Institut für Visual Computing &
Human-Centered Technology
Technische Universität Wien
Favoritenstr. 9-11 / E193-02
A-1040 Wien
AUSTRIA
Tel: +43 (1) 58801-193201
Fax: +43 (1) 58801-193209

Institute of Visual Computing &
Human-Centered Technology
TU Wien
email:
technical-report@cg.tuwien.ac.at
other services:
http://www.cg.tuwien.ac.at/

TECHNICAL REPORT

R-Score: A Novel Approach to Compare Monte Carlo Renderings

C. Freude, H. Sakai, K. Zsolnai-Fehér & M. Wimmer

TR-193-02-2020-4
November 2020

R-Score: A Novel Approach to Compare Monte Carlo Renderings

C. Freude, H. Sakai, K. Zsolnai-Fehér & M. Wimmer

November 9, 2020

Abstract

In this paper, we propose a new approach for the comparison and analysis of Monte Carlo (MC) rendering algorithms. It is based on a novel similarity measure called render score (RS) that is specifically designed for MC rendering, statistically motivated, and incorporates bias and variance. Additionally, we propose a comparison scheme that alleviates the need for *practically* converged reference images (RIs). Our approach can be used to compare and analyze different rendering methods by revealing detailed (per-pixel) differences and subsequently potential conceptual or implementation-related issues, thereby offering a more informative and meaningful alternative to commonly used metrics.

1 Introduction

The accurate simulation of light transport for the synthesis of photorealistic images is of great importance for film production, architectural visualization, product design, and many other application areas. The prevalent approach to this problem is to use a model that is described by the *rendering equation* [Kaj86] and to evaluate its numerous integrals using Monte Carlo (MC) integration.

This type of integration approximates a function's integral by repeated random sampling. Due to the stochastic nature of this approach, initially, the computed results suffer from high variance, which manifests itself as noise in the rendered images. The variance eventually vanishes as the number of samples increases and the integral converges to the correct solution.

A significant amount of research has been dedicated to reduce variance and speed up convergence by using more advanced integration and sampling schemes that try to distribute samples in a more efficient way. However, the variance inherent to all MC-based methods makes the comparison of different rendering techniques difficult, as images are only completely noiseless in the *theoretical* limit, which often cannot be attained in practice.

A common approach to compare renderings is to select a set of representative test scenes, render them using a fixed time or sample budget, and compare the results to a different rendering technique or a *quasi-converged* reference image (RI). Typically, to assess the differences, only qualitative visual comparison or simple metrics, such as the mean squared error (MSE) are used. A limitation of these approach is that the inherent variance may distort the result for both visual or metric-based comparisons. Furthermore, the variable quality of RIs may additionally deteriorate the assessments, as pointed out by Whittle et al. [WJM17]. In Figure 1, we provide an intuitive example of these issues when using common metrics, such as absolute deviation (AD) or MSE. The shown normal distributions exemplify the radiance

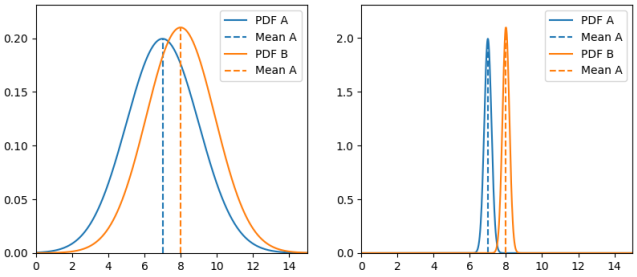


Figure 1: The normal distributions on the left and right only differ in their variance, but have the same distance between the means. Computing the absolute deviation of the means $|\bar{x}_A - \bar{x}_B|$ would yield the same difference for the left and the right case because only the mean difference is taken into account. Our proposed render score (RS) incorporates not only the mean difference but also the variances of the distributions and is therefore able to distinguish those two different cases.

distributions computed by two MC rendering techniques. It is evident that considering only the difference between their means is insufficient, as the variances contain crucial information concerning the similarity of two distributions.

In this paper, we propose a novel approach for the comparison of MC rendering techniques, and our goal is to address the drawbacks of commonly used metrics and to establish a methodology to improve the comparability across different publications, rendering techniques, and their implementations.

Our main contribution is the *render score (RS)*, a novel distance measure for the statistically motivated quantification of the similarity of radiance estimates. It offers a more informative and meaningful alternative to commonly used metrics, e.g., AD or MSE, and can be used to analyze and compare different MC rendering algorithms. Instead of working with less informative per-pixel sample means (i.e., the standard procedure in the state of the art), it additionally incorporates the bias and variance of the radiance sample distribution within a pixel. As discussed in Section 3 and illustrated in Figure 4, we designed the RS specifically for MC rendering cases where the bias and variance are subject to minimization.

To obtain the necessary radiance sample distributions, we compute multiple independent (and non-tonemapped) short

renderings using a relatively low number of samples per pixel (SPP). This enables our approach to work independently of the underlying implementation, as all renderers are capable of image output. It can be used as an *absolute* measure, i.e., a measure that quantifies an image on its own—if high-quality RIs are available. Unfortunately, those are not always feasible to compute and such absolute comparisons against a RI can be problematic [WJM17]. Therefore, we propose to use our RS as a *relative* measure to compare and rank multiple MC rendering techniques against each other and potentially unveil inaccuracies in their implementations that would otherwise be challenging to find. Similarly to the approach proposed by Whittle et al. [WJM17], our relative comparison scheme is based on an ensemble of different sample size combinations.

Our approach can be summarized as follows (Figure 2):

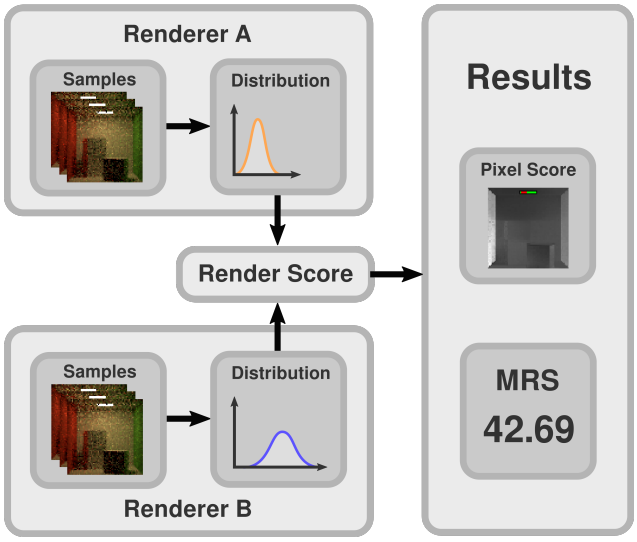


Figure 2: Overview of our proposed comparison scheme based on our novel RS. First, we generate multiple short renderings for each renderer which represent samples of the mean radiance distribution. Based on these distributions, we compute the RS, which enables analysis through the examination of average or per-pixel scores and sample distributions. This provides the users with additional insights about the renderers that may go unnoticed with traditional error metrics.

1. First, we compute a prescribed number of short (i.e., fixed low number of SPP) renderings.

2. Then, we select a reference algorithm, e.g., path tracing (PT), and compute the RS of the techniques of interest relative to it based on their sample distributions.

3. We further compute additional sample sets by aggregating the samples of the original distributions. This enables a more extensive analysis, especially with respect to a relative comparison in the absence of RIs.

4. Ultimately, the results can be analyzed by examining the average score for the whole image or the per-pixel scores and sample distributions.

In order to facilitate the adoption of our approach, we plan to publish our source code in the near future.

The remainder of the paper is structured as follows: in the following section, we discuss related work in order to put our proposed approach into context. Our motivations and the theory behind the RS are explained in Section 3, followed by Section 4, where we present several results to demonstrate the usefulness of our approach. In closing, we discuss the benefits and limitations of the RS in Section 5.

2 Related Work

Perceptual Quality Measures for Monte Carlo Rendering. Surprisingly, there have been only a few attempts to quantify the quality of Monte Carlo (MC) renderings. Many researchers employed a perceptual model that can be used to approximate perceived differences, which in turn can be exploited for rendering. For instance, the visible differences predictor [Dal93] has been employed to approximate and monitor perceived rendering quality in order to use it for a stopping condition [Mys98] or to alternate between complementary rendering techniques [VMKK00]. Ramasubramanian et al. [RPG99] developed a perceptually based error metric for image-space adaptive sampling. Farrugia et al. [FP04] used an existing vision model [PFFG98] in order to achieve the same goal. However, all these works do not aim to provide a solution for the robust comparison of different rendering techniques.

General Image Quality Metrics. When it comes to the comparison of rendering techniques, general image quality metrics, which are popular in the image-processing community, are the predominant choice. Prominent examples are the mean squared error (MSE), the root-mean-square error (RMSE), the structural similarity (SSIM) index [WBSS04], and variants of the high-dynamic-range visual difference predictor (HDR-VDP) [MDMS05, MKRH11, NMDSLC15]. For instance, Meneghel and Netto [MN15] employed SSIM and HDR-VDP2 for the comparison of six different rendering techniques.

Whittle et al. [WJM17] provided a comprehensive overview and analysis of a multitude of general image quality metrics. The problem with these *general* metrics is that they are agnostic to the sample distributions in MC rendering, which could potentially provide a breadth of additional information. We aim to alleviate this problem by deliberately incorporating information about distributions.

Rendering Verification. Several works [GTGB84, MCTG00, SW04, McN06] compare renderings to real-world measurements in order to assess rendering quality. Ulbricht et al. [UWP06] investigated the state of the art for the verification of renderings and pointed out that all approaches have their weaknesses and that the development of robust and practical solutions is still an open task. Nevertheless, the verification of rendering techniques using real-world measurements is orthogonal to our objective of comparing different rendering techniques.

Statistical Testing. Subr and Arvo [SA07] employed statistical tests to compare rendering techniques. However, they use test hypotheses which are not suited to test for equality, but can only show significant differences.

3 Our Approach

We have defined the following desired properties that acted as guiding principles for the development of our approach:

Variance-aware: In contrast to other difference measures, such as the mean squared error (MSE), which only consider one single "snapshot" of the means, we aim to also include the variance of the radiance estimates into account.

Radiance-based: We are primarily interested in the quantitative comparison of linear high dynamic range (HDR) radiance estimates and therefore do not consider visual perception or tone mapping, as the latter may non-linearly distort the mean radiance distribution.

Implementation-independent: Our approach should be as implementation-independent as possible, without the need for significant changes to the rendering system.

Reference-less: Since the synthesis of accurate reference images (RIs) is not always practical or even feasible, we are interested in a way to robustly compare unconverged results.

Based on these principles, we developed our novel comparison scheme and its basis—the render score (RS)—which we describe in the following sections.

3.1 Render Score

The goal behind the RS is the assessment of the quantitative differences between the radiance estimates of different rendering techniques. Each Monte Carlo (MC) rendering technique produces a characteristic radiance sample distribution. The simplest and most prevalent solution is to compare the sample means using metrics such as absolute deviation (AD) or MSE. The sample mean, however, is highly sensitive to the inherent variance of the MC sampling process. Our key insight is to not only consider the sample mean but to also consider the underlying *distribution* in the comparisons. The central limit theorem (CLT) states that the distribution of the sample mean tends towards a normal distribution as its sample size increases. In the context of MC rendering, this size corresponds to the number of samples per pixel (SPP): as the SPP are increased, the distribution of the mean will converge to a normal distribution, regardless of the used rendering technique. This way, we can define the difference between two renderers in terms of their sample mean distributions that are assumed to be approximately normal.

The motivation behind our score is that we would like to assess the likelihood that two renderers compute the same sample mean. Since the sample mean distribution approximately encodes the probability of a renderer generating certain mean values, we are interested in some notion of similarity between distributions, which we can use to measure the difference between rendering techniques.

We have identified the following properties that are desirable for our distribution-based score:

1. Given two radiance distributions, we are interested in a measure that penalizes both bias and variance.
2. Therefore, we are not interested in the exact similarity of the two distributions, as a distribution with a *lower variance* (compared to the other distribution) should be favored and accordingly rewarded with a *higher score*.
3. At the same time, a distribution with a similar mean to the other distribution—i.e., a *lower bias*—should manifest itself in a *higher score* as well.
4. Since we can not expect the distributions to be exactly normally-distributed, we require an additional factor that represents the deviation from normality, preferably a scalar in the interval $[0, 1]$.

With these requirements in mind, we have identified two suitable building blocks, i.e., the product of Gaussians (POGs) and the Kuiper statistic [Kui60], which can be combined to obtain our proposed RS.

3.1.1 Product of Gaussians

The POG is given by $f_A(x)f_B(x)$, where $f_A(x)$ and $f_B(x)$ denote two normal probability density functions (PDFs), as illustrated in Figure 3. By integrating the POG, we obtain

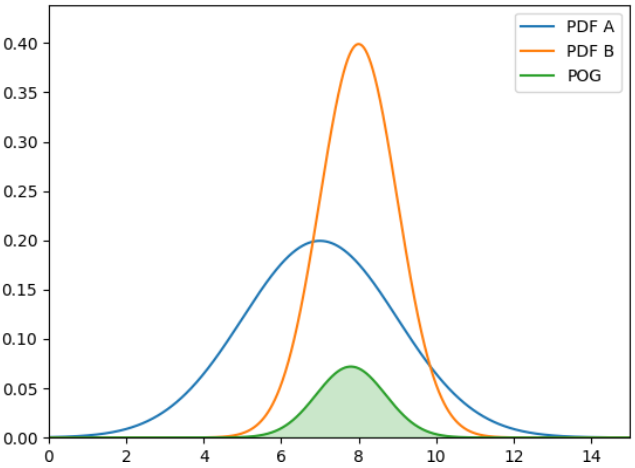


Figure 3: The POG (green) varies with the mean difference and the variance of both PDFs A (blue) and B (orange). The integral of the PG (see Equation 1 and Figure 4) increases as the mean difference and variances decrease, which fulfills the first three of our four main design criteria (see Section 3.1).

PG, the first component, of our RS, which implements our notion of similarity between two distributions:

$$\mathrm{PG} = \int f_A(x) f_B(x)\,\mathrm{d}x. \tag{1}$$

Note that this simple formula fulfills the first three requirements we specified before, as illustrated in Figure 4. It shows

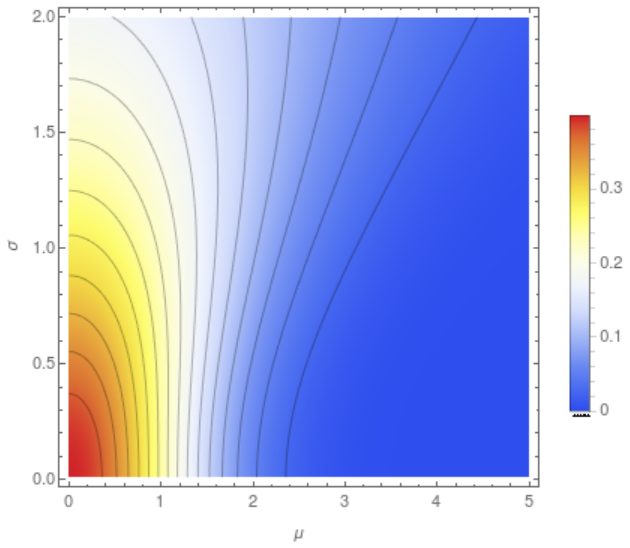


Figure 4: This plot shows the magnitude of PG (Equation 1) between a fixed $N(0,1)$ and a $N(\mu,\sigma)$ normal distribution with varying mean μ and standard deviation σ. This shows how the PG increases with decreasing mean difference and variance.

that the score increases as the difference in means (the bias) and the magnitude of the variance decreases. We refrain from using classic distribution-based similarity measures, because they would penalize a difference in variance, which is not in line with our requirements.

In our case where we approximate f_A and f_B by Gaussians, the analytical solution of PG is given by:

$$\mathrm{PG}(\mu_A,\mu_B,\sigma_A,\sigma_B) = \frac{\exp\left(-\frac{(\mu_A-\mu_B)^2}{2(\sigma_A^2+\sigma_B^2)}\right)}{\sigma_A\sigma_B\sqrt{2\pi}\sqrt{\frac{1}{\sigma_A^2}+\frac{1}{\sigma_B^2}}}, \quad (2)$$

where μ_A, μ_B and σ_A, σ_B denote the population means μ and standard deviations σ of normal distributions A and B, respectively. In practice, we estimate μ and σ by the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and sample standard deviation s and approximate the sample distribution by a normal distribution, i.e., $N(\bar{x}, s)$. For the calculation of the standard deviation s from the samples x_i, we use the conventional formula $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$. These singularity cases of $s = 0$ are handled separately, as explained in Section 4.

3.1.2 Deviation from Normality

Unfortunately, the sample mean is only precisely normal distributed in the limit, thereby violating our fourth rule. To remedy this, we introduce an additional scaling function to "invalidate" the PG values that are computed from samples that deviate from normality. There are multiple statistical approaches to assess normality, however, only a few are bounded between 0 and 1, e.g., the Kuiper statistic [Kui60], an extension of the Kolmogorov-Smirnov (KS) statistic [Kol33]. It describes the difference between two cumulative distribution functions (CDFs) as follows:

$$D_n = D^+ + D^-, \quad (3)$$

$$D^+ = \max_x F(x) - F_N(x), \quad (4)$$

$$D^- = \max_x F_N(x) - F(x), \quad (5)$$

i.e., it is the sum of the upper D^+ and lower D^- maximum difference between the empirical CDF F of the samples to the analytic CDF F_N of the corresponding ideal normal distribution $N(\bar{x}, s)$. Figure 5 further illustrates how we use the Kuiper statistic to compute the deviation from normality. We also use $1 - D_n$ as a weighting factor to scale

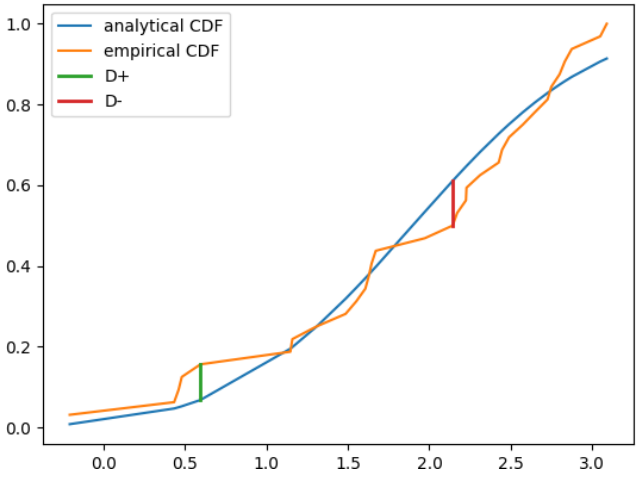


Figure 5: To assess the deviation from normality for a given sample distribution we generate the empirical CDF and the analytical CDF (based on the ideal distribution $N(\bar{x}, s)$). Then we compute the Kuiper statistic using the sum of the upper D^+ and lower D^- maximum difference.

(or "invalidate") the similarity value PG depending on the agreement between the sample distribution and the corresponding ideal normal distribution. The inclusion of this scaling factor successfully fulfills our fourth requirement.

3.1.3 Calculation of the Render Score

As mentioned before, PG is based on the assumption that the mean sample distribution of both renderers is normal-distributed. Any deviation from normality of just one of the renderer's distributions makes PG less reliable. To model this fact, we multiply PG with the weighting factors $1 - D_n$. Since $1 - D_n$ is in the range $[0, 1]$, the multiplication ensures that PG decreases linearly proportionally to the deviation from normality and that it keeps its original magnitude when both distributions are normal-distributed. This multiplicative combination of PG and the normality factors $1 - D_n$ forms our RS between the radiance samples of renderer A and B. It is defined as:

$$\mathrm{RS}(A,B) = \mathrm{PG}(\bar{x}_A,\bar{x}_B,\mathrm{s}_A,\mathrm{s}_B) \\ (1 - D_n(x_A,\bar{x}_A,\mathrm{s}_A))(1 - D_n(x_B,\bar{x}_B,\mathrm{s}_B)), \quad (6)$$

where x denotes the individual samples, $\bar{x}$ the sample mean, and s the sample standard deviation of the chosen renderer in their subscripts.

(l, m)	(11,0)	(10,1)	(9,2)	(8,3)	(7,4)	(6,5)	(5,6)
(n_l, n_m)	(2048,1)	(1024,2)	(512,4)	(256,8)	(128,16)	(64,32)	(32,64)

Table 1: An example of possible set counts and sizes represented as tuples $(n_l, n_m) = (2^l, 2^m)$ that can be created from $n = 2^{11} = 2048$ initial samples. Please note that we restrict the set count by $l \geq 5$.

3.2 Sample Aggregation

In practice, we compute a fixed number n of short renderings for each rendering technique that we seek to compare. The renderings are computed using a specified number of SPP and represent the statistical samples for the respective rendering techniques. These *mean samples* have a particular variance that depends on the used number of SPP. This variance can be reduced further by averaging multiple mean samples. We can exploit this fact by performing the following sample aggregation scheme.

First, we render $n = 2^i$ short renderings, which form our initial samples. The total number of initial samples n can be subdivided into $n_l = 2^l$ sets with $n_m = 2^m$ samples each, where $n = 2^i = 2^{l+m}$. Averaging the samples in each set yields additional samples, each aggregating n_m mean samples (which corresponds a radiance sample count of n_m times the number of SPP used for rendering). Since we need a minimum number of samples to compute meaningful statistics and an empirical CDF, we recommend keeping the number of sets above $n_l = 2^5$ (a common minimal sample size used in statistics for the normal distribution of the mean).

This sample aggregation scheme allows us to investigate the mean sample distribution in different states. As we aggregate, the variance of the mean sample distribution is reduced at the expense of the number of samples that represent the distribution.

4 Evaluation

As a proof of concept, we tested our approach by comparing several popular integrators available in Mitsuba [Jak10]. We chose to compare a set of commonly used integrators, e.g., bidirectional path tracing (BDPT), energy redistribution path tracing (ERPT), Metropolis light transport (MLT), while using path tracing (PT) as the relative reference.

For random number generation we chose the independent sampler in all cases. Furthermore, we slightly adapted Mitsuba to expose the sampler seed to the command line and added an option to successively save images with a chosen number of samples per pixel (SPP) without the need of reloading the scene. In general, our approach can be used with any renderer that is able to output images that were rendered with a specified number of SPP and provides control over the seed of the random number generator. All image samples were computed and processed in linear high dynamic range (HDR) as stored by Mitsuba and were deliberately rendered with a resolution of 64×64 to avoid cluttering the visual analysis with too much image detail. The rendered images and scores are based on $n = 2048$ rendered images that were computed using up to 64 SPP. For the visualization of the rendered images, we chose the global tonemapper by Reinhard et al. [RSSF02], while all other difference and score images were converted to luminance and normalized linearly. Furthermore, in order to be able to compare our render score (RS) to the mean squared error (MSE)—the prevalent approach to compare renderings in the state of the art—we computed the reference images for the results shown in Figure **??** and 6, even though the RS itself is a reference-less approach.

In the following, we discuss various benefits of our approach.

Scoring Renderings. Our approach makes it possible to quantify the visual quality of renderings more faithfully than simple distance-based metrics, such as MSE. This is due to the fact that, additionally to the mean distance, we consider the variance of the involved distributions, which contains important information about similarity. We therefore argue that our approach facilitates *more meaningful comparisons* between different rendering techniques and parametrizations. In Figure **??** and 6, we demonstrate this advantage based on different renderings with similar MSE. Our RS is capable of characterizing the differences in visual quality between the renderings, whereas the MSE fails in this regard.

(a) Path Tracing
MSE: 0.0047, MRS: 0.5743

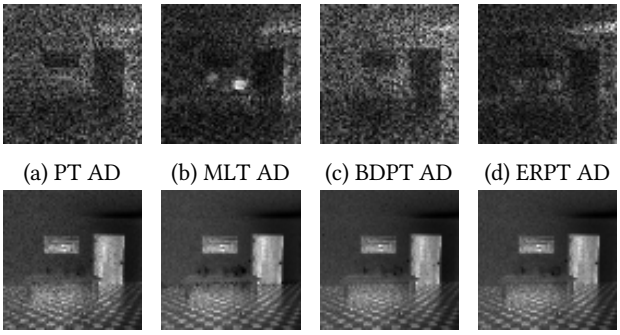
(b) Metropolis Light Transport
MSE: 0.0046, MRS: 0.7125

(c) Bidirectional Path Tracing
MSE: 0.0047, MRS: 0.7207

(d) Energy Redistribution Path Tracing
MSE: 0.0047, MRS: 0.7610

Figure 6: This figure demonstrates how different renderings with equal MSE are scored by the MRS (higher is better). Please note that MLT b had difficulties to sample the glass teapot, hence the lower MRS compared to BDPT c. The MSE clearly fails in distinguishing the visual qualities of these renderings, whereas our proposed MRS scores them more faithfully.

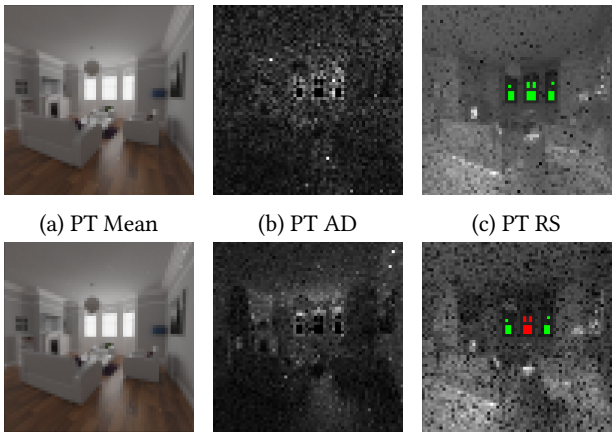
Low-Noise Per-Pixel Analysis. Since our approach considers the variance of the distributions—additionally to the distance—it is also *less susceptible to noise* at low sample counts. This facilitates detailed per-pixel analysis of differences between rendering techniques in order to identify their strengths and weaknesses. Figure 7 demonstrates this aspect by comparing our RS to the absolute deviation (AD), i.e., the absolute difference to the reference image.



(a) PT AD (b) MLT AD (c) BDPT AD (d) ERPT AD

(e) PT RS (f) MLT RS (g) BDPT RS (h) ERPT RS

Figure 7: The noise levels in the per-pixel ADs (top row) are much higher compared to the per-pixel RSs (bottom row). This makes per-pixel analysis more reliable, especially at low sample counts. These images correspond to the renderings shown in Figure **??**.



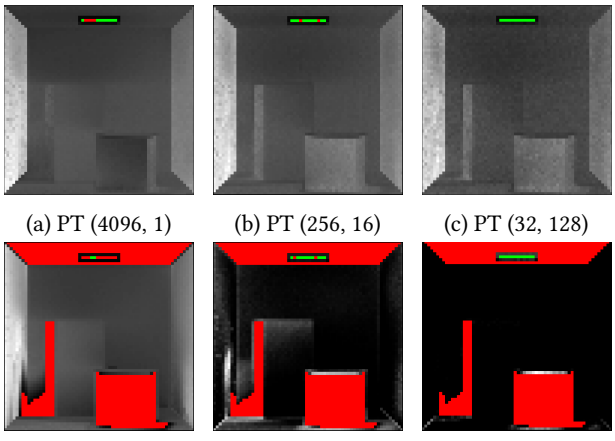
(a) PT Mean (b) PT AD (c) PT RS

(d) BDPT Mean (e) BDPT AD (f) BDPT RS

Figure 8: This figure demonstrates RS singularities for BDPT and PT. The singularities are visible in the windows of the living room. Areas where both distributions have zero variance and equal mean are highlighted in green, whereas areas where exactly one of the distributions has zero variance are highlighted in red. Since the AD does not take variances into account, it fails to identify such cases.

Debugging Renderers. The previous examples demonstrated how the RS can reveal or emphasize subtle differences between integrators. Furthermore, it can also be used as a debugging tool to detect possible implementation issues, as shown in Figure 8. In this figure, there are multiple pixels where at least one of the per-pixel distributions has zero variance, which in turn causes the RS to exhibit a singularity. We chose to highlight such pixels in green for cases where $s_A = s_B = 0$ and $\bar{x}_A = \bar{x}_B$, and red for cases where either s_A or s_B is zero, thereby improving the process of debugging and identifying differences in renderer parametrization. In contrast, the AD cannot identify such cases, as it is not aware of the underlying per-pixel variances.

Variance and Bias Analysis. Increasing the mean size n_m of the sample distributions used to calculate our RS linearly decreases the variance of the compared mean sample distributions (Section 3.2). Intuitively, this means that for a fixed distance between the distribution means (i.e., a fixed bias), the variance has more influence on the score at lower mean sizes than at greater mean sizes. Therefore, at higher mean sizes, the bias manifests itself more in the score, whereas at lower mean sizes, the variance will appear as the primary component in the score. This behavior of our RS facilitates *intricate analyses with regard to the variance and bias behavior* of a particular rendering technique. Figure 9 demonstrates this aspect by showing that an increase in mean size significantly increases the impact of bias in the rendering score.



(a) PT (4096, 1) (b) PT (256, 16) (c) PT (32, 128)

(d) BDPT (4096, 1) (e) BDPT (256, 16) (f) BDPT (32, 128)

Figure 9: This figure illustrates the effect of reducing the variance of the mean sample distribution by increasing the mean size (3.2). The top row shows the per-pixel RSs for different sample set sizes (n_l, n_m) for an unbiased PT. The bottom row shows the corresponding per-pixel RSs for a biased PT that was restricted to a path depth of two. The impact of the bias appears as lower RSs at higher mean sizes. As explained in Figure 8, the pixels that are highlighted in red signify that exactly one of the corresponding distributions exhibit a variance of zero: this is to be expected, as with the biased parametrization, many paths are terminated before they reach the light source.

5 Discussion

We designed our approach according to the principles mentioned in the beginning of Section 3. In order to keep it as independent as possible from the implementation of the renderer, we chose to use short renderings as statistical samples. This statistical approach further enables us to compare the actual sample distributions and incorporate their variances. In order to quantitatively compare different renderers without distracting influences arising from tone mapping or perception-related aspects, we compute our novel scheme on raw linear high dynamic range (HDR) images. To avoid the issue of obtaining high-quality reference images, we use the render score (RS) as a relative measure and compare ensembles consisting of different sample size configurations.

The RS was specifically designed for Monte Carlo (MC) rendering and enables the comparison and analysis of relative differences between rendering techniques. This method incorporates bias and variance in the final score, and improves upon traditional single-value measures (e.g., absolute deviation (AD) or mean squared error (MSE)) by taking per-pixel radiance distributions into account. Thus, using it as a relative distribution-based measure alleviates the need for absolute comparisons against a reference image (RI), while still being able to reveal differences between renderers.

Limitations. Due to our relative comparison scheme, the choice of the reference renderer can have a significant impact on the scores. However, this is a problem that is also inherent to absolute comparisons with other metrics. Although our approach supports the comparison of results from different rendering systems in principle, it is required that one remains vigilant about the differences in scene descriptions, light and material models, and implementation.

Future Work. In this paper, we demonstrated the utility of our RS in creating more informative and meaningful comparisons between MC renderers by enriching the comparisons with bias and variance information. As a result, different rendering methods can now be compared with more confidence than with previous metrics, potentially uncovering implementation-based differences that may otherwise remain concealed. Furthermore, we hope that our proposed approach will be developed further in combination with standardized test scenes to establish more reliable and representative comparisons across publications in photorealistic rendering. We plan to release the source code of our approach in the hope of widespread adoption by the scientific rendering community.

Acknowledgments

Research reported in this technical report was supported by Austrian Science Fund (FWF): ORD 61

References

[Dal93] Scott Daly. Digital images and human vision. chapter The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity, pages 179–206. MIT Press, Cambridge, MA, USA, 1993.

[FP04] Jean-Philippe Farrugia and Bernard Péroche. A progressive rendering algorithm using an adaptive perceptually based image metric. *Computer Graphics Forum*, 23(3):605–614, 2004.

[GTGB84] Cindy M. Goral, Kenneth E. Torrance, Donald P. Greenberg, and Bennett Battaile. Modeling the interaction of light between diffuse surfaces. In *Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '84, pages 213–222, New York, NY, USA, 1984. ACM.

[Jak10] Wenzel Jakob. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org.

[Kaj86] James T. Kajiya. The rendering equation. *SIGGRAPH Comput. Graph.*, 20(4):143–150, August 1986.

[Kol33] A. Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.

[Kui60] Nicolaas H. Kuiper. Tests concerning random points on a circle. *Indagationes Mathematicae (Proceedings)*, 63:38 – 47, 1960.

[McN06] Ann McNamara. Exploring visual and automatic measures of perceptual fidelity in real and simulated imagery. *ACM Trans. Appl. Percept.*, 3(3):217–238, July 2006.

[MCTG00] Ann McNamara, Alan Chalmers, Tom Troscianko, and Iain Gilchrist. Comparing real & synthetic scenes using human judgements of lightness. In Bernard Péroche and Holly Rushmeier, editors, *Rendering Techniques 2000*, pages 207–218, Vienna, 2000. Springer Vienna.

[MDMS05] Rafal Mantiuk, Scott J. Daly, Karol Myszkowski, and Hans-Peter Seidel. Predicting visible differences in high dynamic range images: model and its calibration. In Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Scott J. Daly, editors, *Human Vision and Electronic Imaging X, San Jose, CA, USA, January 17, 2005*, volume 5666 of *SPIE Proceedings*, pages 204–214. SPIE, 2005.

[MKRH11] Rafal Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. HDR-VDP-2: a calibrated visual metric for visibility and

quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4):40:1–40:14, 2011.

[MN15] G. B. Meneghel and M. L. Netto. A comparison of global illumination methods using perceptual quality metrics. In *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 33–40, Aug 2015.

[Mys98] Karol Myszkowski. The visible differences predictor: Applications to global illumination problems. In *Rendering Techniques*, 1998.

[NMDSLC15] Manish Narwaria, Rafal K Mantiuk, Mattheiu Perreira Da Silva, and Patrick Le Callet. Hdr-vdp-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images. *Journal of Electronic Imaging*, 24(1):010501–010501, 2015.

[PFFG98] Sumanta N. Pattanaik, James A. Ferwerda, Mark D. Fairchild, and Donald P. Greenberg. A multiscale model of adaptation and spatial vision for realistic image display. In Steve Cunningham, Walt Bransford, and Michael F. Cohen, editors, *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1998, Orlando, FL, USA, July 19-24, 1998*, pages 287–298. ACM, 1998.

[RPG99] Mahesh Ramasubramanian, Sumanta N. Pattanaik, and Donald P. Greenberg. A perceptually based physical error metric for realistic image synthesis. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pages 73–82, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[RSSF02] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM Trans. Graph.*, 21(3):267–276, July 2002.

[SA07] Kartic Subr and James Arvo. Statistical hypothesis testing for assessing monte carlo estimators: Applications to image synthesis. In *Computer Graphics and Applications, 2007. PG'07. 15th Pacific Conference on*, pages 106–115. IEEE, 2007.

[SW04] Roland Schregle and Jan Wienold. Physical validation of global illumination methods: Measurement and error analysis. *Computer Graphics Forum*, 23(4):761–781, 2004.

[UWP06] Christiane Ulbricht, Alexander Wilkie, and Werner Purgathofer. Verification of physically based rendering algorithms. *Computer Graphics Forum*, 25(2):237–255, 2006.

[VMKK00] Valdimir Volevich, Karol Myszkowski, Andrei Khodulev, and Edward A. Kopylov. Using the visual differences predictor to improve performance of progressive global illumination computation. *ACM Trans. Graph.*, 19(2):122–161, April 2000.

[WBSS04] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004.

[WJM17] Joss Whittle, Mark W. Jones, and Rafał Mantiuk. Analysis of reported error in monte carlo rendered images. *The Visual Computer*, 33(6):705–713, 2017.