

A Neural Quality Metric for BRDF Models

Behnaz Kavoosighafi¹, Rafał K. Mantiuk², Saghi Hajisharif¹, Ehsan Miandji¹, and Jonas Unger¹

¹Department of Science and Technology, Linköping University, Norrköping, Sweden

²Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom

E-mail: behnaz.kavoosighafi@liu.se

Abstract. Accurately evaluating the quality of bidirectional reflectance distribution function (BRDF) models is essential for photo-realistic rendering. Traditional BRDF-space metrics often employ numerical error measures that fail to capture perceptual differences evident in rendered images. In this paper, we introduce the first perceptually informed neural quality metric for BRDF evaluation that operates directly in BRDF space, eliminating the need for rendering during quality assessment. Our metric is implemented as a compact multi-layer perceptron (MLP), trained on a dataset of measured BRDFs supplemented with synthetically generated data and labelled using a perceptually validated image-space metric. The network takes as input paired samples of reference and approximated BRDFs and predicts their perceptual quality in terms of just-objectable-difference (JOD) scores. We show that our neural metric achieves significantly higher correlation with human judgments than existing BRDF-space metrics. While its performance as a loss function for BRDF fitting remains limited, the proposed metric offers a perceptually grounded alternative for evaluating BRDF models.

1 Introduction

Material appearance is a key component of photo-realistic rendering and is influenced by several factors, including surface reflectance, illumination, and geometry. The bidirectional reflectance distribution function (BRDF) describes surface reflectance by defining how incident light is reflected in different outgoing directions at a given wavelength [21]. Since accurately modelling BRDFs with analytical functions often requires substantial domain expertise and careful parameter tuning, measured BRDF datasets [19, 5] have become an essential resource for capturing complex reflectance behaviour without relying on simplified assumptions or isolated perceptual attributes.

Rendering with measured BRDFs often requires approximating them using analytical or learned models. This fitting process typically relies on basic cost functions, such as ℓ_1 or ℓ_2 distances computed in either linear or logarithmic reflectance space. A fundamental property of these metrics is that they operate in the BRDF representation space, treating the reflectance function independently of any scene context, such as geometry or lighting. While this simplifies optimisation, it can lead to fitted models that are numerically accurate but visually inconsistent with the original material, resulting in perceptual artifacts in the final renderings. The mismatch between BRDF-space and image-space quality metrics has been previously observed, but partially addressed, in prior work [2, 10]. For instance, some works [15, 2, 8] have explored evaluating the quality of BRDF fits using image-space metrics, such as PSNR, ΔE_{ITP} , and SSIM, computed over rendered outputs. These metrics inherently capture the influence of scene-dependent factors and have been shown in prior studies [2, 10] to align more closely with human perception. However,

applying these metrics during optimisation is computationally expensive, as it requires rendering the BRDF at each step. This makes them impractical for iterative fitting in most workflows.

To address this limitation, we propose a novel perceptually-based metric for evaluating the quality of BRDF model fits. While operating in the BRDF space, the metric takes scene information into account by learning a mapping from the input BRDFs to the output perceived quality in image space. We use the result of a perceptual study on a stereoscopic high dynamic range (HDR) display to train our neural model. Our evaluation shows that our neural metric correlates with the perceptual quality of rendered images, as assessed by human observers. We conduct evaluations to demonstrate the effectiveness of our proposed metric, comparing it to traditional BRDF-space metrics in terms of correlation with subjective data.

Our key contributions are:

- First neural quality metric for BRDF evaluation that operates directly in the BRDF space and does not require rendered images. The full source code for our metric is available at <https://github.com/behnazkavoosi/neural-brdf-quality-metric>.
- Significantly higher correlation with human judgments than existing BRDF-space metrics, providing a more perceptually reliable evaluation tool.
- Compact and efficient MLP architecture trained on a balanced, augmented BRDF dataset with perceptually aligned pseudo-labels, enabling real-time quality prediction.

2 Background

Since the introduction of measured BRDF datasets aimed at enhancing material realism in rendering applications, the development of perceptually meaningful BRDF metrics has become a central research focus. These metrics play an important role in BRDF fitting and optimisation, as well as in evaluating the performance of BRDF modelling and sampling techniques. Broadly, such metrics are computed either in BRDF space, emphasising representation error, or in image space, where they assess rendering error to evaluate visual accuracy.

BRDF Space. A wide range of cost functions has been proposed for BRDF optimisation, with ℓ_1 and ℓ_2 error measures forming the foundation of many such metrics. These metrics are often weighted by the cosine of the incoming and outgoing directions to account for angular reflectance variations. To improve alignment with human perception, researchers have explored perceptually motivated formulations, such as cube-root and logarithmic transformations. Cube-root cosine-weighted metrics have been shown to better capture perceptual differences, particularly around specular highlights [7, 13]. Logarithmic metrics, on the other hand, are designed to account for perceptual nonlinearity and to reduce sensitivity to grazing angles [16, 4]. While these BRDF-space approaches are simple and provide robust, scene-independent numerical approximations, they often fail to capture perceptual and visual differences in rendered appearance [2]. This limitation was further highlighted in a recent user study evaluating the perceived quality of BRDF models, where BRDF-space metrics demonstrated weak correlation with human judgments [10].

Image Space. To assess BRDF similarity from a perceptual standpoint, many studies have employed image-based metrics evaluated under controlled rendering conditions. Early approaches relied on pixel-wise losses such as the ℓ_2 norm under fixed lighting environments [20], later improved through structural similarity measures applied to tone-mapped images [24, 3]. Subsequent work extended this by exploring a range of perceptual metrics, including colour difference formulas, contrast-based indices, and high dynamic range visibility models, to evaluate differences in appearance, particularly for complex reflectance behaviours like anisotropy [8, 15, 17]. More recent methods leverage similarity metrics and combine them with statistical loss functions to select perceptually plausible BRDF fits from numerically optimised candidates [2]. Large-scale user studies have also been conducted to better understand material appearance across varying lighting and geometric setups, leading to the development of perceptual similarity metrics grounded in human judgments [12, 13]. These efforts have been further expanded to include dynamic content, where user ratings of video sequences indicated that perceptual similarity can be effectively approximated using combinations of correlation-based and norm-based measures [6], as well as through colour difference computations in the ICtCp colour space [10].

3 Neural Quality Metric

Our goal is to develop a neural quality metric that predicts rendering quality given two input BRDFs: one representing the reference and the other the approximation. To train this model in a supervised manner, a comprehensive dataset is required that includes both BRDFs in their representation space along with their corresponding renderings, enabling the computation of image-space error. Once trained, the model can be used to predict image-space error directly from tabulated BRDFs in the test set, without requiring additional renderings.

3.1 Dataset and Pre-Processing

We build on a dataset introduced in prior work [10], which contains rendered video sequences for 159 BRDF material samples sourced from the MERL [19], RGL-EPFL [5], and DTU [22] datasets. The material samples were used to render short video clips showing a rotating Stanford bunny illuminated with the Pixar Campus environment map. Each BRDF was approximated using nine different BRDF models at varying levels of encoding, resulting in a total of 1431 distorted videos. The dataset also includes subjective quality ratings in the form of just-objectable-difference (JOD) scores. However, the perceptual study was conducted on a subset of only 20 BRDF samples, providing ground truth for 180 distorted instances. To enable supervised training across the full dataset, we generated pseudo-labels for the remaining data using an alternative image-space metric.

In the previous study [10], the ΔE_{ITP} metric demonstrated the strongest correlation with human judgments among the evaluated metrics (see Figure 1a). We therefore used this image-space metric to generate labels for training. For the subset of materials used in the perceptual experiment, where ground-truth JOD values are available, we fitted the metric predictions to the corresponding JOD scores using the Levenberg–Marquardt algorithm [14, 18] (`lsqcurvefit` function in MATLAB) to estimate the optimal regression parameters. Once the parameters were obtained, we used them to compute JOD values for all materials in the dataset based on their ΔE_{ITP} values:

$$\text{JOD} = 10 \cdot \left(1 - \frac{1}{1 + \exp(b_1 \cdot (- (\max(\Delta E_{\text{ITP}}, 0)^{b_3}) - b_2))} \right), \quad (1)$$

where $b_1 = -14.11$, $b_2 = -0.47$, and $b_3 = -0.21$ are the fitted parameters.

To prepare the inputs for training the neural network, we reduced the angular complexity of the BRDFs by sampling a subset of values from valid angular regions. We extracted 500 distinct BRDF reflectance samples per material from the available measured datasets. This subsampling facilitates the use of simpler and more efficient network architectures while preserving key reflectance characteristics. All BRDFs were first transformed into the Rusinkiewicz coordinate system [23]. Samples were then selected based on their reflectance magnitude, prioritizing directions with significant energy. For the difference vector, we applied uniform sampling across both the elevation (θ_d) and azimuthal (ϕ_d) angles to ensure diverse angular coverage. For the half-vector elevation angle (θ_h), we adopted a non-uniform sampling strategy to increase sample density near $\theta_h = 0$, thus capturing the structure of specular highlights more effectively (see Figure 1b). To avoid instability due to measurement noise, we discarded grazing angles above 75° prior to sampling [20]. This pre-processing step resulted in a dataset of 159 reference BRDFs and 1431 approximated BRDFs, each represented by 500 samples in the Rusinkiewicz domain. Every BRDF instance was paired with a scalar JOD value, providing the target output for supervised training.

Although our sampling strategy ensures a diverse coverage of BRDF reflectance samples across the angular domain, we observed a notable imbalance in the distribution of quality scores: samples with high JOD values (corresponding to subtle or imperceptible differences) were dominant, while low JOD values (indicating strong perceptual degradation) were significantly under-represented. This imbalance poses a challenge for supervised learning, biasing the network toward specific error levels. To address this, we augmented the dataset by generating additional BRDF pairs that emphasise the low-JOD region. The number of generated samples was informed by the histogram of existing JOD values, such that under-represented regions receive a greater proportion of synthetic examples. We did this by introducing controlled random noise into the BRDF values, ensuring that the relative error between the reference and distorted BRDF remains consistent. We perturbed each BRDF’s reflectance values using additive Gaussian noise drawn from a normal distribution. For each colour channel, the noisy BRDF was computed as $\rho_n = \rho + \mathcal{N}(0, \sigma^2)$, where ρ and ρ_n denote the original and noisy BRDF values, respectively, and σ^2 is the variance of the noise. In our implementation, we set $\sigma = 0.01$ to ensure small but meaningful deviations that simulate low-quality distortions. The noise was added independently to each sampled direction without enforcing angular correlation. While angularly correlated alternatives could be explored in future work, this approach was effective for balancing the dataset and enhancing performance. This augmentation step generated a total of 3340 unique BRDF pairs with a more uniform JOD distribution.

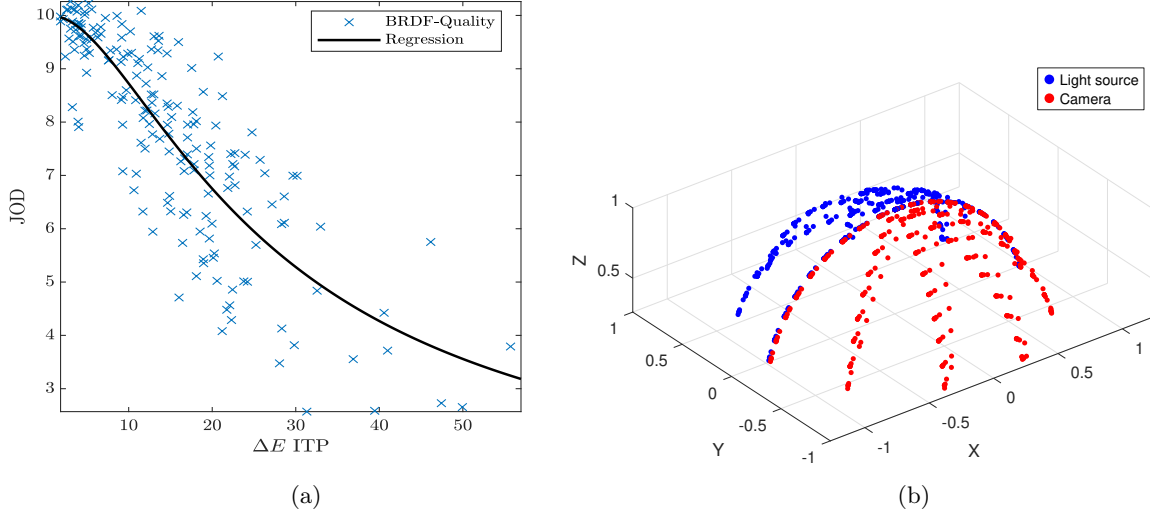


Figure 1: (a) Relationship between JOD values and ΔE_{ITP} in the original dataset, illustrating the correlation between subjective quality assessments and perceptual colour differences. (b) Sampled BRDF locations in the angular domain.

The dataset was randomly split into training (80%) and validation (20%) sets, while the BRDFs used in the perceptual experiment were reserved for testing. This resulted in 2 672 BRDF pairs for training, 668 for validation, and 180 for testing. To further enrich the training set, we applied a second augmentation step in which each pair of test and reference BRDFs was scaled by a random factor sampled uniformly from the range $[0.95, 1.05]$, producing a final training set of 5 344 BRDF pairs.

We standardised the reflectance values through a series of transformations designed to enhance learning and align the input representation with perceptual characteristics. First, a cube root transformation was applied to suppress sharp specular peaks, which are often overemphasised in raw reflectance values [13]. Next, a logarithmic transformation was used to compress the dynamic range and account for the non-linear response of human vision [16], defined as:

$$\rho_t = \log \left(\rho^{1/3} + 1 \right), \quad (2)$$

where ρ denotes the raw reflectance value and ρ_t is the transformed value. Finally, we applied whitening by subtracting the mean and dividing by the standard deviation of each channel, computed over the reference BRDFs in the training set. All transformations were applied identically to both reference and distorted BRDFs to preserve relative differences.

3.2 Network Architecture

Given the ability of multi-layer perceptrons (MLPs) to approximate complex non-linear functions, we design an MLP-based regressor to predict perceptual quality scores (JOD values) from paired reference and approximated BRDF samples (see Figure 2). The MLP architecture comprises four fully connected layers. Each hidden layer is followed by Layer Normalisation [1] to stabilise training dynamics, a GELU (Gaussian Error Linear Unit) activation [9] to introduce smooth non-linearity, and a dropout layer with a rate of 0.2 to mitigate overfitting. To ensure that the predicted JOD values lie within the expected perceptual scale, a sigmoid activation is applied after the final layer, re-scaling the output to the range defined by the minimum and maximum JOD values observed in the dataset.

The input to the network is a concatenation of the reference and distorted BRDF samples, each of dimension 500×3 , resulting in an overall input shape of 500×6 . The first linear layer expands the representation to 1 024 features. The subsequent layers reduce the dimensionality progressively using a decay factor of approximately 0.7, resulting in 716 and 501 features in the second and third layers, respectively. The final output layer produces a single scalar value corresponding to the predicted JOD score. To enhance robustness to outliers while maintaining differentiability and stable gradients, we used

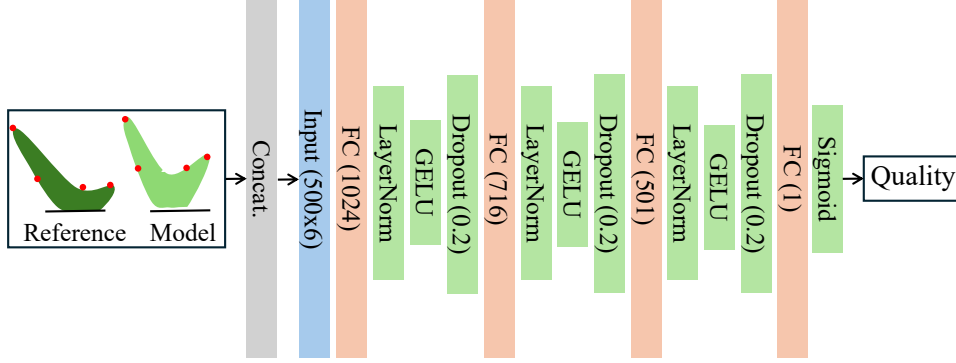


Figure 2: Network architecture: The input layer is the concatenation of reference and model RGB BRDFs with 500 samples, four samples are illustrated with red dots, leading to a vector of size 500×6 . The output is a scalar corresponding to JOD value. Each fully connected layer (FC), except the last, is followed by layer normalisation, Gaussian Error Linear Unit (GELU) activation, and a dropout layer with a probability of 0.2.

the log-cosh loss for training, which is defined as:

$$l = \log(\cosh(x - y)), \quad (3)$$

where x is the predicted JOD value and y is the corresponding ground truth. Training was conducted for 100 epochs using the Adam optimiser [11], with a base learning rate of 10^{-4} for the input layer and 10^{-3} for the deeper layers. We applied a weight decay of 10^{-4} to enhance generalisation and used a batch size of 512. To adaptively reduce the learning rate when validation loss stagnates, we employed PyTorch’s `ReduceLROnPlateau` scheduler with a patience of 5 epochs. On an RTX 4080 GPU, training took approximately 36 seconds, while inference required only 0.04 milliseconds per prediction. The final model comprises 4171125 trainable parameters and requires 15.91 MB of storage when saved in 32-bit floating-point format.

4 Results

We evaluate the performance of our neural quality metric by comparing its predictions with the ground-truth JOD values obtained from the user study conducted in previous work [10]. As discussed, our model was trained using pseudo-labels derived from ΔE_{JTP} scores, which introduces some discrepancy between the training targets and the actual subjective data. Nevertheless, we compare our model against a set of BRDF-space metrics evaluated in the same prior study, selecting the best-performing variant of each metric to ensure a fair comparison.

The baseline metrics include root mean squared error (RMSE) and mean absolute error (MAE), which operate directly on linear BRDF values. In addition to these, we evaluate perceptually motivated variants that apply nonlinear transformations and weighting strategies. These include the root mean squared and mean absolute cubic root-weighted errors (RMS-CRWE and MA-CRWE), which apply a cube root transformation and cosine weighting to emphasise perceptually significant regions. Similarly, the root mean squared and mean absolute logarithmic errors (RMS-LogE and MA-LogE) use a logarithmic transformation to compress the dynamic range in line with perceptual nonlinearity. Finally, root mean squared and mean absolute logarithmic weighted errors (RMS-LogWE and MA-LogWE) combine logarithmic transformation with cosine weighting to better model human sensitivity to directional reflectance variation.

As shown in Figure 3, our proposed method, denoted as *BRDF-NQM*, achieves the highest correlation with subjective JOD scores, reaching an average Spearman correlation of 0.67. A higher correlation value indicates stronger agreement with human perceptual judgments. This result demonstrates a significant improvement over all tested BRDF-space metrics and highlights the effectiveness of learning a perceptually grounded metric directly from BRDF data.

Perceptual metrics are often employed as loss functions in BRDF fitting pipelines, where they are expected to guide optimisation toward perceptually meaningful results. To investigate this use case, we applied our neural metric as a loss function for fitting measured BRDFs. While certain reflectance properties, such as the index of refraction and surface roughness, were predicted with reasonable accuracy, we consistently observed a colour shift toward red in the estimated diffuse component. This artifact likely

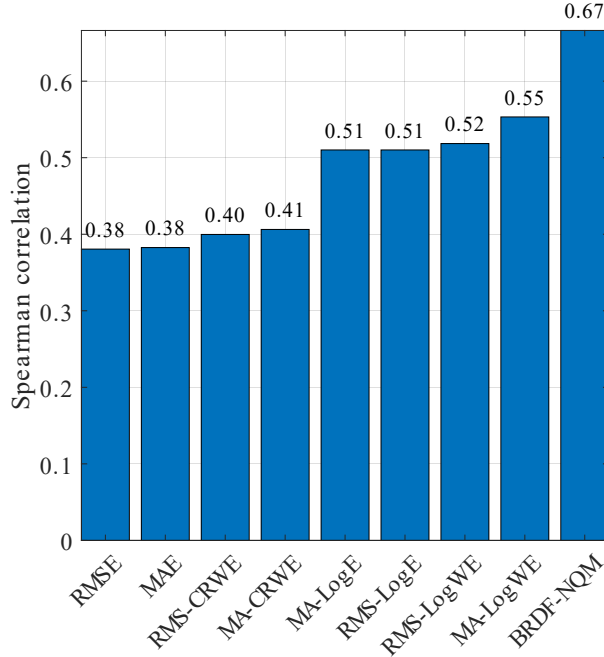


Figure 3: Average Spearman correlation between subjective JOD scores and various BRDF-space metrics. Our neural quality metric (BRDF-NQM) is included for comparison against traditional BRDF-space metrics. Higher values indicate stronger alignment with human perceptual judgments.

results from colour imbalances in the training data and reduced chromatic sensitivity introduced by the whitening step in pre-processing. These findings are consistent with prior work [10], which showed that a metric’s predictive accuracy does not necessarily translate into effective performance as a loss function. Bridging this gap remains an important direction for future research in perceptually guided BRDF modelling.

5 Conclusion

In this paper, we presented the first perceptually-informed neural quality metric for evaluating BRDFs directly in BRDF space. Designed to predict perceptual quality under complex viewing conditions, the metric employs a four-layer MLP that takes as input a pair of reference and approximated BRDFs, sampled to maintain the most perceptually significant reflectance information. We evaluated the proposed metric against the state-of-the-art BRDF-space metrics by measuring their correlation with human judgments and observed a significant improvement, establishing our method as the most effective BRDF-space metric for BRDF evaluation. While the metric shows strong predictive performance, it does not transfer as effectively when used as a stand-alone loss function in BRDF fitting. Nevertheless, our findings highlight the potential of perceptually grounded neural metrics to inform the design of new BRDF-space loss functions that prioritise perceptual fidelity, potentially reducing or eliminating the need for costly rendering in optimisation tasks.

References

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. “Layer Normalization”. In: (July 2016). arXiv: 1607.06450 [stat.ML] (cit. on p. 4).
- [2] J. Bieron and P. Peers. “An Adaptive BRDF Fitting Metric”. In: *Computer Graphics Forum* 39.4 (2020), pp. 59–74. DOI: 10.1111/cgf.14054 (cit. on pp. 1, 2).
- [3] Adam Brady et al. “genBRDF: discovering new analytic BRDFs with genetic programming”. In: *ACM Trans. Graph.* 33.4 (July 2014). ISSN: 0730-0301. DOI: 10.1145/2601097.2601193 (cit. on p. 2).

- [4] O. Clausen, R. Marroquim, and A. Fuhrmann. “Acquisition and Validation of Spectral Ground Truth Data for Predictive Rendering of Rough Surfaces”. In: *Computer Graphics Forum* 37.4 (2018), pp. 1–12. DOI: 10.1111/cgf.13470 (cit. on p. 2).
- [5] Jonathan Dupuy and Wenzel Jakob. “An Adaptive Parameterization for Efficient Material Acquisition and Rendering”. In: *Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 37.6 (Nov. 2018), 274:1–274:18. DOI: 10.1145/3272127.3275059 (cit. on pp. 1, 3).
- [6] Jiri Filip et al. *Material Fingerprinting: Identifying and Predicting Perceptual Attributes of Material Appearance*. 2024. arXiv: 2410.13615 [cs.CV] (cit. on p. 2).
- [7] Adrià Forés, James Ferwerda, and Jinwei Gu. “Toward a Perceptually Based Metric for BRDF Modeling”. In: *Final Program and Proceedings - IS and T/SID Color Imaging Conference CIC’12* (Jan. 2012), pp. 142–148. DOI: 10.2352/CIC.2012.20.1.art00025 (cit. on p. 2).
- [8] V. Havran, J. Filip, and K. Myszkowski. “Perceptually Motivated BRDF Comparison using Single Image”. In: *Computer Graphics Forum* 35.4 (2016), pp. 1–12. DOI: 10.1111/cgf.12944 (cit. on pp. 1, 2).
- [9] Dan Hendrycks and Kevin Gimpel. “Gaussian Error Linear Units (GELUs)”. In: *arXiv preprint arXiv:1606.08415* (2016) (cit. on p. 4).
- [10] Behnaz Kavoosighafi et al. “Perceived Quality of BRDF Models”. In: *Computer Graphics Forum* (2025). ISSN: 1467-8659. DOI: 10.1111/cgf.70162 (cit. on pp. 1–3, 5, 6).
- [11] Diederik Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *International Conference on Learning Representations* (Dec. 2014) (cit. on p. 5).
- [12] Manuel Lagunas et al. “A similarity measure for material appearance”. In: *ACM Trans. Graph.* 38.4 (July 2019). ISSN: 0730-0301. DOI: 10.1145/3306346.3323036 (cit. on p. 2).
- [13] Guillaume Lavoué et al. “Perceptual quality of BRDF approximations: dataset and metrics”. In: *Computer Graphics Forum* 40.2 (2021), pp. 327–338. DOI: 10.1111/cgf.142636 (cit. on pp. 2, 4).
- [14] Kenneth Levenberg. “A method for the solution of certain nonlinear problems in least squares”. In: *Quarterly of Applied Mathematics* 2.2 (1944), pp. 164–168 (cit. on p. 3).
- [15] Ingmar Lissner et al. “Image-Difference Prediction: From Grayscale to Color”. In: *IEEE Transactions on Image Processing* 22.2 (2013), pp. 435–446. DOI: 10.1109/TIP.2012.2216279 (cit. on pp. 1, 2).
- [16] Joakim Löw et al. “BRDF models for accurate and efficient rendering of glossy surfaces”. In: *ACM Trans. Graph.* 31.1 (Feb. 2012). ISSN: 0730-0301. DOI: 10.1145/2077341.2077350 (cit. on pp. 2, 4).
- [17] Rafał Mantiuk et al. “HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions”. In: 30.4 (2011), pp. 1–14 (cit. on p. 2).
- [18] Donald W. Marquardt. “An algorithm for least-squares estimation of nonlinear parameters”. In: *Journal of the Society for Industrial and Applied Mathematics* 11.2 (1963), pp. 431–441 (cit. on p. 3).
- [19] Wojciech Matusik et al. “A Data-Driven Reflectance Model”. In: *ACM Trans. Graph.* 22.3 (July 2003), pp. 759–769. ISSN: 0730-0301. DOI: 10.1145/882262.882343 (cit. on pp. 1, 3).
- [20] Addy Ngan, Frédo Durand, and Wojciech Matusik. “Experimental Analysis of BRDF Models”. In: *Eurographics Symposium on Rendering (2005)*. Ed. by Kavita Bala and Philip Dutre. The Eurographics Association, 2005. ISBN: 3-905673-23-1. DOI: 10.2312/EGWR/EGSR05/117-126 (cit. on pp. 2, 3).
- [21] Fred E. Nicodemus et al. “Geometrical Considerations and Nomenclature for Reflectance”. In: 1977 (cit. on p. 1).
- [22] Jannik Boll Nielsen, Henrik Wann Jensen, and Ravi Ramamoorthi. “On Optimal, Minimal BRDF Sampling for Reflectance Acquisition”. In: *ACM Transactions on Graphics (TOG)* 34.6 (Nov. 2015), 186:1–186:11. DOI: 10.1145/2816795.2818085 (cit. on p. 3).

- [23] Szymon M. Rusinkiewicz. “A New Change of Variables for Efficient BRDF Representation”. In: *Rendering Techniques '98*. Ed. by George Drettakis and Nelson Max. Vienna: Springer Vienna, 1998, pp. 11–22. ISBN: 978-3-7091-6453-2 (cit. on p. 3).
- [24] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861 (cit. on p. 2).