Perceptual quality of BRDF approximations: dataset and metrics

Guillaume Lavoué, Nicolas Bonneel, Jean-Philippe Farrugia, Cyril Soler

EUROGRAPHICS 2021 / N. Mitra and I. Viola
(Guest Editors)

Volume 40 (2021), Number 2

Perceptual quality of BRDF approximations: dataset and metrics

Guillaume Lavoué[1], Nicolas Bonneel[1], Jean-Philippe Farrugia[1], and Cyril Soler[2]

[1]CNRS, Univ. Lyon, LIRIS, France
[2]INRIA, Grenoble University, France

Figure 1: Several analytical approximations of the gold-metallic-paint3 MERL BRDF using classical models, with mean subjective opinion scores reflecting their perceived material similarity with the tabulated reference (between 1–very poor–and 5–excellent).

Abstract

Bidirectional Reflectance Distribution Functions (BRDFs) are pivotal to the perceived realism in image synthesis. While measured BRDF datasets are available, reflectance functions are most of the time approximated by analytical formulas for storage efficiency reasons. These approximations are often obtained by minimizing metrics such as L_2—or weighted quadratic—distances, but these metrics do not usually correlate well with perceptual quality when the BRDF is used in a rendering context, which motivates a perceptual study. The contributions of this paper are threefold. First, we perform a large-scale user study to assess the perceptual quality of 2026 BRDF approximations, resulting in 84138 judgments across 1005 unique participants. We explore this dataset and analyze perceptual scores based on material type and illumination. Second, we assess nine analytical BRDF models in their ability to approximate tabulated BRDFs. Third, we assess several image-based and BRDF-based (L_p, optimal transport and kernel distance) metrics in their ability to approximate perceptual similarity judgments.

CCS Concepts

• ***Computing methodologies*** → ***Reflectance modeling; Perception;***

1. Introduction

Appearance modeling is pivotal to realistic image synthesis. In the case of surface scattering, *bidirectional reflectance distribution functions* (BRDFs) are commonly used to encapsulate the reflection behavior of light. BRDFs of real-world materials can be captured and tabulated, however manipulating such datasets has many practical shortcomings, including inaccuracies due to the capture and sampling process, as well as large memory requirements. As such, more compact models aim to reproduce real-world reflectance as accurately as possible, avoiding exhaustive tabulation of the BRDFs. Common BRDF models include analytical representations [NDM05, WMLT07, BSH12] and models that leverage basis space expansions [GKD07, XSD*13, SBN15]. Each of these representations incurs an intrinsic approximation error, either due to the inability for their closed-form expression to model the variation of real-world reflectance profiles, or due to representational limitations of the basis (e.g., the bandlimiting nature of frequency-space bases). On top of this, additional approximation errors are due to the fitting algorithm that is used [FFG12], and are extrinsic to the model. Papers presenting a specific analytical model always provide visual comparisons but usually lack a perceptual study to compare to previous models over a large set of materials.

Assessing the impact of differences in the BRDF tabulated data on the perceived realism of an image is however not straightforward:

scene geometry, light source geometries and emission profiles, as well as the underlying simulation of light transport, all combine to form this final impression.

Image-based perceptual quality metrics [WBSS04,MKRH11] have been used in several works [HFM16,BP20] for BRDF quality evaluation. Image metrics may indeed represent a good solution to assess *a posteriori* the quality of a BRDF approximation; however, such metrics can hardly be used to *e.g.* predict the visual fidelity of a BRDF approximation *prior* to its use (without the knowledge of lighting, viewpoint and geometry). In addition, the actual correlation of those image metrics with the subjective opinion has never been quantitatively demonstrated. While various metrics can be computed using the BRDF samples themselves in closed form (such as weighted L_2 distances), to our knowledge none has been validated with perceptual user studies.

In this context, perceptual studies are still needed to understand the visual loss introduced by analytical BRDF approximations, in term of *perceived material difference*. Such perceptual evaluation is necessary for the creation of better metrics both for fitting analytical models and for automatically predicting this perceived difference.

In this paper we introduce a large dataset of 2026 isotropic BRDF approximations (including hidden references), associated with human judgments. Those BRDFs were obtained through approximations of 100 reference measured BRDFs [NDM05] by a variety of analytical models; they were then rendered, under two different illuminations [Deb04], to produce 2796 test images. A total of 84138 perceived similarity judgments (roughly 30 per test image) were acquired in a double stimulus rating experiment conducted using crowd-sourcing, where people were asked to rate the perceived similarity between approximated BRDFs and their corresponding unimpaired references, when used to render images. First, the collected data was used to explore the influence of material type (dielectrics or metals) and illumination on material similarity judgment. Second, we leverage the response dataset to benchmark nine analytical models with respect to the perceived quality of their approximations. Finally we use our data to evaluate the performance of multiple image quality metrics and BRDF quality metrics for predicting similarity scores. We summarize below our key contributions:

- we build a large dataset of perceptual isotropic BRDF approximation quality ratings from a crowdsourced experiment.
- we analyze the effect of illumination and material type (dielectric vs. metal) on material perception in light of this dataset.
- we assess the fidelity of nine widely used analytical BRDF models depending on material types.
- we benchmark several metrics computed both in the BRDF sample space and in screen-space for their ability to predict perceptual degradations.
- we introduce new BRDF metrics, notably based on optimal transportation, that outperform state-of-the-art metrics.

The supplementary materials, code and the dataset can be found at http://liris.cnrs.fr/glavoue/data/BRDFs/

2. Previous Work

For their ability to model realistic materials using a simple representation, BRDF models have attracted much attention and are now prevalent despite their lack of support for specific physical phenomena such as subsurface scattering or fluorescence. This section describes related work on the use of perceptual distances in the context of BRDFs, and more generally, a state-of-the-art in material visual perception.

2.1. Perceptual models and BRDF metrics

Since the advent of approximation methods for BRDF models, the need for measuring how well an approximation fits a reference BRDF has kept appearing. Two main directions have been taken: approaches working directly in the space of BRDF samples, and those working on rendered images. While the formers are more flexible and independent of the 3d scene, the latters usually result in more perceptually accurate measures.

BRDF samples space distances. Early methods were based on a simple cosine-weighted L_2 distance between BRDF values [LFTG97,NDM05,LKYU12] or a dimensionality-reduced version of this metric [PL07]. To assess the relevance of this distance, Fores et al. [FFG12] perform a perceptual experiment to determine which one of three metrics in BRDF space (RMS, cosine-weighted L_2, and cube root cosine-weighted RMS) is able to produce the best approximation when used for fitting BRDFs on three different analytical models. They conclude that a cube root cosine-weighted RMS provides higher visual fidelity, especially for very specular materials. More complex metrics have been introduced, such as the E_2 metric of Löw et al. [LKYU12]. The E_2 metric is a log, cosine-weighted L^2 error, which has notably been used for fitting BRDFs generated procedurally with genetic algorithms [BLPW14]. Serrano et al. [SGM*16] propose a control space for predictable editing of captured BRDF data. Relying on a large-scale experiment on an extended version of the MERL dataset [MPBM03], they allow for assessing similarity between BRDFs regarding several aspects of appearance, such as brightness, strength or sharpness of reflections, by mapping these attributes to an underlying PCA-based representation of BRDFs with 5 principal components. The fitting is made using a radial basis function (RBF) network with one hidden layer. However, approximation methods rely on a single value that serves as the objective function, and combining all these aspects into a single number describing the proximity in term of material appearance is not trivial.

Image space distances. Measuring perceptual distances between images is a well-researched area [Dal92, WBSS04, MKRH11]. Measuring perceptual distances on BRDFs using rendered images has thus been extensively used. The simpler of these approaches is that of Ngan et al. [NDM06] who render spheres of a given BRDF lit with a natural environment map (the Grace Cathedral [Deb04]), and use the L_2 distance between the cube root of linear RGB channels. They achieve similar results as the standard L_2 distance between channels in the LAB color space (without cube root). This approach has been extended to instead use a Structural Similarity Image Metric (SSIM) [WBSS04] on gamma-corrected images by Brady et al. [BLPW14]. Very recently, Bieron and Peers [BP20] use the CSSIM color metric [LPU*13] to optimize their fitting, in a two-step process. By using isotropic BRDFs and a symmetric mathematically defined environment map, Pereira and Rusinkiewicz [PR12] obtain a near closed-form expression of a rendered sphere, allowing to render spheres using a fast matrix-vector multiplication. These spheres are, this time, compared using an L_4

norm to obtain a BRDF similarity metric. Havran et al. [HFM16] evaluate different image space metrics in the context of optimizing shapes to optimally depict BRDFs. They automatically design a parametric surface able to represent interesting BRDF variations, and show that ΔE and HDRVDP2 [MKRH11] both discriminate materials comparably well. Bousseau et al. [BCRA11] design an image space metric aimed at optimizing environment maps so that renderings exhibit particular material properties, such as shininess for metals, grazing reflections for Fresnel materials, or grazing highlights for asperity scattering. The closest work to ours is the work of Lagunas et al. [LMS*19]. They conducted a large scale experiment to evaluate the perceived similarity between BRDF models and then used the perceptual judgement to train a new image-based similarity metric. However, their dataset is mostly an image dataset, since it considers the 100 MERL BRDF models rendered using different scenes and illumination. In contrast, our dataset contains 2026 BRDF models.

Other perceptual embeddings. Pellacini et al. [PFG00] obtain a perceptual embedding of BRDFs generated from the Ward isotropic model in a space of gloss by performing a user study. They use Multi-Dimensional Scaling (MDS) on apparent gloss differences between pairs of renderings, and conclude that two dimensions explain most variations and correspond to contrast gloss and distinctness-of-image gloss. The Ward BRDF model is then re-parameterized to achieve perceptually uniform gloss variations. Wills et al. [WAKB09] perform a similar experiment on measured BRDFs. They use non-metric MDS and ask participants to judge which one of two renderings is more similar to a third one. An application is to interpolate between different BRDFs in a perceptual way by linearly interpolating within triangles of a Delaunay triangulation of the embedding. A similar approach could in principle be used to infer perceptual distances from the Delaunay triangulation but for this method to capture subtle differences between a reference material and its approximation would require an extremely dense sampling of the space, that is, an intractable experiment involving more than the hundred BRDFs we have at our disposal [NDM05].

2.2. Psychological evaluation of material perception

Material perception is a high level process that cannot be directly accounted for by only pixel-based or sample-based considerations. Shape and lighting further play a prominent role in the perception of materials, which complicates its study. In fact, for gloss perception, both shape and lighting interact in a joint and hard-to-predict way [OB11]. Fleming [Fle14] suggests the brain works out a statistical model by discovering relationships between samples, e.g., by observing how the image is changing as material, shape, and illumination properties are changing and looking at image features, instead of trying to predict physical parameters themselves.

Shape. Initial experiments by Nishida and Shinya [NS98] where the material of a heightfield surface had to be matched to that of another heighfield showed the difficulty of the task when the heighfields were very different. Instead of matching reflectances directly, Vangorp et al. [VLD07] studies the problem at the higher level of material recognition. They ask participants to tell whether two rendered objects are made of the same material, and they vary the objects shape. Their results provide guidelines for shapes to be used for this task: they show that the commonly used sphere is one of the least discriminating shape and tesselated geometries do not work well, while smooth but curved geometries such as a blob or a Buddha model work much better. While the question investigated in their study is relatively close to ours, it unfortunately does not directly allow for assessing BRDF approximation models or metrics.

Lighting. Environment lighting also plays a significant role. Fleming et al. [FDA03] show that natural illumination is an important factor for depicting materials, and that a human observer can infer materials directly from statistical image features. They design an experiment where a sphere is lit from captured environment maps or artificial point or rectangular light sources, and had participants match materials using a Ward BRDF model. They show that captured environment maps allow participants to more accurately match materials, except for the artificial rectangular light source that performed comparably well for Ward's BRDF roughness parameter. Since this pioneering work, many studies (e.g., [VBF17, TF18, ZdRBP19]) confirmed this major role of environment lighting in our perception of materials. In the same spirit at Vangorp et al., Ramanarayanan et al. [RFWB07] evaluate the high level perception of materials when varying environment maps. They develop the concept of "visual equivalence": two images are deemed visually equivalent if, when seen side-by-side, the objects they depicts are perceived as having the same shape and materials, and one cannot tell which one has been rendered with the reference environment map. From a user study, they are able to design a "Visual Equivalence Predictor" (VEP) using a Support Vector Machine classifier on experimental data. Krivanek et al. [KFB10] study visual equivalence (and image quality) between a reference rendering and a degraded rendering that uses only a limited number of virtual point lights (VPLs). They vary materials, geometry and lighting conditions, and find that more geometrically complex glossy shapes and lighter dielectrics are more forgiving of illumination errors while metals are generally unforgiving. Using actual painted glass samples of different materials, Leloup et al. [LPDH10] confirm the importance of illumination on gloss perception and that commercial gloss meters did not accurately predict gloss perception.

3. BRDF dataset

In this section, we describe the dataset we created to assess BRDF perceptual differences. Our dataset consists in 100 source BRDFs, subject to approximations with different models, producing a total of 2026 BRDFs (including references).

3.1. Source BRDFs

Our source BRDFs are the real-world tabulated data of the MERL database [MPBM03]. The database features a total of 100 measurements of isotropic materials sampled over $90 \times 90 \times 180$ couples of directions. For further analysis, we manually categorized those materials into dielectrics (70 out of 100) and metals (30 out of 100). Our classification is publicly available with the dataset.

3.2. Analytical models

We perceptually evaluate the analytical approximations of MERL tabulated data for the following isotropic analytical models: Blinn-Phong [Bli77], ABC [LKYU12], Ward [War92], Beckmann

with Gaussian normal distribution [BS87], Lafortune [LFTG97], Rational-Chebychev [PSCS*12], Rational-Legendre [PSCS*12], SGD [BSH12], GGX [WMLT07], Bagher [BSN16] and Dupuy tabulated approximation [DHI*15].
Ward, Blinn-Phong, Lafortune, ABC and Beckmann models were fitted using Hooke-Jeeves minimization with $D4.Cl$ metric (see Section 6 for details). For Ward, Blinn-Phong, and Lafortune we considered a second approximation by taking the parameters from Ngan's paper [NDM05] (available for 86 out of the 100 MERL BRDFs). For ABC, we also considered a second approximation by taking the parameters from the authors paper [LKYU12]. The rational Chebychev and Legendre models were fitted using the ALTA library [BCP*15], while SGD, GGX and Dupuy were all fitted using Dupuy's BRDF fitting library (available at https://github.com/jdupuy/dj_brdf). The approximations from Bagher [BSN16] were provided by the authors.
Figure 2 gives for each analytical approximation the number of different materials used in our study. The "Reference" material (at right) refers to the original tabulated data that we also include in our tests for sanity check, as a "hidden reference" during the experiment. As can be seen in the figure, we were not able to obtain rational Chebychev and Legendre fits for all MERL BRDFs, due to inherent limitations of the fitting library.

3.3. Manifold approximations

In addition to analytical models, we included in our experiment a number of BRDF approximations obtained by sampling the MERL manifold (as defined by Soler *et al.* [SSN18]) close to the original BRDFs. Since this manifold exactly interpolates the input data, points that are taken close to the original MERL BRDFs in the latent space provide an interesting set of very realistic-looking tabulated BRDFs, extending the space of tabulated data over which the different metrics will be compared. The approximations of a BRDF at latent position c are chosen to be at positions

$$\forall i \in [1,N], \mathbf{c}_i = \mathbf{c} + \mathbf{d} * \left(\frac{i}{N}\right)^5 r,$$

where $\mathbf{d}$ is a random unit-vector in the parameter space of the manifold, N is the total number of approximations produced for this material, and r is the diameter of the parameter space.
Six MERL Manifold Samples (MMS) c_i were constructed for each reference BRDF (see Figure 2).

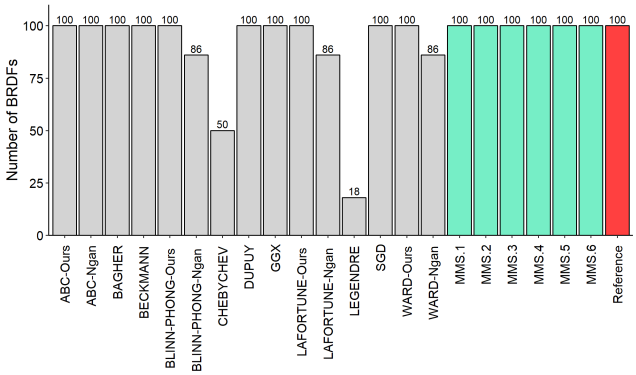


Figure 2: Distribution of our BRDF dataset in terms of approximating methods.

3.4. Stimuli creation

To create stimulus images from our dataset of BRDFs, we apply each of these BRDFs to a 3D shape, and render the corresponding 3D scene under specific illumination conditions.

Shape and scene. Vangorp et al. [VLD07] specifically investigated how the shape of an object influences the perception of its material. They demonstrate that the sphere is not well suited to material discrimination, while more complex geometries such as a blob or the Buddha model are much better suited. We follow their recommendations and selected the Buddha model. This model has the benefit of representing a realistic shape (a statue) and exhibiting a wide range of geometric features: smooth parts, high frequencies as well as creases. We created a 3D scene where this 3D model is placed on a wood table, in order to have a plausibly realistic context. We indeed hypothesise, like many studies related to material perception [VLD07, RFWB07, KFB10], that keeping a realistic setting is an important factor to help naive observers to understand the notion of "material".

Illumination and rendering. As raised by Fleming et al. [FDA03], a natural illumination is a critical factor for depicting materials. We thus used captured environment maps instead of artificial light sources. In order to ensure the generality of our subjective results and evaluate the impact of the lighting environment, we selected two high resolution environment maps from https://vgl.ict.usc.edu/Data/HighResProbes/: *Uffizi* and *Grace* which are respectively low and high frequency (see Figure 8). Our choice was driven by the will to have two maps with opposite characteristics (both in terms of colors and frequency). Our whole dataset (2026 BRDFs) was rendered under the Grace environment map, while approximately a third of it (770 BRDFs) was rendered under the *Uffizi* environment map, which allowed us to effectively reduce the size of the experiment. We privileged *Grace* for rendering all the stimuli because it comparatively spreads over a larger interval of frequencies, hypothesizing this lighting to be less forgiving for the approximation methods and thus stimulating a more significant interval of user responses. A total of 2796 *test* images were thus rendered. All images were rendered at 580×900 using Monte Carlo path tracing. For display, images were tone-mapped using a gamma value of 1.25 using the algorithm of ILM's 'exrdisplay' HDR viewer [ilm] (See pseudo-code in the supplementary material). We also rendered the 100 *reference* BRDFs using both light probes, resulting in 200 *reference* images that will serve as ground-truth material images during the experiment. Note that, as recommended by Krivanek et al. [KFB10], those *reference* images were taken from slightly different camera positions, to make the task object-focused rather than image-focused, i.e., avoiding participants to compare specific pixel values rather than giving their appreciation of the material itself. Figure 3 presents some of our stimuli images.

4. Methods

In this section, we describe our crowdsourced psychophysical study. The objective is to evaluate how similar each approximated BRDF is to its corresponding source BRDF. Participants of our

Reference GGX Reference Dupuy

Figure 3: Examples of stimuli from our dataset. Shown BRDFs are *blue-metallic-paint2* and *gray-plastic*. Reference scenes are rendered using a slightly different viewpoint.

study were asked to rate the perceived similarity between series of *test* images and corresponding *reference* images. This protocol is formally known as the Double Stimulus Categorical Rating (DSCR) commonly used for 2D image and video quality assessment. The experimental procedure is described below.

4.1. Rating protocol

Our goal is to measure the perceived fidelity of approximated BRDFs with respect to their unimpaired reference. We chose to rely on a categorical rating method, the Double Stimulus Categorical Rating (DSCR). Participants were presented with pairs of static images: a *test* image rendered from an approximated BRDF and a *reference* image rendered from the corresponding original BRDF. They were then asked to *rate the similarity between the statues' materials on the left and right images* (see Figure 5). As recommended by the ITU standard BT500-11, we used a five-grade quality scale numbered from 1 (very poor) to 5 (Excellent).

Note that paired comparison methodology have been demonstrated to be more reliable than rating methods in certain conditions [MTM12]. However they are not tractable for large numbers of stimuli, since they require $\binom{n}{2}$ comparisons. This high number of trials could be reduced by using sorting algorithms as recommended by Silverstein et al. [SF01] and Mantiuk et al. [MTM12], but such sorting algorithms cannot be implemented in crowdsourcing experiments where workers do only a small part of the task. For these reasons, categorical rating methods are mostly preferred for large datasets and particularly for crowdsourced studies (e.g., [SGM*16, GB16]).

4.2. Instructions, Training, and Testing

We employ the Appen platform (formerly known as CrowdFlower and Figure 8), which allows, similarly to Mechanical Turk, to provide micro-tasks to selected pools of registered "workers". In our study, the minimal micro-task consists of one page containing 10 pairs of images to rate and is paid $0.20. However, participants can choose to rate several pages in the limit of 150 judgments (i.e., 15 pages).

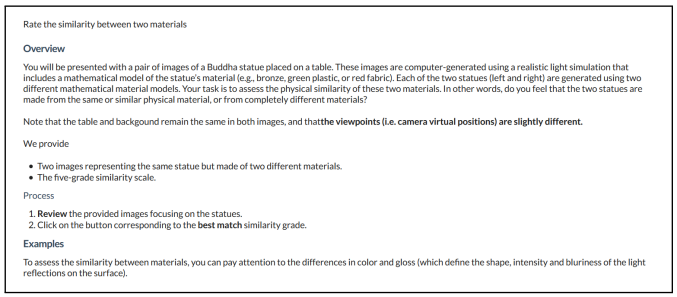
Each participant is initially given instructions, and 6 example pairs of images representative of the approximate range of material qualities are shown along with their corresponding expected answer. Instructions are illustrated in Figure 4, and example pairs are available in the supplementary material.

To filter careless participants, we combine four mechanisms:

1. Only workers of "level 3" (internal rating of CrowdFlower) are allowed to take part in the experiment; this correspond to the highest quality and most experienced workers.
2. Before entering the study, participants first have to complete a qualification test consisting of 10 *gold standard* images with known ground-truth. Participants with less than 85% correct answers are disqualified. To be considered incorrect, a response should be *very* different from the expected answer – the goal being to avoid careless participants while keeping the natural variability within careful participants. These *gold standard* images are randomly selected among a pool of 44.
3. In each micro-task, one (randomly selected) *gold standard* image is inserted. An average accuracy of 85% have to be maintained by the participant all along his/her tasks.
4. Participants who are too quick to respond (less than 10 seconds for 10 ratings) are discarded.

4.3. Participants

In total we collected 84138 judgments from 1005 unique participants after excluding trials that failed the quality checks described above. Each participant rated 83.7 pairs on average (SD = 56.3). Each pair was rated by 30.1 participants on average (SD = 0.4). Participants were very positive about the test, 143 of them gave their feedback reporting an average satisfaction of 4.2/5, with 3.9/5 for *ease of job*.



Rate the similarity between two materials

Overview

You will be presented with a pair of images of a Buddha statue placed on a table. These images are computer-generated using a realistic light simulation that includes a mathematical model of the statue's material (e.g., bronze, green plastic, or red fabric). Each of the two statues (left and right) are generated using two different mathematical material models. Your task is to assess the physical similarity of these two materials. In other words, do you feel that the two statues are made from the same or similar physical material, or from completely different materials?

Note that the table and backgound remain the same in both images, and that**the viewpoints (i.e. camera virtual positions) are slightly different.**

We provide

- Two images representing the same statue but made of two different materials.
- The five-grade similarity scale.

Process

1. **Review** the provided images focusing on the statues.
2. Click on the button corresponding to the **best match** similarity grade.

Examples

To assess the similarity between materials, you can pay attention to the differences in color and gloss (which define the shape, intensity and bluriness of the light reflections on the surface).

Figure 4: Instructions shown at the beginning of the task.

5. Results and observations

5.1. Computing mean opinion scores

A common way of analyzing the opinion scores of a double stimulus subjective test is to compute the Mean Opinion Score (MOS) of each stimulus:

$$MOS_s = \frac{1}{N}\sum_{i=1}^{N} r_s^i \qquad (1)$$

where r_s^i is the rating given by participant i, to stimulus s. Those mean opinion scores MOS_s are associated with 95% confidence intervals.

5.2. Inter-participant reliability

It is essential to analyze the agreement between participants before studying the results from the experiment. As in the work of Ghadiyaram and Bovik [GB16], we split the subjective ratings obtained on every pair of images into two disjoint equal groups, and calculated the correlation between the recovered mean opinion scores

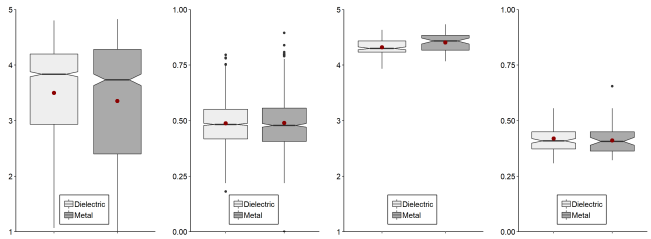
Figure 5: Illustration of the interface of our experiment, asking participants to rate on a scale of 1 to 5 the similarity between the statue's materials on the left and right images.

of the two groups. When repeated over 100 random splits, we obtained an average Pearson linear correlation coefficient of 0.964 (SD = 0.001) and an average Spearman rank order correlation coefficient of 0.929 (SD = 0.002). Those high values indicate that there is a high degree of agreement between the subjects despite the fact that the experiment was conducted via crowdsourcing. We also computed the intraclass correlation coefficient (ICC) [Bar66]; results (ICC= 0.981, 95% CI: 0.980 < ICC< 0.982) confirm the high agreement between raters.

5.3. Effect of material type

Figure 6 illustrates boxplots of MOS values and confidence intervals according to the material type, for all BRDFs (left) and hidden reference BRDFs only (right). No effect of material type on standard deviations was found, meaning that the type of material has no effect on the agreement of observers. However, differences can be observed on MOS values: approximations of dielectrics are globally perceived as of slightly better quality than approximations of metals; this difference is found as statistically significant by Welch's t-test (p-value$<$0.001). We will see in Section 5.5, that this difference of perceived quality largely depends on the approximation models and their ability to reproduce the specular nature of the material. To confirm this effect, we conducted a two-way analysis of variance (ANOVA for material type $\times$ approximation model) on MOS values and found a significant interaction (p-value$<$0.001), meaning that the performance of approximation models depend on material type.

When considering only the 100 *reference* BRDFs, we also found a significant difference of ratings between both types of material (p-value$<$0.001). In this case, metals are rated slightly higher than dielectrics (average MOS=4.41 and 4.32 respectively). This means that people tend to better recognize a same material, under different viewpoints, when it has a metallic nature. This tends to confirm the hypothesis of Fleming [Fle14] that the visual system mostly relies on characteristic signatures of specular reflections to identify a material (see example images in the supplementary material).



MOS - All stimulus | CI - All stimulus | MOS - References | CIs - References

Figure 6: Boxplots of mean opinion scores (MOS) and confidence interval width (CI), according to the type of material.

5.4. Effect of illumination.

Krivanek et al. [KFB10], Leloup et al. [LPDH10] and Fleming et al. [FDA03, Fle14] shown that the nature of illumination is a critical factor for depicting materials; in particular, Fleming et al. [FDA03, Fle14] emphasized the importance of a *natural* illumination. Our dataset allows to evaluate the effect of two radically different natural lighting environments.

We selected the set of 740 BRDFs lit by both *Uffizi* and *Grace* maps and analyzed the MOS values of the corresponding 1540 rendered images. Figure 7 illustrates the correspondences of the two sets of MOS values. Whereas the illumination map used for rendering seems to have a certain impact on the perception of certain BRDFs (i.e., 2D points are not perfectly aligned), no significant and systematic influence on the quality can be observed (the p-value from a paired t-test between the two illumination conditions is 0.93). Pearson and Spearman correlation coefficients between MOS values from both environment maps were found to be 0.936 and 0.910, respectively. No significant effect was found on standard deviations (p-value=0.37).

If we restrict our analysis to the 100 *reference* BRDFs; it is interesting to notice that we observe a significant impact of the environment lighting on the MOS values (p-value$<$0.001). Globally, people tend to better recognize the same material, under different viewpoints, using the Grace environment map (average MOS=4.39) than using Uffizi (average MOS=4.30). Our hypothesis is that the high frequency patterns from Grace (see Figure 8 for the frequency content of both maps) introduce significant reflections that help the user to better recognize the material, despite the slight difference in viewpoint (see example images in the supplementary material).

To further explore the influence of illumination, we conducted a three-way ANOVA (illumination $\times$ material type $\times$ approximation model) on MOS values. The significant interaction between the approximation model and the type of material was confirmed (p-value$<$0.001). We also found a slight interaction between the illumination and the type of material (p-value=0.0151). This interaction is illustrated in Figure 8, right. It seems that the difference of perceived quality between metals and dielectric is dependent on the illumination. Finally, no significant interaction was found between the approximation model and the illumination, meaning that the performance of an analytical model in terms of perceived quality was not found to be dependent on the illumination.

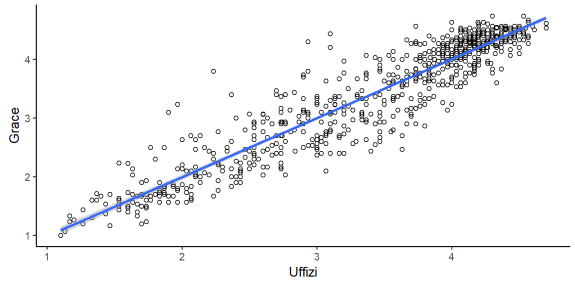


Figure 7: Comparison of quality scores obtained with the two different environment maps.

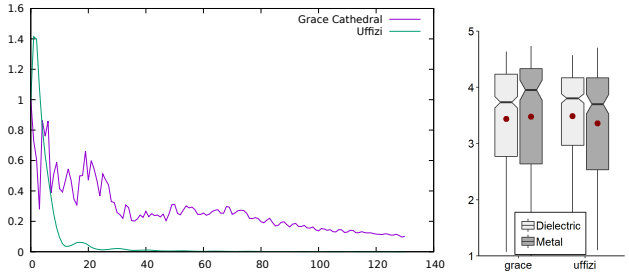


Figure 8: *Left*: Energy in each frequency band for Uffizi and Grace environment maps, normalized to the total energy of the map, computed using spherical harmonics up to order 130. Uffizi clearly shows up as a low frequency profile whereas Grace contains lots of high frequencies. *Right*: Boxplots of MOS values according to material type and environment maps, showing the low dependency of perceived material similarity on the illumination bandwidth.

5.5. Evaluation of analytical fits

Given the lack of impact of the illumination on the MOS values, and the lack of interaction between approximation model and environment lighting, we consider only the 2026 images rendered with the *Grace* environment map for the evaluation of analytical models. This allows for more robust statistics via paired comparisons.

Figure 9 illustrates boxplots summarizing MOS values for each analytical model, and Fig. 10 separates these MOS values by material type. For this analysis we consider nine analytical models plus the hidden references. Each presented model was fitted on the 100 MERL BRDFs. This analysis does not include rational-Chebychev and Legendre models because they were not fitted on all the MERL dataset. For ABC, Ward, Blinn-Phong, and Lafortune, for which we have two versions of each approximation (ours using Hooke-Jeeves method and NGan's parameters [NDM05]), we select the best one (i.e., associated with the highest MOS) for each MERL BRDF. As a preamble to our analysis, it is important to note that, in general, BRDF models with a larger number of degrees of freedom are expected to better fit tabulated data. This is particularly important for semi-tabulated models such as that of Bagher et al. [BSN16], Dupuy et al. [DHI*15], or high-dimensional models such as SGD [BSH12], since the memory usage of BRDF models can be limited in rendering applications, and a quality vs. memory tradeoff has to be found. For instance, a typical BRDF of Dupuy et al. takes 6KB in memory. We believe our analysis is important in this regard, and as such, we first report the number of parameters of the analyzed BRDF models. The Beckmann model has 8 parameters, while GGX has 8, ABC 8, Ward 7, Lafortune 9, Dupuy et al. 8, Blinn-Phong 7, SGD 33. Bagher et al.'s model is non-parametric by design and uses tabulated functions.

Unsurprisingly, semi-tabulated models outperform others, the model of Bagher et al. [BSN16] having the best MOS values overall (p-value$< 10^{-9}$ for all pairwise comparisons), for both dielectrics and metals. Among low-dimensional models (< 10 parameters), Blinn-Phong performs surprisingly well on dielectrics given its simplicity, while metals are best fitted by the GGX model. Most analytical models are more appropriate for a specific type of material: ABC, Ward, Lafortune, Blinn-Phong, and SGD perform better for dielectrics than for metals, while GGX, Bagher et al. [BSN16] and Dupuy et al. [DHI*15] perform better on metals than on dielectrics.

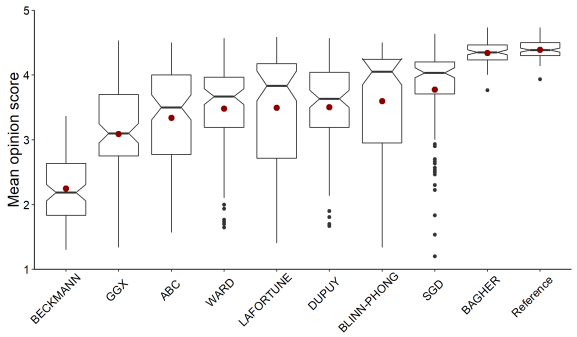


Figure 9: Boxplot of MOS values, for 9 analytical models, for the MERL BRDFs. Results for hidden references are also included.

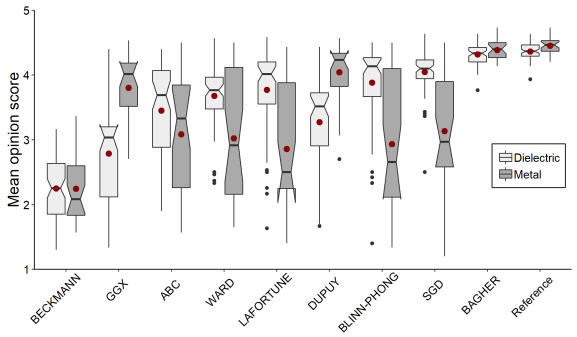


Figure 10: Boxplot of MOS values, for 9 analytical models, according to the material type. Results for hidden references are also included.

In Figure 11, we plot perceptual scores as a function of the distance value used to generate synthetic BRDFs (MERL Manifold Samples) via the Gaussian process latent variables model of Soler et al. [SSN18]. We can see that perceptual quality correlates well with distances on the learned BRDF manifold.

6. Evaluation of objective metrics

For this analysis, we consider all the images rendered using the Grace environment map, except the 100 hidden reference images. We evaluated several metrics, operating on both the BRDF space and the image space.

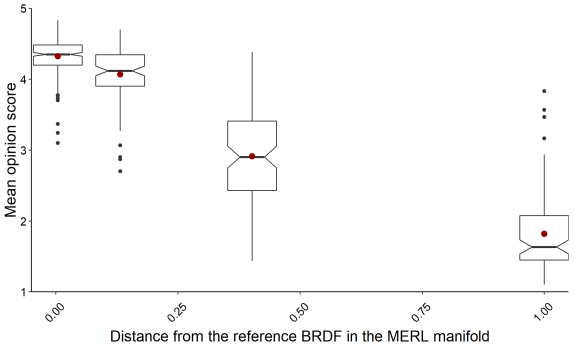


Figure 11: Boxplot of MOS values for the MERL Manifold Samples [SSN18], according to the distance from the reference BRDF (normalized by the diameter of the parameter space). A Pearson correlation coefficient of 0.90 indicates a strong linear correlation between distances in the manifold and perceived quality.

6.1. Selected metrics

Image metrics. The field of image quality assessment has been extremely active over the last decade. Hundreds of quality metrics can be found [Zha12]. For the present study, we selected metrics that have already been used by the computer graphics community for the evaluation and/or fitting of materials:

- HDR-VDP-2 [MKRH11]: this bottom-up technique tries to mimic the low-level mechanisms of the human visual system (HVS) such as the *contrast sensitivity function* (CSF). This metric is able to predict both artifact visibility and global quality. It has been used by Havran et al. [HFM16] to discriminate materials. As output, we selected the quality prediction Q proposed in the extension by Narwaria et al. [NPC*15].
- Structural SIMilarity index (SSIM) [WBSS04]: this is a top-down metric, which does not take into account any HVS model, but relies on some local Luminance statistics, related to the structure of the images. This metric has been used by Brady et al. [BLPW14] to evaluate the quality of analytical BRDFs.
- Color Image Difference (CID) [LPU*13, PFU14]: since the above metrics only consider luminance, we selected this color-based approach, which integrate chromatic information in a feature-based approach similar to SSIM. This metric, also called CSSIM, has been used by Havran et al. [HFM16] to discriminate materials, and by Bieron and Peers [BP20] for their fitting algorithm.
- Learned Perceptual Image Patch Similarity (LPIPS) [ZIE*18]: this is a deep-learning approach which relies on *deep features*, i.e., internal activations of convolutional networks trained for high-level classification tasks. This approach has shown to achieve state-of-the-art results for perceptual image quality assessment. The authors compare multiple different networks and models; we selected the VGG [SZ14] and Alexnet [Kri14] models and used the metrics as specified in the released code, without any re-training.
- Material Appearance Similarity Measure (MatSim): this metric has been introduced by Lagunas et al. [LMS*19] for predicting the similarity between materials, from rendered images. It is a deep-learning approach based on the ResNet network [HZRS16] and trained on a dataset of images of different materials (from MERL), associated with similarity judgments from 2AFC tests. We used the metric as specified in the released code, without any re-training.
- We also included a baseline color distance (CIELAB): the quadratic distance in the CIELab color space.

L_p-based BRDF metrics. For the present study, we used the three BRDF metrics from [FFG12] (D_1, D_2, and D_3), plus six additional metrics from other sources listed below. They are defined as follows using for angles the notations of [BSH12]:

- $D_1 = \sqrt{\frac{\sum_{\theta_i,\theta_o,\phi_d}(f_r(\theta_i,\theta_o,\phi_d)-f_a(\theta_i,\theta_o,\phi_d))^2}{N}}$
  This corresponds to the standard root mean square error over tabulated data
- $D_2 = \sqrt{\frac{\sum_{\theta_i,\theta_o,\phi_d}(f_r(\theta_i,\theta_o,\phi_d)\cos(\theta_i)-f_a(\theta_i,\theta_o,\phi_d)\cos(\theta_i))^2}{N}}$
  This is the cosine weighted version of D_1. The cosine input angle is used to compensate reflection increase at grazing angles.
- $D_3 = \sqrt{\frac{\sum_{\theta_i,\theta_o,\phi_d}\sqrt[3]{(f_r(\theta_i,\theta_o,\phi_d)\cos(\theta_i)-f_a(\theta_i,\theta_o,\phi_d)\cos(\theta_i))^2}}{N}}$
  This is the cubic root of D_2, it intends to attenuate peak values in mirror direction and amplify off-peak values.
- $D_4 = \sqrt{\frac{\sum_{\theta_i,\theta_o,\phi_d}(f_r(\theta_i,\theta_o,\phi_d)-f_a(\theta_i,\theta_o,\phi_d))^2\cos(\theta_o)\sin(\theta_o)}{N}}$
  This is our own implementation of the L_2 distance between BRDFs with a correct Jacobian term to represent spherical integration.
- $D_5 = \frac{\sum_{\theta_i,\theta_o,\phi_d}|f_r(\theta_i,\theta_o,\phi_d)-f_a(\theta_i,\theta_o,\phi_d)|}{N}$
  This is the common L_1 distance.
- $D_6 = \sqrt{\frac{\sum_{\theta_i,\theta_o,\phi_d}(f_{rLab}(\theta_i,\theta_o,\phi_d)-f_{aLab}(\theta_i,\theta_o,\phi_d))^2\cos(\theta_o)\cos(\theta_i)}{N}}$
  This metric is the projected area weighted CIELAB metric of Ryman et al. [Rym18].
- $D_7 = \sqrt{\frac{\sum_{\theta_i,\theta_o,\phi_d}\sqrt[3]{(f_r(\theta_i,\theta_o,\phi_d)\,c(\theta_i,\theta_o)-f_a(\theta_i,\theta_o,\phi_d)\,c(\theta_i,\theta_o))^2}}{N}}$
- $D_8 = \frac{\sum_{\theta_i,\theta_o,\phi_d}\left|\log\left(\frac{f_r(\theta_i,\theta_o,\phi_d)\,c(\theta_i,\theta_o)+10^{-3}}{f_a(\theta_i,\theta_o,\phi_d)\,c(\theta_i,\theta_o)+10^{-3}}\right)\right|}{N}$
- $D_9 = \sqrt{\frac{\sum_{\theta_i,\theta_o,\phi_d}\left(\log\left(\frac{f_r(\theta_i,\theta_o,\phi_d)\,c(\theta_i,\theta_o)+10^{-3}}{f_a(\theta_i,\theta_o,\phi_d)\,c(\theta_i,\theta_o)+10^{-3}}\right)\right)^2}{N}}$
  D_7, D_8 and D_9 are the cubic root, log1 and log2 metric of Sun et al. [SJR18], where $c(\theta_i,\theta_o) = \max(\cos(\theta_i)\cos(\theta_o), 10^{-3})$

In these equations, f_r denotes the reference BRDF and f_a denotes the approximated BRDF, both linearly tabulated along the $\theta_i, \theta_o, \phi_d$ angles. N denotes the number of samples in each BRDF, related to the level of discretization of θ_i, θ_o and ϕ_d (we chose $N = 90 * 90 * 360 * 3$ for the MERL database since it comes parameterized with the same steps on half-angles [MPBM03]). f_{rLab} and f_{aLab} denotes the same functions but with values described in Lab colorimetric space instead of RGB. Note that some of these existing metrics ignore the Jacobian that would be needed to let the distance be a discretised hemispherical integral. Note that using an isotropic parameteriation with $(\theta_i, \phi_i, \theta_o, \phi_o)$ can be restricted to

$\phi_i = 0$. Therefore we prefer to use $\phi_d = \phi_o - \phi_i$ in the above equations, which doesn't impact Jacobians.
We computed the results from these metrics with four variations on the input data:

- **No processing**: the BRDF data is left as it is;
- **Cube root**: we take the cubic root of the input BRDF data in order to attenuate peak values, as suggested by some papers;
- **Clamping**: we discard grazing θ_i and θ_o angles above 80 degrees, as performed by Ngan et al. [NDM05];
- **Clamping+cube root**: This combines the clamping and cube root strategies.

This leads to a total of 36 different metric variations.

Kernel and Optimal Transport-based BRDF metrics. Optimal transport is a well-researched area to compare probability distributions by minimizing the effort required to move a pile of sand shaped as the first distribution towards a hole shaped as the other distribution [PC*19]. This effort is computed as the sum over all mass particles of the cost of moving it from location X to location Y. This cost, $c(X,Y)$ is called the ground metric. Optimal transport has seen many applications including BRDF interpolation [BvdPPH11], though to the best of our knowledge, it has not been used as a metric between BRDFs.
Extensions have been proposed to compare arbitrary functions, and in particular, the transportation-L^p distance [TPK*17] which effectively amounts to computing an optimal transport problem on the graph of the functions being compared. We propose to use this approach to compare BRDFs, and thus compute the optimal transport between two 4-d discrete measures[†] of the form $\sum_i \delta_{X_i}$, where X_i has coordinates (x,y,z,w) defined as

$$x = \alpha\frac{\theta_i}{\pi/2} \qquad y = \alpha\frac{\theta_o}{\pi/2},$$
$$z = \alpha\frac{\phi_d}{2\pi} \qquad w = \varphi(f(\theta_i,\theta_o,\phi_d))$$

where the factor α allows to weight differently the base space (x,y,z) and the function value w, and φ a function that compresses BRDF values f and is detailed next.
Optimal transport computation being a costly minimization, to efficiently solve it, we rely on the GeomLoss GPU library [Fey19] for a fast approximation. We effectively compute a Sinkhorn divergence [FSV*19], which introduces a regularization parameter ε and multiscale scaling factor s. To speed computations, we use a $45\times45\times180$ grid for $(\theta_i,\theta_o,\phi_d)$ instead of $90\times90\times360$ as before. Optimal transport computations were performed for each color channel of the BRDFs. The three resulting values W_r, W_g, W_b were combined into a single scalar value as $(W_r^+ + W_g^+ + W_b^+)^\gamma$, where x^+ denotes $max(x,0)$.
We test several ground metrics $c(X,Y)$, weightings α, regularizations ε, functions φ and scalings s. Ground metrics where chosen as either a square distance $c(X,Y) = \|X-Y\|^2$ or a distance $c(X,Y) = \|X-Y\|$. Values for α were chosen in $\{0.01,0.05,0.1,1\}$, values for ε in $\{0.01,0.02,0.05,0.1\}$, values for s in $\{0.8,0.9,0.95,0.99\}$, values for γ in $\{0.25,0.35,0.5\}$, and functions $\varphi(t)$ were taken in $\{t, t^{1/2}, t^{1/3}, t^{1/4}, \log(t+\eta)\}$ with η in $\{0.001,0.01,0.1,1\}$. Due to the sheer number of possible combinations of optimal transport parameters (992), combined with high computational costs, only a small subset of 45 of these combinations were tested, and computations were interrupted for non-promising metrics.
In addition, the GeomLoss library provides tools for computing kernel Maximum Mean Discrepancies that amount to computing Euclidean distance between blurred signals. Specifically, we tested the energy distance [FSV*19] with the same set of parameter ranges for φ, α, γ and η (this loss does not rely on ε and s), and also restricted to a subset of 12 possible combinations.

6.2. Results

Performance measures. As classically done in image and video quality assessment, the performance of objective metrics is evaluated using the Spearman rank order correlation and the Pearson linear correlation coefficients computed between the objective metric's values and the subjective mean opinion scores. The Pearson correlations are computed after a logistic regression which provides a non-linear mapping between the objective and subjective scores. This mapping allows the evaluation to take into account the saturation effects associated with human senses. For each metric, instead of single Pearson/Spearman correlation values, we compute distributions of correlations using bootstrapping: The correlation is computed 100 times, each time on a random set of BRDFs having the same size as the original dataset; this random set is generated by sampling with replacement. The bootstrap distribution allows to provide an average correlation and a 95% confidence interval.

Evaluation of L_p metrics. Figure 12 show the performance of the 36 variations of the L_p-based BRDF metrics. First, we observe that taking the cubic root of the BRDFs before computing the distance, greatly improves the results for most of metrics. The only ones that are not improved (D8 and D9) are those that already consider such attenuation function in their formulas. Secondly, applying the clamping also consistently improves the results (but in a more moderate way as compared to the effect of the cubic root). The quantitative assessment of those two effects thanks to our data, provides strong hints for the scientific community for good practices concerning the use of those L_p distances for analytical model fitting.
Overall, the metric that provides the best correlation with perceptual measures is $D9.Cl$ (Pearson=0.81). However, we note that with cubic root and clamping, even the simple L_2 distance ($D1.Cl.CR$) provides fairly good results (Pearson=0.79).

Evaluation of optimal transport and kernel metrics. The set of parameters that performs best on average for the optimal transport metrics corresponds to $c(X,Y) = \|X-Y\|^2$, $\alpha = 0.1$, $s = 0.8$, $\varepsilon = 0.05$, $\gamma = 0.25$, $\varphi(t) = t^{1/4}$. The computational time for this metric is 3 seconds per pair of BRDF on an NVIDIA RTX 2080. The attained correlation are 0.85 for both Pearson and Spearman. This showed to be robust with respect to the entropic regularization parameter ε – setting $\varepsilon = 0.1$ merely increases the Pearson correlation by 0.15% while reducing the Spearman correlation by 0.39% and setting $\varepsilon = 0.02$ decreases both Pearson and Spearman correlations by 0.3% and 0.06% respectively. For large majority of tested parameters, correlations where above 0.8.
The best Kernel distance was obtained for $\varphi = t^{1/3}$, $\alpha = 0.1$ and

[†] We also experimented with more principled higher dimensional spherical parameterizations but found no significant differences in term of results.

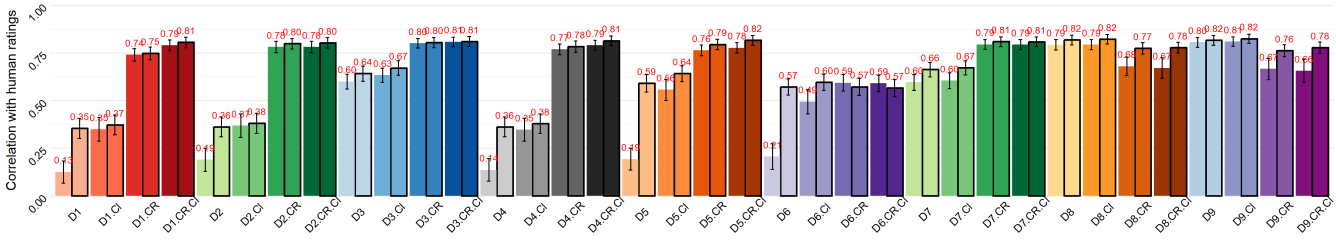


Figure 12: Performance, in terms of Pearson and Spearman correlation with mean opinion scores (MOS), of tested L_p-based BRDF metrics. For each metric *Di*, we present its four variations. *Cl* stands for Clamping and *CR* stands for Cubic Root. Bars without border refer to Pearson and bars with black borders refer to Spearman. The error bars indicate the 95% intervals computed by boostrapping.

$\gamma = 0.25$, takes 9 seconds to compute, and results in Pearson and Spearman correlation coefficients of 0.85 and 0.84 respectively.

Global performance results. The performance of the tested metrics are shown in Figure 13. Plots illustrating subjective MOS vs metric values are shown in Figure 14. For L_p distances, we included only $D1.Cl.CR$ and $D9.Cl$ for the sake of clarity. Note that image-based metrics have an advantage in this comparison, since they are computed on the exact same rendered images as those used in the experiment. Hence, as compared with BRDF metrics they have the knowledge of the illumination, rendering parameters and geometry. Results for image metrics show the high importance of the chroma information, since luminance-only metrics (HDRVDP2 and SSIM) provide the lowest correlations. Our results allow to quantitatively assess the good behavior of the CID metric, used in [HFM16, BP20] for material quality estimation. Metrics based on deep learning (LPIPS and MatSim) do not achieve higher performance. However, they have not been specifically tuned for our task.
When it comes to BRDF metrics, best results are provided by the optimal transport metric, which even attain the performance of CID, followed by the Kernel distance. Nevertheless those metrics are much more costly to compute than the L_p distances.

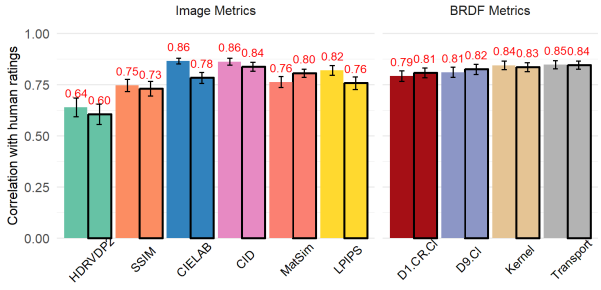


Figure 13: Performance, in terms of Pearson and Spearman correlation with mean opinion scores (MOS), of several Image-based and BRDF metrics. Bars without border refer to Pearson and bars with black borders refer to Spearman. The error bars indicate the 95% intervals computed by bootstrapping.

7. Conclusion

In this paper we presented a database of perceptual measurements that we obtained by comparing images of approximations of measured materials to reference images featuring the original BRDF. We conducted multiple careful statistical evaluations using our data, that lead to interesting new findings as well as confirmation of some previously known (yet not systematically tested) facts.
Concerning material perception in general, (1) we didn't find any global effect of the illumination on the perceived quality of approximated BRDFs. However, when restricting the analysis to the set of hidden references, (2) we found that people tend to better recognize a same material (rendered under different viewpoint) when it has a metallic nature, and/or under high frequency illumination, confirming the hypothesis of Fleming [Fle14]. Our small set of two illumination conditions still renders general interpretations difficult, and our results should thus be taken with caution.
Concerning analytical models, we found that (1) Their respective performance heavily depends on the material type (dielectric or metal). (2) Among parametric models, Blinn-Phong works surprisingly well on dielectrics while GGX outperforms other models on metals. The non parametric model of Bagher *et al.* [BSN16] outperforms all other existing models.
Concerning the metrics, we found that (1) best metrics (i.e. showing the highest correlation with subjective scores) consider a logarithmic distance (e.g. D_9 in Section 6) or weighted L_p distances computed over the cubic root of BRDF data (distances D_1, D_2, D_4); (2) all those L_p distances are improved by clamping grazing angles; (3) the transportation-L^p metric of Thorpe et al. [TPK*17], in spite of being costly to evaluate, shows the best correlation with perceptual measurements among all tested BRDF-space metrics; (4) the CID metric, which includes chromatic information, outperforms other image-space metrics, even deep-learning based ones.

The comparison of perceptual quality of the different analytical models was based on different fitting methods, some being own fitting algorithm, some obtained from previous work data releases. It would be beneficial to confirm this comparison by re-doing all fits using a common optimization method (e.g. Hooke-Jeeves with a cube-root distance).
Our dataset can be used right away as a reference to benchmark future BRDF error metrics, in terms of their performance to predict the perceived fidelity of BRDF approximations.
A challenging future work is to create a data-driven BRDF-space metric by extrapolating perceptual measurements from our database. Such a distance could be used e.g. as a cost function when fitting analytical model to measured data, or for material-preserving gamut mapping of BRDFs [SSGM17]. The relatively low number of such measurements however limits their use for deep learning, justifying the need to either acquire many more measurements, or

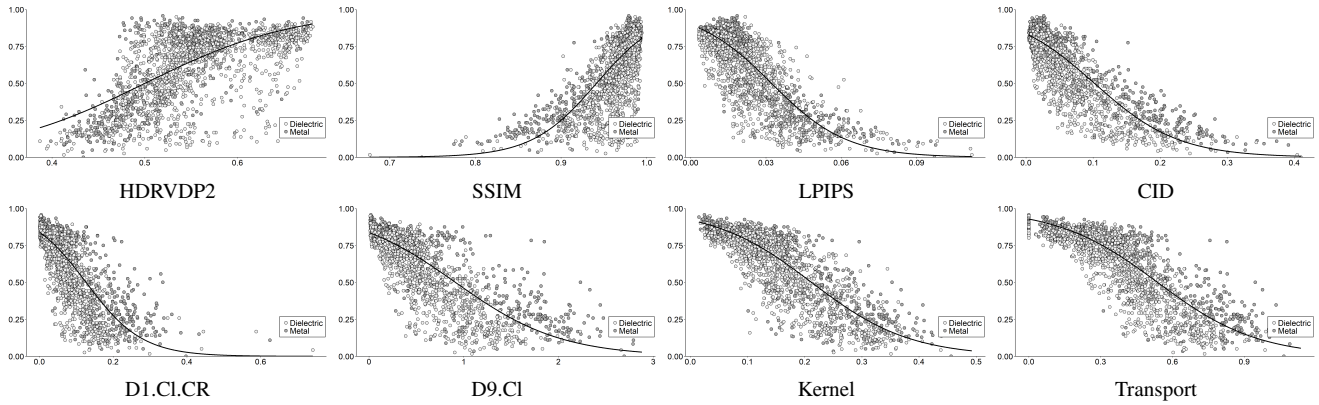


Figure 14: Subjective MOS vs metric values for several BRDF metrics (bottom) and image quality metric (top). The curve shows the logistic regression.

turn to a low-dimensional parameterized metric at the expense of accuracy.

Acknowledgements

We thank Derek Nowrouzezahrai and Mahdi Bagher for providing data from their paper, and Abir Zendagui for helping to generate the fits. This project was partly funded by ANR CALiTrOp (ANR-16-CE33-0026).

References

[Bar66] BARTKO J.: The intraclass correlation coefficient as a measure of reliability. *Psychological Reports 19* (1966), 3–11. 6

[BCP*15] BELCOUR L., COURTES L., PACANOWSKI R., ET AL.: ALTA: A BRDF Analysis Library. http://alta.gforge.inria.fr/, 2013-2015. 4

[BCRA11] BOUSSEAU A., CHAPOULIE E., RAMAMOORTHI R., AGRAWALA M.: Optimizing environment maps for material depiction. *Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering) 30*, 4 (07 2011). 3

[Bli77] BLINN J. F.: Models of light reflection for computer synthesized pictures. *SIGGRAPH Comput. Graph. 11*, 2 (July 1977), 192–198. 3

[BLPW14] BRADY A., LAWRENCE J., PEERS P., WEIMER W.: genbrdf: Discovering new analytic brdfs with genetic programming. *ACM Transactions on Graphics 33*, 4 (July 2014), 114:1–114:11. 2, 8

[BP20] BIERON J., PEERS P.: An Adaptive BRDF Fitting Metric. *Computer Graphics Forum 39*, 4 (2020), 59–74. 2, 8, 10

[BS87] BECKMANN P., SPIZZICHINO A.: *The Scattering of Electromagnetic Waves from Rough Surfaces*. Artech Print on Demand, Norwood, MA, Mar. 1987. 4

[BSH12] BAGHER M. M., SOLER C., HOLZSCHUCH N.: Accurate fitting of measured reflectances using a Shifted Gamma micro-facet distribution. *Computer Graphics Forum 31*, 4 (June 2012), 1509–1518. 1, 4, 7, 8

[BSN16] BAGHER M. M., SNYDER J., NOWROUZEZAHRAI D.: A non-parametric factor microfacet model for isotropic brdfs. *ACM Transactions on Graphics (TOG) 35*, 5 (2016), 1–16. 4, 7, 10

[BvdPPH11] BONNEEL N., VAN DE PANNE M., PARIS S., HEIDRICH W.: Displacement Interpolation Using Lagrangian Mass Transport. *ACM Transactions on Graphics (SIGGRAPH ASIA 2011) 30*, 6 (2011). 9

[Dal92] DALY S. J.: Visible differences predictor: an algorithm for the assessment of image fidelity. vol. 1666, pp. 2–15. 2

[Deb04] DEBEVEC P.: Light probe image gallery. http://www.pauldebevec.com/Probes/ (last accessed April 2017), 2004. 2

[DHI*15] DUPUY J., HEITZ E., IEHL J.-C., POULIN P., OSTROMOUKHOV V.: Extracting microfacet-based BRDF parameters from arbitrary materials with power iterations. *Computer Graphics Forum 34*, 4 (2015), 21–30. 4, 7

[FDA03] FLEMING R. W., DROR R. O., ADELSON E. H.: Real-world illumination and the perception of surface reflectance properties. *Journal of Vision 3*, 5 (2003), 3. 3, 4, 6

[Fey19] FEYDY J.: Geometric loss functions between sampled measures, images and volumes, 2019. URL: https://www.kernel-operations.io/geomloss/. 9

[FFG12] FORES A., FERWERDA J., GU J.: Toward a Perceptually Based Metric for BRDF Modeling. *Color and Imaging Conference* (2012). 1, 2, 8

[Fle14] FLEMING R. W.: Visual perception of materials and their properties. *Vision research 94* (jan 2014), 62–75. 3, 6, 10

[FSV*19] FEYDY J., SÉJOURNÉ T., VIALARD F.-X., AMARI S.-I., TROUVÉ A., PEYRÉ G.: Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics* (2019), pp. 2681–2690. 9

[GB16] GHADIYARAM D., BOVIK A. C.: Massive Online Crowdsourced Study of Subjective and Objective Picture Quality. *IEEE Transactions on Image Processing 25*, 1 (2016), 372–387. 5

[GKD07] GREEN P., KAUTZ J., DURAND F.: Efficient Reflectance and Visibility Approximations for Environment Map Rendering. *Computer Graphics Forum 26*, 3 (2007), 495–502. 1

[HFM16] HAVRAN V., FILIP J., MYSZKOWSKI K.: Perceptually Motivated BRDF Comparison using Single Image. *Computer Graphics Forum 35*, 4 (2016), 1–12. 2, 3, 8, 10

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778. 8

[ilm] Specification and implementation of the EXR file format. URL: https://openexr.com. 4

[KFB10] KŘIVÁNEK J., FERWERDA J. A., BALA K.: Effects of global illumination approximations on material appearance. *ACM Transactions on Graphics 29*, 4 (jul 2010), 1. 3, 4, 6

[Kri14] KRIZHEVSKY A.: One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997* (2014). 8

[LFTG97] LAFORTUNE E. P. F., FOO S.-C., TORRANCE K. E., GREENBERG D. P.: Non-linear approximation of reflectance functions. In *ACM SIGGRAPH* (1997), pp. 117–126. 2, 4

[LKYU12] LÖW J., KRONANDER J., YNNERMAN A., UNGER J.: Brdf models for accurate and efficient rendering of glossy surfaces. *ACM Transactions on Graphics 31*, 1 (Feb. 2012), 9:1–9:14. 2, 3, 4

[LMS*19] LAGUNAS M., MALPICA S., SERRANO A., GARCES E., GUTIERREZ D., MASIA B.: A Similarity Measure for Material Appearance. *ACM Transactions on Graphics 38*, 4 (2019). 3, 8

[LPDH10] LELOUP F., POINTER M., DUTRÉ P., HANSELAER P.: Geometry of illumination, luminance contrast, and gloss perception. *JOSA A 27*, 9 (2010), 2046–2054. 3, 6

[LPU*13] LISSNER I., PREISS J., URBAN P., LICHTENAUER M. S., ZOLLIKER P.: Image-difference prediction: From grayscale to color. *IEEE Transactions on Image Processing 22*, 2 (2013), 435–446. 2, 8

[MKRH11] MANTIUK R., KIM K. J., REMPEL A. G., HEIDRICH W.: HDR-VDP-2. *ACM Transactions on Graphics 30*, 4 (jul 2011), 1. 2, 3, 8

[MPBM03] MATUSIK W., PFISTER H., BRAND M., MCMILLAN L.: A data-driven reflectance model. In *ACM SIGGRAPH* (New York, NY, USA, 2003), ACM, pp. 759–769. 2, 3, 8

[MTM12] MANTIUK R. K., TOMASZEWSKA A., MANTIUK R.: Comparison of four subjective methods for image quality assessment. *Computer Graphics Forum 31*, 8 (2012), 2478–2491. 5

[NDM05] NGAN A., DURAND F., MATUSIK W.: Experimental analysis of brdf models. In *Proceedings of the Eurographics Symposium on Rendering* (2005), pp. 117–226. 1, 2, 3, 4, 7, 9

[NDM06] NGAN A., DURAND F., MATUSIK W.: Image-driven navigation of analytical brdf models. In *Proceedings of the 17th Eurographics Conference on Rendering Techniques* (2006), pp. 399–407. 2

[NPC*15] NARWARIA M., PERREIRA M., CALLET P. L., NARWARIA M., PERREIRA M., SILVA D., CALLET P. L.: HDR-VQM : An Objective Quality Measure for High Dynamic Range Video. *Signal Processing: Image Communication 35* (2015), 46–60. 8

[NS98] NISHIDA S., SHINYA M.: Use of image-based information in judgments of surface-reflectance properties. *J. Opt. Soc. Am. A 15*, 12 (Dec 1998), 2951–2965. 3

[OB11] OLKKONEN M., BRAINARD D. H.: Joint effects of illumination geometry and object shape in the perception of surface reflectance. *i-Perception 2*, 9 (2011), 1014–1034. 3

[PC*19] PEYRÉ G., CUTURI M., ET AL.: Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning 11*, 5-6 (2019), 355–607. 9

[PFG00] PELLACINI F., FERWERDA J. A., GREENBERG D. P.: Toward a psychophysically-based light reflection model for image synthesis. In *ACM SIGGRAPH* (2000), pp. 55–64. 3

[PFU14] PREISS J., FERNANDES F., URBAN P.: Color-image quality assessment: From prediction to optimization. *IEEE Transactions on Image Processing 23*, 3 (2014), 1366–1378. 8

[PL07] PELLACINI F., LAWRENCE J.: Appwand: Editing measured materials using appearance-driven optimization. *ACM Transactions on Graphics 26*, 3 (July 2007). 2

[PR12] PEREIRA T., RUSINKIEWICZ S.: Gamut mapping spatially varying reflectance with an improved BRDF similarity metric. *Computer Graphics Forum (Proc. Eurographics Symposium on Rendering) 31*, 4 (June 2012). 2

[PSCS*12] PACANOWSKI R., SALAZAR CELIS O., SCHLICK C., GRANIER X., POULIN P., CUYT A.: Rational BRDF. *IEEE Transactions on Visualization and Computer Graphics 18* (Nov. 2012), 1824–1835. 4

[RFWB07] RAMANARAYANAN G., FERWERDA J., WALTER B., BALA K.: Visual equivalence: towards a new standard for image fidelity. In *ACM SIGGRAPH* (2007), ACM, pp. 76–es. 3, 4

[Rym18] RYMAN D.: A metric for perceptual distance between bidirectional reflectance distribution functions, 2018. 8

[SBN15] SOLER C., BAGHER M. M., NOWROUZEZAHRAI D.: Efficient and accurate spherical kernel integrals using isotropic decomposition. *ACM Transactions on Graphics 34*, 5 (Oct. 2015), 161:1–161:14. 1

[SF01] SILVERSTEIN D. A., FARRELL J. E.: Efficient method for paired comparison. *Journal of Electronic Imaging 10*, 2 (2001), 394. 5

[SGM*16] SERRANO A., GUTIERREZ D., MYSZKOWSKI K., SEIDEL H.-P., MASIA B.: An intuitive control space for material appearance. *ACM Transactions on Graphics 35*, 6 (2016). 2, 5

[SJR18] SUN T., JENSEN H. W., RAMAMOORTHI R.: Connecting measured brdfs to analytic brdfs by data-driven diffuse-specular separation. *ACM Transactions on Graphics (TOG) 37*, 6 (2018), 1–15. 8

[SSGM17] SUN T., SERRANO A., GUTIERREZ D., MASIA B.: Attribute-preserving gamut mapping of measured BRDFs. *Computer Graphics Forum 36*, 4 (2017), 47–54. 10

[SSN18] SOLER C., SUBR K., NOWROUZEZAHRAI D.: A Versatile Parameterization for Measured Material Man ifolds. *Computer Graphics Forum 37*, 2 (Apr. 2018), 135–144. 4, 7, 8

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014). 8

[TF18] TODD J. T., FARLEY NORMAN J.: The visual perception of metal. *Journal of Vision 18*, 3 (2018), 1–17. doi:10.1167/18.3.9. 3

[TPK*17] THORPE M., PARK S., KOLOURI S., ROHDE G. K., SLEPČEV D.: A transportation lp distance for signal analysis. *Journal of mathematical imaging and vision 59*, 2 (2017), 187–210. 9, 10

[VBF17] VANGORP P., BARLA P., FLEMING R. W.: The perception of hazy gloss. *Journal of Vision 17*, 5 (2017). doi:10.1167/17.5.19. 3

[VLD07] VANGORP P., LAURIJSSEN J., DUTRÉ P.: The influence of shape on the perception of material reflectance. *ACM Transactions on Graphics 26*, 3 (2007). 3, 4

[WAKB09] WILLS J., AGARWAL S., KRIEGMAN D., BELONGIE S.: Toward a perceptual space for gloss. *ACM Transactions on Graphics 28*, 4 (aug 2009), 1–15. 3

[War92] WARD G. J.: Measuring and modeling anisotropic reflection. In *ACM SIGGRAPH* (New York, NY, USA, July 1992), pp. 265–272. 3

[WBSS04] WANG Z., BOVIK A., SHEIKH H., SIMONCELLI E.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4 (2004), 600–612. 2, 8

[WMLT07] WALTER B., MARSCHNER S. R., LI H., TORRANCE K. E.: Microfacet models for refraction through rough surfaces. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques* (2007), pp. 195–206. 1, 4

[XSD*13] XU K., SUN W.-L., DONG Z., ZHAO D.-Y., WU R.-D., HU S.-M.: Anisotropic spherical gaussians. *ACM Transactions on Graphics 32*, 6 (Nov. 2013), 209:1–209:11. 1

[ZdRBP19] ZHANG F., DE RIDDER H., BARLA P., PONT S.: A systematic approach to testing and predicting light-material interactions. *Journal of Vision 19*, 4 (2019), 1–22. 3

[Zha12] ZHANG L.: A comprehensive evaluation of full reference image quality assessment algorithms. *IEEE International Conference on Image Processing* (2012), 1477–1480. 8

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 586–595. 8