RV-TMO: Large-Scale Dataset for Subjective Quality Assessment of Tone Mapped Images

Ali Ak, Abhishek Goswami, Wolf Hauser, Patrick Le Callet, Frédéric Dufaux

RV-TMO: Large-Scale Dataset for Subjective Quality Assessment of Tone Mapped Images

Ali Ak*, Abhishek Goswami†‡, Wolf Hauser†, Patrick Le Callet*, and Frederic Dufaux‡
*Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
Emails: {ali.ak, patrick.lecallet}@univ-nantes.fr
†DxO Labs, France
Emails: {whauser}@dxo.com
‡Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, France
Emails: {abhishek.goswami, frederic.dufaux}@l2s.centralesupelec.fr

Abstract—Tone mapping operators (TMO) are functions that map high dynamic range (HDR) images to a standard dynamic range (SDR), while aiming to preserve the perceptual cues of a scene that govern its visual quality. Despite the increasing number of studies on quality assessment of tone mapped images, current subjective quality datasets have relatively small numbers of images and subjective opinions. Moreover, existing challenges in transferring laboratory experiments to crowdsourcing platforms put a barrier for collecting large-scale datasets through crowdsourcing. In this work, we address these challenges and propose the RealVision-TMO (RV-TMO), a large-scale tone mapped image quality dataset. RV-TMO contains 250 unique HDR images, their tone mapped versions obtained using four TMOs and pairwise comparison results from seventy unique observers for each pair. To the best of our knowledge, this is the largest dataset available in the literature for quality evaluation of TMOs by the number of tone mapped images and number of annotations. Furthermore, we provide a content selection strategy to identify interesting and challenging HDR images. We also propose a novel methodology for observer screening in pairwise experiments. Our work does not only provide annotated data to benchmark existing objective quality metrics, but also paves the path to building new metrics for tone mapping quality evaluation.

Index Terms—image quality evaluation, pairwise comparison, tone mapping operators, crowdsourcing

1 Introduction

High dynamic range images have gained popularity over the last decade, owing to the advancement in image acquisition and display technologies. Modern image sensors with additional techniques, such as exposure bracketing and fusion, allow us to capture high dynamic range (HDR) images with ease. HDR contents provide a more accurate and aesthetic representation of the real world by utilizing a wider range of color and luminance. HDR capable displays have also become widely available in the consumer market. Despite this growing interest, most images are still consumed as low dynamic range (LDR) media, which are represented by eight bits per pixel per channel. Converting HDR contents into pleasing LDR representations often requires employing a tone mapping operator (TMO) or remastering the content by a professional artist. Due to the subjective nature of the tone mapping task, results may vary greatly between TMOs and professional artists.

Quality assessment of tone mapped images is of considerable interest as a result of the increased adoption of HDR technologies and the inevitable need for TMOs. Although quite costly and time consuming, subjective studies are considered as the most reliable way for tone mapped image quality assessment (IQA). Moreover, they provide the necessary data to benchmark objective quality metrics and develop new ones. Various publicly available tone mapped IQA datasets can be found in the literature [1], [2], [3], [4], [5]. However, several of these databases contain low numbers of images and of subjective annotations. A summary of the available datasets is provided in Table 1. Previously, Kundu et al. [2] proposed the largest publicly available dataset till date also containing tone mapped IQA. Although the dataset contains a relatively high number of HDR images, as stated by the authors, the aim of their study was not to evaluate TMO performances nor benchmarking tone mapped IQA metrics. Therefore, by design, most of the tone mapped images are generated from different HDR images making it impossible to compare TMO performances.

Similar to [2], other crowdsourced IQA studies can be found. The last decade brought a surge in the popularity of crowdsourcing platforms like Prolific [6] and Amazon Mechanical Turk (AMT) [7]. They indeed allow researchers to conduct large-scale subjective experiments within a short amount of time, and with reduced cost and effort. Additionally, they provide a large and varied demography thanks to their wide participant pools. Despite the fact that crowdsourcing platforms bring many advantages, uncontrolled experimental conditions may lead to noisy data. A recent study [8] showed that, with proper experimental design, such challenges can be overcome and tone mapped IQA experiments can be conducted on crowdsourcing platforms, more specifically over Prolific [6]. Results indicated that the differences between subjective annotations acquired from laboratory and crowdsourcing experiments are negligible. Moreover, the same level of certainty in subjective annotations acquired from a laboratory experiment can be achieved via crowdsourcing.

As suggested in [8], pair comparison (PC) methodology is a favorable design choice for crowdsourced tone mapped IQA. PC design simplifies the task for observers by asking for a binary preference between two images, therefore eliminating observer's bias based on the understanding of scales in rating tasks. It is also

argued to be more suitable for real-world use cases [9]. One major shortcoming of the PC methodology is the lack of well-established observer screening tools in the literature, mainly due to the binary nature of the subjective annotations collected via PC experiments.

Another important aspect of a large-scale tone mapped IQA dataset is linked to content selection. A proper content selection strategy should ensure a diverse and well represented dataset. In PC tone mapped IQA experiments, stimuli are in the form of a tone mapped image pair. Certain pairs may be more ambiguous than others, *e.g., for a given pair [A, B], some of the observers may prefer image A whereas others may prefer image B*. Therefore, a diverse dataset with PC methodology should contain pairs with varying ambiguity. Having a dataset with only obvious pairs brings negligible benefit to benchmarking and metric development. Current content selection strategies aim to diversify images in datasets with features such as spatial information [10], colorfulness [10], etc. Although these may ensure a diverse dataset in terms of image characteristics, they have no implication on the ambiguity distribution of the image pairs in a PC experiment.

To address these challenges and limitations, we conducted a large-scale crowdsourced subjective experiment to assess the aesthetic quality of tone mapped images. Our contributions are as follows:

- To address the lack of large-scale datasets for TMO evaluation, we present the *RealVision-TMO (RV-TMO)*[1] dataset containing 250 HDR images, each tone mapped with four different state-of-the-art TMOs, as well as their respective subjective annotations using a full-PC methodology. 1500 unique image pairs, each evaluated by seventy unique observers over Prolific crowdsourcing platform, and a total pool of 3500 unique observers attending the experiment make this the largest publicly available dataset for quality assessment of tone mapped images, to the best of our knowledge.
- To address the lack of content selection strategies ensuring pairs with varying ambiguity in PC experiments, we developed a content selection strategy that provides representative HDR images and tone mapped image pairs with varying ambiguity. The proposed strategy provides a challenging dataset that can be used to benchmark objective IQA models and develop new objective quality metrics.
- To address the lack of well-established observer screening tools in PC experiments, we propose a novel approach to analyze the collected subjective pairwise preferences in order to assess observer reliability and reject observers with undesired behaviors.
- Finally, we analyze the performance of existing state-of-the-art IQA metrics for tone mapped images on the collected dataset. Furthermore, we provide recommendations and tools towards objective tone mapped IQA metric development and benchmarking.

2 Related Work

Crowdsourcing is relatively new in the quality of experience (QoE) domain. Indeed, even though crowdsourcing platforms provide researchers with a wider audience, faster turnover, and reduced costs, they also bring additional challenges which differ from traditional laboratory experiments. Qualinet whitepaper [11] discusses these benefits and challenges from the QoE point of view. An early example of subjective IQA on crowdsourcing shows promise by comparing crowdsourcing and laboratory experiment results [12]. Recent works raise concerns on the effects of QoE tasks on crowdsourcing subjective experiments [13]. LIVE In the Wild [14] IQA dataset consists of over 350000 opinion scores on 1162 images. More than 8000 unique participants attended the subjective study to evaluate the quality of images containing a wide set of distortions.

TABLE 1
An overview of tone mapped image quality datasets available in the literature. Note that the dataset proposed by Kundu et al. [2] also contains images processed with multi-exposure fusion (MEF) algorithms and photographic effects which were excluded from the summary made in this table.

	Method	SRC	TMO	Total images	Obs. per stim	Total annotations
RV-TMO	PC	250	4	1000	70	105000
Krasula et al. [1]	PC	20	9	180	20	-
Yeganeh et al. [3]	Ranking	15	8	120	20	-
Kuang et al. [4]	PC	10	8	80	30	8400
Ledda et al. [5]	PC	23	6	138	48	15660
Kundu et al. [2]	SSCQS	605	4	747	110	75000

Several tone mapped IQA datasets collected in controlled laboratory environments are publicly available in the literature. A comprehensive review of existing works can be found in [15], and a summary of tone mapped IQA datasets is presented in Table 1. Columns in the table represent the methodology used by each experiment, number of source content (SRC), number of TMOs applied, total number of tone mapped images, number of unique observers per stimulus, and total number of annotations, respectively. Krasula et al. [1] conducted two separate subjective experiments to measure the effects of having the reference HDR scene on observer preferences. The experiment was conducted in a controlled laboratory environment with twenty HDR images (ten real world, and ten synthetic scenes), tone mapped with nine different TMOs. Twenty naive observers participated in the experiment. In an earlier study [3], Yeganeh et al. conducted another subjective experiment in a controlled laboratory environment to evaluate the objective IQA performances on fifteen HDR images, each one tone mapped with eight different TMOs. Each stimulus is rated by around twenty observers. The dataset from Kuang et al. [4] contains eighty tone mapped images generated from ten HDR images with eight TMOs in a PC experiment conducted in a laboratory with thirty participants. The dataset by Ledda et al. [5] also uses the PC design with twenty-three SRC and six TMOs. Forty-eight unique participants evaluated the image pairs in two sessions under a controlled environment, resulting in 15660 total annotations. As observed, most of the existing datasets collected in controlled laboratory environment suffer from a relatively low number of HDR images.

To the best of our knowledge, only one work on subjective quality evaluation of tone mapped images has been carried out using crowdsourcing. In their work, Kundu et al. [2] conducted a subjective experiment on the AMT platform with more than 5000 observers on 605 HDR images. In Table 1, we only included the TMO part of this dataset since the full dataset does not only contain tone mapped images, but also HDR images processed with multi-exposure fusion (MEF) and visual effect algorithms. As shown in the table, despite having 605 HDR images in the

1. The dataset is available at ftp://ftp.polytech.univ-nantes.fr/RV-TMO

dataset, there are only 747 tone mapped images. The aim of the study, as expressed by the authors, was not to evaluate the TMO performances. Therefore, most of the HDR content were tone mapped using only one TMO.

3 Content Generation

We hypothesize that an HDR image can be characterized by several image features. Therefore, it is essential to compile a dataset of HDR images that cover a significant portion of the image features space. Furthermore, it will help us identify whether certain TMOs enhance certain aesthetic attributes, thereby influencing subjective preference. In the following subsections, we present our proposed strategies to generate meaningful image data.

In our search for available high-resolution HDR images in the literature, two large datasets stood out; Fairchild et al's *HDR Photographic Survey* [16] and Artusi et al. [17]. Both datasets contain images of high to very high spatial resolutions.

Our search uncovered that 1080p full HD (FHD) displays are the most common commercially available and accessible form of displays. Considering our subjective experiment is crowdsourced, we have aimed to create content that is accessible to participants. Furthermore, to ensure that the display devices do not interpolate image content and display the content without any scaling, we decided to use a 480p spatial resolution, such that a 1080×1920 pixel display can present two tone mapped images side by side in landscape mode with a gray space in between. To utilize as much information from the HDR images, and for consistency of operations, we adopted a method of systematically scaling and cropping the full resolution images to a 480p resolution. This allows to not only create a dataset tailored to be used for future learning-based approaches, but it also increases the number of images by natural augmentation.

3.1 Scale and Crop

Fairchild and Artusi datasets respectively contain 105 and 124 images, for a total of 229 high-resolution HDR images. Our content selection strategy involves a process of iterative downscaling of the original image by a factor of 2, 4, and 8; and successive uses of sliding-window with 100 px shifts to generate 480p px crops at each iteration.

Figure 1 illustrates our scale-crop strategy, where each scale corresponds to the factor by which the spatial resolution was scaled down from the HDR image original size. It can be observed that higher scales provide meaningful crops more often and may help reduce homogeneous, redundant, or less spatially informative crops. At each scale, we applied a sliding window crop of resolution 480×640 with a stride of 100 px. Following our strategy over 229 full-resolution HDR images, we computed 167100 candidate crops of 480p resolution. Next, we identified certain image features from each crop and assigned a score based on the extracted features to help with further filtering.

3.2 Feature Extraction

We extracted a set of six perceptual and objective features from the HDR images. We hypothesize that a combination of these features is a good indicator of whether an HDR crop can provide valuable and interesting information necessary for tone mapping evaluation. Each feature objectively provides some information about the crop to classify it as an informative 'good' crop. Our crop selection strategy aims to widen the distribution of such good crops in the aforementioned six-dimensional feature space.

- **ADR** (r)**: A**bsolute **D**ynamic **R**ange of the HDR crop. The ADR helps us identify crops with exposure variations and helps to filter out mostly homogeneous crops.
- **SD** (σ)**: S**tandard **D**eviation of the luminance of the HDR crop. The standard deviation value for the crop can help identify wether a certain patch belongs to a homogeneous region (like a blank patch of sky).
- **MLE** (m)**: M**ulti-**L**evel **E**ntropy of the saliency map of the crop. We use minimum barrier saliency detection [18] to generate a saliency map for each crop. To approximate the information provided by the saliency map, we compute a multilevel entropy (Depth = 4) for each crop [19]. This provides an intuition whether the crops are informative or have salient objects.
- **Scale priority** (s)**:** The scale of the crop (i.e., 2, 4, or 8). As discussed previously, the quality of the crop significantly depends on the scale of the crop. Higher scale crops have the chance to include more spatial information due to the scale-crop technique. Although we generate crops for each scale, we provide simple weights to prioritize the scales.
- **Objective mean score** (O)**:** The mean TMQI [3] score across three tone mapped versions of the crop. Each crop is tone mapped by three state-of-the-art classical TMOs, i.e., *KimKautzTMO* [20], *KrawczykTMO* [21] and *ReinhardTMO* [22]. Our aim, on top of compiling a dataset, is to identify a correlation between subjective and objective assessment of tone mapped images. Mean objective scores help identify crops where the tone mapped versions have acceptable visual quality according to the objective IQA metric.
- **Objective disagreement score** (ΔO)**:** A score representing the difference of tone mapping objective quality across the three aforementioned TMOs. Disagreement scores highlight challenging crops that are difficult to tone map, leading to variation among TMQI scores of the tone mapped crops. In other words, it helps identify HDR crops which generate tone mapped image pairs with varying ambiguity.

Each attribute provides certain understanding about how interesting and informative a crop is. They aim to differentiate between redundant crops, or ones without valuable spatial information, and crops with salient features or variable dynamic range. Following the parameter fusion strategy of Krasula et al. [23], we decided to compute an affine combination of our features to get *Crop Scores*. However, we realised that each feature except scale has a different range of possible values. Consequently, we computed the histogram for each feature and we scaled them such that the feature values lied within the range of their 1^{st} and 99^{th} percentile, followed by normalisation to the range [0, 1]. Normalized attribute values contribute to the crop scores. Hence, for each crop, we get a crop score:

$$S = \frac{s}{4} + \hat{O} + \hat{\Delta O} - \hat{m} + \hat{r} + \hat{\sigma}, \tag{1}$$

where $\hat{\square}$ represents the clipped and normalized feature values for each crop. The scale s is weighted so that higher scale crops are prioritized. For MLE, a negative sign is introduced as we prefer a lower MLE score. This affine combination is empirical; it is not an exhaustive approach to compute crop scores. However,

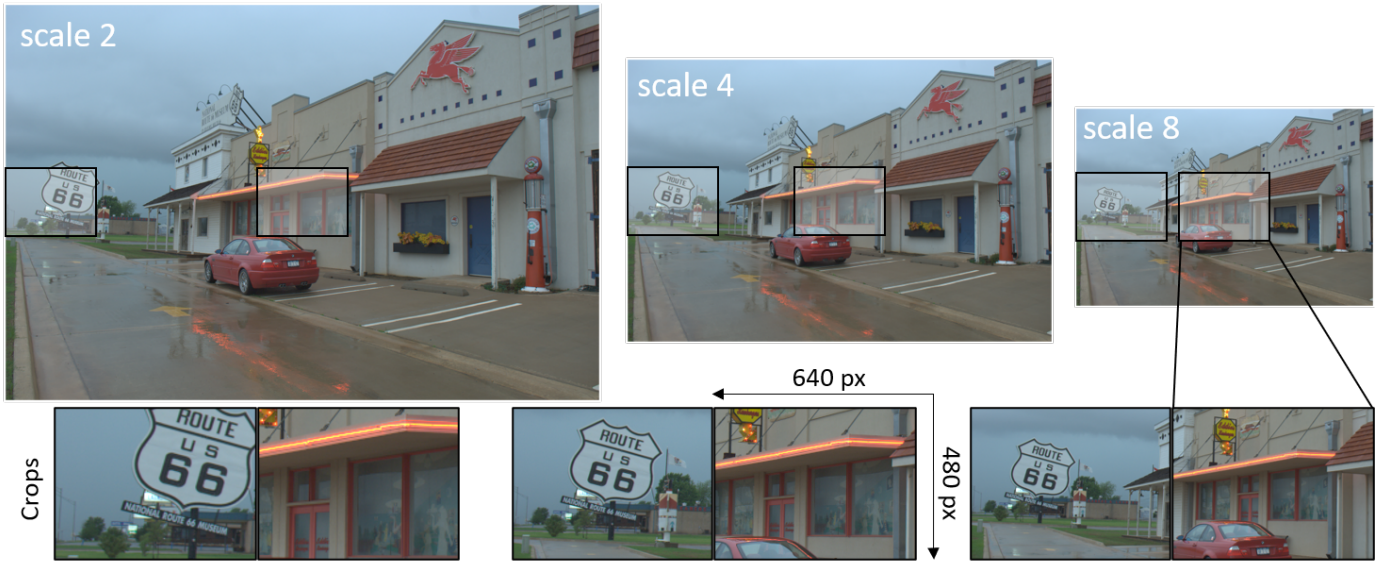

Fig. 1. Scale and Crop Strategy shows *66 Museum* image taken from Fairchild's HDR dataset [16]. The objective behind the strategy is natural augmentation of images for the dataset. The strategy creates three versions of each image, scaling the original size down by a factor of $2, 4$ and 8 respectively. Consequently, for each version, a sliding-window crop of size $480 \times 640px$ is created with a horizontal and vertical stride of $100px$.

the combination of attributes plays a vital role in prioritizing certain crops over others. In our experiment, we found that the crop scoring technique helps distinguish between interesting and non-informative crops.

It is important to note that, while computing the scores, we consider the tone mapping quality of only three TMOs and their disagreements. As a reminder, the choice of the three TMOs was based on a previous comparative study evaluating tone mapping quality across different use cases [24]. KimKautzTMO [20], KrawczykTMO [21], and ReinhardTMO [22] were validated by different research works and shown to produce better overall tone mapping quality in several experiments. Hence, the attribute scores corresponding to the three TMOs were considered in our strategy. We do not consider SemTMO because of two reasons. Primarily, the other three TMOs are established and widely researched upon in literature in comparison to SemTMO, which follows a novel semantic-based approach. Furthermore, unlike the other three TMOs, SemTMO is not well optimized for real-time tone mapping and presents a significant time complexity to tone map 167100 candidate crops. Therefore, we decided to rely on the three established TMOs to determine the agreement and disagreement of quality while computing scores for each crop.

3.3 Overlapping crops

We observe that crops collected from the same spatial neighborhood of an HDR image using our sliding window strategy get similar scores. To reduce the redundancy of overlapping regions among candidate crops, we first apply a threshold of spatial overlap percentage. Consequently, we prioritize based on the score and the scale of the crop which allows to remove redundant candidate crops across all scales for an HDR image. We empirically set the spatial overlap threshold at 60%. Figure 2 presents an example image with four crops. Crop 1 is the reference, while crops 2, 3, and 4 are generated with 80%, 60%, and 40% overlap, respectively. As observed, crop 2 is very similar to the reference crop in the example. Including both of these crops in the dataset brings negligible benefit. On the contrary, crops 2 and 3 provide relatively different images in comparison to the reference.

As an example, when we encounter two crops with an overlap greater than the threshold, if they belong to the same scale, the crop with the higher score is chosen. If they belong to different scales, the candidate with the higher score is preferred unless candidates belong to scales 2 and 8, where a candidate with scale 8 is preferred over scale 2, irrespective of score difference. The reason behind such preference is the more information and natural framing that a higher-scale crop provides. Figure 1 shows that, for a same region, a higher scale crop incorporates more information in the scene due to its scale. Consequently, a higher-scale crop has a higher potential to be more interesting and challenging because of the extra information.

3.4 Clustering based on TMQI scores

Redundancy removal from the 167100 candidate crops leaves us with 19540 crops. We observe that, although spatial redundancy were removed, crops vary in terms of how challenging or informative they are, as suggested by crop scores computed for each crop. The crop scores do not have a fixed bound, but a higher value suggests a more interesting crop. In our case, we find that crop scores lie in $[0.02, 3.85]$. We empirically put a threshold of 1.5 to the score, and select all candidate crops above this threshold. This step provides 9730 candidate crops that are deemed to be interesting, challenging, and informative.

Consequently, we cluster the selected crops on the basis of their objective IQA scores (TMQI [3] scores of each crop, tone mapped by three TMOs, i.e., KimKautzTMO [20], Krawczyk [21], and ReinhardTMO [22]). Each crop can be represented in a three-dimensional space, with their three IQA scores as coordinates. The objective of the task is to identify crops that provide variety in their objective assessment. We observe cases where all three tone mapped versions of an HDR crop are rated highly by TMQI. Conversely, in some other cases, TMQI suggests that one tone mapped crop is significantly better or worse than the others. We group our selected candidates into five clusters—one

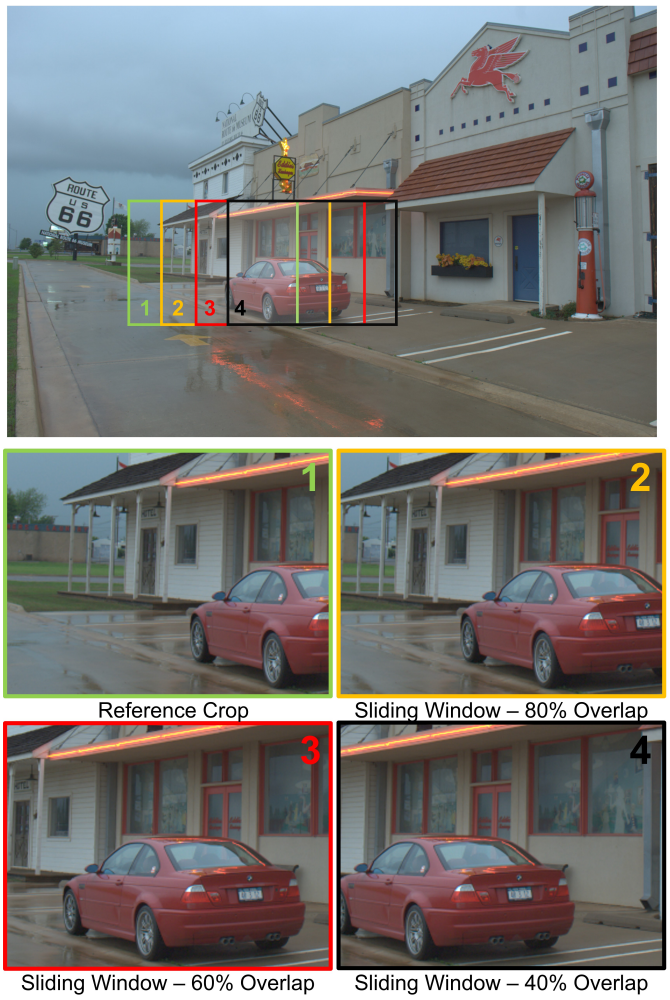

Fig. 2. Sample crops with overlap ratio of 40%, 60%, and 80%. Empirically, we chose to accept crops with less than 60% overlap, as higher overlaps may result in similar, redundant images.

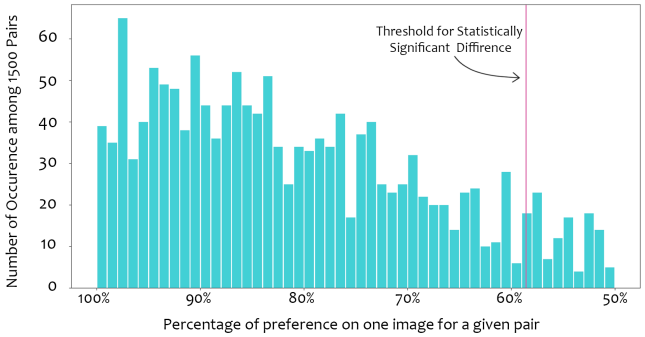

Fig. 3. Histogram of pairwise preference percentages for all 1500 pairs in our dataset. The vertical axis represents the number of pairs that belong to each bin in the histogram, while the horizontal axis represents the percentage of observer preferences for each pair. The vertical magenta line divides the pairs into two groups, where pairs on the left side show statistically significant differences.

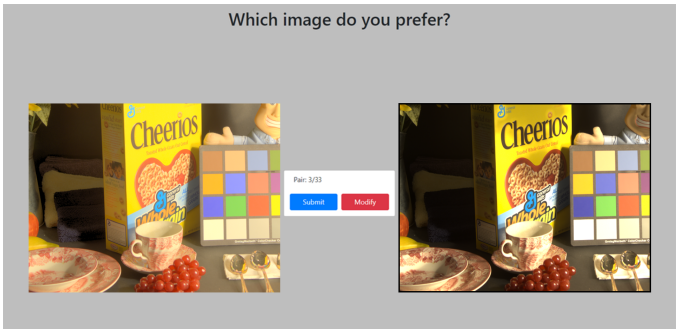

Fig. 4. Sample test screen from the experiment displaying the stimulus presented to observers. Observers are provided with a pair of tone mapped images and asked to choose the one they prefer.

where all three tone mapped crops have high scores, one where all have poor scores, and three other clusters where one tone mapped crop is better than the other two. Finally, we randomly select fifty crops from each cluster to produce a large dataset of 250 SRC HDR crops. Our selection procedure aims to maximize the distribution of crops in the aforementioned feature space. The clustering approach aims to have a diverse ambiguity on the resulting tone mapped image pairs.

3.5 Validation of Content Selection Strategy

The content selection strategy aims to identify HDR crops that can provide image pairs with varying ambiguity. We hypothesize that, by clustering the candidate HDR crops on a three-dimensional space where each axis represents the TMQI scores of a tone mapped image, we can select a subset of HDR crops that will generate pairs with a wide variety of ambiguity. Collecting a dataset with obvious pairs (i.e., where a majority of observers chooses image A over image B in a pair [A, B]) does not bring any value into benchmarking tone mapped IQA metrics, nor for the development of new metrics.

With this in mind, validating our approach is a straightforward procedure after collecting pairwise preferences. Figure 3 depicts the distribution of pairwise preferences in terms of percentages (for a detailed explanation of how to acquire pairwise preferences, please refer to Section 5.1). The vertical axis represents the number of pairs for each bar. The horizontal axis represents the pairwise preference percentages. More specifically, 100% represents the pairs for which every observer preferred the same image, whereas 50% represents the pairs where half of the observers prefer one tone mapped image while the other half prefers the other. In other words, the ambiguity of the pairs increases from left to right on the horizontal axis. We can observe that a balanced distribution of ambiguity exists in the dataset despite being not perfectly uniform. The dataset provides some obvious pairs (i.e., where a majority of observers chooses the same image in a pair), as well as ambiguous pairs. This validates the initial motivation behind using clustering in TMQI metric score space. By providing pairs with varying ambiguity, we aim to provide a more challenging dataset for benchmarking existing tone mapped IQA metrics, as well as a more representative dataset for developing new metrics.

The approximate threshold for statistically significant difference (for a detailed explanation of the statistically significant difference, please refer to Section 5.1) is plotted as a vertical line on the plot. Pairs on the right-hand side of the threshold line show no statistically significant difference in terms of subjective preference. Due to the high number of unique observations per pair, the majority of the dataset presents a statistically significant

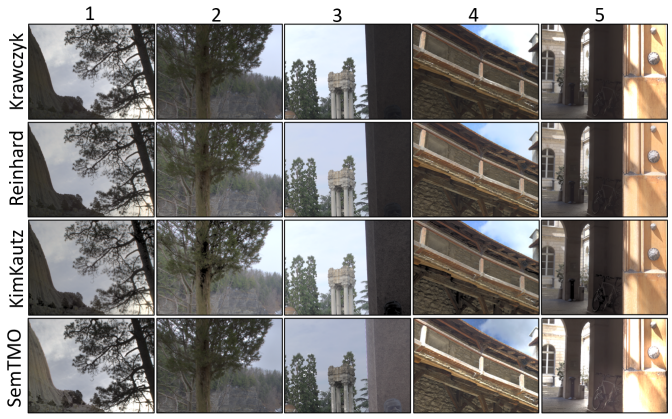

Fig. 5. Playlist #42 containing 20 tone mapped images over 5 unique SRCs. Each row shows the tone mapped images which are tone mapped by the TMOs indicated on the left. Additionally the playlist contains 3 pairs of 'golden unit' stimuli as described in Figure 6.

Fig. 6. Pair of images in each column were used as golden units. Golden units are stimuli for which the preference outcome is known beforehand. For each of the three pairs (column-wise), we expect images at the bottom to be preferred since images at the top are over exposed. Expected answers were confirmed by an in-lab study [8] prior to the dataset creation.

difference in pairwise preferences.

4 Subjective Experiment Design

In the following subsections, we describe in details the experimental design, generated dataset, information about participants, and crowdsourcing platform, as well as the strategies adopted to reject outliers and spammers.

4.1 Experiment Setup & Procedure

Subjective quality evaluation of tone mapped images can be conducted with either full-reference or no-reference methodologies. While full-reference comparison reveals information regarding the test image's fidelity to the HDR image, no-reference scenario reveals the overall aesthetic quality preferred by the observer [25]. In this experiment, we aim to collect aesthetic preferences among tone mapped images. Therefore, a no-reference experiment design is more suitable. A sample test screen can be seen in Figure 4.

4.2 Stimuli & Dataset

The stimuli presented to observers contain two tone mapped images where observers need to submit their preference among the two. The dataset compiled for the subjective study consists of 250 SRCs tone mapped using four TMOs, thus a total of 1000 tone mapped images. This results in 1500 stimuli where observers are asked to submit their preference in a pairwise fashion. For ease of experiments, we divided the dataset into fifty playlists of five SRCs each. Consequently, each playlist had thirty tone mapped image pairs to compare and three golden unit pairs (explained in Section 4.4). The four TMOs used consist of three classical TMOs deliberated by the literature, i.e., *KimKautzTMO* [20], *Krawczyk* [21], and *ReinhardTMO* [22], and a fairly recent semantic-aware operator, *SemanticTMO* [26].

Kim et al. [20] proposed a global TMO based on the log-luminance adaptation of the human visual cortex. As a local approach, Krawczyk et al. [21] introduced a TMO based on a probabilistic model of lightness perception. They decomposed an HDR image into areas of consistent luminance (lightness framework), and mapped each framework by adjusting the perceived 'white' point. Reinhard et al. [22] proposed a TMO considering the photographic practices based on eminent photographer Ansel Adams. Finally, we used SemanticTMO by Goswami et al. [26], which addresses tone mapping as a semantic-aware operation taking semantic labels and a corresponding specific target luminance into consideration.

4.3 Experiment Platform & Participants

We used the Prolific platform to recruit observers and to conduct the subjective experiment [6]. Contrary to other alternatives such as AMT and Microworkers, which try to make crowdsourcing platforms more accessible while compromising on ethical concerns and overall quality, Prolific focuses on the researchers' needs with a platform that maintains standards of recruitment similar to a laboratory experiment [27]. Participants are well informed that they are being recruited for a research study, and recruitment standards are set to benefit both researchers and participants [28]. Therefore, Prolific eliminates ethical concerns and significantly increases the collected data reliability.

In a previous study [8], it was shown that, through a PC experiment design for tone mapped IQA, the difference between subjective annotations acquired from a laboratory experiment and from an identical crowdsourced study on Prolific is negligible. The authors also recommend to have, for each pair, a minimum of fifty to sixty unique observations in crowdsourcing in order to reach the same level of certainty acquired through a laboratory experiment with thirty-five observers. Following these recommendations, we recruited 3500 unique participants where each participant evaluated thirty-three pairs. This allows us to have around seventy unique observations for each pair in our dataset. The majority of participants gave consent to share their demographic information. We have 2311 male participants with a mean age of 28.75 years and a standard deviation of 9.47. Similarly, we have 1154 female participants with a mean age of 31.54 years and a standard deviation of 10.83.

4.4 Rejection Strategies

The following subsections outline the strategies adopted to filter unreliable observer annotations and motivations behind them. Consequently, we report the number of rejected spammers based on each method.

Golden Unit: Golden unit is a filtering technique where a stimulus is presented for comparison with the preference known

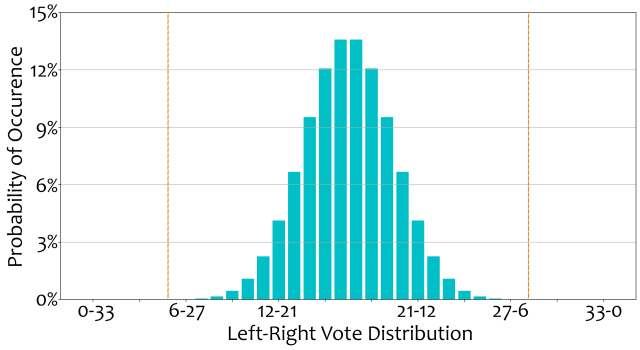

Fig. 7. Probability of occurrence of left-right voting distribution across 33 image pairs. *m-n* voting distribution denotes the observer prefers the left image *m* and the right image *n* times out of 33 comparisons. The x-axis provides the voting distributions and the y-axis provides the percentage of occurrence. Dashed lines represent the limit for rejection.

before the subjective experiment. Participants who provide preferences different from the prior are considered to be spammers. The reason for such behavior can be attributed to a lack of attention, or simply a random selection on the observer's part.

To select a set of golden units, we conducted a pilot subjective experiment in a controlled laboratory environment. As a result, we collected three pairs of tone mapped images where 100% of the participants provided the same preference. Figure 6 presents these golden unit pairs used in the experiment with their known prior preference. Preference towards strongly overexposed images, displayed in the top row, is considered as an indicator of unreliable behavior. Three golden units are included in all the fifty playlists of our experiments. The order of stimuli in each playlist is shuffled to prevent any bias towards the display order. Forty-nine participants out of 3500 failed the golden unit check at least once and therefore were rejected and new participants were recruited.

Vote Position Pattern: Previous studies suggest another behavior that can be observed in pairwise comparison experiments to filter unreliability. We can check for positional bias in terms of participants submitting preference for an image at the same position continually during the experiment. Since image positions are shuffled for each participant, we can calculate the probability that an observer votes for a fixed position. Figure 7 shows the probability distribution for each possible left-right vote share over thirty-three pairs. Orange dashed lines show the threshold for rejection. It can be observed that either position receiving less than six votes is statistically highly unlikely, with a probability of occurring once in 10000 participants. Therefore, among 3500 participants, we rejected thirteen participants who voted less than six times on one position.

Voting Speed: Spammers on online platforms tend to optimize their effort by finishing more tasks and hence minimizing the time spent on each task. We used the timestamp of observer votes to identify participants with unusually fast completion time, which indicates a possible spammer-like behavior, lack of attention, and probable noisy data. We observed an average time spent per pair of 4.08 seconds over the whole experiment. We identify 56 out of 3500 participants who completed the task with a median time of one second per pair, which is far from an expected speed. Consequently, their pairwise preferences are considered to be unreliable and not included in the final results.

Rogers-Tanimoto Dissimilarity: Behavioral analysis may reveal certain spammer profiles but it is not enough to identify all types of unreliable behaviors. After filtering unreliable observers with behavioral methods, statistical measures can be utilized to further improve the reliability of collected data. The literature lacks a well-established methodology to statistically measure the individual observer reliability for PC experiments. We propose a novel methodology based on Rogers-Tanimoto (RT) dissimilarity. Details of the approach are given in Section 5.2. It is shown that the efficiency of RT dissimilarity in measuring inter-observer agreement decreases with the increasing percentage of spammers among observers. Therefore, observers who are rejected with behavioral analysis are omitted from RT dissimilarity analysis.

TABLE 2
Each cell in the table denotes the percentage of pairs (among 250 total) for which the TMO on the row is significantly better than the TMO on the column.

	KimKautz	Krawczyk	Reinhard	SemTMO
KimKautz	-	**60%**	**42%**	**62%**
Krawczyk	19%	-	19%	**52%**
Reinhard	24%	**56%**	-	**62%**
SemTMO	19%	30%	20%	-

5 Analysis of Subjective Preferences

5.1 TMO Performances

In this section, we analyze the collected subjective preferences to evaluate the performance of tone mapping operators in comparison to each other. As previously described, 250 SRCs are tone mapped with four different TMOs, and each tone mapped image is compared in a pairwise fashion. Therefore, we can compare each TMO with the other ones for all compiled HDR contents. Figure 8 presents the result of this evaluation. Each row in the plot contains 250 data points which represent an SRC. It displays the preference in terms of percentage of observers. Points on the x-axis closer to one side of the y-axis indicate a higher preference towards corresponding TMOs on the y-axis. Additionally, the statistical significance of pairwise preferences is color-coded, as labeled on the figure. We use Barnard's Exact Test to estimate the statistical significance of the difference between pairwise preferences.

As Figure 8 indicates, KimKautzTMO has a higher performance compared to the other TMOs evaluated in our experiment. Reinhard performs the second best, while Krawczyk is slightly better than SemTMO, as the third-best performer. Additionally, the number of pairs where one TMO is better/worse than another is calculated for a quantitative evaluation. Results are presented in Table 2. Each cell in the table indicates the percentage of pairs which has a statistically significant preference towards the TMO on the row in comparison to the TMO on the corresponding column. Note that the sum of percentages between two TMOs is not equal to 100% due to pairs with a statistically non-significant difference (points with yellow color in Figure 8).

5.2 Inter-Observer Agreement

The inter-observer agreement is an important indicator of reliability. Although several methodologies have been proposed for inter-observer agreement and outlier detection in rating experiments [29], [30], there are not many well-established methodologies for ranking experiments. Ak et al. [13] showed that inter-observer agreement in pairwise comparison experiments can be measured based on Rogers-Tanimoto (RT) dissimilarity measure. A similar

Fig. 8. Distribution of number of participant preferences for each pairwise comparison. Each data point represents a unique image pair in the dataset. Black lines indicate the mean values of the preference percentages represented on the horizontal axis. Each data point is color coded based on the statistical significance test, i.e. significantly better/worse or no significant difference.

variation to such metric, known as Jaccard index [31], has been developed by Paul Jaccard. We use the Scipy implementation [32] of RT dissimilarity measure which is defined as follows:

$$RT_{ij} = \frac{2 \times (v_d)}{v_a + 2 \times (v_d)} \tag{2}$$

where v_d is the number of pairs for which two participants disagree on their pairwise preference, i.e., one selects the left image over the right one while another observer selects the right image over the left one, whereas v_a is the number of pairs where both participants agree on their preference. In addition to the above equation, a weight vector with the same size can be used to prioritize certain elements. More specifically, we generate the weights by the following equation to emphasize the effect of pairs with higher agreement on dissimilarity calculation:

$$w_{ij} = \frac{|p_{ij} - p_{ji}|}{n} \tag{3}$$

where n is the number of observers who ranked the pair of images $\{i, j\}$. p_{ij} is the number of observers who prefer image i over image j in pair comparison. Similarly, p_{ji} is the number of observers who prefer image j over image i in pair comparison. This allows us to generate weights that are closer to 1 as more observers agree on the preference among image pair $\{i, j\}$, and closer to 0 as the ambiguity of the pair increases.

Figure 9 shows the distribution of mean RT dissimilarities. Each point is a unique observer, and their RT dissimilarities are measured with every other observer in the corresponding playlist and averaged. Since the ambiguity of image pairs is different in each playlist, this might affect the agreement among observers. Therefore, each playlist is represented in a separate column on the figure. Based on the synthetic spammer profiles described in [13], we created an expected spammer RT dissimilarity range for each playlist. For a given playlist, each observer's RT dissimilarity distributions are then compared. Observers who have 90% overlap

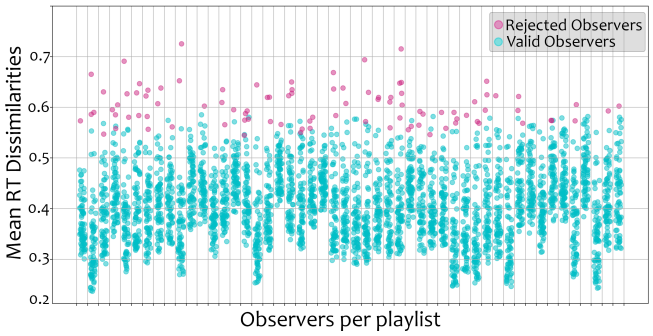

Fig. 9. Mean RT dissimilarity values of observers. Lower values of RT dissimilarity indicates a higher agreement. Observers are grouped on horizontal axis by their corresponding playlists

with an expected spammer RT dissimilarity range are rejected. In total, ninety-six observers, represented in magenta color, are rejected; while valid observers are represented in teal color.

6 Performance Evaluation of Objective Quality Metrics

6.1 Evaluation Criteria

Traditionally, the performance of objective quality metrics has relied on ground truth MOS which are obtained through subjective experiments using rating methodology. Correlation between the MOS and predicted quality scores are computed to evaluate the performance of objective quality metrics. Methods which map pairwise preferences into a continuous scale have been proposed in the literature. Zerman et al. [33] showed that there is a strong linear relation between pairwise preferences and MOS. However, cross-content evaluation is required to reduce the content dependency of mapped pairwise preferences. Cross-content evaluation of tone

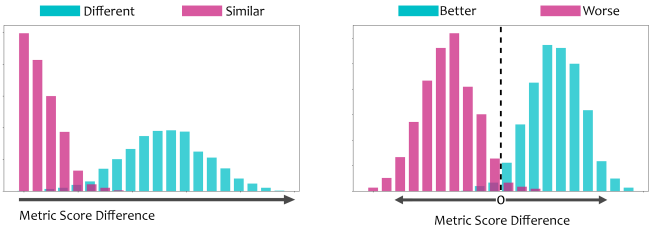

Fig. 10. Ideal distributions for 'Better/Worse' and 'Different/Similar' analysis. For 'Different/Similar' analysis, we expect a greater score difference for 'different' pairs and a much smaller difference for 'similar' pairs. For the 'Better/Worse' analysis, we expect 'worse' pairs to have negative metric score differences and conversely positive score differences for 'better' pairs

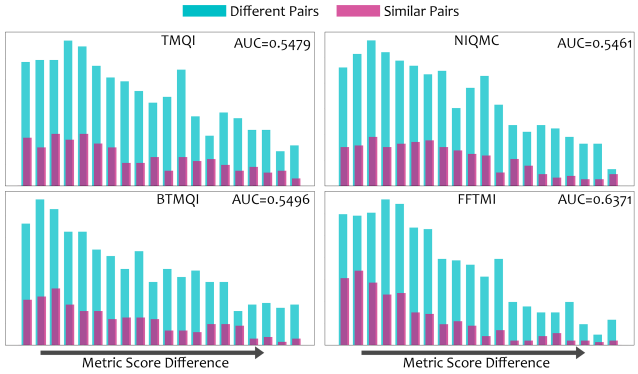

Fig. 11. Histograms of each metric for different vs. similar analysis is presented. AUC values are reported on top right corner of each plot. Blue colour denotes distribution of different pairs while pink denotes similar pairs. FFTMI produces the most desirable distribution.

mapped images does not provide any information in terms of TMO performances. Therefore, cross-content image pairs are not considered in our subjective experiment. This prevents mapping pairwise preferences onto a global continuous scale.

Krasula et al. [9] proposed an evaluation model which does not rely on mapping collected preferences onto a common scale. It also enables merging multiple datasets while allowing to determine the statistical significance of performance differences. In Krasula's model, the performance evaluation of objective quality metrics is conducted for two different aspects. Firstly, the area under the curve (AUC) value is used to determine how well a quality metric can distinguish between significantly different and similar pairs. Secondly, objective metrics are evaluated in terms of AUC and percentage of correct recognition of the qualitatively better image from a pair. Examples of an ideal distribution of metric score differences for each scenario are shown in Figure 10. This allows for an evaluation strategy that is closer to use cases in real applications. Interested readers are recommended to refer to the original paper [9] for more details.

6.2 Selected IQA Metrics

Several metrics dedicated to tone mapped image quality and aesthetic image quality assessment tasks have been collected for evaluation.

TMQI is a full-reference image quality metric to assess the quality of tone mapped images [3]. Structural and naturalness measures are combined to evaluate the quality of a tone mapped image with respect to the HDR image. It is the state-of-the-art quality metric for tone mapped image quality assessment.

NIQMC is a no-reference image quality metric developed to assess the quality of contrast distorted images [34]. It combines local and global features to generate a quality score. Although it is not specifically developed for tone mapped image quality assessment, it has a high correlation with subjective opinions in aesthetic evaluation tasks.

BTMQI is a no-reference image quality metric to assess the quality of tone mapped image by combining eleven features related to information entropy, statistical naturalness, and structural preservation [35].

FFTMI is a full-reference tone mapped image quality metric [23]. It relies on structural similarity, feature naturalness, and feature similarity between the HDR and tone mapped images.

6.3 Pre-processing Subjective Scores

As briefly described in Section 6.1, Krasula's method relies on the statistical significance of the differences between a pair of images. Therefore, we need to determine the statistical significance of the pairwise preference differences for each image pair. Furthermore, significantly different pairs are divided into two different groups as better and worse. This two-step evaluation strategy is highly comparable to real-life applications.

There are several ways to determine the statistical significance of the differences between different distributions [36], [37]. It has been shown that Barnard's exact test is more powerful than alternative statistical tests on 2×2 contingency tables [38]. Therefore, in this work, we use Barnard's exact test, since pair comparison results are represented by 2×2 matrices.

Since each pair was approximately ranked by seventy observers, we use the observers' pairwise preferences to generate the 2×2 contingency tables. Then, we use Barnard's test to determine the significance of the differences. 1154 pairs among the total 1500 are found to be significant with 95% confidence. Significantly different pairs are further divided into two groups as better (736 pairs), and worse (418 pairs). Better pairs indicate pairs where the left image is better than the right one, and conversely worse pairs indicate pairs where the right image is better than the left one. Although any pair can easily be categorized as better or worse by swapping their image positions, we used the initial positions while initialising the dataset.

6.4 Objective IQA Metrics Evaluation Results

We present the results of the objective quality metric evaluations in two steps: whether objective metrics can tell an image pair has a qualitative difference (different vs. similar), and if affirmative, which image has a higher quality (better vs. worse).

6.4.1 Different vs. Similar Analysis

The first analysis in Krasula's method [9] aims to determine how good the objective quality metrics are in distinguishing pairs with and without statistically significant difference. Ideally, the difference between predicted quality scores should be higher for image pairs with a statistically significant difference. Krasula method's uses the receiver operating characteristic (ROC) analysis [39] to determine the metrics different-similar classification performances. The performance of each metric is later represented

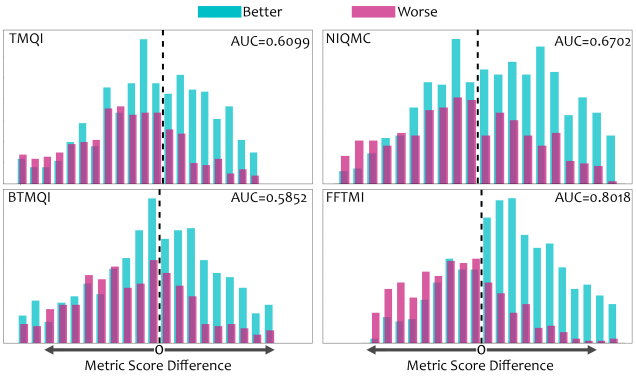

Fig. 12. Histograms of metrics for better vs. worse analysis. AUC values are reported on top right corner of each plot. Blue colour denotes distribution of better pairs while pink denotes worse pairs. Again, FFTMI performs better than the other three metrics.

as the AUC, where higher AUC values indicate a greater performance. To measure the statistical significance between metric performances, Krasula's method relies on a procedure proposed by Hanley and McNeil [40]. It calculates a critical ratio c_{ab} between the AUC of objective quality metrics, and statistical significance is estimated as the cumulative distribution function of c_{ab}. Interested readers are recommended to refer to the original paper [9] for more details.

Result of the different vs. similar analysis can be visualized on a histogram of metric score differences. Ideally, different pairs should be distributed away from 0 metric score difference, while similar pairs should be concentrated around 0 metric score difference. Figure 11 presents the results of the histogram of metric score difference for different and similar pairs. For each plot, blue color represents different pairs while pink color represents similar pairs. By analyzing the histograms, we can observe that FFTMI provides the most desirable distribution among the four, although far from ideal. Note that metric score differences increase from left to right for each plot. AUC values for each metric are reported on the top right corner of the plots. By comparing AUC values, we can observe that TMQI, NIQMC, and BTMQI provide similar and low performance on classifying different and similar pairs. Although FFTMI performs better than the other three metrics, there is still room for improvement. Statistical test results suggest that FFTMI significantly outperforms the other metrics in different vs. similar classification scenarios. It is also observed that performance differences between TMQI, NIQMC, and BTMQI are not statistically significant.

6.4.2 *Better vs. Worse Analysis*

After measuring the performance of metrics on identifying different and similar pairs, we aim to determine whether the metrics are able the recognize the image with higher preference in a pair. We divide different pairs into two groups as better and worse. The metric scores distribution can be visualized similarly to the previous analysis, with AUC values to quantify the performance differences among evaluated metrics. To statistically compare the metric performances in terms of AUC, we rely on the same methodology described in Section 6.4.1. Additionally, a more straightforward way of evaluation is to measure the percentage of correct classifications of better and worse pairs for each metric. In other words, we can check how many times an objective quality metric correctly recognizes the higher quality tone mapped image for each pair in the dataset. To statistically compare the correct classification performance of the metrics, the Krasula method relies on Fisher's exact test [37]. Interested readers are recommended to refer to the original paper [9] for more details about the evaluation procedure.

Figure 12 presents the histogram of metric score differences along with the AUC values. We can observe from the AUC values that FFTMI performs significantly better than the others. Distribution of the FFTMI score differences for better and worse pairs are closer to the desired distribution in comparison to the distribution of other metrics' score differences. We can see that the metric score differences for worse pairs are mostly located on negative values, while better pairs are on the positive side. TMQI, NIQMC, and BTMQI fail to provide a similar distinction between better and worse pairs as numerically represented by the AUC values. In terms of the percentages of correct classification of better and worse groups, we observe a similar outcome. Percentages of correct classifications are 58%, 61%, 56%, and 72% for TMQI, NIQMC, BTMQI, and FFTMI respectively. Statistical analysis with Fisher's exact test suggests that FFTMI performs significantly better than the other three metrics. NIQMC also performs significantly better than TMQI and BTMQI, whereas there is no statistically significant difference between TMQI and BTMQI performances.

7 Discussion & Conclusion

As discussed previously, it is easier and more natural for participants to compare the quality of two images than to assign a quality score to each image individually. Despite the advantages of pairwise comparison over rating tasks, metric development often relies on MOS scores. A method has been proposed to acquire MOS from pairwise preferences [33]. The authors conducted a series of experiments to acquire MOS scores from pairwise preferences and suggest including cross-content comparisons into the experiment to properly scale each image into a global quality scale. However, it is not useful to include cross-content comparisons in many use cases such as ours.

To develop objective IQA models directly on pairwise preferences, alternative objective functions might be incorporated into training. Prashnani et al. used a modified Bradley Terry (BT) [41] model as an objective function to train a deep learning model on probabilistic pairwise preference data [42]. During training, the model predicts quality scores for each image and pairwise preference probabilities are calculated from the predicted scores with modified BT. After training, the model is able to predict quality scores for individual image (in comparison to a pristine reference image).

To the best of our knowledge, current tone mapped IQA metrics are often developed with handcrafted features as can be seen in the evaluated metrics. Despite the advancement of learning based algorithms, due to lack of publicly available datasets, there are not well established tone mapped IQA metrics in the literature. In comparison to previous works, as presented in Table 1, we propose the largest publicly available dataset and allowing research community to build upon. We believe that, by providing the largest publicly available tone mapped IQA dataset (RV-TMO), we open a new route for researchers to develop tone mapped IQA metrics. Thanks to the proposed content selection strategy, we ensured

a well distributed ambiguity in tone mapped image pairs which learning-based algorithms can benefit from.

We conducted a large-scale experiment on tone mapped image quality evaluation via crowdsourcing. To the best of our knowledge, this is the largest publicly available TMO evaluation dataset: 250 unique HDR images used to generate 1000 tone mapped images which provide 1500 pair comparisons. 3500 observers participated in the subjective experiment where each pair was evaluated by approximately seventy unique observers. Four state-of-the-art TMO performances were evaluated, where KimKautzTMO [20] was preferred most often. ReinhardTMO [22] performed the second best while KrawczykTMO [21] came in third place, performing slightly better than the SemTMO [26] in fourth.

Moreover, we developed a content selection strategy to select representative and challenging HDR crops from high-resolution HDR images. We further developed an objective quality metric based clustering method to balance the ambiguity of the pairs in the experiment. It is crucial to have such balance to develop new metrics, specifically for machine learning-based models. To the best of our knowledge, there is a lack of a well-established methodology for observer reliability for pairwise experiments. In addition to behavioral tools used for the observer analysis, we proposed a novel approach to statistically evaluate the observer reliability for pairwise experiments.

Finally, we provide a benchmark for well-known tone mapped image quality metrics based on Krasula's method [9]. We discussed how to utilize collected data to develop novel objective quality metrics, and how to benchmark existing metrics. Collected pairwise preferences, tone mapped images used in the experiment, HDR images used for tone mapping, and scripts are made publicly available to aid further research.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 765911 (RealVision)

REFERENCES

[1] L. Krasula, M. Narwaria, K. Fliegel, and P. Le Callet, "Preference of Experience in Image Tone-Mapping: Dataset and Framework for Objective Measures Comparison," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 64 – 74, Feb. 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01633843

[2] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "Large-Scale Crowdsourced Study for Tone-Mapped HDR Pictures," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4725–4740, 2017.

[3] H. Yeganeh and Z. Wang, "Objective Quality Assessment of Tone-Mapped Images," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 657–667, 2013.

[4] J. Kuang, H. Yamaguchi, G. M. Johnson, and M. D. Fairchild, "Testing hdr image rendering algorithms," in *Color Imaging Conference*, 2004.

[5] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, "Evaluation of tone mapping operators using a high dynamic range display," *ACM Trans. Graph.*, vol. 24, no. 3, p. 640–648, jul 2005. [Online]. Available: https://doi.org/10.1145/1073204.1073242

[6] "Prolific," https://www.prolific.com/, accessed: Apr. 2021. [Online].

[7] "Amazon Mechanical Turk," https://www.mturk.com, accessed: Apr 2021. [Online].

[8] A. Goswami, A. Ak, W. Hauser, P. L. Callet, and F. Dufaux, "Reliability of crowdsourcing for subjective quality evaluation of tone mapping operators," in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, 2021, pp. 1–6.

[9] L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma, "On the accuracy of objective image and video quality models: New methodology for performance evaluation," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6.

[10] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625, 2012.

[11] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.

[12] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 3097–3100.

[13] A. Ak, M. Abid, M. P. Da Silva, and P. Le Callet, "On Spammer Detection in Crowdsourcing Pairwise Comparison Tasks: Case Study on Two Multimedia QoE Assessment Scenarios." in *ICME 2021 - First International Workshop on Quality of Experience in Interactive Multimedia*, Virtual, China, Jul. 2021. [Online]. Available: https://hal.archives-ouvertes.fr/hal-03236236

[14] D. Ghadiyaram and A. C. Bovik, "Massive Online Crowdsourced Study of Subjective and Objective Picture Quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016.

[15] J. Petit and R. K. Mantiuk, "Assessment of video tone-mapping: Are cameras' s-shaped tone-curves good enough?" *J. Vis. Commun. Image Represent.*, vol. 24, pp. 1020–1030, 2013.

[16] M. Fairchild, "The HDR photographic survey," pp. 233–238, Jan. 2007.

[17] A. Artusi, R. K. Mantiuk, T. Richter, P. Hanhart, P. Korshunov, M. Agostinelli, A. Ten, and T. Ebrahimi, "Overview and evaluation of the JPEG XT HDR image compression standard," *Journal of Real-Time Image Processing*, vol. 16, no. 2, pp. 413–428, 2019.

[18] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1404–1412.

[19] W. Zhang, R. R. Martin, and H. Liu, "A saliency dispersion measure for improving saliency-based image quality metrics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1462–1466, 2018.

[20] M. H. Kim, J. Kautz *et al.*, "Consistent tone reproduction," in *Proceedings of the Tenth IASTED International Conference on Computer Graphics and Imaging*. ACTA Press Anaheim, 2008, pp. 152–159.

[21] G. Krawczyk, K. Myszkowski, and H.-P. Seidel, "Lightness perception in tone reproduction for high dynamic range images," in *Computer Graphics Forum*, vol. 24, no. 3. Citeseer, 2005, pp. 635–646.

[22] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 2002, pp. 267–276.

[23] L. Krasula, K. Fliegel, and P. Le Callet, "FFTMI: Features Fusion for Natural Tone-Mapped Images Quality Evaluation," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 2038–2047, 2019.

[24] X. Cerdá-Company, C. A. Párraga, and X. Otazu, "Which tone-mapping operator is the best? A comparative study of perceptual quality," *CoRR*, vol. abs/1601.04450, 2016. [Online]. Available: http://arxiv.org/abs/1601.04450

[25] L. Krasula, M. Narwaria, K. Fliegel, and P. Le Callet, "Influence of HDR reference on observers preference in tone-mapped images evaluation," *2015 7th International Workshop on Quality of Multimedia Experience, QoMEX 2015*, 07 2015.

[26] A. Goswami, M. Petrovich, W. Hauser, and F. Dufaux, "Tone Mapping Operators: Progressing Towards Semantic-awareness," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2020, pp. 1–6.

[27] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti, "Beyond the turk: Alternative platforms for crowdsourcing behavioral research," *Journal of Experimental Social Psychology*, vol. 70, pp. 153–163, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022103116303201

[28] S. Palan and C. Schitter, "Prolific.ac—a subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2018.

[29] R. K. P. Mok, R. K. C. Chang, and W. Li, "Detecting low-quality workers in qoe crowdtesting: A worker behavior-based approach," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 530–543, 2017.

[30] ITU-R, "Methodology for the Subjective Assessment of the Quality of Television Pictures," ITU-R Recommendation BT.500-13, 2012.

[31] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bulletin del la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.

[32] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[33] E. Zerman, V. Hulusic, G. Valenzise, R. K. Mantiuk, and F. Dufaux, "The Relation Between MOS and Pairwise Comparisons and the Importance of Cross-Content Comparisons," in *Human Vision and Electronic Imaging 2018, Burlingame, CA, USA*, 2018. [Online]. Available: https://doi.org/10.2352/ISSN.2470-1173.2018.14.HVEI-517

[34] K. Gu, W. Lin, G. Zhai, X. Yang, W. Zhang, and C. W. Chen, "No-Reference Quality Metric of Contrast-Distorted Images Based on Information Maximization," *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4559–4565, 2017.

[35] K. Gu, S. Wang, G. Zhai, S. Ma, X. Yang, W. Lin, W. Zhang, and W. Gao, "Blind Quality Assessment of Tone-Mapped Images Via Analysis of Information, Naturalness, and Structure," *IEEE Transactions on Multimedia*, vol. 18, pp. 1–1, 03 2016.

[36] G. A. Barnard, "A new test for 2× 2 tables," *Natur*, vol. 156, no. 3954, p. 177, 1945.

[37] R. A. Fisher, "On the interpretation of χ 2 from contingency tables, and the calculation of P," *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87–94, 1922.

[38] C. R. Mehta and P. Senchaudhuri, "Conditional versus unconditional exact tests for comparing two binomials," *Cytel Software Corporation*, vol. 675, pp. 1–5, 2003.

[39] J. A. Swets, "Book Reviews : Signal Detection Theory and ROC Analysis in Psychology and Diagnostics : Collected Papers." *Medical Decision Making*, vol. 19, no. 2, pp. 217–217, 1999. [Online]. Available: https://doi.org/10.1177/0272989X9901900216

[40] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases." *Radiology*, vol. 148, no. 3, pp. 839–843, 1983, pMID: 6878708. [Online]. Available: https://doi.org/10.1148/radiology.148.3.6878708

[41] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: The method of paired comparisons," *Biometrika*, vol. 39, no. 3-4, pp. 324–345, 12 1952. [Online]. Available: https://doi.org/10.1093/biomet/39.3-4.324

[42] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual Image-Error Assessment through Pairwise Preference," 2018.