UNIVERSITY OF
CAMBRIDGE

# Active sampling, scaling and dataset merging for large-scale image quality assessment

Aliaksei Mikhailiuk

Sidney Sussex

This dissertation is submitted on September, 2020 for the degree of Doctor of Philosophy

# Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the prescribed limit of 60 000 words.

Aliaksei Mikhailiuk
September, 2020

# Abstract

**Active sampling, scaling and dataset merging for large-scale image quality assessment**

**Aliaksei Mikhailiuk**

The field of subjective assessment is concerned with eliciting human judgements about a set of stimuli. Collecting such data is costly and time-consuming, especially when the subjective study is to be conducted in a controlled environment and using a specialized equipment. Thus, data from these studies are usually scarce. One of the areas, for which obtaining subjective measurements is difficult is image quality assessment. The results from these studies are used to develop and train automated or objective image quality metrics, which, with the advent of deep learning, require large amounts of versatile and heterogeneous data.

I present three main contributions in this dissertation. First, I propose a new active sampling method for efficient collection of pairwise comparisons in subjective assessment experiments. In these experiments observers are asked to express a preference between two conditions. However, many pairwise comparison protocols require a large number of comparisons to infer accurate scores, which may be unfeasible when each comparison is time-consuming (e.g. videos) or expensive (e.g. medical imaging). This motivates the use of an active sampling algorithm that chooses only the most informative pairs for comparison. I demonstrate, with real and synthetic data, that my algorithm offers the highest accuracy of inferred scores given a fixed number of measurements compared to the existing methods. Second, I propose a probabilistic framework to fuse the outcomes of different psychophysical experimental protocols, namely rating and pairwise comparisons experiments. Such a method can be used for merging existing datasets of subjective nature and for experiments in which both measurements are collected. Third, with a new dataset merging technique and by collecting additional cross-dataset quality comparisons I create a Unified Photometric Image Quality (UPIQ) dataset with over 4,000 images by realigning and merging existing high-dynamic-range (HDR) and standard-dynamic-range (SDR) datasets. The realigned quality scores share the same unified quality scale across all datasets. I then use the new dataset to retrain existing HDR metrics and show that the dataset is sufficiently large for training deep architectures. I show the utility of the dataset and metrics in an application to image compression that accounts for viewing conditions, including screen brightness and the viewing distance.

# Acknowledgements

First and foremost, I am thankful to my supervisor, Dr. Rafal K. Mantiuk, who has provided guidance and encouragement throughout the past three years. I am also grateful to my assessors, Prof. Alan Blackwell, Dr. Damon Wischik, Dr. Hatice Gunes, Prof. Peter Robinson and Prof. Patrick Le Callet for their constructive feedback presented in the format of friendly discussions. I am grateful to the European Research Council (ERC) for making my PhD happen.

I would like to thank all my friends and colleagues in Cambridge. Collaborations at the Computer Lab were certainly fun. In particular, I would like to extend my gratitude to Maria Perez-Ortiz, Gyorgy Denes, Andrei Iliescu, Param Hanji, Kuba Maruszczyk, Fangcheng Zhong, Dingcheng Yue, Akshay Jindal, Aamir Mustafa, Maryam Azimi, Minjung Kim, Nanyang Ye, Marwa Mahmoud, Nikhil Churamani, and Andy Rowlands.

Last but not least, I am grateful for the support I have received both from my family and all my friends. In particular, Victor Prokhorov, Anton Stankevich, Vladislav Baboshko, Vitali Petsiuk, Uladzimir Liasun, Nikolay Mazurenko, Vlad Tereshenko, Palina Malash, Aliaksandra Shysheya, Khrystsina Darapei, Kirill Piasotski, Nickolay Yaskevich, Irina Palto, Robert Clucas, Callum J. Court, Francois Chalus, Francesca Farrington and Yeva Volkava.

I would love to extend my gratitude to my mentors and supervisors who helped me grow at different stages of my life, Leokadia Efremkova, Stas Buben, Dr. Martin Cook, Dr. Naim Dahnoun, and Dr. Anita Faul.

# Contents

# Chapter 1

# Introduction

## 1.1 Background

Automatic or *objective* assessment of image and video quality is a stepping stone for accurate compression, reconstruction, enhancement and tone-mapping algorithms [120, 17, 8, 68, 122]. Since the ultimate consumer of visual content is a human observer the results of these algorithms must be perceptually pleasing. The final model built for this type of applications should show high correlation with *subjective* quality as perceived by human observers. In my dissertation I focus on image quality assessment and, thus, in most experiments and discussions I refer to image quality scale and datasets. However, many of the findings can also be applicable in other problems, where human judgements are elicited.

A typical objective image quality metric takes as input a test image, with or without the reference counterpart, and produces a single number – the quality score, where the quality scale has been fixed beforehand. Objective image quality assessment is a multi-faceted problem and achieving high correlation with subjective quality is challenging. Generally, two types of objective quality assessment methods are distinguished: model-based and data-driven [59]. The former relies on assumptions about the human visual system and has no or very few trainable parameters, whereas the latter often has millions of trainable parameters and relies heavily on large quantities of versatile training data. It has been shown that perceived image quality depends on the content of the image. Here, distortions on the common fixation points [95], such as human faces, are more noticeable [110]; and the viewing conditions – distance from the display and display brightness [145]. Traditionally, image quality metrics relied solely on the differences in pixel values and ignored the physical specification of the display and viewing environment. This limitation can be partially attributed to the standards [127] developed in the era of standardized displays. The standards stipulate in what conditions and what displays the content are to be viewed. However, these are outdated, given various ways in which visual content is displayed nowadays – mobile devices of different sizes and resolution, SDR and very bright HDR displays.

In this dissertation, I define an ideal *perceptual* quality metric as the one that: (i) has a high

correlation with the quality as perceived by human observers; (ii) is content-driven, so that it accounts for common fixation points; (iii) is differentiable, enabling its ubiquitous use as an optimization objective; (iv) is viewing condition dependent, considering the parameters of the plethora of modern displays; and (v) has a meaningful scale, important for image compression algorithms, where an improvement in quality can be quantified and interpreted. To the best of my knowledge, no metric that would adhere to all these requirements exists. Nevertheless, certain metrics have successfully incorporated some of the features defined above. Model-based, HDRVDP-2.2 [82], based on the assumptions about the human visual system (HVS), accounts both for the display brightness and the viewing distance. However, it is not differentiable, does not use an interpretable quality scale and, is not content-driven. Data-driven metrics, on the other hand, are content driven, differentiable and can learn a meaningful scale. To make them viewing condition dependent, they need to be trained on a dataset with image quality evaluated at different viewing conditions.

The major obstacle to developing an "ideal" metric is, thus, the lack of a sufficiently diverse training dataset that considers different content and distortions, viewing distance and display brightness. Training data for image and video quality assessment is hard to obtain, as it requires running expensive and time-consuming *subjective* quality assessment experiments with human observers. The problem of collecting such data for a set of viewing conditions is even more acute, as it may require specialized equipment and a controlled experimental environment, preventing the use of crowd-sourced image quality assessment studies. Thus, existing subjective image quality datasets are homogeneous and fragmented, covering only a small number of image contents and impairments. Using these datasets together is not possible as the data coming from different quality assessment experiments might be scaled differently, often resulting in very different quality scores for images of similar perceived quality. For example, an image rated four on a five-point scale in one experiment could be rated two in another experiment because of differences in the training of the participants, range, and type of considered distortions. Dealing with widely different scales when training quality metrics is problematic, often requires using rank-order correlation, which has limited expressive power, as a measure of prediction accuracy, and makes difficult the use of multiple datasets for training [102, 148].

## 1.2 Objective

In my dissertation, I focus on efficient data collection and aggregation for image quality assessment experiments. More specifically, I concentrate on image fidelity metrics within the wider topic of image quality. Here the test image of highest quality would be indistinguishable its' pristine counterpart. The ultimate goal of my work is creating a dataset that would be: (i) of high quality, containing accurate estimates of the ground truth scores, and (ii) large enough for training data-driven metrics. I look into how to minimize the number of required measurements

for obtaining an accurate representation of perceived subjective quality by proposing a new active learning method for pairwise comparison experiments. By noting that many datasets can be re-used, I argue for the consolidation of existing datasets and propose a new procedure for merging existing datasets coming from both types of experiments. Using the methods proposed in this dissertation, I construct the largest subjective photometric image quality dataset and show the utility of such a dataset by training a deep perceptual image quality assessment metric that adheres to the criteria set for an "ideal" metric and re-train existing quality metrics. To show the utility of the new dataset, I present examples of novel applications enabled by the metrics trained on it.

Although I mainly consider image quality in my dissertation, my findings are far-reaching and can be applied to many problems where data collection may involve human participants. Examples of such problems are: i) user preferences (i.e. recommendation systems, information retrieval and relevance estimation) [61]; ii) matchmaking in gaming systems such as TrueSkill for Xbox Live [41] and Elo for chess and tennis tournaments [33]; iii) psychometric experiments for behavioural psychology [19].

## 1.3 Structure of the dissertation

In Chapter 2 I cover the fundamentals of objective and subjective quality assessment and psychometric scaling, focusing on pairwise comparisons and rating experiments. In Chapter 3 I use the concepts discussed in Chapter 2 to improve the scores of one of the largest subjective image quality assessment (IQA) dataset, TID2013 [106]. I then discuss active sampling for pairwise comparisons in Chapter 4, which can greatly decrease experimental effort with an increasing accuracy. In Chapter 5 I propose a new scaling method for unifying the results of pairwise comparisons and rating experiments. In Chapter 6 I use the scaling method from Chapter 5 for merging together four existing IQA datasets, producing the largest photometric IQA dataset to date. In Chapter 7 I show the utility of the new dataset by re-training existing objective HDR IQA metrics and show that the dataset is sufficiently large for deep convolutional neural network (CNN)-based metrics by training the first deep photometric image quality metric. I also present novel applications, enabled by the metrics trained on my dataset. In Chapter 8 I verify whether the state-of-the-art objective IQA metrics can be used for finding the threshold for viewing condition dependent visually lossless image compression. Chapter 9 concludes my dissertation.

## 1.4 Publications, talks, demos

### Publications that are part of this dissertation

- **Mikhailiuk A.**, Wilmot C., Perez-Ortiz M., Mantiuk R., 2020. "Active Sampling for Pairwise Comparisons via Approximate Message Passing and Information Gain Maximization". International Conference on Pattern Recognition (ICPR).

  *Contribution:* Developed the state-of-the-art active sampling procedure for pairwise comparisons. Analyzed the performance with respect to other methods.

- **Mikhailiuk A.**, Perez-Ortiz M., Yue D., Suen W. and Mantiuk R., 2020. "Consolidated Dataset and Metrics for High-Dynamic-Range Image Quality". Transactions on Multimedia (TMM). *Under review*.

  *Contribution:* Performed subjective image quality assessment user study, scaled several subjective IQA datasets together to obtain the largest photometric IQA dataset to date, developed a new CNN architecture for photometric image quality assessment, proposed an application for the newly build metric.

- **Mikhailiuk A.**, Ye N., Mantiuk R., 2020. "The effect of display brightness and viewing distance: a dataset for visually lossless image compression". Human Vision and Electronic Imaging (HVEI). *Under review*.

  *Contribution:* Performed analysis of quality threshold for viewing condition dependent visually lossless image compression. Performed validation of image quality and visibility metrics.

- Perez-Ortiz M.*, **Mikhailiuk A.***, Zerman E., Hulusic V., Valenzise G., Mantiuk R., 2019 "From pairwise comparisons and rating to a unified quality scale". IEEE Transactions on Image Processing (TIP). *(*) Authors had an equal contribution.*

  *Contribution:* Performed a user study with pairwise comparison experiments. Proposed a new model for combining datasets with ranking and rating. Analyzed the results.

- **Mikhailiuk A.**, Perez-Ortiz M., Mantiuk R., 2018 "Psychometric scaling of TID2013 dataset". 2018 IEEE Conference on Quality of Multimedia Experience (QoMEX2018).

  *Contribution:* Performed subjective image quality assessment study, scaled and analyzed the data.

### Publications that are not part of this dissertation

- Andrein Iliescu, **Mikhailiuk A.**, Mantiuk R.,2020. "Domain content disentanglement." Computer Vision and Pattern Recognition (CVPR). *Under review*.

*Contribution:* Implemented and trained benchmark generative adversarial networks, and developed validation metrics.

- Gyorgy Denes, Akshay Jindal, **Mikhailiuk A.**, Mantiuk R.,2020. "A perceptual model of motion quality for rendering with adaptive refresh-rate and resolution" (SIGGRAPH).

  *Contribution:* Designed an experimental study based on pairwise comparisons and active sampling procedure proposed in this dissertation.

- **Mikhailiuk A.**, Faul A., 2018. "Deep Learning Applied to Seismic Data Interpolation". In: European Association of Geoscientists and Engineers (EAGE) Conference 2018.

  *Contribution:* Proposed a deep neural network architecture for seismic data interpolation.

Work presented in this dissertation does not include results obtained or code written by my collaborators (Maria Perez-Ortiz, Wilson Suen, Dingcheng Yue, Nanyang Ye, Clifford Wilmot, Emin Zerman, Vedad Hulusic, Guizeppe Valenzisse), unless stated otherwise. However, I acknowledge their input in conversations and discussions.

## Talks

- "Photometric image quality assessment". Computer Laboratory, University of Cambridge, 8th May 2019.
- "Assessing the quality of experience". Computer Laboratory, University of Cambridge, 4th Feb 2019, (workshop).
- "Psychometric scaling of TID2013 dataset". QoMEX2018, Sardinia, Italy, 31st May 2018.
- "Psychometric scaling of TID2013 dataset". Computer Laboratory, University of Cambridge, 17th May 2018.
- "Machine learning for image quality assessment". Sidney Sussex College, University of Cambridge, 17th Feb 2018.

# Chapter 2

# Background

## 2.1 Introduction

The field of subjective assessment is concerned with measuring and modeling human judgments. Here, the ultimate goal is to map the subjective degree of belief or strength of human participants' experiences to a scale. The field stems from social and psycho-physical studies. Here the perceptual processes are analyzed by studying the subject's experience or behaviour conditional on the systematic changes in the the properties of a stimulus along one or more physical dimensions [123]. However, since in my dissertation I am concerned with image quality assessment, most examples and explanations are focused on it. In typical subjective image quality assessment experiments, participants *rate* a set of stimuli or conditions according to some criteria, or *rank* a subset of them. Rating is inherently more difficult for participants than ranking. Before the rating experiment participants need to undergo training, as the notion of the scale depends on the selected conditions for the experiment and varies from observer to observer. Whereas a binary choice for the question *"which condition is better?"* or *"what is the order of the conditions in terms of quality?"* carries a much lower cognitive load and as a result is less biased [125, 114, 84, 146]. Because of that advantage over rating, comparative judgment experiments gain attention in subjective assessment and crowd-sourced experiments. The simplest form of ranking experiments is comparing conditions in pairs (pairwise comparison protocol), and hence it is the most common ranking choice. Here observers are asked to choose one out of two conditions according to some criteria. To construct a scale from pairwise comparisons, unlike in rating, in which conditions are mapped directly to a scale by averaging the observers' scores, we need to model and infer the latent scores. This problem is known as psychometric scaling. Once the quality scale is available, the scores can be used to develop objective quality assessment algorithms.

In this chapter, I give an overview of subjective image quality assessment methods, talk about advantages and disadvantages of uni-dimensional and multi-dimensional scaling, explain the observer model, show the procedural prerequisites for obtaining an accurate and interpretable

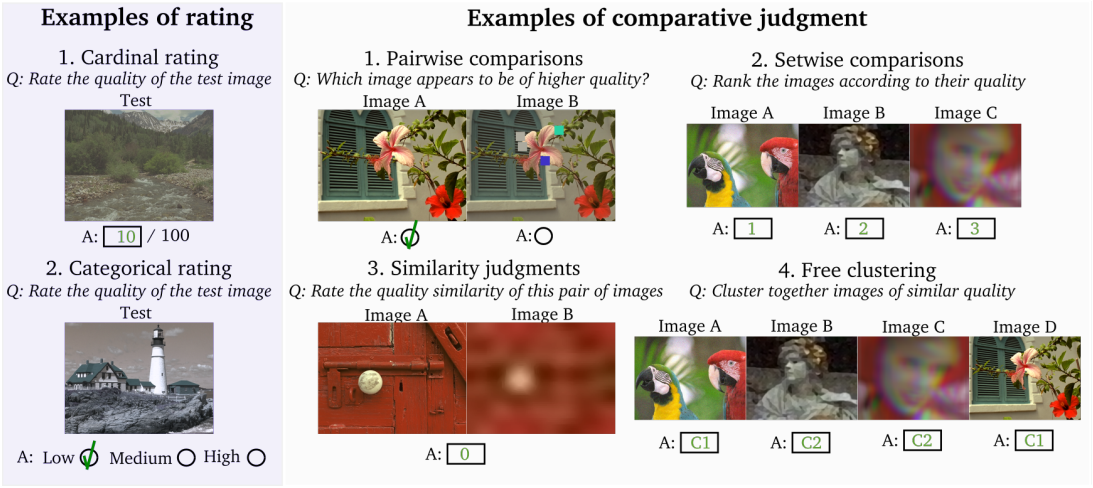



Figure 2.1: Examples of different subjective judgment experiments.

scale and discuss different procedures of scale construction from collected measurements. I then introduce various quality assessment criteria and objective IQA methods, I also talk about the ways of making them account for display brightness.

## 2.2 Subjective quality assessment

Methodologies for subjective quality assessment [127, 46, 27] can be generally classified as rating and ranking (or comparative judgment) methods. Figure 2.1 shows some examples of rating and comparative judgment experiments. Rating methods can be single, double, or multi-stimuli [128] (the latter can only be applied to audio signal and, thus are out of the scope of this dissertation), depending on the presentation of the test stimuli. Users are asked to rate the stimuli using either a categorical or continuous interval scale. The most commonly used rating methodologies for single-stimulus are: absolute category rating (ACR) [127], where only test stimuli are rated; or absolute category rating hidden reference (ACR-HR), in which the reference stimuli is mixed among test ones and in the experiment observers are prompted to rate both of them. For double stimulus the common methodologies are: double stimulus impairment scale (DSIS) – with reference stimuli preceding the corresponding test stimuli, but only test stimuli being rated; or double stimulus continuous quality scale (DSCQS) [46] for double-stimulus cases – where the order in the pair of test and reference stimuli is randomized. Rating methods generally work better when stimuli are easily distinguishable from one another.

Comparison methods require observers to rank two or more stimuli [26]. These are more suitable for cases in which the visual difference between two stimuli is small. The most commonly used comparative approach is referred to as pairwise comparisons (PWC) when only two stimuli are compared at a time. The main advantage of this approach is its simplicity. The weaknesses and strengths of these strategies were compared in several studies [103, 124, 111, 84].

Essentially, rating has the advantage to provide an interpretable, supra-threshold scale of quality or distortion impairment, but it also requires careful training of subjects, who might have a different interpretation of the scale adjectives. As a consequence, the rating scale is, in general, not universal and may require further calibration to adjust the scores obtained from individual observers [125]. On the other hand, pairwise comparison experiments have a lower cognitive load, require little training, and generally eliminate the observer's bias and are therefore well suited for non-expert participants and crowd-sourcing experiments. However, the total number of possible comparisons increases quadratically with the number of stimuli. In practice, not all comparisons are equally useful, e.g., comparing stimuli with too distant impairment levels is generally uninformative [138]. Pairs of stimuli to be compared can be sampled iteratively based on: (i) the previously compared stimuli; (ii) the heuristics [106] or; (iii) information-theoretic criteria [146]. Recently, Shah et al. [114] compared rating and pairwise comparison experiments by conducting a series of subjective experiments in which ground truth was available – e.g. the correct radius of the presented circle or the word count in a paragraph. Comparison experiments were found to be more accurate in most cases and took less time compared to rating. However, the authors also found that the performance of rating and pairwise comparison experiments depends on the measurement noise of each experiment (standard error of the estimation of the mean), due to a limited sample size.

### 2.2.1 Uni-dimensional and multi-dimensional scaling

A natural question of scale dimensionality arises in many psychophysical experiments [86]. Uni-dimensional and multi-dimensional models are distinguished. Both have their advantages and disadvantages. However, their usage is experiment and application-specific. For example, where data possess a higher number of latent dimensions, uni-dimensional models would be insufficient in explaining them, as such, distances between the capitals of the largest countries on different continents require to be projected on multiple dimensions. At the same time, uni-dimensional scaling can be sufficient in cases where the experimental procedure explicitly measures a single data dimension, for example, when measuring the skill of a game player or perceived brightness of a display. Furthermore, interpreting multi-dimensional models, where the dimensions are not clearly defined beforehand, can be challenging [86]. In my work, I am interested in the quality of impaired images judged relative to the reference image, where the produced scale must be interpretable – with distances in the scale mapped to physical, explainable quantities. Many applications require one-dimensional scale – for example, image compression algorithms [109, 117], which need to be fine-tuned, or neural network based images-translation algorithms, which require a one-dimensional loss function [122, 8, 152]. Although I focus on uni-dimensional scaling, I explore the possibility of using multi-dimensional scaling on image quality datasets in Chapter 3.

### 2.2.2 Observer model

To build a quality scale, certain assumptions need to be made about how observers respond and perceive quality. Such assumptions are encapsulated in the observer model. The model, described in this section, will be used to formulate the dataset-merging procedure in Chapter 5.

It is often assumed, that quality is a one-dimensional variable, i.e., observers assign a scalar quality value to each condition. However, observers might vary in their notions of quality (inter-observer variance), and their opinions are also likely to change when they repeat the same experiment (intra-observer variance). Thus, quality is not a deterministic value, but a random variable, which accounts for the subjective nature of these experiments.

Suppose we aim to form a scale for a set of $n$ conditions $S = \{o_1, \ldots, o_n\}$ (conditions being images, players, etc.) that are evaluated according to a feature or characteristic (subjective measurements such as aesthetics, relevance, quality, etc.) with unknown underlying ground truth scores $\mathbf{q} = (q_1, \ldots, q_n)$, $q_i \in \mathbb{R}$. Here I simply refer to these as quality scores.

**Rating experiments** In rating experiments the random variable associated with the quality can be expressed using the following mixed-model [30] of observer rating behavior [52]:

$$\pi_{ik} = q_i + \delta_k + \xi_{ik}, \tag{2.1}$$

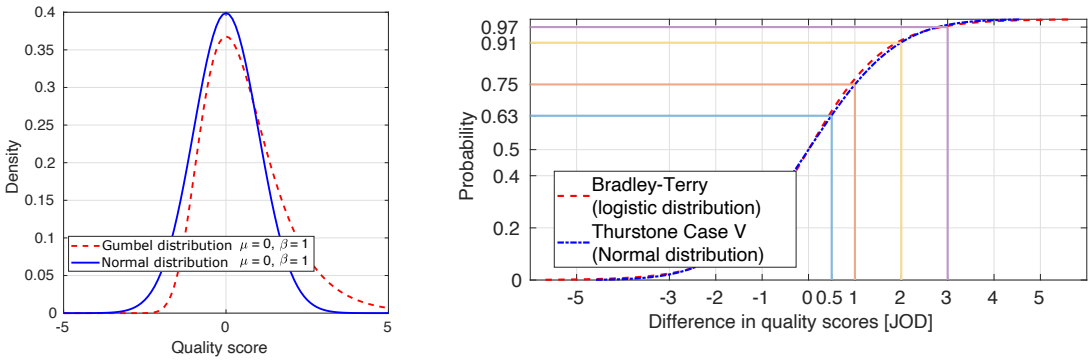
meaning that the rating $\pi_{ik}$ for observer $k$ and condition $i$ depends on: the scalar $q_i$, the ground truth quality score; random variable $\delta_k$, the subject bias; and random variable $\xi_{ik}$ the subject inaccuracy and stimulus scoring difficulty. Bias and inaccuracy components in the model are assumed to be independent random variables that are normally distributed and $\xi_{ik}$ is assumed to have a zero mean and $\delta_k$ has a zero mean when observed across all subjects. This makes rating $\pi_{ik}$ also normally distributed.

**Pairwise comparison experiments** Two most widely used observer models for pairwise comparisons are Thurstone [123] and Bradley-Terry [10]. In practice, both lead to similar solutions. Within the Thurstone model, the perceived quality of condition $i$ is modeled as a random variable $\omega_i$:

$$\omega_i \sim \mathcal{N}(q_i, \beta_i^2), \tag{2.2}$$

where the mean of the distribution is assumed to be the true quality score $q_i$ and the standard deviation $\beta_i$ accounts for combined inter- and intra-observer variance. Individual quality scores of compared conditions can be inferred from the relative distances, calculated as:

$$\omega_j - \omega_i \sim \mathcal{N}(q_{ij}, \beta_{ij}^2), \tag{2.3}$$




(a) (b)

Figure 2.2: (a) Thurstone observer model assuming normal distribution and Bradley-Terry observer model assuming Gumbel distribution. (b) Cumulative distribution functions mapping probabilities to distances in the scale. Parameters for Thurstone Case V and Bradley-Terry models were chosen such that the difference in 1 unit correspond to 75% probability of one condition being better than another.

where $\beta_{ij}$ is the standard deviation of a new distribution obtained from the difference between two quality distributions, its exact formulation, presented below, depends on the Case of the Thurstone model; and $q_{ij}$ is the new mean of the distribution, i.e., $q_{ij} = q_i - q_j$.

Five cases of the original Thurstone model are distinguished, based on simplifying assumptions imposed on $\beta_{ij}$:

1. The original Thurstone model, referred to as Case I, assumes that only one participant is performing the experiment and the standard deviation of the difference between random variables $\omega_i - \omega_j$ is $\beta_{ij} = \sqrt{\beta_i^2 + \beta_j^2 - 2\rho\beta_i\beta_j}$, where $\rho$ is the correlation between individual scores. Despite being general, Thurstone Case I is insolvable, as every new observation will introduce a new unknown, making the number of unknowns always greater than the number of equations [123].

2. Thurstone Case II assumes that the model can be applied to a group of participants, i.e., the results of individual participants can be aggregated.

3. Thurstone Case III assumes that $\beta_{ij} = \sqrt{\beta_i^2 + \beta_j^2}$, that is $\rho = 0$.

4. Thurstone case IV further assumes that $\beta_i$ and $\beta_j$ are approximately equal, resulting in further simplification $\beta_{ij} = \frac{\beta_i + \beta_j}{\sqrt{2}}$.

5. Thurstone Case V assumes $\beta_i$ and hence $\beta_{ij}$ are constant across all conditions. For simplicity of notation, I refer to $\beta_i$ as $\beta_*$ and $\beta_{ij}$ as $\beta$ throughout the whole dissertation. Thus, for a single random variable $w_i$ the following holds: $\omega_i \sim \mathcal{N}(q_i, \beta_*^2)$

When Thurstone Case V is compared to the rating model in Equation 2.1, it can be noticed that it eliminates the observer bias $\delta_i$ (since pairwise comparisons are relative) and that it assumes

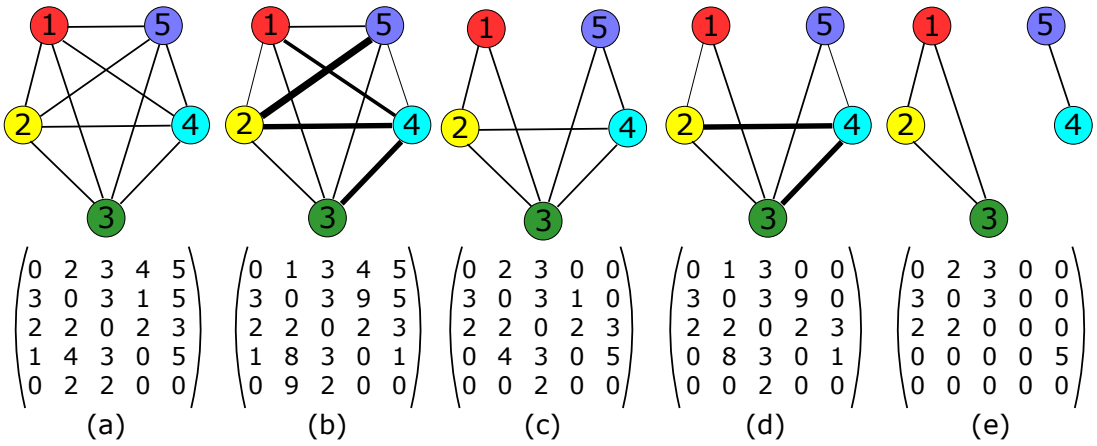



Figure 2.3: Bottom row – examples of a matrix with the results of pairwise comparison experiments ($\mathbf{C}$); top row visualization of a comparison graph corresponding to the matrix $\mathbf{C}$, the thickness of lines corresponds to the number of comparisons between conditions for: (a) full balanced design, where everything is compared to everything the same number of times (in this case five); (b) full unbalanced design, where everything is compared to everything not necessarily the same number of times; (c) incomplete balanced design, where some pairs are not compared, but those that are compared, have the same number of comparisons (five in this case); (d) incomplete unbalanced design, where some comparisons are omitted, and those that are compared have a different number of comparisons; (e) disconnected design, the graph of comparisons is not connected (conditions one, two and three are connected, but disconnected from conditions four and five).

the same standard deviation $\beta_*$ for different comparisons. It is important to note that this standard deviation describes the inherent inter- and intra-observer variations, and it is not an estimate of the measurement noise due to the limited sample size (standard error of the mean). As both are often confused in the context of pairwise comparison experiments, I will discuss these differences in greater detail in Section 5.5.

**Thurstone and Bradley-Terry models** Another frequently used observer model is a Bradley-Terry model [10]. The main difference between Thurstone Case V and Bradley-Terry models is that in the latter, the difference between quality scores is expressed using a logistic distribution instead of a normal distribution. Logistic distribution allows for a more efficient numerical solution when optimizing quality scores [125], however, it has limitations when used in the probabilistic formulation, as it is not conjugate to Gaussian prior. When a logistic distribution describes the difference, individual quality measurement are described by the Gumbel distribution [125], shown in Figure 2.2a. It can be seen that the Bradley-Terry observer model is not symmetric. However, it leads to a very similar description of the difference in quality scores, as shown in the next section. In my dissertation I focus on the Thurstone Case V, however, the findings also generalize to the Bradley-Terry model.

### 2.2.3 Pairwise comparisons and psychometric scaling

To project pairwise comparison data to one dimension, while maintaining as accurately as possible individual relationships between conditions, one requires to run an optimization procedure. In this subsection, I describe the ways of mapping the results of pairwise comparison experiments to a quality scale. Procedures described here are applied in Chapter 3 to map pairwise comparisons of one of the largest datasets to an interpretable scale and in Chapter 4, for building an active sampling procedure.

The results of a pairwise comparison experiment are usually arranged in a pairwise comparison matrix $\mathbf{C}$, in which element $c_{ij}$ counts the number of times stimulus $i$ was chosen as better than $j$. Depending on what and how many times pairs of conditions are compared, full, incomplete, balanced, and unbalanced matrices $\mathbf{C}$ are distinguished (Figure 2.3). Pairwise comparison experiments can be viewed as a graph, in which conditions represent nodes and comparisons edges. A necessary condition for mapping pairwise comparisons to a scale is a connected graph of pairwise comparisons. An example of a disconnected graph is given in Figure 2.3e (top).

Figure 2.4 shows a graphic representation of four steps in scaling pairwise comparisons.

1. First pairwise comparisons are aggregated into a matrix of comparisons.
2. Comparisons are mapped to an empirical probability of one condition being better than another.
3. Probabilities are then converted to distances in the quality scale.
4. Then, assuming an observer model, such as those described in Section 2.2.2, individual quality scores are recovered from distances with an optimization procedure.

Below I discuss each step in more detail.

Obtaining empirical probabilities $\hat{p}_{ij}$ of a condition $o_i$ being better than condition $o_j$, from the matrix of pairwise comparisons is straightforward:

$$\hat{p}_{ij} = \frac{c_{ij}}{c_{ij} + c_{ji}}, \quad i \neq j. \tag{2.4}$$

To obtain the scores, their connection with empirical probabilities needs to be established. Since scores are relative, when scaling pairwise comparison data, we can only recover the distance $q_{ij}$:

$$q_{ij} = q_i - q_j, \tag{2.5}$$

between underlying quality scores $q_i$ and $q_j$.

In the Thurstone observer model, we rely on the fact that the difference between two normal random variables $\omega_i \sim \mathcal{N}(q_i, \beta_i^2)$ and $\omega_j \sim \mathcal{N}(q_j, \beta_j^2)$ representing perceived quality, is also a

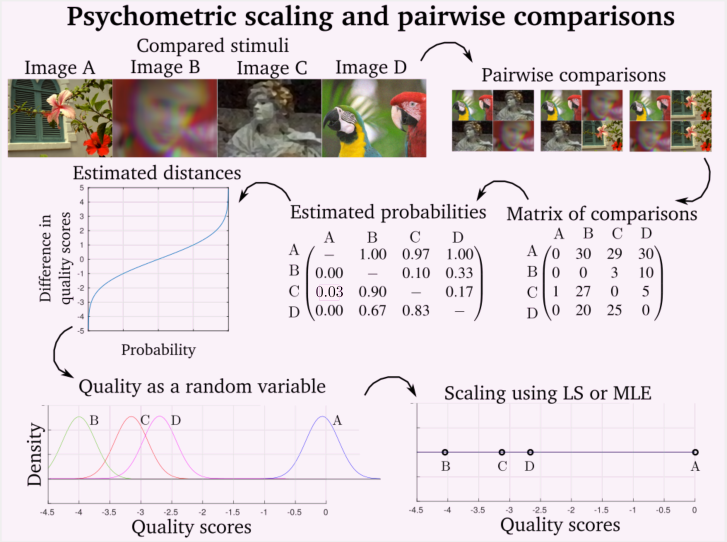



Figure 2.4: Graphic representation of scaling with pairwise comparisons.

normal random variable, as shown in Equation 2.3. The probability of choosing condition $o_i$ over $o_j$ can thus be computed using the cumulative distribution over the difference $\omega_i - \omega_j$:

$$P(\omega_i > \omega_j | q_i, q_j, \beta^2) = \Phi(q_{ij}, \beta^2) \approx \hat{p}_{ij}, \tag{2.6}$$

where $\Phi(\cdot)$ is the cumulative distribution function associated to the chosen observer model, i.e., the normal distribution in Thurstone model and the logistic function in Bradley-Terry model. Note that $\beta$ determines the relationship between distances in the quality scale and probabilities of better-perceived quality. Parameter $\beta$ is often set to $1.4826$ for Thurstone Case V (to $0.9102$ for Bradley-Terry) so that when conditions are 1 unit apart in the quality scale, 75% of observers select one condition over another ($p_{ij} = 0.75$). Such units are referred to as Just-Objectionable-Differences (JOD)s or Just-Noticeable-Differences (JND)s [100] and are discussed later in Section 2.2.7. The mapping between the probabilities and distances for JOD/JND units is shown in Figure 2.2b. To obtain the distances $q_{ij}$ from the probabilities $\hat{p}_{ij}$, we can use the inverse cumulative distribution $q_{ij} = \Phi^{-1}(\hat{p}_{ij}, \beta^2)$. In the following sections, I will explain the procedures used to elicit the scores.

#### 2.2.3.1 Maximum likelihood estimation

Psychometric scaling aims to find estimated scores $\hat{\mathbf{q}}$ such that distances between scores closely resemble distances $q_{ij}$. The simplest way is to solve a least square optimization [28] of the form:

$$\hat{\mathbf{q}} = \underset{q_1,\ldots,q_n}{\operatorname{argmin}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} ((q_i - q_j) - \Phi^{-1}(\hat{p}_{ij}, \beta^2))^2. \tag{2.7}$$

Since scores are relative, the optimization needs to be constrained. In practice, either the first condition, $q_1$ or the average of all conditions is set to 0. The least squares solution is simple but presents limitations when: (i) unanimous answers, in which all observers agree ($p_{ij} = 0$ or $p_{ij} = 1$), result in an infinite distance between $A_i$ and $A_j$, and (ii) confidence in the measurements is not considered, where the total number of performed measurements $c_{ij} + c_{ji}$ is not accounted for.

A more elegant solution is provided by the maximum likelihood estimation (MLE), where the probability of observing pairwise comparisons $c_{ij}$ given latent quality scores $q_i$ is explained by the Binomial distribution:

$$P(\mathbf{C}|\mathbf{q}, \beta) = \prod_{i,j} \binom{n_{ij}}{c_{ij}} \Phi\left((q_i - q_j), \beta^2\right)^{c_{ij}} \left(1 - \Phi\left((q_i - q_j), \beta^2\right)\right)^{n_{ij} - c_{ij}}, \tag{2.8}$$

where $n_{ij} = c_{ij} + c_{ji}$ and $\Phi$ is the cumulative distribution from Equation 2.6. Given the probability in Equation 2.8, the latent quality scores can be found using the maximum likelihood estimation:

$$\hat{\mathbf{q}} = \underset{q_1,\ldots,q_N}{\operatorname{argmax}} \mathcal{L}(\mathbf{q}|\mathbf{C}, \beta). \tag{2.9}$$

More information on this formulation can be found in [100].

#### 2.2.3.2 Maximum a posteriori estimation

When prior information about the scale is available, one can use it to improve the accuracy of the solution. This prior is included in the objective function forming a maximum a poseteriori estimation (MAP). The probability from Equation 2.9 becomes:

$$\hat{\mathbf{q}} = \underset{q_1,\ldots,q_N}{\operatorname{argmax}} \mathcal{L}(\mathbf{q}|\mathbf{C}, \beta) p(\mathbf{q}), \tag{2.10}$$

where $p(\mathbf{q}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\beta^2}} e^{-\frac{(\mu_{\mathbf{q}} - q_i)^2}{2\beta^2}}$ and $\mu_{\mathbf{q}} = \frac{1}{N}\sum_1^N q_i$, is the mean of all scores $\mathbf{q}$. One of the cases where MAP significantly improves the estimation accuracy is when unanimous answers are present. These put no upper bound on the distance between two conditions. In these cases either, normal prior acting like ridge regularization [38] or bounded prior can be used [100].

#### 2.2.3.3 Bayesian approach to psychometric scaling

MLE and MAP solutions provide point-estimates of the quality. However, it is possible to recover the distribution of the scores when the Bayesian approach to psychometric models is used. In Chapter 4, I use the score distribution, obtained following the steps described in this section, to actively sample pairwise comparisons.

**Estimation Model** The model is similar to Thurstone's model Case V [123], with unobserved normally distributed independent random variables. However, given the approach is fully Bayesian, and so instead of point value scores $q_i$ for each condition $o_i$, it is assumed that each score is a random variable $r_i$ with a distribution $r_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Analogous to Thurstone's model, $r_i$ represents the distribution of the score $q_i$, with the mean $\mu_i$ and the uncertainty $\sigma_i^2$ in an estimate of $q_i$. $\sigma_i^2$ is not explicitly expressed in Thurstone's model (it can be obtained, for example by bootstrapping [100]). To avoid confusion in the subsequent chapters I refer to $q_i$ as a point estimate and to $r_i$ as a random variable. Similar to Equation 2.6, by noting that $r_i - r_j \sim \mathcal{N}(\mu_i - \mu_j, \sigma_{ij}^2)$, the probability that $o_i$ is better than $o_j$ is given by:

$$P(o_i \succ o_j | r_i, r_j) \triangleq \Phi\left(\frac{\mu_i - \mu_j}{\sigma_{ij}}\right), \tag{2.11}$$

where $\Phi$ denotes the cumulative density of a standard normal distribution function and $\sigma_{ij}^2 = \sigma_i^2 + \sigma_j^2 + \beta^2$, with $\beta$ representing an observer/comparison noise. I further assume Thurstone Case V model in which $\beta$ is constant across all conditions. The difference here is that this formulation explicitly models both the variance due to measurements error ($\sigma$) and the variance due to variability within and across observers ($\beta$).

For a pair of compared conditions $A_t = (o_i, o_j)$ for $t \in \{1, \ldots, T\}$, where $T$ is the total number of comparisons measured so far, I denote the comparison outcome as $y_t \in \{-1, 1\}$, where $1$ indicates that $o_i$ was preferred and $-1$ indicates that $o_j$ was preferred, with no draws allowed. In the inference step, we want to estimate the distribution of score variables $\mathbf{r}$ given $\mathbf{y}$ and $\mathbf{A} \triangleq \{A_1, \ldots, A_T\}$. The posterior distribution is:

$$P(\mathbf{r}|\mathbf{y}, \mathbf{A}) = \frac{P(\mathbf{y}|\mathbf{A}, \mathbf{r}) \cdot p(\mathbf{r})}{P(\mathbf{y}|\mathbf{A})}, \tag{2.12}$$

where a factorizing normal prior distribution over scores $p(\mathbf{r}) \triangleq \prod_{i=1}^{n} \mathcal{N}(r_i; \nu_i, \alpha_i^2)$ is assumed, $\nu_i$ and $\alpha_i^2$ being the parameters of the prior, usually set to $0$ and $0.5$, respectively. The likelihood $P(\mathbf{y}|\mathbf{A}, \mathbf{r})$ of observing comparison outcomes $\mathbf{y}$ given the ground truth scores is modelled as:

$$P(\mathbf{y}|A, \mathbf{r}) = \prod_{t=1}^{T} P(y_t|A_t, \mathbf{r}), \tag{2.13}$$

where individual likelihoods can be defined as $P(y_t|A_t, \mathbf{r}) = \mathbb{I}\left(y_t = \text{sign}(r_i - r_j)\right)$, i.e., equal to $1$ if the sign of $y_t$ is the same as that of the difference $r_i - r_j$ and 0 otherwise.

**Posterior Estimation Inference** Figure 2.5 shows a factor graph implementing the distribution $P(\mathbf{r}|\mathbf{y}, A)$, used as the basis for efficient inference, and built based on the TrueSkill algorithm [41]. A factor graph is a bipartite graph describing a joint probability with variable (circles) and factor nodes (squares). It is used to factorize the probability distribution functions to enable more efficient computations. In the general case of $n$ conditions and $T$ comparisons, we will have $n$ score variables and prior factors, $T$ difference factors, difference variables, output factors,

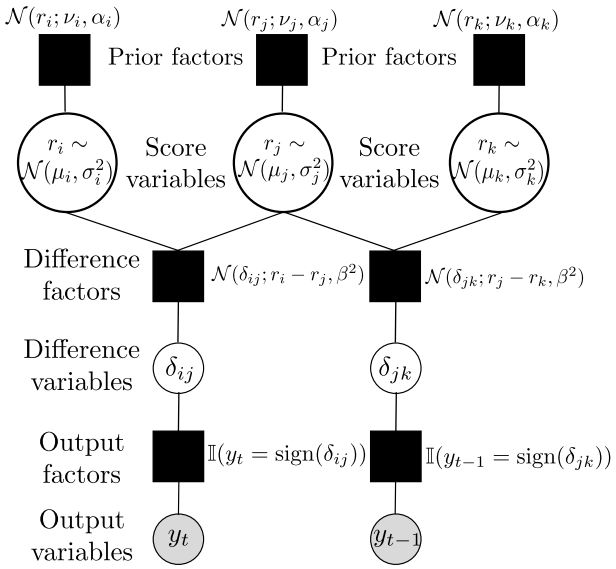



Figure 2.5: Factor graph for two comparisons of three condition implementing the distribution $P(\mathbf{r}|\mathbf{y}, A)$. Real problems may have hundreds of conditions and thousands of comparisons.

and output variables. The posterior over $r_i$ is inferred via message passing between nodes on the graph, with messages computed using the sum-product algorithm [66]. The sum-product algorithm can be described by:

- variable node messages: the product of all messages received on the edges;
- factor node message: the product of all messages received on the other edges multiplied by the factor function and summed over all variables except the one being sent to;
- observed node message: a point mass at the observed value.

The algorithm begins with messages sent from leaf nodes and factors. If the graph is acyclic and the messages can be expressed and calculated exactly each message is computed once, otherwise several iterations are required to reach convergence. In our case, although the score posterior can be written via Bayes rule, the binary nature of the output factor means that the likelihood in Equation 2.13 is not conjugate to the Gaussian prior ($p(\mathbf{r})$). This would lead to a non-Gaussian posterior for $r_i$, and result in challenging, high-dimensional integrals for the inference. A Gaussian approximation to messages yields a multivariate Gaussian posterior with a diagonal covariance matrix, resolving both issues. To approximate messages from the binary output factors as Gaussians, Expectation Propagation via moment matching [91] is used, where the mean and variance of the Gaussian is matched to those in the original distribution. That requires iterations of the algorithm.

#### 2.2.3.4 Limitations

As any model psychometric scaling discussed above has several limitations: (i) observers or repetitions effects are not considered (specific scaling models can account for this); (ii) as

discussed above, the model represents quality in a one-dimensional scale, however, due to transitivity violations present in the data, where, for example, for three conditions $o_a$, $o_b$ and $o_c$, $o_a \succ o_b$ and $o_b \succ o_c$ and $o_c \succ o_a$, a one-dimensional scale might not be enough to represent quality scores [13] and (iii) case V of Thurstone model was used, however, quality scores may have different variances. I discuss these in greater detail in the next Chapter 3.

### 2.2.4 Vote counts

If $\Phi$ in Equation 2.6 is set to the uniform distribution, Equation 2.7 corresponds to ranking conditions according to $\hat{q}_i = (1/n)\sum_{j=1}^{n} c_{ij}$. This last idea is usually referred to as vote or borda counts [23] – the number of times one condition was selected as better than any other condition. However, this approach has limitations, i.e., $q_i \geq q_j$ doest not imply $p_{ij} \geq \frac{1}{2}$, when there are transitivity violations in the data or when conditions are not compared the same number of times. In those situations, psychometric scaling algorithms are usually preferred. These algorithms will produce correct ranking and capture the magnitude of the differences between conditions in a principled way. Moreover, vote counts do not explicitly account for the relative quality difference between the conditions, whereas psychometric scaling infers the scores by considering relationships among all compared conditions.

In this regard, Zerman et al. compared psychometric scale and vote counts to the scores obtained in a direct rating experiment [149]. They showed that psychometric scaling scores are stronger related to rating scores than vote counts, confirming that quality magnitudes are better captured when pairwise comparison data are scaled. In Chapter 3 I conduct a set of experiments to verify that observation.

### 2.2.5 From ratings to a scale

Obtaining a scale from rating experiments does not require an optimization procedure and relies on averaging the scores from individual observers per condition. This average is called Mean Opinion Scores (MOS). The scores provided by MOS, when test stimuli correspond to different reference stimuli, are not comparable. To take the relative ratings into account they are assumed to be anchored at the same values. In this case Differential Mean Opinion Scores (DMOS) [46, 127] is used. To obtain DMOS scores, the MOS scores given to the test stimuli are subtracted from the scores given to the corresponding reference stimuli. For a test stimulus $j$ with a corresponding reference $i$ and judged by observer $k$, with ratings $m_{ijk}$ and $m_{ik}$ respectively, the DMOS is computed as follows:

$$d_{ijk} = m_{ik} - m_{ijk}. \tag{2.14}$$

Essentially, DMOS also represents the amount of impairment from the reference stimulus, similar to the psychometric scaling results. As a concept, the ‘distance’ found by DMOS to the

undistorted reference stimulus is very similar to the scaling solution with the reference anchored at 0.

It is also common to account for the difference among observers in their perception of the scale by computing Z-score of the ratings per observer:

$$z_{ijk} = \frac{d_{ijk} - \mu_k}{\sigma_k}, \tag{2.15}$$

where $\sigma_k$ and $\mu_k$ are the mean and standard deviation of the scores provided by observer $k$ [84].

### 2.2.6 Pairwise comparisons, requirements for a meaningful scale

The vast majority of image quality assessment studies, employing pairwise comparisons, compare only images depicting the same content, e.g., comparing different distortion levels applied to the same original image. This "apple-to-apple" comparison simplifies the observers' task, but it comes with some limitations. Assessing and scaling each content independently makes it impossible to obtain scores that correctly capture quality differences between conditions across different contents on a common quality scale. Secondly, pairwise comparisons capture only relative quality relations. Therefore, to assign an absolute value to such relative measurements, the experimenter needs to assume a fixed quality for a particular condition, which is then used as a reference for the scaling.

To scale the pairwise comparisons in a consistent manner, there should be no disconnected components in the graph of comparisons. However, when each content is assessed individually, this forms a set of disconnected graphs, each with its relative quality scale. We could potentially anchor each content by assuming that each component's reference has a fixed quality score, for example, 0. However, the accuracy of the scale then suffers from the lack of relative information between the conditions far away in quality from the reference. Thus, connecting these disconnected parts is an essential step for a unified quality scale.

To address these problems, cross-content pairs can be used to connect the disconnected 'nodes' and to eliminate the error accumulation. Additionally, assuming that all the undistorted reference stimuli are equivalent to each other (i.e., having pristine quality with "0" quality score), this graph can be connected at the reference 'node'. All the distorted images would then have negative quality values after scaling, corresponding to the distortions compared to the undistorted reference stimuli (unless enhancement is considered).

### 2.2.7 JNDs and JODs

Results of pairwise comparisons are typically scaled in Just-Noticeable-Difference (JND) units [28]. Usually, the scale is constructed such that two stimuli are 1 JND apart when 75% of observers can see the difference between them. However, considering measured differences as

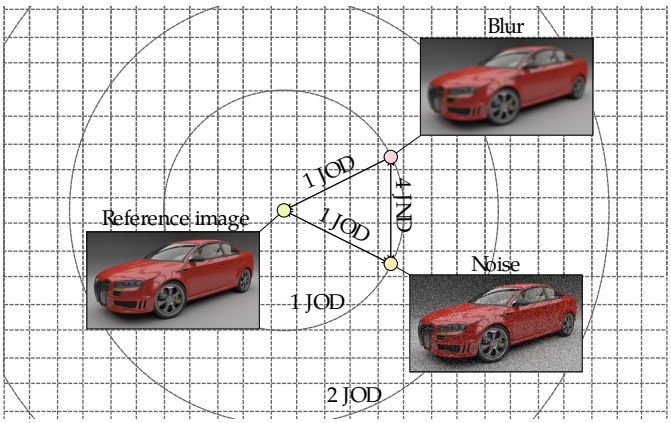



Figure 2.6: Illustration of the difference between just-objectionable-differences (JODs) and just-noticeable-differences (JNDs). The images affected by blur and noise may appear to be similarly degraded in comparison to the reference image (the same JOD), but they are noticeably different and, therefore, several JNDs apart. The mapping between JODs and JNDs can be very complex and the relation shown in this plot is just for illustrative purposes.

"noticeable" leads to an incorrect interpretation of the experimental results. Let us take as an example two distorted images shown in Figure 2.6: one image is distorted by noise, another by blur. They are noticeably different, and, intuitively, should be more than 1 JND apart. However, the question asked in an image quality experiment is not whether they are different, but, which one is closer to the perfect quality reference. Note that a reference image does not need to be shown to answer this question as we usually have a mental notion of how a high-quality image should look like. Therefore, the collected data is not related to noticeable differences between images, but rather to image quality difference in relation to a perfect quality reference. For that reason, this quality measure is often described as JOD rather than JNDs [149]. Note that JOD is the measure of impairment and not overall image aesthetics and, therefore, is related to DMOS rather than to MOS. Note also that JOD does not replace JND, and the term JND is still more appropriate for the tasks that involve direct discrimination between a pair of conditions. Since JOD scores are relative to the anchored reference condition, they are particularly suitable for image fidelity metrics, the focus of my dissertation. For the same reason these are more suitable for scaling conditions that are not substantially different from the pristine reference condition. With no bound on the distance from the reference, estimated scores for the conditions far away from the anchor are likely to be inaccurate.

The relation between JOD values and the probability of selecting condition A over condition B is illustrated in Figure 2.2b. When an equal number of observers vote for both conditions, the probability is 0.5 and JOD difference between the conditions is 0. The differences of 1 JOD, 2 JOD and 3 JOD correspond to the probabilities $P(A > B)$ of 0.75, 0.91, and 0.97. The negative JOD values indicate that more observers preferred B over A. In all examples through out the dissertation I assume that the reference condition is at 0 JOD. Because of that, most reported JOD scores are negative (worse than the reference).

## 2.3 Objective quality assessment

In the previous section, I discussed how to elicit subjective quality from psychometric experiments with human participants. The scores obtained from subjective experiments are useful in training and testing automated or objective quality metrics. Thus, this section gives an overview and the background of modeling the perceived image quality. I first discuss various quality assessment criteria. I then talk about the dynamic range of a scene and the ways of modeling the perceived dynamic range. Finally, I discuss different approaches to objective IQA.

### 2.3.1 Quality assessment criteria

There are at least three standard criteria related to image quality: aesthetics, visibility, and impairment. Aesthetic judgments are concerned with the quality of an image as judged by commonly established photographic rules, such as appropriate lighting, contrast, and image composition. Here, the quality may be perceived in terms of creative composition and execution of an image, rather than artifacts [25]. Visibility assessment predicts whether a difference between a pair of images is going to be visible, but does not assess the magnitude of a distortion [141]. It also produces visual difference maps rather than a single quality score. Impairment assessment, which is the focus of this work, assesses the quality of distorted test images, with artifacts such as noise, blur, compression, etc., with reference to its original undistorted version. As a specific case, image enhancement, treated as part of image aesthetics falls outside the scope of this dissertation. However, when enhancement introduced to an image is treated as the distortion of the reference image, it falls into the impairment assessment category, discussed within this work.

### 2.3.2 Existing approaches to objective IQA

Image quality metrics can be divided into full-reference, where all information about the undistorted reference image is available, reduced-reference, where partial information is available, and no-reference, where no information about the reference image is available. The metrics can be either model-based or data-driven. Model-based metrics aim to capture image difference statistics relevant to quality. These metrics do not require extensive training and their performance depends on assumptions made about the human visual system or image statistics [14, 148]. Data-driven metrics, on the other hand, such as those based on CNNs [60, 31, 151, 107, 59, 77], are capable of extracting features in an automated way, learning complex relationships without the need to make assumptions about the human visual system. However, these desirable properties come with limitations – complex machine learning models are susceptible to the quality and quantity of the training data. If data are scarce, the model will fail to generalize. To alleviate the problem of insufficient and noisy data, transfer learning is often used. For example, authors in [31, 2, 72, 7]

pre-trained the CNN model on image classification tasks, arguing that learned features would capture image statistics important for IQA. Others [151] pre-trained the network on the quality predictions of the model-based quality metrics, and then, fine-tuned on the subjective image quality scores. Authors in [80] exploited yet another approach — they first pre-trained the network to classify distortions. The risk of this approach, however, is that it can overfit the model to the given set of distortions. Other works were based on training on image ranking [77, 107, 151]. The advantage of that approach is that training can be performed directly on the pairwise comparison data. But the shortcoming of this approach is that it discards meaningful information by converting the quality scores to a binary classification problem. Although all those approaches can improve the ability of the ML-metrics to generalize, they do not substitute the need for a larger and diverse IQA dataset, which was acknowledged in most works. Collecting such a dataset is one of the main objectives of my dissertation.

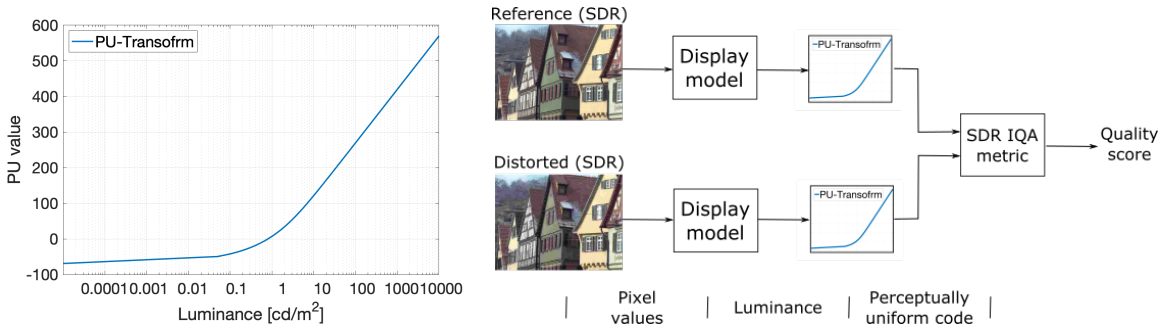
### 2.3.3 Dynamic range of an image

Perceived image quality depends on the brightness of the display [145]. The human eye can perceive luminance in the range from $10^{-5}$ to over $10^4$ cd/m$^2$ [43]. However, in typical SDR displays, the colour is encoded in 8 bits, bound in a rather small range of integer numbers between 0 and 255. This range was sufficient in the era of cathode ray tube (CRT) monitors, where the difference between the brightest and the darkest points of a displayed image (its dynamic range (DR)) was between 1 and 100 cd/m$^2$ [83]. Modern display technologies allow for much larger DR, and thus a higher amount of transmitted information. The aim of expanding DR of gamma-corrected images is to achieve that of the real world. Not only are the images brighter, but they are also perceived as more colorful [83] and realistic [57]. Human observers also reported better depth perception for brighter images, with average luminance of 1000 cd/m$^2$, than darker, with average luminance of 50 cd/m$^2$ [131].

An image captured by a camera is represented in radiance. It then undergoes post-processing steps, including compression via gamma-encoding and tone-mapping [43], resulting in an SDR image. During this process brighter regions of the images are compressed more – the human eye is less susceptible to high spatial frequency variations in the brighter areas [116]. Whereby this can described as:

$$I_{\mathrm{gamma-encoded}} = (I_{\mathrm{linear}})^{\frac{1}{\gamma}}\,, \tag{2.16}$$

where $\gamma$ is usually set to $2.2$. SDR images are thus represented with gamma-encoded values. This process of transferring linear luminance values to digital representation is also referred to as forward display model. The inverse display model is the backward process of mapping digital values to linear luminance values. Here some model of the display is assumed. The most




(a) PU encoding (b) Extending SDR metrics to HDR content

Figure 2.7: (a) PU-transform used to transform absolute physical values (in cd/m$^2$) into approximately perceptually uniform units that can be used with existing quality metrics. Note, some values are negative – to ensure that the luminance range of a typical (sRGB) monitor of 1 cd/m$^2$ to 80 cd/m$^2$ is mapped to the range 0–255. (b) Pipeline for extending SDR metrics to photometric values. An SDR image is first converted from gamma-corrected pixel values to linear luminance values, via a display model. For HDR images the "Display model" is omitted, as those already store luminance values. Luminance values are then passed via the PU-transform to obtain perceptually uniform code. This code is then passed to an SDR image quality metric.

practical one is gamma-offset-gain (GOG) display model [35]:

$$C_{\text{displayed}} = (L_{\text{peak}} - L_{\text{black}})\, C_{\text{original}}^{\gamma} + L_{\text{black}} + L_{\text{refl}}, \tag{2.17}$$

where $\gamma = 2.2$, $C_{\text{lin}}$ is linear, $C_{\text{sRGB}}$ is the gamma-encoded colour value for one of the channels (R, G or B), $L_{\text{peak}}$ is the the peak luminance of a given display, $L_{\text{black}}$ is the black level and $L_{\text{refl}}$ is the ambient light that is reflected from the surface of a display (usually assumed to be 0).

Different from SDR images, which are represented with gamma-encoded values, HDR images store linear luminance values. There are several approaches to producing an HDR image – either inverse tone-mapping, where a SDR image is mapped to luminance values, or multiple SDR images with different exposures are processed together for wider luminance range [43].

### 2.3.4 Dynamic range in objective IQA

As the range of luminance a display can cover affects the visibility of distortions of a viewed image, a reliable quality metric should account for it. I will refer to the metrics that operate on physical photometric/luminance values as *photometric quality metrics*. HDR quality metrics, such as HDR visual difference predictor (HDR-VDP) [82] or HDR video quality measure (HDR-VQM) [99]), are photometric to account for a large range of luminance produced by HDR displays. SDR metrics can also be adapted to operate on photometric quantities [3]. For that, luminance values of HDR images need to be converted into Perceptually Uniform (PU) or logarithm values, with the former achieving better results [3, 62]. Perceptually uniform values are then passed to an SDR image quality metric.

PU transform is derived from the contrast sensitivity function (CSF) that predicts detection thresholds of the human visual system for a broad range of luminance adaptation conditions. This transformation is necessary as the response of the human eye to luminance is not linear. We perceive relative, rather than absolute difference in luminance. To adapt quality metrics accordingly, we transform linear luminance values of images to the perceived luminance via PU encoding. The transform is given by:

$$PU(L) = \int_{L_{min}}^{L} \frac{1}{T(l)} dl, \tag{2.18}$$

where $L_{min}$ is the minimum luminance to be encoded, $T(l)$ is the detection threshold of absolute luminance $l$ defined as:

$$T(l) = S\left(\left(\frac{C_1}{l}\right)^{C_2} + 1\right)^{C_3}, \tag{2.19}$$

where $S$ is the absolute sensitivity constant and $C_1$, $C_2$, $C_3$ are scaling parameters. Throughout my dissertation I use the values for parameters from [82]. The transformation is further constrained to map luminance values typically reproduced on SDR monitors (0.8-80 cd/m$^2$) to a range $0 - 255$. Thus, resulting quality values for SDR images converted to photometric units assuming a display with (0.8-80 cd/m$^2$) dynamic range correspond to those without such transformation. The shape of the PU-transform is given in Figure 2.7a.

Similar procedure can be applied to evaluate quality of SDR images when these are shown on displays with different brightness. Before passing through the PU-transform an SDR image is first transformed to luminance emitted from a display, assuming a model of that display, such as the one in Equation 2.17. The full pipeline of making SDR image quality metrics photometric is given in Figure 2.7b.

## 2.4 Summary

In this Chapter, I provided background on subjective and objective quality assessment methods. I described two most commonly used subjective assessment protocols, rating, and ranking. These are the core experimental procedures used in my dissertation and will be used to collect data in Chapter 3, Chapter 4, and Chapter 5. While the construction of the scale from rating experiments is trivial, scaling pairwise comparisons requires several assumption on how individual judgements are distributed (an observer model). I discussed the two widely used observer models for scale construction from pairwise comparisons, Bradley-Terry [10], and Thurstone [123]. I have also discussed the requirements for constructing a meaningful scale from pairwise comparisons in image quality assessment. Although in many subjective studies involving human participants pairwise comparisons are scaled in JND (just-noticeable difference) units, I showed that a more appropriate choice for image quality assessment is JOD (just-objectionable-difference) units. I

then discussed areas forming the field of IQA, objective IQA metrics, and the influence of the dynamic range on the perceived quality and ways of embrasing it within objective quality metrics. These topics are relied upon in Chapter 6, where I propose a new subjective photometric IQA dataset, in Chapter 7, where I test objective IQA metrics and Chapter 8, where I present a new dataset with quality threshold for viewing condition dependent visually lossless compression.

# Chapter 3

# Psychometric image quality assessment

## 3.1 Introduction

In Chapter 2 I introduced the protocols for conducting subjective image quality assessment experiments and ways of obtaining an accurate and interpretable scale from the collected measurements. In practice, however, many of the requirements for obtaining such a scale are neglected. An example dataset, where these requirements are violated, is a widely used benchmark subjective image quality dataset, TID2013 [106]. The scores in this dataset were obtained from pairwise comparisons with vote counts, lacking an observer model, and, important for an accurate scale, cross-content and with-reference comparisons were not performed (Section 2.2.6). The dataset is used for training and testing objective quality metrics. Thus, the limitations present in the dataset, if not addressed, propagate, resulting in poorly tuned objective quality metrics and misinterpreted results. In this chapter, following the procedures for an accurate and interpretable scale, I improve the scores of the TID2013 dataset[1]. Doing so required running additional psychometric experiments with human participants. Using the original data from the TID2013 dataset and newly collected measurements, I analyze the requirements necessary for scaling image quality datasets and validate assumptions set in Chapter 2.

The work in this chapter is based on my publications at IEEE Transactions on Image Processing [108] and IEEE Conference on Quality of Multimedia Experience [88]. However, I also extend the published work with Section 3.6.

## 3.2 TID2013 dataset

One of the most recent and extensively evaluated image quality datasets, TID2013 [105, 106], uses a crowd-sourced experiment with pairwise comparisons to measure image quality. The quality scale, in this dataset, was obtained from vote counts. TID2013 presents 3000 distorted

[1] I make the re-scaled TID2013 available online: `https://doi.org/10.17863/CAM.21517`

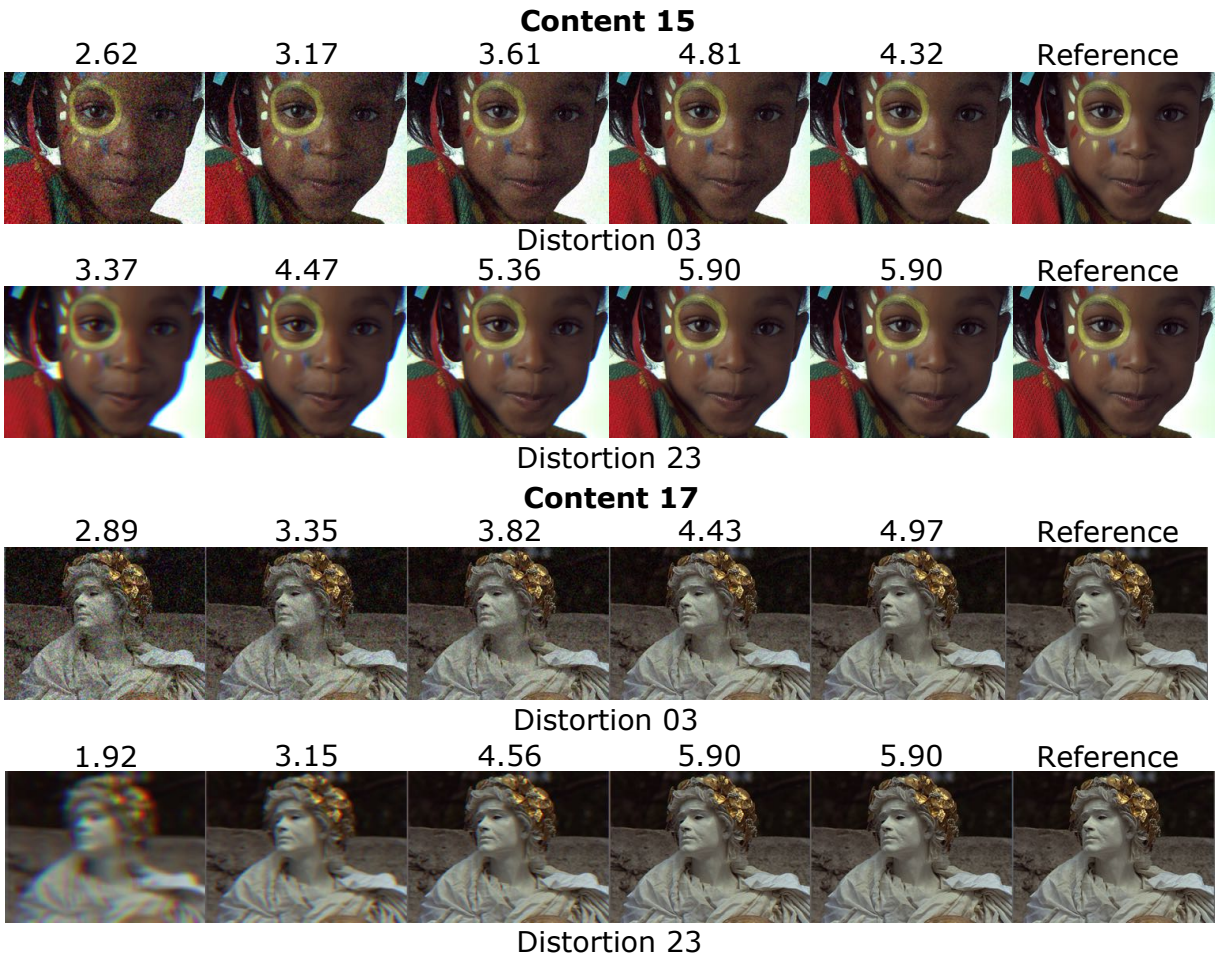



Figure 3.1: Example images from the TID2013 dataset for two contents and two distortion types at five distortion levels. Numbers on top of the images, are the corresponding vote count based quality scores, found in the original dataset.

conditions (25 contents, 24 distortion types, and 5 levels of distortions). Example images from the TID2013 dataset are given in Figure 3.1 Approximately 30 observers were involved in the measurement of every content. In total, more than 900 observers participated in the experiments, completing over 400,000 comparisons.

Each observer in one experiment only performed comparisons within one content, i.e., no cross-content comparisons were performed. In the pairwise comparison experiments, the less distorted image was chosen with an aid of the reference image, displayed alongside. The pairs of conditions to compare were chosen using the Swiss chess system [20]. With this method, all conditions are compared the same predefined number of times. The first comparisons are chosen at random. In later stages, conditions are sorted based on the number of times they were previously selected by an observer and conditions having similar quality compete in pairs.

Subjective image quality scores presented in TID2013 are given in vote counts. Vote counts were obtained for each content separately by taking the total number of times a condition was selected as better and dividing by the number of observers. Every observer compared each

condition within one content in nine pairwise comparisons, producing a scale between 0 and 9. Therefore, the matrix of comparisons in TID2013 has an incomplete, unbalanced design with 25 disconnected components and no comparison to the reference. As discussed in Chapter 2, the absence of cross-content comparisons and comparisons with the reference makes it impossible to construct a unified quality scale for all contents.

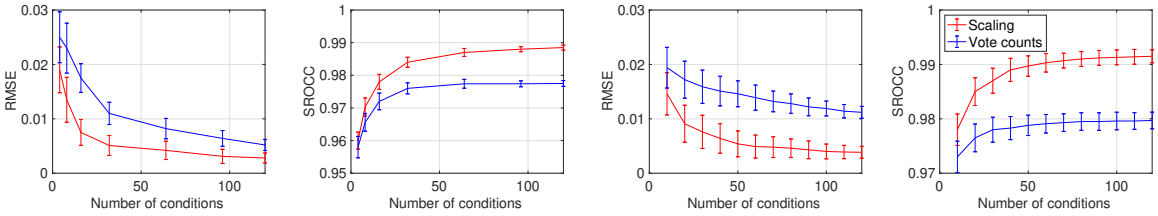
## 3.3 Psychometric scaling and vote counts simulation

To compare the scores produced by vote counts with those produced by psychometric scaling, introduced in Section 2.2.3.1, I use a Monte Carlo simulation. Since it is nearly impossible to obtain the ground truth scores in quality assessment experiments, simulation of these experiments is necessary to draw conclusions about the consistency of both vote counts and psychometric scaling. Two types of data were used as the ground truth in the simulation: i) randomly generated quality scores within a fixed range and ii) TID2013 vote counts for content 1.

### 3.3.1 Simulation procedure

The simulation of an experiment was designed to mimic the experimental procedure from TID2013 i.e. every condition was compared nine times in three random and six sorted rounds using the Swiss system by each of the 30 observers. In the simulation of a comparison between two conditions, every simulated observer chooses condition $o_i$ over $o_j$ with a probability defined by $P(o_i \succ o_j) \sim \Phi(\frac{q_i - q_j}{\beta})$, following Equation 2.6 where I set $\beta = 1.4826$, i.e., assuming that the scores for conditions are given in JOD units. Randomly generated true quality scores $\mathbf{q}$ for the first simulation were uniformly sampled at random from the [0, 9] interval, to match the range of the scale in TID2013. For the simulation with the data from the TID2013 score distribution, true quality scores were randomly sampled from the vote counts for the first content. Comparison matrices produced by the simulated observers were aggregated.

Vote counts are produced by summing elements along the rows of the resultant matrix, i.e., the number of times every image was preferred divided by the number of observers. Psychometric scaling is produced using MLE with the Thurstone Case V model in JOD unites (described in Section 2.2.3.1) and the Matlab code provided in [100]. To compare the scales, I calculate the spearman rank order correlation coefficient (SROCC) and root-mean-squared error (RMSE) between the standardized scores used for the simulation, and those produced by vote counts and psychometric scaling (also standardized). SROCC and RMSE values are averaged for the simulation repeated 1000 times.




(a) (b) (c) (d)

Figure 3.2: (a), (b) Simulation of the experiment with random data; (c), (d) Simulation of the experiment with TID data. Lower is better for RMSE, higher is better for SROCC. Confidence intervals are computed for 75% level.

### 3.3.2 Simulation results

The results of the simulation are depicted in Figure 3.2. Regardless of the number of conditions, psychometric scaling consistently outperforms vote counts in estimating both the ranking (SROCC) and the scale (RMSE). The difference in SROCC between the psychometric scaling and vote counts increases with the number of conditions, and so psychometric scaling is preferable as the number of conditions increases.

## 3.4 Psychometric scaling of TID2013

The experimental procedure in TID2013 presents two important limitations. First, although reference images were used to help observers choose between distorted images, comparisons of the distorted images to the corresponding references were never performed. A common quality anchor for every content, i.e., a reference image, is necessary for constructing a fully connected graph of comparisons. Second, comparisons across different contents were not performed. Without cross-content comparisons, contents cannot be accurately scaled [149]. Therefore original TID2013 scores cannot be compared across different contents. These required types of comparisons are shown in Figure 3.3.

### 3.4.1 Experimental setup

I extend the data collected in TID2013 with five additional experiments. The first experiment was used to include reference images. The next three include cross-content comparisons and further comparisons to the reference to improve the scale. The last experiment is a validation experiment, used to evaluate the quality of scales produced by vote counts and psychometric scaling. I later use comparisons from all five experiments to construct the final scale. Each experiment had ten participants. In a single experiment each participant completed 300 trials. Overall additional 15,000 comparisons were collected. For the design of the first four experiments, I consider that it is often more informative to compare conditions that are of similar quality (which is the

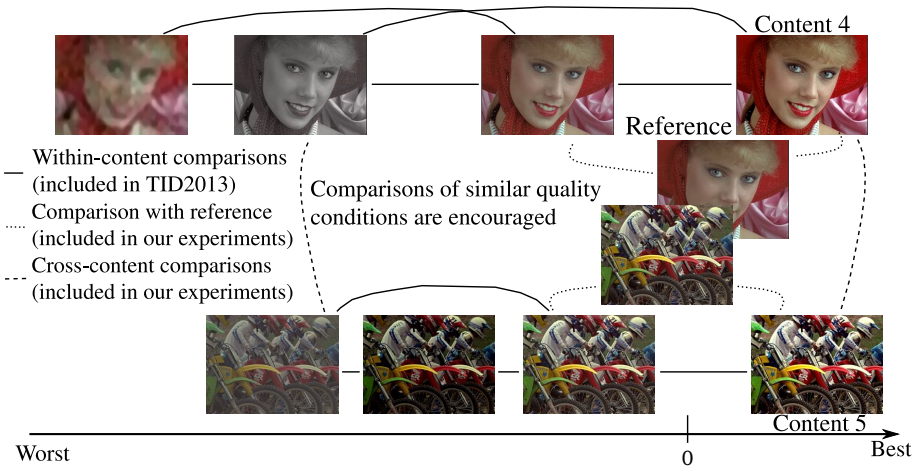



Figure 3.3: Representation of different types of comparisons, necessary to position all contents on a common quality scale.

primary motivation for the use of the Swiss system in TID2013). I asked observers to select the better quality image among distorted ones. The order of comparisons in every experiment was randomized.

Overall 14 observers participated in all experiments, 11 male and three female with most participants performing all three experiments. All participants were research members of the Rainbow group. The mean age of the group was 32 years. The youngest participant was 20 years old and the oldest was 47 years old. Participants were represented by three ethnic groups with four participants from Asian group, eight from white and two from Arab groups. All observers were paid for the participation with 10£ Amazon vouchers per hour. Ethical approvals are provided in Appendix A.1.

#### 3.4.1.1 Inclusion of reference

To select comparisons for the first experiment, I scaled pairwise comparisons from the original TID2013 dataset for each content separately using the MLE based scaling procedure from Section 2.2.3.1. I then compared each reference image to four conditions (within the same content) with the best quality score. For $25$ contents present in TID2013, $25 * 4 = 100$ pairs were selected to include reference images in the quality scale. Each measurement was repeated three times by each of the ten observers. I then included the newly collected data to the dataset and scaled it. I assumed that all reference images have a common quality score of zero. This assumption holds within the image fidelity assessment, where the quality of the test condition is judged by the proximity to the undistorted reference condition. For other quality assessment criteria, for example, aesthetics or where there is no access to the ground truth image, such as evaluation of the tone-mapping algorithms, this assumption might not hold due to the dependency on the content of the image. In these cases validation of the assumption with a benchmark study on the reference conditions is necessary.

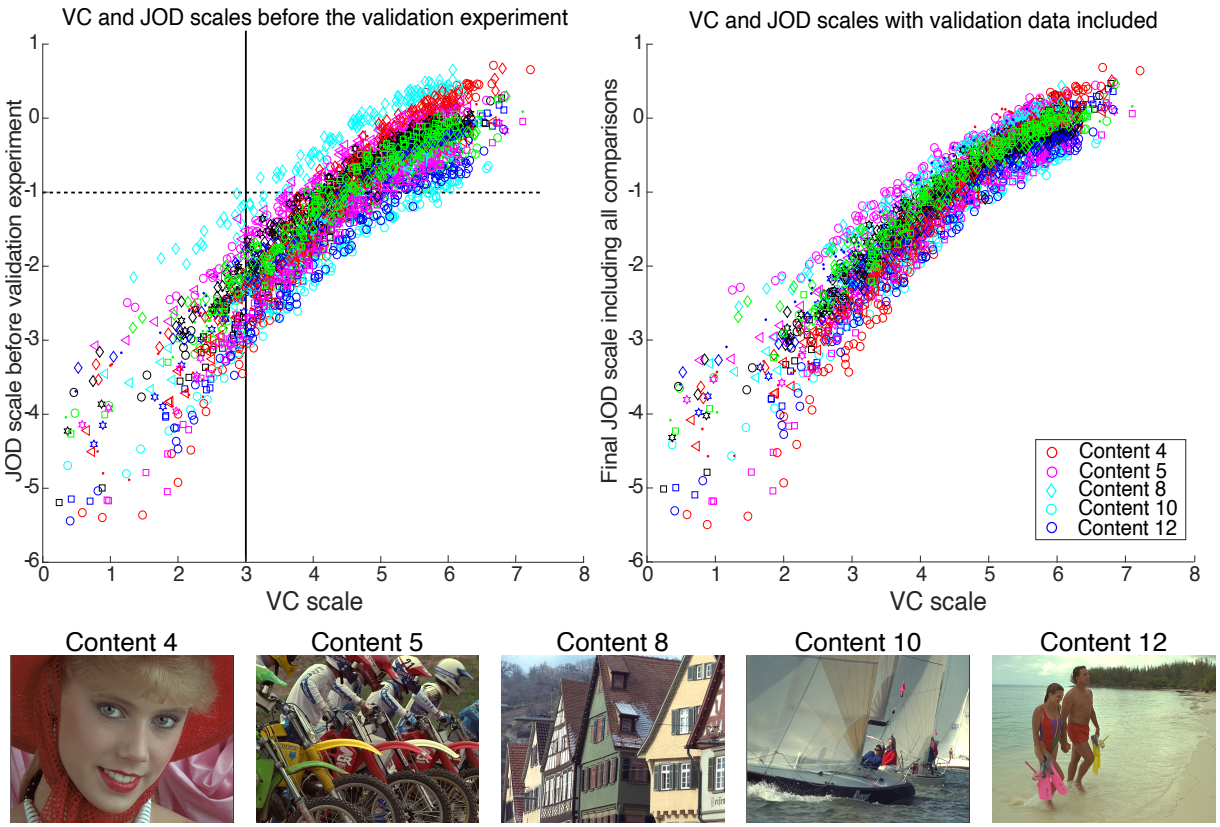


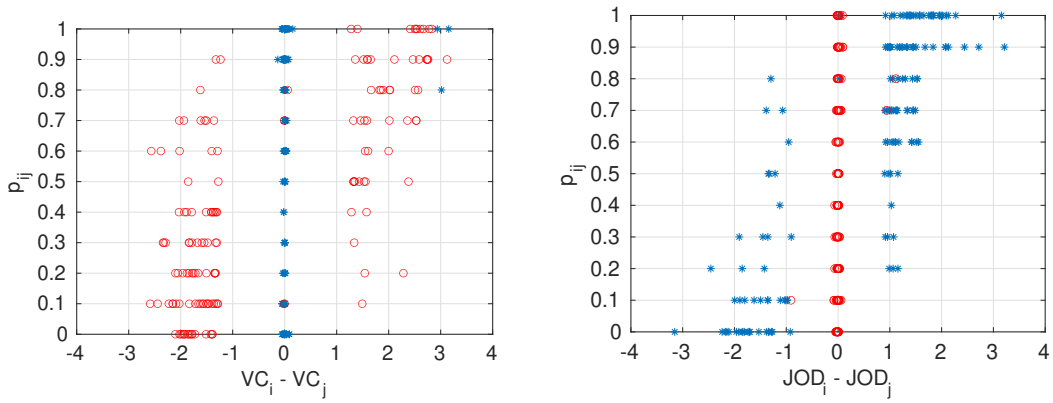

Figure 3.4: Relationship between vote counts (VC) and JOD scales when adding new data. The left part shows the scatter plot of the VC scores versus JOD values using the data from the first four experiments only. The dashed horizontal and the bold vertical lines illustrate the procedure for the maximum differentiation competition. The method is used to select the validation comparisons. To select the pairs of images one scale is kept constant e.g. JOD (horizontal line), and the leftmost and the rightmost conditions are selected to form a pair. Similarly for the vertical line (VC scaled set constant) the bottom and the top condition are chosen for comparison. The plot on the right represents the final scale after including comparisons from all five experiments.

#### 3.4.1.2 Inclusion of cross-content comparisons and scale refinement

For the next three experiments, I included 300 more pairs of conditions, most of which were cross-content, i.e., 40 comparisons to the reference, 240 cross-content comparisons, and 20 within content comparisons. After each experiment, I re-scaled the data and used the produced new scale to select comparisons for the next experiment.

#### 3.4.1.3 Validation experiment

To compare the consistency of both scales on real data, I conducted an additional experiment. Using both scales (as represented in the left part of Figure 3.4), I found the cases in which VC and JOD differ the most. This strategy is referred to as maximum differentiation (MAD) competition [153]. It is used to compare subjective and objective image quality metrics. The hypothesis of the procedure is that for the selected pairs $o_i, o_j$ the empirical probability of one




(a) (b)

Figure 3.5: Scatter plot of the empirical probabilities of one condition being better than another found from the validation experiment versus the distance between these conditions in both VC and JOD scales. Red circles show pairs of conditions for which $VC_i \gg VC_j$ and $JOD_i \approx JOD_j$ and blue starts show pairs of conditions for which $JOD_i \gg JOD_j$ and $VC_i \approx VC_j$.

condition being better than another $p_{ij} = P(o_i \succ o_j) = \frac{c_{ij}}{c_{ij}+c_{ji}}$ would be better reflected in the distances of the more accurate scale. For example, if two images $o_i$ and $o_j$ are close in the VC scale ($VC_i \approx VC_j$) and far apart in the JOD scale ($JOD_i \gg JOD_j$) and the empirical probability of selecting one over another is $p_{ij} \approx 0.5$, and similarly for conditions $VC_i \gg VC_j$ ($JOD_i \approx JOD_j$) the probability $p_{ij} \approx 1$, VC scale would also be more accurate.

Thus, I selected 150 pairs of conditions $o_i, o_j$ for which JOD is as different as possible and VC scores were as close as possible i.e. $\text{argmax}_{ij}(|JOD_i - JOD_j| - |VC_i - VC_j|)$. And similarly, for $o_i, o_j$ for which VCs were different and JODs were similar. To promote diversity, I allowed each content to participate only in 50 comparisons. Overall I selected 300 pairs of images with 150 images in each group and asked ten observers to perform a pairwise comparison experiment.

### 3.4.2 Results and discussion

**Correlation comparison** To compare both scales, I computed the SROCC between the probability $p_{ij} = P(o_i \succ o_j)$, of an image being better, inferred from the pairs of comparisons in the validation experiment (following Equation 2.4) and the difference in quality scores in both VC and JOD scales for the selected pairs. Figure 3.5 shows the scatter plot of the $p_{ij}$ versus differences in both scales. The SROCC between $VC_i - VC_j$ and $p_{ij}$ is 0.52, and between $JOD_i - JOD_j$ and $p_{ij}$ is 0.69, indicating that the output of the new JOD scale better correlated with ranking of the conditions. I then included the data from the validation experiment into the psychometric scale. The SROCC for $JOD_i - JOD_j$ and $p_{ij}$ improved to 0.84, indicating that psychometric scaling can successfully include the information from collected comparisons. Nevertheless, the SROCC is still far from one, as the psychometric scaling finds the best one-

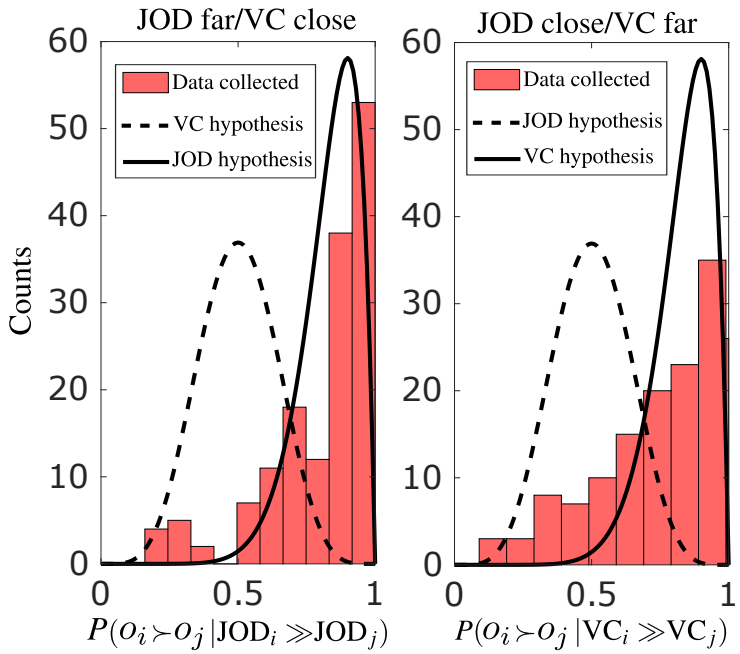



Figure 3.6: Histograms of the probability counts for the validation experiment under two scenarios: (i) selected pairs of conditions $o_i$ and $o_j$ are far in the JOD scale, and close in the VC scale ($JOD_i \gg JOD_j, VC_i \approx VC_j$), left; (ii) selected pairs of conditions $o_i$ and $o_j$ are far in the VC scale, and close in the JOD scale ($VC_i \gg VC_j$, $JOD_i \approx JOD_j$), right. The horizontal axis is the probability of image $o_i$ being better than $o_j$, $p_{ij}$, inferred from the validation experiment. Dashed and solid black curves represent the hypothesis, from which the data might be coming and are re-scaled binomial distributions with parameters $N = 10$ and $p = 0.5$, $N = 10$ and $p = 0.9$ respectively.

dimensional scale, taking into account all relationships in the data. Due to the transitivity violations present in the data, the mapping cannot not be optimal.

**Histogram comparison** The histogram on the left of Figure 3.6 shows probability counts for 150 pairs of conditions $o_i$ and $o_j$ for which $JOD_i \gg JOD_j$ and $VC_i \approx VC_j$. Similarly, the histogram on the right shows 150 probability counts for pairs of conditions $o_i$ and $o_j$ for which $JOD_i \approx JOD_j$ and $VC_i \gg VC_j$. Dashed and solid black curves represent the hypothesis and are re-scaled binomial distributions with parameters $N = 10$ and $p = 0.5$, $N = 10$ and $p = 0.9$ respectively (not 1 since the average distance between the images for the chosen pairs is 2 JOD, corresponding to $\approx 90\%$ of observers choosing one image over another). Ideally, the distribution of the probability counts for a better scale would follow the black solid line. The frequency around $P(o_i \succ o_j) \approx 1$ is higher for JOD (left plot of Figure 3.6). The frequency for the VC scale gradually increases with $p_{ij}$, whereas for the JOD, the change is abrupt, with a sharp rise of the counts (right plot of Figure 3.6). The frequency for high probability $P(o_i \succ o_j) > 0.7$ is also greater for the JOD scale. Nevertheless, the plot on the right does not exactly follow the JOD hypothesis (bell shaped dashed distribution centered at 0.5), meaning that it can be improved.

Table 3.1: Log-likelihood of observing the data collected in a validation experiment, assuming binomial distribution.

| Data | Assumption | Log-likelihood |
|---|---|---|
| $JOD_i \gg JOD_j$ and $VC_i \approx VC_j$ | $o_i \approx o_j$ | -672.98 |
| $JOD_i \gg JOD_j$ and $VC_i \approx VC_j$ | $o_i \gg o_j$ | -393.34 |
| $VC_i \gg VC_j$ and $JOD_i \approx JOD_j$ | $o_i \approx o_j$ | -546.98 |
| $VC_i \gg VC_j$ and $JOD_i \approx JOD_j$ | $o_i \gg o_j$ | -543.47 |

**Log-likelihood comparison** Another way of evaluating the consistency of both scales is to compute the log-likelihood of observing the data collected in the validation experiment, assuming binomial distribution, for two hypotheses: $o_i \approx o_j$, and hence $P(o_i \succ o_j) = 0.5$ and $o_i \gg o_j$, and hence $P(o_i \succ o_j) = 0.9$:

$$\mathcal{L} = \sum_{k=1}^{K} logBin(c_{ij}^k, N, p), \tag{3.1}$$

where $p = \{0.5, 0.9\}$ corresponding to two hypothesis, $N = 10$, since I performed 10 comparisons per pair and $c_{ij}$ is the number of times condition $i$ was selected over $j$ in a pair $k$.

Table 3.1 suggests that the JOD scale better explains the validation data. The bottom two rows indicate that conditions far apart in the VC scale are equally likely to come from both hypotheses.

**Scatter plot comparison** Figure 3.4 shows the relationship between vote counts and psychometric scaling. The plot in the left part of Figure 3.4 shows the relationship after including comparisons from the first four experiments (without the validation experiment). It can be seen that there are some cases, for example, content 5 and 8, which are consistently ranked better on the JOD quality scale than the rest, and others, such as contents 4, 10, and 12, which are consistently ranked worse. There are several reasons for this effect. First of all, only a small number of cross-content experiments were performed, and the selection of compared conditions might not be sufficient to accurately capture all variations in the quality. Secondly, annoyance caused by different distortions is conditional on the content to which they are applied, for example, they are more noticeable on human faces [95]. The relationship between VC and the final psychometric scaling, which includes the data from all experiments, is shown in the right part of Figure 3.4. Here the contents are more mixed in the scale. Both scales have a substantial positive correlation for conditions within the same content. I hypothesize that this is because the original TID dataset contains many more comparisons than the ones I collected, thus significantly impacting the psychometric scaling.

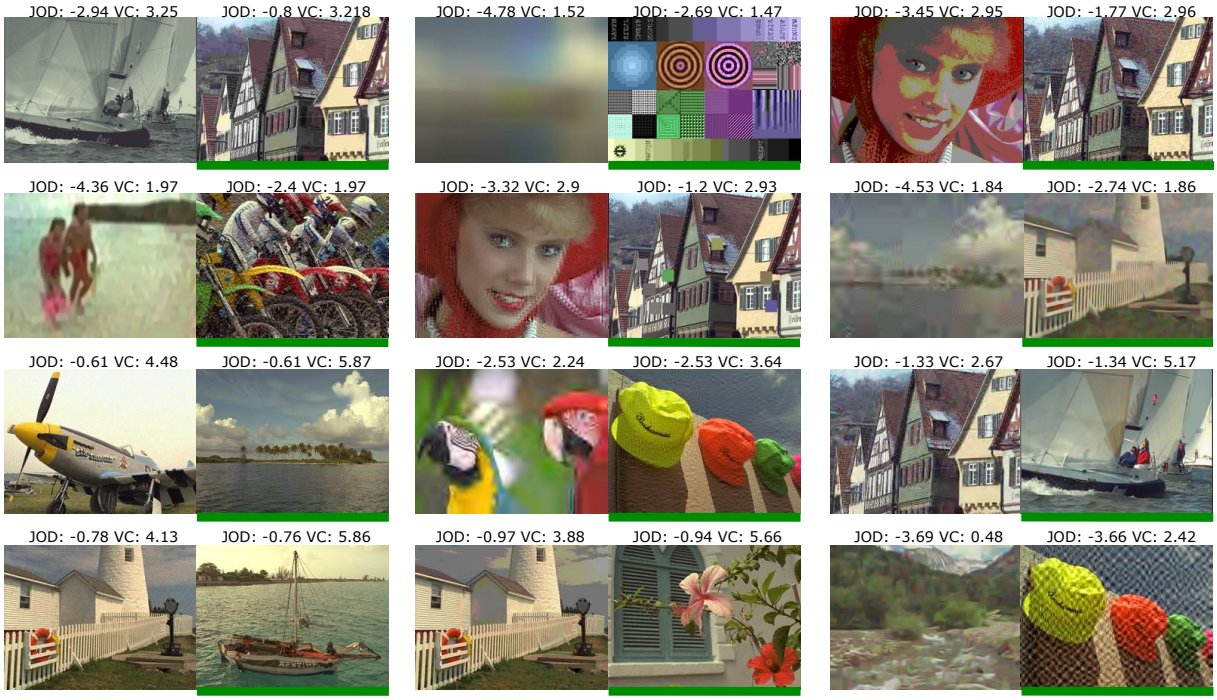



Figure 3.7: Representation of comparisons where VC and JOD scales differ the most, and empirical probabilities are unanimous (all observers agreed). In each pair, the left image was not chosen by any of the observers, and the right was chosen by all observers (highlighted with a green line). First six pairs represent cases where VC failed to correctly rank conditions (but JOD succeeded) and last six comparisons depict failure cases in JOD (where VC ranking succeeds).

**Inconsistencies** A selection of pairs of images used for the validation experiment with the largest inconsistencies in both scales is plotted in Figure 3.7. Interestingly, most of the cases are cross-content and cross-distortion – the first two rows of pairs in Figure 3.7 show obvious failures of the VC scale, which are resolved by the JOD scale. The last two rows show failures in the JOD scale, which are, however, less obvious.

### 3.4.3 Limitations

Even though the new psychometric scale is more accurate and thus is an appealing alternative to vote counts, the dataset can still be improved. Unanimous answers (in which all observers agree) represent 56% of comparisons in the TID2013 dataset. These may introduce a bias in the scaling, as no upper bound is imposed on the distance between compared conditions [100]. The majority of these answers are due to some conditions being compared only once by one observer because of the use of the Swiss system.

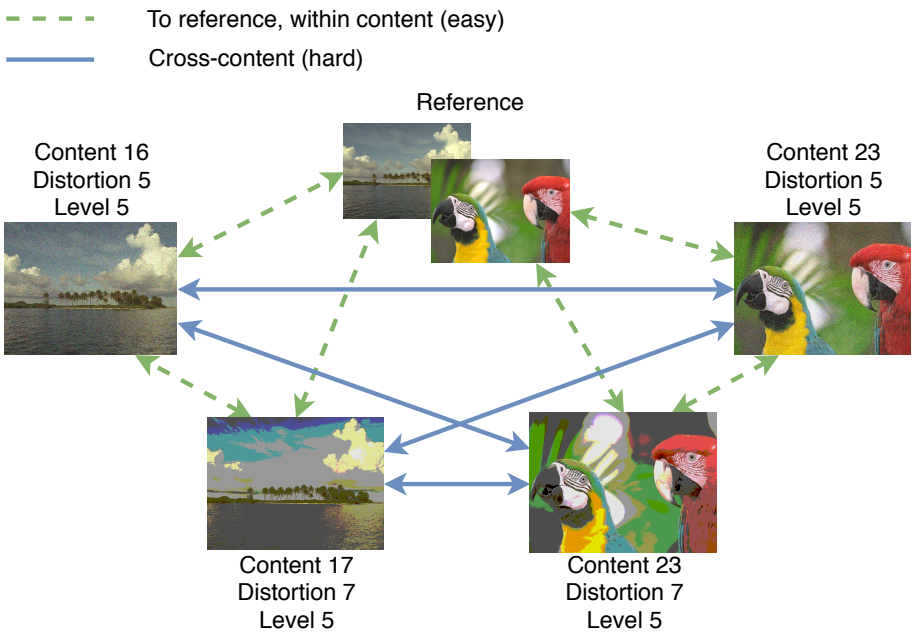


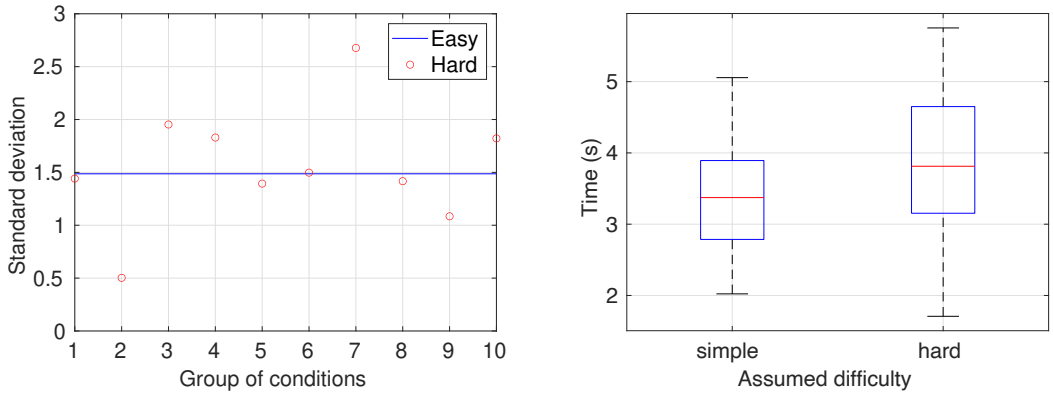

Figure 3.8: An example of different comparisons types for images selected from the TID2013.

## 3.5 Validation of Thurstone Case III vs. V

In Section 2.2.2, I stated that the most commonly used assumption for the observer model, Thurstone Case V, stipulates that the standard deviation for each pair of measured conditions is the same. This would imply that the difficulty of assessing each pair of conditions and the level of confusion is the same across all pairs. However, cross-content comparisons are clearly more difficult for observers to perform than within-content comparisons [149]. Thus, it is reasonable to expect that more difficult types of comparisons will have higher variability in human judgments and the Case V model assumption will no longer be valid.

To determine whether Thurstone Case V assumption is valid for cross-content and within content comparisons, I ran an additional experiment. I selected ten groups of six conditions. Each group included two contents from the TID2013 dataset, and consisted of all possible comparisons: with-reference, within-content, cross-content, within-distortions and cross-distortions. The types distortions and distortion levels were the same across two contents. I then asked ten participants to perform ten comparisons: six within-content and four cross-content, on every group of six conditions, as illustrated in Figure 3.8.

To validate whether the type of comparisons has an effect on the level of confusion ($\beta_{ij}$ in Equation 2.3), I performed a MLE-based scaling in which $\beta_{hard}$ for all "hard" comparisons (shown as solid lines in Figure 3.8) was a free parameter. The standard deviation for all "easy" comparisons was fixed to the usual value of $\beta_{easy} = 1.4826$. The estimated value of $\beta_{hard}$ for all ten groups is shown in Figure 3.9a. The result of a one-sample t-test, with a null hypothesis of the $\beta_{hard} = \beta_{easy}$, ($p_{0.05} = 0.72$) indicates that I do not have evidence to suggest that the comparisons of different difficulty result in a different standard deviation $\beta$. Therefore, contrary




(a) (b)

Figure 3.9: (a) The estimated standard deviation of the "hard" comparisons ($\beta_{hard}$) for ten groups of conditions. The blue line represents the fixed standard deviation of the "easy" comparisons. (b) Time to complete each comparison for both difficulty levels. The box-plot shows 95% confidence intervals with median values marked in red and whiskers showing the outliers.

to the expectations, I cannot reject the assumptions of the Thurstone Case V model.

Figure 3.9b shows the average time spent on easy and hard comparisons. Although the result for 10 groups of a two-sample Kolmogorov-Smirnov test [119] (null hypothesis: data in vectors $x_1$ and $x_2$ are from the same continuous distribution) does not indicate a statistically significant difference ($p_{0.05} = 0.36$), the observers spent, on average, 3.9s on hard and 3.3s on easy comparisons.

I do not have sufficient evidence that the harder difficulty of comparisons results in a higher level of confusion. Therefore, even though the standard deviations could potentially be different, Case V is a good simplifying assumption and a pragmatic choice.

## 3.6 Multidimensional scaling

In this section, I explore multidimensional scaling applied to the TID2013 dataset. For ease of presentation, I use only two dimensions. Here I use the metric based multidimensional scaling, where the metric is the Euclidean distance. I treat the empirical probability $p_{ij} = P(o_i \succ o_j) = \frac{c_{ij}}{c_{ij}+c_{ji}}$, for one condition being better than another converted to JODs $d_{ij} = \Phi^{-1}(p_{ij}, \beta^2)$ as the distance. Thus, the goal is to find the position of each condition $[o_1 \dots o_n]$ in the two-dimensional scale $[\mathbf{x}_1 \dots \mathbf{x}_n]$, where $\mathbf{x}_i = [x_{i1}, x_{i2}]$, minimizing the loss function (also called Stress):

$$Stress = \Big( \sum_{ij=1, i \neq j}^{n} (d_{ij} - ||x_i - x_j||)^2 \Big)^{1/2} \tag{3.2}$$

Figure 3.10a shows the scatter plot of all conditions in the TID2013 dataset with contents marked with different colors. The conditions are uniformly spread around the reference. Figure

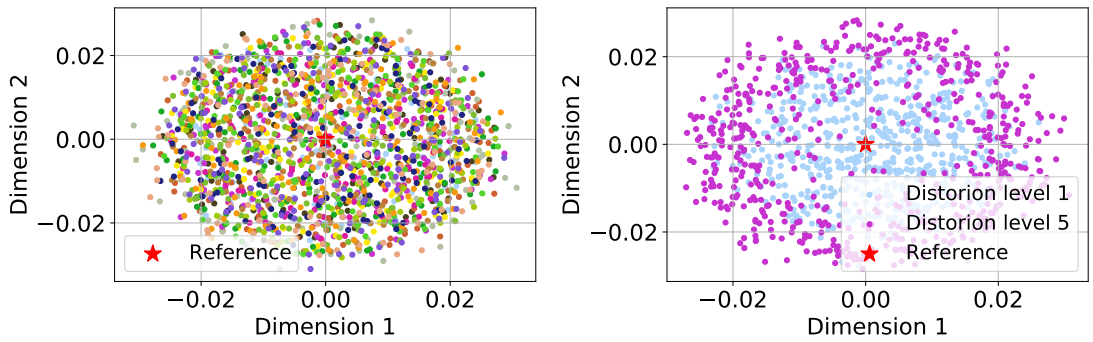



Figure 3.10: Two-dimensional scaling of the TID2013 dataset, the data were re-scaled so that the reference condition has a score of zero in both axes. (a) Scatter plot of the whole dataset with contents displayed in different colors; (b) scatter plot of the images with the lowest and highest distortion levels given in different colors.

3.10b shows conditions with the lowest (level 1) and the highest distortion (level 5) levels applied to them in different colors. The data was re-scaled so that the reference has a score zero in both dimensions. Several trivial observations can be drawn from the plots. First, all contents are similarly distributed in both dimensions. Second, less distorted conditions are placed close to the reference condition, which is at the center of both scales, whereas distorted ones are placed further away.

Since the scatter plot of all contents, distortion types, and distortions levels can be hard to analyze and interpret, I plot the scores of the conditions with the first three distortion types at two impairment levels from the first two contents in Figure 3.11. Consider, for example, distortions one and two with the small impairment levels for both contents. Visual impairment level for both conditions is similar and is almost indistinguishable from the reference. This is reflected in the one-dimensional JOD scale, however, not so in the two-dimensional case. For content 1, distortion one is far away from the reference, whereas distortion two is close. For content 2 the reverse holds. Similar pattern can be observed across all conditions.

Although multidimensional scaling is a useful tool for analyzing distance relationships in the dataset, it is hard to interpret. Furthermore, with a growing number of dimensions, grows the degree of freedom – the data can be represented in multiple ways without violating the constraints. This property might be desirable in many problems, however, it is not suited for image quality assessment, where the goal is to identify a single easy to interpret scale. The growing degree of freedom presents problems for sparse comparison graphs, where conditions are connected to a few or a single other conditions. One advantage of one-dimensional scale is that it provides additional regularization - the space of solutions is much smaller than in the case of multidimensional scaling. Ultimately the scale depends on the task in the subjective experiment, and if the task is: “Assign a single quality value to an image.”, a one-dimensional scale must suffice.

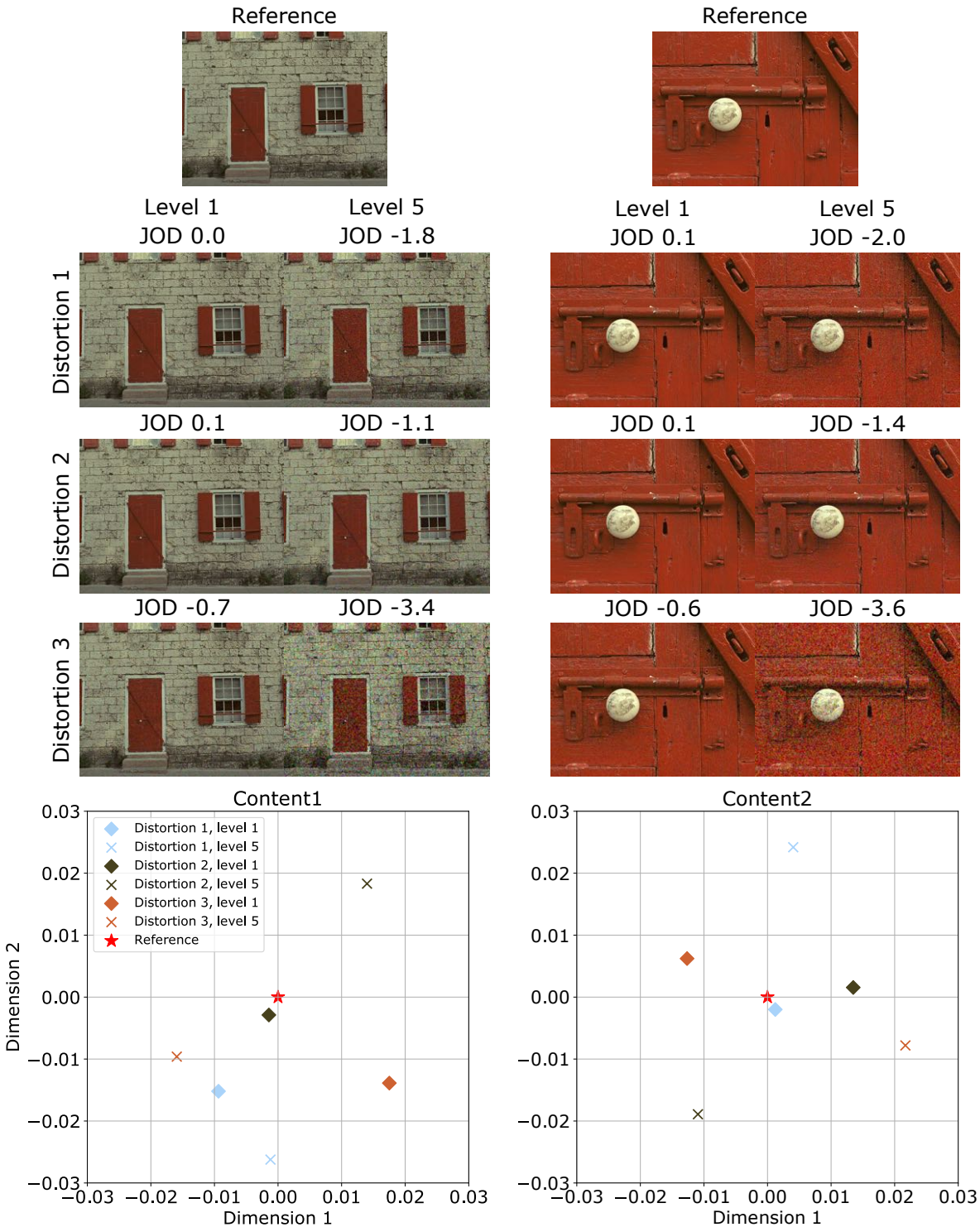



Figure 3.11: Two-dimensional scaling of the TID2013 dataset. Same distortion types are given in the same color. Same distortion level across all distortions is given the same marker. Star shows the reference.

## 3.7 Summary

In this chapter, I investigated how quality scores change when computed using psychometric scaling instead of vote counts. I showed that psychometric scaling produces more accurate results than vote counts in a simulated experiment, especially as the number of conditions in the experiment increases. I also demonstrated that the additional set of comparisons and psychometric scaling improve the consistency of quality scores of the TID2013. I validated that the assumptions of Thurstone Case V are sufficient for modeling image quality in this specific case and did not find sufficient evidence for using Thurstone case III. Finally, I investigated the results of applying multidimensional scaling to modeling image quality. Findings suggest that although some of the data patterns are reflected in the scale produced by multidimensional scaling, difficult to interpret distances is a substantial limitation of the scale with more than one dimension.

The procedure used to select pairs of conditions for the experiments described in this chapter has an important limitation. The pairs were chosen manually, based on the heuristic that similar in quality images form informative pairings. However, this approach is far from ideal, and a more effective and elegant way of selecting pairs can be used. I discuss a new method for active sampling of pairwise comparisons in the next chapter.

# Chapter 4

# Active sampling for pairwise comparisons

## 4.1 Introduction

The main limitation of pairwise comparison experiments is that for $n$ conditions, there are $\binom{n}{2} = n(n-1)/2$ possible pairs to compare, which makes collecting all comparisons too costly for large $n$. However, *active sampling* can be used to select the most informative comparisons, minimizing experimental effort while maintaining accurate results. In the previous chapter, I used a simple heuristic to manually select pairs to compare, to improve the TID2013 dataset. In this chapter, I talk about active-sampling of pairwise comparisons, where pairs for comparisons are selected automatically.

State-of-the-art active sampling methods are typically based on information gain maximization [101, 34, 15, 146, 70, 144], where pairs in each trial are selected to maximize the weighted change of the posterior distribution of the scale. However, these are computationally expensive for a large number of conditions, as they require computing the posterior distribution for $n(n-1)/2$ pairs at every iteration of the algorithm. To make active sampling computationally feasible, most existing techniques update the posterior distribution only for the pairs selected for the next comparison. I show that this leads to a sub-optimal choice of pairs and worse accuracy as the number of measurements increases. To address this problem, I substantially reduce the computational cost of active sampling by using approximate message passing for inference, described in Section 2.2.3.3, and by computing the expected information gain only for the subset of the most informative pairs. The reduced computational overhead allows to update the full posterior distribution at every iteration, thus significantly improving the accuracy. To ensure balanced design and allow for a batch sampling mode, I sample the pairs from a minimum spanning tree as in [70]. The proposed technique (ASAP - Active SAmpling for Pairwise comparisons) results in the most accurate psychometric scale, especially for a large number of measurements. Moreover, the algorithm has a structure that is easy to parallelize, allowing for a fast GPU implementation. I show the benefit of using the full posterior update by comparing it to an approximate version of the algorithm (ASAP-approx), which, similar to other

methods, relies on the online posterior update.

The work in this chapter is based on my publication at IEEE International Conference on Patter Recognition [1]. The proposed algorithm was also used for data collection in the SIGGRAPH publication [24] I co-authored.

## 4.2 Related work

This section discusses related work, divided into four groups, based on the type of an approach: passive, sorting, information-gain, and matchmaking. The methods tested in the experiments are highlighted in boldface. I also distinguish between sequential methods —where the next pair is generated only upon receiving the outcome for the preceding pair —and batch, or parallel methods —where a batch of comparison pairs is generated and outcomes can be obtained in parallel. Batch methods are preferred in crowd-sourcing, where multiple conditions are distributed to participants in parallel. Although other works exist, e.g., estimating total or partial order rankings [39, 40, 51, 147, 121], my dissertation is focused on active sampling for psychometric scale construction, which uses pairwise comparisons to estimate quality scores $\mathbf{q}$.

**Passive approaches** When every condition is compared to every other condition the same number of times, the experimental design is referred to as full pairwise comparisons (FPC). Such an approach is practical only for a small number of compared conditions, as it requires $n(n-1)/2$ comparisons per participant. Another approach, nearest conditions (NC), relies on the idea that conditions that are similar in quality are more informative for constructing the quality scale. Thus, if the approximate ranking is known in advance, one can compare only the conditions that are neighbors in the ranking. Such initial ranking, however, may not be available in practice.

**Sorting approaches** Similar to NC, sorting-based methods first sort the conditions, based on some criteria, and then compare conditions that are of similar quality. Authors in [21] proposed an active sampling algorithm using a binary tree. Every new condition descends the tree, branching depending on whether it is better or worse than the current node's condition. Authors in [85] applied **Quicksort** [42] using pairwise comparisons as the comparison operator.

Recently, [106] used the **Swiss system** in chess to rank subjective assessment of visual quality. The Swiss system first chooses random conditions to compare, then sorts the conditions to find similar pairings. This method was used by the authors in [106] to collect the TID2013 dataset, discussed in the previous chapter. A related method is the Adaptive Rectangular Design (ARD) [48], which allows comparison of conditions far apart on the quality scale in the later stages of an experiment. The work of [16] takes a different approach, where active sampling (**AKG**) is based on the Bayesian decision process maximizing Kendall's tau rank correlation

coefficient [58].

Sorting approaches are praised for their simplicity and low computational complexity and are thus often employed in practice. However, these approaches use heuristics that often result in suboptimal comparison choices, and in general, perform worse than the methods that rely on information gain.

**Information-gain approaches** These methods are based on information maximization. That is, the posterior distribution of quality scores is computed, and the next comparison is selected according to a utility function, e.g., Kullback-Leibler (KL) divergence [67] between the current distribution and the distribution assuming any possible comparison [112]. This group is the most relevant to the new method. Methods listed in this section are sequential unless stated otherwise.

A greedy Bayesian approach, **Crowd-BT**, was proposed in [15]. The entropy for every pair of conditions is computed using the posterior distribution of each pair individually rather than jointly. The method also explicitly accounts for the reliability of each annotator: scores and annotator quality are updated using an alternating optimization strategy.

Authors in [101] derive the score distribution from the maximum likelihood estimation and the negative inverse of the Hessian of the log-likelihood function. Since the original implementation was not provided by the authors, and my implementation suffered from numerical instability, I did not include it in the tests.

Authors in [34, 146] develop a fully Bayesian framework to compute the posterior distribution of the quality scores. **Hybrid-MST** [70] extends this idea by selecting batches of comparisons (instead of single pairs) to maximize the information gain in the minimum spanning tree [18] of a comparison graph. The time efficiency of the method over its predecessor is improved by computing the information gain locally —within the compared pair.

A different approach is taken by [144], where authors propose to solve a least-squares problem to elicit a latent global rating of the conditions using the Hodge decomposition of pairwise comparison data. Like other methods, the information gain is computed using the posterior of only the pair of compared conditions. I refer to this approach as **HR-active**.

**Matchmaking** A matchmaking system, which I refer to as **TS-sampling**, was proposed for gaming, together with the TrueSkill algorithm [41]. The aim is to find the pairs of players with similar skills. The skill distribution of two players is used to predict the match outcome.

**My work** In contrast to the previous work, my method (i) allows for batch and sequential modes, which only one other method allows for [70]; (ii) estimates the posterior using the entire set of comparison outcomes that has been collected so far; and (iii) computes the utility function for a subset of pairs, saving computations without compromising on performance.

## 4.3 Sampling algorithm: ASAP

The algorithm consists of two main steps: (i) computing the posterior distribution of score variables $\mathbf{r}$ using the pairwise comparisons collected following TrueSkill based Bayesian score estimation in Section 2.2.3.3; (ii) using the posterior of $\mathbf{r}$ to estimate the next comparison to be performed based on a criterion of maximum information gain.

### 4.3.1 Pair selection

Given the scores $\mathbf{r}$, where each score $r_i$ is a random variable $r_i \sim \mathcal{N}(\mu_i, \sigma_i)$ estimated using the procedure from Section 2.2.3.3, we can find the next pair of conditions to compare. For that we would look for the pair of conditions for which the result of a trial improves the scores $\mathbf{r}$ the most.

Several utility functions can be used to compute the expected information gain (EIG). A commonly used choice is Kullback-Leibler (KL) divergence [67] between the distribution at a time step $t$ and the one at $t+1$, assuming all possible comparison outcomes. For two continuous distributions $G$ and $J$ the KL divergence is given by:

$$D_{KL}(G||J) = \int G(x) log\left(\frac{G(x)}{J(x)}\right) dx. \tag{4.1}$$

For discrete distribution the integration is replaced with a summation.

More specifically, my active sampling strategy picks conditions $(o_i, o_j) = A_t$ to compare in measurement $t$, such that they maximize a measure of information gain $I^{ij}_{t-1}$:

$$A_t = \underset{(o_i,o_j)\in S^2, i\neq j}{\operatorname{argmax}} I^{ij}_{t-1}, \tag{4.2}$$

where $S$ is the set of all conditions and subindex $t-1$ indicates that we use all measurements collected up to the point in time $t$. For simplicity, I define $\hat{\mathbf{r}}_{t-1}$ as the estimated posterior after measurement $t-1$.

For each possible pair $A_t$, let $P(\hat{\mathbf{r}}_t|y_t = +1, A_t)$ and $P(\hat{\mathbf{r}}_t|y_t = -1, A_t)$ denote the updated posterior distributions (i.e., including comparison $A_t$) if $o_i$ is selected over $o_j$ ($y_t = +1$ for $A_t = (o_i, o_j)$) and vice versa. Since we cannot guarantee the outcome of the next pairwise comparison, i.e., which condition will be selected, similarly to other active sampling methods [70, 144, 15, 101, 34], for the expected information gain computation I weight the information gain from each outcome by the probability of each outcome. I compute this probability using Equation 2.11, $P(o_i \succ o_j|\hat{\mathbf{r}}_{t-1}) \triangleq \Phi\left(\frac{\hat{\mu}_i - \hat{\mu}_j}{\sqrt{2}\hat{\sigma}_{ij}}\right)$ where $\hat{\sigma}^2_{ij} = \hat{\sigma}^2_i + \hat{\sigma}^2_j + \beta^2$; for condition $o_i$ selected over $o_j$ and vice versa, EIG is then defined as:

$$\begin{aligned} I^{ij}_{t-1} =& P(o_i \succ o_j|\hat{\mathbf{r}}_{t-1}) \cdot \mathrm{D}_{\mathrm{KL}}\left(P(\hat{\mathbf{r}}_t|y_t = +1, A_t) \parallel p(\hat{\mathbf{r}}_{t-1})\right) \\ &+ P(o_i \prec o_j|\hat{\mathbf{r}}_{t-1}) \cdot \mathrm{D}_{\mathrm{KL}}\left(P(\hat{\mathbf{r}}_t|y_t = -1, A_t) \parallel p(\hat{\mathbf{r}}_{t-1})\right). \end{aligned} \tag{4.3}$$

### 4.3.2 Efficiency considerations

At every iteration $t$, there are $n(n-1)/2$ comparisons to consider, where $n$ is the total number of compared conditions. The complexity of the posterior evaluation is $O(n+t)$, thus the complexity of selecting the next comparison is $O(n^2(n+t))$. This may be very costly when the number of conditions is large. Here, I discuss two modifications that reduce the computational cost, and a batch mode, which also improves the accuracy.

**Approximate (online) posterior estimation (ASAP-approx)** To quantify the improvement in accuracy brought by the full posterior update, I follow the common approach and use an online posterior update with assumed density filtering (ADF) [91]. Here, unlike the full posterior update described in Section 2.2.3.3, where the posterior is computed based on all observed outcomes $\mathbf{t}$, we update the posterior only with the most recent outcome, using the posterior from the previous step as the prior. That is, the posterior $\hat{\mathbf{r}}_{t-1}$ is used as the prior when computing the posterior for the $t^{\text{th}}$ comparison, allowing the algorithm to run in an online manner [92]. The update for two conditions (winning $o_W$ and loosing $o_L$) is formulated as follows:

$$\mu_W = \mu_W + \frac{\sigma_W}{\zeta}\upsilon(\frac{\tau}{\zeta}),$$

$$\mu_L = \mu_L - \frac{\sigma_L}{\zeta}\upsilon(\frac{\tau}{\zeta}),$$

$$\sigma_{W/L} = \sigma_{W/L}(1 - \frac{\sigma^2_{W/L}}{\zeta^2}\chi(\frac{\tau}{\zeta})), \tag{4.4}$$

where $\zeta^2 = 2\beta^2 + \sigma_W^2 + \sigma_L^2$, $\tau = \mu_W - \mu_L$, $\upsilon = \frac{\mathcal{N}(t)}{\Phi(t)}$ and $\chi = \upsilon(t)(\upsilon(t)+t)$ and $\mathcal{N}$ and $\Phi(t)$ are probability density function and cumulative density function of a standard normal distribution respectively.

Thus, for every $o_i$ and $o_j$ pair, I update only the scores $r_i$ and $r_j$, resulting in $O(1)$ complexity per pair. No additional ADF-projection step is required since expectation propagation has already yielded a Gaussian approximation to the posterior. The time complexity of selecting the next comparison is thus decreased to $O(n^2)$. However, computational efficiency comes at the cost of accuracy in posterior estimation [92]. I refer to the algorithm using the approximate posterior update as *ASAP-approx*.

**Selective EIG evaluations** EIG evaluations are computationally expensive. It is thus desirable to compute EIG only for a subset of the pairs. To select the pairs for EIG evaluation I rely on the fact that some comparisons are less informative than others [112], such as between conditions far apart on a scale where the outcome $y_t$ is obvious [101, 34]. Thus, I want to prioritize EIG computation for the pairs of conditions which are closer in the scale and which

are thus more informative. For that I use a simple criterion from Equation 2.11 and compute the probability $Q_{ij}$ that conditions $o_i$ and $o_j$ are selected for EIG evaluation. Since $p_{ij}$ in the Equation 2.11 is the probability of condition $o_i$ being better than $o_j$, to identify obvious outcomes I set $Q_{ij} = min(p_{ij}, p_{ji})$, where $p_{ji} = 1 - p_{ij}$. Thus, the probability is large when the difference between the scores and their standard errors are small. To ensure that at least one pair including $o_i$ is selected, I scale $Q_{ij}$ per condition, i.e., $Q^*_{ij} = \frac{Q_{ij}}{\max_{\forall j}(Q_{ij})}$. Since computing the probability $Q^*_{ij}$ is computationally less expensive than EIG, I compute $Q^*_{ij}$ for each pair of conditions.

**Minimum spanning tree for the batch mode** When a sampling algorithm is in the sequential mode, one pair of conditions is scheduled in every iteration of the algorithm. However, selecting a batch of comparisons in a single iteration of the algorithm is computationally more efficient and can yield superior accuracy [70]. To extend the algorithm to the batch mode, I treat pairwise comparisons as an undirected graph. Vertices are conditions, and edges are pairwise comparisons. I follow the approach from [70], where the minimum spanning tree (MST) is constructed from the graph of comparisons. The MST is a subset of the edges connecting all the vertices, such that the total edge weight is minimal. The edges of the graph are weighted by the inverse of the EIG, i.e., for an edge $E_{ij}$ connecting conditions $A_i$ and $A_j$ the weight is given by $w(E_{ij}) = \frac{1}{I^{ij}}$. $n - 1$ pairs are selected for the MST, allowing us to compute the EIG every $n - 1$ iterations, greatly improving speed. Since each condition is compared at least once within the batch, detrimental unbalanced designs [113], where a subset of conditions is compared significantly more often than the rest, are eliminated.

## 4.4 Evaluation

To assess different sampling strategies, I run a Monte Carlo simulation on synthetic and real datasets. SROCC and RMSE between the ground truth and estimated scores are used for performance evaluation. I report the results as multiples of standard trials, where **1 standard trial** corresponds to $n(n-1)/2$ measurements (the number of possible pairs for $n$ conditions). For clarity, I present RMSE on a log-scale, and SROCC after a Fisher transformation ($y' = \operatorname{arctanh}(y)$). The same method, based on the MLE-based from Section 2.2.3.1, was used to produce the scale from pairwise comparisons for each method.

### 4.4.1 Simulated data

To generate synthetic data, I run a Monte Carlo simulation. I note that the strongest influence on the results is the proximity of compared conditions in the target scale. When conditions have comparable scores, they are confused more often in comparisons. In contrast, when conditions are far apart in the scale, they are easily distinguished, resulting in different performances for

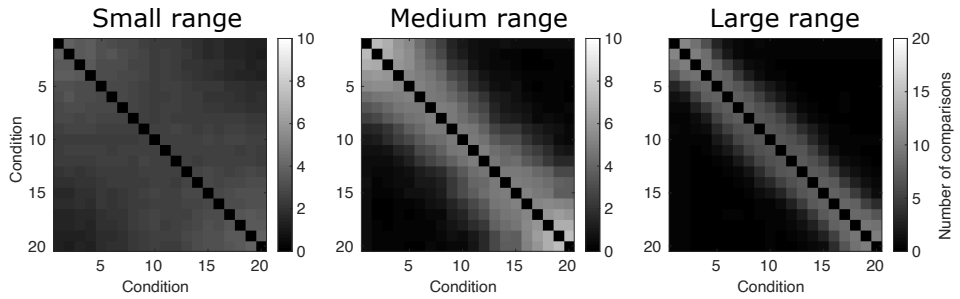


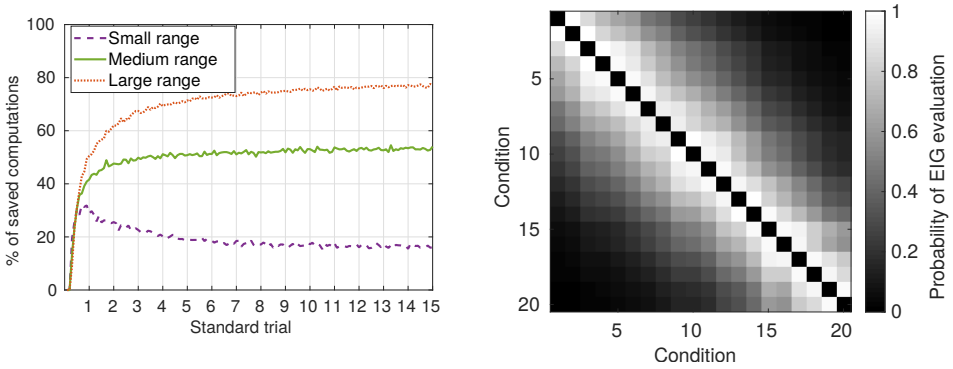

Figure 4.1: Heatmaps for three standard trials of comparisons of 20 ordered conditions from small, medium and large ranges

sampling methods. Hence, I consider three scenarios for 20 conditions with scores $\mathbf{s}$ sampled uniformly from: (i) large range $[0, 20]$ (scores far apart); (ii) medium range $[0, 5]$; (iii) small range $[0, 1]$ (scores close together). Results for larger numbers of conditions are given in Section 4.4.3. I run the simulation 100 times for comparisons ranging from 1 to 15 standard trials. In the simulation, I use $P(o_i \succ o_j) \sim \Phi(\frac{q_i - q_j}{\beta})$ from Equation 2.6 and $\beta = 1.4826$ to draw the outcome of the comparison between conditions $o_i$ and $o_j$, which are determined by each algorithm. Accuracy of scales obtained from comparisons chosen by active-sampling strategies were computed with respect to the ground-truth quality scores $\mathbf{q}$ drawn from a uniform distribution and the given range.

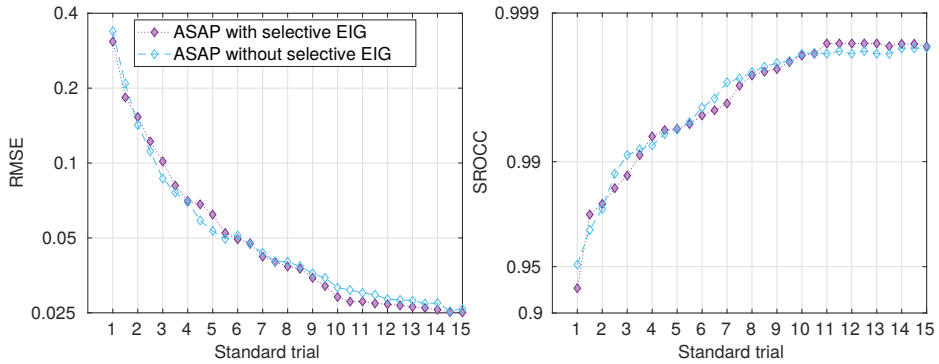
#### 4.4.1.1 Ablation study

**Sampling patterns** To better understand which conditions are favored by ASAP, I produce heatmaps of pairings for conditions sampled from the small, medium, and large ranges. I use three standard trials. The heatmaps are given in Figure 4.1. For better visualization, conditions are ordered ascending in their ground truth scores in the consecutive rows and columns. For conditions sampled from the small range, all pairs of conditions are compared approximately the same number of times. However, the number of comparisons gradually decreases for conditions further away in the scale, i.e., further away from the diagonal on the heatmap. For conditions sampled from the medium and large ranges, most comparisons are selected for conditions close in the quality scale, i.e., along the diagonal on the heatmap. This is expected, as pairs of conditions that are far away in the scale are less likely to be confused by observers and are therefore less informative.

**Selective EIG evaluations** Figure 4.2a shows the proportion of saved evaluations with selective EIG computations, where EIG is updated only for the most informative comparisons. Since I initialize the algorithm with all scores set to 0, all possible pairs have their EIG computed at first (0 standard trials in the plot), as all conditions are close to each other. As more data are collected, conditions move away from each other on the scale, and the EIG is computed for a subset of pairs only. Computational saving is greater for large-range simulations than for




(a) Saved evaluations (b) Probability of EIG evaluation




(c) Performance with and without selective EIG

Figure 4.2: (a) Percentage of saved evaluations with selective EIG evaluations; (b) probability of EIG evaluation after ten standard trials for medium range; and (c) RMSE and SROCC with and without selective EIG;

small-range simulations. In small-range simulations, conditions first move away from each other, as, in the first few iterations, their relative distances are likely to be overestimated, decreasing the overall number of computations; however, with more measurements, the conditions move closer, and the proportion of saved evaluations decreases. Figure 4.2b shows the probability of the EIG being evaluated after 10 standard trials for 20 conditions sampled from the medium range. For visualization purposes, conditions were ordered ascending in the quality scale. Pairs of conditions along the diagonal, i.e. close in the scale, have a higher chance of their EIG being computed. Figure 4.2c shows performance of ASAP with and without selective EIG evaluations. Since pairs chosen by selective EIG evaluation are likely to be the most informative, the number of computations is greatly reduced, while maintaining the accuracy. In the following sections, I only present the results with selective EIG computations.

**Minimum spanning tree for the batch mode** Figure 4.3 shows the results of ASAP with and without batch mode for medium-range simulations. Without MST batch mode, the method is likely to result in an unbalanced sampling pattern, where certain conditions are compared

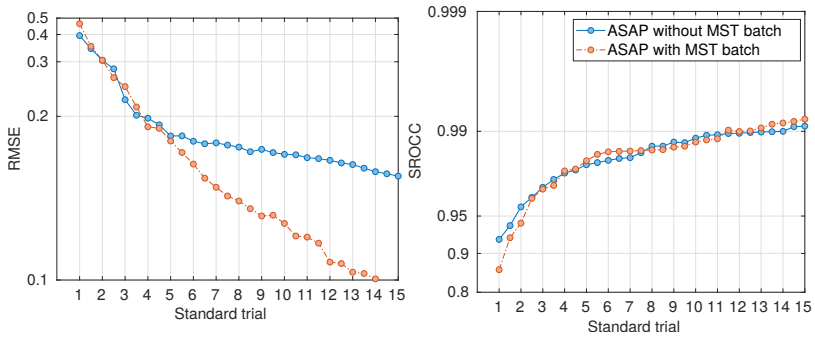



Figure 4.3: Simulation with 20 conditions sampled from the medium range with and without MST. I observe similar pattern for conditions sampled from small and large ranges.

significantly more often than others. This has a detrimental effect on performance, deteriorating the results with a growing number of comparisons [113]. Below, I only present results with MST batch mode.

#### 4.4.1.2 Simulation results

**Algorithms compared** I implement and compare different active sampling strategies using original authors' codes where possible: AKG [16], Crowd-BT [15], HR-active [144] and Hybrid-MST [70]. My own implementation was used for Quicksort [85], Swiss System [106], and TS-sampling [41].

Figure 4.4a-c shows the results of the simulation for the implemented strategies. At all tested ranges, EIG-based methods have lower RMSE, and therefore higher accuracy, than the sorting methods (Quicksort and the Swiss System). While TS-sampling and Crowd-BT have good accuracy for the large range, these are among the worst methods for the small range. ASAP-approx exerts performance similar to the methods with the online posterior update. However, it offers a modest but consistent improvement in accuracy over Hybrid-MST and HR-active. Of all tested methods, ASAP, employing the full posterior update, is the most accurate by a substantial margin and across all ranges.

For SROCC, EIG-based methods do not show a clear advantage over sorting methods; however, it should be noted that EIG-based methods are designed to optimize for RMSE rather than ranking. Even so, ASAP still performs the best for small and medium range simulations, and one of the best for large range, reaching SROCC of 0.99 within five standard trials. However, it should be noted, that the problem of ordering conditions from the large range is trivial, and the best methods compete at 0.99+ SROCC levels (almost perfect ordering). Figure 4.4d presents 75% confidence interval of the RMSE and SROCC distributions for the top five methods in SROCC for conditions sampled from the large range. Results for SROCC are noisier than for RMSE, making it hard to identify the best performing method. In terms of RMSE for a number of standard trials less than two, ASAP shows similar to others' performance, however with the

number of standard trials growing, it significantly outperforms the compared methods. Because of the poor performance of the sorting-based methods, I do not consider them in the following experiments.

### 4.4.2 Real data

I validate the performance of sampling strategies on two real-world datasets: i) Image Quality Assessment (IQA) LIVE dataset [115], with pairwise comparisons collected by Ye and Doermann [146]; and ii) Video Quality Assessment (VQA) dataset [143]. Each dataset contains complete and balanced matrices of pairwise comparisons, with each condition compared to every other condition the same number of times. The empirical probability of one condition being better than another is obtained from the measured data following Equation 2.4 and used throughout the simulation. I compute RMSE and SROCC between scores produced by each method and scores obtained by scaling the original matrices of all comparisons.

**IQA dataset** To allow multiple Monte Carlo simulation runs, I randomly select 40 conditions from the 100 available. In the original matrix, each condition is compared five times with each other (5 standard trials), yielding 24750 comparisons.

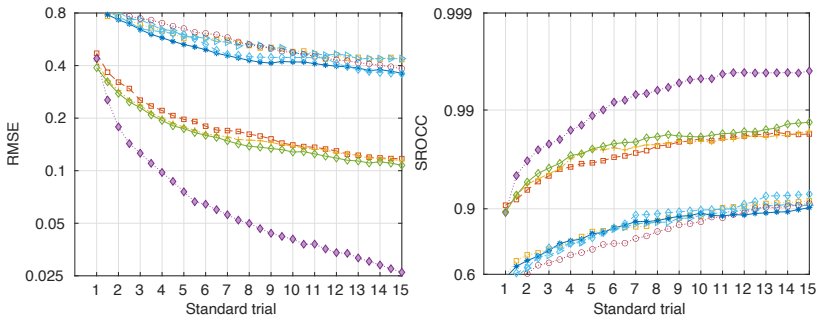
Figure 4.5 shows the results. The performance trends are consistent with the results for the simulated data for the medium range. ASAP has the best performance both in terms of SROCC and RMSE. ASAP-approx, Hybrid-MST, and TS-sampling follow it, each having roughly the same performance in terms of both RMSE and SROCC, with ASAP-approx performing slightly better in ranking. Crowd-BT and HR-active have the worst performance in terms of both RMSE and SROCC.

**VQA dataset** The dataset contains ten reference videos with 16 distortions. Each $16 \times 16$ matrix contains 3840 pairwise comparisons, i.e., each pair was compared 32 times.

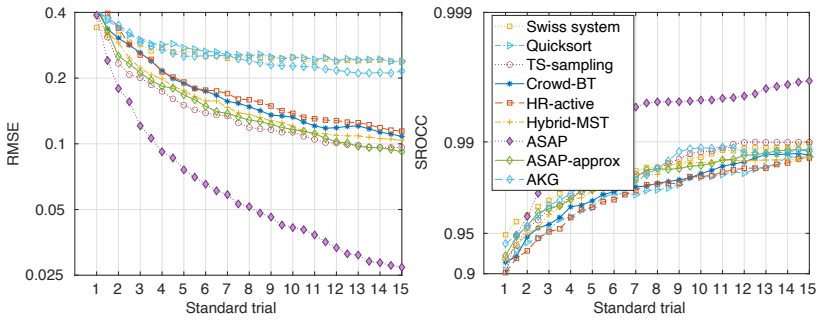
Figure 4.6 shows the results of running simulations. The performance trends are again, in general, consistent with the results for the simulated data sampled from the medium range, except that TS-sampling performs substantially worse, and Hybrid-MST outperforms ASAP-approx for small numbers of trials. ASAP consistently outperforms other methods. The results for the remaining eight reference videos follow the same trend and are given in Appendix A.4.

### 4.4.3 Large scale experiments

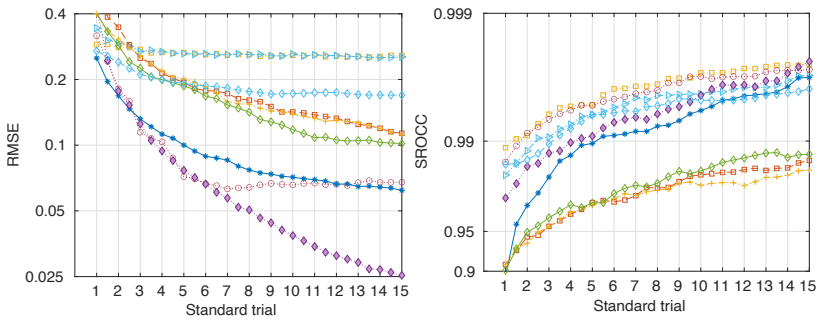
It is often considered that 15 standard trials are the minimum requirement for FPC to generate reliable results [127, 11]. However, this is rarely feasible in practice. Real-world, large-scale datasets barely reach one standard trial. To make experiments with a large number of conditions



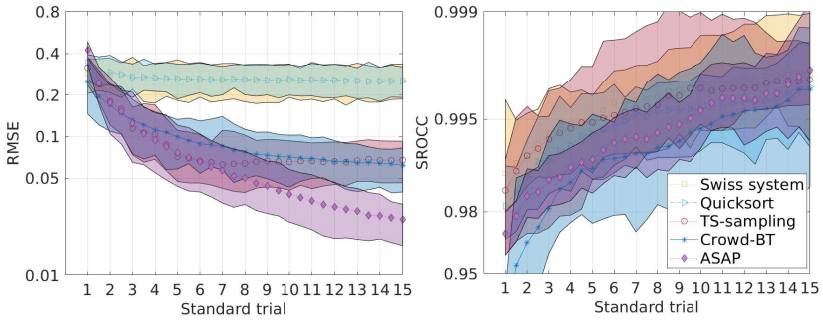

(a) Small range




(b) Medium range




(c) Large range




(d) Large range confidence interval

Figure 4.4: (a-c) Simulation results with 20 conditions for the compared sampling strategies. (d) 75% confidence interval of the RMSE and SROCC distributions for 20 conditions sampled from the large range and five best performing methods in terms of SROCC.

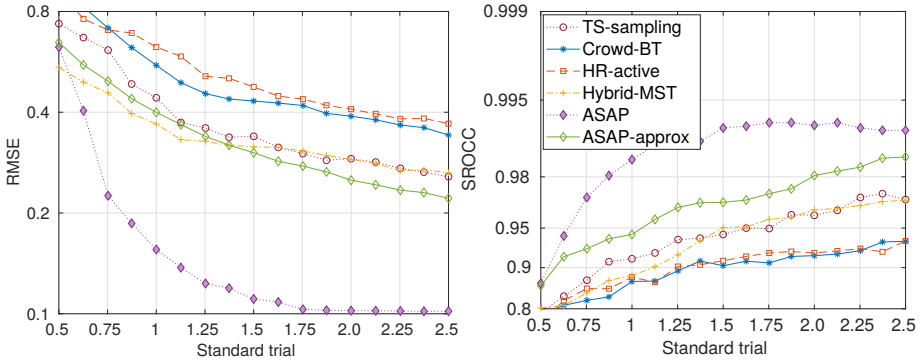
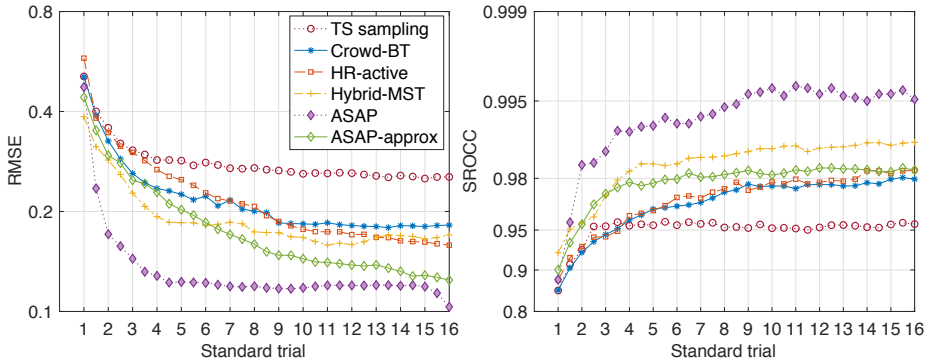
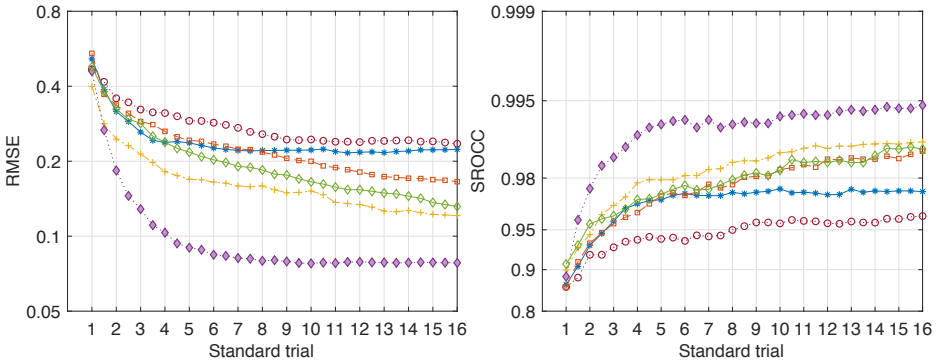



Figure 4.5: Compared sampling strategies on LIVE dataset.




(a) Reference 1




(b) Reference 2

Figure 4.6: Compared sampling strategies on VQA dataset.

feasible, individual reference scenes or videos are often measured and scaled independently, missing important cross-content comparisons. However, the lack of cross-content comparisons yields less accurate scores [149]. I investigated this problem on the example of the scale of TID2013 dataset in the previous chapter. Active sampling techniques, such as ASAP, should accurately measure a large number of conditions while saving a substantial amount of experimental effort. To test such a scenario, I simulate the comparison of 200 conditions with scores distributed in the medium range. The results, shown in Figure 4.7, demonstrate that even with a small number of standard trials, ASAP outperforms existing methods; ASAP-approx and Hybrid-MST follow it.

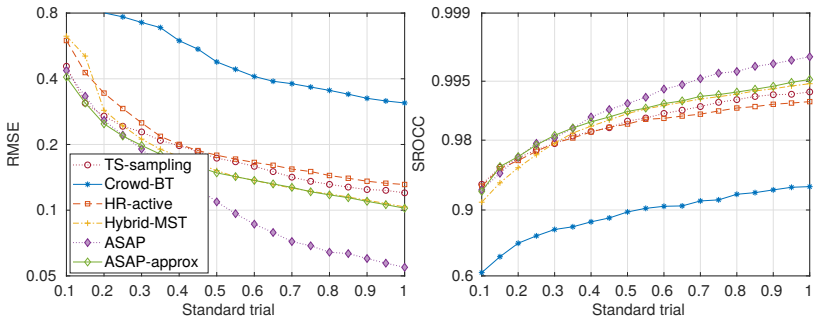



Figure 4.7: Large scale experiment simulation with 200 conditions sampled from medium range.

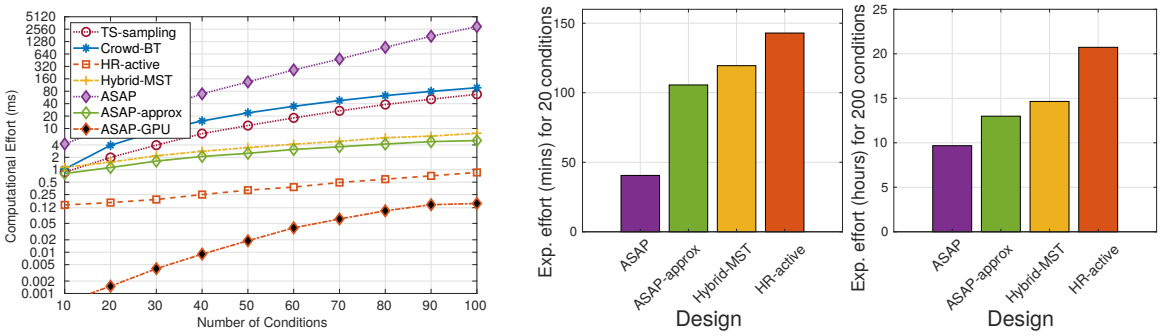
### 4.4.4 Running time and experimental effort

A practical active sampling method must generate new samples in an acceptable amount of time. Hence, in Figure 4.8a, I plot the time taken by each method as the number of conditions grows. The reported times are for generating a single pair of conditions, assuming that five standard trials have been collected. CPU times were measured for MATLAB R2019a code running on a 2.6GHz Intel Core i5 CPU and 8GB 1600MHz DDR3 RAM. GPU time was measured for Pytorch 1.4 with CUDA 9.2, running on GeForce GTX1080. I omit sorting methods as they do not offer sufficient accuracy. Although ASAP is the slowest method when running on a CPU, it can be effectively parallelized on a GPU and deliver the results in a shorter time than other methods running on a CPU.

In Figure 4.8b I show the experimental effort required to reach an acceptable level of accuracy for 20 and 200 conditions, where I define experimental effort as the time required to reach an RMSE of 0.15. I assume that each comparison takes 5 seconds, which is typical for image quality assessment experiments [108, 106]. ASAP offers the biggest saving in the experimental effort for both small and large scale experiments. In an experiment with 200 conditions, ASAP achieves an accuracy of 0.15 RMSE in 0.355 standard trials. Thus, the total experimental time is 9.8h (7065 comparisons), which is significantly better than the 14.6h (10550 comparisons) for Hybrid-MST. Similarly, for 20 conditions, the entire experiment would take 40 min for ASAP and 120 min for Hybrid-MST to reach the same accuracy of score estimates. For experiments with longer comparison times (e.g., video comparison) or high comparison cost (e.g., medical images), ASAP's advantage is even more significant.

## 4.5 Summary

In this chapter, I showed the importance of choosing the right sampling method when collecting pairwise comparison data, and proposed a new active sampling strategy for pairwise comparisons – ASAP. Commonly used sorting methods perform poorly compared to the state-of-the-art methods based on the EIG, and even EIG-based methods are sub-optimal, as they rely on a




(a) Computational time

(b) Experimental effort

Figure 4.8: (a) Average time to select the next comparison for a varied number of conditions and 5 standard trials. (b) Experimental effort (amount of time, assuming 5 second decision time, required to reach 0.15 RMSE) for experiments with 20 and 200 conditions.

partial update of the posterior distribution. ASAP computes the full posterior distribution, which is crucial to achieving accurate EIG estimates, and thus the accuracy of active sampling. Fast computation of the posterior, important for real-time applications, was made possible by using a fast and accurate factor graph approach, which is new to the active sampling community. Besides, ASAP only computes the EIG for the most informative pairs, reducing the computational cost of ASAP by up to 80%, and selects batches using a minimum spanning tree method, allowing to avoid unbalanced designs.

I recommend ASAP, as it offers the highest accuracy of inferred scores compared to existing methods in experiments with real and synthetic data. The computational cost of the technique is higher than for other methods in the CPU implementation. However, it is still in the range that makes the technique practical, with a substantial saving of experimental effort. For large-scale experiments, ASAP-GPU offers both accuracy and speed.

ASAP is a useful tool when collecting large-scale pairwise comparison datasets from scratch, however, many real-world datasets exist, that contain both rating and ranking experiments. For example one of the largest image quality datasets, TID2013 [106] (3000 images), contains pairwise comparisons, whereas LIVE image quality dataset [114] contains both ratings and pairwise comparisons (780 images). This data can be efficiently merged together to a dataset of 3780 images, without the need for exhaustive large-scale experiments. In the next section I propose a method, which, with a relatively low experimental effort can combine datasets with different experimental procedures.

The strength of ASAP is, however, limited to problems where the outcomes of pairwise comparisons are assumed to be free of errors. Methods with error tolerance or those accounting for observer accuracy/decency [15, 71] would be preferred for applications where this condition cannot be ensured or validated with a benchmark test.

# Chapter 5

# Unified subjective quality scale

## 5.1 Introduction

Multiple protocols for data collection, such as rating and ranking, impede using homogeneous data together. As such, an image rated four on a five-point scale in one experiment could be rated as two in another experiment because of the participant training procedure. In this chapter, I propose a probabilistic model and a scaling procedure that can bring qualities from different subjective experiments to a unified scale. The scores produced by the scaling model are interpretable and given in the JOD units, with the unit distance between two conditions corresponding to 75% of observers selecting one condition over another. The proposed method builds on a well-established psychophysics and sensory evaluation field and scales together the two most commonly used protocols: rating and pairwise comparisons.

Such scaling can be used to merge existing datasets of subjective nature and for experimental protocols in which both rating and pairwise comparisons are collected. Existing quality datasets and newly collected data are used to justify the assumptions made in the model, such as the linear relation between rating and scaled pairwise comparison data. The utility of the method is demonstrated by re-scaling three existing datasets: TID2013 IQA dataset [106], LIVE IQA dataset [115] and the HDR video compression dataset [149].

The side-benefit of the joint scaling is that sensitivity and time effort can be compared and analyzed for both experimental protocols. Findings from several analyzed real-world datasets show that the standard deviation of the observer model for rating and pairwise comparisons depends on the task and the dataset. However, generally for image quality assessment tasks, observers confuse measured conditions more often in rating experiments. Finally, I demonstrate in simulations that, given the mean times required to rate and compare image quality and the standard deviations found for the observer model, pairwise comparisons, on average, provide better estimates. I also demonstrate that both protocols can be used together to avoid time-consuming cross-content comparisons, discussed in Chapter 3 and to create larger datasets with relatively low experimental effort.

The work in this chapter is based on my publication at IEEE Transactions on Image Processing [108] and reviewed manuscript at IEEE Transactions on Multimedia [90].

## 5.2 Related work

It is useful in practice to aggregate quality scores obtained from different quality evaluation experiments, for example, to create larger annotated datasets. While this aggregation of subjective quality scores is usually done for rating (i.e., mean opinion scores) [102, 104] or pairwise comparisons [101, 118] individually, little work has been done to study the fusion of scores obtained by these two methodologies.

In this regard, Ye and Doermann [146] proposed a unified probabilistic model, aggregating rating and pairwise comparisons together. The model is extended to an active sampling framework and uses the information gain for choosing either a pairwise comparison or a rating protocol for the next measurement. The model allows an observer to rate conditions on a continuous scale, however, these continuous scores are then converted to categorical values using cutoff values for these categories. This makes the optimization an iterative two-step alternating procedure, where in the first step cutoff values are found, and in the second step these are used in an optimization procedure for finding the score distribution. The method relies on heavy computations and is not feasible for large scale datasets. Moreover, the authors did not consider the relationship between both scales, meaning that the final mixed scale could not be interpreted in terms of probabilities.

Authors in [148] proposed a procedure for aligning subjective scales based on the objective scores. The method takes as input only the resultant scores obtained from scaling pairwise comparisons or MOS, without considering individual measurements or underlining protocols. The method assumes that the quality predictions from multiple objective metrics can be used to transform quality scores from one subjective dataset to another. However, this approach relies on the objective scores, which might not be available in practice. Furthermore, the accuracy of the obtained subjective scale depends on the accuracy of objective metrics, which in their turn are developed based on the subjective scores. I compare my method to the one from [148] in the next chapter.

Some works have studied the relationship between the psychometric scale obtained from pairwise comparisons and MOS. As such, Watson [138] studied the correlation between rating scales and the results of pairwise comparisons in the context of psychometric scaling of pairwise preference probabilities. He found that the degree of agreement between two scales, for video compression, is relatively high. The work reports a quadratic relationship between MOS and scaled pairwise comparisons, with a very small quadratic coefficient ($JND = 1.917 + 0.125 * DMOS + 0.0012 * DMOS^2$ for $DMOS \in [0; 50]$). Similarly, [149] shows a strong linear relationship between MOS and pairwise comparison scaling results.

The model proposed in this section does not rely on objective metrics, it explicitly models

the relationship between rating and ranking scores, based on their respective observer models, and takes into consideration the number of measurements performed for each protocol. The model is computationally lightweight and can be used for large scale experiments. Although the model does not allow for a dynamic choice of the protocol, I believe, however, that the proposed approach can be beneficial when combined with pilot studies. More specifically, since my scaling method explicitly accounts for the standard deviation of each protocol's observer model, after the pilot study the protocol with the smaller standard deviation can be chosen.

## 5.3 From pairwise comparisons and rating to a unified scale

When the results of both ranking and rating experiments are available for the same contents, it may be desirable to use all information when constructing the quality scale. In this section, I propose a simple way of combining both types of measurements.

I assume a linear relationship between a random variable $\omega_i$ representing quality scores obtained from a pairwise comparison experiment (Equation 2.2), and the random variables obtained from a rating experiment $\pi_i$ (Equation 2.1):

$$\omega_i = a \cdot \pi_i + b. \tag{5.1}$$

I could instead assume a more complex relationship between the quality scores, for example, quadratic [138]. However, I found that a linear assumption is sufficient for large-scale quality datasets (more details in Section 5.4). Nevertheless, the model can easily be extended to more complex functional forms, provided that this relation is known. I further assume that the standard deviation of the observer model may differ between both experimental protocols: people can confuse two conditions more often in one protocol than the other. Similar to Section 2.2.2 for every condition $o_i$ I set the standard deviation of the observer model $\beta_i$ constant to $\beta_*$. This is to contrast it with the standard deviation $\beta = \frac{\sqrt{2}\beta_*}{2}$ of the normal distribution obtained from the difference of two normal distribution with standard deviation $\beta_*$. The relationship can then be written as:

$$\mathcal{N}(q_i, \beta_*^2) = a \cdot \mathcal{N}(m_{ik}, \eta^2 \cdot \beta_*^2) + b = \mathcal{N}(a \cdot m_{ik} + b, a^2 \cdot \eta^2 \cdot \beta_*^2), \tag{5.2}$$

where $m_{ik}$ is the collected opinion score for the condition $i$ and observer $k$. $q_i$ is the latent quality score, which I want to recover. $a$, $b$ and $\eta$ (accounting for the difference in the variance of observer models of rating and pairwise comparisons) are the unknown parameters that control the relationship between the rating and pairwise comparison data. The goal is to find the values of the latent variables given the observed opinion scores and pairwise comparisons.

Since opinion scores are generally continuous, I express the probability of observing $m_{ik}$

using the density function of the normal distribution:

$$p(m_{ik}|q_i, \beta_*, a, b, \eta) = \frac{1}{\sqrt{2\pi a^2 \eta^2 \beta_*^2}} e^{-\frac{((a \cdot m_{ik}+b)-q_i)^2}{2a^2\eta^2\beta_*^2}}. \tag{5.3}$$

Assuming independence between observers, the likelihood of observing the whole set of opinion scores $\mathbf{M}$ is:

$$P(\mathbf{M}|\mathbf{q}, \beta_*, a, b, \eta) = \prod_{i=1}^{N}\prod_{k=1}^{J} f(m_{ik}|q_i, \beta_*, a, b, \eta). \tag{5.4}$$

Similarly, the likelihood of observing pairwise comparisons $P(\mathbf{C}|\mathbf{q}, \beta_*)$ is given in Equation 2.8. One advantage of this probabilistic formulation is that missing data, for example when observers rate only a portion of all conditions, can be simply omitted from the above product.

To recover latent quality scores $\mathbf{q}$ from both measurements, I use the MAP estimator with the posterior probability:

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q},a,b,c} P(\mathbf{q}, a, b, \eta|\mathbf{C}, \mathbf{M}, \beta_*), \tag{5.5}$$

where

$$P(\mathbf{q}, a, b, \eta|\mathbf{C}, \mathbf{M}, \beta_*) \propto P(\mathbf{C}|\mathbf{q}, \beta_*) \cdot P(\mathbf{M}|\mathbf{q}, \beta_*, a, b, \eta) \cdot P(\mathbf{q}), \tag{5.6}$$

and $P(\mathbf{q})$ is a Gaussian prior included to enforce convexity:

$$P(\mathbf{q}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\beta_*^2}} e^{-\frac{(\mu_\mathbf{q}-q_i)^2}{2\beta_*^2}}, \tag{5.7}$$

$\mu_\mathbf{q}$ being the mean of quality scores $\mathbf{q}$.

Likelihood functions are scale-invariant, i.e., $P(\mathbf{M}|q, \beta_*) = P(\mathbf{M}|tq, t\beta_*)$ for a constant $t \neq 0$. Thus, without loss of generality, we can fix $\beta_*$ to an arbitrary value. As before, since scales are relative, I set an anchor to $q_1 = 0$. Similar to other chapters, I fix $\beta_* = 1.0484$ ($\beta = 1.4826$), so that a distance of 1 unit between two conditions indicates that 75% of observers can see the difference between two conditions, allowing the interpretation of distances in the scale.

Note, that to mix different datasets, for example, several datasets for which rating measurements have been collected, one would need to collect pairwise comparisons that link the data and run the optimization procedure. In this case, different standard deviation of the observer model and scaling parameters ($a$, $b$, and $\eta$) should be assumed for different datasets. I show how my model can be used to mix different datasets in Chapter 6. I made the code for mixing both types of measurements available online [1].

[1] `https://github.com/gfxdisp/pwcmp_rating_unified`

## 5.4 Experiments: scaling existing datasets

In this section, I compare the experimental effort and validate the model assumptions on two real-world image quality assessment datasets: LIVE [115] and TID2013 [106], and one video compression dataset [149]. I test the linear relationship between subjective quality scores coming from pairwise comparisons and rating and estimate the time effort and the standard deviation of the observer model in both measurements. To scale the pairwise comparisons data, I use psychometric scaling with maximum likelihood estimation using the Thurstone Case V model, described in Section 2.2.3 and Matlab code from [100].

### 5.4.1 Datasets

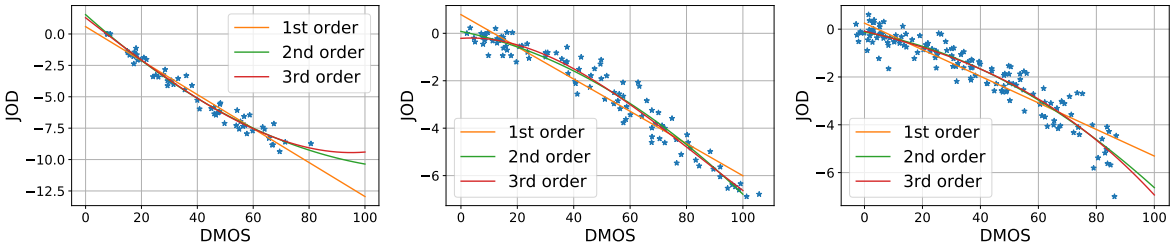
**HDR video compression dataset** As one of the real-world examples, I use a HDR video compression dataset [149]. This dataset contains 60 compressed HDR videos. As it was created to analyze the relationship between rating and pairwise comparison scaling, this dataset includes rating (DSIS) and pairwise comparison experiments with and without cross-content pairs. The authors explored the effect of additional cross-content comparisons in pairwise comparison experiments on the scaling. The results show a strong linear relationship between MOS and pairwise comparison scaling results, the results also reveal, that, adding cross-content comparisons is beneficial in two different ways: reducing the content dependency and increasing the linear relationship between MOS values and pairwise comparison scaling results.

**LIVE image quality assessment dataset** The original LIVE dataset contained only MOS values from 20 observers for 779 conditions. Subsequently authors in [146] complimented the dataset with pairwise comparisons. Here the authors selected 100 conditions and performed full comparison design, comparing each condition to each other five times. The relationship between MOS and pairwise comparison scaling for this dataset have not been analyzed before.

**TID image quality assessment dataset** In my experiments I also consider TID2013 dataset [106], which I have also discussed in Chapter 3. Since the original dataset contained only pairwise comparisons, we, in our work at Transactions on Image Processing [108] complemented it with MOS values, The data was collected by Emin Zerman and the detailed experimental procedure is described in Appendix A.2. Overall 175 conditions were rated by 21 participants.

### 5.4.2 Model complexity

In this section, I validate if the linear relationship is indeed sufficient to explain the relationship between rating and ranking scores.




(a) HDR video (b) LIVE (c) TID2013

Figure 5.1: Polynomial fits into the JOD and MOS scores of the three subjective image and video quality datasets.

To compare the goodness of fit I report adjusted $R^2$ statistic – $R^2_{adj}$, which, unlike simple $R^2$ accounts for the number of model parameters in explaining the variance in the data [96]:

$$R^2_{adj} = 1 - (1 - R^2)\frac{(n-1)}{n-p-1}, \tag{5.8}$$

where $R^2$ is defined as:

$$R^2 = \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y}_i)^2}, \tag{5.9}$$

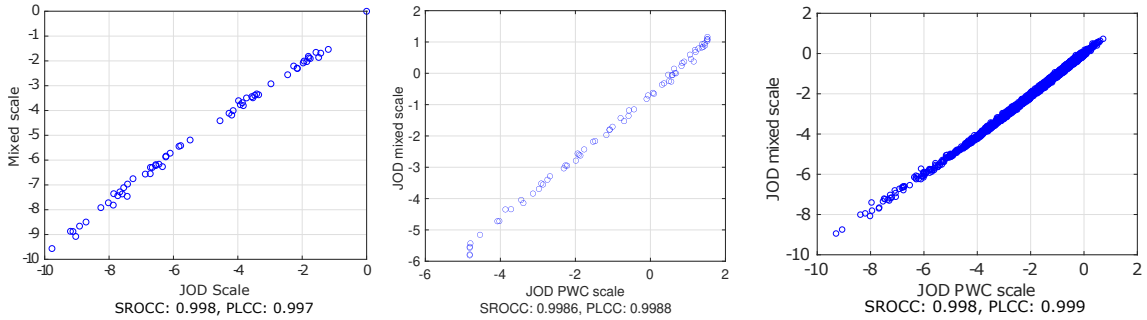
and $n$ is the number of data points in the dataset, $p$ is the number of parameters, excluding the constant term, $\mathbf{y}$ and $\hat{\mathbf{y}}$ true and predicted response variables and $\bar{y}$ is the mean of $\mathbf{y}$.

I fit $1^{st}$ $2^{nd}$ and $3^{rd}$ order polynomials into the JOD and MOS from HDR video quality [149], TID2013 image quality [106] and LIVE image quality [115] datasets. Figure 5.1 shows the scatter plot of the scores and fitted polynomials. The model describing the relationship between JOD and MOS for image quality must be monotonic – an increase/decrease in the quality of an image should result in the increase/decrease of the scores in both scales. Violation of this requirement is possible for $2^{nd}$ and $3^{rd}$ order fits, which is a problem.

$R^2_{adj}$ statistic is given in Table 5.1. There is only a slight increase in $R^2_{adj}$ for TID2013 and LIVE datasets for $2^{nd}$ and $3^{rd}$ order polynomials, for HDR Video dataset $R^2_{adj}$ stays constant. Thus, a higher degree relationship is hard to justify, given the need for additional constraints on the function to be monotonous. Visually, the relation between DMOS and JOD values can be well explained by a linear function, except a few values at the extreme end of the quality scale. For those extreme points, the JOD scale predicts stronger quality degradation than the DMOS scale. This effect can be attributed to the fixed nature of the DMOS scale, where the scale is constrained within a predefined range, e.g, from one to five. The JOD scale, on the contrary, is not constrained, and where a poor or very good quality image cannot be assigned a score beyond the range in the DMOS scale, its' perceived quality is reflected in the JOD scale.

Table 5.1: $R^2_{adj}$ statistic for polynomial fits describing the relationship between MOS and JOD.

| Dataset | $1^{st}$ order | $2^{nd}$ order | $3^{rd}$ order |
|---|---|---|---|
| HDR Video | 0.92 | 0.92 | 0.92 |
| LIVE | 0.87 | 0.89 | .89 |
| TID2013 | 0.77 | 0.79 | 0.79 |




(a) HDR video (b) LIVE (c) TID2013

Figure 5.2: Relationship between the scaling with the mixture of MOS and pairwise comparison data (JOD) and pairwise comparisons (PWC) scale (JOD) of the three subjective image and video quality datasets.

Table 5.2: Average time per trial for rating and ranking experiments and the value of parameter $\eta$ from Equation 5.2, for the three compared datasets.

| Dataset | Time rating | Time Pairwise comparisons | $\eta$ |
|---|---|---|---|
| HDR video | 6.1 | 1.2 | 1.5 |
| LIVE | - | - | 1.02 |
| TID2013 | $7.7 \pm 0.9$ | $3.4 \pm 1.8$ | 1.24 |

### 5.4.3 Experimental effort and observer model

In this section I provide the scaling of three real-world datasets. I also report the response time as well as the value of the parameter $\eta$ per dataset. Overall, in the compared datasets rating scores tend to have little influence on the final scale, while generally having larger variance of the observer model and also longer decision times.

**HDR video compression dataset** For the HDR video dataset the value of $\eta > 1$ (Table 5.2) means that the standard deviation of the observer model in rating experiments is 50% higher for this problem than with pairwise comparisons. The relationship between the JOD scale, incorporating both rating and ranking, and pairwise comparison JOD with only ranking is shown in Figure 5.2a. The relation shows that rating data has little influence on the final scale with the mixture of MOS and pairwise comparison data, which could be explained by the higher standard deviation of the observer model in the rating data. In rating experiments, on average, observers

took five times longer to respond than in pairwise comparison experiments.

**LIVE image quality assessment dataset** The JOD values obtained from pairwise comparisons and scale with the mixture of MOS and ranking data for LIVE dataset have very strong linear relation (Figure 5.2b). However, unlike the HDR video compression dataset, for which the parameter $\eta$ was found to be greater than 1, for LIVE dataset both pairwise comparisons and mean opinion scores are approximately the same. I do not provide the experimental effort for LIVE dataset, as these details are not provided in the original studies [115, 146].

**TID2013 image quality dataset** The value of the parameter $\eta$ from Equation 5.2, for TID2013 is 1.24. Figure 5.2c shows that adding rating data (JOD from the mixture of pairwise comparison and rating scale) has little impact on the final scale, as the rating experiment contains much fewer measurements than the original set of pairwise comparisons. For the TID2013 dataset the response time for pairwise comparisons is lower than for rating experiments.

## 5.5 Comparison of quality scales

I show the differences between the JOD, DMOS, and vote count (VC) quality scales in Figure 5.3. The figure shows three images from the TID2013 dataset and their corresponding quality scores in each scale. I plot on top of each scale the distribution associated with the observer model as a solid line and the one associated with the distribution of the estimate of the mean as a filled area. The observer distribution explains how the quality estimates vary across the population, and it combines inter- and intra-observer variations. The standard deviation of this distribution is fixed for the JOD scale so that the difference of 1 unit corresponds to 75% of the population selecting one condition over another. Since the DMOS scale is approximately linearly related to the JOD scale (as I show in Figure 5.1), the observer model distribution for DMOS also has approximately constant standard deviation across all conditions. However, its value is larger than for the JOD scale ($\eta = 1.24$ found for TID2013). This means that the observer model and its distribution differ between experimental procedures. Observers are more likely to confuse image quality in rating experiments than in a pairwise comparison experiments. The main difference between JOD and DMOS scales is that the distances in the JOD scale are well defined and directly related to the standard deviation of the observer model. In contrast, such distances are arbitrary for the DMOS scale and vary between experiments. This is because there is no strict definition of quality ratings, such as “poor” or “excellent” used in those experiments. Their interpretation depends on the type of distortions that are considered, the training of the participants, and other factors.

The filled-shape distributions in Figure 5.3 tell how confident we are in the estimate of the mean quality score associated with the observer model. If we were to run the experiment multiple

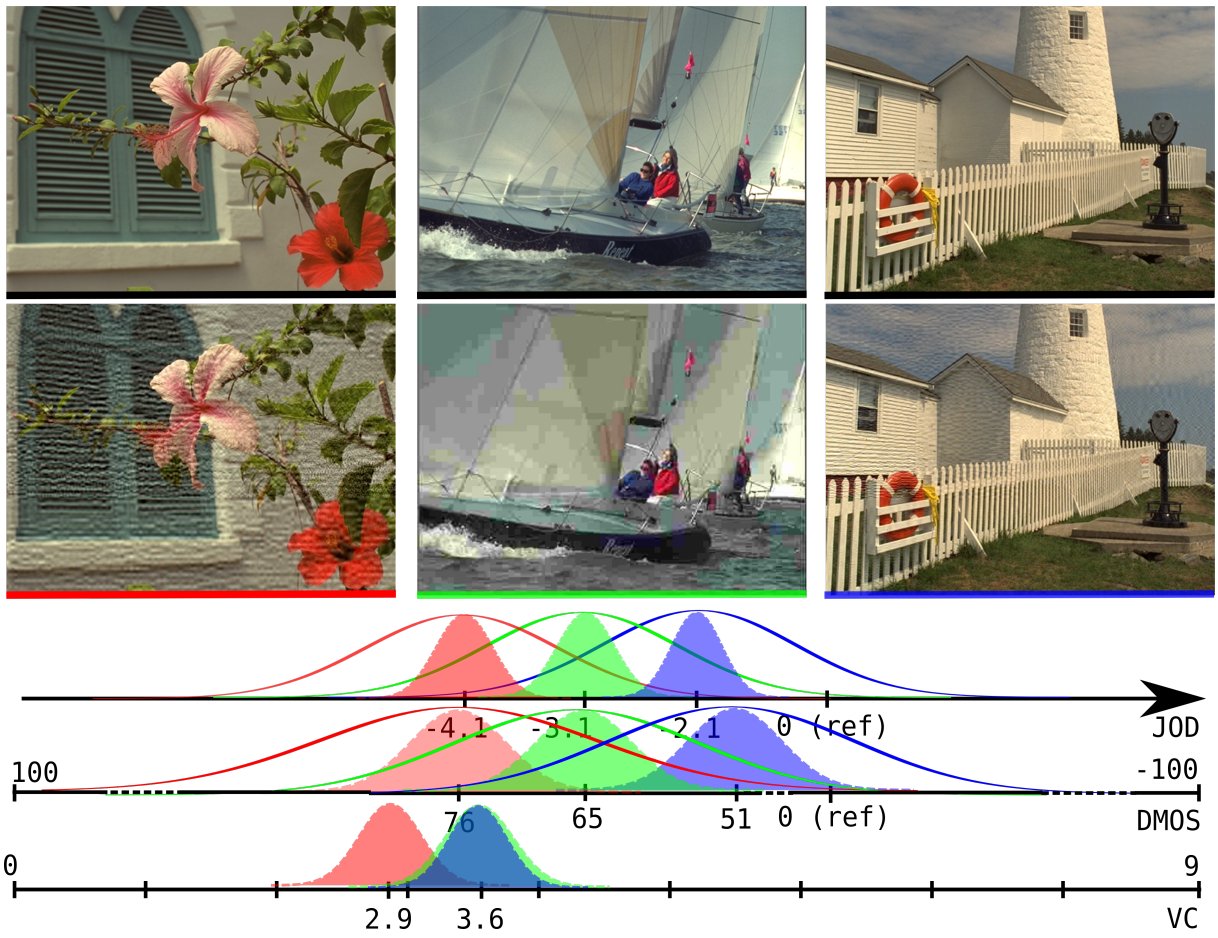



Figure 5.3: The comparison of three quality scales (JOD, DMOS, VC) on the example of images from the TID2013 dataset, underlying observer model distributions (lines) and estimate distributions (filled shapes). Colors used in scales correspond to the underlines below each image. The top row shows reference images, which correspond to (ref) condition on the scale.

times with the same number of observers, the mean quality values across all repetitions would be distributed according to the filled shapes. Such estimate distribution can be readily calculated for the DMOS scale as the standard error of the mean. Finding such distribution for the JOD scale is more complex and can be obtained, for example, by bootstrapping [100] or by Bayesian optimization as in Section 2.2.3.3. As we collect more data, the standard deviation of that estimate distribution decreases, while the standard deviation of the observer model converges to the same constant value of $\beta$. The estimation distribution is typically used to determine whether we have enough data to say that the quality means are different from each other (statistical significance). The distribution of the observer model (inter- and intra-observer variance) tells us about the practical significance of the difference between two quality scores: what portion of the population will make a particular judgment.

Figure 5.3 also shows limitations of vote counts used as a quality measure. Firstly, it does not have an associated observer model. Secondly, the scale does not have the absolute 0 point assigned to reference images.

## 5.6 Experiments: validation

In this section, I analyze the effect of combining rating and pairwise comparison through a set of experiments on benchmark datasets and simulations, for which ground truth is available. I use two measures for performance evaluation: 1) SROCC, which accounts for the ranking, and 2) RMSE, which takes the distance between conditions into account. For some experiments, I also report pearson linear correlation coefficient (PLCC).

### 5.6.1 Berkeley datasets

In order to find the relationship between rating scores and estimations from pairwise comparisons, Shah et al. [114] conducted seven different experiments for various tasks. The tasks were estimating areas of *circles*, *age* of people from photos, *distances* between cities, number of *spelling* mistakes in text, finding the frequency of *piano* sounds, rating *tag-lines* for a product and rating the *relevance* of image search results. Some of these datasets (*distances*, *age*, *piano*, *spelling*) include ground truth, I use those for the analysis.

The measurements from each dataset were used to estimate scores for a) rating data alone, b) pairwise comparison data alone using the scaling procedure from Section 2.2.3, and c) mixed measurements, combining both rating and pairwise comparison data using the scaling method from Section 5.3. When both protocols were combined, I could also estimate factor $\eta$, explaining by how much observer variance differs between rating and pairwise comparisons (Equation 5.2). I also include the total time effort spent collecting each type of experimental measurement. Note that since time effort differs, we cannot directly compare both protocols in terms of accuracy. However, since the standard error decreases as sample size increases, the estimated parameter $\eta$ takes into account both the observer variance and the number of measurements.

I could not scale pairwise comparison results for the *Age* dataset as it contained disconnected components. However, I could use pairwise comparisons when the data from both protocols were combined. This illustrates one of the benefits of mixing both types of data: It allows us to have disconnected components in the graph of comparisons, as long as conditions from both components are rated.

Results of scaling all four datasets are shown in Table 5.3, together with the total time needed to collect the data. Several conclusions can be drawn from these results. Firstly, SROCC and PLCC are similar for both rating and pairwise comparisons. This indicates that both protocols are capable of estimating the ranking between conditions correctly. However, with pairwise comparisons, these ranking results are achieved with less time effort. Secondly, when RMSE is considered, the performance of both protocols depends on the standard deviation of the observer model associated with each protocol, as suggested in [114]. Note that if the $\eta$ parameter is greater than 1, the rating protocol results in a larger standard deviation of the observer model than the pairwise comparison protocol. For example, since $\eta$ is greater than 1 in the *Piano* dataset,

Table 5.3: Results obtained by rating, pairwise comparisons, and mixed experiments in four publicly available datasets. The table shows PLCC, SROCC, and RMSE measures and the fitted $\eta$ parameter explaining the relation between the standard deviation of the observer model for both protocols. Total time for data collection for each type of experiments is also shown.

| | PLCC | | | SROCC | | | RMSE | | | $\eta$ | Total time (secs.) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Rating | PWC | Mix | Rating | PWC | Mix | Rating | PWC | Mix | Mix | Rating | PWC |
| Distances | **0.982** | 0.951 | 0.981 | **0.982** | 0.977 | 0.979 | 0.258 | 0.304 | **0.189** | 0.911 | 15176 | **12844** |
| Age | 0.886 | - | **0.913** | 0.805 | - | **0.875** | 0.442 | - | **0.388** | 0.762 | 6462 | **2790** |
| Piano | 0.889 | **0.944** | 0.938 | 0.830 | 0.927 | **0.939** | 0.602 | **0.316** | 0.334 | 1.737 | 7431 | **5218** |
| Spelling | **0.568** | 0.481 | 0.546 | 0.667 | 0.667 | 0.667 | **0.785** | 0.953 | 0.892 | 0.810 | **9706** | 17505 |

pairwise comparisons result in the smaller RMSE. In the rest of the cases, $\eta$ was lower than 1, which meant that the rating had better results. Finally, concerning the mixing of both protocols, in most cases, this approach has better performance or achieves a good trade-off between both measures. This is expected, as the total amount of measurements is significantly increased when mixing both sources. However, the result of mixing strongly depends on the accuracy of both types of measurement, achieving worse results in cases when one of the protocols was significantly less accurate than the other (for example, the case of *Spelling* for RMSE).

### 5.6.2 Simulations

The goal now is to analyze which measurement is more appropriate given the same time budget. In this section, I rely on Monte Carlo simulations, which assume ground truth quality scores, and can be used to easily test a range of experimental strategies. For every method, the simulation was set to run 100 times. I found this number of Monte Carlo iterations sufficient due to the stability of the results. The first 30 conditions of TID2013 (i.e., associated with content 1) were used as underlining true quality scores for the simulation. I use the Thurstone case V observer model, described in Section 2.2.2, to generate simulated pairwise comparison data. Swiss system was used to guide the search for the pairs to compare using nine rounds, as done in TID2013 [106]. This means that each observer of pairwise comparison experiments measured $9 \cdot (N/2)$ comparisons in total. To generate simulated ratings, I add Gaussian-distributed noise to ground truth data, i.e., assuming that the same observer model is used for both pairwise comparisons and ratings. Each observer measured $N$ conditions for rating. In the simulation, I test how the standard deviation of the observer model for each protocol (related to $\eta$ in the model) affects the results.

I simulated pairwise comparison, rating, and mixed experiments with a varying number of measurements. In the mixed scale case, half of the observers performed a pairwise comparison experiment, and the other half performed rating. In the simulations, I tested i) $\eta = 0.5$ (rating results in less confusion than pairwise comparisons), ii) $\eta = 1$ (both measurements result in the same confusion), iii) $\eta = 1.24$ (the ratio found in TID2013) and iv) $\eta = 2$ (rating has double the standard deviation of pairwise comparisons). The error measures are plotted according to the total time effort needed in Figure 5.4, where time effort corresponds to the number of

Table 5.4: Results for the experiment with data missing (DM) and disconnected components (DC) for RMSE, SROCC and total time effort (in secs).

| | Obs = 10 | | | Obs = 20 | | | Obs=30 | | |
|---|---|---|---|---|---|---|---|---|---|
| Type of measurement | RMSE | SROCC | Time effort | RMSE | SROCC | Time effort | RMSE | SROCC | Time effort |
| Rating | 0.367 | 0.926 | 2310 | 0.277 | 0.958 | 4620 | 0.220 | 0.973 | 6930 |
| Rating with DM | 0.415 | 0.908 | **1848** | 0.311 | 0.947 | **3696** | 0.249 | 0.966 | **5544** |
| PWC | **0.200** | **0.978** | 4590 | **0.143** | **0.988** | 9180 | **0.116** | **0.991** | 13770 |
| Mix with DM and DC | 0.207 | 0.976 | 4677 | 0.151 | 0.987 | 9333 | 0.126 | 0.990 | 13956 |

measurements multiplied by the average time required per measurement found with TID2013.

From the figures, it can be concluded that the measurement with the lowest standard deviation of the observer model achieves better performance and is preferred in all scenarios. However, most measurements converge with enough time effort. When measurement noise is unknown, scaling with the mixture of MOS and pairwise comparison data represents a suitable approach, achieving reasonable performance and a trade-off between both experimental protocols. Mixing also behaves well when data coming from rating is noisier, achieving performance close to pairwise comparisons. It can also be seen that for the case of $\eta = 1.24$ (found with TID2013) pairwise comparisons are more efficient.

Next, I study disconnected components in the graph of comparisons and missing rating data when mixing both scales. Here I do not assume the same budget of comparisons but instead, use a fixed number of observers. The same configuration for the simulation explained at the beginning of this subsection, is used. Table 5.4 shows the case of four approaches: (i) rating; (ii) rating with data missing at random (20% of the rating data is missing); (iii) pairwise comparisons with connected components (PWC); and (iv) mixing with data missing at random (again, same 20%); and disconnected components (here I break the graph of comparisons so that there are always two disconnected components). I perform 100 runs for each method and test it with 10, 20, and 30 observers. I report RMSE, SROCC, and total time effort. The same standard deviation of the observer model, as in TID2013 ($\eta$=1.24) is assumed. Analyzing these results, it can be concluded that mixing is possible even when dealing with disconnected components and missing rating data, showing similar performance to the sole use of pairwise comparisons at a similar time cost. Being able to handle such experimental designs is a highly desirable feature, given that this can simplify the pairwise comparison experimental procedure for large-scale datasets or when mixing different quality assessment datasets, for which missing rating data is common.

Although results presented in terms of RMSE and SROCC provide the means for evaluating the dataset merging procedure in terms of the ability to recover the ranking and regression accuracy, the uncertainty in the recovered scores is not accounted for. Furthermore, for the narrow quality range the reliability of these metrics can be affected by the range effect [87]. Whenever the mapping to the scale is not necessary these shortcomings can be remedied by the method proposed in [63, 64, 65, 36]. The method was developed for evaluation of objective quality metrics and is based on their ability for two compared stimuli reliably identity: if the stimuli are qualitatively different and, if they are, what is their correct ranking.

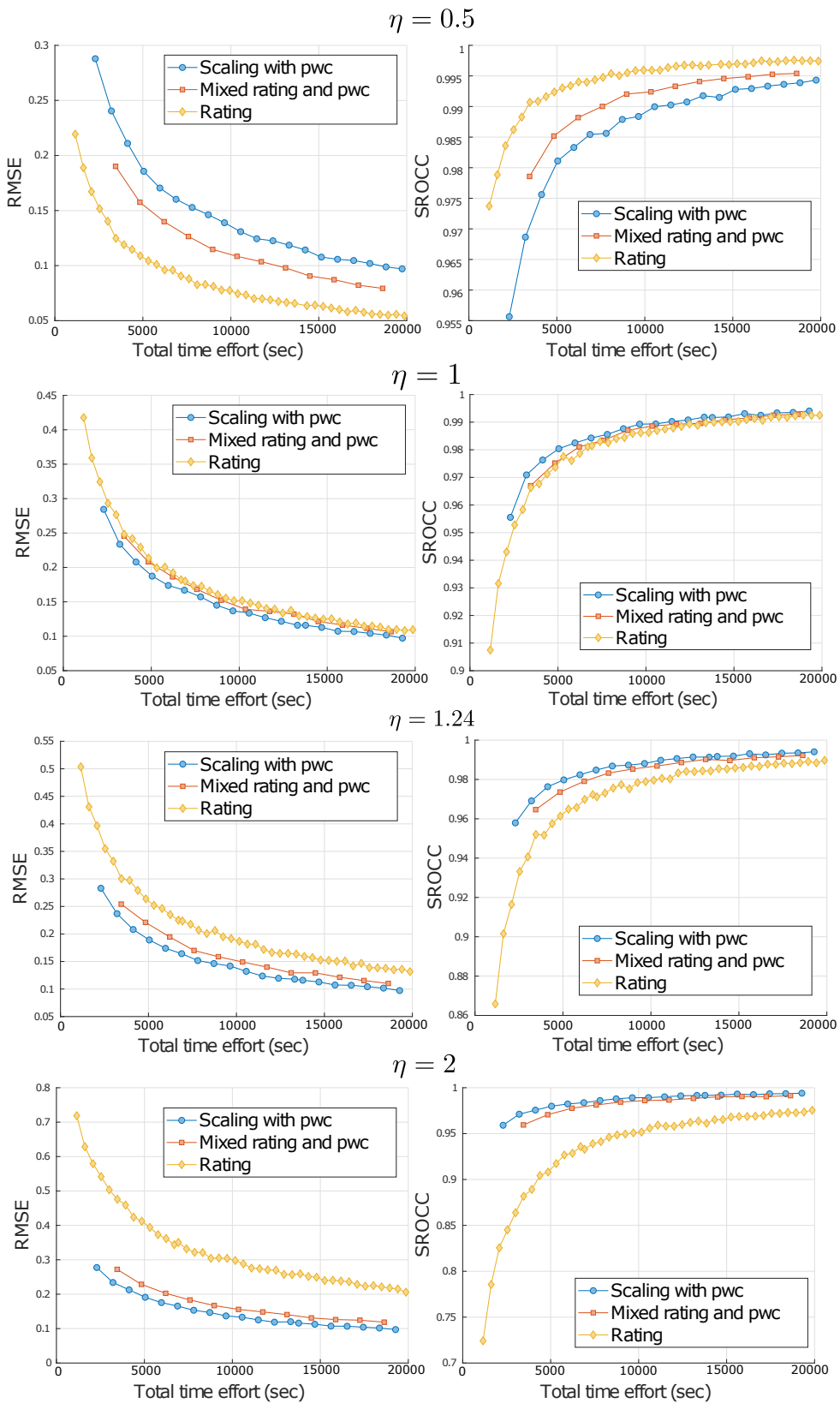



Figure 5.4: Simulation of mixed scale for different values of standard deviation of the observer model (parameter $\eta$).

## 5.7 Summary

In this chapter, I proposed a probabilistic model that can bring pairwise comparison and rating experiments into a unified quality scale. The units in that scale, are scaled accordingly to the combined inter- and intra-observer variations so that 1 unit corresponds to 75% of observers selecting one condition over another (JOD units). The model can estimate observer variation for each experimental protocol and bring measurements to the scale determined by the variation in a side-by-side pairwise comparison experiment.

I test the model on several real datasets and in several simulations. Tests have confirmed the assumption and further revealed interesting patterns in the two experimental protocols. Given the same time effort, there is no clear conclusion about what experimental protocol to use. The decision should rely on the noise of both scales, measured by the parameter $\eta$ in the model. I also found that mixing both protocols can be beneficial in several ways: i) to mix datasets that use either rating or pairwise comparisons, ii) to avoid disconnected components in pairwise comparison experiments, iii) if cross-content comparisons must be avoided and iv) if both types of measurements were previously collected. My model is aimed at recovering the relative relationship between the scores coming from both rating and ranking experiments and does not explicitly account for subject reliability. Other methods [74, 73, 71] have been proposed to recover subjective quality scores from noisy opinion score measurements and would be more suitable for that task.

In the examples above I showed the utility of the proposed method for improving the scores of a single dataset. In the next chapter I apply the proposed method to combining several datasets together. I particularly focus on image quality datasets, for which, the features of the proposed method are particularly useful. Image quality datasets are often fragmented, lacking cross-content comparisons (Chapter 3), collected with different experimental protocols.

# Chapter 6

# Unified photometric image quality dataset

## 6.1 Introduction

Objective image quality assessment metrics, such as peak signal to noise ratio (PSNR) and structural similarity index measure (SSIM) are widely used in image compression, reconstruction, and enhancement [122, 152, 22, 55, 9]. However, most IQA metrics do not account for display characteristics such as the dynamic range of the display, which can influence the perceived image quality. For example, compression artifacts are more visible on a HDR display, than on a dimmed mobile phone [145]. Recent development of HDR displays motivates the need for a new, *photometric* IQA model that accounts for absolute image luminances and can operate in both HDR and SDR images.

The primary limitation to developing such a metric has been the lack of a large-scale subjective image quality dataset. Although attempts have been made to adapt and verify the performance of SDR metrics on HDR content [3, 130, 5, 37, 148] those have not been thoroughly tested due to the lack of a unified dataset. There are methods [63, 133, 36] for metric validation on disjoint datasets, which do not require mapping to a scale and which also provide the means of evaluating statistical significance in differences in performance of the tested metrics. These, however, require access to the raw data, are more suitable for the datasets with large score ranges and test the discriminative power of the metrics, rather than the ability to recover the exact scale. The absence of a large unified dataset also prevented the development of metrics based on machine learning for HDR images, which require large amounts of versatile and heterogeneous data to train. While current machine-learning-based SDR image quality metrics relied on large crowd-sourcing studies [107, 151, 44], these are not straightforward to conduct for HDR content as it requires an HDR display and a controlled viewing environment. Methods overcoming this shortcoming have been proposed. As such [65] proposed an objective function and an algorithm for training machine learning methods on disjoint datasets. The method is, however, more suitable for the applications where the recovery of the exact scale is not necessary.

The available subjective image quality datasets [106, 97, 148, 115, 62, 53, 107, 151, 32],

are insufficient in isolation, as they are limited in terms of the number of images, diversity of distortion types, and image sizes. These datasets also cannot be easily combined due to the use of different experimental protocols and the relative nature of the quality scores. Moreover, incomparable quality scales across datasets prevent the use of absolute scores as a mean of benchmarking IQA metrics, forcing to rely on correlation coefficients, such as SROCC or PLCC. Instead of following a common practice of collecting a dataset from scratch, I argue for the consolidation of existing datasets and focus on combining SDR and HDR image quality datasets to create the largest photometric subjective IQA dataset with a unified quality scale. I perform a set of subjective assessment experiments and construct the largest subjective HDR IQA dataset to date (Unified Photometric Image Quality (UPIQ)) using the psychometric scaling procedure from Chapter 5. The dataset contains *3779* SDR and *380* HDR images from four existing IQA datasets. I show the necessity and advantages of psychometric scaling by comparing it to other strategies to merge datasets.

The contributions of this chapter can be summarized as follows: (i) I perform a series of subjective image quality assessment experiments; (ii) using the psychometric scaling described in the previous chapter construct the largest subjective HDR IQA dataset to date (UPIQ); and (iii) I show the necessity and advantages of the psychometric scaling by comparing it to other strategies for merging datasets. The work in this chapter is based on my manuscript submitted to IEEE Transactions on Multimedia [90].

The rest of the chapter is organized as follows: I first describe existing IQA datasets; I then talk about the datasets that I have selected for alignment in the new UPIQ dataset; after that, I talk about the required experiments for accurate alignment of the datasets and experimental procedure; I then validate the dataset by comparing the produced scaling to the existing re-alignment technique [148].

## 6.2 Existing IQA datasets

To train and validate image quality metrics, one requires a dataset where image quality scores are obtained from human observer judgments. The ideal dataset would contain a large number of psychometric measurements over a range of image content, spanning all dynamic ranges, along with a variety of distortions at different levels of impairment.

Although many subjective IQA datasets exist, they are far from ideal. For example, the largest currently available SDR dataset BAPPS [151] offers only a single distortion type per content — therefore, machine learning based metrics may struggle to learn how to scale the magnitude of distortion. Moreover, image quality scores were not measured extensively, with only two judgments per $64 \times 64$ pixel patches rather than full-sized images. Another recently collected large-scale SDR dataset [107], contains pairwise preference probability, *i.e*. the likelihood of that one image in a pair is more similar to the reference than another. Even though authors

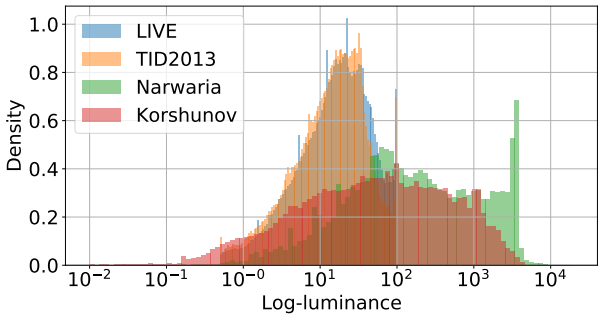



Figure 6.1: Distribution of log-luminance per dataset.

collected a large number of comparisons per image ($> 10$), only within-content comparisons were performed, thus making it impossible to judge relative quality across different contents [149]. The dataset is not publicly available. Existing HDR IQA datasets [148, 62, 97] are significantly smaller than SDR datasets, more homogenous in the versatility of their contents and distortions, and are thus insufficient for the applications outlined in this paper.

Since collecting large amounts of IQA data is time-consuming and expensive, it is preferable to reuse existing datasets. The idea of combining subjective IQA datasets has been considered before. Authors in [148] align subjective scores of HDR datasets using objective quality metrics. The method assumes that the quality predictions from multiple metrics can be used to transform quality scores from one dataset to another. However, this approach cannot be used to combine datasets with different dynamic ranges as no metric can reliably predict the quality of both HDR and SDR images. In this work I propose a different approach to combining IQA datasets. Instead of using predictions from objective metrics, I conduct a set of subjective experiments to measure the relative cross-dataset quality and then use psychometric scaling procedure to place the datasets on a unified quality scale.

## 6.3 Unified photometric IQA dataset

The goal was to create a large dataset consisting of both SDR and HDR images, with the image quality scores on a unified quality scale with JOD units. This was achieved by selecting existing SDR and HDR datasets, collecting additional cross- and within-dataset comparisons, and scaling all the measurements together. I call the dataset UPIQ ("You Pick") — Unified Photometric Image Quality. Before my work, the largest HDR IQA dataset contained only 240 conditions [62]. My dataset has 4159 images, making it the largest HDR dataset to date. Unlike most IQA datasets, images in my dataset are provided in absolute photometric units cd/m$^2$, and scores are given in interpretable JOD units.

Table 6.1: Characteristics of the chosen IQA datasets

| Name | Dynamic range | Experiment | No. images | No. distortions | No. contents | Image sizes (h×w pixels) |
|---|---|---|---|---|---|---|
| LIVE [115] | SDR | MOS | 779 | 5 | 29 | 512×768 |
| TID2013 [106] | SDR | PWC | 3000 | 24 | 25 | 348×512 |
| Narwaria [97] | HDR | MOS | 140 | 2 | 10 | 1080×1920 |
| Korshunov [62] | HDR | MOS | 240 | 3 | 20 | 1080×944 |

### 6.3.1 Selected datasets

Despite a large number of available IQA datasets, only a few of them meet my criteria and could be included in UPIQ. Some datasets were constructed for no-reference quality assessment and do not contain reference images [32]. Other datasets contained a single distortion per content. Thus they provided no means to scale the magnitude of a distortion [32, 53]. For some datasets, the image size was too small for a proper judgment of image quality [151]. While I attempted to scale some datasets, I found their quality scores to be too inconsistent with my measurements to be included in UPIQ [148]. I selected four existing datasets—two SDR (TID2013 [106] and LIVE [115]) and two HDR (Korshunov [62] and Narwaria [97]), which I summarize in Table 6.1. All four datasets span very large dynamic range, as shown in Figure 6.1.

### 6.3.2 Dataset alignment experiments

Subjective measurements for the four original datasets were collected with different experimental protocols, scales, observers, and images. This makes quality scores from different datasets incomparable, ite.*g*., a score of four may indicate high quality in one dataset, but low in another.

To align quality scores from different datasets, I need to perform several types of pairwise comparisons, illustrated in Figure 6.2. Comparisons within a single dataset (within-content, cross-content, and with-reference) are needed to bring the quality values to a common scale of JOD units. This is especially important for the datasets with only MOS (rating) values, as these are provided in an arbitrary scale. I need to find the relationship between MOS and JOD values by estimating the associated parameters ($a, b$, and $\eta$ in Equation (5.5)). The cross-dataset comparisons are necessary to ensure that the quality values are comparable across the datasets. Because different datasets usually do not share the same content, cross-dataset comparisons also tend to be cross-content comparisons. Cross-content comparisons have been shown to be of similar difficulty as within-content comparisons in Chapter 3 and significantly improve the accuracy of a quality scale [149].

**Displays and stimuli** The data necessary for alignment were collected on two different displays. Comparisons of SDR to SDR images were performed on a color-calibrated SDR Samsung

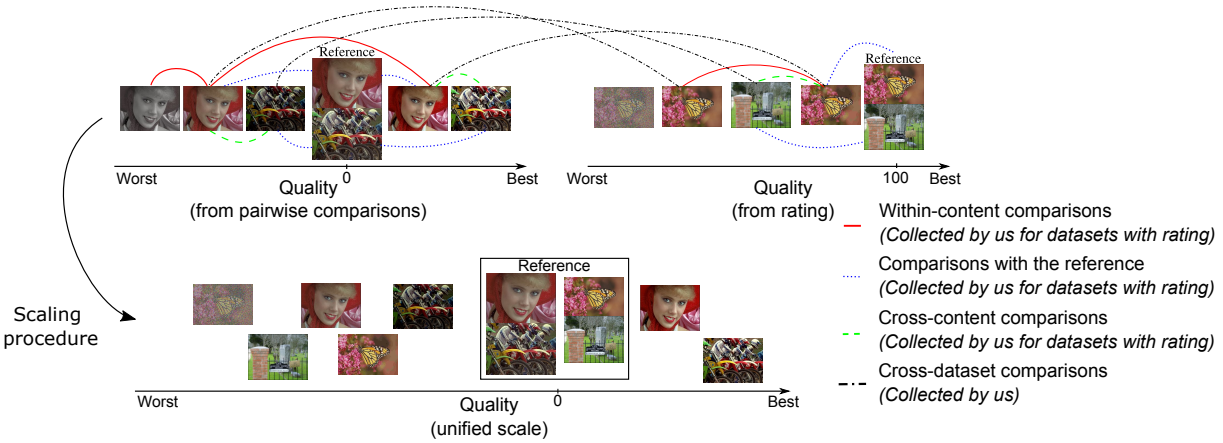



Figure 6.2: Types of comparisons necessary for dataset alignment. The lines link pairs of images selected for pairwise comparisons. Within-content comparisons (solid red lines) are most commonly used in pairwise comparison experiments. However, such datasets often lack comparisons with reference (blue dotted lines), which are useful to provide an absolute anchor of quality. Cross-content comparisons (green dashed lines) are less common, but can substantially improve the scale [149]. Finally, cross dataset comparisons (black dash-dotted lines) are necessary to scale the datasets together.

S32D850T display. The comparisons involving HDR images were presented on a custom-built, color-calibrated 10" HDR display with $2048 \times 1536$ pixels, 15,000 cd/m$^2$peak luminance, and a black level below 0.01 cd/m$^2$. Because I had no information on the displays used in the SDR image quality studies, I used the typical parameters of an SDR display in gain-gamma-offset display model from Equation 2.17. To reproduce SDR images $I_{\text{original}}$ on the HDR display, I used gamma $\gamma = 2.2$, the peak luminance, $L_{\text{peak}} = 100$ cd/m$^2$, and the black level, $L_{\text{black}} = 0.5$ cd/m$^2$, and $L_{refl} = 0$. For HDR images, I reproduced the absolute luminance values used in the original studies. The viewing distance was 90 cm for both the HDR display (164 pixels per degree) and the SDR display (51 pixels per degree). When the image size exceeded the size of the display, I provided a simple panning interface in which observers could use a trackball to inspect different portions of the image.

**Experimental procedure and participants** In order to produce a meaningful unified quality scale using pairwise comparisons for a specific single IQA dataset, one needs a) comparisons of distorted to the pristine quality reference image, b) within-content comparisons to scale different levels of distortion for the same distortion type, c) cross-content comparisons [149], to connect all contents and put them on the same quality scale and d) cross-dynamic-range comparisons (which are also cross-content comparisons), to build a unified scale capturing quality relationships across all luminance levels. For rating, this would be equivalent to having observers rate across all distortions and distortion levels during the same session, instead of having separate experiments. In the case of selected datasets, all of these considerations were taken into account when original data was collected, i.e., each dataset has a self-contained unified quality scale. To align these

data, I need to connect disjoint datasets through pairwise comparisons and find the relationship between rating and pairwise comparison judgments within each of the datasets. This means that for every disjoint rating dataset, I need to collect within dataset comparisons and link all datasets with across dataset comparisons.

The observers were presented with two distorted images and asked to select the image of better quality with respect to the reference. Observers could press and hold a space-bar to view the reference images. Each participant saw images in a different order. Each selected pair of images was compared by 6 participants, with each participant completing approximately 300 trials. Overall, 6000 new comparisons were collected from 20 participants. Note that this required moderate experimental effort as compared to collecting the data from scratch (3000 images in the TID2013 dataset required over 500,000 comparisons). The order of comparisons in every experiment was randomized. I ensured that ITU recommendations were met and that the time for performing one experiment did not exceed 30 mins, to prevent observer tiredness from influencing the experiment outcomes.

Similar to Chapter 3, all participants were research members of the Rainbow group. The mean age of the participants was 29 years, with the youngest participant being 19 years old and the oldest 47 years old. 15 male and five female subjects participated. Six participants belonged to the Asian ethnicity and 14 participants belonged to white ethnic group. All observers were paid for the participation with 10£ Amazon vouchers per hour. Ethical approval granted by the ethics committee with the details of the experiment is provided in Appendix A.1.

I extended the data collected in original datasets and follow-up studies for TID2013 [88, 108] and LIVE [146] datasets with two additional pairwise comparison experiments for all four datasets. In all cases, comparisons were selected so that the two compared conditions were of similar quality to improve the information gain of the collected data and to exclude obvious comparisons [146].

In the first experiment, I collected only comparisons within the dataset, *i.e.* comparing images of the same dataset. This is necessary for finding the relationship between rating measurements and pairwise comparisons. It is only necessary for rating-based datasets, which means I excluded TID2013 from this experiment since I use previously collected pairwise comparisons and rating measurements [88]. I ensured that all three types of previously mentioned comparisons were covered: to reference, within-content, and cross-content. After this first experiment, all the data could be scaled, since I had comparisons to a common reference for all four datasets.

For the second experiment, I compared conditions exclusively from different datasets, connecting every dataset to the rest. Images were chosen to cover the whole quality scale uniformly. I performed several iterations of the pair selection. This is, after conducting a pairwise comparison experiment on a small batch of comparisons, I re-scaled the combined dataset and selected the next batch from the new scale. ASAP algorithm, described in Chapter 4 was not ready at the time of the experiment, however, it also cannot be used for the rating data.

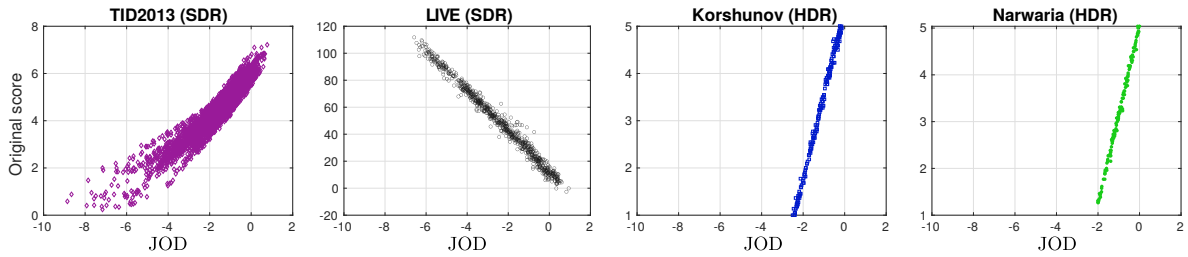



Figure 6.3: Original quality scores versus the results of my scaling in JODs (UPIQ dataset)

### 6.3.3 UPIQ dataset scaling

I combined the newly-collected comparisons with the original data from the four datasets and from the two follow-up studies on TID2013 [88] and LIVE [146]. In total, the combined dataset consists of 571,215 individual pairwise comparison and 27,676 rating measurements, which were input to the scaling procedure from Chapter 5.

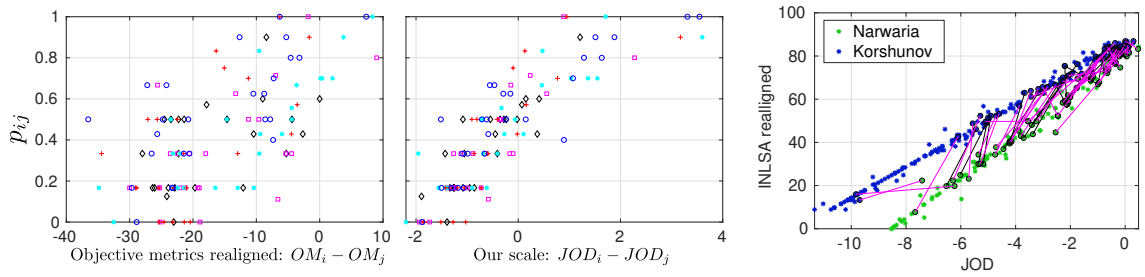
Figure 6.3 shows the relationships between the original quality values of each dataset and the new JOD values from a unified dataset. For all datasets, but for TID2013, the plots show strong linear relationship between the original scores and re-scaled JOD units. Note that the original scores of the TID2013 dataset were obtained with vote counts, reliant only on within-content comparisons. This approach has proven to be less accurate as compared to psychometric scaling in chapter 3.

### 6.3.4 Examples of the UPIQ dataset

Figures 6.6 and 6.7 show sample images from the unified dataset at $JOD = -1$ and $-2$ and are intended to be a visual subjective validation of the final scale. These levels were selected to show images from all four datasets, as images from the HDR datasets (Korshunov and Narwaria) have quality scores above -2 JOD only. Each figure contains four separated sections, each associated to a different dataset. Each section has two rows: distorted and reference images. For display purposes HDR images were converted to SDR with gamma encoding:

$$I_{\mathrm{SDR}} = I_{\mathrm{HDR}}^{\frac{1}{2.2}}. \tag{6.1}$$

As the perceived image quality depends on the display luminance, the SDR images in the figures might be masking or amplifying some image distortions. Thus figures are intended to be an approximate demonstration of the final image quality scale. Nevertheless, images from different datasets at the same JOD level have similar distortion severity. Without the unified photometric image quality dataset (UPIQ) it would be impossible to compare image scores across datasets. Most of the HDR images are distorted only locally, with the overall image quality not deteriorating significantly, as opposed to images from SDR datasets that had uniform distortions applied to them. Narwaria mostly has panorama images, where local distortions are




(a) Scatter plots for five folds

(b) Scatter plot of the two scales

Figure 6.4: (a) Scatter plots for five folds (distinguished by colors and shape); difference in quality scores in the scale constructed with objective quality metrics [148] $\mathrm{OM}_i$ - $\mathrm{OM}_j$ versus empirical probability, $p_{ij}$, of one image $i$ being selected over image $j$ (left); and difference in quality scores in my scale $\mathrm{JOD}_i - \mathrm{JOD}_j$ versus $p_{ij}$ (right). (b) Scatter plot of the two considered quality scales for the HDR datasets. The plot also shows an example of the data used for one of the cross-validation folds. Purple lines represent the training comparisons and black lines the test comparisons (right)

less noticeable due to the size of the image.

### 6.3.5 UPIQ dataset validation

In the following subsections, I compare my scaling with the metrics-based dataset alignment and then demonstrate the improvement in pairwise accuracy.

#### 6.3.5.1 Comparison to previous re-scaling work

Multiple IQA datasets can be merged using an iterated nested least-squares (INLS) algorithm [134]. The algorithm uses existing objective quality metrics to find the relationship between conditions in different datasets. The assumption made is that a weighted combination of metrics should have a high correlation with human judgments. The algorithm iteratively finds weights for the combination of objective quality metrics and aligns subjective quality scores from each of the datasets until convergence. Since no metric exists that has been exhaustively tested on both SDR and HDR images, I validate the results using two HDR datasets (Korshunov and Narwaria), aligned with INLS in the previous work [148]. Figure 6.4b shows that my scaling procedure and the one from [148] lead to substantially different scores. To determine which alignment is more consistent with the subjective judgments, I compute the rank-order correlation between the unprocessed human subjective measurements and scaled values. Since the collected human judgment data comes in the form of pairwise comparisons, I compute the correlation between the empirical probability of selecting one condition over another and differences in quality scores.

The method proposed in [134] relies exclusively on quality scores (regardless of the method used to obtain them) and objective metrics to re-align datasets. At the same time, my approach uses psychometric scaling, which requires additional pairwise comparisons to build the unified

Table 6.2: SROCC between scaled quality scores and empirical probabilities, for my and metric-based scaling. The values are reported for each fold of the cross-validation

| Validation Fold | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Psychometric scaling (our) | 0.77 | 0.72 | 0.62 | 0.74 | 0.71 |
| Objective-metric-based [148] | 0.67 | 0.60 | 0.52 | 0.52 | 0.53 |

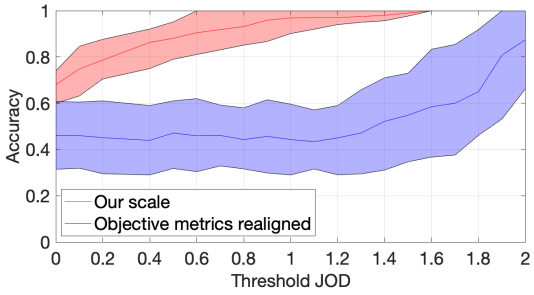



Figure 6.5: Accuracy of classifying cross-dataset conditions into better/worse after alignment with the proposed method. Higher value of Threshold JODs means that more conditions are excluded from training and testing sets. Shaded region is 95% confidence interval.

quality scale. Thus, to ensure a fair comparison, I perform five-fold cross-validation on the collected cross-dataset comparisons. I split cross-dataset comparisons into five equal-sized partitions. In each fold of the cross-validation, I scale the data from four partitions and use the fifth partition for validation. The cross-validation results are given in Table 6.2. My model correlates better with the subjective judgments for each fold, with a mean SROCC of 0.71 versus 0.56 for the method from [148]. It should be noted that the correlation values computed in this manner cannot reach high values because of the measurement noise in the pairwise comparison data. Here it is important to consider that projecting pairwise comparisons to one unique quality dimension with perfect precision is often impossible. For example, previous work has shown how pairwise comparison data could be represented with higher accuracy in higher dimensional spaces, however, with much reduced interpretability of the scale [129]. Figure 6.4a also shows that the relationship is closer to the expected cumulative normal function for my method.

Although the scale in my dataset is significantly better than the one constructed with the method from [148], interpreting correlation coefficients might be challenging. To answer whether an SROCC of 0.71 is accurate enough, I validate the scale by computing the pairwise ranking accuracy for comparisons of varying difficulty.

#### 6.3.5.2 Measuring pairwise accuracy

In this section, I provide an interpretation of the SROCC results from my validation experiment. I will demonstrate that the scale correctly ranks 97% of the pairs that are at least 1 JOD apart.

I first transform the collected data and the produced scale into pairwise rankings. This is, if the quality of $i$ is higher than that of $j$ (as measured in the collected pairwise comparison matrix $\mathbf{C}$) then I set the binary target label $t_{ij}$ to +1, otherwise I set the target $t_{ij}$ to -1. This represents the ground truth pairwise rank averaged across the population. I then compare this ground truth binary label to my predicted binary labels $\hat{t}_{ij}$, following the same procedure but using the output of the scaling algorithm instead of probabilities. Having ground truth and predictions, I compute ranking accuracy. For this, I ran 10-fold cross-validation. In each iteration, I withheld 10% of the compared cross-dataset pairs of conditions for validation. The remaining 90% of compared pairs were for scaling. To compute the ranking accuracy, I assume the minimum threshold distance (in terms of JODs) required for a pair of conditions to be considered, then report the ratio of the number of correctly ranked considered pairs to the total number of considered pairs.

Figure 6.5 shows the accuracy scores for different thresholds of reliable JOD differences, for both my scale and that of [134]. For conditions $> 0.75$ JODs apart (where 63% of observers agreed on the highest quality image, only 13% more than random choice), my scale has 90% accuracy. That is, 90% of the pairs, which are more than 0.75 JODs apart, are correctly ranked by my psychometric scaling. The difference with [134] is very significant, with my scale being consistently much more accurate across different thresholds. Note that the confidence intervals decrease with the value of the JOD threshold increasing. For larger thresholds the scaling method is less likely to make mistakes. Thus, after the threshold value of 1.6 JODs the scaling methods had 100% accuracy in my set of experiments.

## 6.4 Summary

A large scale photometric image quality dataset would enable the development of deep learning based image quality metric. However, existing HDR image quality datasets are small in size and expensive to collect. I remedy this limitation and increase their size by merging together a mixture of both HDR and SDR datasets. My merging procedure, presented in the previous chapter, requires collecting additional data (cross-dataset comparisons), however, the experimental effort is much smaller compared to collecting the dataset from scratch. The accuracy of the resulting dataset is much higher than that of alternative procedures [134, 148]. Unified Photometric Image Quality dataset (UPIQ), is the first large-scale dataset that can be used for training and testing HDR image quality metrics. Images in my dataset are represented in absolute photometric and colorimetric units and their quality scores are provided in the interpretable JOD units [100]. In the next chapter I show the utility of my new dataset by re-training existing quality metrics and show that the dataset is large enough for training deep architectures.

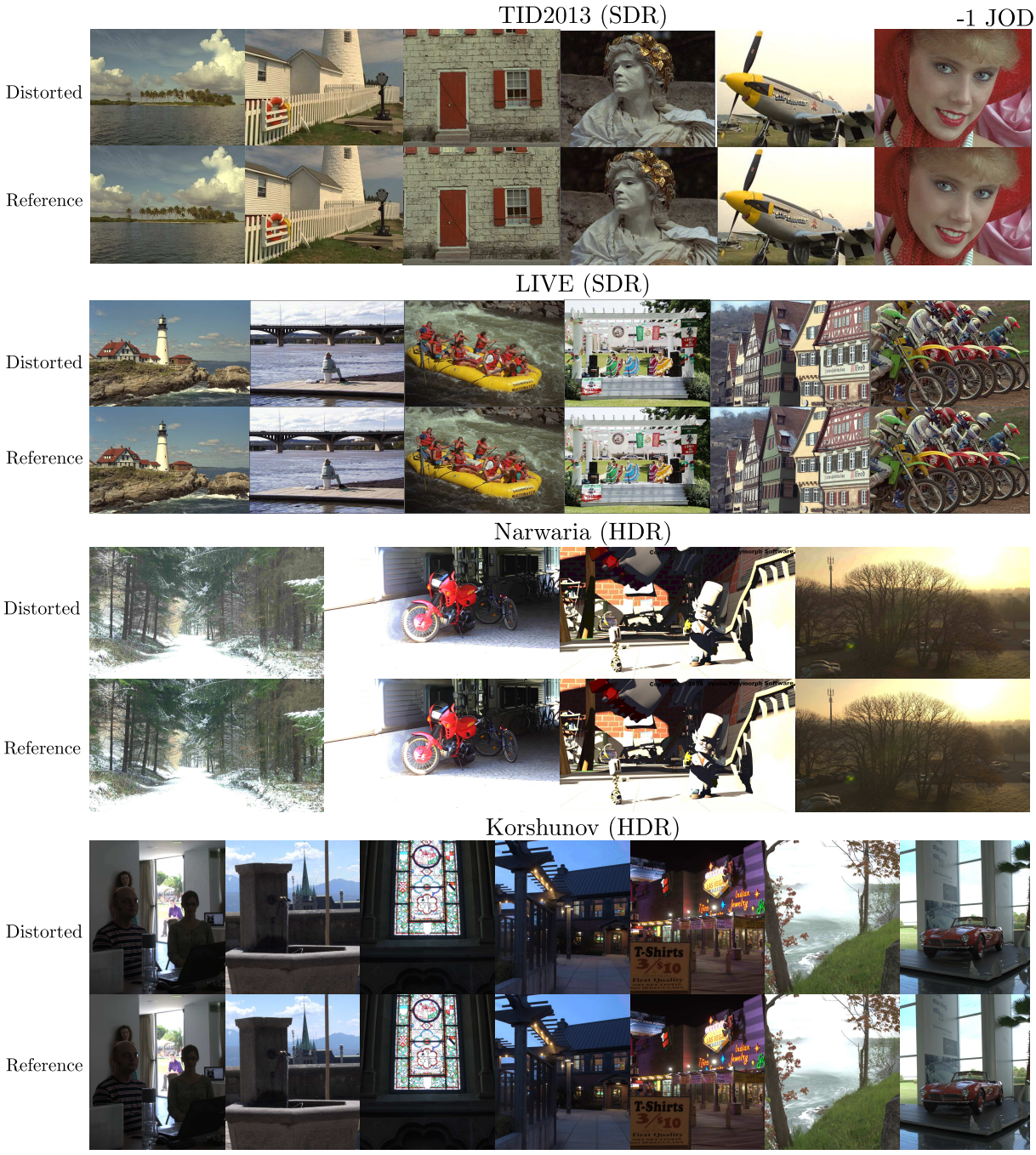



Figure 6.6: A selection of images from the four combined datasets at approximately $-1$ JOD level. Each dataset has two rows: distorted and reference images. I converted HDR images to SDR with gamma correction and gamma 2.2. Images from different datasets at the same JOD level have similar distortion severity. Without a unified dataset it would be impossible to compare image scores across datasets.

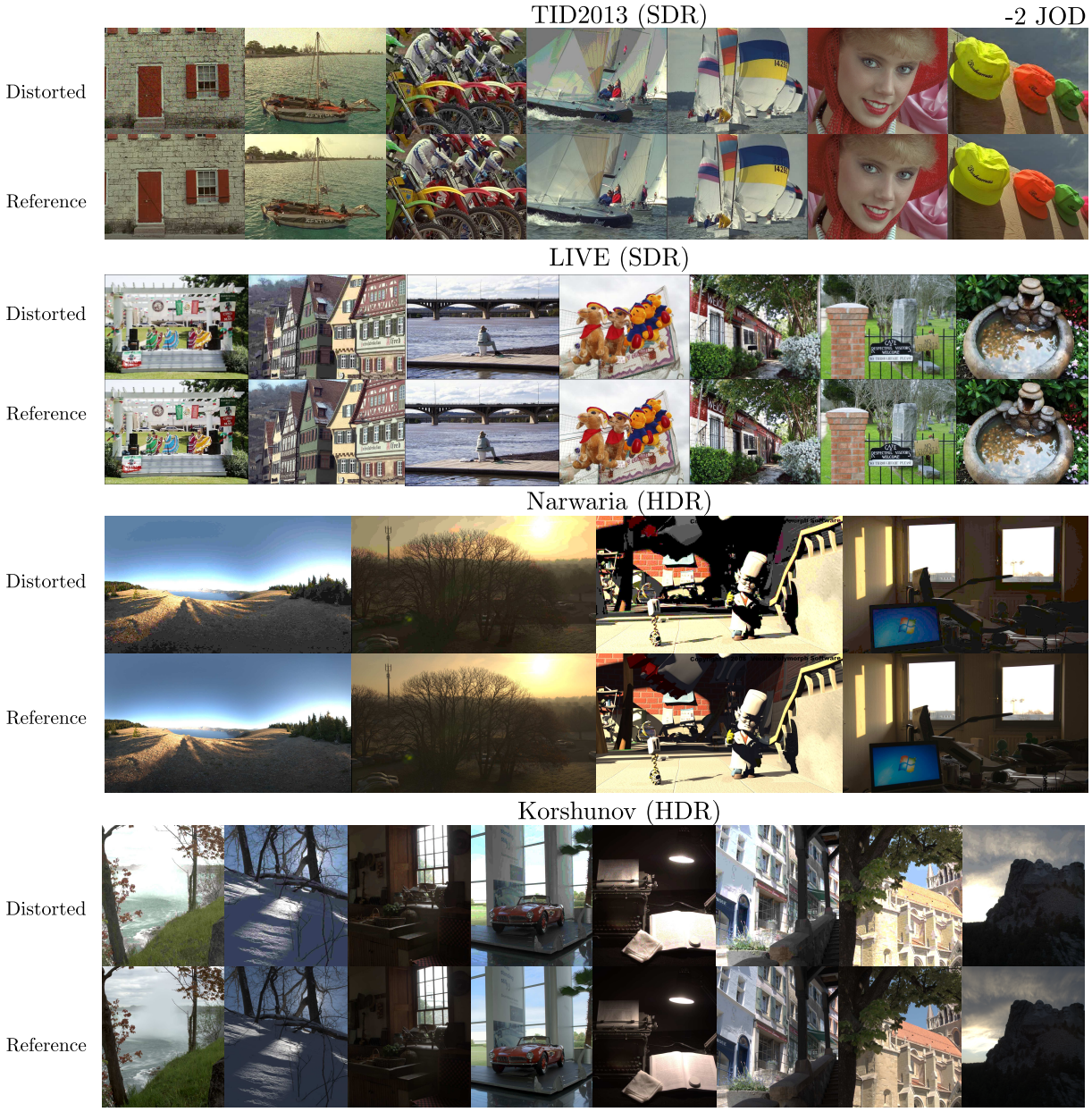



Figure 6.7: A selection of images from the four combined datasets at approximately $-2$ JOD level. Each dataset has two rows: distorted and reference images. I converted HDR images to SDR with gamma correction and gamma 2.2. Images from different datasets at the same JOD level have similar distortion severity. Without a unified dataset it would be impossible to compare image scores across datasets.

# Chapter 7

# Photometric objective quality metrics

## 7.1 Introduction

This chapter shows how the large scale dataset, presented in Chapter 6 can be used to re-train and benchmark existing HDR image quality metrics. I show that the proposed dataset is sufficiently large for deep architectures by training a CNN-based full-reference *photometric* image quality metric. The advantage of training on the unified dataset is shown in comparison with training on a single dataset and multi-task learning on disjoint datasets. The utility of training HDR metrics on the new dataset is shown in an application to image compression. The new dataset, code, and metrics are available online[1]. The work presented in this chapter is based on my publication at IEEE Transactions on Multimedia [90].

The rest of the chapter is organized as follows: first I discuss existing approaches to IQA and how existing IQA metrics can be adapted to the varied dynamic range of a display; I then show how the unified dataset can be used to train data-driven metrics; following that I compare the performance of existing HDR quality metrics; I conclude by showing the importance of a unified dataset in training data-driven metrics by comparing to a multi-task learning approach.

## 7.2 Related work

Authors in [3] proposed a simple method to adapt the standard dynamic range metrics to HDR contents. Where an SDR image quality metric is applied to an image transformed to a perceptually uniform domain, through either PU or logarithmic transform. I discussed the general pipeline for that in Section 2.3.4. PU-transform was also used to adapt a no-reference deep SDR IQA metric to operate on HDR images [54]. Unlike [54], which, due to the absence of a sufficiently large HDR dataset had to rely on a metric trained on SDR images, my work enables training of a deep HDR IQA metric on both SDR and HDR images.

[1] https://www.cl.cam.ac.uk/research/rainbow/projects/upiq/.

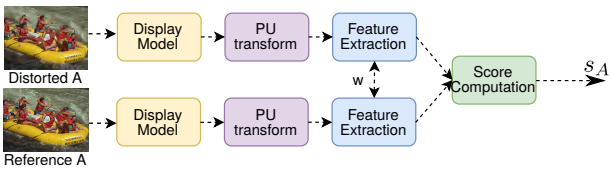



Figure 7.1: The pipeline used to train PU-PieAPP on absolute scores. Images first pass through the display model and are then fed to the PU transform. The feature extraction network with shared weights extract representations, which are passed to the score computation network, producing a score per image.

Another HDR quality metric was proposed in [69], while developing a rendering algorithm for displaying an HDR image on an SDR display. Two linear HDR image are first transformed to SDR via inverse gamma encoding. They are then decomposed into frequency channels, using the Laplacian Pyramid [12]. The resultant decomposition is then normalized by a weighted sum of the localized element-wise amplitudes. The $L_\alpha$-norm of the differences between the coefficients within each frequency channel (that is, the absolute value of each coefficient difference is raised to the power $\alpha$, and then summed over the entire channel, an overall sum is then raised to the poser $\frac{1}{\alpha}$), combined across channels using an $L_\beta$-norm was used as quality predictor.

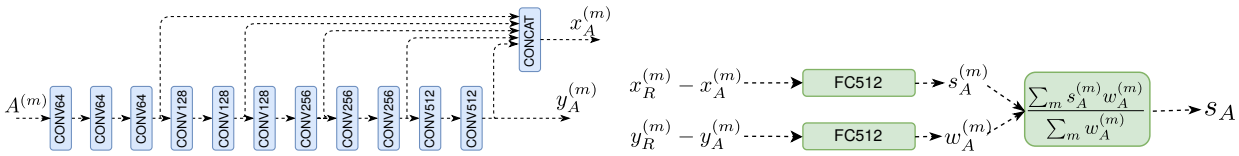
High dynamic range video quality metric (HDR-VQM) was proposed in [99]. The authors first transform HDR values, to emitted luminance, assuming a display model. The emitted luminance values are then transformed to the perceptually uniform space. Log-gabor filters are then employed in the frequency domain to extract features related to image quality. Features from the reference and test images are compared. Although the metric is tailored for video quality it can also be used for image quality assessment.

Authors in [82] proposed a metric based on comprehensive model of an early visual system (HDR-VDP2). The model accounts for the intra-ocular light scatter, photoreceptor spectral sensitivities, separate rod and cone pathways, contrast sensitivity across the full range of visible luminance, intra- and inter-channel contrast masking, and spatial integration. The metric achieves state-of-the-art performance across the non-deep learning based metrics.

## 7.3 Training a data-driven HDR metric

UPIQ is sufficiently large to allow us to train from scratch a CNN-based image quality metric to predict the quality of both SDR and HDR images. The metric combines the ideas behind PU encoding [3] (Section 2.3.4) and a recently proposed CNN architecture for image quality assessment (PieAPP) [107]). I will refer to this metric as PU-PieAPP.

**Architecture** The diagram of the deep metric architecture is shown in Figure 7.1. The metric takes as input a pair of test and reference images and produces a single quality score $s_A$ in




(a) Feature extraction (b) Score computation

Figure 7.2: (a) The feature extraction network takes image patches as an input and has two outputs: one from a patch passing through the whole network and another formed from skip connection. The network has 11 convolutional layers with $2 \times 2$ max-pooling after every even layer. (b) The score computation network computes patch-wise weights and scores, the weighted average produces the final score

JOD units. The input images need to be transformed into the display domain to account for the dynamic range of the displayed images. This is achieved by a *display model* from Equation 2.17 for SDR images, or by scaling color values according to the presentation conditions from the original papers for HDR images. Then, the resulting trichromatic color values (with ITU Rec. 709 primaries [47]) are converted into approximately perceptually uniform units using the PU transform (Section 2.3.4), which is applied individually to each color channel. Such encoded images are fed into the PieAPP architecture, which combines a pair of feature extraction networks with shared weights with the score computation network, identical to those in [107].

The detailed architecture of the PieAPP network is shown in Figure 7.2. For every input patch $m$ of reference $R$ and distorted $A$ images, the feature extraction (FE) network has two outputs: $y^{(m)}$ from the input passing through the whole network and $x^{(m)}$ formed by the concatenation of the flattened outputs of layers at different depths of the network. The score computation (SC) network takes two inputs: the difference between $x_R^{(m)} - x_A^{(m)}$, which is passed through a fully connected layer, predicting patch-wise error $s^m$ and the difference $y_R^{(m)} - y_A^{(m)}$, which is passed through another fully connected layer, producing the patch-wise weight $w^{(m)}$. The two outputs $s^{(m)}$ and $w^{(m)}$, are then used to produce the weighted average of all per patch scores – a quality score of the entire image $s_A$. Note that passing two reference images through the network will result in the $x_R^{(m)} - x_A^{(m)} = 0$, thus the output of the quality estimation function $f(A, B)$, will be constant, defined by the bias of the score computation network.

**Alternative Architectures** I experiment with several CNN architectures to find the one that generalizes the best. Since the CNN-based metric can be trained end-to-end, it could potentially learn the PU-transform. I replaced the PU transform with a logarithmic function followed by scaling to the 0-1 range and then trained the network. The prediction error was much higher for the logarithmic function (RMSE 0.68) compared to the PU transform (RMSE 0.47). This confirms that the PU is beneficial for quality predictions in SDR/HDR images, even for CNN-based metrics.

**Training** In contrast to the original PieApp implementation [107], I train the network as regression rather than learning-to-rank. The scaling procedure achieves the same goals as learning-to-rank, but offers a more accurate observer model and allows us to split the problem into two separate scaling and learning steps. I train the network from scratch, using Adam optimizer on 4 NVIDIA P100 GPUs. Every tested architecture was run for 500 epochs and the model with the best performance on the validation set was saved (using 60-20-20 split into training, validation and test sets). I train the network on $64 \times 64$ patches. To densely cover the whole image, the image is stratified by a uniform grid and patches are sampled at random positions in each grid cell (jittered sampling). The grid size is selected to give approximately square cells. In training, I extract 1024 patches per image. I found that 1024 was the largest number of patches that I could process on my GPU. When testing, I sampled twice the number of $64 \times 64$ patches needed to cover the image. This number was optimal in terms of the time versus performance trade-off.

## 7.4 Benchmark of HDR quality metrics

Although HDR image quality metrics have been compared in many studies [62, 148, 5], none of them could test the metrics on an extensive dataset such as UPIQ. Therefore, I use UPIQ to test existing HDR metrics.

Here I consider full-reference metrics, which are either adapted to HDR content using PU-transform: PU-PSNR, PU-SSIM [137], PU-FSIM [150], or are designed to work with HDR data: HDR-VQM [99], HDRVDP-2.2 [82, 98] and NLP [69]. I also evaluate no-reference metrics, adapting them to the HDR content with PU-transform: PU-BRISQUE [93], PU-PIQE [132] and PU-NIQE [94], due to their widespread use and competitive performance. Finally, I adapted existing SDR CNN-based metrics to HDR content using the PU-transform: PU-KonCept512 [44] (no-reference) and original PU-PieApp (original) [107] (full-reference). I did not re-train deep metrics on UPIQ but used weights provided by the authors. For comparison, I also include full reference PSNR and FSIM metrics, not adapted to the HDR content.

Most objective metrics predict values that are non-linearly related to absolute quality in JOD units. The scatter plot of the considered metrics predictions versus those of the JOD's is provided in Figure 7.3. Since my goal is to predict the absolute quality, I need to map metric predictions to JODs. I follow a standard approach [115] and fit a logistic function mapping objective quality $o$ into absolute JOD units $q(o)$:

$$q(o) = \frac{a_1}{1 + e^{a_2(o-a_3)}} + a_4 o + a_5, \tag{7.1}$$

where $a_1, \ldots, a_5$ are fitted parameters. Fitting a logistic function is necessary for computing performance measures, RMSE and PLCC, but it also helps to scale objective metric results into

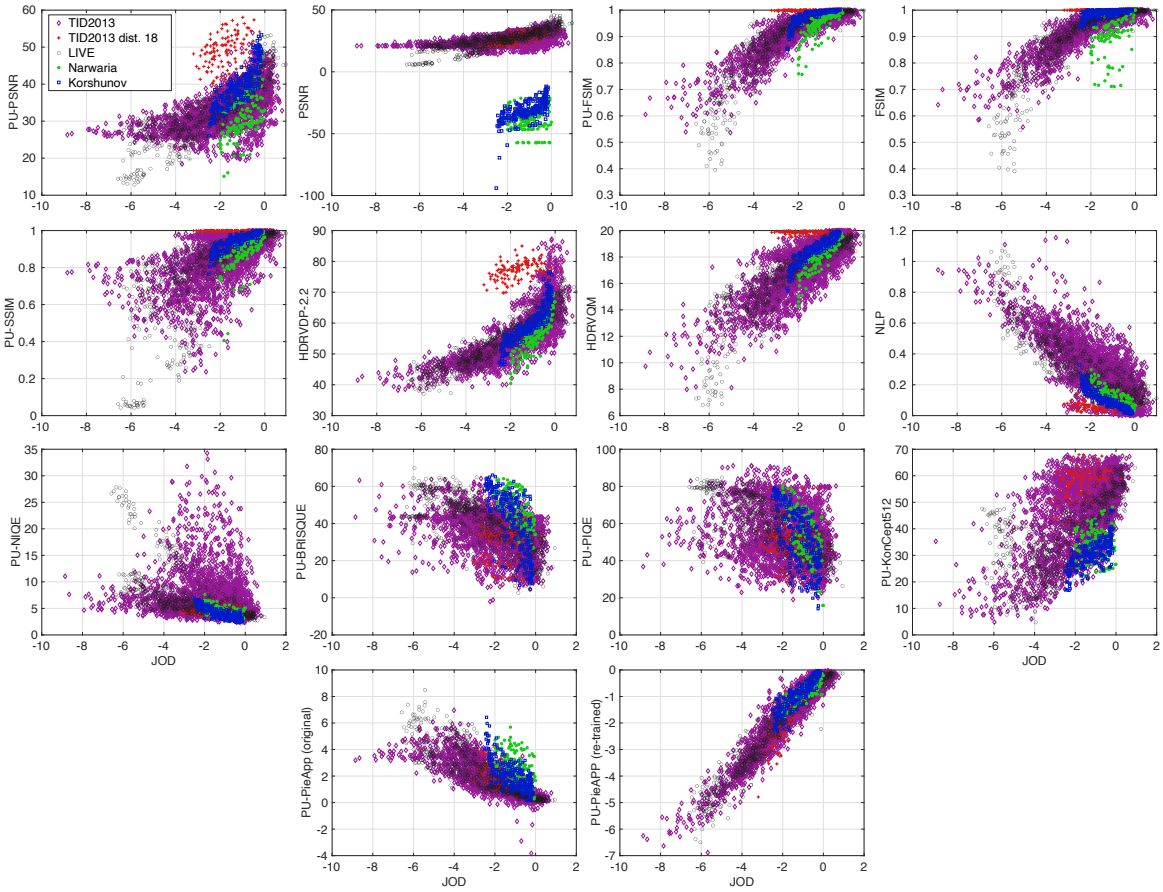



Figure 7.3: Objective metrics predictions vs. JOD quality values. I separately label distortion 18 from TID2013 (change of color saturation) as it introduces the biggest prediction error for the metrics that operate on luma/luminance values and ignore color information.

interpretable and comparable units of JODs. For example, while the result of PU-SSIM of 0.98 is difficult to interpret, the result of -1 JODs tell us that 75% of the population will select reference condition over the distorted one.

For a fair comparison, I use the same 5-fold split into 80-20% training and testing dataset when fitting psychometric function for the tested metrics. In each fold, a different portion of the entire dataset is tested while ensuring that no content is shared between training and testing sets. I also ensure that each subset (TID2013, LIVE, Narwaria, and Korshunov) was split in the same 80-20 ratio. Note that since PU-PieAPP (re-trained) is trained on the quality scores from the UPIQ dataset, I do not need to fit the logistic function into its prediction.

Described above training approach is suitable only for the case where training is performed on a single dataset with unified scores. When there is no access to a unified dataset, a different approach, proposed in [65] would be suitable for training on multiple disjoint datasets. The approach builds on the framework developed in [63]. The overall cost function for training machine learning algorithms is based on the ability of the algorithm to discriminate conditions in different versus similar and better versus worse scenarios. Furthermore this approach would be

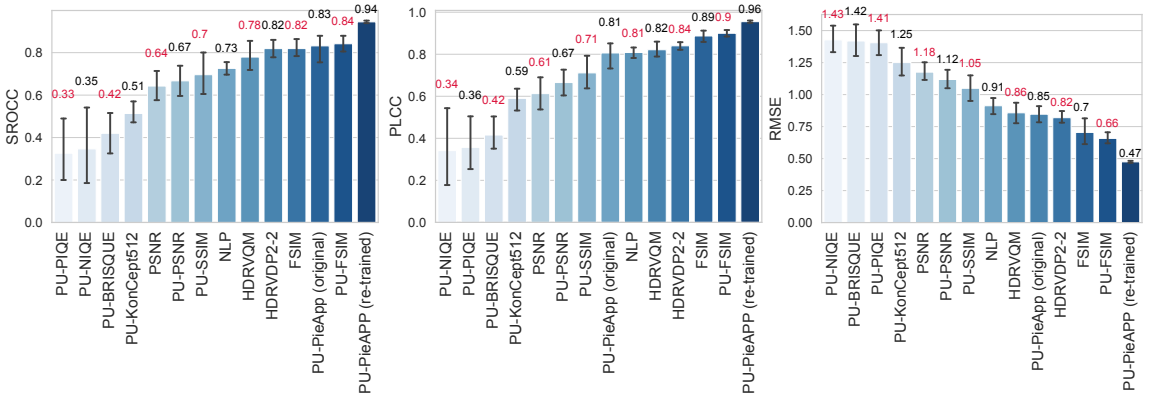



Figure 7.4: Cross-validation results for all trained metrics, expressed as SROCC, PLCC and RMSE. Error bars denote 95% confidence intervals.

more suitable for applications where the recovery of the exact scale is not necessary and thus, would allow to avoid unnecessary mapping.

### 7.4.1 Cross-validation

**Cross-content validation** The most common approach to the validation of learning-based quality metrics is the split into training and testing sets that contain different content but share distortion types. Note that I took extra care to isolate the same content in LIVE and TID datasets as those share some of the reference images. The results for the 5-fold cross-validation on such cross-content splits, shown in Figure 7.4, indicate that PU-PieAPP (re-trained) outperforms existing hand-crafted metrics. PU-PieAPP shows 30% improvement to the second-best performing metric, PU-FSIM, followed by FSIM without the PU-transform. I later show that the performance difference between PU-metrics and original metrics is much higher when tested on HDR datasets (images from SDR datasets dominate UPIQ).

No-reference metrics, based on hand-crafted features, exhibit the worst performance — the PU-transformation applied to the images distort the statistics that these metrics rely on. Deep learning-based no-reference metric PU-KonCept512 does not perform well either.

Original PieApp adapted to my dataset with PU-transform, performs reasonably well on SDR images (SROCC: 0.8764). However, it exhibits poor performance on both HDR datasets (SROCC: 0.5791). This is expected, as the metric was trained on SDR images, and the range of PU-transformed HDR images is greater than that of SDR.

**Cross-validation schemes** To understand what mixture of data is required to train quality metrics robustly, I experiment with different data partitioning schemes. For this experiment, I selected 5 best performing metrics from Figure 7.4. Table 7.1 lists the training and test data combinations I tested and the corresponding results.

PU-PieAPP generalizes well when trained cross-content (C-C), i.e., the training and test set

Table 7.1: Test RMSE and SROCC for different data partitioning schemes and the best performing metrics. (C-C – cross-content, C-D – cross-dataset, C-DR – cross-dynamic-range). I remove the listed test portion of the UPIQ from training and test on it. Test PLCC for different data partitioning schemes and the best performing metrics. (C-C – cross-content, C-D – cross-dataset, C-DR – cross-dynamic-range).

| Metric | C-C Test: sel. cont. | C-D Test: TID2013 | C-D Test: LIVE | C-D Test: Narwaria | C-D Test: Korshunov | C-DR Test: HDR | C-DR Test: SDR |
|---|---|---|---|---|---|---|---|
| **RMSE** | | | | | | | |
| PU-PieAPP | **0.47** | 0.92 | 0.70 | 0.68 | 0.62 | 0.72 | **1.29** |
| PU-FSIM | 0.66 | **0.65** | **0.50** | 0.26 | 0.29 | 0.68 | 1.39 |
| FSIM | 0.70 | 0.65 | 0.51 | 0.45 | 0.52 | 1.17 | 1.61 |
| HDRVDP | 0.82 | 0.88 | 0.64 | 0.24 | 0.21 | 0.78 | 1.34 |
| HDRVQM | 0.86 | 1.04 | 0.68 | **0.23** | **0.20** | **0.39** | 1.43 |
| **SROCC** | | | | | | | |
| PU-PieAPP | **0.94** | 0.78 | 0.87 | 0.82 | 0.79 | 0.74 | 0.65 |
| PU-FSIM | 0.90 | **0.80** | **0.96** | 0.87 | 0.93 | 0.71 | 0.77 |
| FSIM | 0.89 | 0.80 | 0.96 | 0.54 | 0.52 | 0.45 | 0.54 |
| HDRVDP | 0.84 | 0.78 | 0.94 | 0.94 | 0.94 | 0.81 | **0.82** |
| HDRVQM | 0.82 | 0.71 | 0.92 | **0.95** | **0.95** | **0.87** | 0.60 |
| **PLCC** | | | | | | | |
| PU-PieAPP | **0.96** | 0.78 | 0.89 | 0.78 | 0.75 | 0.73 | 0.67 |
| PU-FSIM | 0.90 | **0.89** | **0.96** | 0.87 | 0.90 | 0.66 | 0.77 |
| FSIM | 0.89 | 0.89 | 0.96 | 0.53 | 0.66 | 0.34 | 0.51 |
| HDRVDP | 0.84 | 0.83 | 0.93 | 0.89 | 0.95 | 0.72 | **0.78** |
| HDRVQM | 0.82 | 0.78 | 0.92 | **0.89** | **0.95** | **0.86** | 0.62 |

Table 7.2: SROCC between the difference in quality scores $s_A - s_B$, where $A$ and $B$ are images from different datasets and empirical probability $p_{ij}$ for the multitask network.

| TID2013 LIVE | Narwaria Korshunov | TID2013 Narwaria | LIVE Narwaria | TID2013 Korshunov | LIVE Korshunov |
|---|---|---|---|---|---|
| 0.46 | 0.27 | 0.33 | 0.46 | 0.26 | 0.10 |

overlap in distortion types but not in content. However, the performance of this deep-learning metric drops significantly if one or more datasets are missing from the training set. This, and the poor performance of no-reference metrics in Figure 7.4, show that learning-based metrics are prone to overfitting when the training dataset is not sufficiently large.

As expected, SDR metrics exhibit better performance tested on SDR datasets. The same holds for metrics for HDR content – they perform better on HDR datasets. PU-FSIM and FSIM have similar performance when tested on SDR datasets. However, tested on HDR, PU-FSIM performs significantly better than FSIM, clearly demonstrating the need for the PU-transform.

## 7.5 Value of cross-dataset measurements and a unified scale

Collecting data is time-consuming and expensive. Hence a method capable of learning the implicit unified quality without the need for additional data is desirable. To verify if the network

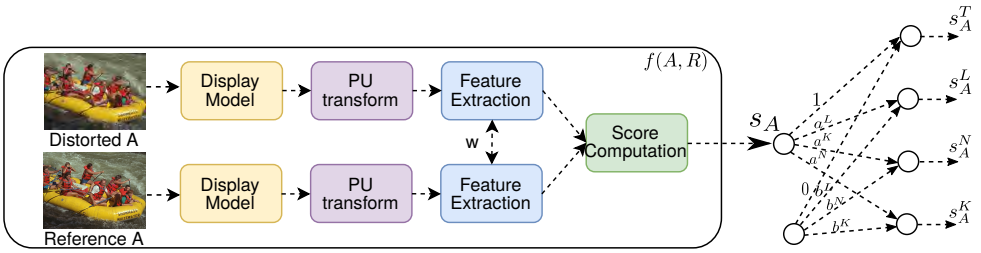


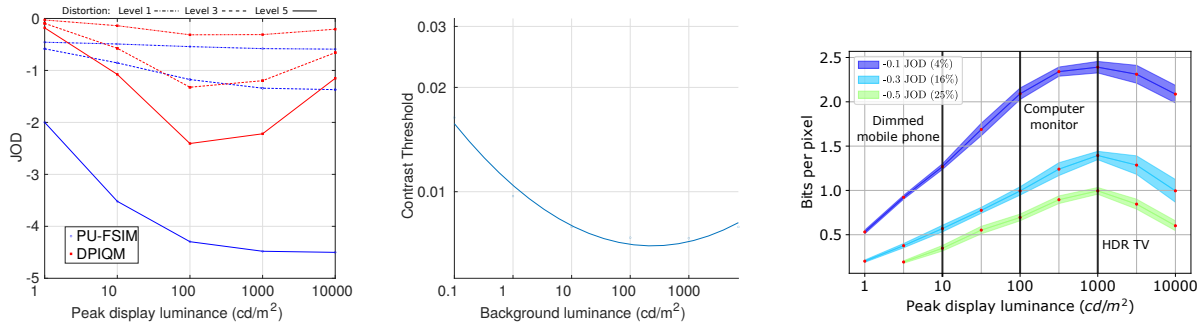

Figure 7.5: Multitask network. The network is trained to predict original scores from individual datasets. Similar to the scaling procedure the network learns the implicit quality $s_A$ and parameters $a$ and $b$ for each dataset. To constraint the scores I set parameters of TID dataset $a^T = 1$, $b^T = 0$.

can learn this implicit quality and cross-dataset relationship, I train the network using a multi-task learning approach, where it is assumed that all datasets share the same feature representation for quality. The architecture of the network is given in Figure 7.5. The $f(A, B)$ part of the network is the same as PU-PieApp and produces a score $s_A$, which is a unified, underlying quality for disjoint datasets. Similar to my scaling procedure from Chapter 5, the scores from individual datasets are linked with the unified $s_A$ via a linear relationship. For example, the quality score $s_A^L$ for the LIVE dataset would be predicted with $a^L * s_A + b^L$, where $a^L$ and $b^L$ are learned parameters. These parameters from individual datasets are treated and learned as individual tasks. Since quality scores are relative, I constraint them by setting parameters of the TID2013 dataset $a^T = 1$ and $b^T = 0$. To allow for faster convergence, I standardized scores from the separate datasets. The training procedure for the multi-task network was the same as for the PU-PieApp.

Similar to Section 6.3.5 of Chapter 5 I compute the correlation between the difference in quality scores $s_A - s_B$, where $A$ and $B$ are images from different datasets and empirical probability $p_{ij}$. The detailed results are given in Table 7.2. Neither of the cross-dataset relationships is well captured by the multitask network, as no information about the relative quality relations across datasets is provided. This information is however available to the scaling method, which uses cross-dataset comparisons to link the scales.

## 7.6 Brightness-adaptive image quality and coding

I investigate how PU-PieAPP and PU-FSIM (the two best-performing metrics) predict the quality of images shown on displays of different brightness. Figure 7.6a left shows the quality predictions averaged across all contents for JPEG distortion from the TID2013 dataset as a function of peak display luminance. Both metrics predict improvement in image quality as the display gets darker than 100 cd/m$^2$. However, the predictions diverge at luminance levels above 100 cd/m$^2$: PU-FSIM predicts a decrease in image quality, whereas PU-PieAPP predicts an improvement. Interestingly, PU-PieAPP's U-shaped curve is consistent with the recent measurements [142] of human contrast detection thresholds. I show an example of these measurements in the right




(a) Prediction across luminance range (b) Contrast thresholds (c) Brightness-Adaptive Coding

Figure 7.6: (a) The quality predictions as the function of display peak luminance. The predictions are shown separately for three distortion levels of JPEG and averaged across contents from the TID2013 dataset. (b) Contrast threshold function for a varied display brightness. (c) Bits per pixel for JPEG compression to achieve constant perceived quality at different luminance levels. Different colors represent different quality levels. Shaded regions are 75% confidence intervals.

plot of Figure 7.6a. It is notable that PU-PieAPP implicitly learned human contrast sensitivity function across the luminance range.

The ability of PU-PieAPP to learn the human contrast sensitivity across the luminance range and predict the effect of absolute image brightness on image quality enables novel applications. Here I control the compression rate of a standard JPEG codec (the ”quality” parameter) to achieve a distortion at a desirable JOD level. The selected levels signify that about 4% (-0.1 JOD), 16% (-0.3 JOD), or 25% (-0.5 JOD) of the population will correctly indicate a compressed image from a test and reference pair (discounting 50% guess rate). Figure 7.6c shows the distribution of the required bit-rate to compress 200 pristine test images from the LocVis dataset [140] at the desired JOD level. The vertical bars in the plot denote the peak luminance levels of three displays, typical for an HDR TV, computer monitor, and a dimmed mobile phone. The plot shows that the bit-rate could be substantially reduced when images are shown on a dimmed mobile phone, but it should be increased for HDR TV. Furthermore, the difference is larger for images encoded with high quality. Such information could be useful, for example, for internet caches that attempt to reduce the amount of data sent to mobile web browsers. Only photometric metrics, trained on both SDR and HDR images, can be used for such applications as they can capture the effect of absolute luminance on image quality. A more detailed validation of such brightness-adaptive image coding can be found in [145].

## 7.7 Summary

Using the large scale dataset, presented in Chapter 6, with diversity of content, distortions, and luminance of the assessed images, I trained and tested existing HDR image quality metrics as well as a CNN-based image quality metric (PU-PieAPP) capable of predicting quality of both

HDR and SDR images. My results indicate that the dataset is sufficiently large for PU-PieAPP to outperform the state-of-the-art HDR quality metrics and generalize across different content and image distortions. The new dataset and trained metrics could be used to test HDR reconstruction and coding methods or to control their performance adaptively, as I show on the example of brightness-adaptive image coding.

In the next chapter I validate the ability of the metrics trained and tested in this chapter to predict visually lossless threshold (VLT) for display brightness and viewing distance dependent compression on a new VLT dataset.

# Chapter 8

# The effect of display brightness and viewing distance on image quality

## 8.1 Introduction

In the previous chapter I have shown that IQA metrics, adjusted to account for display brightness can enable novel applications. In this chapter I verify the performance of the metrics on the task of finding a visually lossless threshold for image compression under varied display brightness and viewing distance.

Finding a threshold at which the human eye cannot perceive changes introduced to an image can be beneficial for computer vision, computer graphics, and image processing algorithms. Such a threshold can be used, for example, to adjust the image/video compression level so that the size of the bit-stream is minimized while the distortions remain mostly invisible. I will refer to such quantization or quality level of an image/video codec as a visually lossless threshold (VLT). Such a VLT depends not only on the image content, but will also vary with the viewing conditions: the viewing distance and display brightness. While all image quality metrics account for image content, very few of them account for viewing conditions [4, 82, 62, 145]. Such lack of accountability for viewing conditions makes VLT prediction unreliable across different displays and different viewing distances.

The goal of this chapter is to provide a dataset that could be used to evaluate quality and visibility metrics on the task of finding VLTs under different viewing conditions. VLT was measured at viewing distances corresponding to 30 ppd (pixels per degree) and 60 ppd and two peak brightness levels: 220 $\mathrm{cd/m^2}$, common to computer displays, and 10 $\mathrm{cd/m^2}$, replicating the brightness of a dimmed phone. The dataset was collected for two popular compression methods: JPEG [135] and WebP [117]. I then benchmark both hand-crafted and data-driven image metrics on the dataset.

The main contributions of this chapter are: (i) a visually lossless image compression dataset with varying peak display brightness and viewing distance and (ii) performance evaluation of

image quality and visibility metrics on the new dataset.

The work in this chapter is based on my manuscript submitted to the Human Vision and Electronic Imaging conference [89]. I am grateful to my collaborator, Nanyang Ye, who has designed and conducted the experiment.

## 8.2 Related work

In this section I will go over the related work on most commonly used methods for image compression, ways of finding a VLT and talk about existing datasts with VLTs.

### 8.2.1 Image compression

Image and video compression methods can be divided into three types: lossless, lossy, and visually lossless. Lossless compression methods preserve all information in the decompressed image [45]. However, their compression rates are much lower than for lossy compression. Lossy compression methods allow for much higher compression rates, but they compromise on the visual image quality, often introducing noticeable artifacts [109, 117]. Two commonly used lossy image compression standards are JPEG and WebP. Both methods have an adjustable quality factor (QF). The QF varies between 0 (lowest visual quality and highest compression rate) and 100 (highest visual quality and lowest compression rate). Visually lossless compression methods introduce distortions but it is ensured that they are unlikely to be noticed [79]. Visually lossless compression requires finding a VLT — the maximum compression level at which distortions are invisible to most observers. In this work, I choose the VLT for which at most 25% of observers can perceive the compression artifacts.

Most image compression algorithms rely on hand-crafted methods with a relatively small number of adjustable parameters. With the advent of machine learning, deep neural networks, capable of learning complex relations in an automated manner without the need for explicit assumptions, have also been used for image compression [81]. Their results, however, rely on the quality and quantity of training data.

### 8.2.2 VLT prediction

Several works have focused solely on predicting the VLT. For example, authors in [29] proposed SUR-Net, a deep Siamese-CNN architecture predicting the satisfied-user-ratio (SUR) curve. For each compressed image, their model predicts the proportion of the population, which would not notice the distortion. The authors first pre-trained the network to predict image quality and then fine-tuned on a smaller dataset to predict SUR. The authors later extended the work to SUR-FeatNet [75], where the pre-trained Inception-V3-CNN is used to extract the features from the images, which are then fed into a smaller network trained to predict the SUR curve. Authors

in [76] took a different approach and trained a deep binary classifier. Which, given the distorted and undistorted source image, predicts whether the distortion would be noticed or not.

In my work I do not focus on developing a new VLT predictor, but focus on validating existing approaches. I include image quality and image visibility metrics, discussed in the previous chapter, for comparison, which can also be used for the purpose of finding a VLT.

### 8.2.3 Existing datasets

To train and test image metrics to predict VLT, one requires a benchmark dataset where, at varied quality levels, the visibility of distortions is judged by a group of observers. While several image quality and visibility datasets are available, they either do not contain VLTs [106, 141, 145], or present images compressed only with JPEG codec at a single luminance level at a fixed viewing distance. As such, the authors in [56] collected a subjective dataset, called MCL-JCI, with visually lossless thresholds for images compressed with JPEG codec. The dataset contains 50 source images, with VLTs identified by 30 observers for each image. Overall the dataset was collected with 150 observers. Similarly, authors in [78] created a dataset with VLTs for panoramic images compressed with JPEG codec. Each of the 40 panoramic source images was inspected by at least 25 observers, using a head-mounted display.

Unlike exisiting datasets, the presented dataset in this work consist of VLTs for images compressed by not only JPEG but also WebP standard. More importantly, the proposed dataset includes VTLs based on different viewing distance and display brightness.

## 8.3 Proposed dataset

The goal was to create a dataset with VLTs for images depicting a varied selection of contents and compressed with two codecs (JPEG and WebP), which were viewed on monitors of different peak luminance (10 cd/m$^2$and 220 cd/m$^2$), and at different viewing distance (30 ppd and 60 ppd). I used 20 captured by Dr. Rafal Mantiuk images with $1920 \times 1281$ resolution, which were obtained from DSLR RAW images and stored in a lossless format. Half of the images were compressed with JPEG (libjpeg[1]) and the other half with WebP (libwebp[2]). Since it was more important to capture the variety of content than to compare both codecs, I did not attempt to collect VLT for the same contents and both codecs. To ensure that the observers could find the distortions in a reasonable amount of time, I cropped the stimuli to $512 \times 512$ pixels. Examples of the images from the dataset are given in Figure 8.3. To uniformly cover the entire range of compression quality values at a reasonable number of points, I incremented QF from 2 to 100 in steps of 2, where 100 is an image compressed with the highest quality and also the highest bit-rate.

[1] `https://github.com/LuaDist/libjpeg`
[2] `https://github.com/webmproject/libwebp`


(a) VLT experiment stage 1 (b) VLT experiment stage 2

Figure 8.1: Two stages of visually lossless compression experiment. The images were displayed at the native display resolution without upscaling (images are shown out of scale for better legibility).

### 8.3.1 Procedure

To improve the accuracy of the VLT measurements, the method of adjustment and 4-alternative-forced-choice (4AFC) protocol, controlled by an adaptive procedure were combined.

In the first method-of-adjustment stage, observers were presented with reference and compressed images side-by-side and were asked to adjust the QF of the compressed image until they could not distinguish it from the reference (Figure 8.1a). A half-a-second blank frame with a middle-gray background was displayed when changing compression levels to prevent observers from relying on temporal changes to guide their choice. The compression level found in the first stage was used as an initial guess for the 4AFC procedure [50], in the second stage. Observers were shown three copies of an undistorted reference image and one distorted image in a random quarter, as shown in Figure 8.1b. They were then asked to choose the distorted image by clicking on the spots where they could see the distortions. The location of those mouse clicks for all four viewing conditions are shown in Figure 8.3. The next value of the QF parameter for the distorted image was then selected with the QUEST adaptive procedure [139]. Between 20 and 30 4AFC trials per observer for each image were collected. To find the VLT for each image I fitted a psychometric function to the collected data.

#### 8.3.1.1 Display

The experiments were conducted in a dim room ($\sim$10 Lux). The screen was positioned to minimize ambient light reflections. The viewing distance was controlled with a chin-rest. Observers viewed a 27” Asus PG279Q display (2560$\times$1440) from a distance of 80 cm (angular resolution: 60 pixels per degree) or of 40 cm (angular resolution: 30 pixels per degree). For the bright condition, display brightness was set to its maximum value (220 cd/m$^2$). For the dark condition, a 1.2 neutral density filter (Rosco E-Colour 299) was placed in front of the monitor and

adjusted the display brightness so that the effective peak luminance was 10 cd/m$^2$. The display color calibration conformed with ITU-R recommendations [47] and sRGB transfer function.

#### 8.3.1.2 Observers

The data were collected from 19 observers aged between 20 and 30 years old, with normal or corrected-to-normal vision. All observers were trained and paid for their participation and were naïve to the purpose of the experiment.

### 8.3.2 Data analysis

Before analyzing the data, I removed outliers. For each scene I followed the standard z-score procedure and removed VLT measurements which were more than two standard deviations away from the mean.

#### 8.3.2.1 VLT distribution

I estimate VLT distribution across the population assuming it to be normally distributed, similarly as in [29]. The proportion of the population that can detect the distortion is, thus, described by the function:

$$P_{det}(l) = 1 - \Phi(l; \mu, \sigma^2)\,, \tag{8.1}$$

where $l$ is the JPEG/WebP quality factor and $\Phi(l; \mu, \sigma)$ is the cumulative normal distribution with the estimated mean and variance of VLT distribution for each condition. The plots of those functions for each image are shown in Figure 8.2.

The plots of probability of detection in Figure 8.2 show a clear pattern, with the brighter display (220 cd/m$^2$) and shorter viewing distance (30 ppd) requiring the highest quality factor. The opposite is shown for the darker display (10 cd/m$^2$) seen from a larger distance (60 ppd). The slope of those curves indicates how the VLT varied between the observers. The slope is similar for most tested conditions, with a few exceptions. For example, the slope is shallower for *i7webp* with 220 cd/m$^2$and 60 ppd (red), indicating a higher variance between the observers. Figure 8.3 (red dots) shows that for this image, the distortions were spotted in different parts of the image for different observers (sky, trees, grass, people), which could explain the variability. Therefore, I opt to use a lower value of $P_{det}$ to find a VLT. This way the VLT reflects the results for the most attentive observers, who could spot the most critical part of an image. Another interesting case is image *i17jpeg*, for which the curves are close together and the slopes are steep (low inter-observer variance). As seen in Figure 8.3, the distortions for that image were consistently detected by most observers in a large smooth area of the sky.

For each condition the VLT is found by selecting the quality factor $l$ for which $P_{det} = 0.25$ (less than 25% of the population can see the difference). Such *population* VLT values are shown

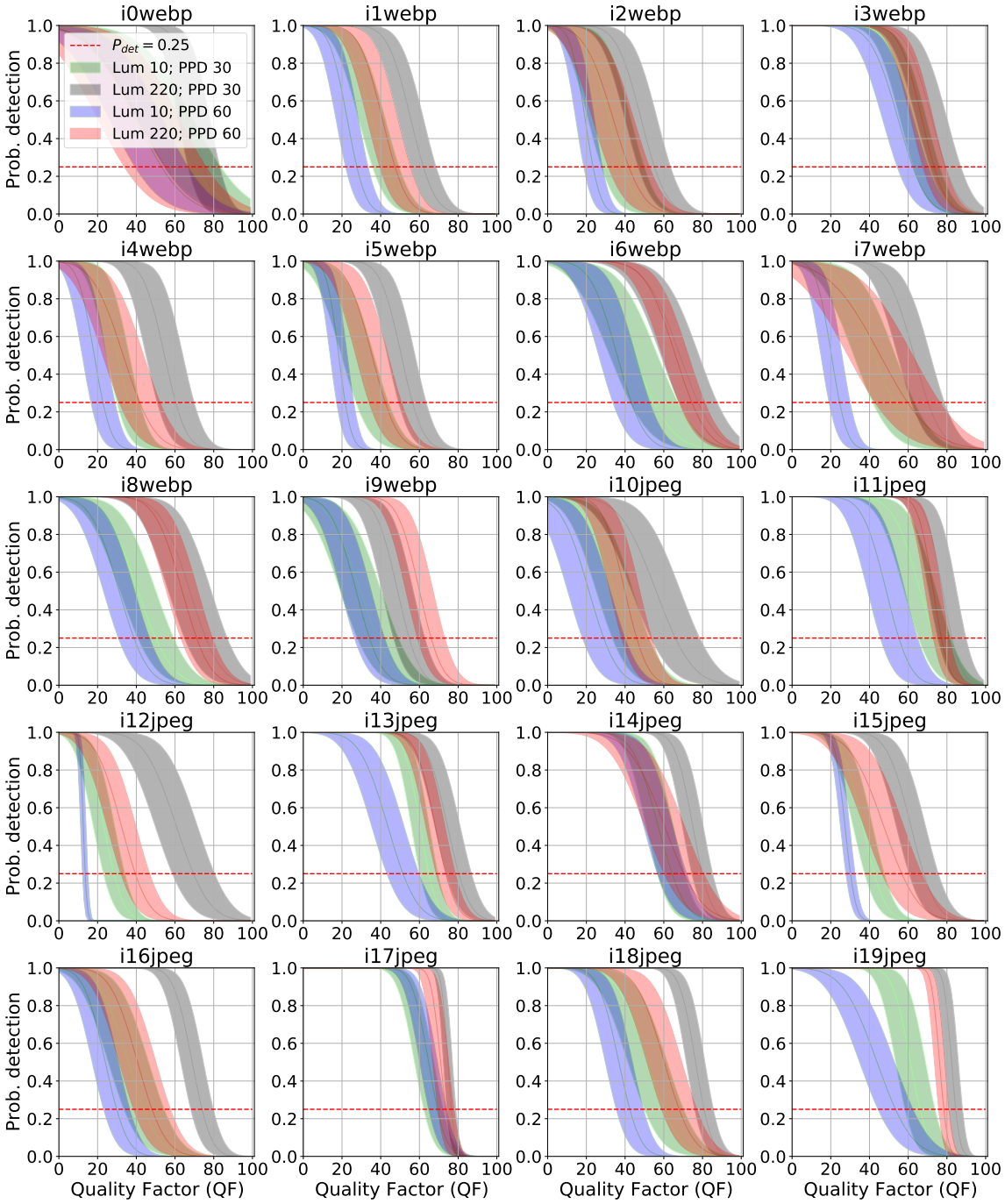



Figure 8.2: Distribution of visually lossless thresholds across the population. The shaded regions are 99% confidence intervals (due to limited sample size). The dashed red lines denote the detection threshold used to find the VLT.

in Figure 8.4. Another interesting observation is that some images were less affected by the display brightness (*i3webp*, *i11jpeg*, *i17jpeg*) than the others (*i12jpeg*, *i16jpeg*). As shown in Figure 8.3, the distortions in less affected images were typically spotted in bright and smooth regions, for which Weber's law can compensate for the loss of display brightness. It is also important to note that the changes in VLTs between the viewing conditions are different for each


Figure 8.3: The images used in the experiment and the distributions of observer clicks (blue for 10 cd/m$^2$, 60 ppd, green for 10 cd/m$^2$, 30 ppd, black for 220 cd/m$^2$, 30 ppd and red for 220 cd/m$^2$, 60 ppd). Colors are consistent with Figures 8.2 and 8.4.

image, suggesting a strong interaction between image content and viewing conditions.

## 8.4 Evaluation of image metrics

In this section, I evaluate how accurately image quality and visibility metrics can predict VLTs. The predicted VLTs then can be used to automatically adjust compression parameters to achieve a trade-off between image visual quality and bit-rate. I evaluated both hand-crafted image quality metrics (PSNR, SSIM [137], MS-SSIM [136], FSIM [150], HDRVQM [99]) and a deep photometric image quality metric, PU-PieApp, trained on UPIQ dataset and presented in the

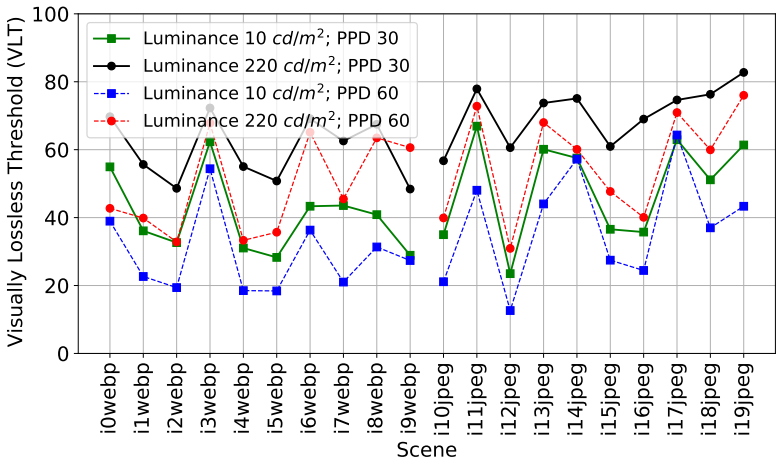



Figure 8.4: Visually lossless threshold (VLT) for all four viewing conditions shown in different colors. Note that the color used are consistent with Figures 8.2 and 8.3.

previous chapter. Additionally, I evaluate two visibility metrics: a high dynamic range visible difference predictor (HDRVDP3) [82], and a CNN-based deep photometric visibility metric (DPVM) [145]. Note that for the HDRVDP3, I report the results of both the visibility predictor (labelled HDRVDP3V) and the quality predictor (labelled HDRVDP3Q). I could not include metrics intended for VLT prediction in [76] and [29], as I could not access the trained models. Furthermore, correspondence with the authors revealed that the models were trained on very small datasets and thus could not generalize well beyond the training data.

### 8.4.1 Luminance-aware metrics

Similar to the previous chapter, I extend the hand-crafted image quality metrics, PSNR, SSIM, MS-SSIM and FSIM to account for display luminance, by transforming the image to luminance emitted from a display assuming a model of that display and then converting it into the Perceptually Uniform (PU) units [4].

To transform standard-dynamic-range (SDR) images from gamma-encoded sRGB colors to linear RGB values shown on the high-dynamic-range (HDR) display, I used gain-gamma-offset display model from Equation 2.17. In the experiments I set the peak luminance of the display to $L_{\text{peak}} = 10\,\text{cd/m}^2$, or $L_{\text{peak}} = 220\,\text{cd/m}^2$, and the black level $L_{\text{black}}$ was set to $0.001\,L_{\text{peak}}$ (assuming that the contrast of the display was 1000:1 and there were no ambient light reflections) for both cases. I used PU encoding (Section 2.3.4) to adapt PSNR, SSIM and FSIM to different luminance conditions. Other metrics (HDRVDP-3, DPVM, HDRVQM and PU-PieApp,) are photometric by design, and thus do not require the application of the PU-transform.

### 8.4.2 Viewing distance-aware metrics

To account for viewing distance in VLT prediction, I followed the procedure in [145]. I re-sampled images with angular resolution of 30 ppd, to the angular resolution of 60 ppd with Lanczos filter [126] before passing them to the metrics. 60 ppd is the highest resolution in the dataset and also a reasonable limit for most visual tasks, since the sensitivity of visual system drops rapidly below 30 cpd [6]. I did not perform the re-sampling for HDRVDP3, as it automatically accounts for the viewing distance.

### 8.4.3 Metrics validation

To validate the metrics, I performed 5-fold cross-validation. The goal was to find a mapping between the quality score predicted by the metrics and the VLTs for $P_{det} = 0.25$. I experimented with different threshold $P_{det}$ values and obtained similar results (not reported here).

I assume that a metric prediction can be mapped to the predicted probability of detection $\tilde{P}_{det}$ by a monotonic function $f$:

$$\tilde{P}_{det} = f(M(T_l, R)) \,, \tag{8.2}$$

where $M$ is the quality metric, $R$ is the reference image and $T_l$ is the test image encoded at the quality level $l$. Therefore, the predicted VLT $\tilde{l}$ can be found as:

$$\tilde{l} = \underset{l}{\operatorname{argmin}} \, ||f(M(T_l, R)) - 0.25||_2 \quad \wedge \quad l \in 2, 4, .., 100 \,. \tag{8.3}$$

Since I am interested in the single value of VLT rather than finding the function $f$, I instead estimate the VLT using a single (per metric) value $q_{VLT}$, corresponding to the metric objective quality at the true VLT level:

$$\tilde{l} = \underset{l}{\operatorname{argmin}} \, ||M(T_l, R) - q_{VLT}||_2 \quad \wedge \quad l \in 2, 4, .., 100 \,, \tag{8.4}$$

which I optimize per metric so that $||\tilde{l} - l||_2$ is minimized (where $l$ is the measured VLT). The individual $q_{VLT}$ was found for each fold so that the results are reported for the testing set.

Since HDRVDP3V and DPVM produce a map of detection probabilities, rather than a single quality value, I consider a percentile value from the probability map to be a prediction. For DPVM, similar to [141], I search for the optimal percentile that minimizes root-mean-squared-error (RMSE) between the predicted and measured VLT. The best percentile for DPVM and HDRVDP3V were 86 and 97 respectively.

### 8.4.4 Results and discussion

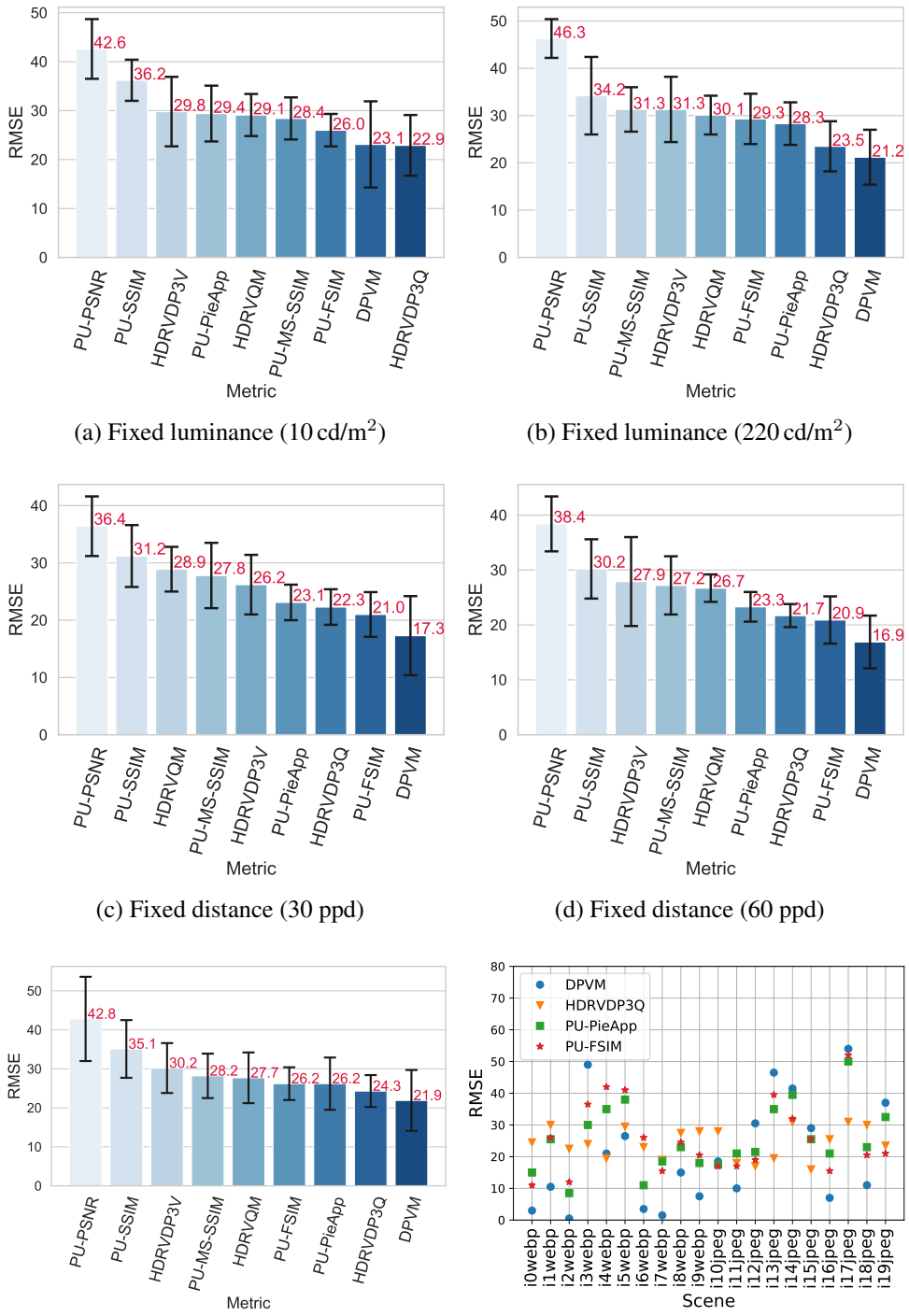
I first explore how well each of the metrics can account for the viewing distance and the display brightness. The metrics that use PU transform are prefixed with PU-. For each experiment I fix one of the viewing conditions and let the metric predict the VLT, while accounting for another viewing condition (e.g., how well the metric can account for the viewing distance at a fixed luminance level of 10 cd/m$^2$).

The results of the 5-fold cross-validation are shown in Figure 8.5. Metrics, which were designed to account for the viewing distance (HDRVDP3Q and DPVM) match the experimental VLT better across different viewing distances compared to the metrics that used a simple resampling.

When viewing distance is fixed and metrics are trained to account for the display brightness, DPVM performs the best among the tested metrics. It is followed by three metrics with comparable performance: PU-FSIM, HDRVDP3Q and PU-PieApp. In general, all metrics better account for the display luminance changes, compared to changes in viewing distance. In all experiments the visibility predictor HDRVDP3V performed unexpectedly worse than HDRVDP3Q. This implies that HDRVDP3V requires fine-tuning for more accurate perfomance.

Figure 8.5e shows the results of the 5-fold cross-validation for all viewing conditions, i.e., each metric needed to account for both display brightness and viewing distance. The results show that DPVM is the best performing metric, followed by five metrics of comparable performance (verified in the paired ttest): HDRVDP3Q, PU-PieApp, PU-FSIM, HDRVQM and PU-MS-SSIM. The visibility predictor HDRVDP3V performed worse than its quality counterpart. It is also worth noting a relatively good performance of PU-FSIM, as this is a hand-crafted metric that has not been trained on this task. The RMSE of the DPVM is 21.9, which is slightly larger, than the average variation of the VLT across the population – 15.9 (refer to Figure 8.2). This difference can be acceptable, and the metric should be robust enough to adaptively encode images displayed at different luminance levels and at a different viewing distance in a visually lossless manner.

Figure 8.5f shows the RMSE per image for the four best performing metrics. It is worth noting that while DPVM resulted in substantially smaller RMSE for some images, it was the worst performing metric for others. The performance of the DPVM did not correlate with the folds. The hand-crafted metrics with a few trainable parameters, HDRVDP3Q and PU-FSIM, tend to be more consistent and vary less in RMSE than machine-leaning based metrics, PI-PieApp and DPVM. Most metrics (except HDRVDP3Q) showed a very large error for images with large smooth gradient areas (e.g., sky), *i3webp,i4webp,i5webp,i13jpeg, i14jpeg and i17jpeg*, suggesting that those metrics could be worse at modeling contrast masking.




(a) Fixed luminance (10 cd/m$^2$)

(b) Fixed luminance (220 cd/m$^2$)

(c) Fixed distance (30 ppd)

(d) Fixed distance (60 ppd)

(e)

(f)

Figure 8.5: (a-d) Five-fold cross-validation results for the compared metrics under varied viewing conditions. I fix one the viewing conditions and make the metric account for the remaining condition. In the first row I isolate distance as the component impacting the VLT and test whether metrics can take it into account under fixed display luminance. Similarly, in the second row, I fix the distance, and let the metrics account for the display luminance. (e) Five-fold cross-validation results for the compared metrics. Error bars show the standard deviation. (f) RMSE per image.

## 8.5 Summary

I analyzed the carefully collected by Nanayng Ye novel dataset for visually lossless image compression under varying display brightness and viewing distance conditions and compared the performance of the state-of-the-art image quality and visibility metrics on this dataset. The results indicate that the display brightness and the viewing distance can significantly affect the compression level required for visually lossless coding. I found that, recently proposed deep photometric visibility metric (DPVM), although, is the best at matching the experimental thresholds for the collected dataset, its' results have a substantial room for improvement.

# Chapter 9

# Conclusion

With the plethora of modern display types, a metric capable of predicting image quality under varied viewing conditions is a necessity. Such a metric can enable new and exciting applications. An ideal metric: (i) has a high correlation with the quality as perceived by human observers; (ii) is content-driven; (iii) differentiable; (iv) is viewing condition dependent; and (v) has a meaningful scale. Perfect candidates for an ideal metric are data-driven metrics requiring large amounts of versatile training data. However, for IQA such data are expensive and time-consuming to collect. In my work I focused on the methods facilitating data collection, such as active sampling, scaling and merging image quality assessment data. With the help of these methods I collected the largest photometric image quality dataset to date and showed that it is sufficient to train deep image quality assessment metrics. The utility of the dataset and metrics is then shown on visually lossless image compression under varied image brightness and distance.

## 9.1 Contributions

In my work I have explored the ways of efficient collection and aggregation for accurate subjective image quality datasets, I have demonstrated that the accuracy of scores in existing datasets often can be improved. I have also shown that existing methods for data collection can be substantially refined and proposed a new method to collect subjective datasets more efficiently. I also demonstrated how to merge existing datasets to ameliorate the diversity of the content and ensure that they are large enough for training deep image quality metrics. I have also shown that many existing image quality datasets do not account for the display brightness. To remedy this limitation, I have collected a new dataset. Results in my dissertation have been achieved through:

**Scaling of one of the largest datasets with pairwise comparisons** I have presented in Chapter 3 scaling of the large scale image quality assessment dataset (TID2013 [106]). I showed that psychometric scaling produces more accurate results than vote counts in a simulated experiment, especially as the number of conditions in the experiment increases. I also demonstrated that

the additional set of comparisons and psychometric scaling improve the consistency of quality scores of the TID2013. I validated that the assumptions of Thurstone Case V are sufficient for modeling image quality and did not find sufficient evidence for using Thurstone case III.

**Active-sampling procedure for pairwise comparisons** A new state-of-the-art active sampling algorithm was presented in Chapter 4. Commonly used sorting methods perform poorly compared to the state-of-the-art methods based on the EIG, and even EIG-based methods are sub-optimal, as they rely on a partial update of the posterior distribution. ASAP computes the full posterior distribution, which is crucial to achieving accurate EIG estimates, and thus the accuracy of active sampling. Fast computation of the posterior, important for real-time applications, was made possible by using a fast and accurate factor graph approach, which is new to the active sampling community. Besides, ASAP only computes the EIG for the most informative pairs, reducing the computational cost of ASAP by up to 80%, and selects batches using a minimum spanning tree method, allowing to avoid unbalanced designs.

**Scaling and data merging for datasets with mean opinion scores and pairwise comparisons** In Chapter 5, I proposed a probabilistic model that can bring pairwise comparison and rating experiments into a unified quality scale. The units in that scale, are scaled accordingly to the combined inter- and intra-observer variations so that 1 unit corresponds to 75% of observers selecting one condition over another (JOD units). The model can also estimate observer variation for each experimental protocol.

**Large scale unified photometric image quality dataset and photometric metrics** In Chapter 6, I performed a series of subjective image quality assessment experiments and constructed the largest subjective HDR IQA dataset to date (UPIQ) using psychometric scaling from Chapter 5. The dataset contains *3779* SDR and *380* HDR images from four existing IQA datasets. I showed the necessity and advantages of the psychometric scaling by comparing it to other strategies for merging datasets. In Chapter 7 I used the new dataset to retrain and benchmark existing HDR metrics. I showed that the proposed dataset is sufficiently large for deep architectures by training a CNN-based full-reference *photometric* image quality metric. The advantage of training on the unified dataset is shown in comparison with training on a single dataset and performing multi-task learning on disjoint datasets. The utility of training HDR metrics on the new dataset is shown in the application to brightness-adaptive image compression.

**Dataset and metrics for visually lossless image compression under varied viewing distance and display brightness** I have analyzed a new dataset for viewing distance and display brightness dependent visually lossless compression in Chapter 8. Along with the dataset I have presented the analysis and validation of the state-of-the-art metrics for finding visually lossless threshold.

## 9.2 Future work

Although a significant amount of work has been done, this dissertation can be extended in several ways. First, the dataset merging procedure presented in Chapter 5 presents a maximum likelihood solution. One of the drawbacks of such a model is inability to predict the distribution, rather than point estimates. Making the model Bayesian can greatly improve the utility. As such an active sampling procedure based on such a model would be very useful for many applications. Secondly, the presented dataset, UPIQ, could be substantially improved by including the distance dimension, impacting the quality. A larger dataset accounting for viewing distance could enable the deep metrics also take into account the distance to the display. Finally, the trained metrics, could be verified as a perceptual loss for image restoration and enhancement algorithms.

# Bibliography


[1] Mikhailiuk A., Wilmot C., Perez-Ortiz M., and Mantiuk R. Asap: Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *International Conference on Pattern Recognition (ICPR)*, 2020.

[2] Seyed Ali Amirshahi, Marius Pedersen, and Stella X Yu. Image quality assessment by comparing CNN features between images. *Journal of Imaging Science and Technology*, 60(6), 2016. doi: http://dx.doi.org/10.2352/J.ImagingSci.Technol.2016.60.6.060410.

[3] Tunc O. Aydin, Rafal Mantiuk, and Hans-Peter Seidel. Extending quality metrics to full luminance range images. In *Human Vision and Electronic Imaging*, pages 68060B–10. Spie, 2008. doi: 10.1117/12.765095. URL `http://link.aip.org/link/PSISDG/v6806/i1/p68060B/s1{&}Agg=doi`.

[4] TunÇ Ozan Aydın, Rafał Mantiuk, and Hans-Peter Seidel. Extending quality metrics to full dynamic range images. In *Human Vision and Electronic Imaging XIII*, Proceedings of SPIE, pages 6806–10, San Jose, USA, January 2008.

[5] Maryam Azimi, Amin Banitalebi Dehkordi, Yuanyuan Dong, Mahsa Pourazad, and Panos Nasiopoulos. Evaluating the performance of existing full-reference quality metrics on high dynamic range (HDR) video content. *International Conference on Multimedia Signal Processing*, 03 2018.

[6] Peter Barten. *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*. SPIE Press, 1999.

[7] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12(2):355–362, Feb 2018. ISSN 1863-1711. doi: 10.1007/s11760-017-1166-8. URL `https://doi.org/10.1007/s11760-017-1166-8`.

[8] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Alan Bovik. *Handbook of Image and Video Processing*. Academic Press, 2010.

[10] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029.

[11] ITU-R BT.500-13. Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union*, 2012.

[12] P. Burt and E. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

[13] Manuela Cattelan. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, 27(3):412–433, 08 2012. doi: 10.1214/12-STS396.

[14] Damon Chandler. Seven challenges in image quality assessment: Past, present, and future research. *ISRN Signal Processing*, 2013, 02 2013. doi: 10.1155/2013/905685.

[15] Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 193–202, 2013.

[16] Xi Chen, Kevin Jiao, and Qihang Lin. Bayesian decision process for cost-efficient dynamic ranking via crowdsourcing. *Journal of Machine Learning Research*, 17(216):1–40, 2016.

[17] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto. Deep convolutional autoencoder-based lossy image compression. In *2018 Picture Coding Symposium (PCS)*, pages 253–257, 2018.

[18] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009. ISBN 0262033844, 9780262033848.

[19] Robert A. Cribbie and H. J. Keselman. Pairwise multiple comparisons: A model comparison approach versus stepwise procedures. *British Journal of Mathematical and Statistical Psychology*, 56(1):167–182, 2003. doi: 10.1348/000711003321645412.

[20] László Csató. Ranking by pairwise comparisons for swiss-system tournaments. *Central European Journal of Operations Research*, 21(4):783–803, 2013.

[21] Joyce E. Farrell D. Amnon Silverstein. Efficient method for paired comparison. *Journal of Electronic Imaging*, 10:10 – 10 – 5, 2001. doi: 10.1117/1.1344187. URL `https://doi.org/10.1117/1.1344187`.

[22] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, Aug 2007. doi: 10.1109/TIP.2007.901238.

[23] Jean-Charles de Borda. *Memoire sur les elections au scrutin, Memoire de l'Académie Royale. Histoire de l'Academie des Sciences*. L'Academie des Sciences, Paris, 1781.

[24] Gyorgy Denes, Akshay Jindal, Aliaksei Mikhailiuk, and Rafal K. Mantiuk. A perceptual model of motion quality for rendering with adaptive refresh-rate and resolution. *SIGGRAPH*, 2020.

[25] Y. Deng, C. C. Loy, and X. Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, July 2017. ISSN 1053-5888. doi: 10.1109/MSP.2017.2696576.

[26] P. Dunn-Rankin. *Scaling methods*. L. Erlbaum, 1983.

[27] EBU. SAMVIQ - subjective assessment methodology for video quality. Technical report, European Broadcasting Union, 2003. BPN 056.

[28] Peter G. Engeldrum. *Psychometric scaling: a toolkit for imaging systems development*. Imcotek Press, 2000.

[29] C. Fan, H. Lin, V. Hosu, Y. Zhang, Q. Jiang, R. Hamzaoui, and D. Saupe. Sur-net: Predicting the satisfied user ratio curve for image compression with deep learning. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, June 2019. doi: 10.1109/QoMEX.2019.8743204.

[30] R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919. doi: 10.1017/S0080456800012163.

[31] Fei Gao, Yi Wang, Panpeng Li, Min Tan, Jun Yu, and Yani Zhu. Deepsim: Deep similarity for image quality assessment. *Neurocomputing*, 257:104 – 114, 2017. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2017.01.054. Machine Learning and Signal Processing for Big Multimedia Analysis.

[32] D. Ghadiyaram and A. C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, Jan 2016. ISSN 1057-7149. doi: 10.1109/TIP.2015.2500021.

[33] Mark E. Glickman. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):377–394, 1999. doi: 10.1111/1467-9876.00159.

[34] Mark E. Glickman and Shane T. Jensen. Adaptive paired comparison design. *Journal of Statistical Planning and Inference*, 127(1):279 – 293, 2005. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2003.09.022.

[35] Rolf R. Hainich. *Perceptual display calibration*. CRC Press, 2016.

[36] P. Hanhart, L. Krasula, P. Le Callet, and T. Ebrahimi. How to benchmark objective quality metrics from paired comparison data? In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2016. doi: 10.1109/QoMEX.2016.7498960.

[37] Philippe Hanhart, Marco V. Bernardo, Manuela Pereira, António M. G. Pinheiro, and Touradj Ebrahimi. Benchmarking of objective quality metrics for HDR image quality assessment. *EURASIP Journal on Image and Video Processing*, 2015(1):39, Dec 2015. ISSN 1687-5281. doi: 10.1186/s13640-015-0091-4. URL https://doi.org/10.1186/s13640-015-0091-4.

[38] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[39] Reinhard Heckel, Max Simchowitz, Kannan Ramchandran, and Martin J Wainwright. Approximate ranking from pairwise comparisons. *arXiv preprint arXiv:1801.01253*, 2018.

[40] Reinhard Heckel, Nihar B. Shah, Kannan Ramchandran, and Martin J. Wainwright. Active ranking from pairwise comparisons and when parametric assumptions do not help. *Ann. Statist.*, 47(6):3099–3126, 12 2019. doi: 10.1214/18-AOS1772.

[41] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: A bayesian skill rating system. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 569–576. MIT Press, 2007.

[42] C. A. R. Hoare. Quicksort. *The Computer Journal*, 5(1):10–16, 01 1962. ISSN 0010-4620. doi: 10.1093/comjnl/5.1.10.

[43] Bernd Hoefflinger. *High-Dynamic-Range (HDR) Vision*. Springler, 2007.

[44] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. ISSN 1941-0042. doi: 10.1109/tip.2020.2967829. URL http://dx.doi.org/10.1109/TIP.2020.2967829.

[45] A.J. Hussain, Ali Al-Fayadh, and Naeem Radi. Image compression techniques: A survey in lossless and lossy algorithms. *Neurocomputing*, 300:44 – 69, 2018. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2018.02.094. URL http://www.sciencedirect.com/science/article/pii/S0925231218302935.

[46] ITU-R. Methodology for the subjective assessment of the quality of television pictures. ITU-R Recommendation BT.500-13, Jan 2012.

[47] ITU-R. Parameter values for the hdtv standards for production and international programme exchange. ITU-R Recommendation BT.709-6, Mar 2015.

[48] ITU-R. Subjective assessment methods for 3d video quality. ITU-R Recommendation P.915, Mar 2016.

[49] ITU-T. Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. ITU-T Recommendation P.1401, Jul 2012.

[50] Frank Jakel and Felix A. Wichmann. Spatial four-alternative forced-choice method is the preferred psychophysical method for naive observers. *Journal of Vision*, 6(11):13–13, 11 2006. ISSN 1534-7362. doi: 10.1167/6.11.13. URL `https://doi.org/10.1167/6.11.13`.

[51] Kevin G. Jamieson and Robert D. Nowak. Active ranking using pairwise comparisons. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 2240–2248, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

[52] Lucjan Janowski and Margaret Pinson. The accuracy of subjects in a quality experiment: A theoretical subject model. *IEEE Transactions on Multimedia*, 17(12):2210–2224, Dec 2015. ISSN 1520-9210. doi: 10.1109/TMM.2015.2484963.

[53] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik. Objective quality assessment of multiply distorted images. In *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 1693–1697, 2012. doi: 10.1109/ACSSC.2012.6489321.

[54] S. Jia, Y. Zhang, D. Agrafiotis, and D. Bull. Blind high dynamic range image quality assessment using deep learning. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 765–769, 2017. doi: 10.1109/ICIP.2017.8296384.

[55] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017. doi: 10.1109/TIP.2017.2713099.

[56] Lina Jin, Joe Yuchieh Lin, Sudeng Hu, Haiqiang Wang, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C.-C. Jay Kuo. Statistical study on perceived JPEG image quality via MCL-JCI dataset construction and analysis. *Electronic Imaging*,

pages 1–9, 2016. ISSN 2470-1173. doi: doi:10.2352/ISSN.2470-1173.2016.13.IQSP-222. URL https://www.ingentaconnect.com/content/ist/ei/2016/00002016/00000013/art00026.

[57] Ronald G. Kaptein, Andre Kuijsters, Marc T. M. Lambooij, Wijnand A. IJsselsteijn, and Ingrid Heynderickx. Performance evaluation of 3D-TV systems. *Proc.SPIE*, 6808:6808 – 6808 – 11, 2008. doi: 10.1117/12.770082. URL https://doi.org/10.1117/12.770082.

[58] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 06 1938. ISSN 0006-3444. doi: 10.1093/biomet/30.1-2.81.

[59] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Processing Magazine*, 34(6):130–141, Nov 2017. ISSN 1053-5888. doi: 10.1109/MSP.2017.2736018.

[60] Jangyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. *Computer Vision and Pattern Recognition (CVPR)*, pages 1969–1977, 2017. doi: 10.1109/CVPR.2017.213.

[61] Y. Kim, W. Kim, and K. Shim. Latent ranking analysis using pairwise comparisons. In *2014 IEEE International Conference on Data Mining*, pages 869–874, Dec 2014. doi: 10.1109/ICDM.2014.77.

[62] P. Korshunov, P. Hanhart, T. Richter, A. Artusi, R. Mantiuk, and T. Ebrahimi. Subjective quality assessment database of HDR images compressed with JPEG XT. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 1–6, May 2015. doi: 10.1109/QoMEX.2015.7148119.

[63] 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)*, pages 1–6, 2016. doi: 10.1109/QoMEX.2016.7498936.

[64] L. Krasula, P. Le Callet, K. Fliegel, and M. Klíma. Quality assessment of sharpened images: Challenges, methodology, and objective metrics. *IEEE Transactions on Image Processing*, 26(3):1496–1508, 2017. doi: 10.1109/TIP.2017.2651374.

[65] Lukáš Krasula, Yoann Baveye, and Patrick Le Callet. Training objective image and video quality estimators using multiple databases. *Trans. Multi.*, 22(4):961–969, April 2020. ISSN 1520-9210. doi: 10.1109/TMM.2019.2935687. URL https://doi.org/10.1109/TMM.2019.2935687.

[66] F. R. Kschischang, B. J. Frey, and H. . Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[67] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22 (1):79–86, 03 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214/aoms/1177729694.

[68] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans. No-reference quality assessment of tone-mapped HDR pictures. *IEEE Transactions on Image Processing*, 26(6):2957–2971, 2017.

[69] Valero Laparra, Alexander Berardino, Johannes Ballé, and Eero P. Simoncelli. Perceptually optimized image rendering. *J. Opt. Soc. Am. A*, 34(9):1511–1525, Sep 2017. doi: 10.1364/JOSAA.34.001511. URL http://josaa.osa.org/abstract.cfm?URI=josaa-34-9-1511.

[70] Jing Li, Rafal Mantiuk, Junle Wang, Suiyi Ling, and Patrick Le Callet. Hybrid-mst: A hybrid active sampling strategy for pairwise preference aggregation. *NIPS, 31st Conference on Neural Information Processing Systems*, 2018. URL https://nips.cc/Conferences/2018/Schedule?showEvent=11349.

[71] Jing Li, Suiyi Ling, Junle Wang, and Patrick Le Callet. A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 3339–3347, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413619. URL https://doi.org/10.1145/3394171.3413619.

[72] Y. Li, L. M. Po, L. Feng, and F. Yuan. No-reference image quality assessment with deep convolutional neural networks. In *2016 IEEE International Conference on Digital Signal Processing (DSP)*, pages 685–689, Oct 2016. doi: 10.1109/ICDSP.2016.7868646.

[73] Z. Li and C. G. Bampis. Recover subjective quality scores from noisy measurements. In *2017 Data Compression Conference (DCC)*, pages 52–61, 2017. doi: 10.1109/DCC.2017.26.

[74] Zhi Li, Christos G. Bampis, Lucjan Janowski, and Ioannis Katsavounidis. A simple model for subject behavior in subjective experiments, 2020.

[75] Hanhe Lin, Vlad Hosu, Chunling Fan, Yun Zhang, Yuchen Mu, Raouf Hamzaoui, and Dietmar Saupe. Sur-featnet: Predicting the satisfied user ratio curve for image compression

with deep feature learning. *Quality and User Experience*, 5(1), May 2020. ISSN 2366-0147. doi: 10.1007/s41233-020-00034-1. URL http://dx.doi.org/10.1007/s41233-020-00034-1.

[76] H. Liu, Y. Zhang, H. Zhang, C. Fan, S. Kwong, C. . J. Kuo, and X. Fan. Deep learning-based picture-wise just noticeable distortion prediction model for image compression. *IEEE Transactions on Image Processing*, 29:641–656, 2020. ISSN 1941-0042. doi: 10.1109/TIP.2019.2933743.

[77] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. RankIQA: Learning from rankings for no-reference image quality assessment. *CoRR*, abs/1707.08347, 2017. URL http://arxiv.org/abs/1707.08347.

[78] Xiaohua Liu, Zihao Chen, Xu Wang, Jianmin Jiang, and Sam Kowng. JND-Pano: database for just noticeable difference of JPEG compressed panoramic images. In Richang Hong, Wen-Huang Cheng, Toshihiko Yamasaki, Meng Wang, and Chong-Wah Ngo, editors, *Advances in Multimedia Information Processing – PCM 2018*, pages 458–468. Springer International Publishing, 2018.

[79] Vladimir Lukin. Performance analysis of visually lossless image compression. *Conference on Video Processing and Quality Metrics for Consumer Electronics*, 01 2012.

[80] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, March 2018. ISSN 1057-7149.

[81] Siwei Ma, Xinfeng Zhang, Chuanmin Jia, Zhenghui Zhao, Shiqi Wang, and Shanshe Wanga. Image and video compression with neural networks: A review. *IEEE Transactions on Circuits and Systems for Video Technology*, PP:1–1, 04 2019. doi: 10.1109/TCSVT.2019.2910119.

[82] Rafal Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4):40:1–40:14, July 2011. ISSN 0730-0301. doi: 10.1145/2010324.1964935. URL http://doi.acm.org/10.1145/2010324.1964935.

[83] Rafal Mantiuk, Karol Myszkowski, and Seidel Hans-Peter. High dynamic range imaging. *Wiley Encyclopedia of Electrical and Electronics Engineering*, 2015.

[84] Rafał K. Mantiuk, Anna Tomaszewska, and Radosław Mantiuk. Comparison of four subjective methods for image quality assessment. *Computer Graphics Forum*, 31(8):2478–2491, 2012.

[85] Lucas Maystre and Matthias Grossglauser. Just sort it! A simple and effective approach to active preference learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2344–2353. PMLR, 06–11 Aug 2017.

[86] John McIver and Edward Carmines. *Unidimensional Scaling*. Thousand Oaks, California, New York, NY, USA, 1981. doi: 10.4135/9781412986441.

[87] Saad Michele, Patrick Le Callet, and Philip Corriveau. Blind image quality assessment: Unanswered questions and future directions in the light of consumers needs. *The Video Quality Experts Group (VQEG) eLetter*, 1:62–66, 12 2014.

[88] Aliaksei Mikhailiuk, María Pérez-Ortiz, and Rafał K. Mantiuk. Psychometric scaling of TID2013 dataset. *International Conference on Quiality of Multimedia Experience (QoMEX)*, 2018.

[89] Aliaksei Mikhailiuk, Ye Nanyang, and Rafał K. Mantiuk. The effect of display brightness and viewing distance: a dataset for visually lossless image compression. *Human Vision and Electronic Imaging (under review)*, 2020.

[90] Aliaksei Mikhailiuk, Maria Perez-Ortiz, Dingcheng Yue, Wilson Suen, and Rafal K. Mantiuk. Consolidated dataset and metrics for high-dynamic-range. *Image Quality IEEE Transactions on Computational Imaging, (under review)*, 2020.

[91] Thomas P. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2001. AAI0803033.

[92] Tom Minka, Ryan Cleven, and Yordan Zaykov. Trueskill 2: An improved bayesian skill rating system. Technical Report MSR-TR-2018-8, Microsoft, March 2018.

[93] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, Dec 2012. ISSN 1057-7149. doi: 10.1109/TIP.2012.2214050.

[94] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20:209–212, 2013.

[95] Hassan Momtaz and Mohammad Daliri. Predicting the eye fixation locations in the gray scale images in the visual scenes with different semantic contents. *Cognitive Neurodynamics*, pages 31–47, 2016.

[96] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 1 edition, 2012.

[97] Manish Narwaria, Matthieu P. Da Silva, Patrick Le Callet, and Romuald Pepion. Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality. *Optical Engineering*, 52(10), 2013. doi: 10.1117/1.OE.52.10.102008.

[98] Manish Narwaria, Rafal K. Mantiuk, Mattheiu Perreira Da Silva, and Patrick Le Callet. HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images. *Journal of Electronic Imaging*, 24(1):010501, jan 2015. ISSN 1017-9909. doi: 10.1117/1.JEI.24.1.010501.

[99] Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet. HDR-VQM: An objective quality measure for high dynamic range video. *Signal Processing: Image Communication*, 35:46 – 60, 2015. ISSN 0923-5965. doi: https://doi.org/10.1016/j.image.2015.04.009.

[100] Maria Perez-Ortiz and Rafal K. Mantiuk. A practical guide and software for analysing pairwise comparison experiments. *CoRR*, 2017.

[101] Thomas Pfeiffer, Xi Alice Gao, Yiling Chen, Andrew Mao, and David G Rand. Adaptive polling for information aggregation. In *The 26th Conference on Artificial Intelligence (AAAI'12)*, 2012.

[102] Margaret Pinson and Stephen Wolf. Techniques for evaluating objective video quality models using overlapping subjective data sets. Technical report, US Department of Commerce, National Telecommunications and Information Administration, 2008. NTIA Technical Report TR-09-457.

[103] Margaret H Pinson and Stephen Wolf. Comparing subjective video quality testing methodologies. In *Visual Communications and Image Processing (VCIP)*, pages 573–582. International Society for Optics and Photonics, 2003.

[104] Yohann Pitrey, Ulrich Engelke, Marcus Barkowsky, Romuald Pépion, and Patrick Le Callet. Aligning subjective tests using a low cost common set. In *Euro ITV*, 2011.

[105] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti. TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009.

[106] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, and Benoit. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57 – 77, 2015.

[107] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[108] M. Pérez-Ortiz, A. Mikhailiuk, E. Zerman, V. Hulusic, G. Valenzise, and R. K. Mantiuk. From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing*, pages 1–1, 2019. doi: 10.1109/TIP.2019.2936103.

[109] A. Raid, Wael Khedr, Mohamed El-dosuky, and Wesam Ahmed. Jpeg image compression using discrete cosine transform - a survey. *International Journal of Computer Science and Engineering Survey*, 5, 05 2014. doi: 10.5121/ijcses.2014.5204.

[110] Subramanian Ramanathan, Harish Katti, Nicu Sebe, Mohan Kankanhalli, and Tat-Seng Chua. An eye fixation database for saliency detection in images. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 30–43, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[111] David M. Rouse, Romuald Pépion, Patrick Le Callet, and Sheila S. Hemami. Tradeoffs in subjective testing methods for image and video quality assessment. In *IS&T/SPIE Electronic Imaging, Human Vision and Electronic Imaging XV*, pages 75270F–1–75270F–11. International Society for Optics and Photonics, 2010. doi: 10.1117/12.845389.

[112] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin–Madison, 2010.

[113] Nihar B. Shah and Martin J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 18(199):1–38, 2018. URL `http://jmlr.org/papers/v18/16-206.html`.

[114] Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17 (58):1–47, 2016. URL `http://jmlr.org/papers/v17/15-189.html`.

[115] H.R. Sheikh, M.F. Sabir, and A.C. Bovik. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Transactions on Image Processing*, 15(11): 3440–3451, 2006. ISSN 1057-7149. doi: 10.1109/TIP.2006.881959.

[116] S Shlaer. The relation between visual acuity and illumination. *The Journal of general physiology*, 21(2):165—188, November 1937. ISSN 0022-1295. doi: 10.1085/jgp.21.2.165. URL `https://europepmc.org/articles/PMC2141937`.

[117] Zhanjun Si and Ke Shen. *Research on the WebP Image Format*, pages 271–277. Springer, 12 2016. ISBN 978-981-10-0070-6. doi: 10.1007/978-981-10-0072-0_35.

[118] Lea Skorin-Kapov, Martín Varela, Tobias Hoßfeld, and Kuan-Ta Chen. A survey of emerging concepts and challenges for QoE management of multimedia services. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14 (2s):29:1–29:29, 2018.

[119] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.*, 19(2):279–281, 06 1948. doi: 10.1214/aoms/1177730256. URL `https://doi.org/10.1214/aoms/1177730256`.

[120] G. J. Sullivan and T. Wiegand. Rate-distortion optimization for video compression. *IEEE Signal Processing Magazine*, 15(6):74–90, 1998.

[121] Baiundefinedzs Szörényi, Róbert Busa-Fekete, Adil Paul, and Eyke Hüllermeier. Online rank elicitation for plackett-luce: A dueling bandits approach. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 604–612, Cambridge, MA, USA, 2015. MIT Press.

[122] H. Talebi and P. Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

[123] L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927. ISSN 0033-295X. doi: 10.1037/h0070288.

[124] Toshiko Tominaga, Takanori Hayashi, Jun Okamoto, and Akira Takahashi. Performance comparisons of subjective quality assessment methods for mobile video. In *2nd International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, Jun 2010. doi: 10.1109/QOMEX.2010.5517948.

[125] Kristi Tsukida and Maya R Gupta. How to Analyze Paired Comparison Data. Technical Report UWEETR-2011-0004, Department of Electrical Engineering University of Washington, 2011.

[126] Ken Turkowski. *Filters for Common Resampling Tasks*, page 147–165. Academic Press Professional, Inc., USA, 1990. ISBN 0122861695.

[127] International Telecommunication Union. Subjective video quality assessment methods for multimedia applications. ITU-T Recommendation P.910, Apr 2008.

[128] International Telecommunication Union. Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment. ITU-T Recommendation P.913, March 2016.

[129] Satoshi Usami. Individual differences multidimensional bradley-terry model using reversible jump markov chain monte carlo algorithm. *Behaviormetrika*, 37(2):135–155, 2010.

[130] Giuseppe Valenzise, Francesca De Simone, Paul Lauga, and Frederic Dufaux. Performance evaluation of objective quality metrics for HDR image compression. *Proceedings of SPIE - The International Society for Optical Engineering*, 9217, 09 2014. doi: 10.1117/12.2063032.

[131] Peter Vangorp, Rafat K Mantiuk, Bartosz Bazyluk, Karol Myszkowski, Radosław Mantiuk, Simon J Watt, and Hans-Peter Seidel. Depth from HDR: depth induction or increased realism? In *ACM Symposium on Applied Perception - SAP '14*, pages 71–78. ACM Press, 2014. ISBN 9781450330091. doi: 10.1145/2628257.2628258. URL `http://dl.acm.org/citation.cfm?doid=2628257.2628258`.

[132] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, S. S. Channappayya, and S. S. Medasani. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6, Feb 2015. doi: 10.1109/NCC.2015.7084843.

[133] T. Vigier, L. Krasula, A. Milliat, M. P. Da Silva, and P. Le Callet. Performance and robustness of hdr objective quality metrics in the context of recent compression scenarios. In *2016 Digital Media Industry Academic Forum (DMIAF)*, pages 59–64, 2016. doi: 10.1109/DMIAF.2016.7574903.

[134] Stephen Voran. An iterated nested least-squares algorithm for fitting multiple data sets. *NASA STI/Recon Technical Report N*, 10 2002.

[135] Gregory K. Wallace. The JPEG still picture compression standard. *Communications of the ACM*, pages 30–44, 1991.

[136] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003.

[137] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13 (4):600–612, April 2004. ISSN 1057-7149. doi: 10.1109/TIP.2003.819861.

[138] Andrew B. Watson and Lindsay Kreslake. Measurement of visual impairment scales for digital video. *SPIE Electronic Imaging, Human Vision and Electronic Imaging VI*, 4299: 79–89, 2001. doi: 10.1117/12.429526.

[139] Andrew B. Watson and Denis G. Pelli. Quest: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, 33(2):113–120, Mar 1983. ISSN 1532-5962. doi: 10.3758/BF03202828. URL https://doi.org/10.3758/BF03202828.

[140] K. Wolski, Ye Nan-yang, D. Giunchi, P. Didyk, K. Myszkowski, and Rafał K. Mantiuk. Locvis: Local visibility maps of artifacts and distortions in images. In *Dataset*, 2018.

[141] Krzysztof Wolski, Daniele Giunchi, Nanyang Ye, Piotr Didyk, Karol Myszkowski, Radosław Mantiuk, Hans-Peter Seidel, Anthony Steed, and Rafał K. Mantiuk. Dataset and Metrics for Predicting Local Visible Differences. *ACM Transactions on Graphics*, 37(5):1–14, nov 2018. ISSN 07300301. doi: 10.1145/3196493. URL http://dl.acm.org/citation.cfm?doid=3278329.3196493.

[142] Sophie Wuerger, Maliha Ashraf, Minjung Kim, Jasna Martinovic, Maria Perez-Ortiz, and Rafal K. Mantiuk. Spatio-chromatic contrast sensitivity under mesopic and photopic light levels. *Journal of Vision*, in review, 2019.

[143] Qianqian Xu, Tingting Jiang, Yuan Yao, Qingming Huang, Bowei Yan, and Weisi Lin. Random partial paired comparison for subjective video quality assessment via hodgerank. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 393–402. ACM, 2011.

[144] Qianqian Xu, Jiechao Xiong, Xi Chen, Qingming Huang, and Yuan Yao. Hodgerank with information maximization for crowdsourced pairwise ranking aggregation. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 4326–4334, 2018.

[145] N. Ye, K. Wolski, and R. K. Mantiuk. Predicting visible image differences under varying display brightness and viewing distance. *Conference on Computer Vision and Pattern Recognition*, 2019.

[146] Peng Ye and David Doermann. Active sampling for subjective image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4249–4256, 2014.

[147] Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 241–248, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

[148] Emin Zerman, Giuseppe Valenzise, and Frederic Dufaux. An extensive performance evaluation of full-reference HDR image quality metrics. *Quality and User Experience*, 2(1):5, Apr 2017. ISSN 2366-0147. doi: 10.1007/s41233-017-0007-4. URL https://doi.org/10.1007/s41233-017-0007-4.

[149] Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, Rafał Mantiuk, and Frédéric Dufaux. The relation between MOS and pairwise comparisons and the importance of cross-content comparisons. In *IS&T/SPIE Electronic Imaging, Human Vision and Electronic Imaging XXII*. International Society for Optics and Photonics, 2018.

[150] L. Zhang, L. Zhang, X. Mou, and D. Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, Aug 2011. ISSN 1057-7149. doi: 10.1109/TIP.2011.2109730.

[151] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep networks as a perceptual metric. In *CVPR*, 2018.

[152] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, March 2017. ISSN 2333-9403. doi: 10.1109/TCI.2016.2644865.

[153] Wang Zhou and Simoncelli Eero. Maximum differentiation competition: A methodology for comparing quantitative models of perceptual discriminability. *Journal of Vision*, 5(8), 2005. doi: 10.1167/5.8.230.

# Appendix A

# Extra Information

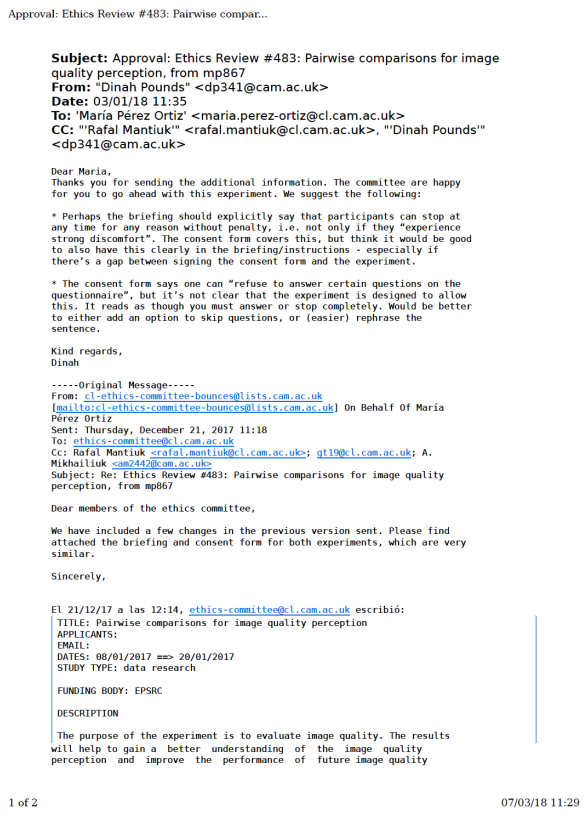
## A.1 Ethical Approvals

All experiments involving data collection with human subjects discussed in this dissertation underwent review of the ethics committee. Figure A.1 shows the ethical approval from the ethics committee of the University of Cambridge for the pairwise comparison experiment reported in Chapter 3. Ethical approval for the experiment reported in Chapter 6 is given in Figure A.2.

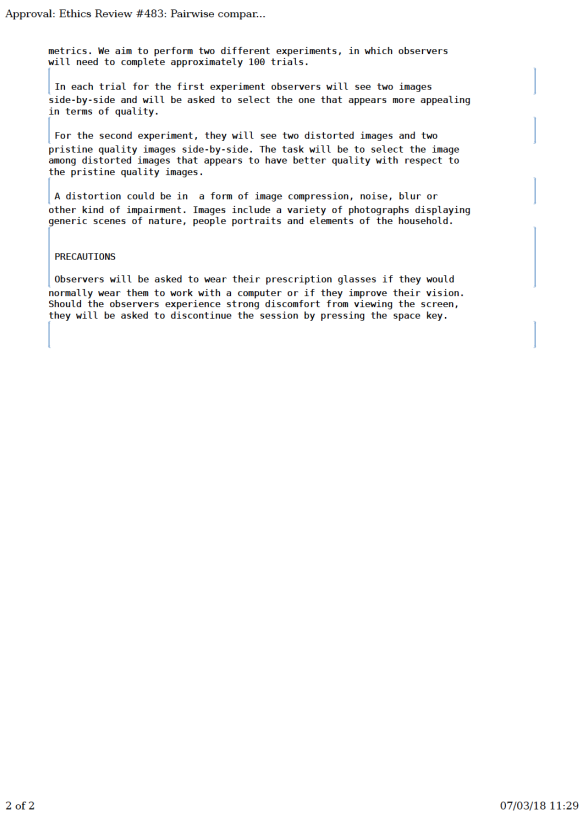
## A.2 TID 2013 rating experimental procedure

In order to obtain mean opinion scores, an experiment was conducted using the absolute category rating with hidden reference (ACR-HR) methodology [127]. In this experiment, a subset of color images from TID2013 color image dataset [106] were presented with a mid-grey background on a standard display in a dark room, following the ITU recommentations [46]. The participants were seated at the distance equal of 3 display heights ($\sim$1m). The stimuli were shown for 5 seconds and the observers were allowed to confirm their answer either during or after displaying the stimulus. The participants were then asked to rate the quality of the color image presented on the display using a continuous scale ([0,100], 100 corresponding to the best quality). ACR-HR was selected to take also the reference images and some quality enhancements (e.g. increase in the contrast for 'contrast change' distortion type). The participants spent on average $3.9\pm1.5$ seconds on viewing an image and $3.8\pm2.3$ seconds on assigning a score.

In order to avoid fatigue and to keep the experiment under 30 minutes, a subset of images was used. Two distortion types were selected for each content through random permutation of the 24 different distortion types. A total of 175 images (25 contents $\times$ 2 distortion types $\times$ 3 distortion levels + 25 original images) were voted during the experiment. Looking at the quality values provided with TID2013 color dataset, we notice that some of the distortion types (e.g. non-eccentricity pattern noise and contrast change) have different behavior compared to the other compression types. In order to capture the uncommon behavior of these distortion methods,




(a) Page 1

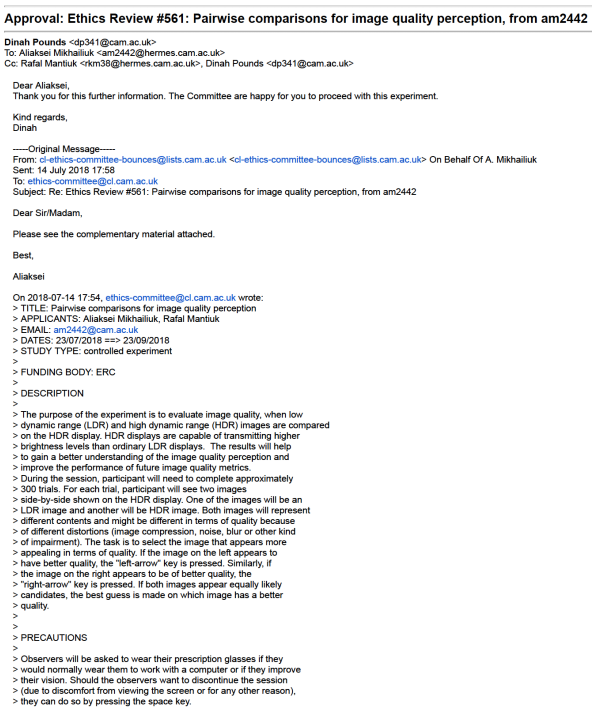



(b) Page 2

Figure A.1: Ethical approval for the pairwise comparison experiment from Chapter 3.

distortion levels of $\{2, 4, 5\}$ were used for non-eccentricity pattern noise and contrast change distortion type, as well as JPEG compression to have a more varying quality values. For the rest of the distortion types, distortion levels of $\{1, 3, 5\}$ were selected. To minimize context effects, the images were ordered randomly for each subject, and consequent images were selected from different contents.

Before the experiment, participants were screened for visual acuity and correct color vision using Snellen and Ishihara charts, respectively. A training session was conducted prior to the experiment to familiarize the subjects with the test procedure and distortion levels. Images used for training were not used in the experiment. Subjects were asked to rate "the overall quality of the presented image". In total, 22 people (4 female and 18 male) with the average age of 30.6 participated in the experiment. After outlier detection [49], 1 of the 22 subjects was removed. MOS, standard deviation, and confidence intervals are calculated for each stimulus as described in ITU-T Rec. P.1401 [49].



**Approval: Ethics Review #561: Pairwise comparisons for image quality perception, from am2442**

**Dinah Pounds** <dp341@cam.ac.uk>
To: Aliaksei Mikhailiuk <am2442@hermes.cam.ac.uk>
Cc: Rafal Mantiuk <rkm38@hermes.cam.ac.uk>, Dinah Pounds <dp341@cam.ac.uk>

Dear Aliaksei,
Thank you for this further information. The Committee are happy for you to proceed with this experiment.

Kind regards,
Dinah

-----Original Message-----
From: cl-ethics-committee-bounces@lists.cam.ac.uk <cl-ethics-committee-bounces@lists.cam.ac.uk> On Behalf Of A. Mikhailiuk
Sent: 14 July 2018 17:58
To: ethics-committee@cl.cam.ac.uk
Subject: Re: Ethics Review #561: Pairwise comparisons for image quality perception, from am2442

Dear Sir/Madam,

Please see the complementary material attached.

Best,

Aliaksei

On 2018-07-14 17:54, ethics-committee@cl.cam.ac.uk wrote:
> TITLE: Pairwise comparisons for image quality perception
> APPLICANTS: Aliaksei Mikhailiuk, Rafal Mantiuk
> EMAIL: am2442@cam.ac.uk
> DATES: 23/07/2018 ==> 23/09/2018
> STUDY TYPE: controlled experiment
>
> FUNDING BODY: ERC
>
> DESCRIPTION
>
> The purpose of the experiment is to evaluate image quality, when low
> dynamic range (LDR) and high dynamic range (HDR) images are compared
> on the HDR display. HDR displays are capable of transmitting higher
> brightness levels than ordinary LDR displays. The results will help
> to gain a better understanding of the image quality perception and
> improve the performance of future image quality metrics.
> During the session, participant will need to complete approximately
> 300 trials. For each trial, participant will see two images
> side-by-side shown on the HDR display. One of the images will be an
> LDR image and another will be HDR image. Both images will represent
> different contents and might be different in terms of quality because
> of different distortions (image compression, noise, blur or other kind
> of impairment). The task is to select the image that appears more
> appealing in terms of quality. If the image on the left appears to
> have better quality, the "left-arrow" key is pressed. Similarly, if
> the image on the right appears to be of better quality, the
> "right-arrow" key is pressed. If both images appear equally likely
> candidates, the best guess is made on which image has a better
> quality.
>
>
> PRECAUTIONS
>
> Observers will be asked to wear their prescription glasses if they
> would normally wear them to work with a computer or if they improve
> their vision. Should the observers want to discontinue the session
> (due to discomfort from viewing the screen or for any other reason),
> they can do so by pressing the space key.

Figure A.2: Ethical approval for the pairwise comparison experiment from Chapter 6.

## A.3 ASAP Pseudo-code

I provide a detailed description of the ASAP method in Algorithm 1. The algorithm requires as input a list of comparisons performed so far, $\mathbf{y}$ and a matrix with the probabilities of the EIG being computed —required for the selective evaluations —$\mathbf{Q}$. The algorithm outputs a batch of pairs to be compared $\mathbf{C}$ and an updated probability of being selected for the EIG evaluations, $\hat{\mathbf{Q}}$.

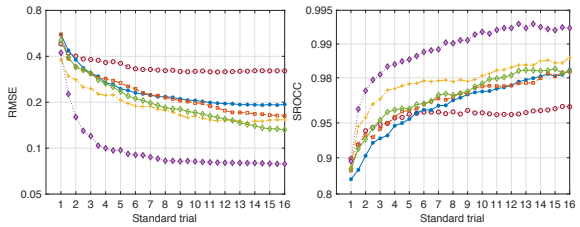
## A.4 ASAP additional results

This dataset contains 10 reference videos with 16 distortions applied to them. Each $16 \times 16$ matrix contains 3840 pairwise comparisons - each pair was compared 32 times. Figure A.3 shows the results for reference videos 3 to 10. Results for the first two reference videos are presented in the main paper.

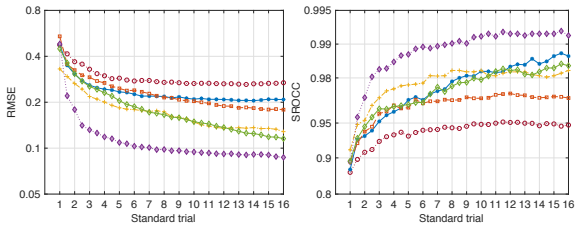
Consistent with other tests in the main paper, ASAP shows superior results to other methods. ASAP-approx. has average, similar to other EIG based methods, results. Hybrid-MST tends to perform better for small numbers of standard trials.

**Algorithm 1:** ASAP

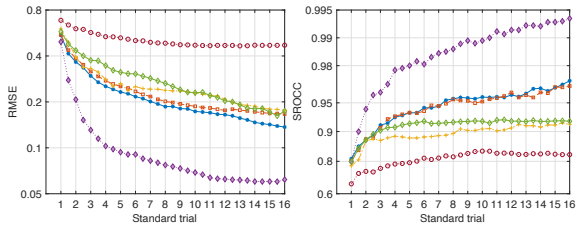
**Input**: $\mathbf{y}, \mathbf{Q}$
**Output**: $\mathbf{C}, \hat{\mathbf{Q}}$
*# Calculate the posterior for the given state of the comparison matrix*
$\boldsymbol{\mu}, \boldsymbol{\Sigma}$ = approxPosterior($\mathbf{y}$)
*# Iterate over the rows of the expected information gain matrix* $\mathbf{I}$
**for** $i \leftarrow 1$ **to** $n$ **do**
  *# Iterate over the columns of the expected information gain matrix* $\mathbf{I}$
  **for** $j \leftarrow 1$ **to** $(i-1)$ **do**
    *# Probability of selecting* $o_i$ *over* $o_j$
    $p_{ij} = P(o_i \succ o_j | r_i, r_j) = \Phi\left(\frac{\mu_i - \mu_j}{\sqrt{2}\sigma_{ij}}\right)$
    *# Selective EIG evaluations with roulette*
    **if** $Q_{ij} > U[0,1]$ **then**
      *# Posterior given all comparisons and assuming* $o_i$ *is selected over* $o_j$
      $\mu_{ij}, \Sigma_{ij}$ = approxPosterior($\mathbf{y}$; $o_i \succ o_j$)
      *# Posterior given all comparisons and assuming* $o_j$ *is selected over* $o_i$
      $\mu_{ji}, \Sigma_{ji}$ = approxPosterior($\mathbf{y}$; $o_j \succ o_i$)
      *# KL divergence between current distribution and distribution assuming the two possible outcomes*
      $KL_{ij}$ = KLDivergence($\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mathcal{N}(\mu_{ij}, \Sigma_{ij})$)
      $KL_{ji}$ = KLDivergence($\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mathcal{N}(\mu_{ji}, \Sigma_{ji})$)
      *# Weighted information gain*
      $I_{ij} = p_{ij} \times KL_{ij} + (1 - p_{ij}) \times KL_{ji}$
    **end**
    *# Update the probability of being selected for the comparison*
    $\hat{Q}_{ij} = min(p_{ij}, 1 - p_{ij})$
  **end**
  *# Scale* $q$ *per condition*
  $\hat{Q}_{ij} = \frac{\hat{Q}_{ij}}{\max_{\forall j}(\hat{Q}_{ij})}, \forall j$
**end**
*# Make the EIG matrix symmetric and find reciprocal of each entry*
$I = \mathbf{1}/(I + I^T)$
*# Create the minimum spanning tree from the matrix* $\mathbf{I}$
G = minspantree($\mathbf{I}$)
*# Nodes connected by an edge are pairs of conditions to compare*
$\mathbf{C}$ = getConnectedNodes(G)
*# Note if batch mode is not used pairs to compare are selected by* $\mathbf{C}$ = *argmax(*$I_{ij}$*)*



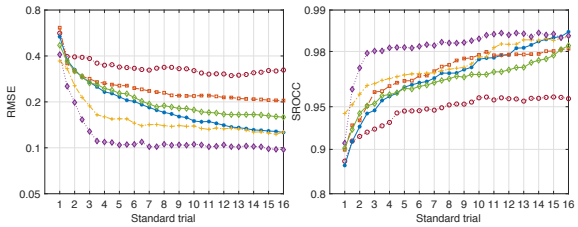

(a) Reference 3



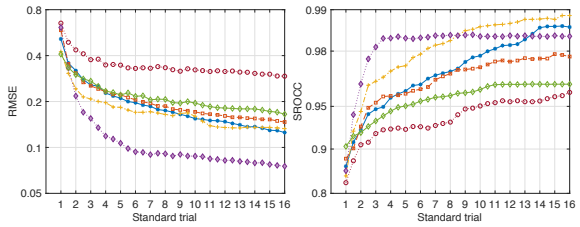

(b) Reference 4



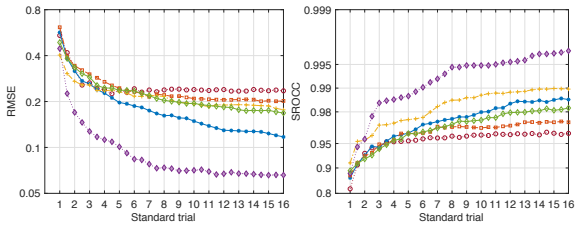

(c) Reference 5



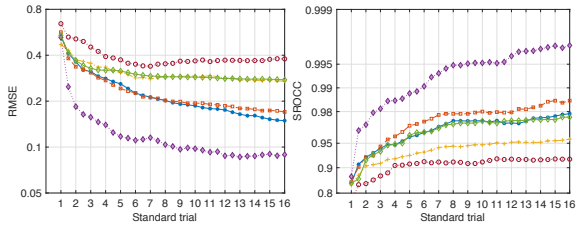

(d) Reference 6



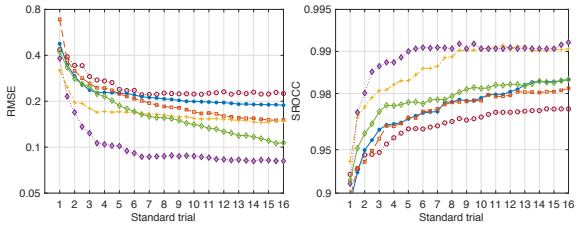

(e) Reference 7




(f) Reference 8




(g) Reference 9




(h) Reference 10

Figure A.3: Compared sampling strategies on VQA dataset.