

Towards a Perceptual Evaluation Framework for Lighting Estimation

Justine Giroux¹ Mohammad Reza Karimi Dastjerdi¹ Yannick Hold-Geoffroy²
 Javier Vazquez-Corral^{3,4} Jean-François Lalonde¹

¹Université Laval, ²Adobe Research, ³Computer Vision Center, ⁴Universitat Autònoma de Barcelona



Figure 1. We pit image comparison metrics, used to quantify the performance of lighting estimation algorithms, against human perception. When asked which render looks most plausible, our controlled psychophysical study reveals that humans preference contradicts image metrics in the vast majority of cases. This paper questions the current practice of employing image quality metrics for evaluating lighting estimation algorithms when used for the task of virtual object insertion: can we do better by considering human perception?

Abstract

Progress in lighting estimation is tracked by computing existing image quality assessment (IQA) metrics on images from standard datasets. While this may appear to be a reasonable approach, we demonstrate that doing so does not correlate to human preference when the estimated lighting is used to relight a virtual scene into a real photograph. To study this, we design a controlled psychophysical experiment where human observers must choose their preference amongst rendered scenes lit using a set of lighting estimation algorithms selected from the recent literature, and use it to analyse how these algorithms perform according to human perception. Then, we demonstrate that none of the most popular IQA metrics from the literature, taken individually, correctly represent human perception. Finally, we show that by learning a combination of existing IQA metrics, we can more accurately represent human preference. This provides a new perceptual framework to help evaluate future lighting estimation algorithms. To encourage future research, all (anonymised) perceptual data and code are available at <https://lvsn.github.io/PerceptionMetric/>.

1. Introduction

Light fashions the appearance of everything we see. From the warm and glowing golden hour to the coldness of neon tubes, lighting determines how scenes look and how we perceive them [35]. This wide variety in illumination conditions creates a challenge for computer vision algorithms attempting to estimate lighting from images. How can it be robustly estimated from images when conditions are so different? Luckily, large lighting datasets [3, 13] and learning-based approaches [7, 13, 18] have enabled significant progress over the past decade. Nowadays, automatic lighting estimation methods allow artists to plausibly relight virtual objects and composite them into real images with ever-increasing ease.

As with many computer vision tasks, progress in lighting estimation is tracked by computing existing image quality assessment (IQA) metrics on standard datasets. Since lighting representations differ (e.g., environment maps [13], parametric lights [14, 57], spherical harmonics [15], spherical Gaussians [26], etc.), it is common practice to render a virtual object with the estimated lighting, and compare it with a render of the same virtual object lit with the ground truth. A variety of existing IQA metrics—ranging from classical approaches (e.g., RMSE and SSIM [52]) to those designed to mimic certain aspects of human perception (e.g., LPIPS [58])—have been used for this purpose.

Whilst this way of quantifying performance appears reasonable, it is unclear whether doing so follows human preference. After all, once a virtual object is relit and inserted into an image, perfectly matching ground truth lighting may not be needed to achieve plausible results. In a practical scenario, one does not have access to the ground truth and must judge the realism of a relighting result solely based on its appearance. For example, consider the images shown in fig. 1. When asked to choose which of the two images looks most realistic, humans prefer the left one. However, IQA metrics establish that the right one most closely matches the ground truth (not shown). In this work, we question the validity of using IQA metrics for evaluating lighting estimation algorithms: beyond the selected examples of fig. 1, do IQA metrics generalise and agree with human perception? Or is another approach necessary?

In contrast to common belief, we show that this strategy of employing existing IQA metrics for determining whether an approach outperforms previous work *does not reflect human preference*. To do so, we perform a controlled psychophysical study where a set of lighting estimation approaches, selected from the recent literature, are pitted against each other by asking observers to judge the accuracy and plausibility of virtual object insertion results. Our perceptual data reveals that no single IQA metric agrees with human perception across all situations. Finally, we show that a learned combination of the same IQA metrics, trained on our perceptual data, achieves results that match human preference.

We summarise our contributions as follows. First, we perform a psychophysical study comparing recent lighting estimation methods by asking observers to judge scenes rendered with lighting estimates. Second, we evaluate the agreement between commonly-used IQA metrics and our perceptual data. In particular, we find that most IQA metrics do not corroborate human perception. Third, instead of considering IQA metrics individually, we show that by *combining* them in a learned fashion, we can more accurately mimic human perception for the evaluation of illumination conditions. This provides a new framework for evaluating future illumination estimation algorithms in a way that more closely matches human perception.

2. Related work

Lighting estimation. Outdoor illumination estimation traditionally relied on manual identification of features such as sky, ground shadows, and surface shading [24], or optimising lighting and reflectance interchangeably [23]. Deep learning replaced these manual features with implicitly learned representations, where two predominant categories emerge. The first category utilises parametric representations for illumination estimation, such as spherical harmonics [15] and spherical Gaussians [14, 26, 27, 53] for indoor, and sky models for outdoor scenarios [18, 57]. While these methods

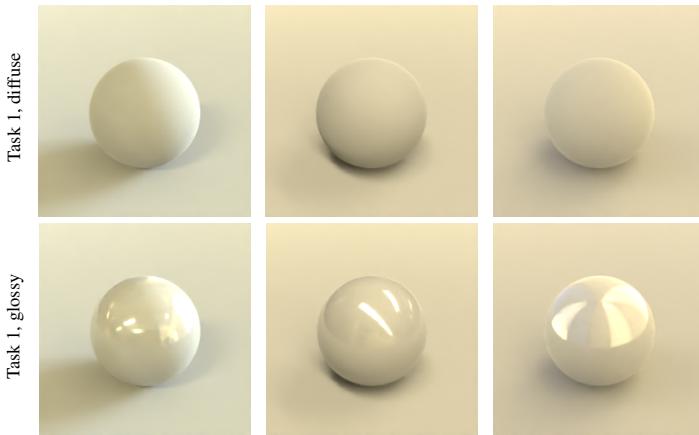
are easily editable, they often fail to produce a plausible environment, limiting its realism. The second category encompasses environment map prediction models [13, 25, 50]. These models predict the complete scene field of view as a 360° environment map. While yielding high-quality and realistic outputs, these works are complex to train and lack editability. Recent research has sought to integrate these categories [8, 54], employing a two-stage approach which first predicts a parametric lighting representation and then uses it to condition environment map generation.

Image quality assessment (IQA) metrics. IQA has been an ongoing research topic for more than 20 years [52]. It particularly flourished with the appearance of mobile phone cameras and their ubiquitous presence in our current daily lives. IQA metrics can be divided into three types: Full-Reference, Reduced-Reference, and No-Reference. Whilst Reduced-Reference metrics such as FID [17] and Inception Score [43], or No-Reference-based metrics such as BRISQUE [32], NIQE [33], UNIQUE [59], HyperIQA [45], and [12] show promising applications, the interest of our work lies in comparing images using a Reference-based metric. Domain-specific IQA metrics were also proposed, such as DISTs [9, 10] for textures; in the remainder of the document, we will focus on generic scene-level metrics. Originally, IQA metrics were mostly based on either physical [16, 19] or perceptual [29, 32, 33, 44, 52, 55] priors, but recently these priors are usually learned by using deep learning approaches from some type of observer data (e.g., mean opinion scores) [2, 40, 45, 58].

A parallel line of research to the development of IQA metrics is the study of their applicability to specific scenarios. In colour constancy [49], authors found that observers did not prefer the image with the smallest angular error against a white reference, but instead the one with a slightly bluish tint—the so-called blue bias [37]. Similarly, [56] showed that no IQA metric correlated well with the observers' choice in colour gamut mapping.

Lighting perception. Humans were shown to perceive lighting direction accurately on both synthetic and real stimuli [21, 39]. [39] demonstrated that human accuracy in lighting direction is correlated to the degree of collimation of the light source, explaining humans' high sensitivity to lighting position outdoors. However, humans have more difficulty detecting lighting inconsistencies as the scene complexity increases [28, 41, 46]. According to [4, 36, 47], observers mainly use cast shadows and highlights to determine the direction and intensity of light sources. [36] confirms that humans use geometry to estimate the lighting direction and that a globally convex shape increases their accuracy. [1, 31] reveal that the perception of gloss is modulated by the complex interaction between the object geometry and the light field, affecting multiple parameters of specular reflections (size, contrast, sharpness, and depth). We use these works as

Task 1: Three renders of {matte, glossy} spheres are presented. Select the one that resembles the most the centre one by pressing the left or right arrow on the keyboard.



Task 2: Two images of {matte, glossy} spheres are presented, select the one that seems the most realistic by pressing on the left or right arrow on the keyboard.

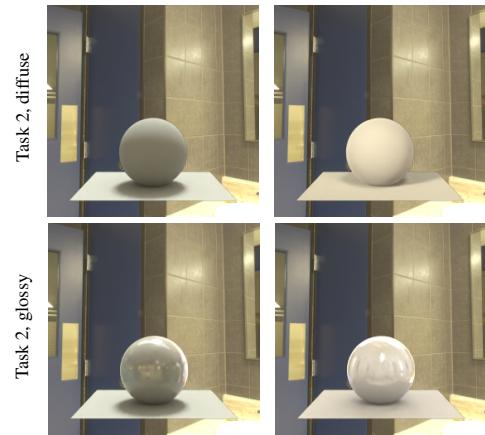


Figure 2. Example of the accuracy (task 1; left) and plausibility (task 2; right) tasks, for the diffuse (top) and glossy (bottom) material, assigned to the observers during the experiment. The question asked to the observers is written above each example.

inspiration to design our tasks and stimuli to bridge the gap between these human perception insights and IQA metrics commonly used to measure visual resemblance.

3. Psychophysical experiment

In this section, we briefly describe our psychophysical experiment, which was approved by the U. Laval institutional Ethics Review Board, file #2023-308.

3.1. Tasks

We devise two tasks to evaluate how humans perceive lighting accuracy and plausibility, as illustrated in fig. 2.

Task 1. Fig. 2 (left) presents each observer with three images of a sphere on a ground plane: one lit with reference ground truth lighting (middle) and two with estimated lighting (left and right). The observer is then asked to choose which of the two renders resembles the reference the most. This task studies how much observers are able to *match lighting to a reference*—in a sense, this is similar to what we ask IQA metrics to do. This will enable us, later on in sec. 5, to evaluate whether existing metrics perform similarly to humans *at the same task*.

Task 2. Fig. 2 (right) presents each observer with two images of a virtual scene (sphere on a plane) embedded within the background image from which lighting was estimated. The observer is then asked to select the most realistic of the two images. As opposed to task 1, no render with ground truth lighting is shown for this task. This task studies how observers *judge the virtual objects in context*, without having to rely on a reference.

To better distinguish the factors impacting human perception of lighting, we formulate four variants for each of those two tasks. The first two variants separate indoor and outdoor

scenes, as both cases are generally addressed differently in the literature. The last two variants use different materials applied to the virtual object: one diffuse and one glossy. The goal is to bring the observer to focus on the low and high frequencies aspects of lighting, respectively.

3.2. Stimuli

Virtual objects. To avoid potential issues with compositing errors, mismatching semantics and/or geometry, social biases of observers, etc., the virtual objects rendered are a simple sphere lying on a plane [36]. The sphere is seen closely (task 1, fig. 2 left) and farther away (task 2, fig. 2 right). The sphere is positioned farther away in task 2 to allow the observers to see the background, as the sphere and plane fill the render in task 1. The Disney Principled BRDF [5] is used for both diffuse (roughness 1.0 and specularity 0.0) and glossy (roughness 0.1 and specularity 1.0) materials, with an albedo of 0.18 in both cases. The images are generated with the Cycles physically-based rendering engine, and then reexposed and tonemapped with $\gamma = 2.4$ to be displayed on the monitor.

Lighting estimation methods. We include recent learning-based lighting estimation methods for both the in-and outdoor scenarios, including the current state-of-the-art by Weber *et al.* [54] (indoor), competitive GAN-based methods EverLight [8] (indoor and outdoor) and StyleLight [50] (indoor), and approaches predicting parametric lights: Gardner *et al.* [14] (indoor) and Zhang *et al.* [57] (outdoor). We also include Khan *et al.* [20] as an example of a simpler, non-learning technique, since it estimates the lighting by projecting the input image onto a sphere and then mirroring it to generate an environment map. See fig. 3 for examples of lighting estimated by the selected methods.

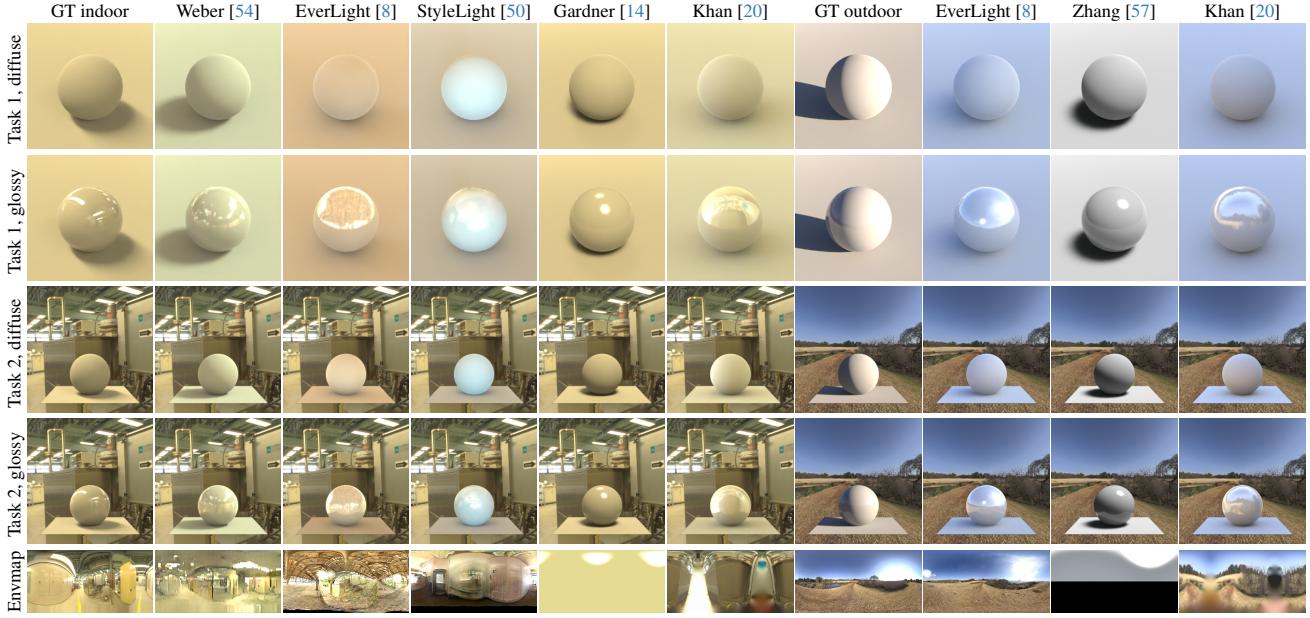


Figure 3. Example of the stimuli produced for a scene by each of the lighting estimation methods (columns), for the different tasks and experiments (rows). The last row corresponds to the estimated lighting (projected to an equirectangular format) by each of the methods and used for the renders. The “GT” columns correspond to the ground truth—note that it is not used in task 2 and shown here only for reference.

Input scenes. To ensure the diversity of the selected scenes, we compute the first-order spherical harmonics coefficients and cluster them using k -means, with $k = 25$. We select the 25 panoramas closest to each cluster centre. Finally, we extract a perspective image of limited field of view from each HDR panorama in our datasets by simulating a camera taking a picture within the environment.

3.3. Experimental setup

Hardware. The experiment was conducted in a controlled lab setting to ensure the uniformity of the data collected. It was carried out in a matte black room with a standard keyboard placed on a desk. The monitor was set to sRGB and was the only light source in the room. The observers were seated ≈ 70 cm away from the monitor, which results in a $11.5^\circ/17^\circ$ visual angle per image for task 1/2 respectively.

Procedure. Two distinct independent two-alternative forced choice (2AFC) tasks are given to the observers sequentially. The pairs of all combinations of the stimuli produced by the lighting estimation methods (5 indoor and 3 outdoor) for the 25 different scenes are presented to the observers. Thus, a total of 250 images (10 method combinations \times 25 scenes) for the indoor case and 75 images (3 method combinations \times 25 scenes) for the outdoor case are used. The order and the placement (left or right) of each method are randomised to reduce potential biases.

Participants. A total of 49 unique observers (33M/16F, ages ranging from 24–63) with normal or corrected-to-normal vision participated in the study. All the observers

were tested for colour blindness using the Ishihara test. Not all observers participated in all experiments, resulting in indoor experiments performed by 30/31 observers for the diffuse/glossy versions respectively and outdoor experiments by 12 observers. Participants were allowed to do each experiment at most once. None of the authors took part in the study. Each experiment takes ~ 25 min for the indoor and ~ 5 min for the outdoor cases. No time restriction is imposed to avoid inducing stress and bias.

4. Evaluating lighting estimation methods with perceptual data

In this section, we evaluate state-of-the-art lighting estimation methods in light of our perceptual data acquired in sec. 3. We begin by describing our statistical analysis approach, then highlight several key observations in the subsequent analysis.

4.1. Data processing

Since observers were asked to select an image produced by a particular lighting estimation method, we assign, for each method, a 1 if its result was selected and a 0 otherwise. For each of the individual experiments, the resulting z -scores for the lighting estimation methods are computed according to the Thurstone Case V Law of Comparative Judgement model [48] and are shown in fig. 4. This method receives as input a matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ of paired comparison data, where element $C_{i,j}$ indicates how many times method i was preferred over method j , and outputs a scaling z -score for

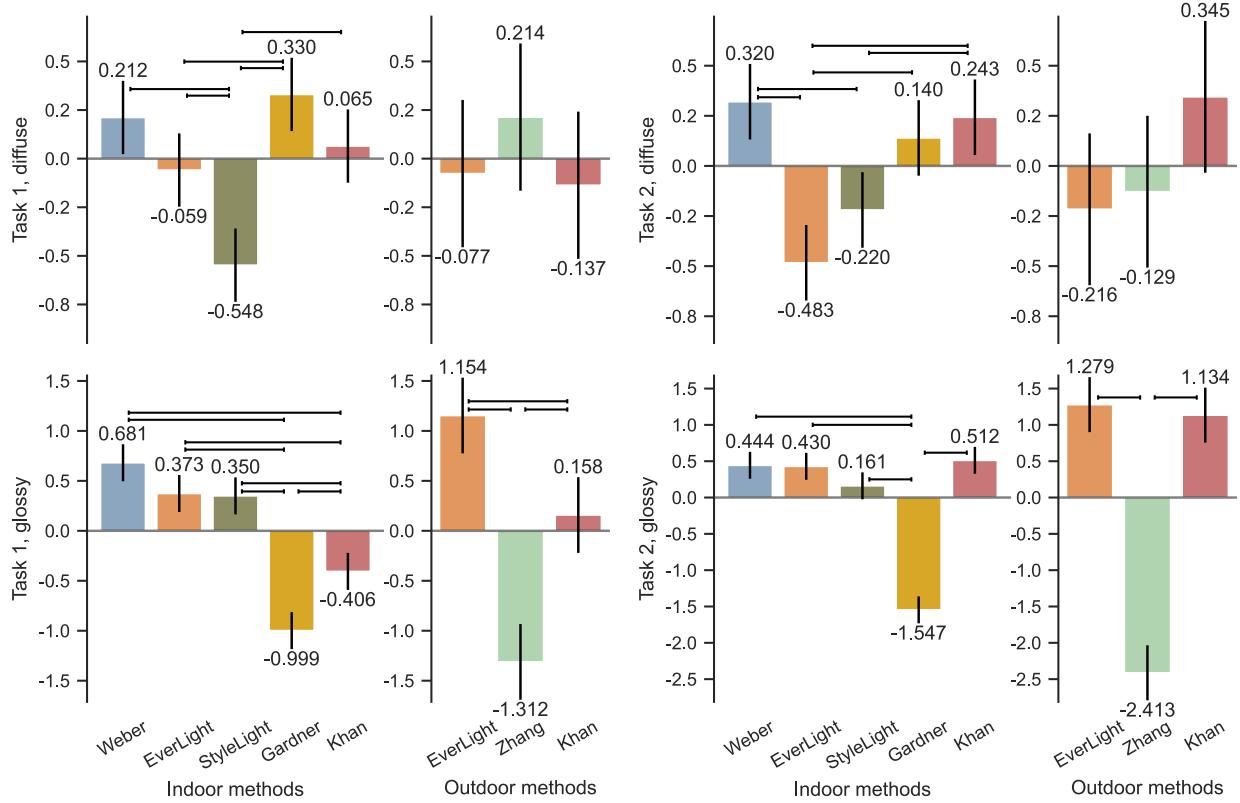


Figure 4. Thurstone Case V Law of Comparative Judgement (z -scores) for all the observers as a function of the different lighting estimation methods (bars), for the different materials (rows) and tasks (columns). A positive score indicates that observers generally prefer the stimuli rendered with the lighting estimation method, and not preferred when the score is negative. The scores of all the methods for an experiment sum to 0. The brackets above indicate pairs of methods for which the perceptual difference is statistically significant. Error bars correspond to 95 % confidence interval.

all N methods compared. This z -score is defined such that a higher value indicates higher observer preference, with a sum over all methods of 0 (so methods that are consistently not preferred get a negative value). We compute the z -score individually for all images from a scene (set of stimuli images produced by all N lighting estimation methods with the same scene given as input). The mean across all scenes is then taken to describe the performance of each method in a specific experiment.

The uncertainties, illustrated as error bars in fig. 4, correspond to the 95 % confidence interval using Montag’s method [34]. These confirm that the number of participants is sufficient, as statistical significance is reached (indicated by the horizontal brackets) for most of the experiments. However, the uncertainty is higher for the outdoor experiments, as it is only proportional to the number of observers. We also compute Fleiss’ κ [11] and Kuder-Richardson-20 (KR20) [22] between the data of each observers to demonstrate the agreement and consistency, respectively, in the supplementary material, which confirm that the number of participants for the outdoor experiment is sufficient.

4.2. Observations

At first glance, fig. 4 reveals that the observers’ judgement varies across tasks and materials. This suggests that judging the accuracy (task 1) vs plausibility (task 2) of lighting are cognitively different tasks. Therefore, humans likely focus on different cues when looking at images. We now highlight several key observations emanating from fig. 4. See the supplemental for more lighting estimation methods rankings.

Solely predicting light sources in parametric format is insufficient. First, we observe that methods focusing exclusively on predicting light sources in parametric format (Gardner *et al.* [14] indoors and Zhang *et al.* [57] outdoors) do perform well when matching diffuse renders to ground truth (task 1, diffuse), but are otherwise disliked by observers when the scene contains glossy objects.

Methods should predict a combination of HDR lighting and plausible textures. Overall, Weber *et al.* [54] seems to be the preferred method across all indoor experiments. It combines HDR lighting estimation that is similarly accurate to that of Gardner *et al.* [14] (task 1, diffuse) with plausible

textures closely matching the ground truth (task 1, glossy). Results from task 2 also reveal that its composites are consistently judged to be plausibly blended with their surroundings. Both EverLight [8] and StyleLight [50] are GAN-based approaches that aim at predicting HDR lighting and textures as well. Whilst they do succeed at predicting visually pleasing textures that create believable reflections on glossy materials, their drop in performance when diffuse materials are used indicates that they seem to struggle at generating accurate (task 1, diffuse) and plausible (task 2, diffuse) HDR lighting. These results indicate that accurate HDR lighting estimation is important in the diffuse spheres experiments, and that plausible textures seem more important for glossy objects.

Matching ground truth lighting matters less when judging the plausibility of composites. The older method of Khan *et al.* [20] is generally expected to perform less well than more recent methods since it creates distorted, implausible LDR textures. Indeed, results from task 1 show that its lighting estimates do not perceptually match the ground truth, especially in the case of glossy materials. However, when considering the plausibility of composited objects in the absence of a reference (task 2), the trend is reversed: the simple method of [20] yields results on par with those of [54] (indoors) and [8] (outdoors, glossy). Surprisingly, it even outperforms all the recent methods in one scenario (outdoors, diffuse)! This suggests that, even if the predicted lighting is perceived to be inaccurate (task 1), the accuracy is not a priority for the observers when judging the realism of inserted virtual objects.

5. Measuring the agreement between IQA metrics and perceptual data

In this section, we pit existing metrics against our perceptual data to determine the degree to which they agree about lighting perception. In particular, we consider Full-Reference metrics (RGB angular error [16], PSNR [19], RMSE, si-RMSE, SSIM [52], VIF [44], ΔE [55], LPIPS [58], PieAPP [40], FLIP [2], HDR-VDP3 [30]) as they are the standard in the community for lighting estimation method evaluation. We also include No-Reference metrics (BRISQUE [32], NIQE [33], UNIQUE [59], Hyper-IQA [45]) to evaluate whether they better match human perception. All metrics are computed on the tonemapped stimuli (as seen by the observers, c.f. sec. 3.2) except HDR-VDP3 which uses HDR renders (before gamma correction).

5.1. Agreement score

We aim to study if our collected perceptual data agrees with the results given by the metrics. To do so, we define an agreement score between a metric (or observer) and a hypothetical *perfect observer*. Larger values of this agreement score will mean a higher agreement with the perfect observer.

Let $\varphi_{a,b}^{(i)}$ be the binary choice of observer $i \in \{1, \dots, n\}$ between stimuli a and b , representing the output of two different lighting estimation methods for the same scene:

$$\varphi_{a,b}^{(i)} \triangleq \begin{cases} 1 & \text{if } a \text{ was selected over } b, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Note that this relation is symmetric, i.e., $\varphi_{b,a} = 1 - \varphi_{a,b}$ but we slightly abuse notation and write $\varphi_{a,b}$ for both cases. The proportion at which all n observers select each stimulus is computed as the mean $\bar{\varphi}_{a,b} = \frac{1}{n} \sum_{i=1}^n \varphi_{a,b}^{(i)}$. We are interested in the comparisons where most of the observers selected a , so all values of $\bar{\varphi}_{a,b} < 0.5$ (i.e. when b is the most selected) are set to 0. The agreement score $\omega^{(i)}$ between observer i and the mean choice of all observers $\bar{\varphi}_{a,b}$ is

$$\omega^{(i)} = \frac{\sum_{(a,b)} \bar{\varphi}_{a,b} \varphi_{a,b}^{(i)}}{\sum_{(a,b)} \bar{\varphi}_{a,b}}. \quad (2)$$

Finally, the *expected observer agreement* $\bar{\omega} = \frac{1}{n} \sum_{i=1}^n \omega^{(i)}$ is the mean agreement over all observers.

A hypothetical *perfect observer* is defined as an observer who makes the exact same choices as the average observer. In this case, $\omega^{(i)} = 1$ and is the upper-bound score. Conversely, a *random observer* is defined as an observer who would randomly select images. Since pairs of images are presented, its agreement score is $\omega^{(i)} = 0.5$. We removed all data from individual observers with an agreement score $\omega^{(i)} < 0.5$ in at least one experiment (5 individuals in our study) since they likely have misunderstood the task.

We also define the agreement score for each IQA metric. For this, we recompute eq. (2) by replacing observer choices φ with the ones made by metrics. Note that metrics choose the best image by selecting the image most (dis)similar to ground truth, even though observers do not always have access to it (e.g., task 2).

5.2. Observations

The agreement scores for each IQA metric are shown alongside the expected observer agreement in fig. 5 for all experiments. The grey lines correspond to the perfect (upper) and random (lower) observers. The orange/blue lines correspond to the expected observer agreement score for the indoor/outdoor experiments, respectively. A higher metric agreement score indicates that it tends to select the same stimuli as the observers.

Note that metrics generally select the same images irrespective of the task since the geometry, material and lighting in the renders are the same. However, perceptual results are task-dependent (c.f. sec. 4.2), which suggests that humans and metrics use different processes when evaluating lighting

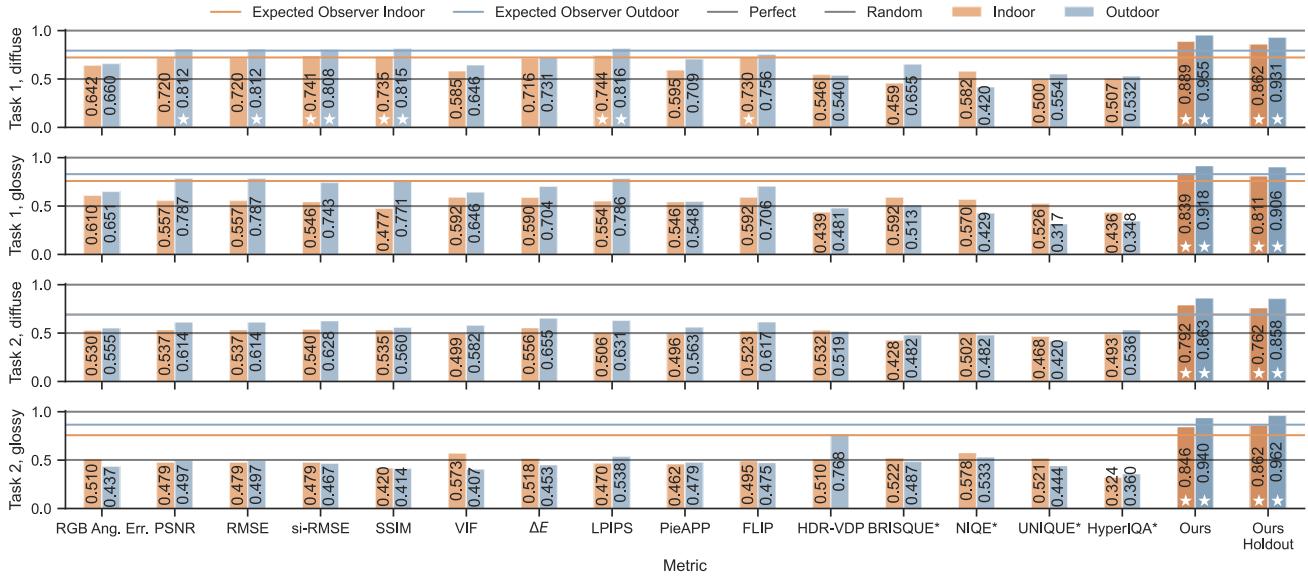


Figure 5. Agreement between the observer scores and the metric scores (columns) for all the lighting estimation methods (indoor: orange bars; outdoor: blue bars), for the different types of experiments (rows). The lower horizontal grey line is set at chance level (~ 0.5) and the higher one corresponds to the perfect observer (set at 1.0). The orange (indoor) and blue (outdoor) lines correspond to the expected observer agreement score (for task 2, diffuse, blue and orange lines are overlapped). The stars indicate methods that have an agreement score equal or superior to the expected observer. “Ours” and “Ours Holdout” refer to our learned metric combination, see sec. 6 for more details. The No-Reference IQA metrics are indicated by asterisks.

in different images, shown by the difference in agreement with the metrics and humans in fig. 5.

Metrics tend to perform well on task 1 with diffuse materials. The stars on fig. 5 indicate the metrics with agreement scores equal or superior to the expected observer, indicating that they agree with human perception. This occurs only with si-RMSE, SSIM, LPIPS, and FLIP for task 1 diffuse (indoor), and PSNR, RMSE, si-RMSE, SSIM, and LPIPS for task 1 diffuse (outdoor). Thus, these metrics may indeed be appropriate for these specific tasks, where observers are asked to match objects with diffuse appearance.

Metrics do not agree with human perception for lighting estimation for the rest of the tasks. Fig. 5 shows that IQA metrics do not usually agree with observers. In many cases, especially for task 1 glossy and task 2 diffuse and glossy, metrics even perform on par with a random observer. Our study demonstrates that we should not rely on such metrics, in the general case, for evaluating lighting estimation algorithms as they do not accurately reflect human judgement.

5.3. Correlation

We computed Spearman’s ρ and Kendall’s τ statistical tests to measure the correlation between the binary observers’ choices and the estimated metric choices, obtained from fitting a logistic curve as recommended by [51]. Tab. 1 shows the best two metrics for each indoor experiment.

Our metric always obtains the highest correlation scores,

Table 1. Correlation scores of our metric and the next best metric (excluding “Ours Holdout”) for Spearman’s ρ and Kendall’s τ tests, for all indoor experiments.

Task	Material	Metric	Spearman’s ρ	Kendall’s τ
Task 1	Diffuse	Ours	0.689	0.572
		LPIPS	0.529	0.440
		Glossy	0.724	0.601
	Glossy	RGB Ang. Err.	0.226	0.187
		Ours	0.711	0.593
		BRISQUE	0.145	0.121
Task 2	Diffuse	Ours	0.753	0.628
		HyperIQA	0.371	0.308

matching the perception data best and corroborating the analysis from the agreement scores in fig. 5. We note there is still room for improvement since scores are lower than 0.8, which is the threshold typically considered as reliable by those tests, thus paving the way for future work.

Overall, no metric can accurately judge the lighting estimation in all contexts, either on accuracy (task 1) or plausibility (task 2), for all materials (diffuse or glossy), in all domains (indoor and outdoor).

6. Learning a metric combination

In this section, we propose to learn task-specific functions which map existing IQA metrics to perceptual data. This is similar to [6] who employed such a strategy for detecting

localised distortions in images. The insight is that these metrics each look for slightly different cues within the image, which provides a signal that a simple learner can leverage. Here, we learn four different functions, one for each of the experiments (two tasks and two materials, indoor and outdoor are considered jointly).

6.1. Formulation and training

We seek a function $f_e : \mathbb{R}^2 \mapsto \mathbb{R}$ for experiment $e \in \{1, \dots, 4\}$ which maps a pair of input images $\mathbf{I}_a, \mathbf{I}_b$ to a perception score $\varphi_{a,b}$ (as in eq. (1)), such that $f_e(\mathbf{I}_a, \mathbf{I}_b) \approx \varphi_{a,b}$:

$$f_e(\mathbf{I}_a, \mathbf{I}_b) \equiv \psi_e(\{\ell_k(\mathbf{I}_1, \mathbf{I}^*) - \ell_k(\mathbf{I}_2, \mathbf{I}^*)\}_{k=1}^K), \quad (3)$$

where \mathbf{I}^* is the ground truth image and $\ell_k, k \in \{1, \dots, K\}$ is the set of all $K = 15$ metrics from sec. 5. ψ_e is implemented using an SVR (ϵ -Support Vector Regression [38]), with $\epsilon = 0.1$ and $C = 1$. As shown in eq. (3), the SVR takes as input the set of differences between all K metrics scores computed on the images and their corresponding GT. The SVR was selected as the best variant amongst a set of simple learners (see the supplementary material for more details).

To train ψ_e , we split our perceptual data into 20 training and 5 validation scenes and consider the comparisons without replacement between 5/3 indoor/outdoor methods (yielding 10/6 combinations), respectively. Considering permutations of the input combinations to make ψ_e agnostic to the order of the input images, the resulting dataset contains 640 and 160 data points for training and validation.

We report the results of our metric on our validation set in fig. 5, where our proposed metric dominates all other metrics. This demonstrates that existing metrics do capture valuable information and that, while a *single* metric does not accurately match perception in all cases, a *learned combination* of them can approximate it much more precisely.

6.2. Generalisation to other methods

We evaluate the generalisation capabilities of our learned metric on other lighting estimation methods in two ways. First, we hold out all perception data of one of the methods (Khan *et al.* [20] was selected arbitrarily) and retrained ψ_e on the remaining methods. Once the metric is trained, we test it on the unseen (Khan) data, and show results in the “Ours Holdout” column of fig. 5. This is done only for the indoor data as outdoor data is too scarce for such analysis.

Second, we run another psychophysical study with 6 observers using the same methodology as in sec. 3, but employing different indoor lighting estimation methods. This time, we use a re-implementation of EverLight [8] at 1024×2048 resolution, Weber *et al.* [54] (used previously but never in paired comparison with the new methods), Garon *et al.* [15], a Stable Diffusion model trained for image outpainting [42], and a simple baseline where the environment map has constant colour equal to the mean colour of the

input image. Our metric, trained on all the data points obtained in the main user study, obtains an agreement score for the diffuse/glossy experiment of 0.786/0.856 for task 1 and 0.787/0.889 for task 2. These agreement scores are superior to what the best standard metrics in sec. 5—VIF, NIQE, RGB Ang. Err., BRISQUE—can provide (0.670/0.809 and 0.686/0.691), which validate the generalisation of our proposed metric. The agreement scores of the best standard metrics are often outliers, with metrics on average performing around chance level.

By releasing all data and code publicly, we hope this learned metric combination will serve as guide for future work in lighting estimation by providing a way to validate its performance in a perceptually meaningful manner.

7. Discussion

This paper explores a new perceptual evaluation framework for lighting estimation algorithms when used for relighting virtual objects into photographs. To do so, it presents a controlled psychophysical study which compares several lighting estimation methods from the recent literature by asking observers to judge scenes rendered with lighting estimates. It then demonstrates that image quality assessment (IQA) metrics typically used to quantitatively evaluate lighting estimation methods seldom coincide with human perception.

Our analysis emanating from this study provides some key insights. Most importantly, human perception is dependent on the task at hand. While IQA metrics generate almost identical results for both our tasks, observer behaviour changes according to whether they must match lighting to a reference, or judge the plausibility of virtual objects in context of a background image. Both scenarios should be considered when evaluating future methods.

To help future lighting estimation approaches in evaluating their performance in a perceptually meaningful way, we propose a learned metric combination which is shown, through generalisation experiments, to accurately predict observers preference on images from unseen methods.

Despite the novel insights our analysis provides and our proposed learned metric combination, understanding the link between human perception and lighting estimation algorithms is still largely uncharted—our paper is but a first step. Further analysis of how specific algorithmic design choices (e.g., choice of lighting representation) affect human perception is an important future direction to explore. It is our hope that this work will pave the way for significant research into effective methods for evaluating the lighting estimation problem, methods that will be more tightly tied to human perception.

This research was supported by Sentinel North and NSERC grant RG-PIN 2020-04799. JVC was supported by Grant PID2021-128178OB-I00 funded by MCIN/AEI/10.13039/501100011033, ERDF “A way of making Europe” and by the Departament de Recerca i Universitats from Generalitat de Catalunya with reference 2021SGR01499.

References

- [1] Barton L. Anderson and Juno Kim. Image statistics do not explain the perception of gloss and lightness. *J. Vis.*, 9(11):10, 2009. [2](#)
- [2] Pontus Andersson, Jim Nilsson, Tomas Akenine-Möller, Magnus Oskarsson, Kalle Åström, and Mark D. Fairchild. FLIP: a difference evaluator for alternating images. *ACM Comp. Graph. Int. Tech.*, 3(2):15:1–15:23, 2020. [2, 6](#)
- [3] Christophe Bolduc, Justine Giroux, Marc Hébert, Claude Demers, and Jean-François Lalonde. Beyond the pixel: a photometrically calibrated HDR dataset for luminance and color prediction. In *Int. Conf. Comput. Vis.*, 2023. [1](#)
- [4] Huseyin Boyaci, Katja Doerschner, and Laurence T Maloney. Cues to an equivalent lighting model. *J. Vis.*, 6(2):2–2, 2006. [2](#)
- [5] Brent Burley. Physically-based shading at disney. 2012. [3](#)
- [6] Martin Čadík, Robert Herzog, Rafał Mantiuk, Radosław Mantiuk, Karol Myszkowski, and Hans-Peter Seidel. Learning to predict localized distortions in rendered images. In *Comput. Graph. Forum*, pages 401–410, 2013. [7](#)
- [7] Dachuan Cheng, Jian Shi, Yanyun Chen, Xiaoming Deng, and Xiaopeng. Zhang. Learning scene illumination by pairwise photos from rear and front mobile cameras. *Comput. Graph. Forum*, 37(7):213–221, 2018. [1](#)
- [8] Mohammad Reza Karimi Dastjerdi, Jonathan Eisenmann, Yannick Hold-Geoffroy, and Jean-François Lalonde. Everlight: Indoor-outdoor editable HDR lighting estimation. In *Int. Conf. Comput. Vis.*, 2023. [2, 3, 4, 6, 8](#)
- [9] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5):2567–2581, 2020. [2](#)
- [10] Keyan Ding, Yi Liu, Xueyi Zou, Shiqi Wang, and Kede Ma. Locally adaptive structure and texture similarity for image quality assessment. In *ACM Int. Conf. Multimedia*, 2021. [2](#)
- [11] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971. [5](#)
- [12] Pedro Garcia Freitas, Luísa Peixoto da Eira, Samuel Soares Santos, and Mylène CQ Farias. Image quality assessment using bsif, clbp, lcp, and lpq operators. *Theoretical Computer Science*, 805:37–61, 2020. [2](#)
- [13] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Trans. Graph.*, 9(4), 2017. [1, 2](#)
- [14] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep parametric indoor lighting estimation. In *Int. Conf. Comput. Vis.*, 2019. [1, 2, 3, 4, 5](#)
- [15] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. Fast spatially-varying indoor lighting estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. [1, 2, 8](#)
- [16] Arjan Gijsenij, Theo Gevers, and Marcel P. Lucassen. Perceptual analysis of distance measures for color constancy algorithms. *J. Opt. Soc. Am. A*, 26(10):2243–2256, 2009. [2, 6](#)
- [17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Adv. Neural Inform. Process. Syst.*, page 6629–6640, 2017. [2](#)
- [18] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. [1, 2](#)
- [19] Alain Horé and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *Int. Conf. Pattern Recog.*, 2010. [2, 6](#)
- [20] Erum Arif Khan, Erik Reinhard, Roland W. Fleming, and Heinrich H. Bülthoff. Image-based material editing. *ACM Trans. Graph.*, 25(3):654–663, 2006. [3, 4, 6, 8](#)
- [21] Jan J Koenderink, Andrea J Van Doorn, and Sylvia C Pont. Light direction from shad(ow)ed random gaussian surfaces. *Perception*, 33(12):1405–1420, 2004. [2](#)
- [22] G Frederic Kuder and Marion W Richardson. The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160, 1937. [5](#)
- [23] Jean-François Lalonde and Iain Matthews. Lighting estimation in outdoor image collections. In *Int. Conf. 3D Vis.*, 2014. [2](#)
- [24] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Estimating the natural illumination conditions from a single outdoor image. *Int. J. Comput. Vis.*, 98(2):123–145, 2012. [2](#)
- [25] Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. [2](#)
- [26] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. [1, 2](#)
- [27] Zhengqin Li, Li Yu, Mikhail Okunev, Manmohan Chandraker, and Zhao Dong. Spatiotemporally consistent hdr indoor lighting estimation. *ACM Trans. Graph.*, 42(3), 2023. [2](#)
- [28] Jorge Lopez-Moreno, Veronica Sundstedt, Francisco Sangorri, and Diego Gutierrez. Measuring the perception of light inconsistencies. In *Symp. App. Percept. Graph. Vis.*, 2010. [2](#)
- [29] Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4):1–14, 2011. [2](#)
- [30] Rafal K Mantiuk, Dounia Hammou, and Param Hanji. Hdr-vdp-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content. *arXiv preprint arXiv:2304.13625*, 2023. [6](#)
- [31] Phillip J. Marlow, Juno Kim, and Barton L. Anderson. The perception and misperception of specular surface reflectance. *Curr. Biol.*, 22(20):1909–1913, 2012. [2](#)
- [32] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012. [2, 6](#)

- [33] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013. [2](#) [6](#)
- [34] Ethan D Montag. Empirical formula for creating error bars for the method of paired comparison. *J. Elec. Imag.*, 15(1):010502–010502, 2006. [5](#)
- [35] Richard F Murray and Wendy J Adams. Visual perception and natural illumination. *Curr. Opinion Behavioral Sc.*, 30: 48–54, 2019. [1](#)
- [36] James P O’Shea, Maneesh Agrawala, and Martin S Banks. The influence of shape cues on the perception of lighting direction. *J. Vis.*, 10(12):21–21, 2010. [2](#) [3](#)
- [37] Bradley Pearce, Stuart Crichton, Michal Mackiewicz, Graham D Finlayson, and Anya Hurlbert. Chromatic illumination discrimination ability reveals that human colour constancy is optimised for blue daylight illuminations. *PloS one*, 9(2), 2014. [2](#)
- [38] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. [8](#)
- [39] Sylvia C Pont and Jan J Koenderink. Matching illumination of solid objects. *Perception & psychophysics*, 69(3):459–468, 2007. [2](#)
- [40] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. [2](#) [6](#)
- [41] Ganesh Ramanarayanan, James Ferwerda, Bruce Walter, and Kavita Bala. Visual equivalence: towards a new standard for image fidelity. *ACM Trans. Graph.*, 26(3), 2007. [2](#)
- [42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. [8](#)
- [43] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Adv. Neural Inform. Process. Syst.*, page 2234–2242, 2016. [2](#)
- [44] Hamid R. Sheikh, Alan C. Bovik, and Gustavo de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Trans. Image Process.*, 14(12):2117–2128, 2005. [2](#) [6](#)
- [45] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. [2](#) [6](#)
- [46] Minghui Tan, Jean-François Lalonde, Lavanya Sharan, Holly Rushmeier, and Carol O’Sullivan. The perception of lighting inconsistencies in composite outdoor scenes. *ACM Trans. Appl. Percept.*, 12(4):1–18, 2015. [2](#)
- [47] Susan F Te Pas, Sylvia C Pont, Edwin S Dalmaijer, and Ignace TC Hooge. Perception of object illumination depends on highlights and shadows, not shading. *J. Vis.*, 17(8):2–2, 2017. [2](#)
- [48] Louis L Thurstone. A law of comparative judgment. In *Scaling*, pages 81–92. Routledge, 1927. [4](#)
- [49] Javier Vazquez-Corral, C Alejandro Párraga, María Vanrell, and Ramón Baldrich. Color constancy algorithms: Psychophysical evaluation on a new dataset. *J. Imaging Sci. Technol.*, 1(3):1, 2009. [2](#)
- [50] Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. Stylelight: HDR panorama generation for lighting estimation and editing. In *Eur. Conf. Comput. Vis.*, 2022. [2](#) [3](#) [4](#), [6](#)
- [51] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conf. Signals Syst. Comp.*, 2003. [7](#)
- [52] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. [1](#) [2](#) [6](#)
- [53] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning indoor inverse rendering with 3D spatially-varying lighting. In *Int. Conf. Comput. Vis.*, 2021. [2](#)
- [54] Henrique Weber, Mathieu Garon, and Jean-François Lalonde. Editable indoor lighting estimation. In *Eur. Conf. Comput. Vis.*, 2022. [2](#) [3](#) [4](#), [5](#), [6](#), [8](#)
- [55] Günther Wyszecki and Walter Stanley Stiles. *Color science: concepts and methods, quantitative data and formulae*. John Wiley & Sons, 2000. [2](#) [6](#)
- [56] Syed Waqas Zamir, Javier Vazquez-Corral, and Marcelo Bertalmío. Vision models for wide color gamut imaging in cinema. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(5): 1777–1790, 2021. [2](#)
- [57] Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenman, and Jean-François Lalonde. All-weather deep outdoor lighting estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. [1](#) [2](#) [3](#) [4](#), [5](#)
- [58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. [1](#) [2](#) [6](#)
- [59] Weixia Zhang, Kede Ma, Guantao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Trans. Image Process.*, 30:3474–3486, 2021. [2](#) [6](#)