Vis Comput (2017) 33:705–713
DOI 10.1007/s00371-017-1384-7

ORIGINAL ARTICLE

# Analysis of reported error in Monte Carlo rendered images

**Joss Whittle**[1] · **Mark W. Jones**[1] · **Rafał Mantiuk**[2]

Published online: 13 May 2017


**Abstract** Evaluating image quality in Monte Carlo rendered images is an important aspect of the rendering process as we often need to determine the relative quality between images computed using different algorithms and with varying amounts of computation. The use of a gold-standard, reference image, or ground truth is a common method to provide a baseline with which to compare experimental results. We show that if not chosen carefully, the quality of reference images used for image quality assessment can skew results leading to significant misreporting of error. We present an analysis of error in Monte Carlo rendered images and discuss practices to avoid or be aware of when designing an experiment.

**Keywords** Image quality assessment · Error metric · Monte Carlo rendering

## 1 Introduction

Monte Carlo rendering algorithms [25] allow for a plethora of photo-realistic and physically based lighting phenomena to be simulated, such as indirect illumination, depth of field, participating media, caustics, and physically based materials. A major problem is slow convergence, and early termination of rendering can leave a large amount of undesirable noise in the images. Many methods have been proposed over the last three decades that attempt to minimize noise using as few samples as possible. These can be roughly classified into path space methods [11,14,19,20,22,24,28,30,31,48] that use extra information available within the renderer to guide sampling in path space and image filtering methods [4,15,18,23,26,27,29,34,36,37,43,44] that attempt to reconstruct the GT from a coarse un-converged image.

New methods need to be evaluated relative to existing ones. Often the increase in quality is not clear cut and is dependent on the test scenes used; while a strong improvement can be observed for suitable scenes, it may be that others are ill-suited. This can cause the relative improvement in image quality to be small, though important none the less. In these cases where small improvements in quality are used to justify a method's performance, the accuracy of these measurements is important.

A commonly accepted methodology for evaluating images is I: to use a known GT which is noise-free; II: that comparisons between the GT and test images use a metric such as mean absolute error (MAE), mean square error (MSE), peak signal-to-noise ratio (PSNR), or more recently Structural Similarity Index (SSIM) [52]; and III: usually equal time and/or equal quality comparisons are reported as results. For these types of metrics to be effective, it is a requirement that the reference image is correct and noise-free.

In this paper, we present analysis of error reported when evaluating Monte Carlo rendered images. We look at the impact of reference image quality on results reported by IQA and highlight practices surrounding sample sets.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00371-017-1384-7) contains supplementary material, which is available to authorized users.

✉ Mark W. Jones
m.w.jones@Swansea.ac.uk

Joss Whittle
csjoss@swansea.ac.uk

Rafał Mantiuk
rafal.mantiuk@cl.cam.ac.uk

[1] Swansea University, Swansea, UK

[2] Cambridge University, Cambridge, UK

## 2 Image quality assessment (IQA)

Thorough analysis of 26 distance metrics applied to image data under varying distortions, spanning from pixel divergence methods such as MSE to those based on pixel correlation, structural features, and spectral measures [2] concluded that MSE most accurately described the level of distortion in images containing additive white noise; while for structural distortions such as blurring or block artefacts measures based on edge similarity or weighted by models of the HVS were more robust. A similar study based on how closely different IQA compare to scores given by human test subjects was conducted [38], with results indicating that the MSE-based metrics can achieve comparable performance to more complex algorithms when images are distorted by additive white noise. From the literature, while there are differing opinions on its effectiveness, image quality metrics based on MSE appear to be most common and trusted when evaluating images corrupted predominantly by additive noise.

Relative quality in error assessment of the MSE and MAE metrics was investigated [55], showing that MSE's nonlinear weighting with divergence can potentially lead to an exaggerated interpretation of error. Recent work [8] has argued that MSE is in fact preferable over MAE when the error distribution is expected to fit a Gaussian model.

Multi-scale geometric analysis (MGA) works by decomposing image signals into subbands of spatial frequency [17]. In the IQA literature, many MGA methods are used to extract structural information from input images. For the IQA considered in this work, MGA appeared repeatedly in the form of Gaussian and Laplacian pyramids [7], steerable pyramids [42], contrast pyramids [46], wavelet transformations [10], the contourlet transformation [13], and the wavelet-based contourlet transformation [49].

Study of the human visual system (HVS) has led to the creation of models that attempt to describe the likelihood that numerical distortions are actually perceivable by human observers under generalized viewing conditions. These models vary from a simple linear weighting of features in a multi-component error measure [52,54] to models based on a nonlinear contrast sensitivity function [33] applied at multiple scales in an MGA decomposition.

Universal Quality Index (UQI) [50] splits image comparison into luminance, contrast, and structural components using statistics over the local neighbourhoods of each pixel. SSIM [52] extends this idea by applying a linear weighting to each component using values derived from the HVS. The size of neighbourhood used in SSIM can alter its effectiveness at evaluating image quality; Multi-Scale Structural Similarity Index (MS-SSIM) [54] addresses this by applying SSIM to each level of a Gaussian pyramid decomposition of images. Further discussion of the drawbacks of MSE-based approaches compared to structural measures such as SSIM [51] shows that in many cases the same MSE score can be achieved for distorted images that are given vastly different quality assessments when viewed by human observers. In such cases, measures that consider structural features were significantly more robust and closely matched the assessments of human observers. More recently, an analysis of the mathematical properties of SSIM (and IQA based on it) compared to MSE derivative metrics showed they share several desirable qualities which make them well suited in the areas of parameter optimization and transform domain noise reduction [6].

Information Weighting provides an interesting extension on several existing image metrics by applying a non-uniform weighting scheme to the pooling stage of IQA [53]. An information map is computed at each pixel that represents its relative importance with respect to visually perceivable distortions in the input. This is performed at multiple scales in a Laplacian pyramid decomposition of the input image. The resulting IW-MSE and IW-PSNR metrics perform comparably with several advanced IQA algorithms that take properties of the HVS into account. A third metric that benefits from information content weighting is IW-SSIM which extends the MS-SSIM algorithm making it an IQA that takes multi-scale and HVS information into account during both the distortion and pooling stages.

Visual signal-to-noise ratio (VSNR) applies knowledge of the HVS to determine if image distortions would be noticeable to a human observer [9]. A spatially varying threshold on visible distortion is used to quickly determine if the comparison needs additional analysis which is performed by measuring perceived contrast and global precedence of structures within the images.

Noise quality measure [12] fits input images to a HVS noise model using a contrast pyramid decomposition which has the effect of filtering out distortions the model which is not sensitive to. Conventional SNR can then be applied to the model-fitted images to provide a quality assessment.

Information fidelity criterion (IFC) [40] and visual information fidelity (VIF) [39] apply MGA by decomposing input images via the wavelet transformation. Statistics applied to the wavelet coefficients attempts to capture the mutual structural information between the inputs. By decomposing the images at multiple spatial subbands, the effects of high frequency impulse noise can be directly measured. VIF can be considered a normalized variant on IFC [5].

Recent work has been targeted at quantifying multichannel image distortions that do not present themselves when images are reduced to a single channel. FSIMc which is an extension of Feature Similarity Index (FSIM) [57] considers images in the YIQ colour space [56]. This representation allows for luminance and chrominance features to be extracted and compared independently. Structural Contrast Quality Index (SC-QI) and Structural Contrast Distortion

Metric (SC-DM) [3] perform feature extraction in the LMN colour space which has similar properties to YIQ. HDR-VDP-2 (visual difference predictor) [32] takes a different approach to multichannel image analysis by looking at the effects of inter-channel contrast masking in the sRGB colour space. The measure makes a per-pixel prediction on the likelihood a human observer would be able to detect the difference between reference and distorted images and is robust to a wide range of illumination conditions seen in natural images.

In full reference (FR) [38] IQA, input images are compared against a GT image that is known to be correct. We also include two methods categorized as reduced reference (RR) IQA in our analysis. These methods are designed with the assumption that the reference image may contain some distortions, but overall is still representative of the GT. Rather than directly measuring per-pixel deviation, these methods measure the structural similarity of images by using the distribution of features extracted by MGA decomposition. The algorithms considered are based on the contourlet transform [45] and wavelet-based contourlet transform [16], respectively.

New IQA methods are often tested against image databases such as LIVE [41] or TID2013 [35] which couple distorted images with mean opinion scores (MOS) on image quality given by human observers. In our exploration of the literature, we have not found an analysis of how these algorithms (both FR and RR) perform when the reference image being used is the product of an un-converged rendering process, still containing impulse noise. We provide an extensive analysis here.

## 3 Computing error

To compute an error value for a given image, it is compared to a GT that is known to be completely noise-free. In computer graphics, error metrics that operate on single-channel (grayscale) images are most widely used in the literature with more recent research working to create IQA measures that operate on multichannel images. To extend single-channel IQA metrics to multichannel (RGB) images, the luminosity [1] of the RGB values is often used for error evaluation (Eq. 1). In this paper, all single-channel IQA are performed on the luminosity channel of images.

$$\mathcal{L} = (0.2989 \cdot r) + (0.587 \cdot g) + (0.114 \cdot b) \tag{1}$$

While IQA measures can use a large variety of methods to compare image similarity, they generally follow a two-stage design pattern. In the first stage, a distortion map is computed by comparing images at each pixel or more generally at a local region around each pixel. Methods can use pixel divergence, structural similarity, statistical models for perceivable difference, or combinations of these and other measures. A secondary pooling stage then consolidates this information to a single representative value which most often takes the form of an average across image space, sometimes weighted further by additional perceptual information based on the HVS.

Other IQA based on natural image statistics leverage decompositions such as the wavelet transformation are more abstract in that image similarity is not compared on a per-pixel basis, but rather on an overall statistical measure of mutual information encoded by the decomposition coefficients.

In our experiment, we chose IQA based on both of the above methodologies and those utilizing a variety of measures on per-pixel distortion to see how these various methods cope under the condition of a degrading and possibly non-representative reference images.

The metrics considered are: (single-channel IQA) MSE, MAE, PSNR, VSNR [9], NQM[12], VIF [39], UQI [50], SSIM [52], MS-SSIM [54], IW-MSE, IW-PSNR, IW-SSIM [53], contourlet [45] and WBCT [16] IQA, IFC [40], FSIM [57]; (multichannel IQA) FSIMc [57], SC-QI [3], SC-DM [3], and HDR-VDP-2 [32].

## 4 Our experiment

Our experiment is motivated by practices we review in the literature. When examining reference images in some literature, we still see impulse noise, and we wish to explore the effect that reference image quality has on the results reported by IQA. Initially, we performed our analysis on images rendered with a bespoke path tracing software developed for our research. We then validated our experiment using the widely trusted Mitsuba Renderer [21], which are the data we show in this work.

We constructed an experiment where test scenes (Fig. 1) were rendered to increasing numbers of independent samples using each of the rendering algorithms considered. Images were generated on a $2^n$ sample per-pixel (*s.p.p.*) sequence $\mathcal{N} \in \mathbb{N} : \{2^n | 2 \leq n \leq \cdots\}$ for each of the test algorithms $\mathcal{A} \in \mathbb{A} : \{PT, BDPT, PSSMLT, MLT, Manifold\text{-}MLT, ERPT, Manifold\text{-}ERPT\}$ and for each scene $\mathcal{S} \in \mathbb{S} : \{Cornell\ Box, Torus, Veach\ Bidir, Veach\ Door, Sponza\}$. This defines a set of images $\mathcal{I}_{\mathcal{SAN}}$ where $(\mathcal{SAN}) \in (\mathbb{S} \times \mathbb{A} \times \mathbb{N})$ parameterized by scene, rendering algorithm, and sample count with which to perform our analysis. For each scene, we chose a rendering algorithm $\mathcal{A}^G$ to be the reference algorithm based upon its rate of convergence and the lack of structural artefacts at low sample counts. Path tracing was chosen as the reference algorithm for the Cornell Box and Sponza scenes, while the caustic illumination in the Torus,

**Fig. 1** Scenes used for error analysis. From left to right: cornell box, torus, veach bidir, veach door, sponza

Veach Bidir, and Veach Door scenes was better sampled using bidirectional path tracing.

For each error metric $\mathcal{E} \in \mathbb{E} : \{MSE, MAE, PSNR, UQI, SSIM, MS\text{-}SSIM, IW\text{-}SSIM, IW\text{-}MSE, IW\text{-}PSNR, VSNR, contourlet, WBCT, NQM, VIF, IFC, FSIM, FSIMc, HDR\text{-}VDP-2, SC\text{-}QI, SC\text{-}DM\}$ we compute the true error values to the GT reference image, and we wish to see how degrading the quality of the reference image affects these true error scores. To do this, we select the next highest sampled image as the reference image and recompute the error values. Only images with lower sample counts than the currently selected reference image are computed. By repeating this for all images in the sequence of the reference algorithm, we end up with a triangular matrix for each error metric, algorithm, and scene, where one row represents the true error values, and the remaining rows represent the error values as the reference image is degraded. Formally, for all configurations $\mathcal{C}$ of an error metric, scene, and rendering algorithm we have a lower triangular matrix $\mathcal{M}^{\mathcal{C}}$ with elements indexed by the number of samples in the test image $\mathcal{N}_j$ and in the reference image $\mathcal{N}_i$, where each element is the error calculated between the reference image $\mathcal{I}_{\mathcal{SA}^G\mathcal{N}_i}$ and the test image $\mathcal{I}_{\mathcal{SAN}_j}$ using an error metric $\mathcal{E}$ (Eq. 2).

$$\mathcal{M}^{\mathcal{C}}_{i,j} = \mathcal{E}\left(\mathcal{I}_{\mathcal{SA}^G\mathcal{N}_i}, \mathcal{I}_{\mathcal{SAN}_j}\right) \quad \text{where } i > j \text{ and } \mathcal{C} = (\mathcal{ESA}) \;\; \forall \mathcal{C} \in (\mathbb{E} \times \mathbb{S} \times \mathbb{A}) \tag{2}$$

To compare the degraded error values to the true values, we use $\ln \mathcal{Q}$ [47] which measures the difference between an observed and expected value. We chose $\ln \mathcal{Q}$ because, like per cent error, it is a measure of relative change that can be used to compare metrics which operate on different scales, and because it is symmetric between positive and negative values which occur frequently within our data. This is applied to our triangular matrices by taking the natural logarithm of the values in each column divided by the true value in the $|\mathbb{N}|$th (bottom) row. This gives a matrix where the bottom row is zeros (referring to the $\ln \mathcal{Q}$ of true values versus themselves) and subsequent rows represent the quality of error evaluations as the reference image is degraded. Formally, from the matrix $\mathcal{M}^{\mathcal{C}}$ for each configuration in the ensemble we define an equally sized matrix $\mathcal{P}^{\mathcal{C}}$ with elements defined by Eq. 3.

$$\mathcal{P}^{\mathcal{C}}_{i,j} = \ln\left(\frac{\mathcal{M}^{\mathcal{C}}_{i,j}}{\mathcal{M}^{\mathcal{C}}_{|\mathbb{N}|,j}}\right) \quad \text{where } i > j \text{ and } \mathcal{C} = (\mathcal{ESA}) \;\; \forall \mathcal{C} \in (\mathbb{E} \times \mathbb{S} \times \mathbb{A}) \tag{3}$$

where $\mathcal{P}^{\mathcal{C}}$ has positive values and this shows the IQA under test has **overestimated** the amount of error while negative values show the error was **underestimated**.

## 5 Results

For all scenes, rendering algorithms, and error metrics, there are 735 separate $\mathcal{P}^{\mathcal{C}}$ matrices in the dataset. We present the full results in supplementary material. Tables 1a–e show $\mathcal{P}^{\mathcal{C}}$ for the Cornell Box scene rendered with bidirectional path tracing and using error metrics VIF (top), MS-SSIM, SC-QI, HDR-VDP-2, and MSE (bottom). A strong increase in values is visible for MSE, showing that overestimation increases as the number of samples in the reference image decreases to the number of samples in the test image. The increase in misreporting also appears for VIF as a strong underestimation. MS-SSIM and SC-QI also exhibit underestimation but at a significantly lower magnitude. HDR-VDP-2 shows both under- and overestimation at magnitudes comparable to VIF.

To condense this to a manageable set of results, Table 2 displays the maximum magnitude of misreporting within a defined region of each matrix. The maximum magnitude is underlined in each table of results (Table 1a–e and supplementary material). The region is defined for reference images having sufficient samples that they exhibit good visual convergence. Reference images outside this region have lower sample counts and consequently more visible noise. The higher sample reference images are representative of image comparisons that are typically seen in the literature when evaluating rendering algorithms. We signify this region in each table by a horizontal rule. For example, configuration (Cornell Box, BDPT), the maximum magnitude of misre-

**Table 1** ln $\mathcal{Q}$ of various IQA measures as reference and test image quality are varied

| BDPT | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 | 2048 | 4096 | 8192 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GT (PT) | | | | | | | | | | | | | |
| 16384 | -0.00220 | -0.00245 | -0.00226 | -0.00297 | -0.00345 | -0.00391 | -0.00436 | -0.00465 | -0.00547 | -0.00540 | -0.00597 | -0.00623 | -0.00681 |
| 8192 | -0.00615 | -0.00624 | -0.00723 | -0.00845 | -0.00899 | -0.01050 | -0.01224 | -0.01361 | -0.01478 | -0.01565 | -0.01678 | -0.01760 | |
| 4096 | -0.01403 | -0.01611 | -0.01684 | -0.01841 | -0.02120 | -0.02472 | -0.02762 | -0.03124 | -0.03331 | -0.03580 | -0.03766 | | |
| 2048 | -0.02817 | -0.02970 | -0.03269 | -0.03682 | -0.04304 | -0.04905 | -0.05489 | -0.06046 | -0.06549 | -0.06902 | | | |
| 1024 | -0.05707 | -0.06009 | -0.06457 | -0.07179 | -0.08089 | -0.09143 | -0.10328 | -0.11357 | -0.12051 | | | | |
| 512 | -0.10132 | -0.10886 | -0.11870 | -0.13149 | -0.14697 | -0.16575 | -0.18370 | -0.19860 | | | | | |
| 256 | -0.17229 | -0.18211 | -0.19971 | -0.22460 | -0.25048 | -0.27817 | -0.30856 | | | | | | |
| 128 | -0.27769 | -0.29942 | -0.32306 | -0.35713 | -0.40423 | -0.44500 | | | | | | | |
| 64 | -0.40747 | -0.44486 | -0.49189 | -0.53931 | -0.59707 | | | | | | | | |
| 32 | -0.58385 | -0.63379 | -0.69616 | -0.76560 | | | | | | | | | |
| 16 | -0.79189 | -0.86063 | -0.94063 | | | | | | | | | | |
| 8 | -1.02780 | -1.11807 | | | | | | | | | | | |
| 4 | -1.28597 | | | | | | | | | | | | |

(a) $\mathcal{P}^{\mathcal{C}}$ for Scene: [**Cornell Box**] Algorithm: [**BDPT**] Metric: [**VIF**] True GT: [**PT @ 32768 spp**].

| BDPT | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 | 2048 | 4096 | 8192 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GT (PT) | | | | | | | | | | | | | |
| 16384 | -0.00010 | -0.00000 | -0.00003 | -0.00006 | -0.00003 | -0.00008 | -0.00005 | -0.00005 | -0.00006 | -0.00005 | -0.00005 | -0.00005 | -0.00006 |
| 8192 | -0.00012 | -0.00004 | -0.00015 | -0.00014 | -0.00013 | -0.00012 | -0.00015 | -0.00016 | -0.00015 | -0.00015 | -0.00016 | -0.00016 | |
| 4096 | -0.00007 | -0.00025 | -0.00025 | -0.00027 | -0.00026 | -0.00035 | -0.00034 | -0.00036 | -0.00035 | -0.00037 | -0.00037 | | |
| 2048 | -0.00010 | -0.00033 | -0.00059 | -0.00063 | -0.00069 | -0.00076 | -0.00073 | -0.00076 | -0.00075 | -0.00076 | | | |
| 1024 | -0.00104 | -0.00105 | -0.00119 | -0.00120 | -0.00141 | -0.00133 | -0.00148 | -0.00155 | -0.00152 | | | | |
| 512 | -0.00179 | -0.00159 | -0.00252 | -0.00266 | -0.00279 | -0.00294 | -0.00305 | -0.00310 | | | | | |
| 256 | -0.00306 | -0.00351 | -0.00468 | -0.00539 | -0.00561 | -0.00583 | -0.00620 | | | | | | |
| 128 | -0.00628 | -0.00800 | -0.00909 | -0.01011 | -0.01146 | -0.01181 | | | | | | | |
| 64 | -0.01118 | -0.01467 | -0.01825 | -0.02007 | -0.02134 | | | | | | | | |
| 32 | -0.02254 | -0.02859 | -0.03358 | -0.03739 | | | | | | | | | |
| 16 | -0.04154 | -0.05122 | -0.05902 | | | | | | | | | | |
| 8 | -0.07407 | -0.08862 | | | | | | | | | | | |
| 4 | -0.12584 | | | | | | | | | | | | |

(b) $\mathcal{P}^{\mathcal{C}}$ for Scene: [**Cornell Box**] Algorithm: [**BDPT**] Metric: [**MS-SSIM**] True GT: [**PT @ 32768 spp**].

| BDPT | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 | 2048 | 4096 | 8192 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GT (PT) | | | | | | | | | | | | | |
| 16384 | 0.00001 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | -0.00000 | -0.00000 | -0.00000 | -0.00000 |
| 8192 | 0.00001 | 0.00001 | 0.00001 | 0.00001 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | -0.00000 | -0.00000 | -0.00000 | |
| 4096 | 0.00003 | 0.00002 | 0.00002 | 0.00001 | 0.00001 | 0.00001 | 0.00000 | 0.00000 | -0.00000 | -0.00000 | -0.00000 | | |
| 2048 | 0.00005 | 0.00004 | 0.00003 | 0.00002 | 0.00001 | 0.00001 | 0.00000 | 0.00000 | -0.00000 | -0.00000 | | | |
| 1024 | 0.00008 | 0.00006 | 0.00004 | 0.00003 | 0.00002 | 0.00001 | 0.00000 | -0.00000 | -0.00001 | | | | |
| 512 | 0.00011 | 0.00009 | 0.00006 | 0.00004 | 0.00002 | 0.00001 | -0.00000 | -0.00001 | | | | | |
| 256 | 0.00016 | 0.00012 | 0.00008 | 0.00005 | 0.00002 | 0.00000 | -0.00001 | | | | | | |
| 128 | 0.00022 | 0.00015 | 0.00009 | 0.00005 | 0.00001 | -0.00002 | | | | | | | |
| 64 | 0.00027 | 0.00017 | 0.00009 | 0.00003 | -0.00003 | | | | | | | | |
| 32 | 0.00032 | 0.00017 | 0.00006 | -0.00003 | | | | | | | | | |
| 16 | 0.00032 | 0.00013 | -0.00004 | | | | | | | | | | |
| 8 | 0.00024 | -0.00003 | | | | | | | | | | | |
| 4 | 0.00002 | | | | | | | | | | | | |

(c) $\mathcal{P}^{\mathcal{C}}$ for Scene: [**Cornell Box**] Algorithm: [**BDPT**] Metric: [**SC-QI**] True GT: [**PT @ 32768 spp**].

| BDPT | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 | 2048 | 4096 | 8192 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GT (PT) | | | | | | | | | | | | | |
| 16384 | 0.00069 | 0.00061 | 0.00023 | 0.00017 | 0.00070 | -0.00122 | -0.00049 | -0.00019 | -0.00438 | -0.00663 | -0.00472 | -0.01012 | -0.01582 |
| 8192 | 0.00135 | 0.00077 | 0.00099 | 0.00059 | 0.00003 | -0.00055 | -0.00104 | -0.00320 | -0.00575 | -0.00867 | -0.01418 | -0.02252 | |
| 4096 | 0.00301 | 0.00112 | 0.00253 | 0.00150 | 0.00180 | -0.00101 | -0.00394 | -0.00516 | -0.01265 | -0.02330 | -0.03381 | | |
| 2048 | 0.00523 | 0.00486 | 0.00434 | 0.00206 | 0.00025 | -0.00148 | -0.00711 | -0.01443 | -0.02351 | -0.03825 | | | |
| 1024 | 0.00934 | 0.01054 | 0.00967 | 0.00628 | 0.00390 | -0.00147 | -0.01202 | -0.02456 | -0.04561 | | | | |
| 512 | 0.02234 | 0.01974 | 0.01844 | 0.01213 | 0.00482 | -0.00383 | -0.02187 | -0.04315 | | | | | |
| 256 | 0.03743 | 0.03500 | 0.02745 | 0.02212 | 0.00886 | -0.01341 | -0.04229 | | | | | | |
| 128 | 0.06009 | 0.05272 | 0.04354 | 0.02893 | 0.00325 | -0.02340 | | | | | | | |
| 64 | 0.08230 | 0.06972 | 0.05534 | 0.03304 | -0.00370 | | | | | | | | |
| 32 | 0.10520 | 0.08631 | 0.06176 | 0.02416 | | | | | | | | | |
| 16 | 0.12428 | 0.09279 | 0.05239 | | | | | | | | | | |
| 8 | 0.13101 | 0.08409 | | | | | | | | | | | |
| 4 | 0.12523 | | | | | | | | | | | | |

(d) $\mathcal{P}^{\mathcal{C}}$ for Scene: [**Cornell Box**] Algorithm: [**BDPT**] Metric: [**HDR-VDP-2**] True GT: [**PT @ 32768 spp**].

| BDPT | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 | 2048 | 4096 | 8192 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GT (PT) | | | | | | | | | | | | | |
| 16384 | 0.00021 | 0.00044 | 0.00035 | 0.00146 | 0.00250 | 0.00604 | 0.01123 | 0.01857 | 0.03963 | 0.06006 | 0.10249 | 0.14676 | 0.20907 |
| 8192 | 0.00082 | 0.00089 | 0.00228 | 0.00439 | 0.00790 | 0.01515 | 0.03214 | 0.05903 | 0.10703 | 0.17634 | 0.27113 | 0.38223 | |
| 4096 | 0.00067 | 0.00301 | 0.00481 | 0.00984 | 0.01821 | 0.03760 | 0.07164 | 0.13536 | 0.23000 | 0.37515 | 0.54666 | | |
| 2048 | 0.00206 | 0.00420 | 0.01003 | 0.01965 | 0.04178 | 0.08150 | 0.14686 | 0.26415 | 0.44191 | 0.66779 | | | |
| 1024 | 0.00577 | 0.01141 | 0.02103 | 0.04097 | 0.08008 | 0.14905 | 0.27787 | 0.48703 | 0.75710 | | | | |
| 512 | 0.01046 | 0.02234 | 0.04352 | 0.08472 | 0.15813 | 0.29235 | 0.50361 | 0.81414 | | | | | |
| 256 | 0.02121 | 0.04110 | 0.08476 | 0.16638 | 0.30385 | 0.52198 | 0.85770 | | | | | | |
| 128 | 0.04228 | 0.08528 | 0.16436 | 0.30002 | 0.54550 | 0.87442 | | | | | | | |
| 64 | 0.08224 | 0.16066 | 0.30742 | 0.53591 | 0.88474 | | | | | | | | |
| 32 | 0.15806 | 0.29905 | 0.53027 | 0.87775 | | | | | | | | | |
| 16 | 0.29062 | 0.52153 | 0.86325 | | | | | | | | | | |
| 8 | 0.50545 | 0.84654 | | | | | | | | | | | |
| 4 | 0.81929 | | | | | | | | | | | | |

(e) $\mathcal{P}^{\mathcal{C}}$ for Scene: [**Cornell Box**] Algorithm: [**BDPT**] Metric: [**MSE**] True GT: [**PT @ 32768 spp**].

The vertical axis represents the number of *s.p.p.* in reference images while the horizontal axis denotes the number of *s.p.p.* in test images. Cells are highlighted from underestimation (blue) to overestimation (orange). The horizontal rule between 2048 and 4096 *s.p.p.* separates ground truths that exhibit good visual convergence (above) from sample counts that result in ground truths with visible noise (below). Maximum magnitude for reference images with good visual convergence is shown with a black underline. The matrix has been flipped vertically, and the zero row of reference values versus themselves has been omitted to aid in visualization

**Table 2** A condensed table of all 735 result tables, showing max $\mathcal{P}^{\mathcal{C}}$ for all algorithms, scenes, and metrics

| | | VIF | IFC | UQI | PSNR | IW-PSNR | SSIM | NQM | IW-SSIM | MS-SSIM | SC-QI | HDR-VDP-2 | FSIMc | FSIM | VSNR | SC-DM | RMSE | MAE | IW-MSE | MSE | Contourlet | WBCT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Cornell Box** | PT | -0.03370 | -0.08914 | -0.03688 | -0.02483 | -0.02508 | -0.00175 | -0.02421 | -0.00299 | -0.00038 | 0.00004 | -0.02242 | 0.00829 | 0.00830 | -0.02707 | -0.09727 | 0.14771 | 0.16635 | 0.31313 | 0.29542 | 0.19501 | 0.24858 |
| **GT (PT)** | BDPT | -0.03766 | -0.16721 | -0.05302 | -0.04385 | -0.05098 | -0.00187 | -0.03961 | -0.00198 | -0.00037 | 0.00003 | -0.03381 | 0.00747 | 0.00747 | -0.02998 | 0.45376 | 0.27333 | 0.35324 | 0.67251 | 0.54666 | 0.08481 | -0.36078 |
| | PSSMLT | -0.03556 | -0.08665 | -0.04487 | -0.01529 | -0.01522 | -0.00181 | -0.01222 | -0.00250 | -0.00037 | 0.00002 | -0.01098 | 0.00694 | 0.00694 | -0.02383 | 0.07454 | 0.08651 | 0.16461 | 0.18188 | 0.17301 | 0.15779 | 0.07700 |
| | MLT | -0.03221 | -0.06861 | -0.03109 | -0.01000 | -0.00561 | -0.00176 | -0.00550 | -0.00266 | -0.00038 | 0.00003 | -0.00911 | 0.00835 | 0.00837 | -0.01501 | -0.05995 | 0.05437 | 0.07013 | 0.06195 | 0.10875 | -0.05600 | 0.06757 |
| | Manifold-MLT | -0.03270 | -0.08660 | -0.03258 | -0.01017 | -0.00606 | -0.00177 | -0.00814 | -0.00255 | -0.00036 | 0.00003 | -0.00505 | 0.00838 | 0.00840 | -0.01564 | -0.06605 | 0.05548 | 0.07346 | 0.06705 | 0.11096 | 0.06355 | 0.06738 |
| | ERPT | -0.02270 | -0.02777 | -0.01161 | -0.00209 | -0.00177 | -0.00164 | -0.00217 | -0.00093 | -0.00035 | 0.00001 | 0.00135 | 0.00307 | 0.00306 | -0.00459 | -0.05302 | 0.00947 | 0.01039 | 0.01547 | 0.01894 | 0.04870 | 0.03389 |
| | Manifold-ERPT | -0.02187 | -0.02349 | -0.01002 | -0.00181 | -0.00157 | -0.00156 | -0.00200 | -0.00094 | -0.00034 | 0.00001 | 0.00162 | 0.00305 | 0.00305 | -0.00511 | -0.05370 | 0.00812 | 0.00894 | 0.01362 | 0.01624 | 0.04506 | 0.03296 |
| **Torus** | PT | -0.25696 | -0.11270 | -0.09195 | -0.04573 | -0.03921 | -0.02489 | -0.06889 | -0.01844 | -0.00672 | 0.00020 | 0.05888 | 0.05853 | 0.05892 | -0.11038 | -0.21958 | 0.19842 | 0.16418 | 0.32761 | 0.39684 | 0.65864 | 0.54952 |
| **GT (BDPT)** | BDPT | -0.24771 | -0.10768 | -0.09545 | -0.04234 | -0.03724 | -0.02300 | -0.05962 | -0.01684 | -0.00625 | 0.00021 | 0.06461 | 0.05934 | 0.05977 | -0.10721 | -0.23365 | 0.18291 | 0.14665 | 0.31065 | 0.36581 | 0.65733 | 0.57969 |
| | PSSMLT | -0.32934 | -0.20121 | -0.18467 | -0.04048 | -0.01267 | -0.03565 | -0.04632 | -0.01518 | -0.00748 | 0.00018 | 0.06946 | 0.01822 | 0.01833 | -0.09096 | 0.29185 | 0.17438 | 0.18690 | 0.09537 | 0.34876 | 0.61894 | 1.10642 |
| | MLT | -0.33488 | -0.21295 | -0.20203 | -0.01372 | 0.00515 | -0.03606 | 0.09016 | -0.01120 | -0.00756 | 0.00020 | 0.05198 | 0.01144 | 0.01169 | -0.07703 | 0.08907 | 0.05004 | 0.07799 | -0.01348 | 0.10008 | 0.38458 | 0.84635 |
| | Manifold-MLT | -0.35628 | -0.25590 | -0.25184 | -0.02246 | 0.00560 | -0.03818 | 0.11864 | -0.01310 | -0.00847 | 0.00015 | 0.03651 | 0.01484 | 0.01525 | -0.33361 | 0.16379 | 0.08771 | 0.12577 | 0.02145 | 0.17542 | 0.49481 | 0.88657 |
| | ERPT | -0.26809 | -0.10343 | -0.12235 | -0.03366 | -0.01447 | -0.03133 | -0.04850 | -0.01580 | -0.00740 | 0.00003 | 0.06267 | 0.01360 | 0.01367 | -0.08538 | -0.05514 | 0.14077 | 0.13070 | 0.11061 | 0.28155 | 0.58266 | 0.77866 |
| | Manifold-ERPT | -0.26922 | -0.10264 | -0.12673 | -0.03137 | -0.01457 | -0.03111 | -0.04444 | -0.01593 | -0.00755 | 0.00002 | 0.02725 | 0.01280 | 0.01289 | -0.08659 | -0.05230 | 0.12942 | 0.12117 | 0.10828 | 0.25885 | 0.61998 | 0.83238 |
| **Veach Bidir** | PT | -0.01659 | -0.04894 | -0.00333 | -0.00107 | 0.00227 | -0.00148 | 0.16618 | 0.00274 | 0.00089 | 0.00009 | 0.00748 | 0.01231 | 0.01231 | 0.01058 | -0.03357 | 0.00399 | 0.00178 | -0.01441 | 0.00798 | 0.03948 | 0.02995 |
| **GT (BDPT)** | BDPT | -0.03412 | -0.08898 | -0.04388 | -0.04562 | -0.07197 | -0.00234 | -0.04159 | 0.00196 | -0.00063 | 0.00006 | -0.08495 | 0.00817 | 0.00817 | -0.09721 | 0.37055 | 0.25648 | 0.18924 | 0.76863 | 0.51296 | -0.20462 | 0.42495 |
| | PSSMLT | -0.03463 | -0.06634 | -0.04196 | -0.02015 | -0.02454 | -0.00242 | -0.01531 | 0.00176 | -0.00060 | 0.00005 | -0.05244 | 0.00832 | 0.00832 | -0.04186 | -0.07394 | 0.10093 | 0.09001 | 0.23824 | 0.20186 | 0.18517 | 0.14017 |
| | MLT | -0.02697 | 0.03898 | -0.01720 | -0.00637 | 0.00964 | -0.00211 | 0.01242 | 0.00212 | 0.00064 | 0.00006 | 0.07547 | 0.01050 | 0.01049 | 0.01921 | -0.05900 | 0.02815 | 0.02496 | -0.06739 | 0.05631 | 0.06778 | 0.10887 |
| | Manifold-MLT | -0.02796 | 0.05100 | -0.02065 | -0.00642 | 0.00935 | -0.00226 | 0.01560 | 0.00233 | 0.00109 | 0.00006 | -0.02304 | 0.01030 | 0.01030 | -0.02151 | -0.05647 | 0.02839 | 0.02692 | -0.06480 | 0.05679 | 0.15042 | 0.09750 |
| | ERPT | -0.02107 | -0.03051 | -0.00509 | -0.00366 | 0.00636 | -0.00216 | 0.00674 | 0.00090 | -0.00061 | 0.00002 | 0.00834 | 0.00438 | 0.00438 | -0.00904 | -0.03695 | 0.01429 | 0.00472 | -0.04387 | 0.02858 | 0.07840 | 0.03530 |
| | Manifold-ERPT | -0.02088 | -0.03062 | -0.00466 | -0.00298 | 0.00571 | -0.00202 | 0.00641 | -0.00098 | -0.00062 | 0.00002 | -0.00631 | 0.00440 | 0.00441 | -0.00922 | -0.03737 | 0.01150 | 0.00477 | -0.03731 | 0.02301 | 0.07422 | 0.03457 |
| **Veach Door** | PT | -0.02103 | -0.05344 | -0.00593 | -0.00573 | -0.00518 | -0.00263 | -0.00632 | -0.00454 | -0.00053 | 0.00011 | -0.04776 | 0.00492 | 0.00491 | 0.18706 | -0.07075 | 0.02645 | 0.02691 | 0.04978 | 0.05291 | 0.14363 | -0.17180 |
| **GT (BDPT)** | BDPT | -0.03288 | -0.05916 | -0.03910 | -0.02856 | -0.02913 | -0.00301 | -0.02221 | -0.00458 | -0.00058 | 0.00011 | -0.08349 | 0.00473 | 0.00473 | 2.02713 | 0.12240 | 0.15684 | 0.16258 | 0.33436 | 0.31368 | -0.13587 | 0.24808 |
| | PSSMLT | -0.03248 | -0.05987 | -0.04727 | -0.02185 | -0.01861 | -0.00295 | -0.01608 | -0.00409 | -0.00056 | 0.00008 | -0.05149 | 0.00392 | 0.00392 | 0.02342 | 0.14725 | 0.11636 | 0.16274 | 0.20014 | 0.23272 | 0.10445 | -0.42433 |
| | MLT | -0.03410 | -0.07112 | -0.06547 | -0.02261 | -0.00967 | -0.00298 | -0.01709 | -0.00404 | -0.00056 | 0.00009 | -0.08106 | 0.00401 | 0.00401 | -0.02349 | 0.21519 | 0.12090 | 0.21554 | 0.10119 | 0.24180 | 0.14322 | -0.36902 |
| | Manifold-MLT | -0.03438 | -0.07323 | -0.06571 | -0.02045 | -0.00787 | -0.00301 | -0.01439 | -0.00424 | -0.00057 | 0.00009 | -0.02645 | 0.00395 | 0.00395 | -0.03055 | 0.21065 | 0.10804 | 0.20825 | 0.07769 | 0.21608 | 0.15131 | -0.31213 |
| | ERPT | -0.01619 | -0.00649 | -0.00410 | -0.00122 | 0.00049 | -0.00227 | -0.00140 | -0.00176 | -0.00056 | 0.00002 | 0.00753 | 0.00303 | 0.00303 | -0.00389 | -0.02886 | 0.00441 | 0.00467 | -0.00301 | 0.00882 | 0.04275 | -0.00764 |
| | Manifold-ERPT | -0.01563 | -0.00862 | -0.00327 | -0.00121 | -0.00047 | -0.00218 | -0.00172 | -0.00203 | -0.00058 | 0.00002 | 0.00702 | 0.00314 | 0.00314 | -0.00388 | -0.02996 | 0.00430 | 0.00415 | 0.00314 | 0.00860 | 0.04368 | 0.00666 |
| **Sponza** | PT | -0.11977 | -0.08352 | -0.05734 | -0.03651 | -0.03375 | -0.01970 | -0.03707 | -0.00850 | -0.00427 | 0.00021 | 0.05730 | 0.02183 | 0.02187 | -0.05280 | 0.16746 | 0.17109 | 0.17006 | 0.33750 | 0.34219 | 0.22184 | 0.39590 |
| **GT (PT)** | BDPT | -0.11455 | -0.07631 | -0.05407 | -0.03133 | -0.02826 | -0.01951 | -0.03274 | -0.00838 | -0.00444 | 0.00022 | 0.05653 | 0.02157 | 0.02160 | -0.04809 | 0.13580 | 0.14373 | 0.14427 | 0.27357 | 0.28747 | 0.21191 | 0.42433 |
| | PSSMLT | -0.10372 | -0.04370 | -0.04225 | -0.01021 | -0.00640 | -0.01793 | -0.01145 | -0.00762 | -0.00397 | 0.00016 | 0.06167 | 0.02050 | 0.02054 | -0.01744 | -0.04242 | 0.04062 | 0.06704 | 0.05275 | 0.08123 | 0.16550 | 0.09456 |
| | MLT | -0.12948 | -0.09366 | -0.07845 | -0.01209 | -0.00628 | -0.01992 | -0.01219 | -0.00856 | -0.00431 | 0.00016 | 0.04427 | 0.02080 | 0.02075 | -0.01973 | -0.03910 | 0.04906 | 0.11132 | 0.05216 | 0.09811 | 0.15318 | 0.10647 |
| | Manifold-MLT | -0.13087 | -0.09648 | -0.08093 | -0.01190 | -0.00723 | -0.01953 | -0.01410 | -0.00716 | -0.00435 | 0.00016 | 0.04745 | 0.02090 | 0.02094 | -0.01777 | -0.03696 | 0.04823 | 0.11457 | 0.06079 | 0.09647 | 0.18738 | 0.11788 |
| | ERPT | -0.08035 | -0.01056 | -0.02362 | -0.00702 | -0.00439 | -0.01611 | -0.00797 | -0.00516 | -0.00391 | 0.00003 | 0.01427 | 0.00583 | 0.00583 | -0.01411 | -0.03573 | 0.02614 | 0.03129 | 0.03308 | 0.05229 | 0.14768 | 0.08338 |
| | Manifold-ERPT | -0.07951 | -0.01462 | -0.02092 | -0.00671 | -0.00428 | -0.01557 | -0.00713 | -0.00545 | -0.00388 | 0.00003 | 0.01591 | 0.00569 | 0.00570 | -0.01438 | -0.03634 | 0.02488 | 0.02913 | 0.03204 | 0.04976 | 0.14297 | 0.08400 |
| | Average | -0.09731 | -0.07349 | -0.05772 | -0.01833 | -0.01251 | -0.01127 | -0.00584 | -0.00532 | -0.00248 | 0.00009 | 0.00683 | 0.01250 | 0.01255 | 0.02372 | 0.02669 | 0.08810 | 0.10273 | 0.13344 | 0.17620 | 0.19058 | 0.19697 |
| | Average Absolute Magnitude | 0.09731 | 0.07863 | 0.05772 | 0.01833 | 0.01506 | 0.01127 | 0.02962 | 0.00600 | 0.00263 | 0.00009 | 0.03760 | 0.01250 | 0.01255 | 0.10585 | 0.11287 | 0.08810 | 0.10273 | 0.14740 | 0.17620 | 0.21324 | 0.29101 |
| | Maximum Magnitude | -0.35628 | -0.25590 | -0.25184 | -0.04573 | -0.07197 | -0.03818 | 0.16618 | -0.01844 | -0.00847 | 0.00022 | -0.08495 | 0.05934 | 0.05977 | 2.02713 | 0.45376 | 0.27333 | 0.35324 | 0.76863 | 0.54666 | 0.65864 | 1.10642 |

This shows the observed worst case estimation in each configuration for IQA. These worse cases are from the high sample count ground truths as underlined in Tables 1a–e and in supplementary materials. A value of zero represents an ideal result, showing that error has not been misreported. Ordered (left to right) by column average. Cells are highlighted from underestimation (blue) to overestimation (orange)

porting in MSE (Table 1e) was ln $\mathcal{Q}$ of 0.54666, in MS-SSIM (Table 1b) −0.00037, and in SC-QI (Table 1c) just −0.00003. Columns of Table 2 have been ordered left to right according to the average magnitude of under- or overestimation for each error metric.

Overall, two of the worst performing metrics were the WBCT and contourlet IQA methods which consistently overestimated error, with an average maximum overestimation across all scenes and rendering algorithms of 0.19697 and 0.19058, respectively. These methods are the same measure performed on the different decompositions of the input images which is simply a distance between two coarse histograms over the proportion of visually important coefficients in a multi-scale image decomposition. These are classified as RR IQA methods, meaning that they are designed to work with the assumption that the reference image may contain errors, but are still representative. However, these results show that the measure is highly sensitive to image distortions such as high frequency impulse noise that are prevalent in Monte Carlo rendered images even at high sample counts. The commonly used MSE measure performs just as poorly, consistently overestimating error with an average maximum of 0.1762 overestimation. MAE performs slightly better with an average overestimation of 0.10273 which is to be expected as MSE weights deviations quadratically while MAE weights deviations linearly.

At the other end of the scale, VIF and IFC consistently underestimate error between images with an average maximum of −0.09731 and −0.07349, respectively. Both methods are based on approximating the two random fields of a GSM noise model. This model assumes that the reference image is correct and does not account of distortions within the reference. Other IQA methods that build off of the GSM model are the information content weighting methods. IW-MSE on average performs slightly better than the standard MSE with an average maximum overestimation of 0.13344; however, due to the poor ability of the GSM to handle noise in the ground truth, this performance is likely due to the addition of multi-scale image analysis rather than because of the GSM noise model. The performance of IW-SSIM which had an average maximum underestimation of −0.00532 supports this theory as it is marginally worse than that of MS-SSIM which scored an average maximum underestimation of −0.00248. These methods only differ in the use of the GSM noise model. UQI and SSIM which do not perform multi-scale image analysis also support this as they perform worse than MS-SSIM with average maximum underestimations by −0.05772 and −0.01127, respectively.

Out of the five scenes the Torus scene showed the largest magnitudes of misreported results, likely due to the slow convergence of caustic illumination. The Veach Bidir and Veach Door scenes also feature caustic illumination; however, these converge comparatively quickly compared to the Torus scene and this can be seen in reduced comparative misreporting between the scenes.

## 6 Recommendations and conclusions

It is difficult to find a balance between the desire for a purely numerical distance metric as we are evaluating the quality of a numerical simulation, and the desire to measure only the perceivable noise as observed by the HVS. We argue that a good balance of these features is for a proposed error metric to be monotonic with respect to a simple numerical divergence like MSE such that a reduction in numerical distance always corresponds to a reduction in reported error. Of the IQA considered in this work that were more advanced than a numerical distance MS-SSIM, SC-QI, SC-DM, and NQM were all monotonic with respect to MSE for the types of distortion that are prevalent in Monte Carlo rendered images. The other IQA tested all showed non-monotonicity in the presence of strong impulse noise, primarily from caustic illumination.

IQA which measured per-pixel structural information seemed to be more robust to the effects of impulse noise in the reference image; however, a stronger divide was seen between methods that applied MGA to those that did not. By isolating high frequency noise in one level of a multi-scale decomposition, its effects on image assessment can be bounded or minimized effectively.

Metrics which used perceptual models of the HVS were highly sensitive to the noise in reference images and quickly became unreliable as the quality of the reference was degraded.

Rendering algorithms such as path tracing and bidirectional path tracing, which uniformly sample path space, are better suited to the task of producing reference images than rendering algorithms which use a Markov based random walk such as Metropolis light transport or energy redistribution path tracing. While in certain situations Markov based algorithms exhibit faster convergence than uniform sampling methods, before the simulation has fully converged a uniform method which will have independently distributed error while a Markov algorithm will exhibit noise distributed deterministically with respect to the trajectory the random walk has followed. The result of this is that when we consider the possibility of noise in reference images, noise from Markov processes is more likely to form structural artefacts in the reference, exacerbating misreported error when IQA consider structural features and similarity.

Our recommendations are that MS-SSIM or SC-QI be used for image quality assessments when evaluating images produced by Monte Carlo rendering algorithms as these methods were the most robust when we consider noise in reference images. Reference images should ideally be rendered

with uniform sampling methods to avoid the introduction of structural artefacts in IQA. It is important that the reference used is not only visually noise-free, but also that it is of sufficiently higher numerical quality than images tested against it. Reference images should therefore be rendered to at least an order of magnitude higher sample count than test images to minimize the possibility of noise in the reference causing a significant deviation in reported error. And finally that the sample count and method of production of the reference image should be stated clearly to give researchers every confidence in reported results.

**Acknowledgements** Joss Whittle is supported by EPSRC Doctoral Training Award EP/K502935/1.

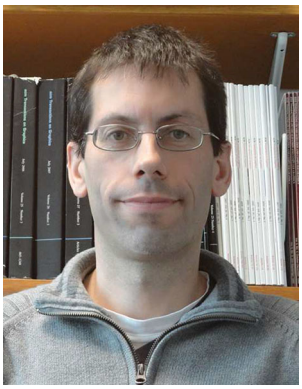


## References

1. Anderson, M., Motta, R., Chandrasekar, S., Stokes, M.: Proposal for a standard default color space for the internet—srgb. Color Imaging Conf. **1996**(1), 238–245 (1996)
2. Avcibas, I., Sankur, B., Sayood, K.: Statistical evaluation of image quality measures. J. Electron. Imaging **11**(2), 206–223 (2002)
3. Bae, S.H., Kim, M.: A novel image quality assessment with globally and locally consilient visual quality perception. IEEE Trans. Image Process. **25**(5), 2392–2406 (2016). doi:10.1109/TIP.2016.2545863
4. Bauszat, P., Eisemann, M., Eisemann, E., Magnor, M.: General and robust error estimation and reconstruction for monte carlo rendering. Comput. Graph. Forum (Proc. of Eurographics EG) **34**(2), 597–608 (2015)
5. Bovik, A.C.: The Essential Guide to Image Processing. Academic Press, Boston (2009)
6. Brunet, D., Vrscay, E.R., Wang, Z.: On the mathematical properties of the structural similarity index. IEEE Trans. Image Process. **21**(4), 1488–1499 (2012). doi:10.1109/TIP.2011.2173206
7. Burt, P., Adelson, E.: The laplacian pyramid as a compact image code. IEEE Trans. Commun. **31**(4), 532–540 (1983)
8. Chai, T., Draxler, R.R.: Root mean square error (RMSE) or mean absolute error (MAE)? Geosci. Model Dev. Discuss. **7**, 1525–1534 (2014)
9. Chandler, D.M., Hemami, S.S.: Vsnr: a wavelet-based visual signal-to-noise ratio for natural images. IEEE Trans. Image Process. **16**(9), 2284–2298 (2007). doi:10.1109/TIP.2007.901820
10. Chui, C.K.: An Introduction to Wavelets. Academic Press Professional Inc., San Diego (1992)
11. Cline, D., Talbot, J., Egbert, P.: Energy redistribution path tracing. ACM Trans. Graph. **24**(3), 1186–1195 (2005)
12. Damera-Venkata, N., Kite, T.D., Geisler, W.S., Evans, B.L., Bovik, A.C.: Image quality assessment based on a degradation model. IEEE Trans. Image Process. **9**(4), 636–650 (2000)
13. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multiresolution image representation. IEEE Trans. Image Process. **14**(12), 2091–2106 (2005). doi:10.1109/TIP.2005.859376
14. Doidge, I., Jones, M.W., Mora, B.: Mixing monte carlo and progressive rendering for improved global illumination. Visual Comput. **28**(6–8), 603–612 (2012). doi:10.1007/s00371-012-0703-2
15. Doidge, I.C., Jones, M.W.: Probabilistic illumination-aware filtering for Monte Carlo rendering. Vis. Comput. (2013). doi:10.1007/s00371-013-0807-3
16. Eslami, R., Radha, H.: Wavelet-based contourlet coding using an spiht-like algorithm. In: Conference on Information Sciences and Systems (CISS), pp. 784–788. Department of Electrical Engineering, Princeton University (2004)
17. Gao, X., Lu, W., Tao, D., Li, X.: Image quality assessment based on multiscale geometric analysis. Trans. Image Proc. **18**(7), 1409–1423 (2009). doi:10.1109/TIP.2009.2018014
18. Hachisuka, T., Jarosz, W., Weistroffer, R.P., Dale, K., Humphreys, G., Zwicker, M., Jensen, H.W.: Multidimensional adaptive sampling and reconstruction for ray tracing. ACM Trans. Graph. **27**(3), 33:1–33:10 (2008)
19. Hachisuka, T., Kaplanyan, A.S., Dachsbacher, C.: Multiplexed metropolis light transport. ACM Trans. Graph. **33**(4), 100:1–100:10 (2014)
20. Hachisuka, T., Ogaki, S., Jensen, H.W.: Progressive photon mapping. ACM Trans. Graph. **27**(5), 130:1–130:8 (2008)
21. Jakob, W.: Mitsuba renderer (2010). http://www.mitsuba-renderer.org
22. Jakob, W., Marschner, S.: Manifold exploration: a Markov Chain Monte Carlo technique for rendering scenes with difficult specular transport. ACM Trans. Graph. (TOG) **31**(4), 58 (2012)
23. Jensen, H., Christensen, N.: Optimizing path tracing using noise reduction filters. In: Proceedings of WSCG95, pp. 134–142 (1995)
24. Jensen, H.W.: Realistic Image Synthesis Using Photon Mapping. A. K. Peters Ltd, Natick (2001)
25. Kajiya, J.T.: The rendering equation. In: Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '86, pp. 143–150. ACM, New York, NY, USA (1986)
26. Kalantari, N.K., Bako, S., Sen, P.: A Machine learning approach for filtering Monte Carlo noise. ACM Trans. Graph. **34**(4), 122:1–122:12 (2015). doi:10.1145/2766977
27. Kalantari, N.K., Sen, P.: Removing the noise in Monte Carlo rendering with general image denoising algorithms. Comput. Graph. Forum **32**(2pt1), 93–102 (2013)
28. Kelemen, C., Szirmay-Kalos, L., Antal, G., Csonka, F.: A simple and robust mutation strategy for the metropolis light transport algorithm. Comput. Graph. Forum **21**(3), 531–540 (2002)
29. Kontkanen, J., Rosonen, J., Keller, A.: Irradiance filtering for Monte Carlo ray tracing. In: Monte Carlo and Quasi-Monte Carlo Methods 2004, pp. 259–272. Springer (2004)
30. Lafortune, E.P., Willems, Y.D.: Bi-directional path tracing. Proc. Comput. Graph. **93**, 145–153 (1993)
31. Li, T.M., Lehtinen, J., Ramamoorthi, R., Jakob, W., Durand, F.: Anisotropic gaussian mutations for metropolis light transport through hessian-hamiltonian dynamics. ACM Trans. Graph. **34**(6), 209:1–209:13 (2015)
32. Mantiuk, R., Kim, K.J., Rempel, A.G., Heidrich, W.: Hdr-vdp-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Trans. Graph. (TOG) **30**(4), 40 (2011)
33. Mitsa, T., Varkur, K.L.: Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms. In: ICASSP-93. 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993, vol. 5, pp. 301–304. IEEE (1993)
34. Pharr, M., Humphreys, G.: Physically Based Rendering, Second Edition: From Theory to Implementation, 2nd edn. Morgan Kaufmann Publishers Inc., San Francisco (2010)
35. Ponomarenko, N., Ieremeiev, O., Lukin, V., Egiazarian, K., Jin, L., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., Kuo,

C.C.J.: Color image database tid2013: Peculiarities and preliminary results. In: European Workshop on Visual Information Processing (EUVIP), pp. 106–111 (2013)
36. Rousselle, F., Knaus, C., Zwicker, M.: Adaptive rendering with non-local means filtering. ACM Trans. Graph. **31**(6), 195:1–195:11 (2012)
37. Rushmeier, H.E., Ward, G.J.: Energy preserving non-linear filters. In: Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '94, pp. 131–138. ACM, New York, NY, USA (1994)
38. Sheikh, H., Sabir, M., Bovik, A.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans. Image Process. **15**(11), 3440–3451 (2006)
39. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Trans. Image Process. **15**(2), 430–444 (2006)
40. Sheikh, H.R., Bovik, A.C., de Veciana, G.: An information fidelity criterion for image quality assessment using natural scene statistics. Trans. Image Proc. **14**(12), 2117–2128 (2005). doi:10.1109/TIP.2005.859389
41. Sheikh, H.R., Wang, Z., Cormack, L., Bovik, A.C.: LIVE Image Quality Assessment Database Release 2 (2014)
42. Shy, D., Perona, P.: Xy separable pyramid steerable scalable kernels. In: CVPR, pp. 237–244 (1994)
43. Suykens, F., Willems, Y.D.: Adaptive filtering for progressive monte carlo image rendering. In: The 8-th International Conference in Central Europe on Computer Graphics, Visualization and Interactive Digital Media 2000 (WSCG' 2000), February 2000. Held in Plzen, Czech Republic (2000)
44. Tamstorf, R., Jensen, H.W.: Adaptive sampling and bias estimation in path tracing. In: Proceedings of the Eurographics Workshop on Rendering Techniques '97, pp. 285–296. Springer, London, UK (1997)
45. Tao, D., Li, X., Lu, W., Gao, X.: Reduced-reference IQA in contourlet domain. IEEE Trans. Syst. Man Cybern. Part B (Cybernetics) **39**(6), 1623–1627 (2009). doi:10.1109/TSMCB.2009.2021951
46. Toet, A., Van Ruyven, L.J., Valeton, J.M.: Merging thermal and visual images by a contrast pyramid. Optic. Eng. **28**(7), 287–789 (1989)
47. Tofallis, C.: A better measure of relative prediction accuracy for model selection and model estimation. J. Oper. Res. Soc. **66**(8), 1352–1362 (2015). doi:10.1057/jors.2014.103
48. Veach, E., Guibas, L.J.: Metropolis light transport. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques. SIGGRAPH '97, pp. 65–76. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA (1997)
49. Vetrivelan, P., Subha, R.: Wavelet based contourlet transform for image compression. In: Proceeding of International conference of Cognition and Recognition. IEEE, pp. 915–919. Citeseer (2005)
50. Wang, Z., Bovik, A.C.: A universal image quality index. IEEE Signal Process. Lett. **9**(3), 81–84 (2002). doi:10.1109/97.995823
51. Wang, Z., Bovik, A.C.: Mean squared error: love it or leave it? A new look at signal fidelity measures. IEEE Signal Process. Mag. **26**(1), 98–117 (2009). doi:10.1109/MSP.2008.930649
52. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. Trans. Image Proc. **13**(4), 600–612 (2004). doi:10.1109/TIP.2003.819861
53. Wang, Z., Li, Q.: Information content weighting for perceptual image quality assessment. IEEE Trans. Image Process. **20**(5), 1185–1198 (2011). doi:10.1109/TIP.2010.2092435
54. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2004, vol. 2, pp. 1398–1402. IEEE (2003)
55. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. **30**(1), 79 (2005)
56. Yang, C.C., Kwok, S.H.: Efficient gamut clipping for color image processing using LHS and YIQ. Opt. Eng. **42**(3), 701–711 (2003)
57. Zhang, L., Zhang, L., Mou, X., Zhang, D.: Fsim: a feature similarity index for image quality assessment. IEEE Trans. Image Process. **20**(8), 2378–2386 (2011). doi:10.1109/TIP.2011.2109730

**Joss Whittle** has received the B.Sc. degree from Swansea University. He is currently persuing a Ph.D. in Computer Science at Swansea University in the Visual Computing Research group. His research interests include global illumination, variance reduction, and high performance computing.



**Mark W. Jones** has received the B.Sc. and Ph.D. degrees from Swansea University. He is a Professor in the Department of Computer Science at Swansea University, where he leads the Visual Computing Research group. His research interests include global illumination, visualization, data science, and associated algorithms and data structures. http://cs.swan.ac.uk/~csmark/.

**Rafał Mantiuk** is a senior lecturer at the Computer Laboratory, University of Cambridge (UK). He received PhD from the Max-Planck-Institute for Computer Science (Germany). His recent interests focus on designing imaging algorithms that adapt to human visual performance and viewing conditions in order to deliver the best images given limited resources, such as computation time, bandwidth, or dynamic range. He contributed to early work on high dynamic range imaging, including quality metrics (HDR-VDP), video compression and tone-mapping. More on his research can be found at: http://www.cl.cam.ac.uk/~rkm38/.