A comparative study of the perceptual quality and aesthetics of Tone-Mapping Operators

Blasco Roca, Pau[a]

[a]*Matemàtica Computacional i Analítica de Dades, Facultat de Ciències, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Catalunya*

Abstract

Tone-mapping operators (TMO) are algorithms designed to generate perceptually similar low-dynamic range images from high-dynamic range ones. We studied the performance of nine TMOs in three psychophysical experiments, where observers compared digitally-generated images both between them (digital to digital) and against the real scene (digital to real). All experiments were performed in a controlled environment and the setups were designed to represent different possible scenes in real life. In the first experiment, we evaluated the local relationships among lightness levels. In the second one, we evaluated global visual similarity and naturalness between physical scenes and tone-mapped images, which were presented side by side. In the third one we ranked digital images in terms of their aesthetics and their visual appeal to the subjects. We ranked the TMOs according to these three experiments: grayscale and intrinsic image properties conservation (e1), faithfulness to reality (e2), and aesthetic appeal (e3). A correlation study involving all metrics revealed a slight positive correlation between better grayscale mapping and Modern TMOs (e1), as well as a strong positive correlation between faithfulness and aesthetics (e2, e3).

Keywords: Computer Vision, Tone-Mapping Operators, HDR imaging, Color Perception, Lightness Perception, HVS, Algorithms, Psychophysics

1. Introduction

In almost every natural scenery, and with the presence of sunlight or strong reflections, we find great differences between the lightest and the darkest regions that we can observe. For context, the difference between the energy emitted by sunlight (equivalent to a daylight scene) and starlight (equivalent to a darker/night scene) is approximately about a hundred million to one [1]. In vision, if we were to capture these scenes, we would label them as HDR, or High Dynamic Range, precisely because of the difference between the darkest and the lightest points. The Human Vision System (HVS from now on) has, through evolution and natural adaptation, solved this problem by processing light in a non-linear manner, and thus reducing this range to about ten thousand to one [2][3]. But what about digital cameras, phones, and computer screens? In the following sections we explain how they are able to capture light and faithfully represent HDR images.

1.1. Historical Context

The problem of translating an HDR scene into a Low Dynamic Range (LDR) image or representation is not new nor recent. For example, during the Renaissance, the *Chiaroscuro* technique appeared, in which the painters dealt with heavily-contrasted images and sources of light. Renowned painters such as Leonardo da Vinci, Caravaggio, Constable, and Rembrandt used it in their art, depicting bright objects in very dark settings and generating beautiful and realistic paintings.

The arrival of photography implied a new set of challenges, due to the strong limitations of early light-sensitive material [4]. First, the technique of *Composition* was used, a method rooted in the same principles as the Chiaroscuro, where different light sources and complicated camera angles allowed the photographer to depict highly contrasted images. Later, the technique known as *Combination Printing* arose, in which the photographer combined different takes at different exposition levels of the same scene, to create a final image. The ultimate exponent of this technique is arguably Rejlander's "The Two Ways of Life" (Figure 1), where 32 different pictures taken at different exposition levels over the course of six weeks were combined into a single image [5][6]. As the art form evolved, other techniques were explored: examples of the first characterization of silver halide films as plots of density vs exposure were made by Hurter and Driffield in 1890 [7].

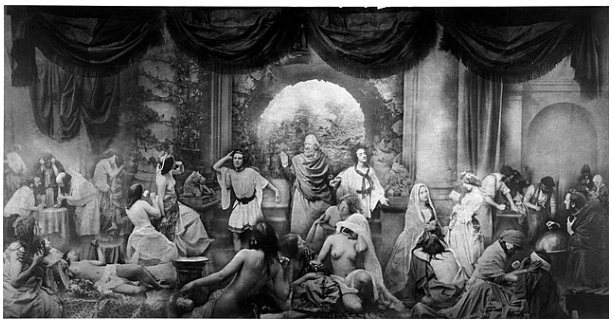

Figure 1: Rejlander's "The Two Ways of Life". From Wikimedia Commons.

1.2. Electronic HDR imaging

Analogue photography in HDR scenes allowed only for limited manipulations of the image (controlling exposure time, chemically altering the physical medium of the images, combination printing...). This gap was closed by the arrival of electronic digital imaging, which opened a whole new world of possibilities –individual pixel data could be utilized for an infinite number of calculations and interactions. At that moment, the focus fell on understanding the functioning of the HVS, and then mimicking it with algorithms.

It has been shown that photopic human vision is the result of highly nonlinear processing of the information captured by the retinal cones. This process includes the inhibition of the output of certain neurons by the output of its neighbouring neurons in its field of view [8], which results in an increased sensitivity for spots and edges compared to uniform, non-varying patches of light. The HVS also combines the visual information in the retina into a series of post-receptoral, chromatically opponent channels to transmit it into the visual cortex via the optic nerve [9]. In the cortex, visual information is mostly processed in terms of its spatial frequency and visual orientation [10]. A series of psychophysical experiments in the 60s dealing with achromatic Mondrians demonstrated that patches reflecting light with the exact same physical properties appear completely different to observers [11]. This implies that digital images cannot be modified using a pixel-wise algorithm to simulate the appearance seen by the HVS: if an algorithm tries to recreate its behaviour, the information in a single pixel should not be processed individually, without taking context into account (a detailed and comprehensive review can be found in [12][13][14][15]). Other effects could be considered, such as how the visual cortex processes local brightness interactions [16][17]. More detailed experiments have shown the effects of edges in illumination perception by matching the appearance of painted wooden facets to that of a painted test target [18], a procedure very similar to our first psychophysical experiment.

To mimic the functioning of the HVS, electronic imaging systems set out to use information not only from single pixels but from the entire scene. This allowed for more flexibility when calculating appearances and applying them to electronic displays or prints. Quite interestingly, the next wave of HDR algorithms reverted to the old multiple exposure methods and techniques of analogue photography.

The latest wave of algorithms, from around 2018 and onward, have used machine learning, most of them working with Convolutional Neural Networks. CNNs work with spatial information and, by tuning their filters, can recognize edges, spots, shapes and noise in images –mimicking the behaviour of human neurons [19][20].

1.3. Tone-Mapping Operators (TMO)

Representing real-world scenes, which almost always have a HDR, into LDR media presents several crucial challenges. Most imaging devices, such as cameras and monitors, are only able to obtain and display images within a range of 100:1 [2] (usually, 256:1), and exceptionally with a slightly broader range of around 1000:1 (for specialized HDR led-based displays) [21]. To solve this issue, multiple non-linear image processing techniques were developed. To construct the HDR image, many LDR images of the same scene are taken, each one at a certain exposure level, thus capturing a much larger dynamic range. Then, the most relevant information is extracted for each exposure, and the HDR image is reconstructed. This image cannot usually be represented in a standard LDR medium, and here TMOs come into play.

The most common solution is to use Tone-Mapping Operators (TMOs) to compress the information from the HDR image to an LDR one, while preserving contrast, color, and other perceptual characteristics. The performance of these algorithms depends on various factors, such as lighting, reflections, viewing conditions, or the amount of detail. They are usually evaluated using computational [22][23] and psychophysical [24][25][26][27] methods.

Although HDR images are usually able to reproduce a wider range of highlights and shadows than LDR ones, the presence of veiling glare both in the camera and in the human eye limits the possible range of accurate luminance measurements [7]. Since HDR images are perceptually closer to reality, there must be reasons other than simply obtaining a larger range of luminance values for this perceived improvement. It has been hypothesized that this improvement could come from better preservation of relative spatial information regarding digital quantization [7].

In this paper we present various experiments and psychophysical analyses to evaluate the performance of our 9 selected TMOs. Some of them are based on past work [28], while others are new. They will allow us to rank the TMOs according both to (perceptual) realism and aesthetics. Like previous studies in the field, our experiments were performed in a controlled environment, minimizing external distractions or changes in lighting, ambiance or perceived colour in our scenes.

1.4. "Classic" vs "Modern" methods

As of January 2024, there have been plenty of implementations of tone-mapping algorithms, many of them coming from the private sector –industry related to electronic imaging and/or optics–, but also from the public sector and open-source communities. Since there has been a recent change of paradigm in the TMO field, we believe it would be useful to understand TMOs by classifying them into two categories: Classic TMOs and Modern TMOs.

Classic methods range from a variety of techniques, which usually consist in mathematic functions or algorithms applied to either individual pixels (Global TMOs) or broader regions of the image (Local TMOs). It is worth noting that, even if counter-intuitive, this has been the conventional classification in the last two decades. In 2007, Yoshida et al. [29] observed a clear distinction between the results they produced. These methods were the first to be utilized in electronic HDR imaging and are still in use, although not as widespread as before.

From 2018 onwards, newer (Modern) methods have been entering the stage. These consist of artificial intelligence, machine learning, and deep learning methods, usually involving convolutional neural networks with various hidden layers. This paradigm is substantially different from the Classic TMOs.

2. State-Of-The-Art

2.1. TMO evaluation: psychophysics and experiments

There exist several paradigms regarding the evaluation of TMOs. As some of them are focused mainly on aesthetic considerations, input/validation from the HVS is not as important. Others are focused on reproducing the HVS responses, and thus their validation experiments revolve around psychophysical experiments. To date, many have been conducted, and can be classified into the following groups:

2.1.1. Non-referenced experiments

These usually consist of pairwise comparisons between images processed by TMOs. Images are shown in pairs to the subject, who has to rate them or choose the one with the best features.

For example, one of the first psychophysical experiments regarding TMOs [24] compared six different algorithms on four different scenes (synthetic and real). The subjects had to rate the images by apparent contrast, level of detail, and naturalness. The ranking concluded that subjects preferred TMOs which produced detailed images with moderate contrast.

Kuang et al. [25] performed pairwise comparisons to rank eight different TMOs applied to ten different sceneries and two modes (B&W and colour). In it, subjects chose their preferred images based on general rendering performance. Their results showed that the grayscale tone-mapping performances are consistent, and very similar, to those in the colour rendering results.

2.1.2. Real-Scene-referenced experiments

There have been several experiments which use the real scene as a reference. In them, the subject is able to directly compare reality to the processed image, which is displayed on an LDR monitor next to the scene.

Yoshida et al. [27] were pioneers of this paradigm, introducing it for the first time by selecting two indoor architectural scenes for their experiments. Fourteen subjects were asked to rate the images on screen according to criteria like realism and appearance. They concluded that none of these attributes had a strong influence on the perception of naturalness by itself.

In another work, Ashikhmin and Goyal [30] also performed three psychophysical experiments to rank TMOs, and one of them involved direct comparisons to the real scene. In it, subjects were asked to rank different tone-mapped images based on how closely they mimicked the real scene, while viewing the real scene next to the processed image as a basis for comparison. They observed that the results obtained with these experiments deferred significantly from the experiments where the subjects didn't see the real image.

Later on, Kuang et al. [31] performed three different experiments: "preference evaluation", "image preference modeling", and "accuracy evaluation". In the accuracy evaluation, an additional rating was performed using the real scene as a basis for comparison, where subjects rated attributes like highlight contrast, shadow contrast, highlight colourfulness, shadow colourfulness, overall contrast, and overall rendering accuracy. In this experiment, though, the processed images and the real scene were not next to each other, so the subjects didn't have immediate access to both of them at the same time. This meant that the subjects had to rely on memory to perform the comparisons.

In another occasion, the authors of the iCAM06 operator [32] performed two psychophysical experiments similar to the ones just described [31]. In the second one, observers had to evaluate the overall rendering accuracy by comparing the images to the real scenes, which were set up in an adjoining room.

With the objective to define a standardized overall image quality measure, Cadík et al. [33] studied the relationships between individual image attributes (brightness, contrast, reproduction of colours and details, etc). They performed two psychophysical experiments, using fourteen different TMOs, so they could propose a scheme of relationships between these attributes. Their first experiment involved real-scene comparisons: ten subjects were asked to rate images based on how their attributes compared to the real HDR scene.

In a new study, Cadík et al. [34] repeated their experiments, adding two new scenes. As these were outdoors, subjects had to perform the experiments at the same time of the day and with similar weather conditions to when the HDR image was taken, in order to make fair comparisons.

2.1.3. HDR-monitor-referenced experiments

In 2005, Ledda et al. [26] performed two different psychophysical experiments comparing six different TMOs to linearly mapped HDR scenes displayed on an HDR device. Their method attempted to replicate real-scene-comparison experiments by giving the subject a closer representation of reality with the HDR monitor. In their experiments, the subjects were asked to select the processed image they believed to be more similar to the HDR reference. They focused on overall appearance and detail reproduction.

In a later work, Akyüz et al. [35] asked subjects to rank six images according to their subjective preferences. The images included an HDR image, three tone-mapped images and two LDR exposures (both good representations of the real scene). It was found that participants did not systematically prefer images processed by TMOs over the LDR exposures.

2.1.4. Machine Learning experiments

A good neural network needs a good training, and this can only be done with the appropriate data. In this case, the data usually consists of a large amount of HDR images, paired with

labels describing various aspects or features, or simply the expected LDR version of it. Other experiments involve GANs [36], a neural network architecture capable of generating validation content and comparing its results against it. As other methods and experiments involved comparing the algorithm to the criterion of the HVS, one could argue that these machine learning experiments also "compare" their results against the real image, fine-tuning and correcting their parameters as they learn. Most of the time, a final experiment based on the HVS is performed, but usually it acts as a "validation" phase.

All the previous studies have been focused on subjective comparisons of various image appearance attributes such as contrast, colourfulness, sharpness, reproduction artifacts, etc. either against the real scene or within TMOs. While there is no doubt that this is extremely important, we believe a good TMO should output a scene that produces a visual sensation as close as possible to the one generated by the real scene. In our work, we will be ranking TMOs both by how closely they represent reality, and by how aesthetically pleasant are their results to the viewer.

2.2. Tone-Mapping Operators

As stated before, TMOs can be classified as Classic or Modern, and the Classic ones can be further separated into Local and Global operators. The line is not as clear as it might appear, as many of the Classic TMOs can behave in both Local and Global mode, or even use techniques from both.

For this study, over 60 different TMOs were considered, their creation dates ranging from 1993 to the current date. Some have been discarded, as they have lost relevance over time or are too simple or similar to each other. Unfortunately, although there have been numerous publications and algorithms created after 2017, many of them were missing the source code necessary to test their results. In one case [37], the code was uploaded without the checkpoints or dataset necessary to run it, making it impossible for other researchers to reproduce its results.

We have been able, however, to find nine TMO algorithms relevant to our study. Here is a brief description of their methodology, as well as their aliases used along the paper.

-*OppoCPH2207* (OPPO) [38]. The HDR function in this mobile phone camera acts discretely as a TMO. By steading the device, it captures several expositions which later joins into a single, LDR image. In our case, the experimental images were taken directly with a phone (50 MP camera), instead of being captured with the Foveon and processed a posteriori.

-*Kaminari* (KAMIN) [39]. The algorithm is strongly based on the Combination Printing. Each exposure is broken down into over and underexposed zones, which are gamma-corrected in relation to their properties. They are then combined, preserving the original colors, to reconstruct the final LDR image.

-*Li* (CLUST) [40]. It is a local TMO based on the Retinex Theory. It decomposes the image in a base and a detail layer, separates by patches, and clusters the information based on different color and lightness metrics.

-*KimKautz* (KIM) [41]. It is a global tone-mapping operator that models the HVS light sensitivity with a Gaussian distribution. It is based on how the HVS adapts to the average log-luminance of the scene.

-*Krawczyk* (KRAWC) [42]. It is a local TMO based on the anchoring theory. It decomposes the image in various chunks (frameworks) based on their luminance, and then calculates their lightness values locally.

-*Liang* (L1L0) [43]. It is a global tone-mapping operator based on an upgrade to the L1-L0 (base/detail) decomposition. Its hybrid model is able to deal with halos and other defects more efficiently.

-*Khan2020* (PRCPT) [44]. Based on the Khan2017 TMO, this algorithm first converts the HDR pixels to a perceptually-quantized image, simulating the HVS. Then a histogram method is applied on the luminance channel, using a uniform distribution model to soften the extreme cases.

-*Reinhard* (RNHRD) [45]. It can function both as a global and a local TMO. The stepping stone for many of its successors, Reinhard's TMO combines a non-linear function to process lightness and a dodge-and-burn post-process.

-*Khan2018* (TIE) [46]. The algorithm focuses on a luminance model based on the HVS light sensitivity. With the TVI-TMO algorithm, it adapts the luminance, and then uses a histogram-based method to map it to LDR.

3. Methodology

We performed three different experiments to compare the TMOs: *segment matching*, *scene reproduction* and *aesthetics ranking*. The aim of the first experiment was to study the internal relationships among gray levels between the processed image and the real scene. The aim of the second one was to evaluate the algorithms according to how similar the overall results were to the real scene. The third experiment was to rank processed images based on personal preference, choosing the most aesthetically pleasing for the subjects. These experiments follow procedures similar to [27] and [28].

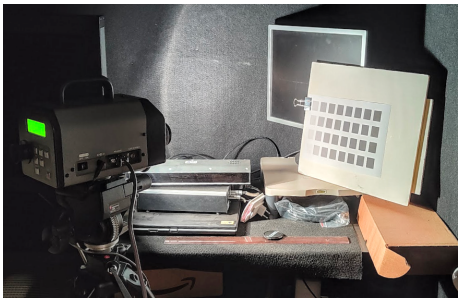

Figure 2: Setup to measure the lightness values of the grayscale chart.

The third experiment is inspired on *Le et al.*'s work [47], where it is shown that human perception of image quality is influenced more by aesthetic judgements than by the image naturalness.

3.1. Materials and Set-up

Our experiments were performed in a darkened room with its walls covered by light-absorbent material, blocking reflections and other uncontrolled illumination. The set up was lit by a 100W incandescent lamp whose glass casing was painted blue, giving it a spectral profile close to the CIE D65 standard illuminant.

We used a Display++ LCD Monitor driven by a ViSaGe MKII Stimulus Generator from Cambridge Research Systems LTD (Rochester, England) to represent the HDR images. This system emmitted light too, but towards the subject and not the scene. The monitor self-calibrated via a customary Cambridge Research Systems Ltd. software for ViSaGe MKII Stimulus Generator. Both the monitor and the real scene were set up so that the objects in both (virtual and real) scenes would subtend approximately the same angle and look similarly positioned to the observer. The pictures were taken with a Sigma Foveon SD10, previously calibrated [28].

We built different scenes, each including a gray-level reference table, two solid cuboids and a variety of elements of different colors and shapes. The reference card was built by printing a series of 32 gray squares (30.7mm x 24.0mm) arranged in a flat 8×4 distribution. The arrangement of rows and columns was labelled A; B; C;...; H for the rows and 1, 2, 3, 4 for the columns (Figure 4). The gray values of the patches were chosen so that their lightness would decrease monotonically from the top (patch A1) to the bottom (patch H4). For this, we calibrated our printer, measuring the printed reference values with our Minolta spectroradiometer (Figure 2) in a separated, controlled environment with uniform light (no shadows, same angle of incidence). After several adjustments, we were able to obtain a new grid and observed that their lightness values were equally spaced, meaning that the lightness (as perceived by the viewer) was linearly increasing from tile to tile.

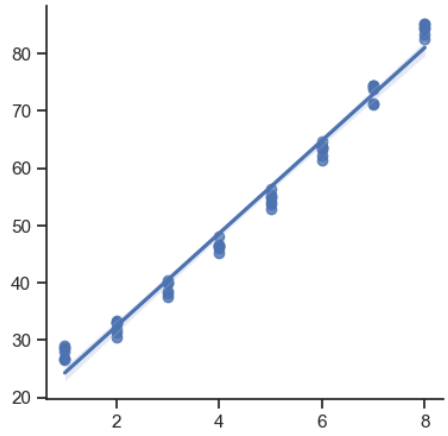

Figure 3: Final lightness measures after the calibration. $R^2 was 0.9849, indicating near-perfect linearity$

We used four parallepipeds of various sizes, whose sides were covered with paper colored with random tones from our graylevel reference card. Apart from this, we also used different toys and natural-looking objects of a broad range of colors. We placed all of them in our scenes, glued them and fixed their positions. Once we were satisfied with their composition and their color variety, we took a picture of each of them from the perspective from which the subjects would be looking. We processed each picture with each TMO in order to display them later.

3.2. Experimentation

To compare the TMO algorithms, we performed three different experiments. The aim of the first two experiments was to study the naturalness of the algorithms, meaning how well they were able to capture and replicate the real scene. The first one focused on the internal relationships among gray-levels between the processed image and the real scene. The second one focused on the overall perception of similarity to the real scene (general naturalness perception), and the third experiment had a different aim: it focused on aesthetics evaluation, studying the overall perception of beauty among all algorithms. In all three cases we obtained a ranking of the different TMOs.

All experiments started with a 1-minute period of subject adaptation to the ambient light. It should be noted that the source code for the TMOs was obtained from different sources (HDR Toolbox [48] and personal GitHub repositories) and ran in the default settings, to ensure impartiality. There were 13 subjects per experiment in total (7 male, 8 female). Subjects were aged 16-50, with standard or corrected-to-standard vision and naïve to the aims of the experiment. They spent about an hour to an hour and a half per session. They were compensated economically for their participation.

3.2.1. Experiment 1: Segment Matching

This experiment consisted on two different tasks:

1. Real scene matching: After adaptation, subjects were asked to match, in the real scenes, the brightnesses of the parallelepipeds' surfaces to the brightnesses of the reference table in the scene.
2. Processed image matching: Here the real scene was not visible and the observers just saw a digital tone-mapped version of it on the monitor. Their task was the same as #1, except that the matchings were conducted entirely on the gray patches shown on the screen.

Although no time constraints were set to perform the tasks, the subjects were advised to take no more than 30 seconds per match.

There were four conditions for the experiment, corresponding to two POVs of the two different scenes created (see Figure 4). Observers matched the 23 surfaces in all 9 different tone-mapped images and in the real scene for each condition, so they performed a total of 230 matchings. Matchings were conducted by (the investigator) logging results in a computer for later analysis. The stimule were always shown in random order, to avoid any possible biases.

Figure 4: Close-up picture of the scene, the two parallelepipeds (one in the back, practically invisible with this LDR image) and the reference chart.

3.2.2. Experiment 2: Scene Reproduction

This experiment consisted on an overall rating on processed images vs. the real scene. Subjects were advised to focus on details such as shadow and color conservation, sharpness and over/under exposition, aswell as an overall sensation of realism. Although there were no time constraints to perform the tasks, they were advised to take no more than 30 seconds per decision.

This experiment also included four different conditions. The subjects gave a rating, from zero to ten, to each stimulus, which was logged for later analysis. The subjects were also shown the images in random order before each section, so they would have prior information on what ratings to give.

3.2.3. Experiment 3: Aesthetics

This last experiment consisted on an overall rating on processed images. The subjects had to rate how ”aesthetically pleasing” or ”beautiful” the displayed images were to them, based solely their own opinion and perception of art and image composition. Although there were no time constraints to perform the tasks, they were advised to take no more than 30 seconds per decision.

This experiment also had four different conditions. The subjects gave a rating, from zero to ten, to each stimulus, which was logged for later analysis. The subjects were also shown the images in random order before each section, so they would have prior information on what ratings to give.

4. Results

4.1. Experiment 1: Segment Matching

For each scene evaluated in Section 3.2.1, we obtained 10 sets of experimental values (the real scene in Task 1 plus 9 tone-mapped images in Task 2) for each visible face of the cuboids. This yielded a total of 10x23=230 gray datapoints per subject. We performed several analyses to evaluate to what extent the local interrelations perceived by the observers in the tone-mapped versions corresponded to those perceived in the real scene.

Figure 5: Overview of the experimental setup.

4.1.1. Linear Model fitting

In the first analysis, we compared the perceived distances from each tone-mapping algorithm to the real scene. The real scene and the TMOs were defined in a 23-dimensional space, where each dimension corresponded to each evaluated surface. To obtain these ten 23-dimensional vectors (one for each presented stimuli and one for the original scene), the psychophysical data was averaged across all subjects. With this information, we fitted a linear regression model for each TMO, as it can be seen on Figure 3.

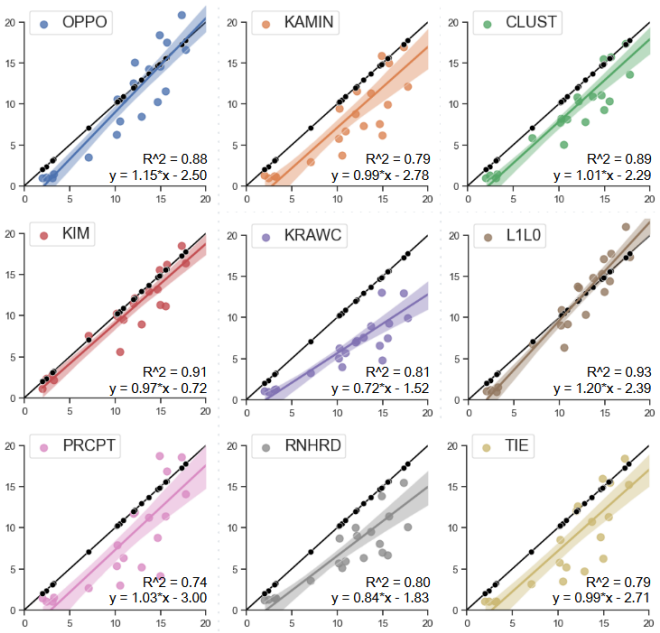

Figure 6: Gray-matching mappings of each TMO from the original scene (in black)

This plot allows us to have a quick overview of the different transformations applied by the TMOs. A priori, we observe that *Krawczyk* (KRAWC) is notoriously deviating from the original distribution, while *OppoCPH2207* (OPPO), *KimKautz* (KIM) and *Liang* (L1L0) are practically on top of the black line.

The image also contains information regarding the equation of the linear model and the R-squared error.

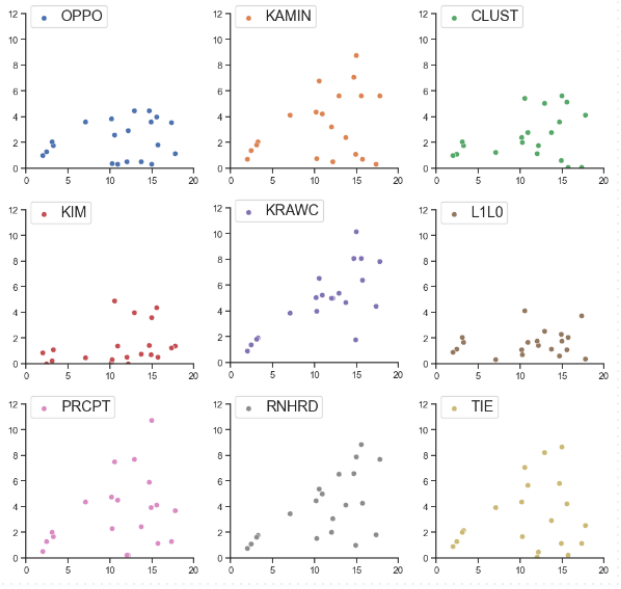

Figure 7: Gray-matching errors of each TMO from the original scene

In this array of error plots we can quickly see that errors tend to accumulate around the lower end (brighter tones) of the mapped space. Towards the beginning, the errors seem to decrease or be practically meaningless. *KimKautz* (KIM) and *Liang* (L1L0) have the least errors, while *Krawczyk* and *Reinhard* have the most.

To further study these distributions, a few statistical tests and estimators will be calculated in the next section-

4.1.2. Statistical comparison

After analyzing the different regression models, a few statistical comparisons were made, and summarized in the following table.

TMO	Slope Difference	RMSE	T-test p-value	R^2	Type
OppoCPH2207	0.149	2.600	0.726	0.880	Modern
Kaminari	**0.012**	4.074	0.108	0.792	Modern
Li2018	**0.011**	2.927	0.225	0.890	Modern
KimKautz	0.027	**2.058**	0.541	0.905	Classic
Krawczyk	0.283	5.613	**0.002**	0.807	Classic
Liang	0.200	**2.122**	0.969	0.933	Modern
Khan2020	0.029	4.315	0.163	0.741	Modern
Reinhard	0.159	4.562	**0.025**	0.797	Classic
Khan2018	**0.011**	4.050	0.117	0.788	Classic

Figure 8: Gray-matching deviations of each TMO from the original scene (in black)

The *Slope Difference* was calculated by taking the absolute value of the difference between each linear model's slope and the original. A lower Slope Difference means a better performance, and viceversa. In this metric, *Li*, *Khan* and *Kaminari* performed best, while *Krawczyk* and *Liang* scored the worst.

The *RMSE* is short for Root Mean Squared Error, and is a measure of how the model's datapoints are spread apart from the original distribution. A lower RMSE means better performance, and viceversa. The top performer was *KimKautz*, closely followed by *Liang*, and the worst was *Krawczyk*.

We also performed a two-sided t-test between each TMO's gray datapoints and the original sample. We assumed independency (the subject annotated the gray coordinates separatedly, and only focusing on either the scene or the screen at the same time) and a similar variance for the two distributions. Our H_0 was that the distribution of the gray points generated by the images had the same mean as the original distribution. We took p=0.05 for our 95% confidence interval. Only *Krawczyk* and *Reinhard* were the ones in which H0 was rejected, so with a 95% of a confidence interval it can be assumed that their gray distributions are NOT similar to the original gray distribution.

4.2. Experiment 2: Scene Reproduction

To better analyze the data from this section, it needed to be able to be compared fairly across subjects and scenes. To do this, the data was normalized with the following transformation:

$$r_{i,s,\alpha} = \frac{o_{i,s,\alpha} - \bar{o}_{i,s}}{\sqrt{\frac{1}{n}\sum_{\alpha'=1}^{9}(o_{i,s,\alpha}) - \bar{o}_{i,s}}} \tag{1}$$

Where $r_{i,s,\alpha}$ is the normalized ranking given by the i-th subject at the scene s for the image processed by the TMO α, $o_{i,s,\alpha}$ is the original ranking given for that stimulus and subject, and $\bar{o}_{i,s}$ is the average ranking given by the i-th subject at the scene s.

This transformation was applied at scene-subject level, in order to make sure any individual bias toward higher or lower rankings was eliminated. The normalized values were centered at 0 and ranged around [-2.5,2.5], each meaning how many standard deviations was it away from the mean (i.e, a normalized ranking of 1.5 meant that the TMO was at the quantile $\mu+1.5\sigma$).

With these transformations, comparisons and metrics were able to be performed fairly across all algorithms, scenes and subjects. The following plot displays the rankings of the Scene Reproduction experiment given by the subjects.

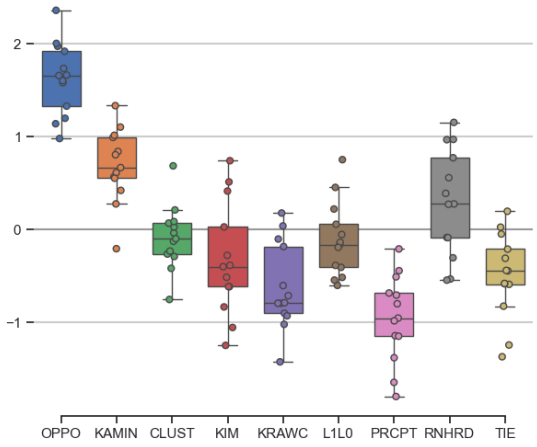

Figure 9: Scene Reproduction box plot containing populational and individual sample information.

In this plot, we can see several boxes of different colors. These wrap the datapoints contained around the percentile interval [25%, 75%], which represent the central half of the sample. The rest of the data points can be either on the whiskers $[\mu - 1.5\sigma, \mu + 1.5\sigma]$ or outside them, marking them as potential outliers. The line in the middle of the boxes describes the median observed.

There is a clear distinction between *OppoCPH2207* (OPPO) and *Kaminari* (KAMIN), which performed well above average and the rest of algorithms. *Reinhard* still seems to be performing confidently above average. The worst performers for this experiment were *Khan2020* (PRCPT), closely followed by *Krawczyk* (KRAWC) and *Khan2018* (TIE). Subjects did not seem to reach a consensus on *Li* (CLUST), *KimKautz* (KIM) and *Liang* (L1L0).

4.3. Experiment 3: Scene Aesthetics

The data was normalized at scene-subject level, as explained in **4.2**. The results for this experiment were very close to the Scene Reproduction rankings, with small differences. Below (Figure 10) is the plot representing the distributions of the subjects' evaluations.

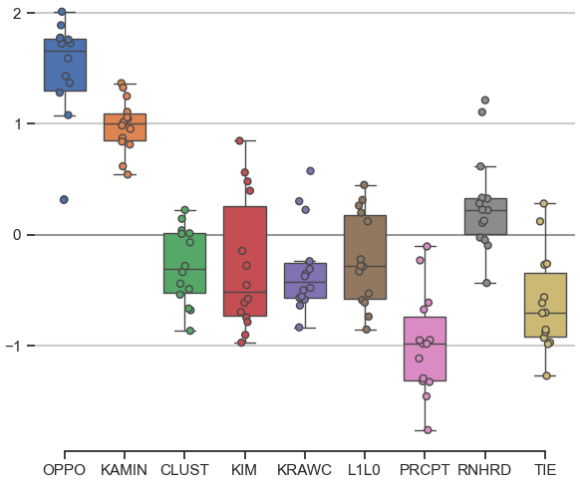

Figure 10: Scene Aesthetics box plot containing populational and individual sample information.

Again, *OppoCPH2207* (OPPO) and *Kaminari* (KAMIN) were the top performers, ranking avobe the rest. While *Oppo* still maintained first place, *Kaminari* was able to improve its ranking and place at $\mu + 1\sigma$. *Reinhard* (RNHRD) still had an above-average performance, but there was a slight decrease from the results in Experiment 2. *Khan2020* (PRCPT) was again the worst performer, followed by *Khan2018* (TIE) and *KimKautz* (KIM). The rest of the algorithms maintained more or less the rankings from Experiment 2, staying around $\mu - 0.5\sigma$ but without dropping below $\mu - 1\sigma$.

4.4. Intra/inter-subject disagreement

Since we based our study on psychophysical tests data, it was crucial to revise subjects' observations and rankings. This allowed us to deep dive in our data and scan for possible outliers or patterns that could be useful to our research.

4.4.1. Metric definition and disagreement calculation

We considered experiments 2 and 3 to be the most subjective, and decided to calculate both *inter* and *intra* subject agreement.

The first metric, **intra-subject disagreement**, describes subject consistency within their own answers. It is defined by how the rankings from each test subject differ between scenes. This measure could help detect subjects that are choosing or ranking algorithms at random. It is calculated as follows:

$$\gamma_i = \frac{1}{6} \sum_{s_1=1}^{4} \sum_{s_2=s_1+1}^{4} \|\vec{p}_{i,s_1} - \vec{p}_{i,s_2}\|_2 \tag{2}$$

Where γ_i is the intra-subject disagreement of the i-th subject, s_1 and s_2 are iterators for scene numbers, and $\vec{p}_{i,s}$ is the 9-dimensional vector with the TMO's normalized rankings given by subject i in scene s. Higher values of γ indicate lower consistency in an individual subject's answers, while lower values (or values close to zero) indicate higher levels of consistency. For example, a subject that gave algorithms (a, b, c) the ranking (8, 2, 5) in *every* scene, would achieve a $\gamma = 0$. On the contrary, a subject that gave completely different rankings in every scene would get a notably high γ.

The second metric, **inter-subject disagreement**, describes disagreement between the overall rankings of two different subjects. It is obtained by calculating the difference of average rankings between each possible pair of subjects. This metric could help find outliers, or subjects who greatly disagree with the general consensus. It is calculated as follows:

$$\lambda_{i,j} = \frac{1}{4} \sum_{s=1}^{4} \|\vec{p}_{i,s} - \vec{p}_{j,s}\|_2 \tag{3}$$

Where $\lambda_{i,j}$ is the inter-subject disagreement between the i-th subject and the j-th subject, s is an iterator for scene numbers, and $\vec{p}_{i,s}$ is the 9-dimensional vector with the TMO's normalized rankings given by subject i in scene s. These λ values are later used as elements of the Λ matrix, represented below as a heatmap. Higher values of λ indicate higher disagreement or discrepancy between subjects. On the contrary, subjects with similar rankings will get lower or near-zero values of λ.

Combining Excel and a Python script, the metrics we just described were calculated and represented in Figure 11.

4.4.2. Scene Representation disagreement

Focusing on plot *a1* from Figure 11, we observe that subject 7 had a noticeably higher γ value than the rest. Regarding inter-subject disagreement (plot *a2*) we notice that subjects 1 and 8 disagree the most from the rest. Still, the variability in the Λ matrix is great enough to absorb these small differences.

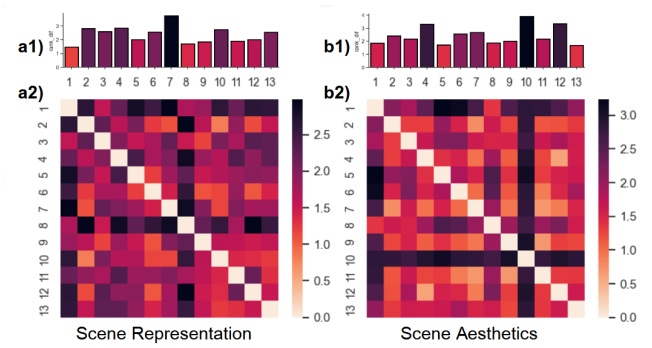

Figure 11: Disagreement metrics. *a1* and *b1* correspond to the intra-subject disagreement metric, while *a2* and *b2* represent the inter-subject disagreement matrices. Both *a1* and *a2* correspond to Experiment 2 (Representation), while *b1* and *b2* describe Experiment 3 (Aesthetics). Along all plots, darker values represent more disagreement, while brighter ones represent more agreement.

Since no subject has high values in both intra and inter-subject disagreement, we can conclude that the general disagreement is standard, and explained simply by a difference in which details are paid the most or the least attention.

4.4.3. Scene Aesthetics disagreement

Looking into *b1*, we again see a few (subjects 4, 10 and 12) higher bars, but none are extraordinarily different from the others. Regarding the inter-subject disagreement heatmap (*b2*) we can clearly see how subject 10 and subject 1 highly disagree from the rest of the subjects rankings.

Although subject 10 has high values in both intra and inter-subject disagreement, they are not discrepant enough to ring any alarms, and can simply be explained by personal preference or just differences in the evaluation of beauty.

4.5. Qualitative evaluation

We also summarized the subjects' general comments and observations on the experiments, since they mostly agreed in a few key points.

- 11 subjects commented at some point of Experiment 1 that some patches in the processed images' grayscale charts were indistinguishable from one another (probably due to the compression of the HDR lightness to the LDR picture).
- 10 subjects pointed out that the darkest colors in some of the images were not displayed in the grayscale chart of the TMO-processed images, or that the darkest color in the screen was darker than A1, the darkest color in the chart.
- 7 subjects noted that some of the parallepipeds sides presented a slight tint or hue. This happened with the brighter colors, which received secondary reflections from some of the objects.
- 2 subjects pointed out that, in the natural-looking scene, none of the images were able to fully and faithfully represent reality, and that all of them lacked something in order to feel truly real.

5. Discussion

A priori, Experiments 2 and 3 seem to have yielded very similar results, while the results from Experiment 1 don't have any obvious link to the other metrics. A correlation matrix between the metrics has been calculated, in order to find any non-apparent relationship between image properties.

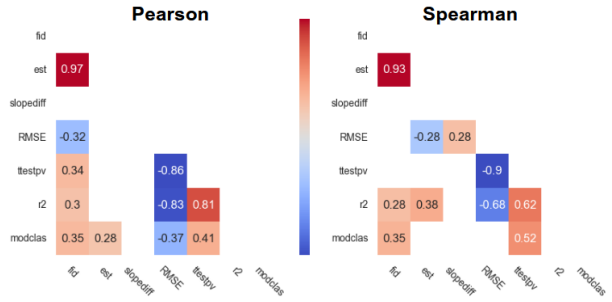

Figure 12: Pearson and Spearman's matrices of correlations, filtered out with a 95% CI (a sample of 13 subjects x 4 scenes = 52 data points was taken into account, even if some metrics have more samples than that). In order of appearance, *fid* is the Scene Representation normalized ranking, *est* is the Scene Aesthetics normalized ranking, *slopediff* is the slope difference, *RMSE* is the Root Mean Squared Error, *ttestpv* is the T-Test p-value, *r2* is the R2 linear model error, and *modclas* is a bool representing their type (Modern = 1, Classic = 0).

We first notice the almost-perfect relationship between Scene Representation and Aesthetics, which is equivalent to saying the ranking given to TMOs in Experiments 2 and 3 is very similar. *ttestpv*, *RMSE* and *r2* seem to also be strongly correlated in both Pearson and Spearman's matrices.

Interestingly enough, there is a low but statistically meaningful positive correlation between *modclas* and both *fid* (in both Pearson and Spearman) and *est* (only in Pearson). This means that there is a slightly positive correlation between being a Modern TMO and performing better in experiments 2 and 3. Modern TMOs also seem to have a lower RMSE (-0.32 Pearson correlation) and a higher consistency in their mappings (a higher R2, 0.3 Pearson correlation and 0.28 Spearman correlation).

With this we have seen that, even if this gray and lightness relationships within images might have an impact on the subject's perception of it, there are still other factors that need to be taken into consideration. Other studies also agree that a good lightness and grayscale mapping is not enough to fully represent a scene, and that other factors such as contrast and luminance gradients should also be taken into account [34] [33] [27] [28] [29].

We also took some time in Section 4.4 to study inter and intra-subject disagreement and possible discrepancies. Although some notable differences were found, we were not able to categorize them as outliers. Those disagreements can simply be explained by the variety and disparity of subjective opinions.

5.1. Comparison to other studies

Regarding the segment matching experiments, we have taken a similar approach as [28], studying the distributions and rela-

tions with the same metrics. We have, however, tried to display individual differences with greater detail, and also introduced correlation matrices to link and tie together the results of all the experiments.

Kuang et al. [25] performed a similar experiment regarding TMOs and psychophysics, but without the reference scene to make comparisons. Still, we agree with them in that *Reinhard* is one of the best ranked TMOs. It is worth noting that *Reinhard* [45] was made in 2002, 22 years ago, and is still one of the top performers, competing against algorithms using much newer and supposedly upgraded methods and technology.

Yoshida et al., in their studies in 2005 [27] and 2007 [29], also claimed that *Reinhard* [45] was one of the top performers. We agree with them in that the algorithm performs well with HDR indoor scenes (as the ones they used for their experiments).

Of course, half of the algorithms of the study are considerably new. *OppoCPH2207*'s last software update (ColorOS 13.1) was on March 21, 2023 [49], *Kaminari*'s latest version was uploaded on January 10, 2024, and the other papers haven't yet accumulated lots of citations. Recent state-of-the-art checks, psychophysical tests and even just review papers are necessary to stay up to date and evaluate the general progress in the field. Because of this, no other studies have compared and evaluated them yet, so we expect to take the first step in that direction.

5.2. Other factors and phenomena

Lately, the term *Uncanny Valley* has caught the attention of the media and the industry. This term is not new, and was coined back in 1970 by Masahiro Mori [50]. Initially defined to deal with robot and prothesis design, its use has been extended to describe digitally generated media (such as processed images or AI generated pictures) and even moved on to the genre of horror. The phenomena is interesting and incredibly relevant to the field of electronic image processing. On top of that, it could perfectly describe the effect caused by some of the TMOs in this study: even if they get to produe an image very close to reality, there's still something missing –something that might be imperceptible to an algorithm, but be detected instantly to the human eye. That close-but-not-quite to the real scene might be what is causing the discomfort in the subjects.

6. Conclusion

Our methods and experiments have shown that *Modern* TMOs are not necessarily better than *Classic* ones. Even if *OppoCPH2207* and *Kaminari* performed substantially better than the rest of algorithms in some experiments, the underperformance of *Liang* and *Khan2020* counter-balance this position. We have also seen that Modern TMOs don't necessarily preserve the intrinsic image properties better than the Classic methods: in the gray matching experiment (4.1.2) we found the top two performers to be a classical and a modern method, followed by a mix of both types. But that has not necessarily impacted the aesthetics or the capability of scene reproduction of the algorithms. The second (*Kaminari*) and third (*Reinhard*) algorithms in terms of aesthetics and faithfulness have placed sixth and eight respectively in the RSME metric of Experiment 1. Additionally, the first (*KimKautz*) and second (*Liang*) scorers in that same experiment haven't done as well, ranking way below *Kaminari* and *Reinhard*. In the discussion, we have claimed that this could be due to the *Uncanny Valley* effect [50], and the lack of other, much more complicated, naturalness features in the images.

6.1. Additional comments about source code

Before finishing the study, one more point has to be made. Up to 60 different TMOs were considered for evaluation, and only nine were able to be used. Specifically, from the 14 algorithms published in journals or papers after 2015, 9 (two thirds) had missing, broken, incomplete or faulty code [37] [51] [52] [53] [54] [55] [56] [57] [58].

These practices have rendered them unusable for this study, since we were unable to reproduce their results. Documentation was also scarce and incomplete, and the lack of demos, checkpoints or environments to run code also made it impossible to use the algorithms, The authors were contacted in several occasions, but the issue was left unresolved.

In a world of lightning-fast digital innovation, **Open Source** is crucial in research. Newer algorithms are inspired in older ones, usually upgrading them and generating better, more adaptive and realistic results. Older research conclusions and results allow other researches and investigators to build on them and to keep on making scientific progress. World-wide known software that we use in our everyday life is based, or entirely made out of, open source code. Without it, and without proper documentation and maintenance of legacy code and applications, all of this falls apart. Researchers should be responsible for delivering usable, user-friendly, well documented and well maintained code, in order for their experiments and research to be reproducible.

Acknowledgements

Special thanks to Fiona Sarola, Biel González, Adarsh Tiwari, Aleix Antón, Mireia Majó and Èric Sánchez.

Also, thanks to the Computer Vision Center - UAB, without their facilities, materials, tools, hardware, and awesome personnel, none of this would have been possible.

Thanks to all the volunteers who participated in the experiments, and thanks to Alejandro Parraga, Lara Blasco, Roger Blasco, Manel-Óscar Blasco, Xim Cerdà-Company, Anna Coderch, Tania Grijalba, Sara Kelley, Xavier Otazu, Xavier Pons and Eva Roca.

I am also thankful to all the investigators and researchers who publicly share and maintain their code.

REFERENCES

[1] S. Luka J. Ferweda. "A high resolution, high dynamic range display system for vision research". In: *Journal of vision* 9.8 (2009), 346a. DOI: 10.1167/9.8.346.

[2] Reinhard et al. *High Dynamic Range Imaging: Acquisition, Display and Image-Based Lighting*. Vol. 1st ed. chap 6. Morgan Kaufmann Publishers Inc., 2005, pp. 187–221.

[3] T. Troscianko R. Snowden P. Thompson. *Basic Vision: an Introduction to Visual Perception*. Oxford University Press, 2006.

[4] C. Mees. *The fundamentals of photography*. Vol. 2nd ed. Eastman Kodak Company, 1921.

[5] Oscar Gustave Rejlander. *The Two Ways of Life*. [Combination Print montage in black and white]. 1857. URL: https://en.wikipedia.org/wiki/Combination_printing#/media/File:Oscar-gustave-rejlander_two_ways_of_life.jpg.

[6] Henry Peach Robinson. *Pictorial Effect in Photography: Being Hints on Composition and Chiaro-oscuro for Photographers. To which is Added a Chapter on Combination Printing*. Piper Carter, 1869.

[7] A. Rizzi J. McCann. *The Art and Science of HDR Imaging*. Vol. 1st ed. chap 13. WILEY, 2012, pp. 119–121.

[8] H. Barlow. "Summation and inhibition in the frogs retina". In: *The Journal of Physiology* 119.1 (1953), pp. 69–88. DOI: 10.1113/jphysiol.1953.sp004829.

[9] P. Lennie A. Derrington J. Krauskopf. "Chromatic mechanisms in lateral geniculate-nucleus of macaque". In: *The Journal of Physiology* 357 (1984), pp. 241–265. DOI: 10.1113/jphysiol.1984.sp015499.

[10] F. Campbell C. Blakemore. "n the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images". In: *The Journal of Physiology* 203.1 (1969), pp. 237–260. DOI: 10.1113/jphysiol.1969.sp008862.

[11] Edwin H. Land. "THE RETINEX". In: *American Scientist* 52.2 (1964), pp. 247–264.

[12] John J. McCann Edwin H. Land. "Lightness and Retinex Theory". In: *Journal of the Optical Society of America* 61.1 (1971), pp. 1–11. DOI: 10.1364/JOSA.61.000001.

[13] John J. McCann. "Capturing a black cat in shade: past and present of Retinex color appearance models". In: *Journal of Electronic Imaging* 13 (2004), pp. 36–47. DOI: 10.1117/1.1635831.

[14] John J. McCann. "Lessons Learned from Mondrians Applied to Real Images and Color Gamuts." In: vol. 14. 1999, pp. 1–8.

[15] John J. McCann. "Retinex at 50: Color theory and spatial algorithms, a review". In: *Journal of Electronic Imaging* 26 (Feb. 2017). DOI: 10.1117/1.JEI.26.3.031204.

[16] Vanrell Maria Otazu Xavier Párraga C. Alejandro. "Multiresolution wavelet framework models brightness induction effects". In: *Vision Research* 48.5 (2008), pp. 733–751. DOI: https://doi.org/10.1016/j.visres.2007.12.008.

[17] Vanrell Maria Otazu Xavier Párraga C. Alejandro. "Toward a unified chromatic induction model". In: *Journal of Vision* 10.12 (2010), pp. 5–5. DOI: 10.1167/10.12.5.

[18] John McCann, Carinna Parraman, and Alessandro Rizzi. "Reflectance, illumination, and appearance in color constancy". In: *Frontiers in Psychology* 5 (2014). DOI: 10.3389/fpsyg.2014.00005. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00005.

[19] Li et al. "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.12 (2022), pp. 6999–7019. DOI: 10.1109/TNNLS.2021.3084827.

[20] Ruohui Wang. "Edge Detection Using Convolutional Neural Network". In: *Advances in Neural Networks – ISNN 2016* (2016), pp. 12–20. DOI: DOI:10.1007/978-3-319-40663-3_2.

[21] Ruppertsberg et al. "Displaying colourimetrically calibrated images on a high dynamic range display". In: *Journal of Visual Communication and Image Representation* Vol.18.No.5 (2007), pp. 429–438.

[22] Aydin et al. "Dynamic Range Independent Image Quality Assessment". In: 27.3 (2008), pp. 1–10. ISSN: 0730-0301. DOI: 10.1145/1360612.1360668. URL: https://doi.org/10.1145/1360612.1360668.

[23] Hojatollah Yeganeh and Zhou Wang. "Objective Quality Assessment of Tone-Mapped Images". In: *IEEE Transactions on Image Processing* 22.2 (2013), pp. 657–667. DOI: 10.1109/TIP.2012.2221725.

[24] Drago et al. "Perceptual Evaluation of Tone Mapping Operators with Regard to Similarity and Preference". In: (Jan. 2002).

[25] Kuang et al. "Testing HDR Image Rendering Algorithms." In: *Final Program and Proceedings - IS and T/SID Color Imaging Conference* (Jan. 2004), pp. 315–320.

[26] Ledda et al. "Evaluation of tone mapping operators using a High Dynamic Range display". In: *ACM Trans. Graph.* 24 (July 2005), pp. 640–648. DOI: 10.1145/1073204.1073242.

[27] Yoshida et al. "Perceptual Evaluation of Tone Mapping Operators with Real-World Sceness". In: *Human Vision and Electronic Imaging X, IST/SPIE's 17th Annual Symposium on Electronic Imaging (2005), SPIE, 192-203 (2005)* 5666 (Mar. 2005). DOI: 10.1117/12.587782.

[28] Xavier Otazu Xim Cerda-Company C. Alejandro Parraga. "Which tone-mapping operator is the best? A comparative study of perceptual quality". In: *Journal of the Optical Society of America A* 35.4 (Apr. 2018), pp. 626–638. DOI: 10.1364/JOSAA.35.000626.

[29] Akiko Yoshida et al. "Testing tone mapping operators with human-perceived reality". In: *Journal of Electronic Imaging, v.16, 1-14 (2007)* 16 (Jan. 2007). DOI: 10.1117/1.2711822.

[30] Goyal Jay Ashikhmin Michael. "A Reality Check for Tone-Mapping Operators". In: *ACM Transactions on Applied Perception* 3.4 (2006), pp. 399–411. ISSN: 1544-3558. DOI: 10.1145/1190036.1190040.

[31] Kuang et al. "Evaluating HDR Rendering Algorithms". In: *ACM Transactions on Applied Perception* 4.2 (2007), 9–es. DOI: 10.1145/1265957.1265958.

[32] Kuang et al. "iCAM06: A refined image appearance model for HDR image rendering". In: *Journal of Visual Communication and Image Representation* 18 (2007), pp. 406–414. DOI: 10.1016/j.jvcir.2007.06.003.

[33] Cadík et al. "Image Attributes and Quality for Evaluation of Tone Mapping Operators". In: 14th Pacific Conference on Computer Graphics and Applications, 2006.

[34] Cadík et al. "Evaluation of HDR tone mapping methods using essential perceptual attributes". In: *Computers Graphics* (2008), pp. 330–349. DOI: 10.1016/j.cag.2008.04.003.

[35] Akyüz et al. "Do HDR Displays Support LDR Content? A Psychophysical Evaluation". In: *ACM Transactions on Graphics* 26.3 (2007), 38–es. DOI: 10.1145/1276377.1276425.

[36] Creswell et al. "Generative Adversarial Networks: An Overview". In: *IEEE Signal Processing Magazine* 35.1 (2018), pp. 53–65. DOI: 10.1109/MSP.2017.2765202.

[37] Rana et al. "Deep Tone Mapping Operator for High Dynamic Range Images". In: *IEEE Transactions on Image Processing* 29 (2019), pp. 1285–1298. DOI: 10.1109/TIP.2019.2936649.

[38] Oppo. *Oppo's product website*. URL: https://www.oppo.com/es/smartphones/series-find-x/find-x3-neo/specs/. (accessed: 9.1.2024).

[39] Blasco Roca Pau. *Kaminari-TMO*. URL: https://github.com/Nerocraft4/KaminariTMO. (accessed: 11.1.2024).

[40] Yang et al. "Clustering based content and color adaptive tone mapping". In: *Computer Vision and Image Understanding* 168 (2018), pp. 37–49. DOI: 10.1016/j.cviu.2017.11.001.

[41] Jan Kautz Min H. Kim. "Consistent Tone Reproduction". In: *"Proceedings of Computer Graphics and Imaging* (2008), pp. 152–159. DOI: 10.5555/1722302.1722332.

[42] H. Seidel G. Krawczyk K. Myszkowski. ""Lightness perception in tone reproduction for high dynamic range images". In: *Proceedings of Eurographics* 24.3 (2005). DOI: 10.1111/j.1467-8659.2005.00888.x.

[43] Liang et al. "A Hybrid l1-l0 Layer Decomposition Model for Tone Mapping". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018). DOI: 10.1109/CVPR.2018.00500.

[44] Seong-O. Shim Ishtiaq Rasool Khan Wajid Aziz. "Tone-Mapping Using Perceptual-Quantizer and Image Histogram". In: *IEEE Access* 8 (2020), pp. 31350–31358. DOI: 10.1109/ACCESS.2020.2973273.

[45] Reinhard et al. "Photographic Tone Reproduction for Digital Images". In: *ACM Transactions on Graphics* 21.3 (2002), pp. 267–276. DOI: 10.1145/566654.566575.

[46] Ishtiaq et al. "A tone-mapping technique based on histogram using a sensitivity model of the human visual system". In: *IEEE Transactions on Industrial Electronics* 65.4 (2017), pp. 3469–3479. DOI: 10.1109/TIE.2017.2760247.

[47] Le et al. "Computational Analysis of Correlations between Image Aesthetic and Image Naturalness in the Relation with Image Quality". In: *Journal of Imaging* 8.6 (2022). DOI: 10.3390/jimaging8060166.

[48] Francesco Banterle. *HDR Toolbox*. URL: https://github.com/banterle/HDR_Toolbox. (accessed: 31.12.2023).

[49] Wikipedia. *ColorOS*. URL: https://en.wikipedia.org/wiki/ColorOS. (accessed: 10.1.2024).

[50] Mori et al. "The Uncanny Valley [From the Field]". In: *IEEE Robotics Automation Magazine* 19.2 (2012), pp. 98–100. DOI: 10.1109/MRA.2012.2192811.

[51] Baek-Kyu et al. "Tone mapping with contrast preservation and lightness correction in high dynamic range imaging". In: *Signal, Image and Video Processing* 10 (2016), pp. 1425–1432. DOI: 10.1007/s11760-016-0942-1.

[52] Cong et al. "Unsupervised HDR Image and Video Tone Mapping via Contrastive Learning". In: (2023). eprint: 2303.07327.

[53] Eilertsen et al. "Real-Time Noise-Aware Tone Mapping". In: *ACM Trans. Graph.* 34.6 (2015). ISSN: 0730-0301. DOI: 10.1145/2816795.2818092.

[54] Panetta et al. "TMO-Net: A Parameter-Free Tone Mapping Operator Using Generative Adversarial Network, and Performance Benchmarking on Large Scale HDR Dataset". In: *IEEE Access* 9 (2021), pp. 39500–39517. DOI: 10.1109/ACCESS.2021.3064295.

[55] Yang et al. "Deep Reformulated Laplacian Tone Mapping". In: *Electrical Engineering and Systems Science* (2021). DOI: 10.48550/arXiv.2102.00348.

[56] Yang et al. "Multi-Scale histogram tone mapping algorithm for display of wide dynamic range images". In: Dec. 2017, pp. 1–5. DOI: 10.1109/CAMSAP.2017.8313070.

[57] Zhang et al. "Multiscale Morphological Tone Mapping Operator for High Dynamic Range Images". In: *Proceedings of the Third International Symposium on Image Computing and Digital Medicine*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 254–258. ISBN: 9781450372626. DOI: 10.1145/3364836.3364887.

[58] Xiuhua Jiang Cheng Guo. "Deep Tone-Mapping Operator Using Image Quality Assessment Inspired Semi-Supervised Learning". In: *IEEE Access* 9 (2021), pp. 73873–73889. DOI: 10.1109/ACCESS.2021.3080331.