New Measurements Reveal Weaknesses of Image Quality Metrics in Evaluating Graphics Artifacts

Martin Čadík* Robert Herzog* Rafał Mantiuk◇ Karol Myszkowski* Hans-Peter Seidel*
*MPI Informatik Saarbrücken, Germany ◇Bangor University, United Kingdom

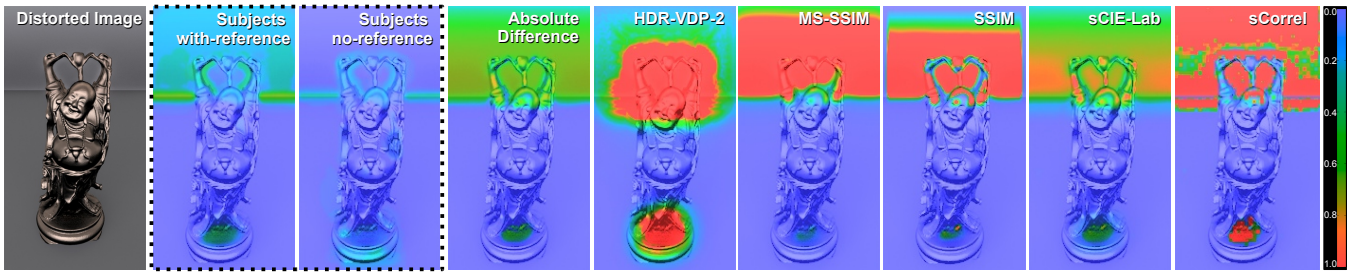


Figure 1: *State-of-the-art image quality metrics often fail in the prediction of the human-perceived distortions in complex images. Here, we show the predicted detection probabilities (color-coded) for gradient-based tone mapping artifacts [Fattal et al. 2002] in a synthetic image.*

Abstract

Reliable detection of global illumination and rendering artifacts in the form of localized distortion maps is important for many graphics applications. Although many quality metrics have been developed for this task, they are often tuned for compression/transmission artifacts and have not been evaluated in the context of synthetic CG-images. In this work, we run two experiments where observers use a brush-painting interface to directly mark image regions with noticeable/objectionable distortions in the presence/absence of a high-quality reference image, respectively. The collected data shows a relatively high correlation between the with-reference and no-reference observer markings. Also, our demanding per-pixel image-quality datasets reveal weaknesses of both simple (PSNR, MSE, sCIE-Lab) and advanced (SSIM, MS-SSIM, HDR-VDP-2) quality metrics. The most problematic are excessive sensitivity to brightness and contrast changes, the calibration for near visibility-threshold distortions, lack of discrimination between plausible/implausible illumination, and poor spatial localization of distortions for multi-scale metrics. We believe that our datasets have further potential in improving existing quality metrics, but also in analyzing the saliency of rendering distortions, and investigating visual equivalence given our with- and no-reference data.

CR Categories: I.3.0 [Computer Graphics]: General;

Keywords: Image quality metrics (IQM), perceptual experiments, global illumination, noticeable and objectionable distortions

Links: DL PDF WEB DATA

*e-mail: mcadik@mpi-inf.mpg.de, the complete dataset is available at: http://www.mpii.de/resources/hdr/iqm-evaluation/

1 Introduction

Rendering techniques, in particular global illumination, are prone to image artifacts, which might arise due to specific scene configurations, imbalanced scene complexity that might lead to a locally varying convergence-rate of the solution, and numerous simplifications in the rendering algorithms themselves. With the proliferation of 3D rendering services, where the user may often arbitrarily interact with the content, the role of automatic rendering-quality control gains in importance. Even in well-established industries such as gaming a massive approach to automatic quality testing is desirable. In practice, objective image quality metrics (IQM) that are successful in lossy image compression and transmission applications [Wang and Bovik 2006] are predominantly used in graphics, including advanced attempts of their adaptation to actively steer rendering [Rushmeier et al. 1995; Bolin and Meyer 1998; Ramasubramanian et al. 1999]. Such objective IQM are trained to predict a single value of mean opinion score (MOS) for image blockiness, noise, blur, or ringing distortions. However, their performance for other distortion types as well as their spatial localization within an image has not been systematically validated so far.

The goal of this work is to generate a new rendering-oriented dataset with localized distortion maps and use it for the evaluation of existing IQM. For this purpose we prepare a set of images with distortions that are typical for popular global illumination and rendering techniques as well as the corresponding distortion-free reference images. Table 1 presents a summary of our stimuli. In two separate experiments (Sec. 3) we ask the observers to locally mark *noticeable* and *objectionable* distortions where the reference image is either shown or hidden, respectively. We demonstrate that the observers can reliably perform both tasks, yielding high coefficients of agreement (Sec. 4.1). In general, our results show a high correlation between the observer marking for the *with-reference* and *no-reference* datasets, but we also indicate the most common sources of discrepancies in such marking (Sec. 4.2).

We use the with-reference dataset to evaluate the performance of state-of-the-art *full-reference* (FR) IQM in detecting and localizing rendering distortions (Sec. 5). We show that even advanced IQM fail for some common computer graphics artifacts (e.g., Fig. 1). Our data shows that in general no IQM performs better than any other, even including the simple absolute difference (AD), which is equivalent to the peak signal-to-noise ratio (PSNR) or mean-

square-error (MSE) given our non-parametric metric performance measures. Moreover, our analysis reveals some interesting weaknesses of FR IQM, including the lack of robustness to brightness and contrast change, the inability to distinguish between plausible and implausible illumination patterns, and poor localization of distortions due to multi-scale processing.

2 Related work

In this section we briefly characterize general purpose full reference (FR) IQM which are central for our comparison against the subjective data. Also, we review major other developments in the evaluation of IQM performance. For more in depth discussion of the image quality problem we refer the reader to the recent textbooks [Wang and Bovik 2006; Wu and Rao 2005], and survey papers [Lin and Kuo 2011; Pedersen and Hardeberg 2011].

2.1 Image quality metrics (IQM)

Full reference IQM can be categorized into different groups based on the principles behind their construction [Wang and Bovik 2006; Pedersen and Hardeberg 2011].

Mathematically-based metrics directly measure the difference of pixel intensity. The root mean square error (RMSE) and peak signal-to-noise-ratio (PSNR) are the most prominent examples of metrics belonging to this category.

HVS-based metrics model early human vision characteristics such as luminance adaptation, contrast sensitivity, visual masking, and visual channels. The most prominent examples of such metrics include the Visible Differences Predictor (VDP) [Daly 1993] and Visual Discrimination Model (VDM) [Lubin 1995]. VDP has also been used in the evaluation of rendered image quality [Rushmeier et al. 1995]. Recently, extensions of VDP have been proposed to handle high dynamic range (HDR) images [Mantiuk et al. 2005; Mantiuk et al. 2011].

Structure-based metrics detect structural changes in the image by means of a spatially localized measure of correlation in pixel values. The Structural Similarity Index Metric (SSIM) is based on this principle. In addition, it is sensitive to the differences in the mean intensity and contrast [Wang and Bovik 2006, Ch. 3.2].

Other metrics combine the strengths of different metric categories. For example, in sCIE-Lab [Zhang and Wandell 1998] spatial color sensitivity is added to a standard color-difference measure in the perceptually-uniform CIE-Lab color-space. In the Visual Signal-to-Noise Ratio (VSNR) metric [Chandler and Hemami 2007] at first an HVS-model is applied to eliminate distortions below the visibility threshold and then a simple mathematically-based metric is used. Other modern metrics, such as the Visual Information Fidelity (VIF) index [Wang and Bovik 2006, Ch. 3.3], rely on natural-scene statistics and employ an information-theoretic approach to measure the amount of information that is shared between two images.

2.2 Evaluation of image quality metrics

The comparison of IQM performance against data collected in experiments with human subjects is required to evaluate metric prediction accuracy and robustness for different types of visual distortions. Standardized procedures for subjective image- and video-quality evaluation have been developed by the International Telecommunication Union [ITU-T-P.910 2008; ITU-R-BT.500-11 2002]. They rely on subjectively collected *mean opinion score (MOS)* data, which is compared against a single number derived from the error pooling over pixels. While such a procedure works well for estimating the overall *magnitude of distortions*, information on different distortion types, their possible interactions and spatial distribution is not captured. In computer graphics applications the prediction of *local distortion detectability* by a human observer is essential, and in this work we favor *image distortion maps*, which capture such spatial information.

Mean opinion score (MOS) data. A number of databases of images with different distortion types and MOS subjective quality scores is publicly available where LIVE [Sheikh et al. 2006] and Tampere Image Database [Ponomarenko et al. 2009] are the most prominent examples featuring both significant variety of distortions and large number of stimuli, which have been judged by many subjects (30–200). Lin and Kuo [2011] present a more complete summary of such databases with detailed characterization of supported distortion types, which arise mostly in image compression and transmission. Distortions covered by those databases that are more relevant for graphics applications include blur, mean intensity shifts, contrast changes, and various types of noise.

Image distortion maps. The spatial aspect of distortion detectability has been addressed in calibration and performance evaluation for HDR image [Mantiuk et al. 2005] and video [Čadík et al. 2011] quality metrics. Thereby, the screen is divided into discrete blocks of about 30×30 pixels and the subjects mark blocks with noticeable distortions. Similar to our work, Zhang and Wandell [1998] used a brush-painting interface for freely marking reproduction artifacts due to half-toning or JPEG compression given the reference image. The marked errors produced by 24 subjects have been pooled for each distorted image and as a result image distortion maps with the probability of error detection have been obtained. In our experiments we enable pixel-precise distortion marking, which we then average in downsampled images that are used in our analysis. This improves the quality of the data compared to the heuristic-driven pixel rejection used in [Zhang and Wandell 1998]. Unlike that study, we focus exclusively on rendering-related artifacts, and we consider both the with- and no-reference experiment scenarios.

In this work we extend our dataset [Herzog et al. 2012], which consists of 10 stimuli exhibiting mostly supra-threshold distortions for 3 selected distortion types, with 27 new stimuli (refer to Table 1 and the supplementary material for a more detailed summary of both datasets). The new stimuli exhibit sub-threshold, near-threshold, and supra-threshold distortions, which are often present in a single image. In comparison to that previous work, the new dataset reduces the subject learning effect by mixing different types of distortions within a single image, restricting their appearance to randomly selected parts of an image and increasing the number of distortion types to 12. This also let us test the metrics in more challenging scenarios, where the distortions are non-uniformly distributed across an image. Moreover, while the previous dataset contained mostly well visible distortions, the new images contain also low amplitude distortions, which are near the visibility threshold. The new dataset reinforces the quality and robustness of the subjective data, which is achieved by stabilizing the distance to the screen using a chin-rest and involving a large number of observers (35).

3 Localized image distortion experiment

The goal of the study is to mark areas in the images that contain *noticeable* distortions and those that contain *objectionable* distortions. The former will let us test how well the IQM predict visibility, while the latter can tell how robust the metrics are to image modifications that are not perceived as distortions. In addition, the analysis of the experimental data alone, for both visible and detectable thresholds, can reveal which image differences are seen as disturbing and which are most likely ignored or interpreted as a part of the original

Scene	Distortion Type	Mask	Method (Ref.)	Tonem.	Settings Artifact (Ref.)
#1 Apartment	VPL Clamp.	no	IGI (LC)	[Rein.]	$0.1 \cdot 10^6$ vpls ($2 \cdot 10^6$ vpls)
#2 CG Figures	Struc-Noise	no	IGI (PT)	[Drag.]	10^6 vpls (97K spp)
#3 Disney	Struc-Noise	no	IGI (PT)	[Drag.]	10^6 vpls (43K spp)
#4 Kitchen	Struc-Noise	no	IGI (PT)	[Drag.]	10^6 vpls (380K spp)
#5 Red Kitchen	Struc-Noise	no	IGI (PT)	[Drag.]	10^6 vpls (200K spp)
#6 Sponza Above T.	Alias. (Shadow)	no	GL (GL)	[Rein.]	Shadowmap-pcf 1024^2 (4096^2)
#7 Sponza Arches	Alias. (Shadow)	no	GL (GL)	[Rein.]	Shadowmap-pcf 1024^2 (4096^2)
#8 Sponza Atrium	Alias. (Shadow)	no	GL (GL)	[Rein.]	Shadowmap 1024^2 (4096^2)
#9 Sponza Tree Shad.	Alias. (Shadow)	no	GL (GL)	[Rein.]	Shadowmap 1024^2 (4096^2)
#10 Sponza Trees	VPL Clamp.	no	IGI (LC)	[Rein.]	$60 \cdot 10^3$ vpls ($2 \cdot 10^6$ vpls)
#11 Apartment II	Struc-Noise	yes	RC,RC (PPM)	[Rein.]	RC+PM phot.: $0.5 \cdot 10^6$ ($3 \cdot 10^9$)
#12 Atrium	Struc-Noise	yes	PPM (PPM)	[Rein.]	$5 \cdot 10^9$ ($20 \cdot 10^9$) photons
#13 Bathroom	Noise	yes	PPM,PPM (PPM)	[Rein.]	custom renderer
#14 Buddha	Tonemap (Halo/Bright.)	yes	[Fat.'02] (γ=3.0)	–	PBRT / pfstools
#15 Chairs	High/Med-freq Noise	yes	MCRT (MCRT)	γ=2.2	backward RT [ITBT/Inspirer]
#16 City-d	Alias. (Downsampl.)	yes	NN (–)	γ=2.2	PBRT / Matlab
#17 City-u	Upsampl. (Lanczos)	yes	NN,Lanczos (–)	γ=2.2	PBRT / Matlab
#18 Cornell	Alias./Struc-Noise	yes	RC (RC)	γ=1.8	PBRT 1 spp (128 spp)
#19 Dragons	Noise	no	RC (RC)	γ=2.2	PBRT 16 spp (128 spp)
#20 Hall	Brightness	no	MCRT (MCRT)	γ=2.2	backward RT [ITBT/Inspirer]
#21 Icido	Struc-Noise	yes	RC (PPM)	[Rein.]	RC+LC vpls: $0.5 \cdot 10^6$ ($3 \cdot 10^9$)
#22 Kitchen II	Struc-Noise/Bright.	yes	RC (PPM)	[Drag.]	RC+PM phot.: $0.5 \cdot 10^6$ ($3 \cdot 10^9$)
#23 Livingroom	Noise	yes	PPM,PPM (PPM)	[Rein.]	custom renderer
#24 MPII	Tonemap. (Grad.)	yes	[Man.'06] (γ=4.5)	–	PBRT / pfstools
#25 Plants-d	Alias. (Downsampl.)	yes	NN (–)	γ=2.2	PBRT / Matlab
#26 Plants-u	Upsampl. (Lanczos)	yes	NN,Lanczos (–)	γ=2.2	PBRT / Matlab
#27 Room Teapot	Struc-Noise	yes	RC (RC)	γ=2.2	PBRT
#28 Sala	Struc-Noise	no	RC (PPM)	[Drag.]	RC+PM phot.: $0.5 \cdot 10^6$ ($5 \cdot 10^9$)
#29 Sanmiguel	Aliasing/Bright	yes	RC,RC (RC)	γ=2.2	PBRT 1 spp (16 spp)
#30 Sanmiguel cam3	Light leaking	yes	PM (RC)	γ=2.2	PBRT
#31 Sanmiguel cam4	Alias./Struc-N./Bright.	yes	RC,RC (RC)	γ=2.2	PBRT 1 spp (16 spp)
#32 Sibenik	VPL Clamp.	no	RC (PPM)	[Rein.]	RC+LC vpls: $0.5 \cdot 10^6$ ($2 \cdot 10^9$)
#33 Sponza	Light leaking	no	PM (RC)	γ=1.8	PBRT
#34 TT	Alias./Noise/Struc-N.	yes	RC,RC (RC)	γ=2.2	PBRT 1 spp (16 spp)
#35 Villa cam1	Noise/Struc-Noise	yes	RC,RC (PM)	[Man.]	PBRT
#36 Villa cam2	Alias./Struc-N./Bright.	yes	RC (PM)	[Man.]	PBRT
#37 Villa cam3	Struc-Noise	yes	RC (PM)	[Man.]	PBRT

Table 1: *Our dataset, from left to right: the scene identifier, distortion type(s), if manually blended by a mask, the rendering method (reference algorithm and settings in parenthesis), tone mapping, and the relevant rendering parameters (if known) used to generate our image dataset (e.g., Fig. 2). The tone mapping operators Fat., Rein., Drag., Man., Man.'06 correspond to [Fattal et al. 2002], global version of [Reinhard et al. 2002], [Drago et al. 2003], [Mantiuk et al. 2008], [Mantiuk et al. 2006], respectively. GL stands for an OpenGL based deferred-renderer using shadow maps with percentage closer filtering (PCF). IGI is an instant global illumination renderer, which supports glossy virtual point lights (VPLs). RC stands for irradiance or radiance caching either in combination with photon-maps (RC+PM) [Křivánek et al. 2005] or lightcuts (RC+LC) [Herzog et al. 2009]. The reference solutions are computed either by pathtracing (PT) or bidirectional pathtracing (Bi-PT) with a constant number of samples per pixel (spp), the lightcuts algorithm (LC) [Walter et al. 2005] with 1% error threshold, or progressive photon mapping (PPM) [Hachisuka et al. 2008]. Some images were blended with artifacts of two different strengths or types, which is indicated by the comma-separated method.*

scene. In the following section we describe the design, procedure, and results of the perceptual experiment that we conducted to gather subjective labeling of artifacts in rendered images.

3.1 Stimuli

Table 1 summarizes the rendering algorithms and the distortions that were introduced to the images. Stimuli #1 – #10 come from our previous *EG'12 dataset* [Herzog et al. 2012], while in this work we performed a similar but more extensive experiment (stimuli #11 – #37) in a more rigorous setup. The key differences between the datasets are outlined in the Section 1, and they are further discussed in the supplementary material.

All scenes were rendered into high-dynamic-range images and tone mapped for display as indicated in Table 1. Each scene was rendered using a high- and low-quality setting. In some cases a few distortions of different character were introduced by varying different rendering parameters. Finally, the high quality image was in some cases manually blended (column *Mask* in Table 1) with the corresponding low quality image to reveal the distortions in randomized regions. This additional level of randomness was necessary, as many distortions appeared consistently either in low-illuminated parts of the scene or near the edges. Without blending, the observers were likely to learn the typical locations for a particular artifact and mark them regardless whether the artifact was noticeable/objectionable or not. Some test scenes were blended with more than one distorted image to contain distortions of very different character (in those cases more than one *Method* appears in Table 1). This was meant to test whether a metric can handle a mixture of heterogeneous distortions and account for their impact on image quality.

We now briefly summarize the distortions we have encountered in various rendering algorithms, which are also listed in Table 1. We restrict ourselves to typical global illumination (GI) related artifacts and do not cover banding, tessellation, shadow bias or other more specific artifacts that mostly arise in real-time rendering. For more details about the nature of the individual rendering-specific artifacts we refer the interested reader to our supplementary material. Furthermore, for the later analysis and readability we manually clustered the numerous distortion types into one of six distortion categories which share a similar subjective appearance.

High-frequency noise is probably the most common error encountered in photo-realistically rendered images, which arises as a by-product of all random sample-based integration techniques (e.g., path-tracing, progressive photon mapping [Hachisuka et al. 2008]).
Structured noise represents the class of distortions with correlated pixel errors, which exhibit both noise *and* bias. These are for example interpolation and caching artifacts commonly encountered in approximate GI algorithms such as photon mapping [Jensen 2001], instant radiosity [Keller 1997] as well as the popular (ir-)radiance caching algorithm [Ward et al. 1988; Křivánek et al. 2005].
VPL clamping and light leaking: approximate GI algorithms systematically introduce local errors, often even intentionally, in order to hide the more visually disturbing artifacts (noise). *VPL clamping* in instant radiosity and *light leaking* in photon mapping and irradiance caching fall into this category.
Brightness: another distortion we have noticed is a consistent change in brightness in large regions of an image. Reasons for this can be of systematic nature (e.g., wrong normalization, incorrect material usage) or approximative nature (e.g., only one-bounce indirect light, no caustics, only diffuse VPLs are computed).
Aliasing is the result of insufficient super-sampling or missing prefiltering during rendering. Our examples comprise aliasing in synthetic images including shadow maps, which we partially generated by downsampling the reference image followed by upsampling to the original resolution using the nearest neighbors approach.
Tone mapping can introduce disturbing artifacts, in particular if local gradient-based tone mapping operators (TMO) are applied. Therefore, we included examples of two gradient TMOs into our test set: typical halo artifacts appear in the *buddha* (#14) scene (refer to Fig. 1) [Fattal et al. 2002], and characteristic gradient "leaking" is demonstrated in the *mpii* (#24) scene [Mantiuk et al. 2006].

3.2 Participants and apparatus

A total of 35 observers (11 females and 24 males; age 19 to 52 years) took part in our experiments, and 21 of them completed the no-reference followed by with-reference sessions. The first group of 17 observers consisted of computer graphics students and researchers (denoted as *Experts* in the further analysis), while 18 observers were naïve to the field of computer graphics (denoted as *Non-experts*). All observers had normal or corrected-to-normal vision, and they were naïve as to the purpose of the experiment.

The evaluated images were displayed on two characterized and calibrated displays: 1) LCD Barco Coronis MDCC 3120 DL display (10-bit, 21-inch, 2048 $\times$ 1536 pixels), and 2) NEC MultiSync PA241W display (10-bit, 24-inch, 1920 $\times$ 1200 pixels). The calibration was performed using the X-Rite i1 Display Pro colorimeter (to D65, 120 $\mathrm{cd/m^2}$, colorimetric characterization by means of measured ICC profiles). The experimentation room was neutrally painted, darkened (measured light level: 2 lux), and the observers sat: 1) 71 cm from the Barco display, and 2) 92 cm from the NEC display, which corresponds to 60 pixels per visual degree. The observing distance was enforced by using a chin-rest.

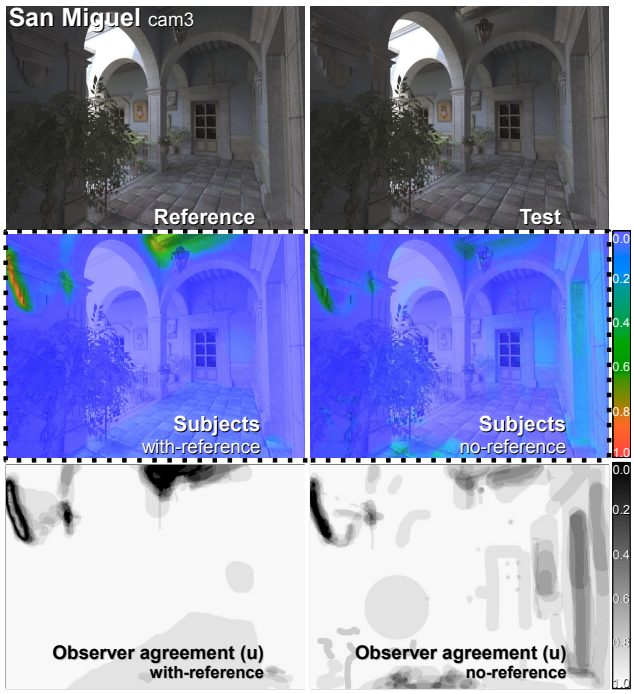


Figure 2: *Example reference and distorted images from our test set along with the mean observer data and maps of Kendall's u for both experiments. (Please refer to supplementary material for all the images.)*

3.3 Experimental procedure

We performed two experiments: in the first (*no-reference*) experiment, the observers saw only a distorted image exhibiting rendering artifacts, while in the second (*with-reference*) experiment the distorted image was presented next to the high-quality reference-image. In each experiment the sequence of images was randomized. We asked the observers to freely mark the image regions where they could see artifacts using a custom brush-paint interface. The brush size could be reduced up to per-pixel resolution by the user.

Each observer was introduced to the problem before the experiment as follows. In the no-reference experiment, the observers were instructed to label all the areas in the image with objectionable distortions. In the with-reference experiment, the observers were asked to mark those regions in the distorted image, where they could notice any differences with respect to the reference image. Each experiment took on average 30 minutes per observer. Note that the subjective distortion maps for images #1 – #10 were taken from our previous dataset [Herzog et al. 2012].

4 Analysis of subjective data

In this section we show that the data indicates high agreement between observers, giving evidence that the experimental method is reliable. Then, we analyze differences between the with-reference and no-reference experiments.

4.1 Inter-observer agreement

The experimental task of marking distortions seems challenging, especially in the no-reference setup, so the variations between observers are expected to be high. If the task is deemed to be impossible, we can expect to see little agreement in the distortion maps produced by individual observers. To test the inter-observer agreement, we compute *Kendall's coefficient of agreement* (u) per pixel [Salkind 2007]. The coefficient u ranges from $u = -1/(o-1)$, which indicates no agreement between o observers, to $u = 1$ indicating that all observers responded the same. An example of such a map of coefficients for the *sanmiguel_cam3* (#30) scene is shown in Fig. 2. The complete set of per-scene coefficients can be found in the supplementary materials.

To get an overall indicator of agreement, an average coefficient $\overline{u}$, is computed for each scene. Such overall coefficient is skewed toward very high values because most pixels did not contain any distortion and were consistently left unmarked by all observers. Therefore, we also compute a more conservative measure $\overline{u}_{mask}$, which is equal to the average u of only those pixels that were marked as distorted by at least 5% of the observers.

The values of $\overline{u}$ and $\overline{u}_{mask}$ averaged across the scenes were 0.78 and 0.41 for the with-reference experiment, and 0.77 and 0.49 for the no-reference experiment. These values are relatively high as compared to the values typically reported in such experiments. For example, Ledda et al. [2005] reported u between 0.05 and 0.43 for the task of pairwise comparison of tone mapping operators. This let us believe that the observers can reliably perform the distortion marking task even without much experience or knowledge of the underlying distortions.

4.2 With-reference and no-reference experiments

The main motivation for two experimental designs was to study the relationship between *noticeable* (the with-reference experiment) and *objectionable* distortions (the no-reference experiment). Fig. 3 shows the correlation of the probabilities of marking distortions for both experiment designs. The Spearman correlation values are very high: 0.88 for EG'12 and 0.85 for the new dataset, though these values can be biased by a larger size of unmarked regions. Such strong correlation is a further evidence that the task is well defined and, even in the no-reference experiment, the observers perform consistently and detect most distortions they would detect in the with-reference experiment. The regression line for our dataset in Fig. 3 indicates that fewer observers are marking the same distortions in the no-reference experiment.

To get further insight, we analyze differences in individual images. To find the regions that were marked systematically different between both with- and no-reference experiments, we perform the non-parametric *Kruskal-Wallis* test between the results of both experiments [Salkind 2007]. The test is run separately for each pixel, resulting in the map of p-values as visualized in Fig. 4. Note that although $p < 0.05$ should indicate that two pixels were marked statistically significantly different in the two experiments, this is only the case if each pixel is considered as an independent measurement. Given the high spatial consistency of the markings, per-pixels measurements are unlikely to be independent. However, such

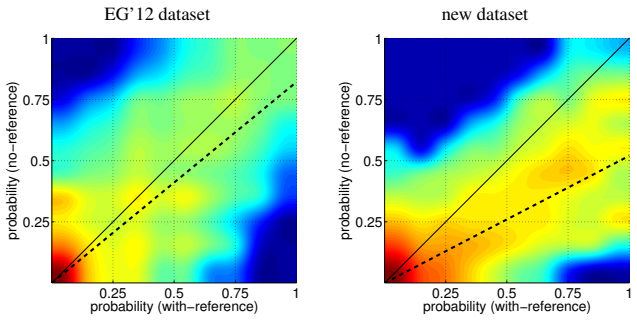


Figure 3: *The relation between the probability of marking a region in the with- and no-reference experiments, plotted separately for each dataset. Similar plots for individual scenes can be found in the supplementary materials. The dashed line shows a linear least-squares regression. The color map was generated from the logarithm of the joint probabilities. The results of with- and no-reference experiments are strongly correlated with fewer observers marking the same regions in the no-reference experiment.*

p-measure is a good indicator of relevant differences in the lack of a suitable statistical test for our dense pixel-based measurements.

The comparison of with- and no-reference results in Fig. 4 shows that the observers sometimes marked regions in the no-reference experiment which were left unmarked in the with-reference experiment. The *buddha* (#14) scene for example exhibits aliasing on the pedestal of the statue (marked in red in Fig. 4 (left)), which was not marked in the with-reference experiment because it was also present in the reference image. However, there were only few such cases in the entire dataset, which were all due to the imperfections of the reference image. In the majority of the cases the observers missed more differences when not seeing the reference image. For example, the brightness change in the background of the buddha statue caused by tone mapping (shown as green in Fig. 4 (left)) was seldom marked in the no-reference experiment.

The number of differences between both experiments indicates that both tasks are different. But at the same time, the high correlation

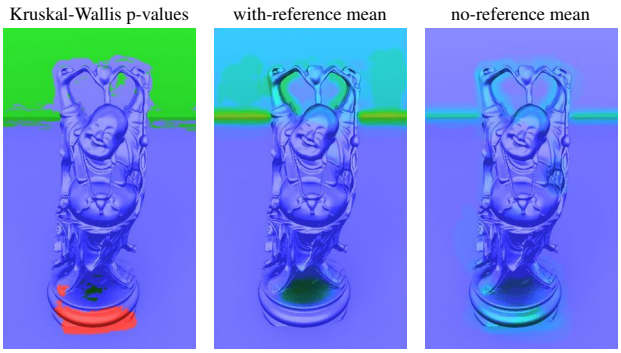


Figure 4: *Differences between with- and no-reference results for the* buddha *scene. The left image shows in green the pixels that were missed in the no-reference experiments (false negatives) and in red those that were marked despite the lack of a difference between the test and reference image (false positives). Only those areas are marked for which the p-values from the Kruskal-Wallis test is less than 0.05. Observers missed in the no-reference experiment the smooth gradient and brightness changes due to tone mapping, but marked aliasing that was also present in the reference image.*

values show that many artifacts are salient enough to be spotted in both with- and no-reference conditions. Note that both experimental designs are still *less conservative* than a typical detection measurement, which involves some form of temporal presentation of co-located test and reference images, for example by sequentially showing the test and reference images in the same screen location. Please refer to the project webpage for such presentation of the differences.

We performed a similar analysis to compare the differences between the expert and non-expert observers. However, we found only a few isolated cases in all scenes where the experts spotted more distortions, such as darkening of corners due to VPL clamping. More extensive discussions of these differences can be found in the supplementary materials.

5 Evaluation of quality metrics

In this section we investigate the performance of existing IQM in detecting distortions marked by the subjects in our experiments. At first, we justify our metric selection and briefly characterize each metric's strength. Then, we present statistical tools that we used for their performance analysis and discuss the outcome.

5.1 Image quality metric selection

Numerous IQM evaluations clearly show that it is impossible to indicate a single metric that performs steadily well for all tested stimuli [Lin and Kuo 2011; Larson and Chandler 2010]. The most problematic cases include images with spatially varying artifacts of different magnitude, as well as mixed distortion types and less common distortions [Lin and Kuo 2011]. Our dataset represents well all such difficult cases. Our choice of metrics in this study is based on the observation that metrics involving perceptual or statistical modeling perform significantly better than PSNR [Wang and Bovik 2006; Lin and Kuo 2011]. Nevertheless, because of its popularity for image quality evaluation in computer graphics, we also consider a simple *absolute difference (AD)* metric that is directly related to the commonly used RMSE and PSNR. We use absolute rather than squared differences because our statistical analysis is robust to any monotonic transformations, such as the quadratic power function.

Another popular choice in graphics is CIE-Lab, but here due to even more favorable conformance with image distortion maps [Zhang and Wandell 1998] we select its spatial extension *sCIE-Lab*. HVS-based metrics are represented by *HDR-VDP-2* [Mantiuk et al. 2011], which provides much improved predictions with respect to its predecessors HDR-VDP [Mantiuk et al. 2005] and VDP [Daly 1993]. Also, we investigate the *SSIM* that is often reported as the most reliable metric [Larson and Chandler 2010], as well as its multi-scale version *MS-SSIM* [Wang et al. 2003], which accounts for structural and contrast changes at different scales to compensate for the variations of image resolution and viewing conditions. MS-SSIM is reported as the best-performer in many IQM comparison studies [Sheikh et al. 2006; Ponomarenko et al. 2009]. Finally, we include as a metric the *Spearman rank-order correlation (sCorrel)* computed over local 8×8-pixel blocks, which can be regarded as a subset of the SSIM functionality, to better understand the importance of eliminated contrast and lightness factors.

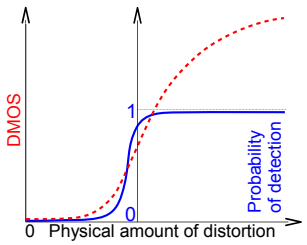
Our metric selection is also representative with respect to computational complexity. AD and sCIE-Lab are attractive due to their mathematical simplicity. On the other hand, HDR-VDP-2 is the most complex but has been shown to successfully predict near-threshold distortions. The medium complexity SSIM has been demonstrated to meaningfully estimate the magnitude of supra-threshold distortions, while its sensitivity to near threshold distor-

tions seems to be more problematic due to the lack of explicit HVS modeling. The sCIE-Lab prediction also conforms to the distortion magnitude and its sensitivity to spatial color patterns is based on the HVS-model. MS-SSIM seems to bridge the gap between SSIM and HDR-VDP-2 by emphasizing on the structural differences while processing at multi-scale.

To account for the differences in viewing distance between our two datasets, the parameters of the metrics that respect a viewing distance (HDR-VDP-2 and sCIELab) were adjusted accordingly, and for the other metrics images were resampled to match the angular resolution of 30 pixels-per-visual-degree.

5.2 Statistical measures of metric performance

It is important to recognize that, in contrast to other image quality experiments, our measurements do not capture the perceived magnitude of distortion. For that reason we need to use different measures for the metric performance. Most image quality assessment experiments measure a single scalar differential-/mean-opinion-score (DMOS/MOS) per test image, shown as the dashed red line in the plot on the right. The non-parametric correlation between a metric and the MOS values is considered as a measure of the metric's performance [Sheikh et al. 2006; Ponomarenko et al. 2009]. Unfortunately, there is no method to measure MOS efficiently for each location in an image. Our experiments capture how likely an average observer will notice *local* distortions, shown as the continuous blue line. It is correlated with MOS in the limited range where the psychometric function (blue line) does not saturate. If this probability of detection is equal or close to either 0 or 1, we have no information about the perceived magnitude of a distortion.



However, our data is well suited to benchmark the metrics ability to spot problematic regions in terms of binary classification: marking the pixels that contain noticeable or objectionable distortions. The performance of such classification is usually analyzed using the receiver-operator-characteristic (ROC) [Baldi et al. 2000]. ROC captures the relation between the size of regions that contain distortions and were correctly marked by a IQM (true positives), and the regions that do not contain distortions but were still marked (false positives). ROC captures the relation of these two quantities for a varying classification threshold. The metric that produces a larger area under the ROC curve (AUC) is assumed to perform better. To simplify considerations it is convenient to assume that a certain percentage of observers need to mark the distortion to consider it noticeable. In Fig. 5 (top-left) we present the results for regions marked by 50% or more observers, but the supplementary materials also include the data for the $\geq$25% and $\geq$75% criteria.

However, AUC values alone may give a wrong impression of the actual metric performance because usually only a small portion of the pixels in the images of our experiments showed distortions. Thus, the reference classification data is strongly unbalanced. For that reason, in addition to ROC, we also plot *Matthews correlation coefficient* [Baldi et al. 2000], which is robust to unbalanced classification data. The coefficient indicates correlation of classification data in the range from -1 to 1, where +1 represents a perfect prediction, 0 no better than random prediction, and -1 indicates total disagreement between prediction and observation.

5.3 Metric performance comparison

The key question is whether any of the IQM performs significantly better than the others in terms of detecting noticeable or objectionable graphics artifacts. The overall metric performance for both datasets and the two experimental designs is summarized in Fig. 5. Such summary, however, requires careful interpretation before any winning or loosing metric can be indicated.

Generalization of ranking. Although the ranking in Fig. 5 is a good summary of metric performance for a particular dataset, care must be taken when extrapolating any conclusions outside our measured data. To test robustness of our ranking to randomization of images, we computed the distribution of AUC by bootstrapping the set of images used for each experiment. The procedure involved computing AUC values 500 times, each time for a random set of images selected from the original set, so that the number of images was the same as in the original dataset and some of them could appear more than once (randomization with repetition) [Howell 2007, ch.18]. The computed 500 AUC values resulted in the distribution, which allowed for statistical testing. After applying Bonferroni's adjustment to compensate for multiple comparisons [Howell 2007, p.377], we found *no statistically significant differences between any pair of the metrics* in the EG'12 dataset, and *only one significant difference* between the metrics on the extreme ranking positions in the new no-reference dataset. This means that neither dataset provides conclusive evidence that any of the metrics is better than the others in a general case, and we cannot generalize the presented rankings to the entire population of images and distortions. The main reason for this is that the individual metric performance differs significantly from image to image depending on the nature of the underlying distortions. Therefore, *no* IQM is robust enough to perform significantly better for the distortions contained in our dataset.

It is important to note that our method of statistical testing differs from the methods used in other IQM comparison studies, such as [Sheikh et al. 2006] and [Ponomarenko et al. 2009]. The statistical testing employed in these studies was meant to prevent false hypothesis only due to the variance in subjective responses. The results of those statistical tests show that the ranking of the metrics is very likely to be the same for a different group of observers while assuming that the *same* set of images and distortions is used. Our testing is much more demanding as it requires the metric to perform better for any set of images (taken from the original population) in order to be considered better in the statistically sense.

Overall metric performance. Due to the unbalanced ratio of the marked and unmarked regions, we refer to *Matthews correlation coefficient* instead of the AUC values to assess the overall performance of the metrics. The average values of the Matthews coefficient for all scenes as shown in Fig. 5 are low, ranging from 0.2 to 0.35 for the EG'12 dataset, and between 0.25 and 0.45 for the new dataset. These values are much lower than Spearman's rank order correlation of 0.953 reported for the LIVE database [Sheikh et al. 2006] and 0.853 reported for the TID2008 database [Ponomarenko et al. 2009] for the best metric (MS-SSIM). However, it must be flagged that Spearman's correlation, although also scaled from -1 to 1, is different to Matthews coefficient, as discussed in Section 5.2. The low correlation values indicate that classifying distortions in the distortion maps is a much more difficult task than correlating a single value per image with the MOS. It also means that our dataset is a more demanding and accurate test for IQM since it can point out the areas where the metric's performance could be improved. Fig. 6 summarizes the Matthews correlation coefficients between the metric predictions and subjective responses in the with-reference experiment. As can be seen the highest correlation is achieved for the high-frequency-noise distortions, while for high-contrast structured

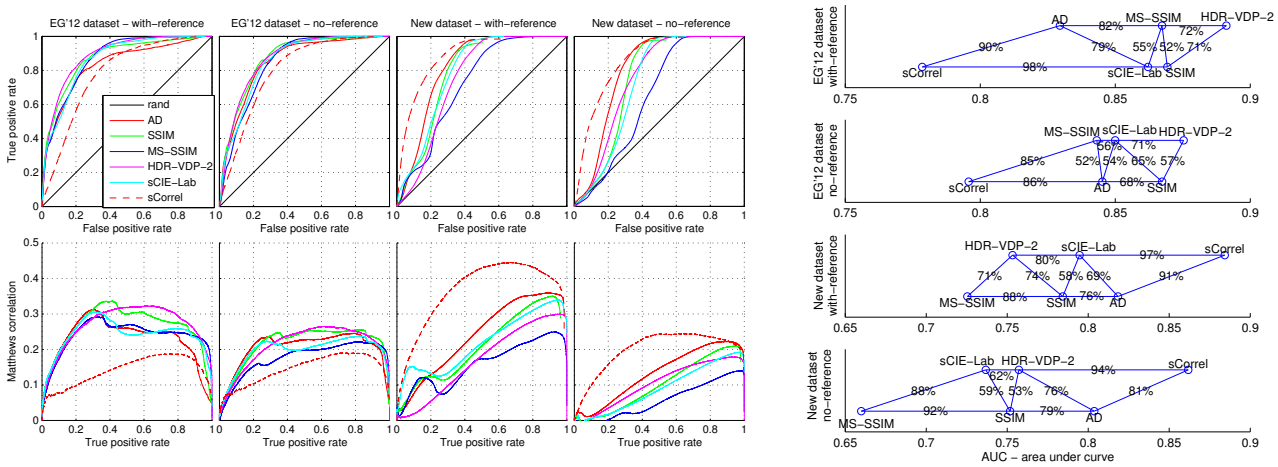


Figure 5: *The performance of quality metrics shown as ROC plots (top-left), Matthews correlation (bottom-left) and ranked according to the area-under-curve (AUC) (right) (the higher the AUC, the better the classification into distorted and undistorted regions). The percentages indicate how frequently the metric on the right results in higher AUC when the image set is randomized using a bootstrapping procedure.*

noise with only localized appearance (e.g., in the *kitchen* (#4) and *red kitchen* (#5) scenes) the correlation drops abruptly. Images with mixed distortions seem to be problematic as well.

5.4 Analysis of image quality metric failures

The ranking plots in Fig. 5 reveal different performance of the metrics for both datasets. HDR-VDP-2 performed the best for the EG'12 dataset, but was the second to the last in the new dataset. Surprisingly, the simple non-parametric correlation metric sCorrel performed the best for the new dataset, but at the same time it was the worst metric for the EG'12 dataset. This unexpected result cannot be easily explained by looking at the aggregated results and requires investigating individual images. In the following we summarize our analysis of individual images and reveal the most pronounced cases of metric failure.

Brightness and contrast change is a very common artifact of many rendering algorithms, as discussed in Section 3.1, and also the cause of failure of most advanced IQM. The best example of that is the *sala* (#28) scene shown in Fig. 7. The brightness differs significantly between the test and reference images for all surfaces, but the observers marked only the floor and in a lesser extent the walls, both affected by low-frequency noise. The noise was more visible on the floor than on the walls because the floor lacked texture and thus did not mask the noise. One metric that excelled in this task was sCorrel, with Matthews correlation exceeding 0.6. This is because non-parametric correlation is also invariant to non-linear transformations of pixel values, including low-frequency brightness changes. The second best performing metric, sCIELab, contains a band-pass model of the CSF, which attenuates low-frequency variations and thus makes this metric more robust to brightness changes. Although HDR-VDP-2 also includes a band-pass CSF model, it is far too sensitive to contrast changes to disregard numerous supra-threshold pixel modifications. Even MS-SSIM, which partially relies on the measure of correlation, did not perform much better than a random guess for this image. This shows that invariance to brightness and contrast changes must be an essential feature of any IQM that needs to reflect the observers' performance in the side-by-side comparison or non-reference tasks.

Visibility of low-contrast differences. For several scenes the test images have been computed using instant global illumination (IGI) while the reference images have been generated by path tracing, which often features certain amount of stochastic per-pixel noise. One example of such an image pair is the *disney* (#3) scene shown in Fig. 8. While the stochastic noise in a well-converged image is usually invisible, and thus remains unmarked in subjective experiments, it clearly affects the absolute pixel values and image struc-

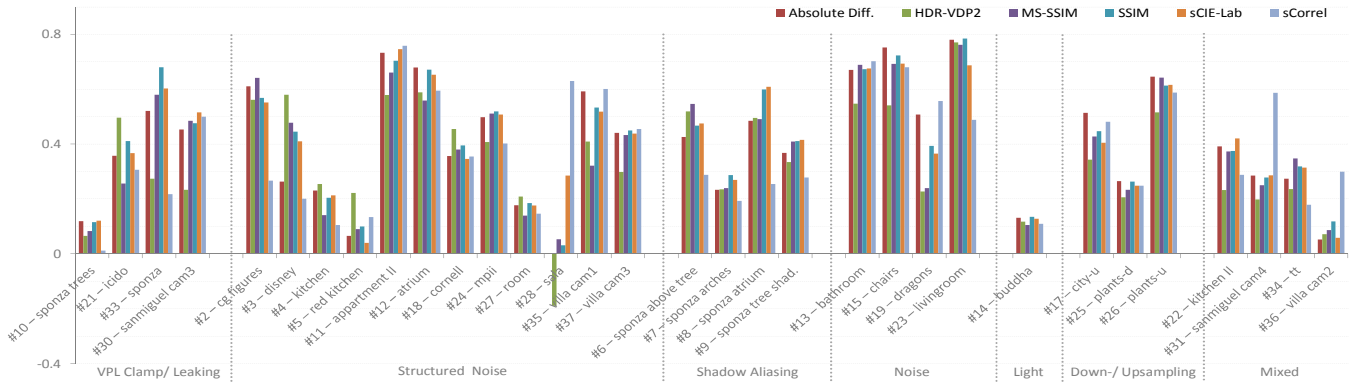


Figure 6: *Matthews correlation coefficient for predictions of HDR-VDP-2, SSIM, MS-SSIM, sCIE-Lab, sCorrel, and Absolute Difference with respect to subjective responses (with-reference experiment). Results are grouped according to the type of artifact as indicated at the bottom.*

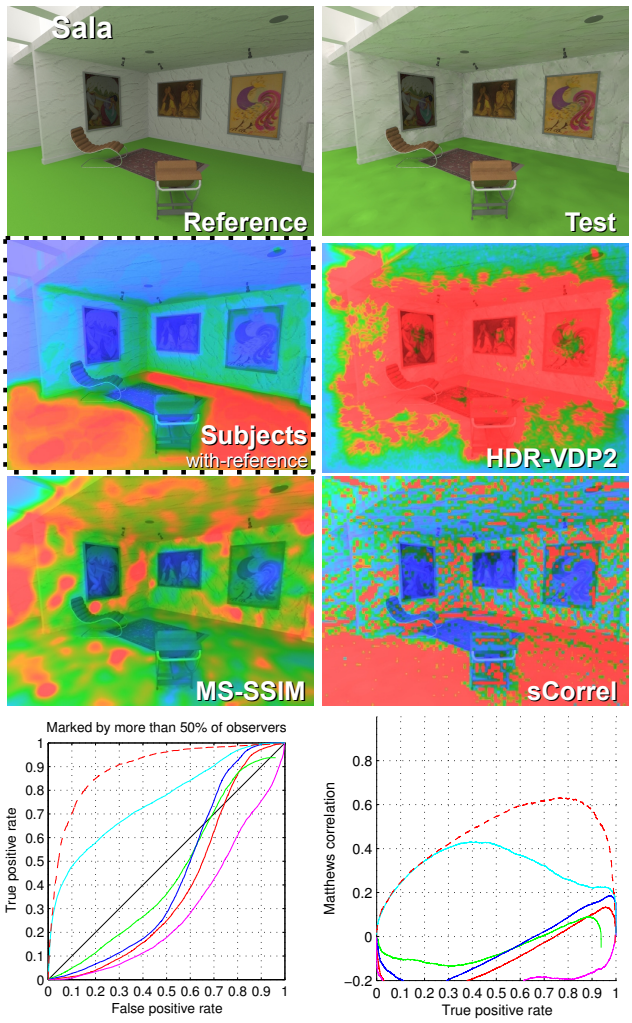


Figure 7: *Scene* sala *(top), distortion maps for selected metrics (2^{nd} and 3^{rd} rows), ROC and correlation plots (bottom). Most metrics are sensitive to brightness changes, which often remain unnoticed by observers.* sCorrel *is the only metric robust to these artifacts. Refer to the legend in Fig. 5 to check which lines correspond to which metrics in the plots.*

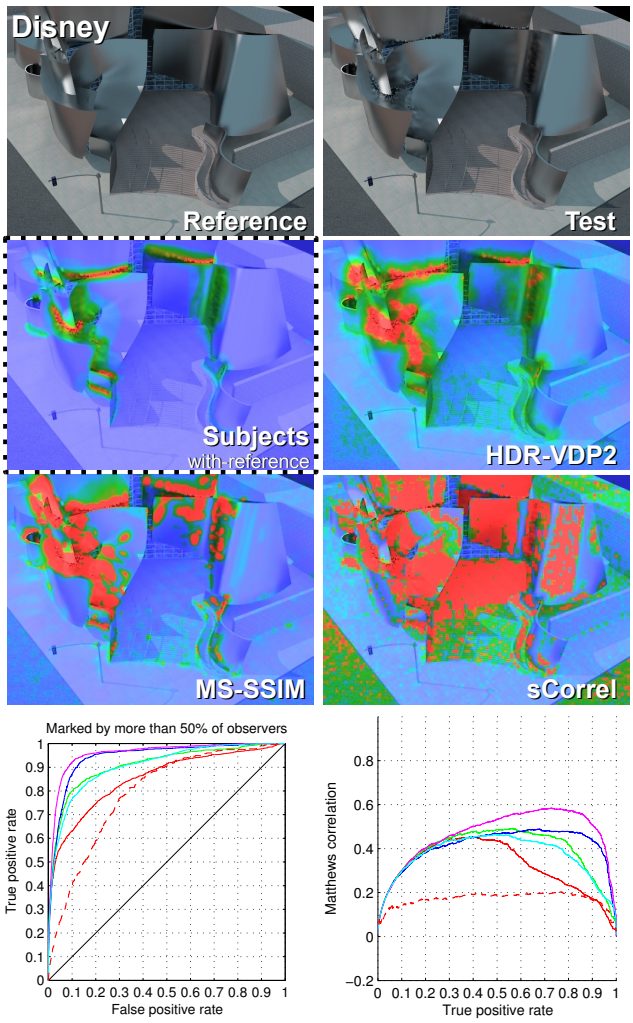


Figure 8: *Scene* disney*: simple metrics, such as sCorrel and AD, fail to distinguish between visible and invisible amount of noise resulting in worse performance.*

ture. Both AD and sCorrel metrics are sensitive to such differences, so they report distortions regardless of their visibility. What makes sCorrel insensitive to global brightness changes, makes it also insensitive to the amplitude of the noise, which prevents this metric from finding a reliable visibility threshold. For that reason both metrics poorly correlate with subjective data, as seen in the plot of Fig. 8. The metrics specifically tuned for near threshold signal detection, such as HDR-VDP-2, performed much better in this task. This stresses the importance of proper visual system modeling, which improves the metric's accuracy for the near-threshold distortions.

Plausibility of shading. A similar kind of distortion can be seen differently depending whether it leads to plausible or implausible shading. For example, two scenes shown in Fig. 9 contain VPL clamping and photon leaking distortions, respectively, near the corners. In the case of the *sponza* (#33) scene photon leaking results in brightening of dark corners. This was marked as distortion by most observers because bright patches are unlikely to be found in dark corners. However, the VPL clamping in the *sibenik* (#32) scene resulted in the opposite effect, the corners were darkened. Such distortion was marked by much fewer observers because darkening could have resulted from self-shadowing and in fact appeared realistic in the given context. All metrics failed to distinguish between these two cases. This suggests that robust IQM may require a higher-level analysis of scene and illumination that could distinguish between plausible and implausible patterns of illumination. This is difficult to achieve if images are the only source of information, but could be possible if information about the 3D scene and its shading were available [Herzog et al. 2012].

Spatial accuracy of the prediction map. Many sophisticated metrics perform often worse than the AD because they are unable to precisely localize distortions. This is well visible in the *dragons* (#19) scene shown in Fig. 10. The distortion maps for MS-SSIM show visible differences that widely disperse from the edges of the dragon figures into the background regions that do not contain any physical difference. This problem affects mostly multi-scale metrics, such as MS-SSIM and HDR-VDP, but SSIM is also affected because of its 8×8 sliding window approach, which limits the effective accuracy of the distortion map. This observation suggests that the metrics should employ techniques that respect object boundaries and thus can produce more accurate distortion maps.

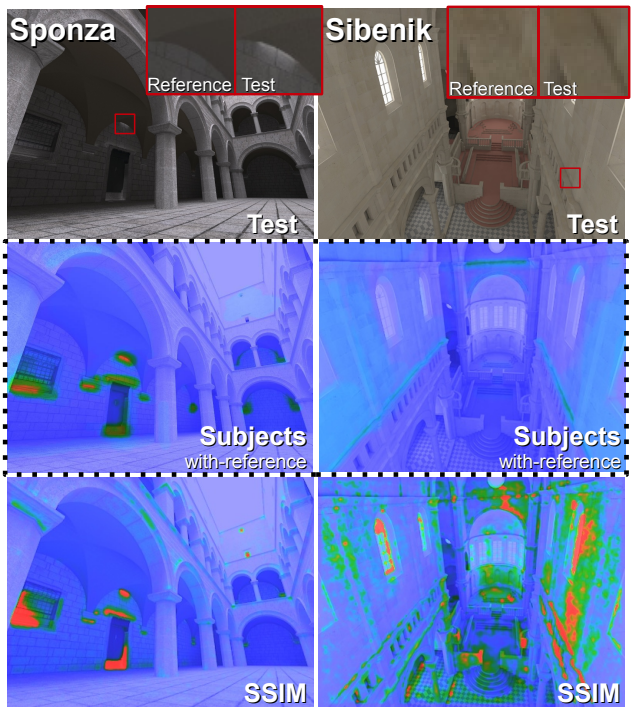


Figure 9: *Photon leaking and VPL clamping artifacts in scenes* sponza *and* sibenik *result in either brightening or darkening of corners. However, darkening is subjectively acceptable, whereas brightening leads to objectionable artifacts.*

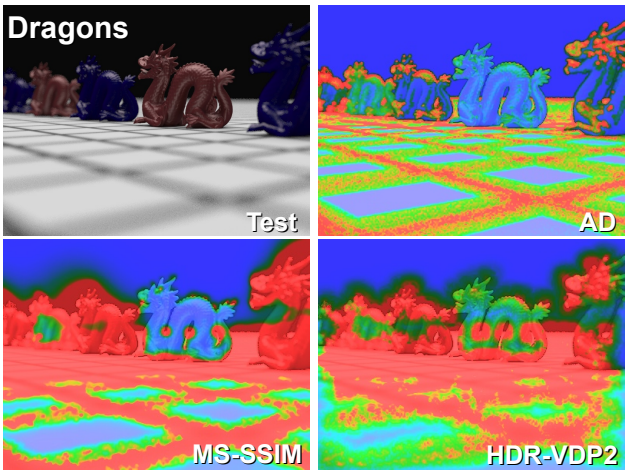


Figure 10: Dragons *scene contains artifacts on the dragon figures but not in the black background. Multi-scale IQM, such as MS-SSIM and HDR-VDP-2, mark much larger regions due to the differences detected at lower spatial frequencies. Pixel-based AD can better localize distortions in this case.*

6 Conclusions and future work

In this work we propose rendering-oriented datasets for image quality evaluation, which provide detailed distortion maps along with the probability of their detection by human observers. We show that *objectionable distortions* marked by the observers that did not see the reference image are strongly correlated in terms of their spatial location with the distortions marked in the presence of the reference image. This may suggest that by further improvement of *full-reference* IQM, we can achieve quality predictions similar to *no-reference* human judgments, which should be an easier task than the development of a *no-reference* IQM that directly mimics the human perception. *Full-reference* perceptual experiments, on the other hand, may potentially be approximated by a no-reference experiment if a reference image is not available.

For existing full-reference IQM our datasets turned out to be very demanding, and our analysis of metric failures suggests directions for improvement. The relatively good performance of the simplistic non-parametric correlation measure (sCorrel) clearly indicates its importance. Although SSIM and MS-SSIM also incorporate a correlation factor their performance is strongly influenced by their excessive sensitivity to brightness and contrast changes. Clearly, near-threshold contrast accuracy is important to disregard all non-noticeable distortions. At the same time proper spatial distortion localization is required, which is the problem for all multi-scale approaches, in particular, in the proximity of high contrast distortions. In general, the performance of state-of-the-art IQM in graphics applications is not very consistent, and one should not be too reliant on them. In particular the IQM originating in the image/video compression community may not be the most suitable for graphics applications where the artifacts are often very distinct.

We believe that all those insights are essential towards improving existing metrics or developing new ones, which we relegate as future work. Upon the public release our datasets should be useful to train such future metrics and compare their performance. However, for a systematic and quantitative study of metric failures further experiments are required.

Our datasets provide the probability of noticing distortions, which could offer interesting insights on the saliency of artifacts in rendering. Such artifact saliency could be investigated in the context of comparing a pair of images, searching for distortions within a single image, as well as task-free image inspection. Similarly to the concept of visual equivalence [Ramanarayanan et al. 2007], objectionable distortions dictate less conservative requirements on image quality, thus enabling further computational savings when used as the measure of desirable quality.

Our published datasets could also be interesting for the broader vision science community, as the complex stimuli presented in our experiments differ significantly from the usual "laboratory" ones and enable inspection of higher-level vision tasks. However, more experiments (based on photo-realistic images) are clearly needed as well as a further study of cognitive factors in the quality assessment, such as inattentional blindness or task fatigue. To this end, a speculative question raised by our results is whether it is beneficial and promising at all to model the early human vision processes (bottom-up modeling) or whether we should concentrate on data-driven approaches that are statistically trained on subjective results (top-down modeling). The bottom-up approach may result in worse than expected predictive power for complex images, while the top-down approach is prone to over-training as image quality databases will offer only very limited sample from the huge population of all potential images and distortions. This study is a step towards combining both approaches that enables training and testing the metrics of any complexity on the per-pixel basis.

Acknowledgements

We thank the creators of the test scenes used in the experiments, in particular to T. Davidovič for VPL renderings (#2, #3, #4, #5), A. Voloboy for MCRT renderings (#15, #20), I. Georgiev for bidirectional pathtracing results (#13, #23), and to the observers at Bangor University and MPII who participated in our experiments. This work was partly supported by the EPSRC research grant

EP/I006575/1 and COST Action IC1005 (STSM-100312-015783).

References

BALDI, P., BRUNAK, S., CHAUVIN, Y., ANDERSEN, C. A. F., AND NIELSEN, H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics 16*, 5 (May), 412–424.

BOLIN, M., AND MEYER, G. 1998. A perceptually based adaptive sampling algorithm. In *Proc. of SIGGRAPH*, 299–310.

ČADÍK, M., AYDIN, T. O., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2011. On evaluation of video quality metrics: an HDR dataset for computer graphics applications. In *SPIE HVEI XVI*.

CHANDLER, D., AND HEMAMI, S. 2007. VSNR: a wavelet-based visual signal-to-noise ratio for natural images. *IEEE Trans. on Image Processing 16*, 9, 2284–2298.

DALY, S. 1993. The Visible Differences Predictor: An algorithm for the assessment of image fidelity. In *Digital Images and Human Vision*, MIT Press, 179–206.

DRAGO, F., MYSZKOWSKI, K., ANNEN, T., AND N.CHIBA. 2003. Adaptive logarithmic mapping for displaying high contrast scenes. *Proc. of Eurographics 22*, 3, 419–426.

FATTAL, R., LISCHINSKI, D., AND WERMAN, M. 2002. Gradient domain high dynamic range compression. In *Proc. of SIGGRAPH*, 249–256.

HACHISUKA, T., OGAKI, S., AND JENSEN, H. W. 2008. Progressive photon mapping. In *Proc. of SIGGRAPH Asia*, 130:1–130:8.

HERZOG, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2009. Anisotropic radiance-cache splatting for efficiently computing high-quality GI with lightcuts. *Proc. of Eurographics*, 259–268.

HERZOG, R., ČADÍK, M., AYDIN, T. O., KIM, K. I., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2012. NoRM: no-reference image quality metric for realistic image synthesis. *Computer Graphics Forum 31*, 2, 545–554.

HOWELL, D. C. 2007. *Statistical Methods for Psychology*, 6th edition ed. Thomas Wadsworth.

ITU-R-BT.500-11, 2002. Methodology for the subjective assessment of the quality of television pictures.

ITU-T-P.910. 2008. Subjective audiovisual quality assessment methods for multimedia applications. Tech. rep.

JENSEN, H. W. 2001. *Realistic Image Synthesis Using Photon Mapping*. AK, Peters.

KELLER, A. 1997. Instant radiosity. In *Proc. of SIGGRAPH*, 49–56.

KŘIVÁNEK, J., GAUTRON, P., PATTANAIK, S., AND BOUATOUCH, K. 2005. Radiance caching for efficient global illumination computation. *IEEE TVCG 11*, 5, 550–561.

LARSON, E. C., AND CHANDLER, D. M. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electron. Imaging 19*, 1, 011006:1–21.

LEDDA, P., CHALMERS, A., TROSCIANKO, T., AND SEETZEN, H. 2005. Evaluation of tone mapping operators using a high dynamic range display. *Proc. of SIGGRAPH 24*, 3, 640–648.

LIN, W., AND KUO, C.-C. J. 2011. Perceptual visual quality metrics: A survey. *JVCIR*, 297–312.

LUBIN, J. 1995. *Vision Models for Target Detection and Recognition*. World Scientific, ch. A Visual Discrimination Model for Imaging System Design and Evaluation, 245–283.

MANTIUK, R., DALY, S., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2005. Predicting visible differences in high dynamic range images - model and its calibration. In *SPIE HVEI X*.

MANTIUK, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2006. A perceptual framework for contrast processing of high dynamic range images. *ACM Trans. on Applied Perception 3*, 3, 286–308.

MANTIUK, R., DALY, S., AND KEROFSKY, L. 2008. Display adaptive tone mapping. In *Proc. of SIGGRAPH*, vol. 27(3), #68.

MANTIUK, R., KIM, K. J., REMPEL, A. G., AND HEIDRICH, W. 2011. HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. In *Proc. of SIGGRAPH*, #40.

PEDERSEN, M., AND HARDEBERG, JON, Y. 2011. Full-Reference Image Quality Metrics: Classification and Evaluation. *Foundations and Trends in Computer Graphics and Vision 7*, 1, 1–80.

PONOMARENKO, N., LUKIN, V., ZELENSKY, A., EGIAZARIAN, K., CARLI, M., AND BATTISTI, F. 2009. TID2008 - A database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics 10*, 30–45.

RAMANARAYANAN, G., FERWERDA, J., WALTER, B., AND BALA, K. 2007. Visual equivalence: towards a new standard for image fidelity. In *Proc. of SIGGRAPH*, #76.

RAMASUBRAMANIAN, M., PATTANAIK, S., AND GREENBERG, D. 1999. A perceptually based physical error metric for realistic image synthesis. In *Proc. of SIGGRAPH*, 73–82.

REINHARD, E., STARK, M. M., SHIRLEY, P., AND FERWERDA, J. A. 2002. Photographic tone reproduction for digital images. In *Proc. of SIGGRAPH*, 267–276.

RUSHMEIER, H., WARD, G., PIATKO, C., SANDERS, P., AND RUST, B. 1995. Comparing real and synthetic images: some ideas about metrics. In *Rendering Techniques '95*, 82–91.

SALKIND, N., Ed. 2007. *Encyclopedia of measurement and statistics*. A Sage reference publication. SAGE, Thousand Oaks.

SHEIKH, H., SABIR, M., AND BOVIK, A. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. on Image Processing 15*, 11, 3440–3451.

WALTER, B., FERNANDEZ, S., ARBREE, A., BALA, K., DONIKIAN, M., AND GREENBERG, D. 2005. Lightcuts: A scalable approach to illumination. *Proc. of SIGGRAPH*, 1098–1107.

WANG, Z., AND BOVIK, A. C. 2006. *Modern Image Quality Assessment*. Morgan & Claypool Publishers.

WANG, Z., SIMONCELLI, E. P., AND BOVIK, A. C. 2003. Multiscale structural similarity for image quality assessment. In *Proc. IEEE Asilomar Conf. on Signals, Systems & Comp.*, 1398–1402.

WARD, G. J., RUBINSTEIN, F. M., AND CLEAR, R. D. 1988. A ray tracing solution for diffuse interreflection. In *Proc. of SIGGRAPH*, 85–92.

WU, H., AND RAO, K. 2005. *Digital Video Image Quality and Perceptual Coding*. CRC Press.

ZHANG, X., AND WANDELL, B. A. 1998. Color image fidelity metrics evaluated using image distortion maps. *Signal Proc. 70*, 3, 201 – 214.