Towards a Perceptual Evaluation Framework for Lighting Estimation Supplementary Material

Justine Giroux[1] Mohammad Reza Karimi Dastjerdi[1] Yannick Hold-Geoffroy[2]
Javier Vazquez-Corral[3,4] Jean-François Lalonde[1]
[1]Université Laval, [2]Adobe Research, [3] Computer Vision Center, [4]Universitat Autònoma de Barcelona

This document complements our main paper, providing the following supplementary information regarding the conducted experiments and analysis:

- Additional description of the lighting estimation methods (see sec 3.2. of the main paper) used in sec. 1.1;
- A more in-depth description of the scene selection (see sec 3.2. of the main paper) given as input to the lighting estimation methods in sec. 1.2;
- Details on the geometry, materials, and rendering of the stimuli used in the psychophysical study (see sec 3.2. of the main paper), in sec. 1.3;
- Description of the hardware used for the psychological experiment (see sec 3.3. of the main paper) in sec. 1.4;
- Additional information on the procedure during the psychophysical experiments (see sec 3.3. of the main paper), in sec. 1.5;
- A more in-depth analysis of the participants in the psychological experiment (see sec 3.3. of the main paper), in sec. 1.6;
- Supplementary statistical tests on the psychophysical results (see sec 4.1. of the main paper) are computed in sec. 2.1;
- Additional analysis of the scores obtained in the psychophysical experiments (see sec 4.2. of the main paper), with examples (in sec. 2.2), per image score (in sec. 2.3), and agreement for individual observers (in sec. 2.4);
- Additional analysis of the scores of the various metrics (see sec 5. of the main paper), in sec. 3.1, statistical testing in sec. 3.2, and for FID in sec. 3.3;
- Additional comparisons between different network architectures and an analysis of the selected network (see sec 6.1. of the main paper), in sec. 4.1;
- Additional content regarding the generalisation study conducted with the selected architecture (see sec 6.2. of the main paper), in sec. 4.2.

1. Psychophysical experiment

The various steps of our psychophysical study are described in the following sections. The lighting estimation methods used are described in sec. 1.1, and the selected scene given as input to them is detailed in sec. 1.2. The design of the stimuli is explained in sec. 1.3. The hardware and the procedure of the experiment are discussed in sec. 1.4 and sec. 1.5, respectively. The study participants are described in more detail in sec. 1.6.

1.1. Lighting estimation methods

In the following sections, we describe the lighting estimation methods used in the indoor and outdoor psychophysical studies (see sec 3.2. of the main paper).

Environment map. We consider three state-of-the-art non-parametric lighting estimation methods to light our virtual scene. Weber *et al*. [29] proposes a two-stage indoor-only approach, where a dominant light source and scene layout are estimated and given as input to a texture network to predict the entire environmental texture based on the input image. EverLight [6] proposes a different two-stage method working simultaneously indoors and outdoors. It first estimates the lighting parameters as spherical gaussians, and then integrates them into environment map generation via guided co-modulation [5]. StyleLight [27] is a recent method that leverages the training of StyleGAN [13] with GAN inversion [22] to predict complete 360° environment maps from input images.

Parametric. We consider Gardner *et al*. [8] to provide parametric indoor lighting estimations. This method predicts three light sources parameterised by their direction, distance, angular size, and colour. For the outdoor parametric lighting model, we chose Zhang *et al*. [30], which trains a network to directly estimate the Lalonde-Matthews sky parameters [15] from a given outdoor image.

Classical. In contrast to these recent sophisticated learning-based methods, we also include Khan *et al*. [14]. This technique lacks any learning components and instead determines the lighting conditions by projecting the background image onto a sphere and then mirroring it to generate a complete LDR environment map.

All the environment maps generated by the indoor and outdoor lighting estimation methods for the user study are

presented in fig. 1. The first and ninth columns display the input given to the lighting estimation methods (more details in sec. 1.2), extracted from the ground-truth panoramas, shown in the third and eleventh columns. The second and tenth columns show the reconstructed first-order spherical harmonics, used to select the scenes (more details in sec. 1.2).

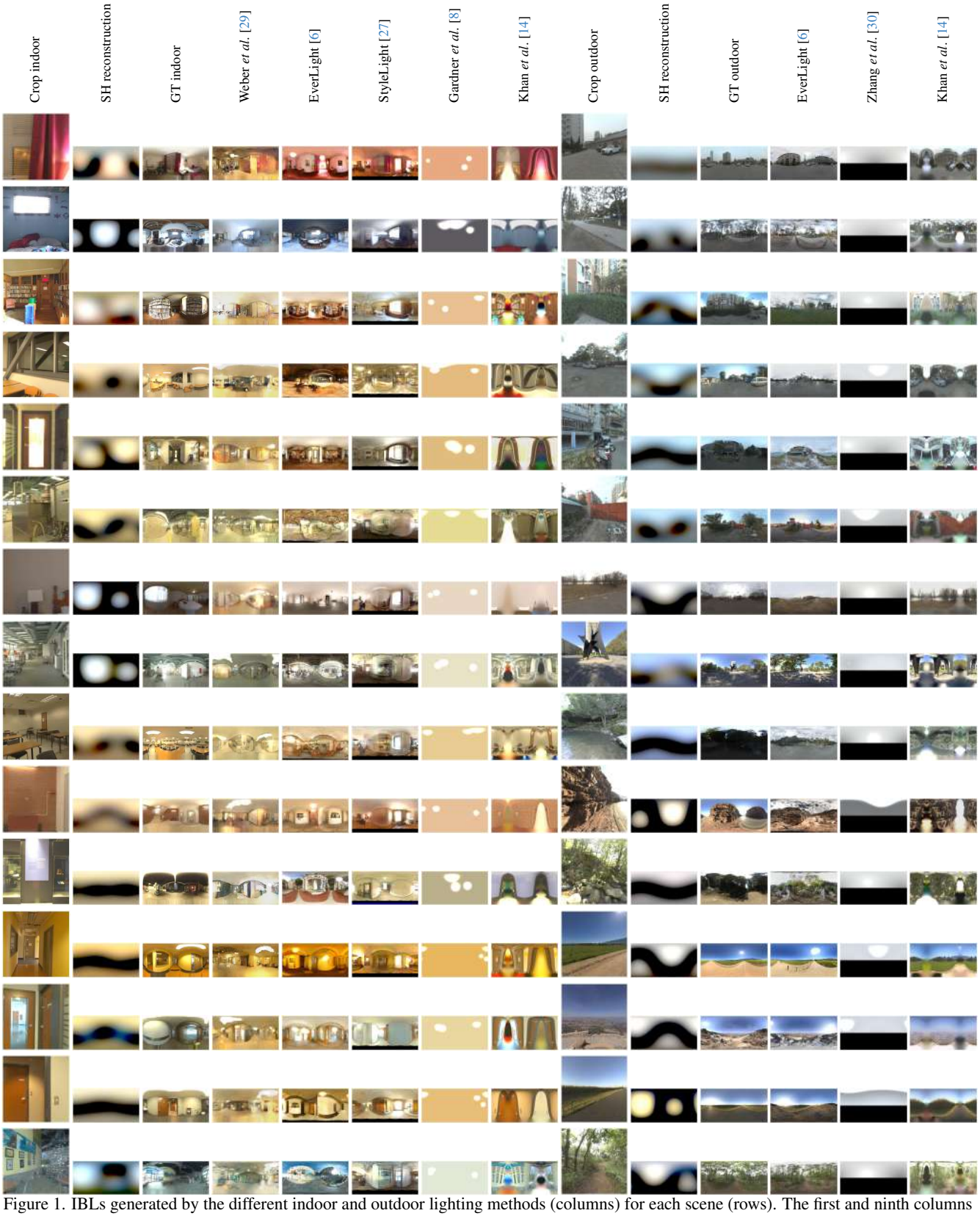

Figure 1. IBLs generated by the different indoor and outdoor lighting methods (columns) for each scene (rows). The first and ninth columns correspond to the region extracted from the indoor/outdoor scene, corresponding to a $50^{\circ}/90^{\circ}$ FoV. This region is taken from the centre of the full GT panorama (for most scenes), shown in the third and eleventh columns. The second and tenth column correspond to the reconstruction of the first-order spherical harmonics, showing the variety of the lighting in the selected scenes. The IBLs are reexposed and tonemapped with $\gamma = 2.4$ for display.

Figure 1. (contd) IBLs generated by the different indoor and outdoor lighting methods (columns) for each scene (rows). The first and ninth columns correspond to the region extracted from the indoor/outdoor scene, corresponding to a 50°/90° FoV. This region is taken from the centre of the full GT panorama (for most scenes), shown in the third and eleventh columns. The second and tenth column correspond to the reconstruction of the first-order spherical harmonics, showing the variety of the lighting in the selected scenes. The IBLs are reexposed and tonemapped with $\gamma = 2.4$ for display.

We utilise the output generated by these models as environment maps to illuminate the synthetic scene (sec. 1.3). For the parametric models, we initially convert the output parameters into environment maps and employ them in rendering stimuli.

1.2. Lighting estimation input scenes

High dynamic range (HDR) panorama images are used to extract limited FoV low dynamic range (LDR) regions to give as input to the lighting estimation methods (see sec 3.2. of the main paper). In our assessment of indoor lighting estimation techniques, we adhered to the procedures outlined in Weber *et al*. [29]. Our evaluation was conducted on the test set of Laval indoor dataset [7], comprising 224 high-resolution HDR panoramas. Within this test set, we systematically extracted 10 LDR images from each of the panoramas using the sampling distribution identical to Weber *et al*. [29]. This process yielded a grand total of 2240 images for our evaluation. We adopted the approach detailed in [6] to assess outdoor lighting. This method leverages 839 distinct outdoor HDR panoramas sourced from SHLight dataset [3]. From these, it derives three LDR images according to the sampling distribution of [6], resulting in an evaluation set comprising 2517 images.

Region extraction. To obtain the images given as input to the lighting estimation methods, an FoV of $50°/90°$ is extracted from the centre of the indoor/outdoor HDR panorama, which is tonemapped with $\gamma = 2.4$ and reexposed. The extracted regions have a resolution of 512×512 for the indoor images and 256×256 for the outdoor images. Examples of the extracted regions used in this study are shown in the first and ninth columns of fig. 1.

Scene selection. 25 scenes are selected from the indoor and outdoor datasets. We limited the number of scenes in our study to keep the experiment time below $\sim$30 min, in order to avoid errors caused by observers' fatigue. 25 scenes are considered sufficient to represent different types of environments with diverse lighting. In order to have a great variety of lighting environments, the coefficients of the first-order spherical harmonics are extracted from the HDR panoramas, using the `skylibs` Python library. The scenes are selected by taking the medoid of the clusters obtained using the `k-means` algorithm, where $k = 25$, from the `sklearn` Python library. The resulting clusters for the indoor (left) and outdoor (right) are shown in fig. 2. Examples of the first-order spherical harmonics reconstruction of the selected scenes are shown in the second and tenth columns of fig. 1, which indeed demonstrates a great variety. The scenes that contain too much noise are removed from the dataset, as they would potentially distract the observers from judging the realism of the inserted virtual object.

1.3. Stimuli

The stimuli for tasks 1 and 2 (see sec 3.2. of the main paper) have different geometries (described below), and each task has a diffuse and glossy variation, with details regarding the materials given subsequently. The rendering details are also indicated.

For reference, the stimuli used in the indoor and outdoor psychophysical experiments, for the diffuse/glossy experiments are shown in fig. 3/fig. 4 for task 1 and fig. 5/fig. 6 for task 2.

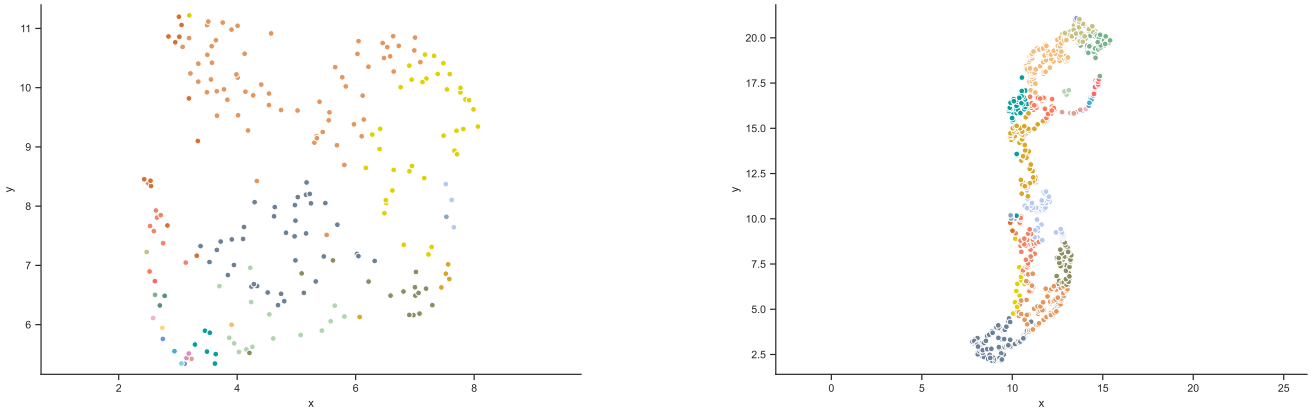

Figure 2. Clusters obtained `k-means` algorithm, where $k = 25$, for the indoor (left) and outdoor (right) high dynamic range (HDR) panorama datasets. The projection to $\mathbb{R}^2$ is done using UMAP [18]. The different colours indicate the different clusters. The axes are in arbitrary units.

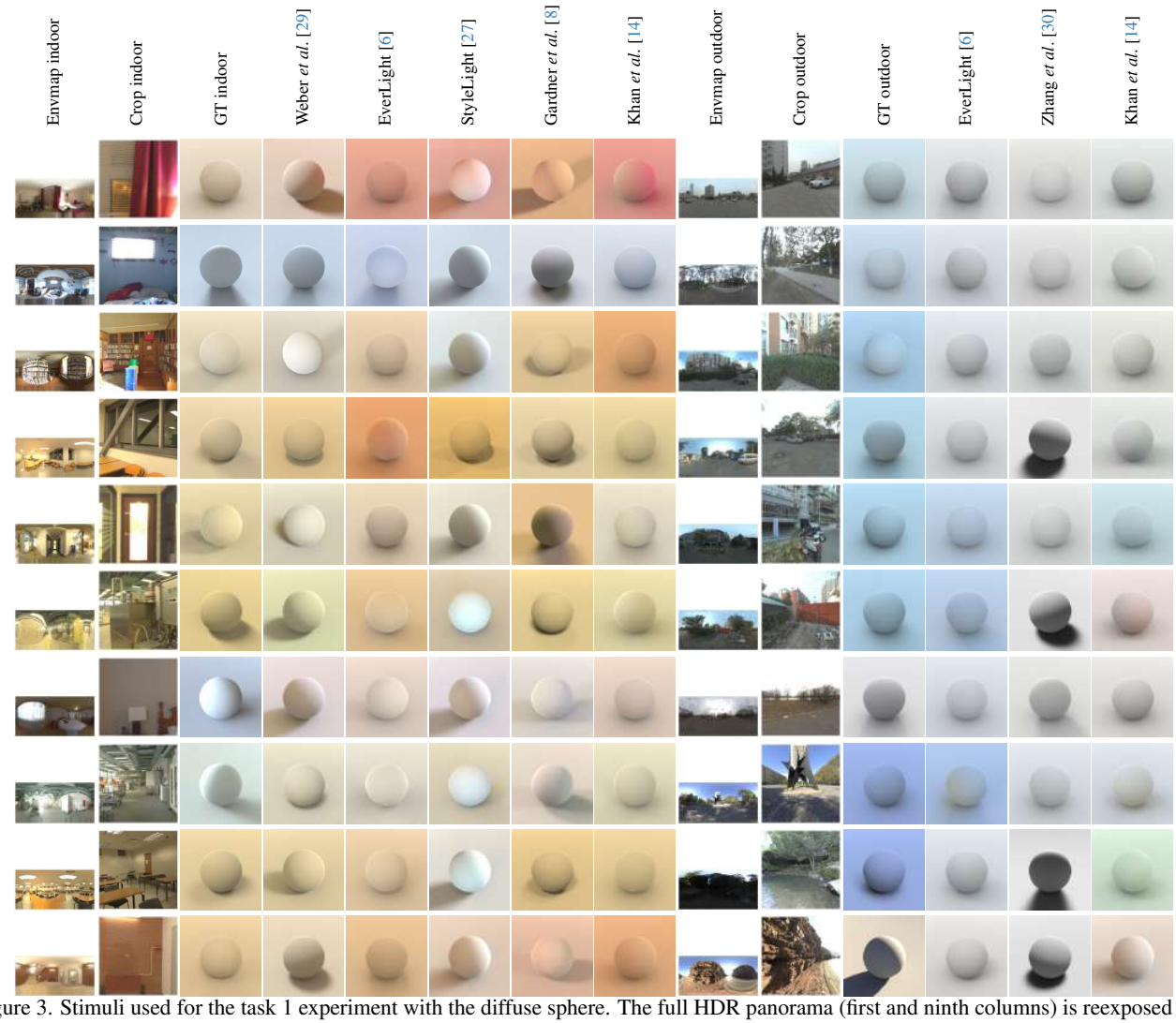

Figure 3. Stimuli used for the task 1 experiment with the diffuse sphere. The full HDR panorama (first and ninth columns) is reexposed and tonemapped with $\gamma = 2.4$ for display, for the indoor and outdoor cases. The region extracted from the scene (second and tenth columns), corresponding to a $50°/90°$ FoV, taken from the centre of the full indoor/outdoor panorama (for most scenes). The rendered stimuli using the ground truth IBLs (first/ninth columns) are shown in the third/eleventh columns. The other columns are the rendered stimuli using the IBLs (shown in fig. 1) produced by the different lighting estimation methods.

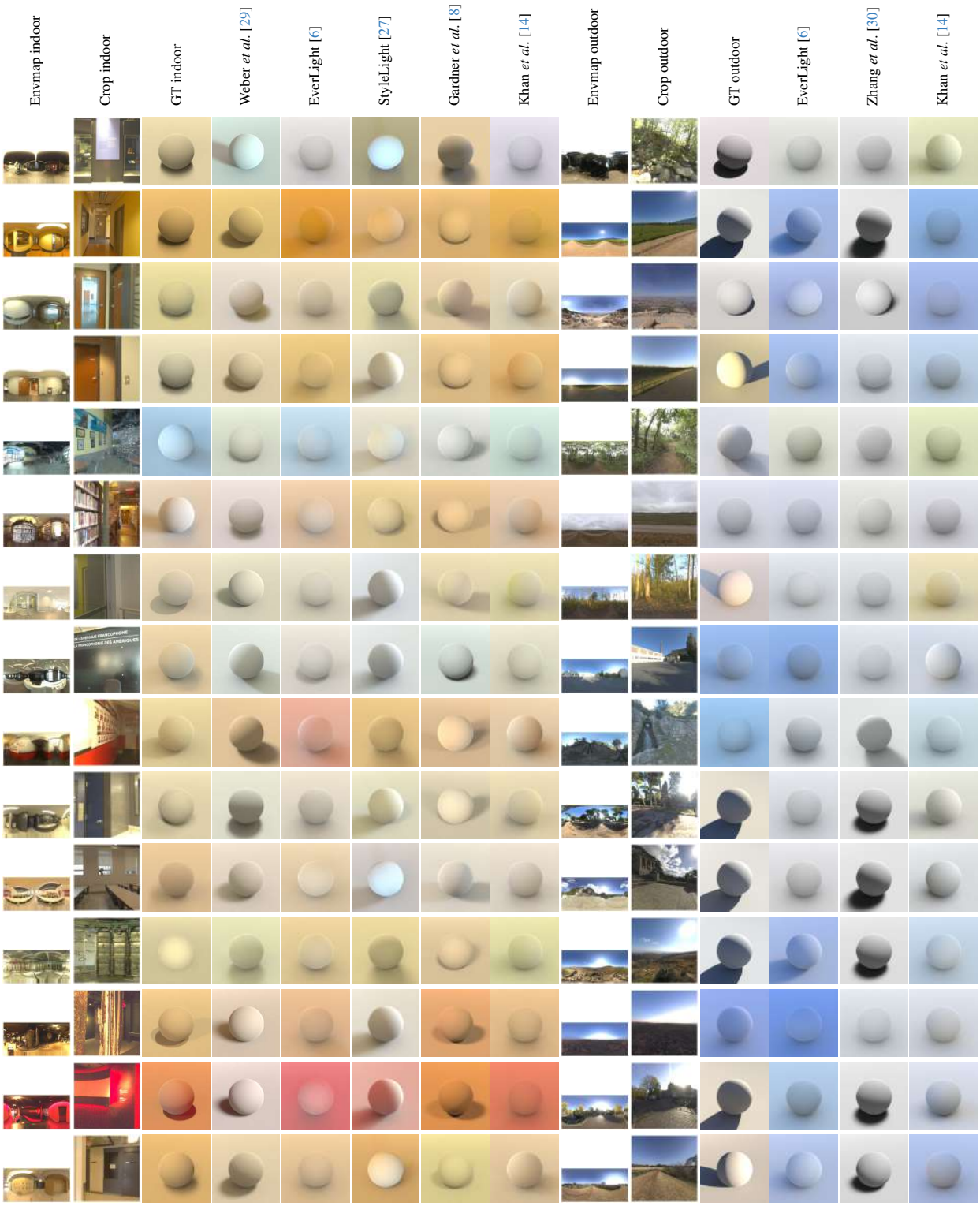

Figure 3. (contd) Stimuli used for the task 1 experiment with the diffuse sphere. The full HDR panorama (first and ninth columns) is reexposed and tonemapped with $\gamma = 2.4$ for display, for the indoor and outdoor cases. The region extracted from the scene (second and tenth columns), corresponding to a $50°/90°$ FoV, taken from the centre of the full indoor/outdoor panorama (for most scenes). The rendered stimuli using the ground truth IBLs (first/ninth columns) are shown in the third/eleventh columns. The other columns are the rendered stimuli using the IBLs (shown in fig. 1) produced by the different lighting estimation methods.

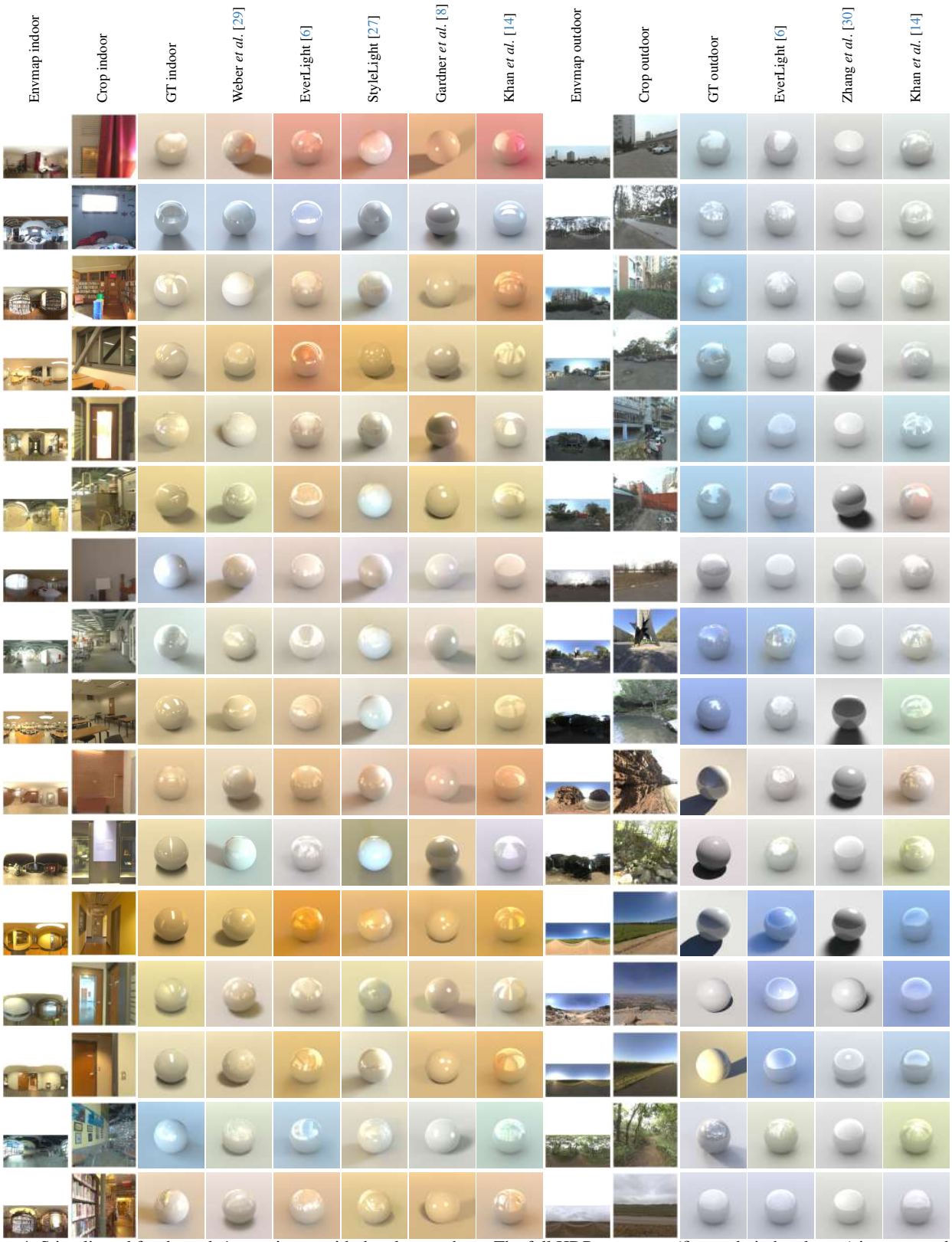

Figure 4. Stimuli used for the task 1 experiment with the glossy sphere. The full HDR panorama (first and ninth columns) is reexposed and tonemapped with $\gamma = 2.4$ for display, for the indoor and outdoor cases. The region extracted from the scene (second and tenth columns), corresponding to a $50°/90°$ FoV, taken from the centre of the full indoor/outdoor panorama (for most scenes). The rendered stimuli using the ground truth IBLs (first/ninth columns) are shown in the third/eleventh columns. The other columns are the rendered stimuli using the IBLs (shown in fig. 1) produced by the different lighting estimation methods.

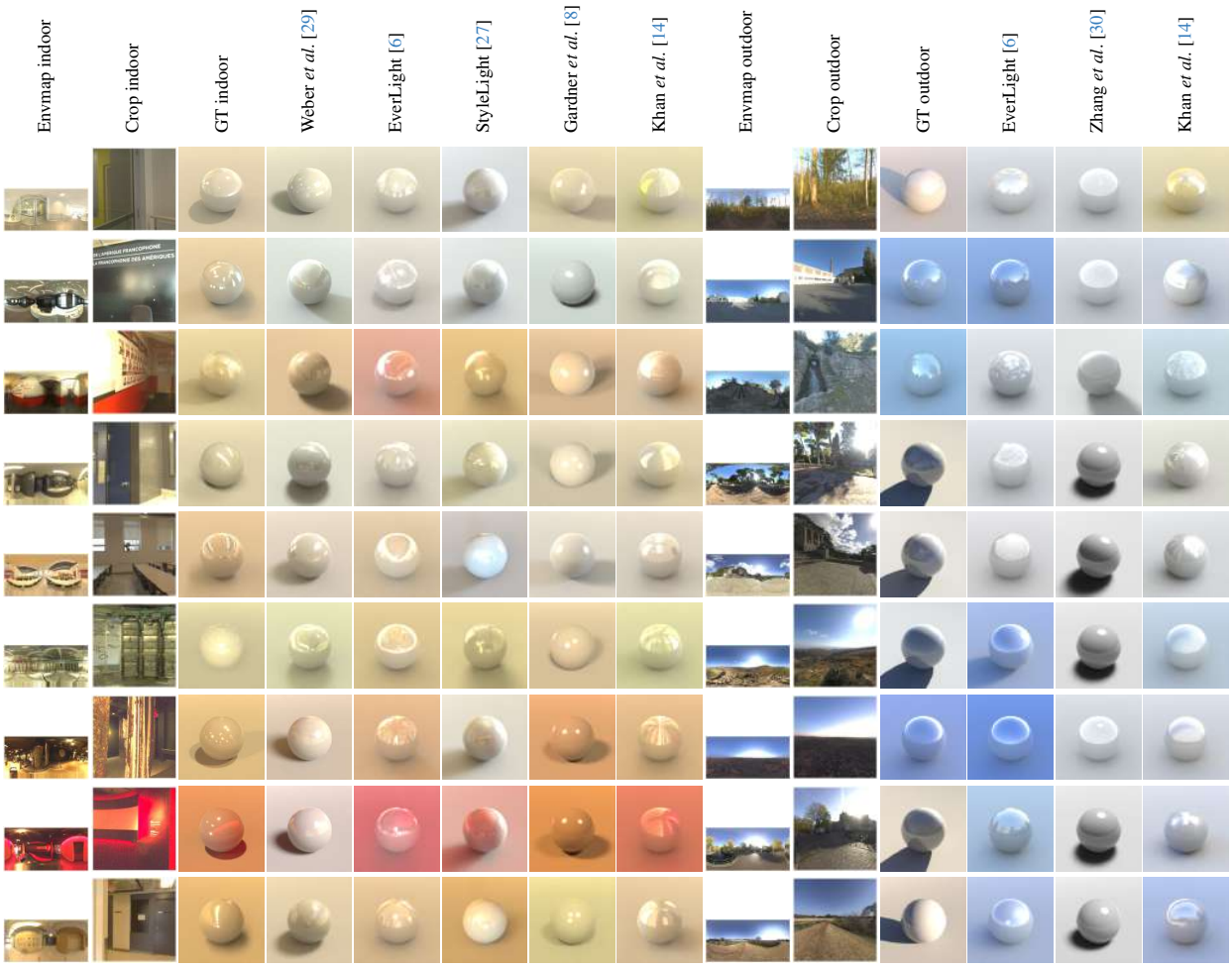

Figure 4. (contd) Stimuli used for the task 1 experiment with the glossy sphere. The full HDR panorama (first and ninth columns) is reexposed and tonemapped with $\gamma = 2.4$ for display, for the indoor and outdoor cases. The region extracted from the scene (second and tenth columns), corresponding to a $50^\circ/90^\circ$ FoV, taken from the centre of the full indoor/outdoor panorama (for most scenes). The rendered stimuli using the ground truth IBLs (first/ninth columns) are shown in the third/eleventh columns. The other columns are the rendered stimuli using the IBLs (shown in fig. 1) produced by the different lighting estimation methods.

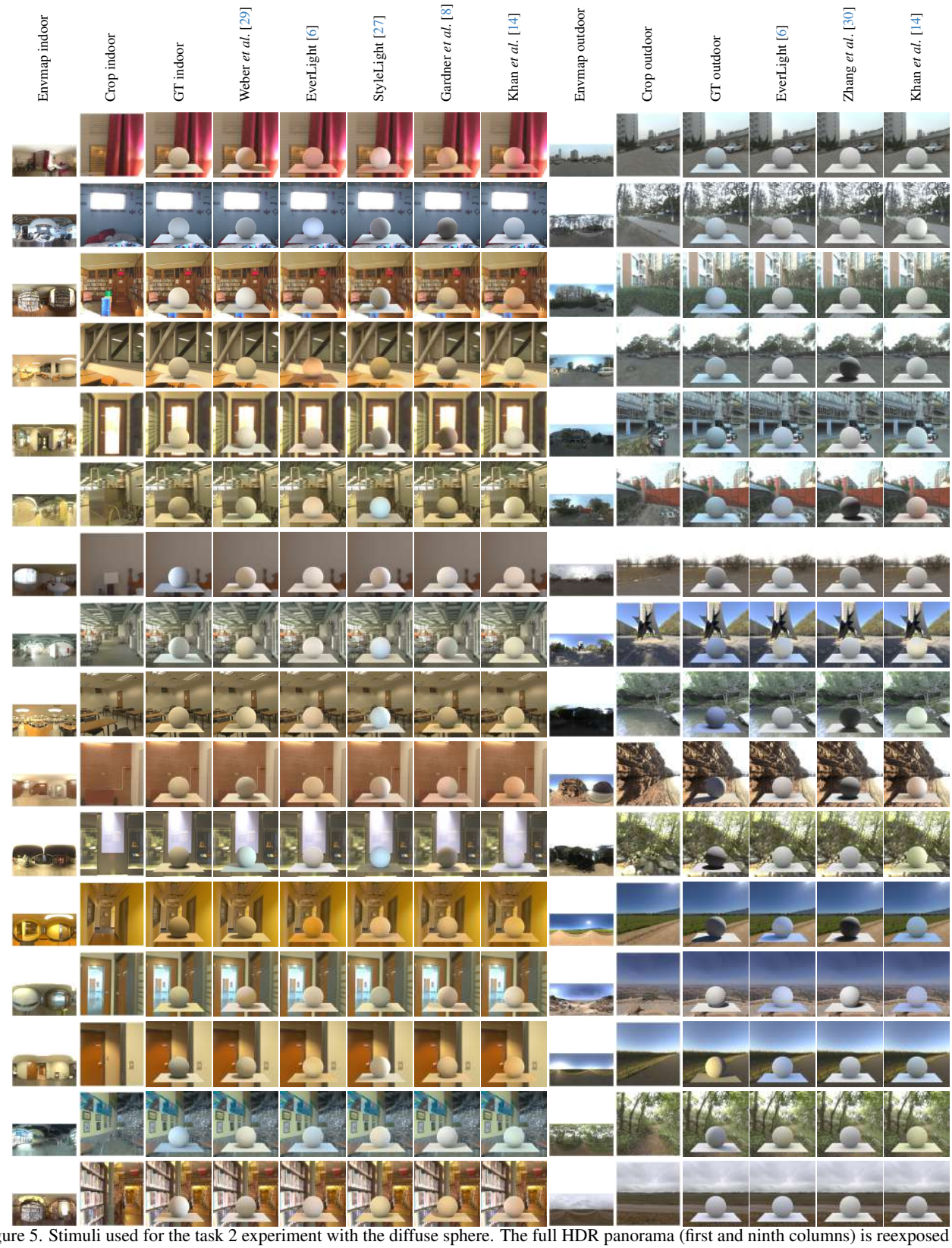

Figure 5. Stimuli used for the task 2 experiment with the diffuse sphere. The full HDR panorama (first and ninth columns) is reexposed and tonemapped with $\gamma = 2.4$ for display, for the indoor and outdoor cases. The region extracted from the scene (second and tenth columns), corresponding to a $50°/90°$ FoV, taken from the centre of the full indoor/outdoor panorama (for most scenes). The rendered stimuli using the ground truth IBLs (first/ninth columns) are shown in the third/eleventh columns. The other columns are the rendered stimuli using the IBLs (shown in fig. 1) produced by the different lighting estimation methods.

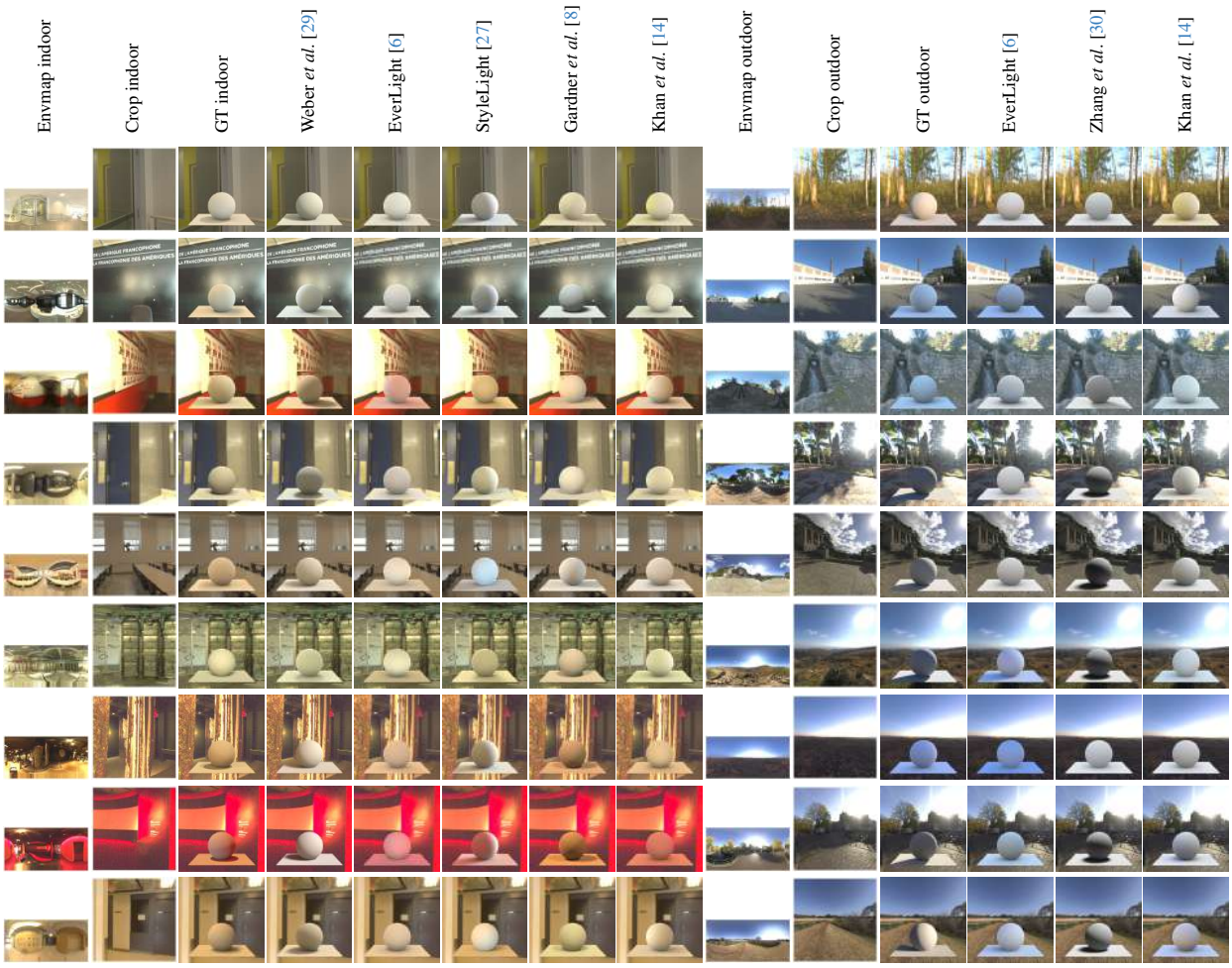

Figure 5. (contd) Stimuli used for the task 2 experiment with the diffuse sphere. The full HDR panorama (first and ninth columns) is reexposed and tonemapped with $\gamma = 2.4$ for display, for the indoor and outdoor cases. The region extracted from the scene (second and tenth columns), corresponding to a $50°/90°$ FoV, taken from the centre of the full indoor/outdoor panorama (for most scenes). The rendered stimuli using the ground truth IBLs (first/ninth columns) are shown in the third/eleventh columns. The other columns are the rendered stimuli using the IBLs (shown in fig. 1) produced by the different lighting estimation methods.

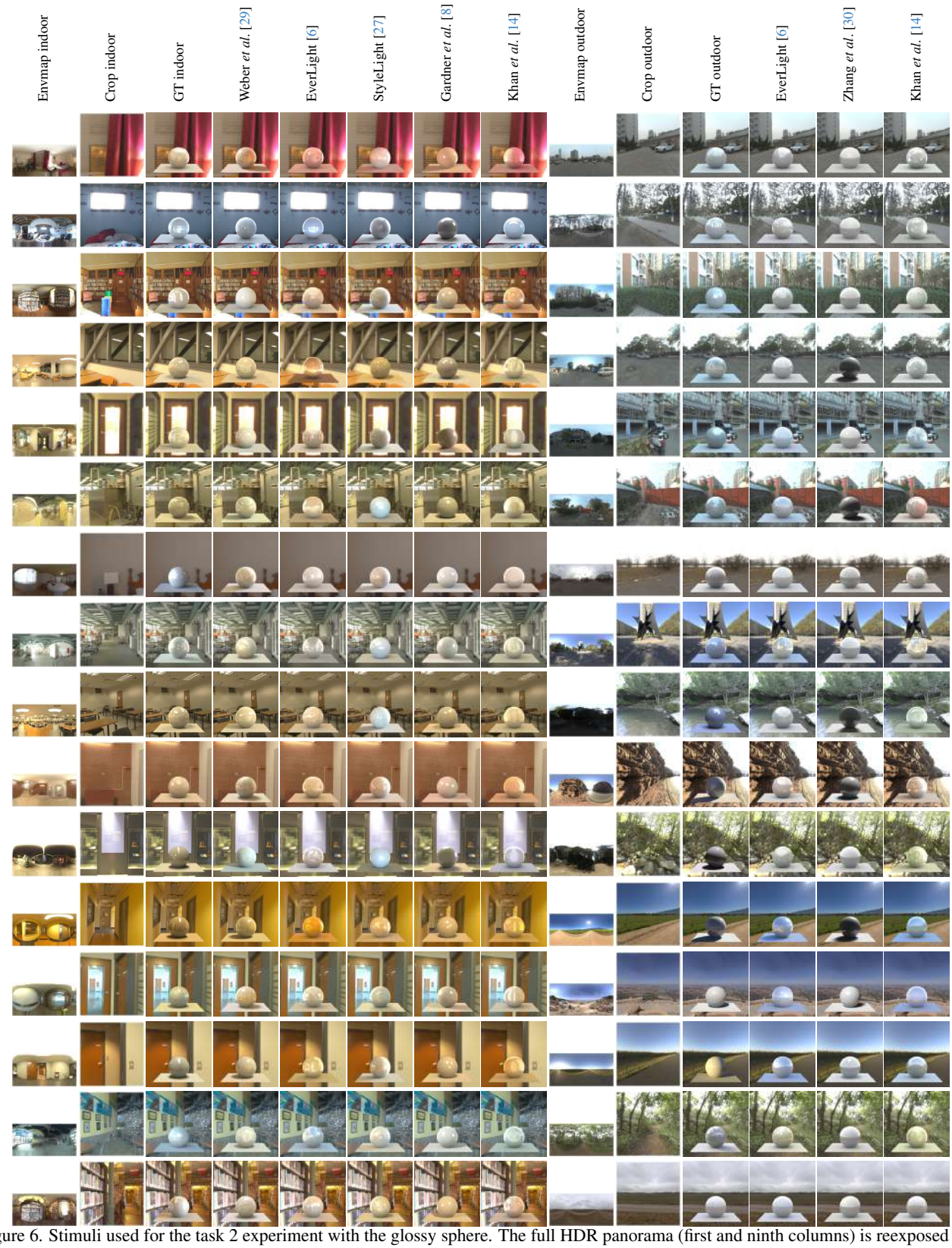

Figure 6. Stimuli used for the task 2 experiment with the glossy sphere. The full HDR panorama (first and ninth columns) is reexposed and tonemapped with $\gamma = 2.4$ for display, for the indoor and outdoor cases. The region extracted from the scene (second and tenth columns), corresponding to a $50°/90°$ FoV, taken from the centre of the full indoor/outdoor panorama (for most scenes). The rendered stimuli using the ground truth IBLs (first/ninth columns) are shown in the third/eleventh columns. The other columns are the rendered stimuli using the IBLs (shown in fig. 1) produced by the different lighting estimation methods.

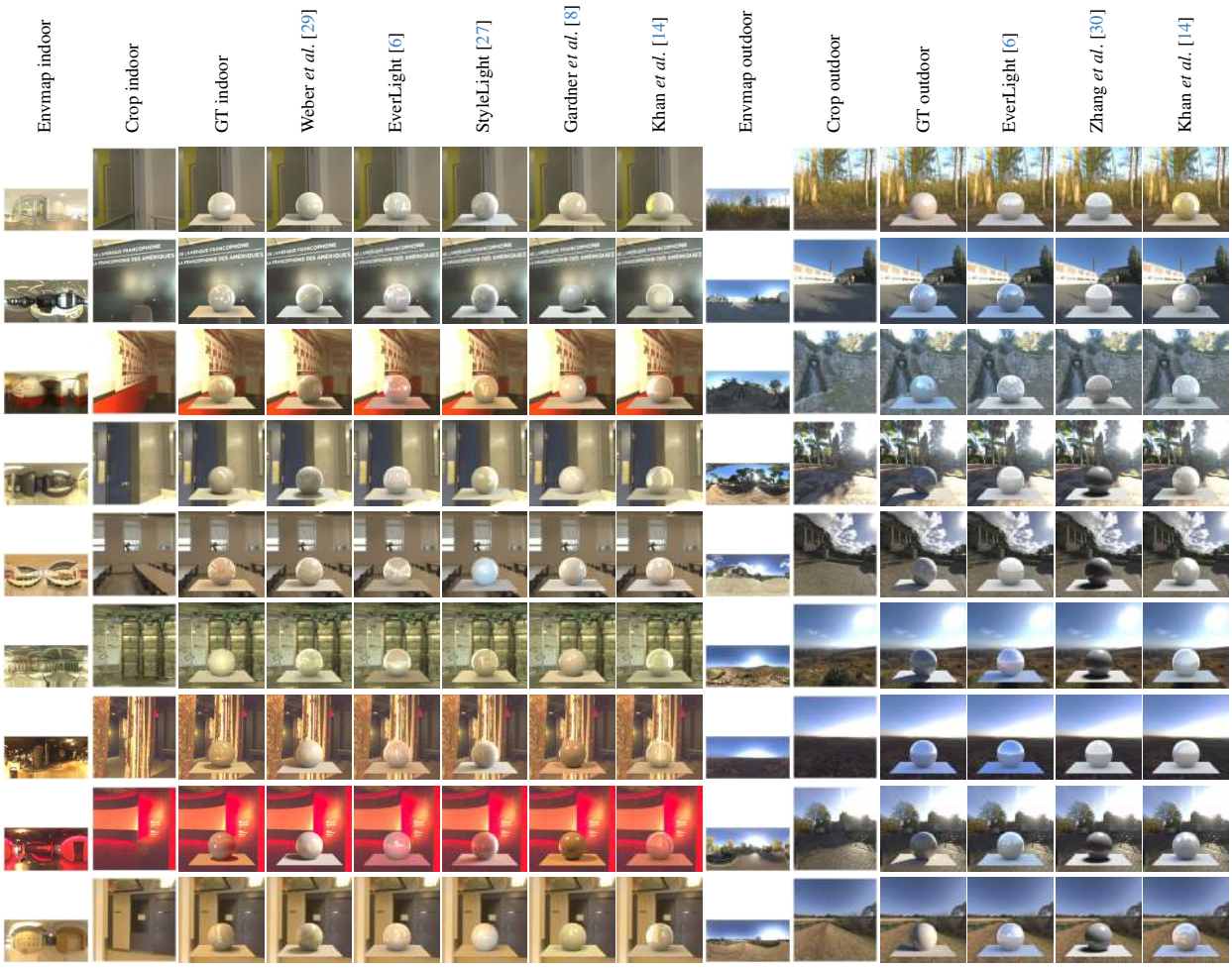

Figure 6. (contd) Stimuli used for the task 2 experiment with the glossy sphere. The full HDR panorama (first and ninth columns) is reexposed and tonemapped with $\gamma = 2.4$ for display, for the indoor and outdoor cases. The region extracted from the scene (second and tenth columns), corresponding to a $50°/90°$ FoV, taken from the centre of the full indoor/outdoor panorama (for most scenes). The rendered stimuli using the ground truth IBLs (first/ninth columns) are shown in the third/eleventh columns. The other columns are the rendered stimuli using the IBLs (shown in fig. 1) produced by the different lighting estimation methods.

Geometry. For both tasks, the stimulus corresponds to a sphere, with a radius of $1.5\,\mathrm{m}$, on a plane, to act as a shadow catcher. In task 1, the plane is $30\,\mathrm{m} \times 30\,\mathrm{m}$ to cover the entire FoV and in task 2, the plane is $2.5\,\mathrm{m} \times 3\,\mathrm{m}$ to allow the composited background to be seen. The background image used in task 2 corresponds to the extracted region from scenes given as input to the indoor lighting estimation methods (sec. 1.2), to give the virtual objects context.

A virtual camera is positioned parallel to the $x - y$ plane, facing the $+x$ axis, thus capturing in its FoV the $y - z$ plane, where y points towards the left and z upwards. The virtual camera is raised by $1.6\,\mathrm{m}$ from the perpendicular axis of the horizontal plane (z) with regard to the origin, to simulate the standard height of humans. In task 1, the FoV is inclined by $30°$ with regards to the horizon, to include the shadows produced by the sphere on the plane and for task 2, the FoV is inclined by $90°$, to align realistically with the composited background image.

Materials. For both tasks, two separate experiments are done on spheres with two different materials (diffuse and glossy), which use the Disney Principled BRDF [2]. The diffuse material has a roughness of 1.0 and a specularity of 0.0, whilst the glossy material has a roughness of 0.1 and a specularity of 1.0. The Lambertian sphere allows the observers to evaluate the lower frequency light cues, such as the colour, intensity, and degree of collimation. The opaque glossy sphere includes the high frequency light cues and the texture to be judged by the observer. For both versions, the plane maintains a grey Lambertian material with the same parameters as the diffuse sphere. All the objects have an albedo of 0.18.

Rendering. The synthetic objects are rendered using the physically based rendering engine `Cycles` in `Blender` [4]. The rendered stimuli have a resolution of $256\,\mathrm{px} \times 256\,\mathrm{px}$, as the extracted regions from scenes given as input to the indoor lighting estimation methods (sec. 1.2). The renders are saved in the `exr` format and then tonemapped with $\gamma = 2.4$ and reexposed to be displayed on the monitor used during the experiment (sec. 1.4).

1.4. Hardware

The experiment (see sec 3.3. of the main paper) is conducted in a controlled lab setting to ensure the data collected is uniform. The experiment is carried out in a matte black room (painted walls and ceilings, with black rug flooring) with a standard keyboard placed on a desk and the monitor set to sRGB. The monitor was the only light source. The observers are seated at $\sim 70\,\mathrm{cm}$ from the monitor, which gives a $11.5°/17°$ visual angle for task 1/2. The experimental setup is shown in 7.

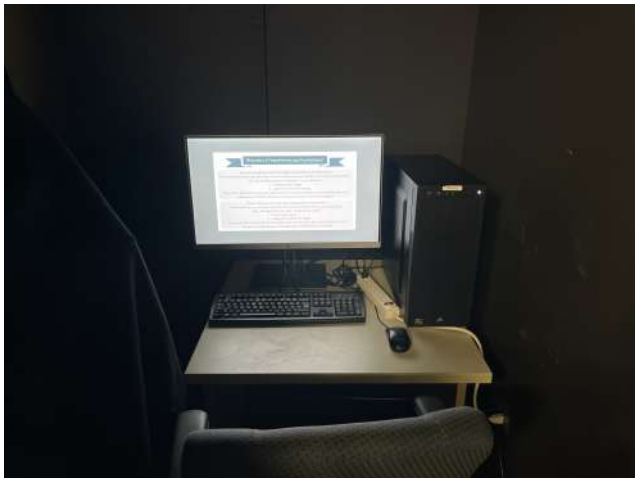

Figure 7. Photograph of the experimental setup of the psychophysical experiment.

The experiment runs on `MATLAB` (version R2023a) and uses the `Psychophysics Toolbox`. An example of the screen displayed to the observers is shown in 8 for all four experiments.

1.5. Procedure

During the experiment (see sec 3.3. of the main paper), the images are selected using the arrows on the keyboard. The background is middle grey.

Observers are asked to participate in two of the four experiments (task 1 and task 2), with randomly assigned material for the first task and the opposite material for the second task, to avoid potentially causing bias. A break is offered between the experiments to avoid fatigue. Each experiment takes 10–35 min to complete. No time restriction is imposed on the observers to avoid inducing stress and bias. The observers are advised to follow their intuition to determine their preference and that each combination of stimuli shown should be analysed in around than $5\,\mathrm{s}$, so the experiment would not last too long. This is done to avoid the fatigue or boredom they experience when doing the task for too long.

At the beginning of each experiment, a short tutorial is shown to the observer with an example not included in the dataset. The observers are informed that the images always contain the same sphere (same geometry) made of the same material, for all the stimuli they see during that specific experiment, and that only the lighting has changed. They are also informed that there is no right answer, and that we are only trying to measure their preferences. The participants are unaware that different lighting estimation methods have been used to produce the stimuli. To confirm that the observers are not colourblind, an Ishihara test is conducted for each participant before starting the experiments.

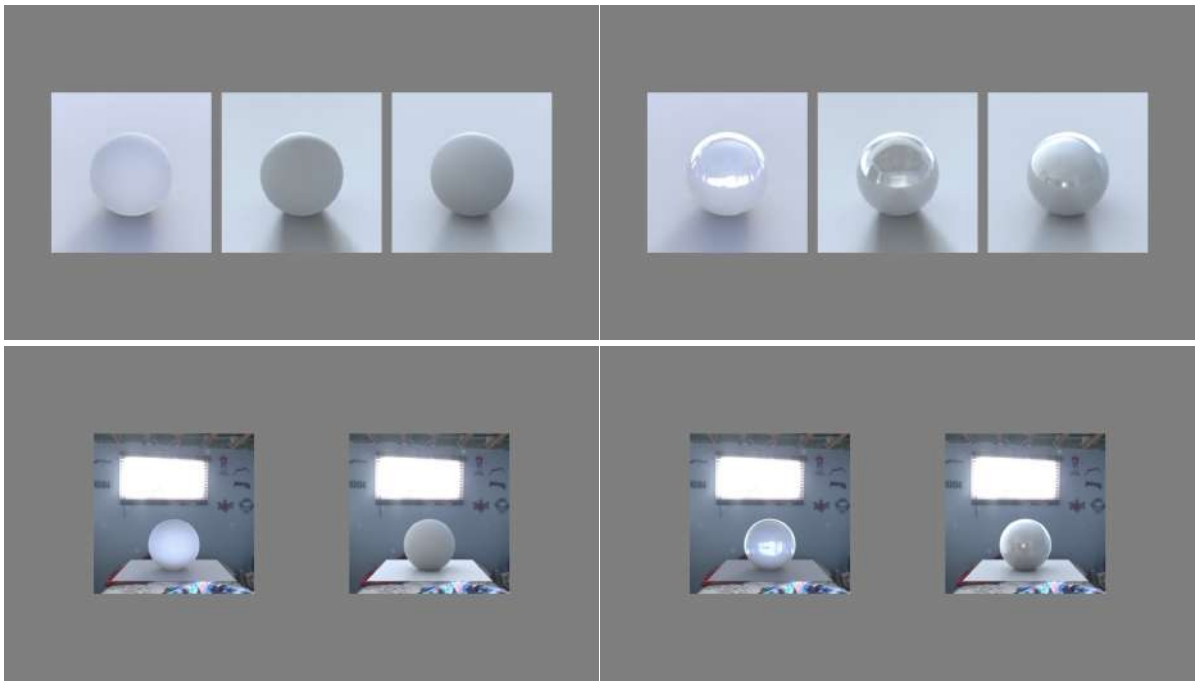

Figure 8. Examples of the stimuli presented during the experiments. Top row are the stimuli fr the comparison experiments for the diffuse (left) and glossy (right) spheres and the bottom row are the stimuli for the realism experiments.

1.6. Participants

A total of 49 unique observers (33M/16F) participated in the study (see sec 3.3. of the main paper). Some observers were asked to participate twice (in different sessions) and were assigned the two remaining experiments they had not previously done. 12 observers participated in all four outdoor experiments.

Fig. 9 shows the temporal evolution of the score based on the number of participants, to validate the convergence of the preferred methods by the observers. The curves show little variance as more participants are added after $\sim$15 for the indoor experiment (left) and $\sim$8 for the outdoor experiment (right), which confirms the number of participants included in the study is sufficient to describe a general tendency.

None of the participants were authors. 11 observers were students from the computer vision department (labelled *experts*), who were unaware of the project.

The scores obtained for the indoor lighting estimation methods for the expert (yellow) and naive (teal) observers are shown in fig. 10 (the procedure for computing score is described in sec. 4.1. of the main paper). It is possible to see that the scores for all the experiments are similar, which confirms that both samples of observers have the same trends and do not use different light cues in the stimuli.

The agreement score (described in sec. 5.1. of the main paper) between the expert and naive observers and the metrics, shown in fig. 11, also displays the same trends between each group, which further confirms that their behaviour is similar.

2. Psychophysical results

Additional analysis of the psychophysical results (see sec 4.2. of the main paper) is done in this section. Further statistical testing is done on the data in sec. 2.1, an example of the observers' ranking of a set of stimuli for an input scene is shown in sec. 2.2, and the trends of the preferred methods per image is shown in sec. 2.3. The agreement score of the individual observers is discussed in sec. 2.4

2.1. Agreement and consistency of the psychophysical results

In addition to the Thurstone Case V Law of Comparative Judgement z-score [26], computed in sec. 4.1. of the main paper, we demonstrate the agreement and consistency of the psychophysical data by computing Fleiss' κ and Kuder-Richardson-20 (KR20), respectively. The outdoor results are likely to be less statistically robust, as there are $60\,\%$ fewer participants than for the indoor experiments. However, the results in tab. 1 show that the results for the outdoor experiment exhibit agreement and consistency. Noting as (task 1/task 2), the scores are higher for glossy ($\kappa = 0.437/0.536$, KR20$= 0.907/0.936$) than diffuse ($\kappa = 0.345/0.141$, KR20=$0.868/0.675$). KR20 is higher than 0.85 in all but task 2 diffuse (where it is still

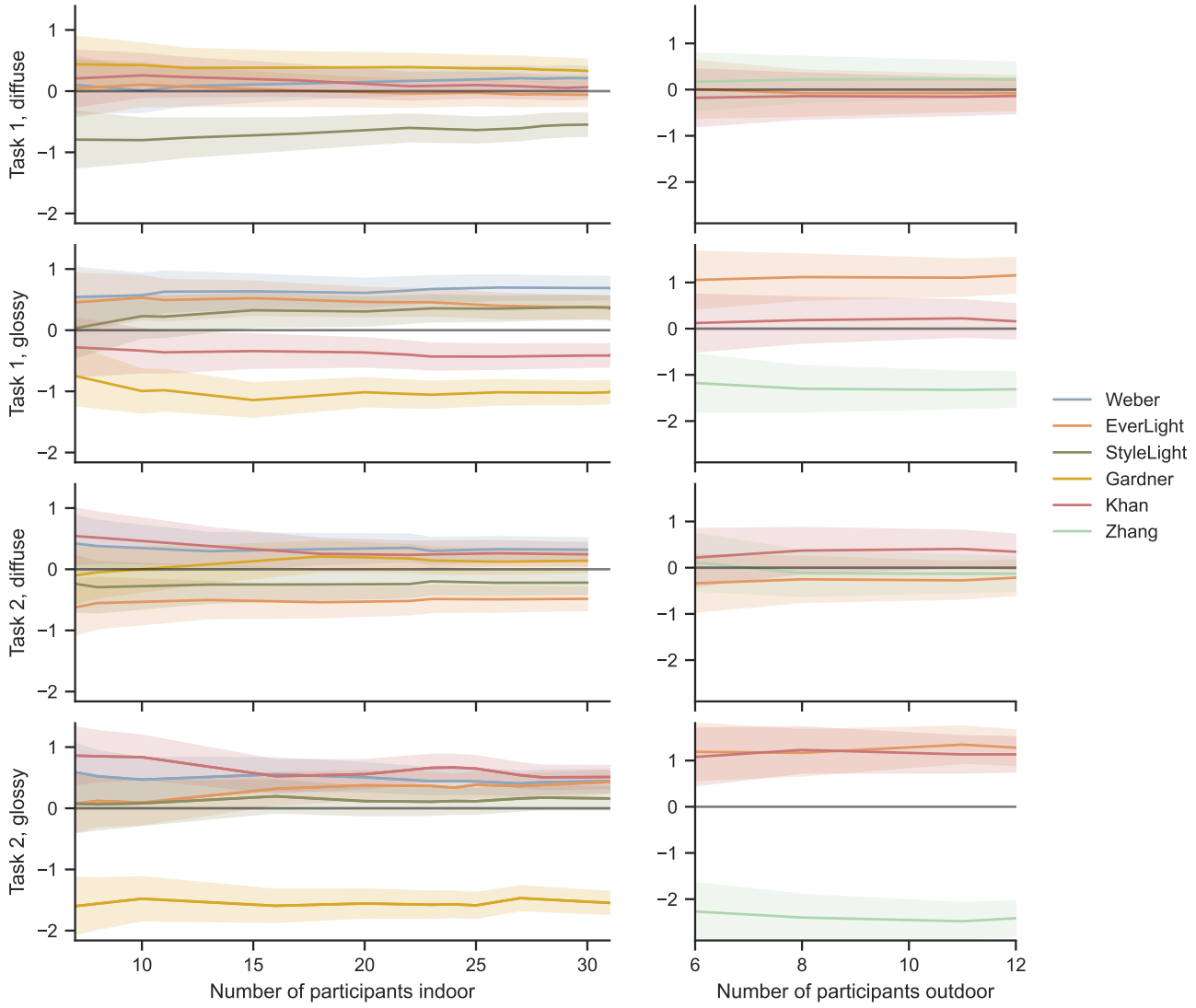

Figure 9. Convergence of the score for the different lighting estimation methods as a function of the number of participants. The line uncertainties correspond to 95 % confidence interval.

at a remarkable 0.67), showing internal consistency. Fleiss' κ values indicate moderate agreement for glossy [16], fair agreement for task 1 diffuse and slight agreement for task 2 diffuse.

2.2. Example of stimuli ranking

An example of each indoor lighting estimation method's ranking (in decreasing order) and the associated score for each stimulus for the same input scene is shown in fig. 12, for each experiment. When comparing against the ground truth stimulus (task 1) for the diffuse sphere (first row), observers seem to agree—at least in essence—to what IQA metrics are trying to achieve: having an image as close as possible to the ground truth reference. E.g. the lighting estimation of Gardner *et al.* [8] does not accurately match the ground truth,

Table 1. Fleiss' κ and Kuder-Richardson 20 scores computed on all the observers' results for each experiment.

	Task	Material	Fleiss' κ	Kuder-Richardson 20
Indoor	Task 1	Diffuse	0.210	0.890
		Glossy	0.255	0.915
	Task 2	Diffuse	0.149	0.842
		Glossy	0.269	0.922
Outdoor	Task 1	Diffuse	0.345	0.868
		Glossy	0.437	0.907
	Task 2	Diffuse	0.141	0.675
		Glossy	0.536	0.936

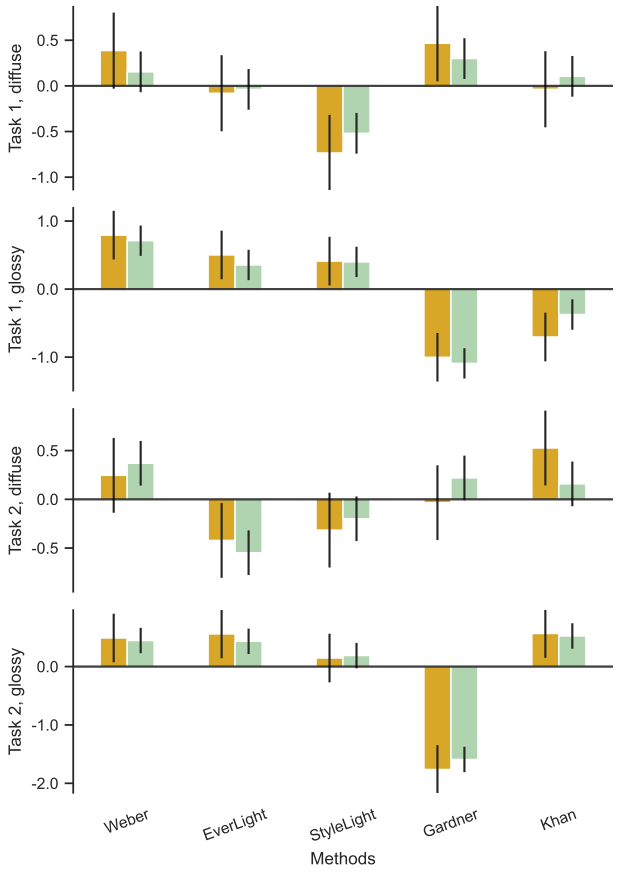

Figure 10. Thurstone Case V Law of Comparative Judgement scores from the expert (yellow) and naive (teal) observers as a function of the different indoor lighting estimation methods (columns), for the different types of sphere materials used in the experiments (rows). Error bars correspond to 95 % confidence interval.

while the one produced by Weber *et al*. [29] resembles the ground truth very closely. Yet, when the resulting lighting estimations are put into context (task 2, third row), lighting accuracy based on the ground truth does not seem to matter as much to be considered plausible. In this case, the preferred estimated lighting does not seem to match well the ground truth.

Fig. 12 also illustrates an example of the trend observed in fig. 4 of the main paper for the glossy experiments produced by Gardner *et al*. [8]. The observers seem to consider texture primordial when observing a more reflective surface. They seemingly consider it more important than having a plausible lighting intensity and direction, when judging the plausibility of an inserted object.

2.3. Scores per image

The scores from all the observers for each stimulus are shown in fig. 13. This figure clearly shows that glossy stimuli from Gardner *et al*. [8] (indoors) and Zhang *et al*. [30] (outdoor) are, in general, rarely picked. For the other lighting estimation methods, we can see a greater variance in the scores, with some images performing very well or very poorly. This indicates some of the limitations of a given method in some specific case.

2.4. Individual observer agreement

The individual observer agreement scores $\omega^{(i)}$ are shown in fig. 14, for the indoor (left) and outdoor (right) lighting estimation methods, for all experiments (rows). The observers are anonymised by assigning them a random number and a random order between each experiment (i.e. the observer labelled "1" in the first row is not necessarily the same observer "1" in the second row), as not all the observers participated in the same experiments. The observers labelled as "P" and "R" correspond to the perfect and random observers (defined in sec. 5.1. of the main paper), respectively. This score is how the observers removed from the study are determined (see sec. 5.1. of the main paper for more details). The orange/blue lines are determined by taking the average of the individual observer agreement scores $\omega^{(i)}$ (excluding the perfect and random observers), to obtain the expected observer agreement score (defined in sec. 5.1. of the main paper).

3. Measuring the scores of the IQA metrics

Additional information regarding the scores of the IQA metrics (see sec 5. of the main paper) is provided in this section. In this subsection we aim analyse the Thurstone Case V Law of Comparative Judgement z-scores of the individual metrics, using the same approach as the one used in sec. 4.1. of the main paper, in sec. 3.1. We also compute the correlation score between the IQA metrics and the observers in sec. 3.2 using Spearman's ρ and Kendall's τ statistical tests, to confirm the results obtained by the agreement score, computed in sec. 5.1. of the main paper. We will consider the same metrics as in the main paper RGB angular error [10], PSNR [12], RMSE, si-RMSE and more recent ones, such as SSIM [28], VIF [24], LPIPS [31], PieAPP [21], FLIP [1], HDR-VDP3 [17], BRISQUE [19], NIQE [20], UNIQUE [32], and HyperIQA [25]. The FID [11] metric is also studied in sec. 3.3.

3.1. Scores of the IQA metrics

Broadly speaking, our idea is to consider the metrics as if they were observers in our pair comparison test. In order to do so, we will perform all the same pair-wise comparisons of our experiment, and for each comparison, we will assign a 1 to the image that the metric picks as better and a 0 to the other image. Once we have all the selections, we can compute the scores for each of the metrics. More in detail, the score

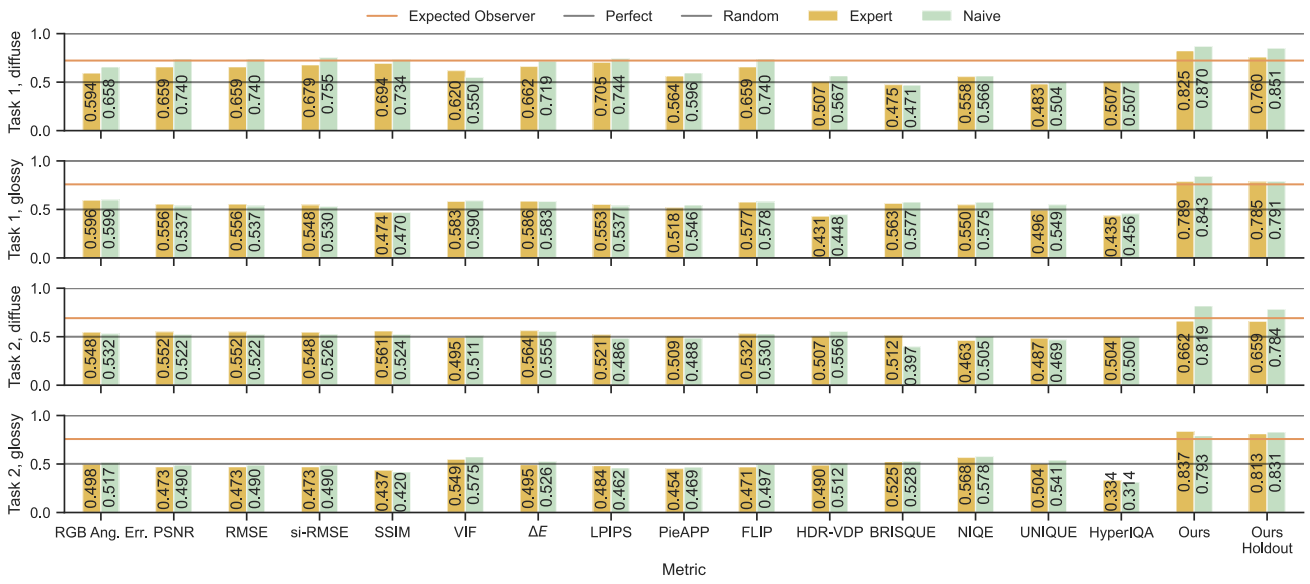

Figure 11. Agreement between the observer scores (expert: yellow left bars; naive: teal right bars) and the metric scores (columns) for all the indoor lighting estimation methods, for the different types of experiments (rows). The lower horizontal grey bar is set at chance level ($\sim$0.5) and the higher one corresponds to the perfect observer (set at 1.0). The orange line corresponds to the expected observer agreement score for all the observers for the indoor methods (same as the one in fig. 5 of the main paper).

is then computed for each metric, for each experiment, the same way as described in sec. 4.1. of the main paper.

The Thurstone Case V Law of Comparative Judgement scores obtained for each metric computed on the stimuli from the indoor (left) and outdoor (right) lighting estimation methods are shown in fig. 15, for all the experiments. This figure clearly highlights that the different metrics do not agree with each other for a given set of images. This fact reinforces our main arguments in two different ways. Firstly, it shows that current metrics are not feasible enough as by selecting a specific metric the ranking of the methods is also modified. Secondly, it also proves that our approach of considering an ensemble of different metrics to derive our new metric is the avenue to pursue.

3.2. Correlation scores between the IQA metrics and the observers

To study more in-depth the variations of the correlation between each experiment, the mean Spearman's ρ and Kendall's τ scores are computed over all metrics for the indoor experiments and shown in tab. 2. As it is possible to observe, the correlation scores are indeed lower for the task 2 experiments compared to task 1. The considerably higher correlation score for the task 1 diffuse experiment also confirms the trend observed with the agreement score (fig. 5 of the main paper), with a few metrics agreeing with human perception.

Table 2. Average Spearman's ρ and Kendall's τ correlation scores for all the metrics, including "Ours", but excluding "Ours Holdout" for the Spearman's ρ and Kendall's τ tests, for all the indoor experiments.

	Spearman's ρ		Kendall's τ	
	Task 1	Task 2	Task 1	Task 2
Diffuse	0.305	0.057	0.253	0.047
Glossy	0.134	0.103	0.111	0.085

3.3. Agreement scores between FID and the observers

The FID metric [11]—commonly used to evaluate generative models—is not included in our analysis. As a Reduced-Reference metric, it is computed on a distribution of images and not on individual images, contrary to the rest of the metrics considered in this study. Hence, obtaining a value for each pair of images is impossible. It is impossible to get the judgement of human perception for a whole distribution of images due to the prohibitively time-consuming experimental setup used in our study. Despite this, we provide FID results as an extra analysis following the procedure described in [6]. FID also contrasts with the other metrics in our study as it is computed directly on IBLs instead of renders. This is due to the nature of FID, which starts by feeding the images to a neural network trained on natural images. We chose to run it on IBLs to minimise the domain gap of this neural

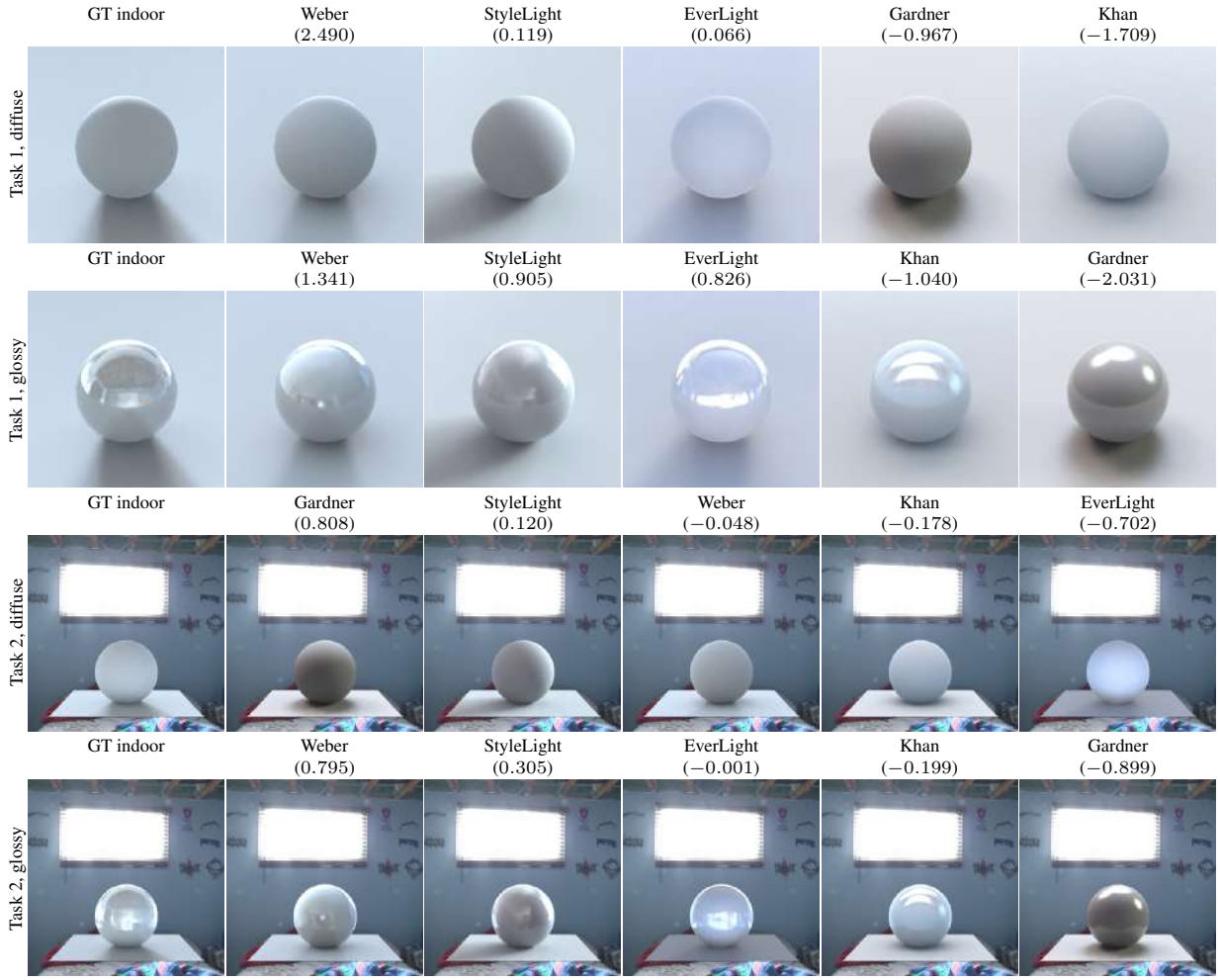

Figure 12. Example of the preferences of all the observers of the stimuli generated by the different lighting estimation methods (columns) ordered decreasingly as a function of their score, for the different types of experiments (rows). The first column is the ground truth (not judged by the observers) associated to the scene for comparison.

network.

To properly compare the observers' Thurstone Case V Law of Comparative Judgement scores with the FID scores, both scores are normalised, so they have similar scales. The human scores range from negative to positive values, unlike FID, so both scores (x') are mapped between -1 and 1, using

$$x' = \frac{2(x - \min\{x\})}{\max\{x\} - \min\{x\}} - 1\,, \tag{1}$$

where x corresponds to the scores. Unlike the human score, a low FID score indicates a better quality image; thus, the normalised FID scored is reversed to match the human score. This means that a method with a value of 1 is considered to be the best performing method and a value of -1 is considered the worst performing method. The normalised scores are presented in fig. 16.

It is important to keep in mind that the scores compared in fig. 16 are obtained from different stimuli; the scores of the human observers were obtained on the rendered stimuli, while the FID is computed directly on the IBLs. This also implies that the normalised FID scores are the same for all experiments (rows) in fig. 16. Also, as previously mentioned, the human scores are obtained for each individual images, and do not represent a global statistic for the method itself, as is the case for the FID. It is thus harder to robustly compare the normalised scores from both sources, but it is still possible to observe general trends. When looking at the similarity of the distributions for the human score (green) and normalised FID scores (pink), the experiments using glossy materials seem to match better.

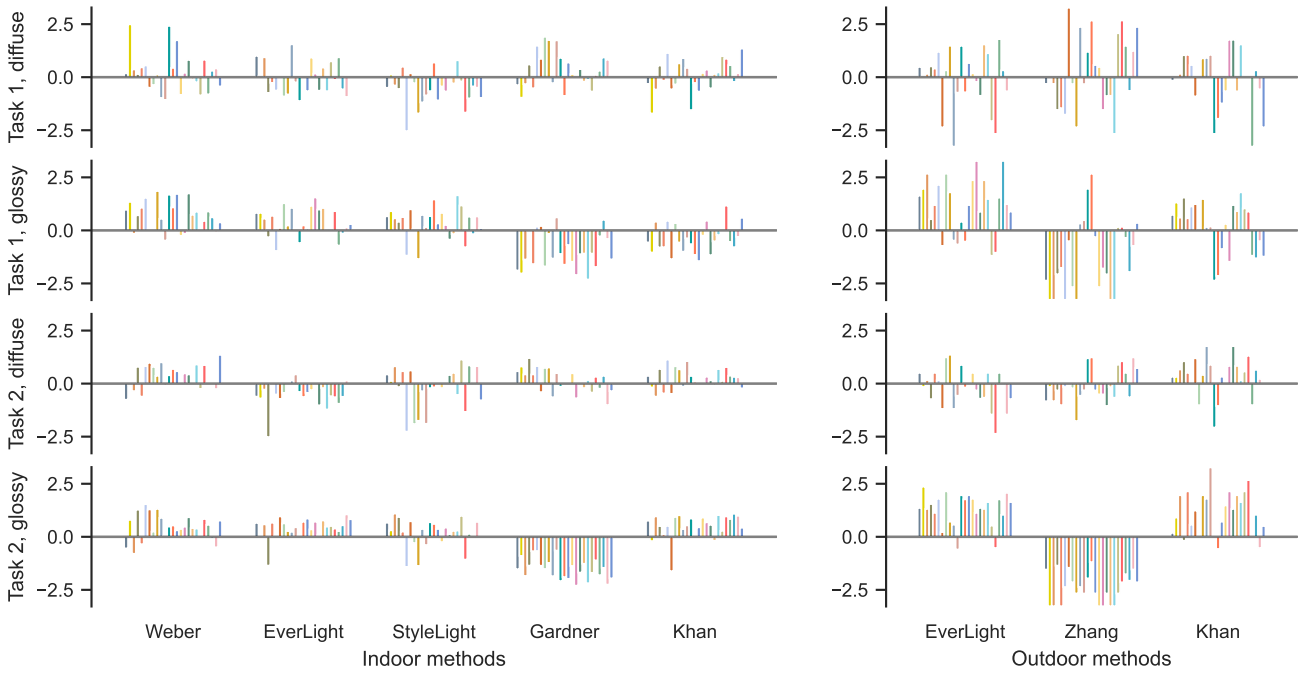

Figure 13. Thurstone Case V Law of Comparative Judgement scores from all the observers for each of the sets of images as a function of the different lighting estimation methods (columns), for the different types of sphere materials used in the experiments (rows).

4. Learning a metric combination

Additional details on the training of our proposed metrics (see sec 6.1. of the main paper) are given in sec. 4.1, and a supplementary analysis of the generalisation psychophysical study (see sec 6.2. of the main paper) is done in sec. 4.2.

4.1. Formulation and training

In tab. 3, the mean and standard deviation of the training accuracies for the k-fold approach ($k = 10$) are displayed, for various classical learners. The parameters are all the default values. In some cases, *BernoulliNB* performs better than the SVR, however that methods outputs discrete values, and not a continuous ranking like the SVR, thus this method is preferred. Regarding the other methods outperforming the SVRs for specific experiments, the SVR is the method that performs the best across all experiments. Thus, the SVR is chosen for uniformity.

A representation of the learnt metric by the SVR is shown in fig. 17 for the indoor and outdoor validation cases (shown in "Ours" column of fig. 5 in the main paper) and for the holdout and new methods generalisation tests in fig. 18 (which corresponds to the "Ours Holdout" column of fig. 5 in the main paper and sec. 6.2. of the main paper, respectively). The projection of the $\mathbb{R}^{15}$ space of the classical metrics is projected to $\mathbb{R}^2$ to illustrate the decision boundaries (white lines) of the learnt metrics. The data points with the black contours indicate data points used in validation and the grey contours correspond to the support vectors.

4.2. Generalisation to other lighting estimation methods

All the environment maps generated by the generalisation lighting estimation methods for the user study are presented in fig. 19.

The stimuli used in the indoor and outdoor user study, for the diffuse/glossy experiments, are shown in fig. 20/fig. 21 for task 1 and fig. 22/fig. 23 for task 2.

The scores for all observers for the generalisation lighting estimation methods are shown in fig. 24. Only Weber *et al*. [29] is a method also included in the main psychophysical study; however, it has never been compared to the new methods. The methods are described in sec. 6.2. of the main paper.

Unlike Khan *et al*. [14], which includes texture but lacks a proper HDR lighting estimation, the Average method only includes the average pixel colour of the background image given as input (sec. 1.2). Thus, it includes no texture nor lighting estimation. The fact that for all experiments this method is not preferred demonstrates that excluding texture and lighting estimation cannot produce accurate (task 1) nor realistic (task 2) results. We hypothesize that this lack of texture causes the method to be disliked for the glossy experiments, as it has been empirically shown in the main

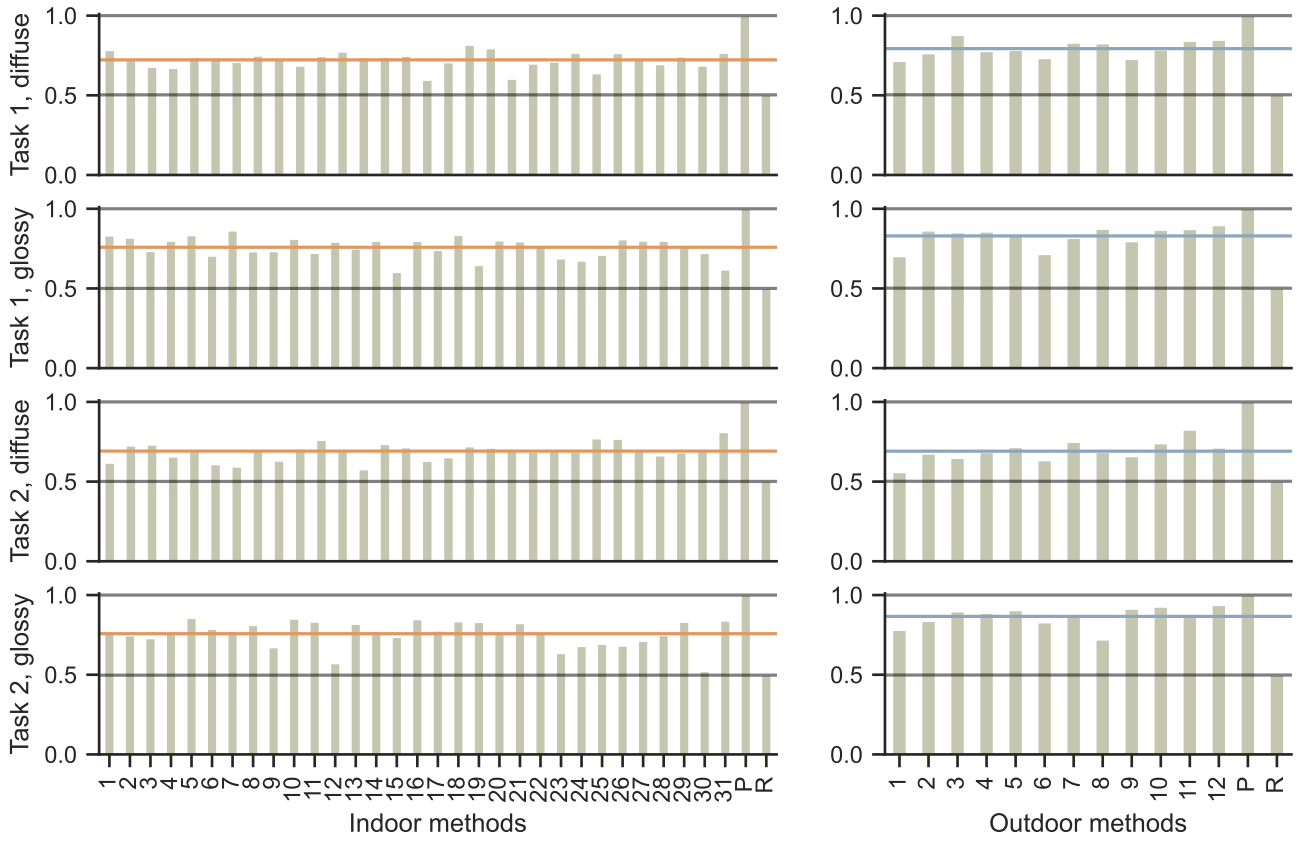

Figure 14. Agreement between the expected observer scores and the individual observer scores (columns) for all the lighting estimation methods (indoor: left bars; outdoor: right bars), for the different types of experiments (rows). The lower horizontal grey bar is set at chance level (labelled as "R"; $\sim$0.5) and the higher one corresponds to the perfect observer (labelled as "P"; set at 1.0). The orange (indoor) and blue (outdoor) lines corresponds to the expected observer agreement score.

study with the results from fig. 4 in the main paper for Gardner *et al*. [8] and Zhang *et al*. [30]. However, the naive approach to lighting estimation from the Average method is similar to Khan *et al*. [14], yet the Average method does not perform well in diffuse spheres, especially on task 2, like Khan *et al*. [14]. When comparing the stimuli from Khan *et al*. [14] (fig. 5) and the Average method (fig. 22), it is possible to see that there are low-frequency lighting variations on the renders done with Khan *et al*. [14] (e.g. spatial colour variations), which we believe contribute to the judgement of plausibility by the observers.

For the same reasons as Gardner *et al*. [8] and Zhang *et al*. [30] in fig. 4 of the main paper, Garon *et al*. [9] seems not to be preferred by observers as it lacks textures, adding empirical evidence that textures are an important part of lighting estimation, especially for the glossy experiments.

However, it is important to remember that this generalisation study has been done on a small number of participants and fewer images, yielding an increased statistical uncertainty. Nevertheless, fig. 9 shows that even with 6 observers, the general trends do not seem to change much.

References

[1] 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. 18

[2] Brent Burley. Physically-based shading at disney. 2012. 15

[3] 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. 5

[4] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 15

[5] M. Dastjerdi, Y. Hold-Geoffroy, J. Eisenmann, S. Khodadadeh, and J. Lalonde. Guided co-modulated GAN for 360° field of view extrapolation. In *Int. Conf. 3D Vis.*, 2022. 1

[6] Mohammad Reza Karimi Dastjerdi, Jonathan Eisenmann, Yannick Hold-Geoffroy, and Jean-François Lalonde. Ev-

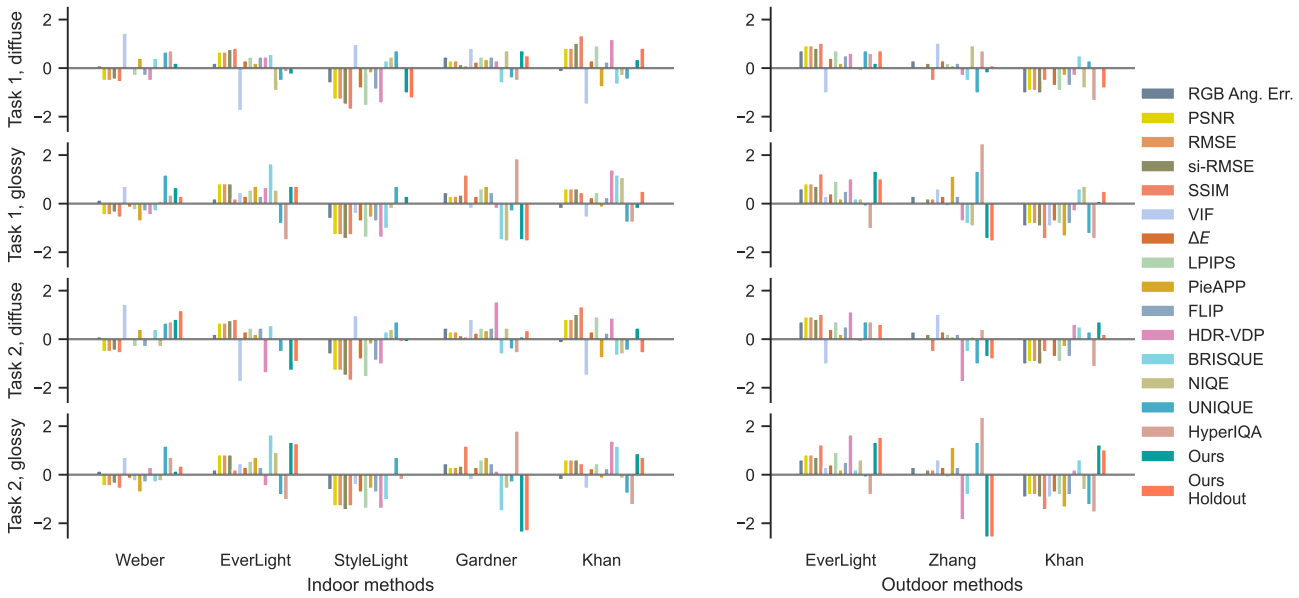

Figure 15. Thurstone Case V Law of Comparative Judgement scores for the different metrics as a function of the different lighting estimation methods (columns), for the different types of sphere materials used in the experiments (rows).

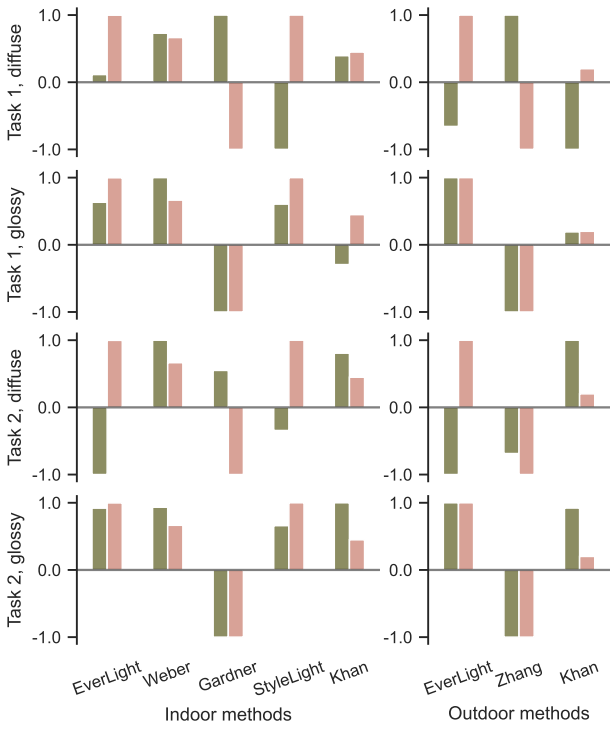

Figure 16. Normalised Thurstone Case V Law of Comparative Judgement scores (green) of the observers compared to the normalised FID score (pink), for all the lighting estimation methods (columns), for the different types of experiments (rows).

erlight: Indoor-outdoor editable HDR lighting estimation. In *Int. Conf. Comput. Vis.*, 2023. 1, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 19, 27, 28, 29, 30, 31

[7] 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. 5

[8] 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, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 17, 18, 22

[9] 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. 22, 27, 28, 29, 30, 31

[10] 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. 18

[11] 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. 18, 19

[12] Alain Horé and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *Int. Conf. Pattern Recog.*, 2010. 18

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1

[14] Erum Arif Khan, Erik Reinhard, Roland W. Fleming, and Heinrich H. Bülthoff. Image-based material editing. *ACM*

Type Scene	Material	Method	Mean Accuracy	STD Accuracy
Task 1	**Diffuse**	**Lasso**	72.20	6.40
		ElasticNet	72.20	6.40
		Ridge	72.10	6.30
		LogisticRegression	73.10	6.10
		Perceptron	66.70	8.30
		ARDRegression	72.40	6.70
		BayesianRidge	72.00	6.20
		BernoulliNB	74.30	6.20
		TheilSenRegressor	71.00	5.80
		BayesianGaussianMixture	52.60	1.40
		GaussianMixture	52.60	1.40
		SVR	71.40	6.50
	Glossy	**Lasso**	61.60	8.60
		ElasticNet	61.70	8.50
		Ridge	61.80	8.60
		LogisticRegression	61.30	6.90
		Perceptron	54.50	7.10
		ARDRegression	60.80	8.70
		BayesianRidge	62.40	8.60
		BernoulliNB	61.00	7.40
		TheilSenRegressor	61.10	9.30
		BayesianGaussianMixture	50.80	0.70
		GaussianMixture	50.80	0.70
		SVR	62.00	8.40
Task 2	**Diffuse**	**Lasso**	55.10	6.80
		ElasticNet	55.10	6.80
		Ridge	54.80	7.10
		LogisticRegression	55.20	6.10
		Perceptron	52.20	7.00
		ARDRegression	54.60	6.10
		BayesianRidge	54.60	7.00
		BernoulliNB	52.90	7.60
		TheilSenRegressor	57.00	6.50
		BayesianGaussianMixture	53.70	1.60
		GaussianMixture	53.70	1.60
		SVR	55.70	7.70
	Glossy	**Lasso**	65.70	6.10
		ElasticNet	65.60	6.00
		Ridge	65.20	6.30
		LogisticRegression	63.10	5.90
		Perceptron	58.40	8.20
		ARDRegression	65.30	6.40
		BayesianRidge	65.10	6.20
		BernoulliNB	57.90	5.40
		TheilSenRegressor	66.40	6.60
		BayesianGaussianMixture	50.40	0.50
		GaussianMixture	50.40	0.50
		SVR	63.20	6.80

Table 3. Results of the different classical models trained with each of the psychophysical experiments data. The mean and standard deviation accuracy results on the k-folds ($k = 10$) experiments are given of all the models for each experiments.

Trans. Graph., 25(3):654–663, 2006. 1, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 21, 22

[15] Jean-François Lalonde and Iain Matthews. Lighting estimation in outdoor image collections. In *Int. Conf. 3D Vis.*, 2014. 1

[16] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. 17

[17] 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. 18

[18] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, 2018. 6, 25, 26

[19] 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,

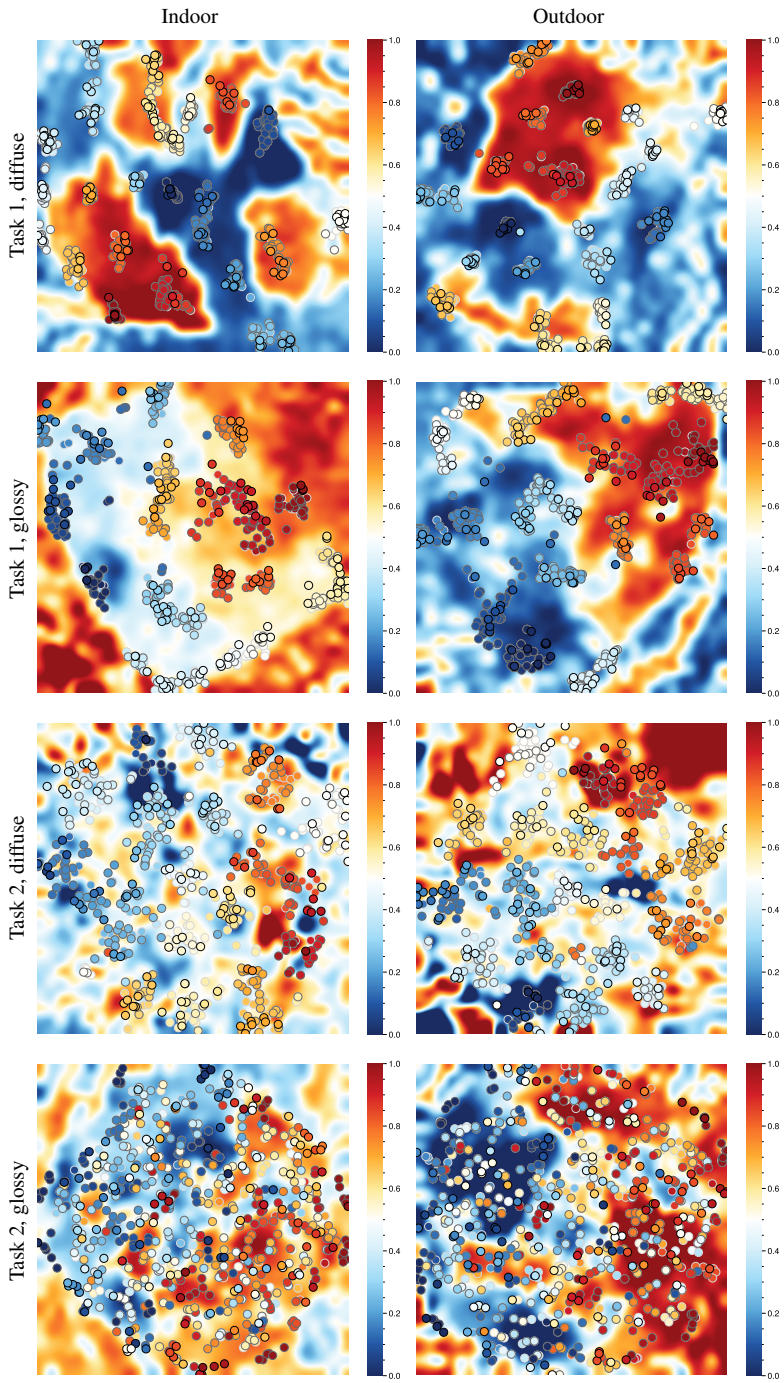

Figure 17. Visualisation of the learnt SVR metric for the four different experiments (rows), applied to an indoor/outdoor (columns) validation dataset. The projection of the fifteen classical metrics to $\mathbb{R}^2$ is done using UMAP [18]. The point corresponds to the pair comparisons of the stimuli (given as input to the network) and the background corresponds to the learnt function by the SVR. The white line corresponds to the boundary between the left and right choice of image. The colour of the data points and the background correspond to the choice of picking the left or right image given as input to the metric. The data points with the black contours indicate data points used in validation, and the grey contours correspond to the support vectors.

2012. 18

[20] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013. 18

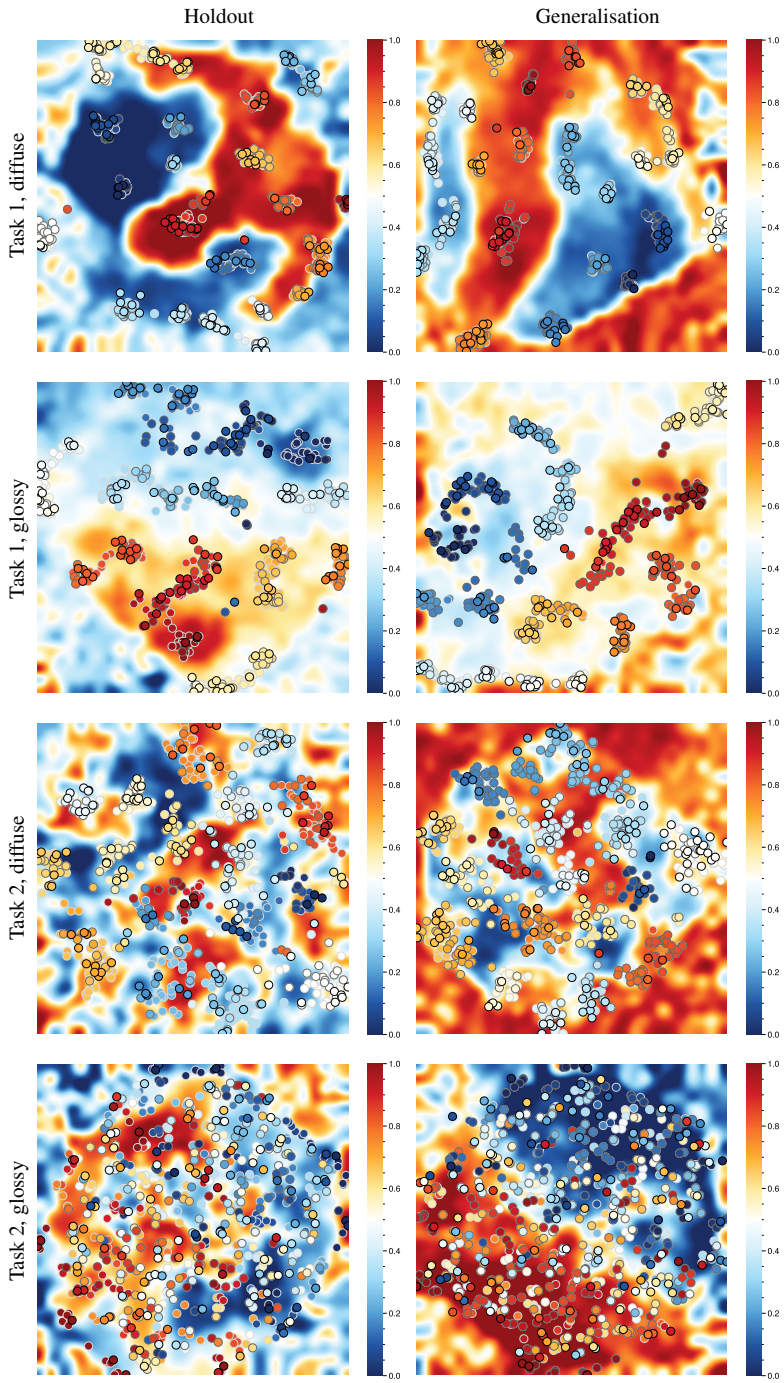

Figure 18. Visualisation of the learnt SVR metric for the four different experiments (rows), applied to the metric trained with the holdout approach (see sec. 5.1. of the main paper for details) in the left column and for the metric trained for the generalisation test (see sec. 5.2. of the main paper for details) in the right column. The projection of the fifteen classical metrics to $\mathbb{R}^2$ is done using UMAP [18]. The point correspond to the pair comparisons of the stimuli (given as input to the network) and the background corresponds to the learnt function by the SVR. The white line corresponds to the boundary between the left and right choice of image. The colour of the data points and the background correspond to choice of picking the left or right image given as input to the metric. The data points with the black contours indicate data points used in validation and the grey contours correspond to the support vectors.

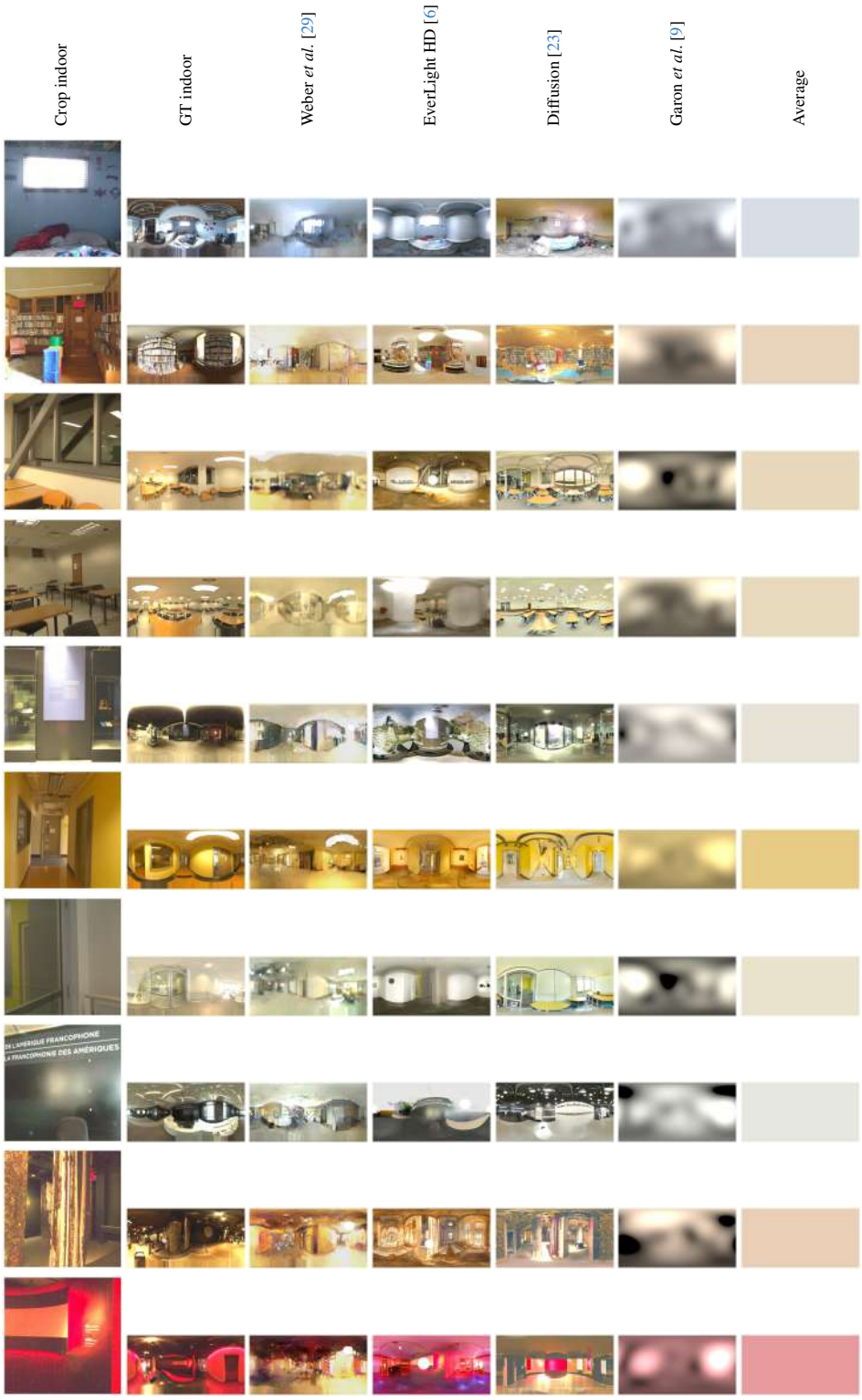

Figure 19. IBLs generated by the different generalisation lighting methods (columns) for each scene (rows). The first column corresponds to the region extracted from the indoor scene, corresponding to a 50° FoV, is taken from the centre of the full GT panorama (for most scenes), shown in the second column. The IBLs are reexposed and tonemapped with $\gamma = 2.4$ for display.

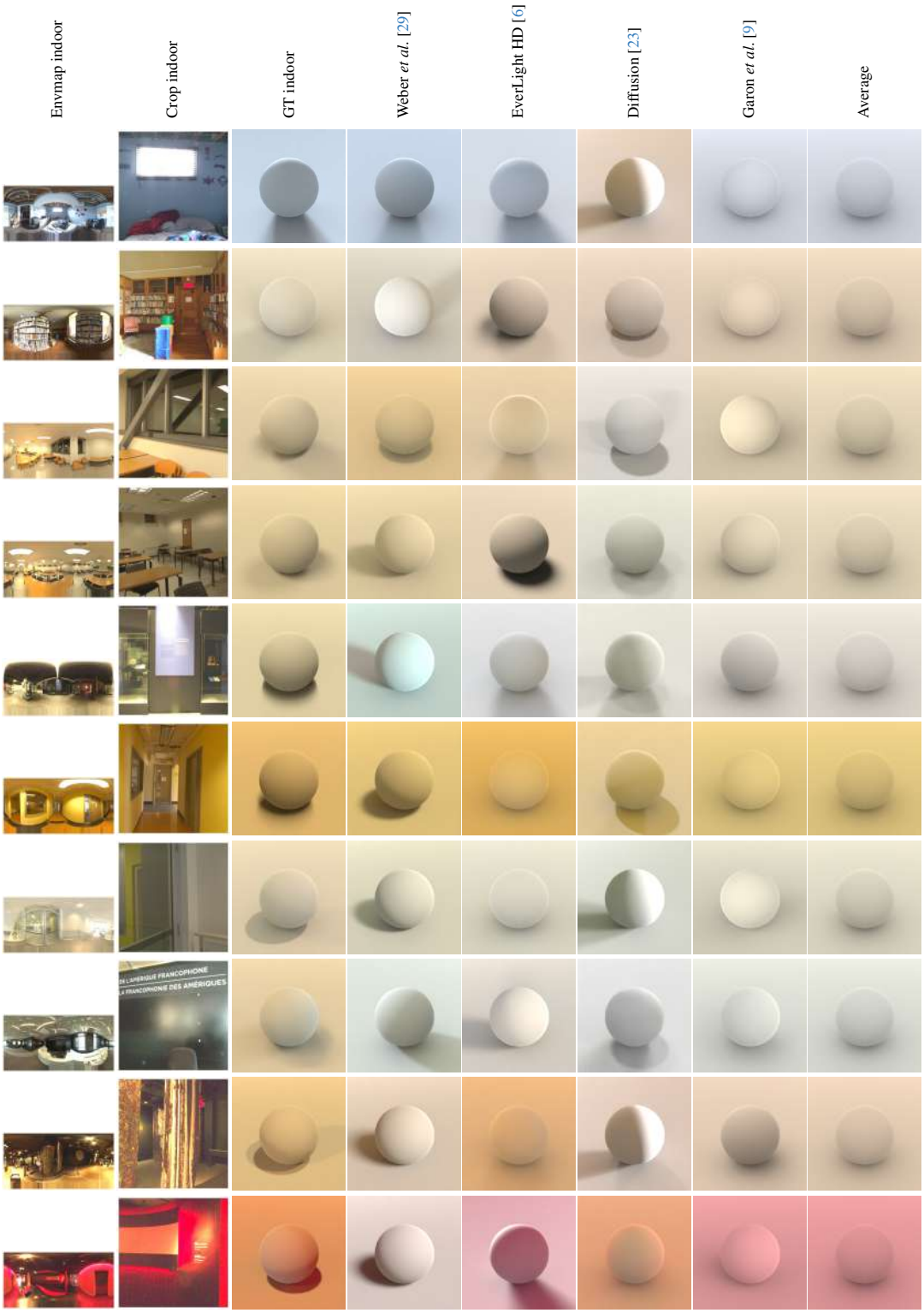

Figure 20. Stimuli used for the generalisation task 1 experiment with the diffuse sphere. The full HDR panorama (first column) is reexposed and tonemapped with $\gamma = 2.4$ for display. The region extracted from the scene (second column), corresponding to a 50° FoV, taken from the centre of the full panorama (for most scenes). The rendered stimuli using the ground truth IBLs (first columm) are shown in the third column. The other columns are the rendered stimuli using the IBLs (shown in fig. 1) produced by the different lighting estimation methods.

Figure 21. Stimuli used for the generalisation task 1 experiment with the glossy sphere. The full HDR panorama (first column) is reexposed and tonemapped with $\gamma = 2.4$ for display. The region extracted from the scene (second column), corresponding to a 50° FoV, taken from the centre of the full panorama (for most scenes). The rendered stimuli using the ground truth IBLs (first columm) are shown in the third column. The other columns are the rendered stimuli using the IBLs (shown in fig. 1) produced by the different lighting estimation methods.

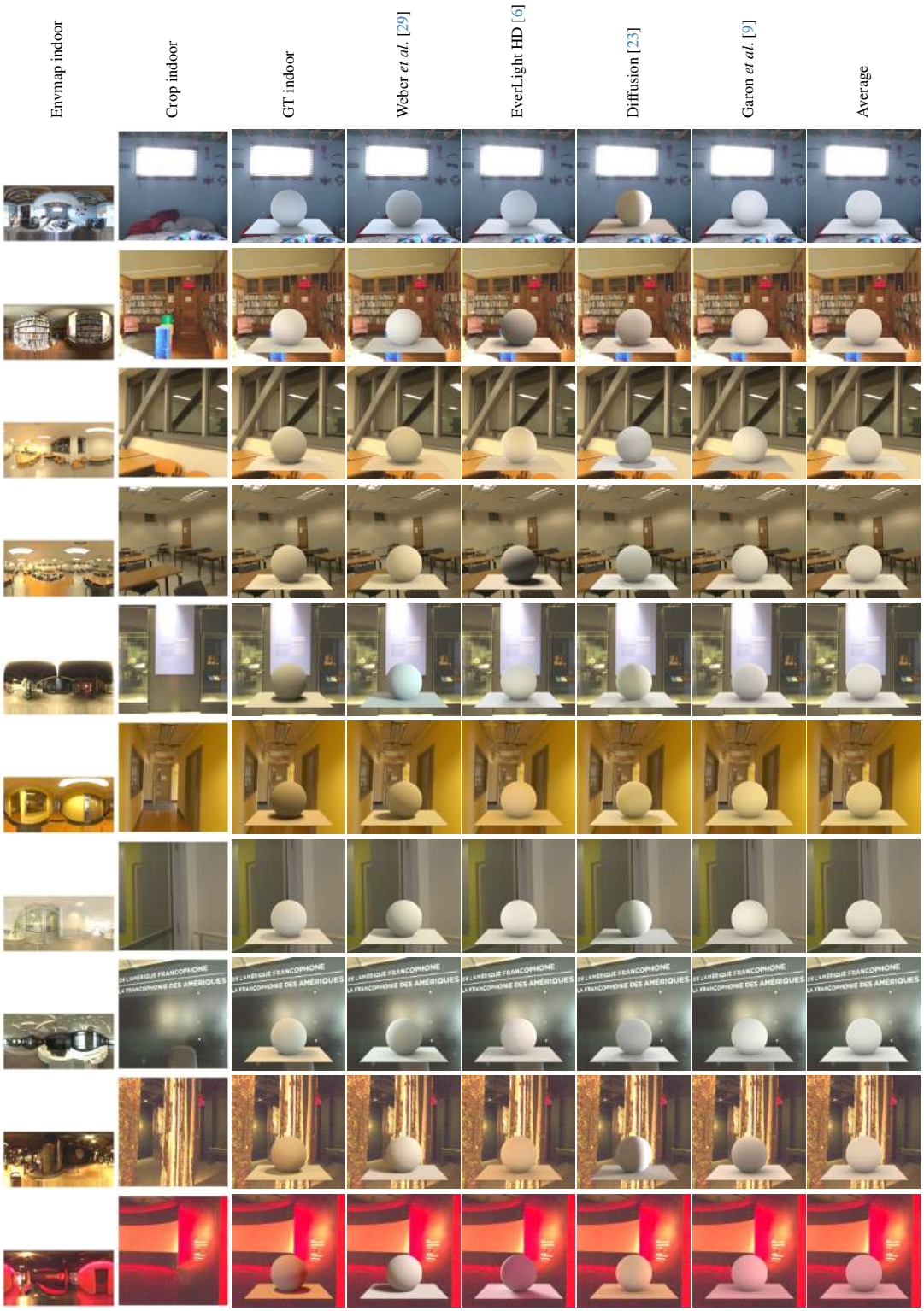

Figure 22. Stimuli used for the generalisation task 2 experiment with the diffuse sphere. The full HDR panorama (first column) is reexposed and tonemapped with $\gamma = 2.4$ for display. The region extracted from the scene (second column), corresponding to a 50° FoV, taken from the centre of the full panorama (for most scenes). The rendered stimuli using the ground truth IBLs (first columm) are shown in the third column. The other columns are the rendered stimuli using the IBLs (shown in fig. 1) produced by the different lighting estimation methods.

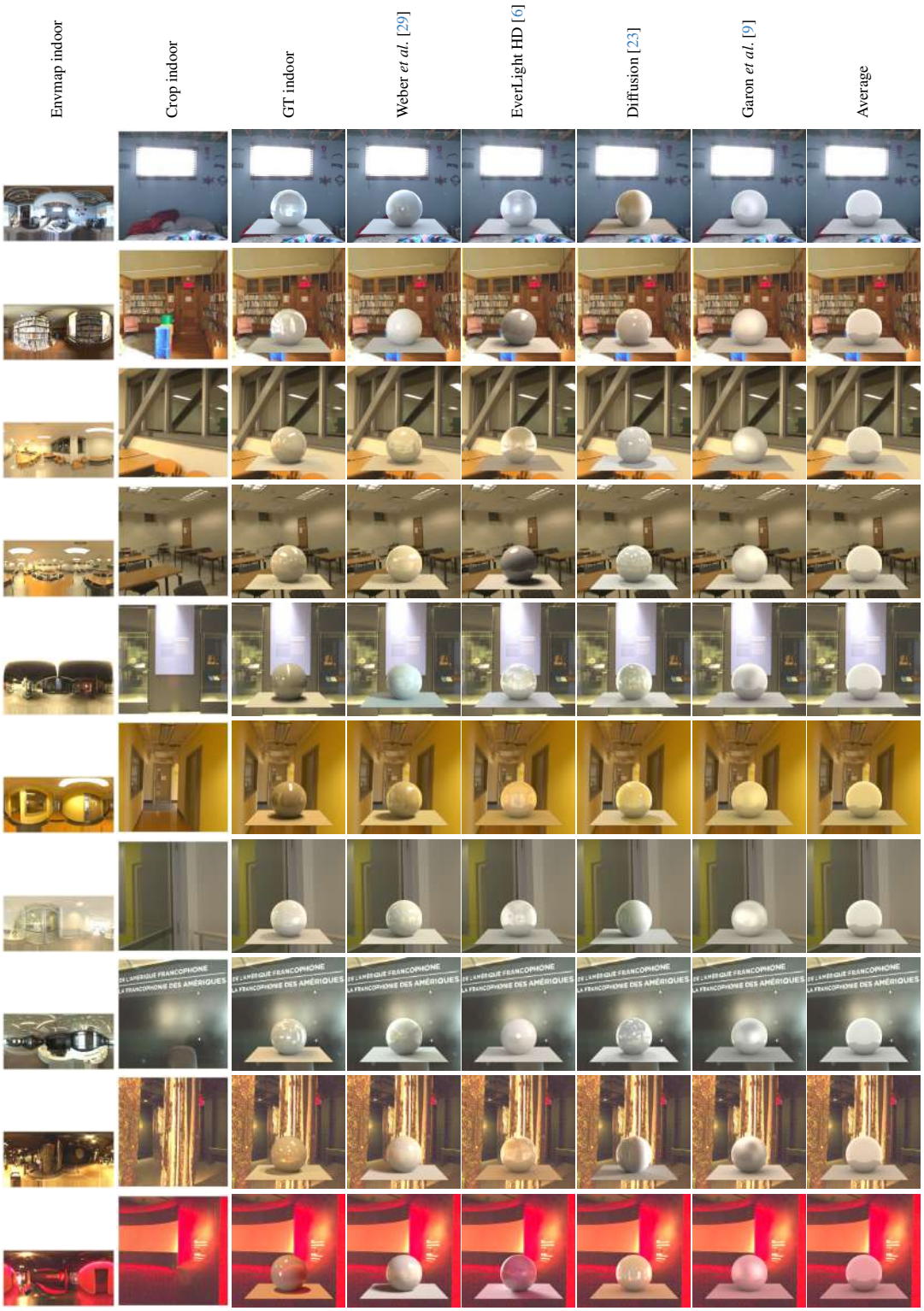

Figure 23. Stimuli used for the generalisation task 2 experiment with the glossy sphere. The full HDR panorama (first column) is reexposed and tonemapped with $\gamma = 2.4$ for display. The region extracted from the scene (second column), corresponding to a 50° FoV, taken from the centre of the full panorama (for most scenes). The rendered stimuli using the ground truth IBLs (first columm) are shown in the third column. The other columns are the rendered stimuli using the IBLs (shown in fig. 1) produced by the different lighting estimation methods.

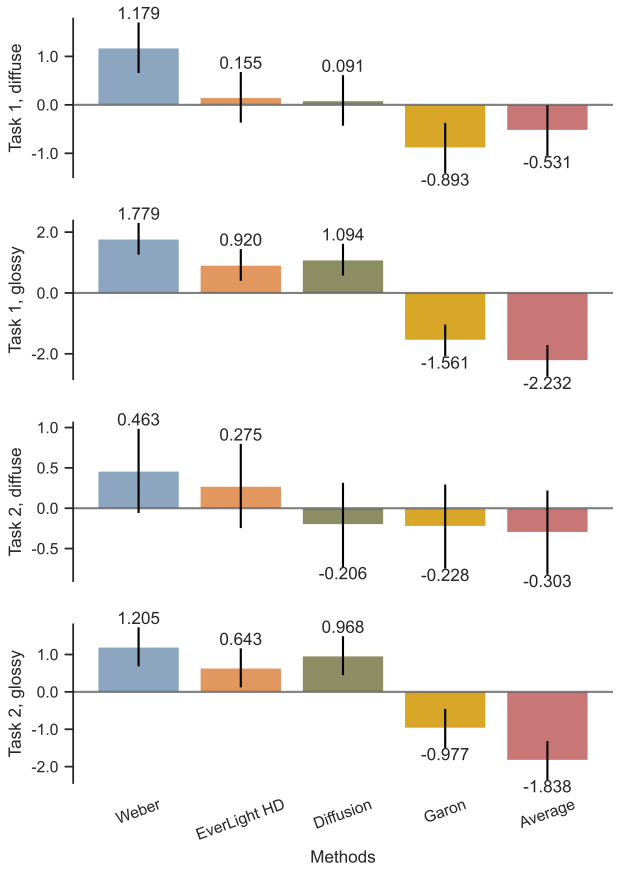

Figure 24. Thurstone Case V Law of Comparative Judgement scores from the observers as a function of the different generalisation lighting estimation methods (columns), for the different types of sphere materials used in the experiments (rows). Error bars correspond to 95 % confidence interval.

[21] 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. 18

[22] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 1

[23] 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. 27, 28, 29, 30, 31

[24] 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. 18

[25] 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. 18

[26] Louis L Thurstone. A law of comparative judgment. In *Scaling*, pages 81–92. Routledge, 1927. 16

[27] 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. 1, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14

[28] 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. 18

[29] Henrique Weber, Mathieu Garon, and Jean-François Lalonde. Editable indoor lighting estimation. In *Eur. Conf. Comput. Vis.*, 2022. 1, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 18, 21, 27, 28, 29, 30, 31

[30] 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, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 18, 22

[31] 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. 18

[32] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Trans. Image Process.*, 30:3474–3486, 2021. 18