing 2D metrics when it comes to 360-degree image quality evaluation. This opens several future work such as the development of dedicated quality metrics and the construction of reliable databases.

References

[1] David R. Bull. Measuring and Managing Picture Quality. In *Communicating Pictures*, pages 317–360. Elsevier, 2014.

[2] S. Chen, Y. Zhang, Y. Li, Z. Chen, and Z. Wang. Spherical structural similarity index for objective omnidirectional video quality assessment. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, July 2018.

[3] M. Huang, Q. Shen, Z. Ma, A. C. Bovik, P. Gupta, R. Zhou, and X. Cao. Modeling the perceptual quality of immersive images rendered on head mounted displays: Resolution and compression. *IEEE Transactions on Image Processing*, 27(12):6039–6050, Dec 2018.

[4] ITU-R. *Methodology for the subjective assessment of the quality of television pictures BT Series Broadcasting service*, volume 13. 2012.

[5] Rafael Reisenhofer, Sebastian Bosse, Gitta Kutyniok, and Thomas Wiegand. A haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication*, 61:33 – 43, 2018.

[6] F. Ribeiro, D. Florencio, and V. Nascimento. Crowdsourcing subjective image quality evaluation. In *2011 18th IEEE International Conference on Image Processing*, pages 3097–3100, Sep. 2011.

[7] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, Feb 2006.

[8] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1633–1642, April 2018.

[9] W. Sun, K. Gu, S. Ma, W. Zhu, N. Liu, and G. Zhai. A large-scale compressed 360-degree spherical image database: From subjective quality evaluation to objective model comparison. In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, Aug 2018.

[10] Z. Xiao. A multi-scale structure similarity metric for image fusion qulity assessment. In *2011 International Conference on Wavelet Analysis and Pattern Recognition*, pages 69–72, July 2011.

[11] W. Xue, X. Mou, L. Zhang, and X. Feng. Perceptual fidelity aware mean squared error. In *2013 IEEE International Conference on Computer Vision*, pages 705–712, Dec 2013.

[12] W. Xue, L. Zhang, X. Mou, and A. C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, Feb 2014.

[13] M. Yu, H. Lakshman, and B. Girod. A framework to evaluate omnidirectional video coding schemes. In *2015 IEEE International Symposium on Mixed and Augmented Reality*, pages 31–36, Sep. 2015.

[14] Vladyslav Zakharchenko, Kwang Pyo Choi, and Jeong Hoon Park. Quality metric for spherical panoramic video. In Khan M. Iftekharuddin, Abdul A. S. Awwal, Mireya García Vázquez, Andrés Márquez, and Mohammad A. Matin, editors, *Optics and Photonics for Information Processing X*, volume 9970, pages 57 – 65. International Society for Optics and Photonics, SPIE, 2016.

[15] K. Zhang and Z. Chen. Video saliency prediction based on spatial-temporal two-stream network. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(12):3544–3557, Dec 2019.

[16] L. Zhang and H. Li. Sr-sim: A fast and high performance iqa index based on spectral residual. In *2012 19th IEEE International Conference on Image Processing*, pages 1473–1476, Sep. 2012.

[17] L. Zhang, Y. Shen, and H. Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, Oct 2014.

[18] L. Zhang, L. Zhang, and X. Mou. Rfsim: A feature based image quality assessment metric using riesz transforms. In *2010 IEEE International Conference on Image Processing*, pages 321–324, Sep. 2010.

[19] L. Zhang, L. Zhang, X. Mou, and D. Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, Aug 2011.

[20] X. Zhang, W. Lin, S. Wang, J. Liu, S. Ma, and W. Gao. Fine-grained quality assessment for compressed images. *IEEE Transactions on Image Processing*, 28(3):1163–1175, March 2019.

[21] Zhou Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, March 2002.

[22] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[23] H. Ziaei Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *IEEE Access*, 4:5579–5590, 2016.

Table 8: Performance evaluation (LCC, SRCC) of the improved 2D-IQA metrics on individual impairment for Huang et al.

Metric	E-Metric								W-Metric								Sal-Metric								ESal-Metric								WSal-Metric							
	4k		2K		1K		720p		4k		2K		1K		720p		4k		2K		1K		720p		4k		2K		1K		720p		4k		2K		1K		720p	
	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC
PSNR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MAE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.73	0.717	-	-
SSIM	0.738	0.704	**0.833**	0.796	0.851	0.814	**0.866**	**0.752**	0.744	**0.716**	-	-	0.704	0.805	0.865	**0.759**	-	-	**0.83**	**0.803**	0.845	0.8	0.861	**0.77**	-	-	0.828	**0.799**	0.849	0.801	0.86	0.76	-	-	0.824	0.794	**0.848**	0.792	0.858	**0.758**
MS-SSIM	**0.761**	**0.709**	0.83	**0.798**	**0.861**	0.81	0.863	0.748	**0.761**	**0.716**	**0.83**	**0.797**	**0.863**	0.809	0.864	0.754	**0.778**	**0.731**	0.822	0.793	0.847	0.807	**0.866**	0.766	**0.766**	**0.726**	0.822	0.788	0.849	**0.812**	0.862	**0.763**	**0.763**	**0.731**	0.817	0.789	0.846	**0.805**	0.861	0.756
PAMSE	-	-	0.815	0.778	0.847	0.8	0.864	0.716	-	-	0.814	0.783	0.807	0.797	**0.866**	0.722	0.71	0.636	0.75	0.765	**0.867**	**0.818**	**0.866**	0.712	0.713	0.658	**0.838**	0.798	**0.868**	**0.812**	**0.867**	0.717	0.713	0.654	**0.836**	**0.796**	0.833	0.802	**0.865**	0.72
UQI	-	-	0.733	0.793	0.699	0.743	-	-	-	-	-	0.729	0.77	-	-	-	-	-	-	-	0.738	0.783	-	-	-	-	-	-	0.728	0.767	-	-	-	-	-	0.726	0.768	-	-
VSI	0.732	0.694	0.83	0.796	0.859	**0.816**	0.863	0.733	0.732	0.694	**0.83**	0.796	0.859	**0.816**	0.863	0.733	-	-	0.775	0.747	0.822	0.772	0.793	0.705	0.721	0.609	0.82	0.76	0.839	0.778	0.863	0.733	-	-	0.805	0.752	0.833	0.775	0.863	0.733
SR-SIM	0.732	0.694	0.83	0.795	0.86	0.815	0.864	0.732	0.732	0.693	**0.83**	**0.797**	0.86	0.815	0.864	0.731	-	-	0.727	0.747	0.817	0.742	0.865	0.739	0.715	0.649	0.809	0.765	0.85	0.763	**0.867**	0.74	-	-	0.779	0.751	0.838	0.751	0.863	0.735

Table 9: Performance evaluation (LCC, SRCC) of the improved 2D-IQA metrics on individual impairment for CVIQD2018.

Metric	E-Metric						W-Metric						Sal-Metric						ESal-Metric						WSal-Metric					
	AVC		HEVC		JPEG		AVC		HEVC		JPEG		AVC		HEVC		JPEG		AVC		HEVC		JPEG		AVC		HEVC		JPEG	
	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC	LCC	SRCC
PSNR	0.76	0.76	0.7	0.693	0.885	0.775	0.751	0.751	0.687	0.682	0.88	0.765	0.782	0.781	0.758	0.75	0.884	0.781	0.787	0.787	0.775	0.767	0.868	0.782	0.796	0.796	0.786	0.776	0.882	0.781
MAE	0.779	0.778	0.657	0.656	0.894	0.796	0.766	0.768	0.64	0.636	0.895	0.785	0.77	0.772	0.7	0.697	0.89	0.793	0.77	0.77	0.709	0.716	0.885	0.789	0.776	0.774	0.733	0.728	0.888	0.787
SSIM	**0.947**	**0.946**	**0.927**	**0.925**	0.849	0.932	**0.949**	**0.947**	**0.933**	**0.933**	0.846	0.93	**0.951**	**0.951**	**0.939**	**0.938**	0.818	0.927	**0.95**	**0.95**	**0.937**	**0.935**	0.813	0.923	**0.949**	**0.949**	0.943	**0.941**	0.809	0.919
MS-SSIM	0.832	0.883	0.82	0.868	0.963	0.911	0.832	0.884	0.823	0.87	0.962	0.909	0.848	0.898	0.894	0.892	0.967	0.918	0.844	0.899	0.849	0.897	0.965	0.914	0.848	0.901	0.903	0.901	0.872	0.911
PAMSE	0.849	0.85	0.822	0.819	0.94	0.876	0.701	0.835	0.809	0.805	0.934	0.865	0.858	0.857	0.844	0.838	0.813	0.877	0.766	0.859	0.791	0.846	0.818	0.88	0.867	0.868	0.86	0.856	0.803	0.877
UQI	0.902	0.892	0.861	0.854	0.927	0.867	0.902	0.892	0.861	0.854	0.927	0.867	0.902	0.892	0.861	0.854	0.927	0.867	0.902	0.892	0.861	0.854	0.927	0.867	0.902	0.892	0.861	0.854	0.927	0.867
VSI	0.925	0.922	0.892	0.888	0.97	**0.933**	0.926	0.923	0.892	0.889	0.971	**0.934**	0.927	0.926	0.9	0.895	**0.969**	**0.936**	0.929	0.928	0.898	0.894	0.971	0.935	0.927	0.926	0.899	0.895	**0.97**	**0.937**
SR-SIM	0.936	0.935	0.914	0.913	**0.973**	**0.933**	0.936	0.935	0.914	0.913	**0.973**	0.933	0.936	0.936	0.92	0.918	0.968	0.935	0.94	0.94	0.92	0.919	**0.974**	**0.937**	0.938	0.937	0.919	0.918	**0.972**	0.936

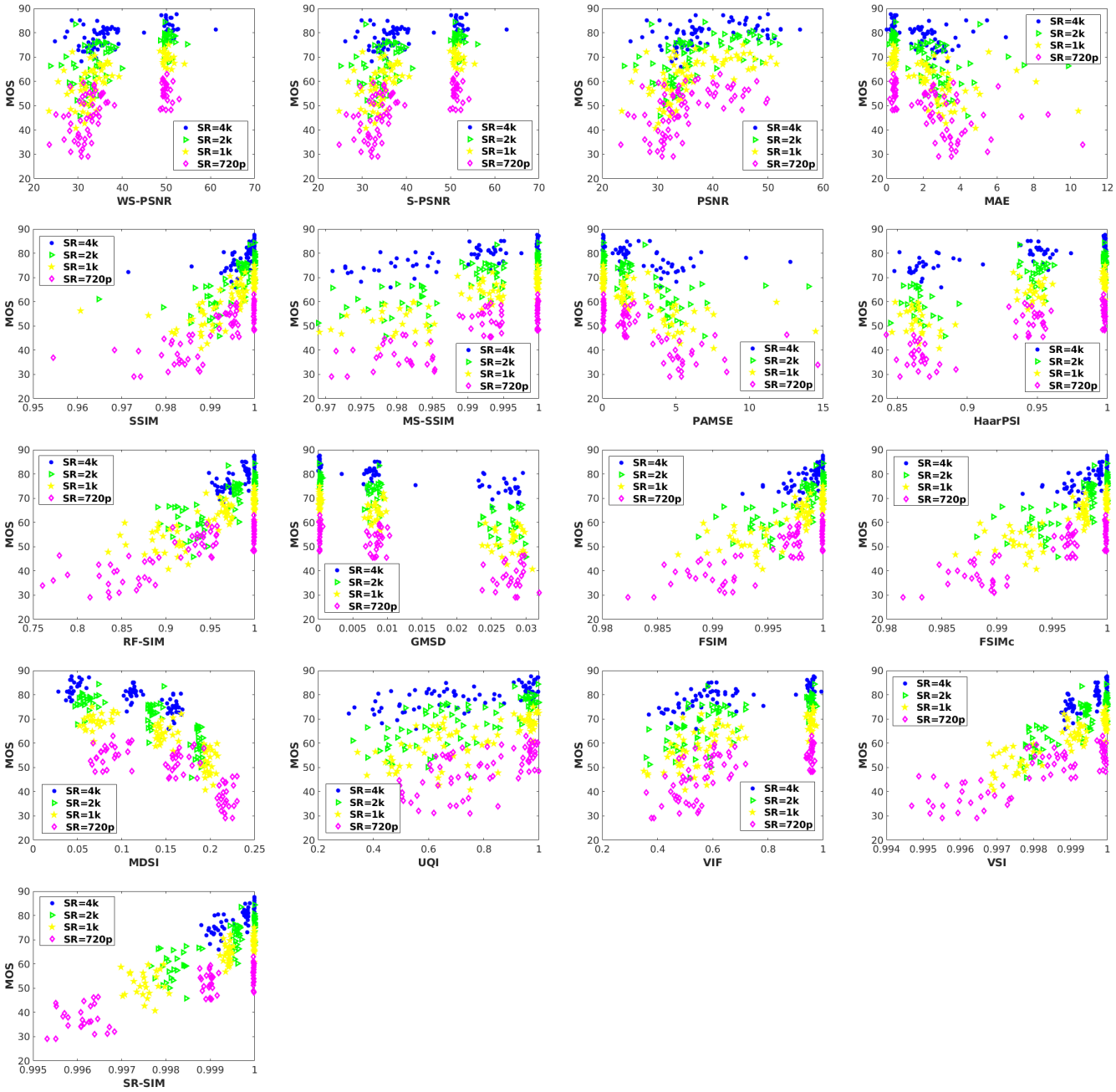


Figure 3: Scatter plots of predicted quality scores by 2D metrics against subjective scores (MOS) for Huang et al. [3].

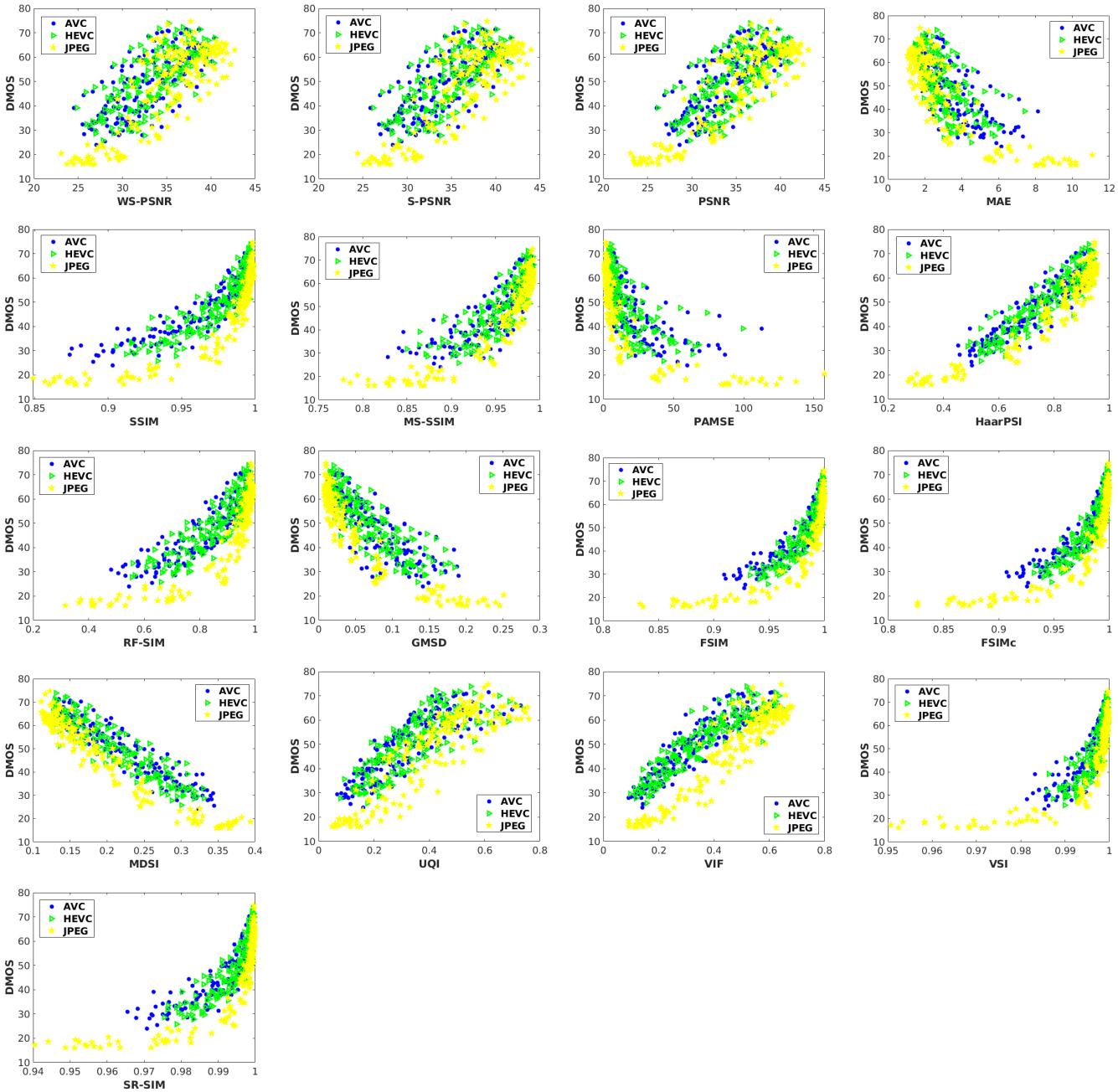


Figure 4: Scatter plots of predicted quality scores by 2D metrics against subjective scores (DMOS) for CVIQD2018 [9].