Subjective image quality assessment with boosted triplet comparisons.

MEN, H., LIN, H., JENADELEH, M. and SAUPE, D.

2021

Subjective Image Quality Assessment With Boosted Triplet Comparisons

HUI MEN, HANHE LIN, MOHSEN JENADELEH, (Member, IEEE), AND DIETMAR SAUPE

Department of Computer and Information Science, University of Konstanz, 78464 Konstanz, Germany

Corresponding author: Hui Men (hui.3.men@uni-konstanz.de)

ABSTRACT In subjective full-reference image quality assessment, a reference image is distorted at increasing distortion levels. The differences between perceptual image qualities of the reference image and its distorted versions are evaluated, often using degradation category ratings (DCR). However, the DCR has been criticized since differences between rating categories on this ordinal scale might not be perceptually equidistant, and observers may have different understandings of the categories. Pair comparisons (PC) of distorted images, followed by Thurstonian reconstruction of scale values, overcomes these problems. In addition, PC is more sensitive than DCR, and it can provide scale values in fractional, just noticeable difference (JND) units that express a precise perceptual interpretation. Still, the comparison of images of nearly the same quality can be difficult. We introduce boosting techniques embedded in more general triplet comparisons (TC) that increase the sensitivity even more. Boosting amplifies the artefacts of distorted images, enlarges their visual representation by zooming, increases the visibility of the distortions by a flickering effect, or combines some of the above. Experimental results show the effectiveness of boosted TC for seven types of distortion (color diffusion, jitter, high sharpen, JPEG 2000 compression, lens blur, motion blur, multiplicative noise). For our study, we crowdsourced over 1.7 million responses to triplet questions. We give a detailed analysis of the data in terms of scale reconstructions, accuracy, detection rates, and sensitivity gain. Generally, boosting increases the discriminatory power and allows to reduce the number of subjective ratings without sacrificing the accuracy of the resulting relative image quality values. Our technique paves the way to fine-grained image quality datasets, allowing for more distortion levels, yet with high-quality subjective annotations. We also provide the details for Thurstonian scale reconstruction from TC and our annotated dataset, *KonFiG-IQA*, containing 10 source images, processed using 7 distortion types at 12 or even 30 levels, uniformly spaced over a span of 3 JND units.

INDEX TERMS Subjective quality assessment, full-reference, artefact amplification, zooming, flicker test, triplet comparisons, scale reconstruction, just noticeable difference.

I. INTRODUCTION

Full-reference image quality assessment (FR-IQA) quantifies the perceptual image qualities of distorted versions of pristine reference images. In addition, FR-IQA quantifies the trade-off between the bitrate and perceived quality in perceptual image compression, which helps in the optimisation of encoding parameters. Similarly, the development of other

The associate editor coordinating the review of this manuscript and approving it for publication was Mingbo Zhao.

image processing applications such as image restoration and enhancement may profit from knowing the expected perceptual quality of their output images.

Since it is not feasible to assess perceptual image quality by a subjective study each time in such applications, automated FR-IQA algorithms must be used that estimate the quality from the image data without any human interaction. To develop and train such FR-IQA algorithms, annotated image datasets, derived from subjective studies, are required. In such studies, images are judged by subjects according

to their perceived quality, either individually or in comparison with one or more other images. This paper contributes boosting methods for the presentation of the image stimuli in subjective studies that improve the accuracy and sensitivity of the perceptual measurements.

A. SUBJECTIVE FULL-REFERENCE IMAGE QUALITY ASSESSMENT

In subjective studies, test stimuli may be presented one at a time and rated according to a 5-point absolute category ratings (ACR) scale, i.e., Bad (1), Poor (2), Fair (3), Good (4), and Excellent (5). For each stimulus, the integer values of the ratings from many subjects are averaged, yielding corresponding mean opinion scores (MOS), which serve as scalar perceptual image qualities [1]. ACR is likely to lead to low sensitivity in distinguishing among stimuli of similar qualities. A modified version of ACR, degradation category rating (DCR), provides higher sensitivity [2], [3]. In a DCR test, distorted stimuli are presented with their references, either sequentially or simultaneously side by side. Stimuli are rated according to the 5-point DCR scale, namely Very Annoying (1), Annoying (2), Slightly Annoying (3), Perceptible but not Annoying (4), and Imperceptible (5). Average ratings are called degradation mean opinion scores (DMOS).

The approaches mentioned above, although straightforward, have some limitations.

(1) Observers may have different understandings of the quality categories [4], [5], leading to large variances of ratings and therefore requiring a large number of ratings to achieve the desired precision of the mean opinion scale of ACR or DCR.

(2) There is the danger of a saturation effect. If a subject scores an image in the best (or worst) quality category, another item to be judged may come up with a perceived quality that is even significantly better (or worse). Then this item can only be scored with the same category as before. There is no way to correct previously assigned quality values to accommodate the overall larger than expected dynamic range in quality.

(3) ACR and DCR scales should be regarded as ordinal, not interval scales [4]–[6], even though in practice the categories Bad, Poor, etc. are linked to the numerical values 1, 2, and so on. This means that pairs of stimuli with an equal difference in MOS, resp. DMOS are not generally perceived to have the same perceptual distance.

(4) Given the mean opinion scores s and $s + \Delta s$ of two images, there is no meaningful interpretation for the difference in perceptual quality based on s and Δs. It would be desirable to have a scaling property similar to that provided by the peak-signal-to-noise ratio (PSNR) for the case of objective image quality. For example, if two distorted images have a difference of 1 dB in PSNR, then we know that the mean-square-error in one image is $10^{0.1} \approx 1.259$ as large as that in the other one.

The pair comparison method (PC) is an alternative to ACR and DCR. In the 2-alternative forced-choice (2AFC) setting, observers are presented with pairs of test images and asked to identify the image in each pair with less distortion, i.e., the image with better quality. The PC method is an indirect measurement, and scale values cannot be generated simply by averaging ratings. Instead, an algorithm that "reconstructs" the latent quality scale values is required

Many reconstruction methods have been proposed. In psychophysics, the field that studies the relationship between physical stimuli and the perceived experiences they evoke, methods based on probabilistic models have become the de facto standard for this purpose. In Thurstonian models, the perception of each stimulus is modelled quantitatively as a scalar Gaussian random variable. The random variables corresponding to sequences of stimuli under investigation are most commonly taken to have the same variance of 0.5 so that quality differences in PC have a unit variance when the independence of the random variables is assumed. This setting defines the units of measurement. Initially, following Thurstone's pioneering work [7] in 1927, a least-squares approach was taken to solve for the reconstructed scale values based on his model. Nowadays, the method of choice is maximum likelihood estimation (MLE) [8].

The PC method overcomes all of the above-listed limitations of ACR and DCR.

(1) Such a comparison does not rely on the particular interpretation of a nominal category of quality. Therefore the task is clear and more natural than ACR and DCR.

(2) By design, the saturation effect is eliminated.

(3) The reconstruction for PC yields quality values on an interval scale. Specifically, according to the Thurstonian model, pairs of stimuli with an equal difference in scale value are perceived as having the same perceptual distance in the sense of the old, famous psychological rule of thumb: *Equally often noticed differences are equal, unless always or never noticed* [9]. Thus, for a difference of Δs on the perceptual quality scale, the proportion of subjects who consider the stimulus with the larger latent scale value to be the one with better quality is a function of Δs alone.

(4) By appropriately choosing the variance of the Gaussian distribution in the Thurstonian model, we can define the perceptual scale such that one unit difference between the two values of a pair corresponds to a fraction of 75% of the subjects choosing the correct better quality item among the pair. This corresponds to the usual definition of the just noticeable difference (JND): The JND is the perceptual quality difference for which the probability of detecting the better quality image is 50%. In PC (2AFC) for this case, therefore, half of the subjects will detect and choose the correct stimulus as the better one, while the other half has to guess and will be correct half of the time, leading to the 75% ratio.

Based on the first point above, forced-choice PCs are easier to decide in subjective trials and require less time than the

ACR or DCR categorization task. In addition, the PC method yields the lowest measurement variance and thus provides the most accurate results [10].

In a slight variation of PC, the reference stimulus is placed in the middle of the pair of stimuli to be compared. The task of the observers is to select the one that looks more (or less) similar to the reference [11]–[14]. Similar to DCR, assessing the (dis)similarity of the distorted stimuli to the reference should lead to a more appropriate, informed choice of the subject than a PC without reference. Such an approach can be seen as a special case of triplet comparisons (TC). In TC tests, three stimuli are displayed simultaneously, and the (dis)similarity to the stimulus placed in the middle, called the *pivot*, is asked to be compared. In general, the pivot can be any element of the sequence of distorted stimuli (*general TC*), and the PC with reference corresponds to TC where the pivot is fixed and equal to the reference for all comparisons. In our main experimental study, we applied the latter approach, which we call *baseline TC*. In a secondary experiment, we used general triplets and showed their potential to further increase the performance of FR-IQA compared to DCR and baseline TC.

Like PC, TC avoids the problems discussed above of ACR and DCR. The subject responses to baseline TC can be interpreted as answers to the corresponding 2AFC questions that show only the two distorted stimuli with the reference. Therefore, the same Thurstonian reconstruction may be applied. However, for general triplet comparisons with arbitrary pivot stimuli, this is not possible. Triplet comparisons had already been introduced in psychophysics by Torgerson [15] in 1958, and recently found much interest in vision science and machine learning [16]. Many scale reconstruction methods for TC have been proposed. However, with just one exception, none of them are based on the Thurstonian model allowing to give scale values in meaningful JND units as discussed above. Hence, in this paper, we propose a complete method for scale reconstruction from triplet comparisons, based on Thurstone's model and MLE, to produce scale values expressed in perceptual JND units.

B. CURRENT VISUAL QUALITY DATASETS

In addition to FR-IQA, there are other applications in which a visual reference stimulus is distorted to various levels of severity and compared. For example, in full-reference video quality assessment (FR-VQA), short image sequences of a few seconds are viewed and evaluated for visual quality. Since video data transmission is the dominant load on the Internet traffic, FR-VQA for compressed video streaming is the most relevant application of visual quality assessment methods. Recently, it has been proposed to replace the quality assessment scale based on MOS or DMOS from subjective categorical or nominal ratings by the just noticeable difference [36]. In JND assessments, the distorted reference images or videos are also compared to the reference, and the minimal distortion level that leads to a perceivable difference of the stimulus is reported. In all of these cases of visual quality

assessment, the boosting techniques that we have developed can be applied to increase the sensitivity, allowing for a more fine-grained visual analysis of the range of distortions.

In Table 1 we present an overview of the currently available datasets for subjectively assessed visual quality for FR-IQA, FR-VQA, and JND. In this paper, we contribute a new dataset, KonFiG-IQA (Konstanz Fine-Grained IQA), which is also listed in the table. Several points are noteworthy:

(1) Quality and JND assessment techniques of the current datasets are dominated by the classical approaches such as ACR, DCR, and their variations where a discrete or continuous ratio scale replaces the categorical scale. For TID2008, TID2013, and MDID, baseline triplet comparisons were carried out. For the scale reconstruction, the Swiss-system tournament style point scoring resp. a ranking procedure based on insertion sort was used. The only dataset for which a probabilistic MLE-based reconstruction from comparisons was carried out is MCL-V, without final conversion into JND units. Here, the Bradley-Terry model was used, which is very similar to the Gaussian Thurstonian model. Thus, in all current IQA and VQA datasets, possibly except for MCL-V, artefacts due to the nonlinear scaling of perceptual quality may be present.

(2) In all current IQA and VQA datasets, only a small number of distortion levels were applied, up to 6 for images and up to 11 for video. This choice corresponds to the small number of only 5 nominal quality values available in ACR and DCR. It would be desirable to introduce IQA and VQA datasets with a larger number of distortion levels, especially at the high end of quality. This would allow for training machine learning techniques for objective quality assessment aimed at applications delivering high quality imagery and streaming video over the Internet at a minimal but sufficient bitrate.

(3) Two recent trends can be observed. In 2019, the first crowdsourced FR-IQA dataset was introduced (KADID-10k), and more are likely to come, like the two sets included in this paper. Moreover, since 2016 the first datasets for JND were established both for images and video, and more of them can be expected, including crowdsourced JND datasets.

The new dataset from our study, KonFiG-IQA, stands out from the rest of the FR-IQA datasets in the following aspects:

(1) The number of distortion levels is larger and designed by perceptual consideration, namely 12, resp. 30 distortion levels, perceptually equally spaced over a range of 3 JND.

(2) The number of ratings, resp. triplet comparisons, averaged per distorted image, is much larger, with 97 per image in Part A and up to 875 in Part B. This allowed for an extensive analysis of the reliability and convergence of the resulting scale values.

(3) The scale values are derived from the probabilistic Thurstonian MLE process and converted to give

perceptually linear quality scale values in meaningful JND units.

C. BOOSTING FOR VISUAL QUALITY ASSESSMENT: MOTIVATION

When designing visual quality datasets, the creators usually try to cover the complete range of visual quality with only a few samples taken from a very large pool of images or videos. In the case of FR-IQA or FR-VQA, for each source stimulus, a large range of values for the distortion parameters of the chosen distortion types may be applied, yielding stimuli that are very close to the original as well as others with severe distortions. It may be hard to reliably assess the resulting small and large quality differences in subjective quality assessment experiments. Let us illustrate this, assuming the Thurstonian model for visual quality impairment.

Consider a sequence of $M + 1$ images $I_0, \ldots, I_M$, where I_0 is a pristine source image and $I_1, \ldots, I_M$ are increasingly distorted versions of the source. Let their image qualities on the impairment scale be modelled by random variables with Gaussian distributions with means $\mu_0, \ldots, \mu_M$ and standard deviations equal to $\sigma = \sqrt{0.5}/\Phi^{-1}(0.75) \approx 1.0484$, where Φ is the normal cumulative distribution function (CDF). This particular choice of the variance scales the quality values to be expressed in convenient JND units as pointed out in Subsection I-A.

When comparing two such stimuli that are close to each other in their mean values, the corresponding effect size Θ determines the difficulty of assessing their difference. A common way to define the effect size is the standardized mean difference. In our case,

$$\Theta = \frac{|\mu_i - \mu_j|}{\sigma} \approx 0.9539 \, \Delta\mu_{i,j}.$$

Smaller effect sizes indicate the necessity of larger sample sizes. Effect sizes $\Theta \in [0.2, 0.5), [0.5, 0.8), [0.8, 1.3)$ and $\Theta \geq 1.3$ are called small, medium, large, and very large, respectively. For example, at $\Delta\mu = 1 \, \text{JND}$, we have $\Theta = 0.9539 \in [0.8, 1.3)$, and thus, a large effect size. This is in line with the detection rate of 50% at 1 JND quality difference. However, for differences $\Delta\mu < 0.2097 \, \text{JND}$ we have $\Theta < 0.2$ and therefore a very small effect size.

Typically, image quality datasets have hundreds of images with perceptual qualities ranging over just a few JND. If one were to compare these images to each other, such small effect sizes would become relevant. Therefore, quality assessment techniques that enlarge the effect size would be beneficial, allowing one to distinguish image qualities with small differences with a smaller number of samples. For this purpose, we propose and study our boosting methods in this contribution.

It is quite natural that small visual differences are difficult to assess. It has been observed, in addition, that large quality differences are hard to quantify: Stimulus differences larger than about 1.5 JND cannot be reliably assessed by the human

visual system, presumably due to a kind of saturation effect by overwhelming noise [37].

In addition to the problem of subjectively quantifying large distortions reliably, there is the numerical problem of reconstructing such large quality differences from paired comparisons with the subsequent Thurstonian scale reconstruction. In the Thurstonian model, a quality difference is given by a normal random variable with unit variance and the mean equal to the quality difference on the perceptual quality scale. Thus, a fraction $p \in (0, 1)$ of observations that correctly identify the better quality image in the given pair gives rise to the reconstructed quality difference $\Phi^{-1}(p)$, where Φ again denotes the normal CDF. However, when the quality difference is large, most observers (k out of n) will agree on which image is of better quality. Therefore, the fraction $p = k/n$ is close to 1 and $\Phi^{-1}(p)$ depends very sensitively on p when p is near 1 or 0. For $k = n$, we even have $p = 1$ and $\Phi^{-1}(1) = \infty$. So, if just one observer would change his/her response, the reconstructed quality difference between the stimuli would drastically change.

We analyse this effect by simulating subjective PC using the probabilistic model to compute E_{rms}, the root of the expected square error of the reconstruction as follows. We assume a quality difference of $\Delta\mu$ and collect n responses for the corresponding pair comparison. Then we make use of the binomial distribution with probability $\Phi(\Delta\mu)$ to get the result.[1]

$$E_{\text{rms}}(\Delta\mu, n) = \left[\sum_{k=0}^{n} \binom{n}{k} \Phi(\Delta\mu)^k (1 - \Phi(\Delta\mu))^{n-k} \right.$$
$$\left. \times \left(\Phi^{-1}\left(\frac{k}{n}\right) - \Delta\mu \right)^2 \right]^{1/2}$$

Figure 1 illustrates this root-mean-square error (RMSE) as a function of $\Delta\mu$ for $n = 5, 10, 20, 40$. For increasing quality differences, we see that E_{rms} is stable and nearly constant until about 2 or 3 JND. From then on, the error increases approximately linearly. This gives another reason to restrict paired comparisons (resp. triplet comparisons) to cases where image quality differences are not too large, i.e., up to about 2 or 3 JND.

With our boosting methods implemented to enlarge the distortions applied to source images, this effect of decreased psychovisual sensitivity and increased reconstruction noise at large distortion levels can be overcome partially, as our experiments will show. However, by boosting image differences also between distorted images, we will show that also for large distortions, fine-grained quality scaling can be achieved reliably.

---

[1] The straightforward implementation of this formula will lead to the so-called zero-frequency problem when $k = 0$ or $k = n$. In these cases, $\Phi^{-1}\left(\frac{k}{n}\right)$ will be $\pm\infty$, rendering a reconstruction infeasible. To avoid this problem, it is common practice to install a 'prior' by adding half a vote to either option, thus computing $\Phi^{-1}\left(\frac{k+0.5}{n+1}\right)$, which is what we have done here as well.

TABLE 1. IQA/VQA/JND datasets with artificial distortions.

IQA Datasets	Year	SRC[a]	DST[b]	DST Types	DST Levels	Method	Scale Range[c]	Average Responses[d]	Environment
LIVE IQA [17]	2006	29	779	5	5–6	ACR	[0, 100]	23	Lab
CSIQ IQA [18]	2010	30	866	6	3–5	Customized[e]	[0, 1]	5–7	Lab
TID2008 [11]	2009	25	1700	17	4	Baseline TC[f]	[0, 9]	33	Lab
VCL@FER [19]	2012	23	552	4	6	SS-NS[g]	[1,100]	16–36	Lab
TID2013 [12]	2013	25	3000	24	5	Baseline TC[f]	[0, 9]	30	Lab
CID:IQ [20]	2014	23	690	6	5	ACR	[1,9]	17	Lab
MDID [13]	2016	20	1600[h]	5	4	Baseline TC[f]	[0, 8]	33–35	Lab
KADID-10k [21]	2019	81	10 125	25	5	DCR	[1, 5]	30	Crowdsourcing
KonFiG-IQA (Part A)	2021	10	840	7	12	Baseline TC	$[0, \infty)$	97	Crowdsourcing
KonFiG-IQA (Part B)	2021	10	300	1	30	General TC	$[0, \infty)$	582	Crowdsourcing
VQA Datasets									
EPFL-PoliMI [22]	2009	6	72[i]	2	6	SS-CS[j]	[0,5]	17–23	Lab
LIVE VQA [23]	2010	10	150	4	3–4	ACR	[0, 100]	38	Lab
IVP [24]	2011	10	128	4	4–5	ACR	[1, 5]	42	Lab
CSIQ VQA [25]	2014	12	216	6	3	SS-NS[g]	[0, 100]	35	Lab
MCL-V [26]	2015	12	96	2	4	PC	[0, 8]	32	Lab
NFLX [27] [28]	2016	9	70	2	7–11	DCR	[0, 100]	N/A	Lab
JND Datasets							JND Range		
MCL-JCI [29]	2016	50	5000	1	100	PC[k]	$\{1, ..., 100\}$	30	Lab
VVC-JND [30]	2020	202	7878	1	39	PC[k]	$\{13, ..., 51\}$	20	Lab
MCL-JCV [31]	2016	30	1530	2	51	PC[l]	$\{1, ..., 51\}$	50	Lab
VideoSet [32]	2017	220	44 800[m]	1	51	PC[k]	$\{1, ..., 51\}$	30	Lab
SIAT-JSSI [33]	2019	10	3510	2	51/300	PC[k]	$\{1, ..., 51\}, \{1, ..., 300\}$[n]	36	Lab
JND-Pano [34]	2018	40	4000	1	100	PC[o]	$\{1, ..., 100\}$	25	Lab
QAD-HEVC [35]	2017	40	2040	1	51	PC[k]	$\{1, ..., 51\}$	30	Lab

[a] SRC: number of source images/videos
[b] DST: number of distorted images/videos
[c] Range of MOS/DMOS (possibly scaled to a larger interval), reconstructed scale from comparisons, or JNDs
[d] Total number of quality ratings, resp. comparisons, divided by the number of images or videos. For JND datasets, it is the number of observers per source stimulus.
[e] All distorted images of one sequence were displayed simultaneously. Observers were asked to place the images in relation to each other according to the overall quality they perceived.
[f] Baseline TC: PC with the reference image provided.
[g] Single stimulus with numerical scales (SS-NS), observers are asked to assign each stimulus an integer from a given range.
[h] For each source 20 source images, 80 distorted images with combined distortions were generated by concatenating 4 types of distortions, in the order of Gaussian blur, contrast change, compression (JPEG or JPEG 2000), and Gaussian noise. Each distortion was randomly distorted at one of the four levels.
[i] Each of the 6 source videos was distorted by a simulation of packet loss distortion at 6 different packet loss rates with 2 channel realizations, resulting in a total of 12 distorted versions.
[j] Single stimulus with continuous scale (SS-CS), subjective quality scores of the videos were obtained using the single stimulus method with a 5-point continuous scale.
[k] Observers were asked to compare two images displayed side by side and determine whether the differences between them are noticeable.
[l] Observers were asked to determine whether the differences between the two video clips displayed one after another are noticeable.
[m] Four different resolutions were generated from each source video. The whole process of encoding and JND evaluation was carried out for each resolution. Therefore, the total number of distorted/encoded sequences in the VideoSet is 4 × 220 × 51.
[n] Reference stereoscopic images were compressed using H.265 intra coding with the quantization parameter ranging from 1 to 51 and JPEG 2000 with the compression ratio ranging from 1 to 300.
[o] Subjects wore a head-mounted display device and were free to control the field of view to compare two panoramic images whether they could notice a difference.

D. BOOSTING BY ARTEFACT AMPLIFICATION AND ZOOMING

The approach of boosting for visual quality assessment is to enlarge the differences between stimuli artificially. In the basic setting, using baseline triplets for comparison, we are asking subjects to identify the distorted image of the presented pair that most closely resembles the reference image, which is equal to the source image for generating the distorted versions. Here it is the distortions in each derived image that need to be enlarged for the boosting effect. Distortions typically occur in two main aspects, in terms of greyscale intensity, respectively color, or spatially. Therefore, the first two boosting techniques are as follows:

(1) Artefact amplification (A): In a triplet comparison, the similarity of two distorted images with respect to the pivot image has to be judged. Artefact amplification scales the pixel-wise differences of each distorted image in the three color channels linearly.

(2) Spatial zooming (Z): A linear scaling of the size of the image enlarges the visual representation of image differences. Due to limitations on the available screen size, spatial zooming may make it necessary to crop the distorted and zoomed image.

In Figure 2 we show an example of how this boosting reveals otherwise invisible or hard to detect distortions caused by JPEG 2000 compression.

E. BOOSTING BY THE IMAGE FLICKER METHOD

In PC, the image stimuli are usually displayed on a screen side by side. To detect a small detail that differs between two images, the observer must search over both images and memorize the last examined detail of one image when the eye fixation point moves to the corresponding location in the other image. This task can be difficult. It is at the core of the

RMSE of Reconstructed Difference of Means
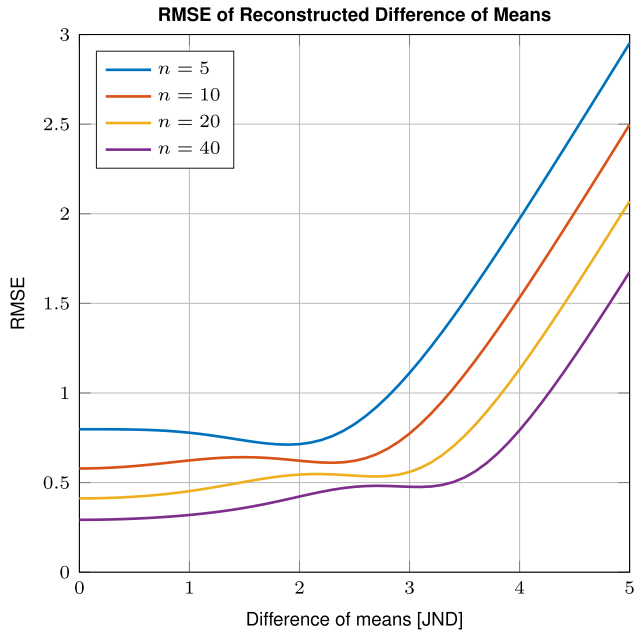
FIGURE 1. In this simulation, we consider the reconstruction of a quality difference from a set of *n* pair comparisons of two stimuli with a given difference of mean observed qualities, shown on the horizontal axis. On the vertical axis, the resulting RMSE of the reconstruction of the difference is shown. This simulation confirms that it is not advisable to apply paired comparisons when the difference to be assessed is so strong that nearly all observers agree on which of the two stimuli should be chosen as the better (or stronger) one.

FIGURE 2. This figure shows an example of boosting by artefact amplification and zooming. The top left image is an original, undistorted source image. On the right, a zoomed crop of it is pictured. The source image was compressed by JPEG 2000 with a compression ratio of 1:47.34, resulting in the lower-left image. The distortions due to JPEG 2000 compression are barely visible. However, on the lower right, the same compressed image is shown after artefact amplification and zooming. Now, by this boosting technique, the distortions in the flower petals and stigma became emphasized and clearly visible.

popular fun game for kids, where differences between two seemingly identical comic drawings are to be found.

The change detection task can be simplified by displaying the two images in the same screen space, one in the foreground and the other one invisible in the background. The observer can toggle the view between the two images using keystrokes or mouse c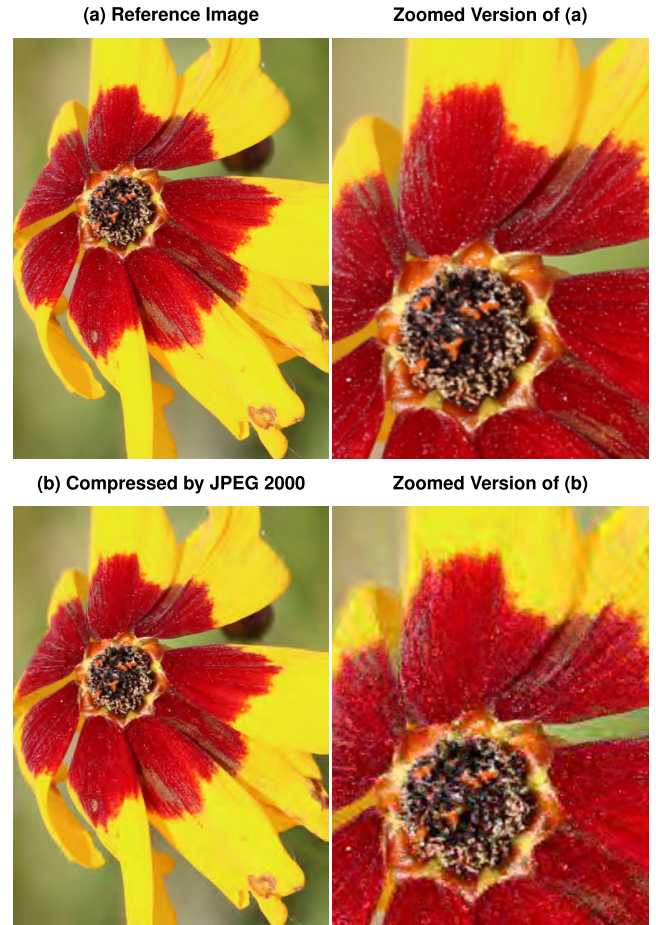licks. In this setting, the eye needs to scan only half of the screen space, and the saccadic eye movements between the two images are replaced by key clicks. Moreover, from an evolutionary perspective, it is plausible that the fundamental ability of human perception to detect a change in the visual environment has developed to a high level. Thus, small changes in the visual field should be easier to detect than the differences between two objects (or images) next to each other.

The visual sensitivity to contrast change has been researched for a long time [38]. For simple test scenes, contrast is defined as the ratio of the target intensity to the background intensity. When the contrast changes periodically, e.g., in a sinusoidal fashion, the change becomes visible when its amplitude surpasses a certain threshold, called the contrast threshold. Its inverse is the contrast sensitivity, and its variation as a function of temporal frequency can be described by the temporal contrast sensitivity function (TCSF). For a sufficiently high luminances, the contrast sensitivity reaches a maximum of about 200 near a frequency of 8 Hz.

In 2014, the above concepts were applied for the first time in an image flicker viewing method for subjective assessment

of barely visible image compression artefacts [39]. Observers were presented with an original reference image, temporally interleaved with a test image, which was reconstructed from the compressed reference. The flicker frequency was chosen as 7.5 Hz, close to the maximum of the TCSF. This method was expected to make even subtle artefacts visible that would be undetectable in a side-by-side comparison. The paper did not compare the performance with that achievable by the side-by-side display, however.

In our work, we provide such studies by including the flicker viewing method as our third option of boosting techniques:

(3) Image flicker (F): Two images to be compared are displayed temporally, interleaved at a frequency of 8 Hz.

The application of the flicker viewing technique in a triplet comparison requires adapting the visual appearance of the displayed scene. The triplet (i, j, k) has the pivot image I_j, and we are asking the observer to determine whether the

perceived differences in the left image pair (I_i, I_j) are larger or smaller than those in the right pair, (I_k, I_j). Therefore, with the flicker viewing technique, we show two flickering images side by side: On the left, image I_i alternates with the pivot, and on the right, it is I_k that alternates with the pivot.

F. CONTRIBUTIONS

We expect (and we will show) that the boosting methods, outlined in the previous subsections, increase the measurement sensitivity on the visual impairment scale, enabling the detection of more subtle artefacts. In our experiments, we investigated the performances by measuring sensitivities with respect to the magnitude of the applied distortion. We selected ten source images from the MCL-JCI dataset [29], each of which was distorted by seven types of distortion: color diffusion, jitter, high sharpen, JPEG 2000 compression, lens blur, motion blur, and multiplicative noise. Figure 3 shows the flowchart of our boosted triplet comparison method for subjective IQA.

Along with the boosting of sensitivity, however, we have to accept that the absolute values of impairment, given in JND units, will be different and typically larger than those obtained using plain pair or triplet comparison or by using the DCR method. For example, if a particular distortion produces an impairment of 1.5 JND, measured by plain comparison, we may obtain a much larger impairment of perhaps as much as 3 JND when using one of the boosting methods. Therefore, we have also devised a method to adaptively transform the boosted impairment quality values back so that they approximately match the range of impairment scales as measured by plain comparison, however, without sacrificing the better discrimination ability.

To summarize, in this paper, we present the first study on the potential of perceptual boosting techniques in the context of subjective image quality assessment. The main contributions are the following:

1) We propose three boosting strategies (artefact amplification, zooming, and flicker) that can enlarge the sensitivity of pair and triplet comparisons, as well as increase the accuracy of visual quality assessment.
2) We propose a method based on Thurstone's model and MLE to reconstruct the perceptual qualities of images from triplet comparisons.
3) We generate an IQA dataset of 1140 images with distortions from 10 source images. We provide the responses to all triplet comparisons from a large subjective crowdsourcing campaign together with the reconstructed quality values from plain triplet comparison and seven types of boosted triplet comparison. This dataset will be made available after publication.
4) We provide an extensive performance analysis of boosted triplet comparisons for image quality assessment, including measures of the true positive responses, detection rates, sensitivity, effect size, convergence, correlation, and time complexity.

G. GLOSSARY

Sequence. A sequence is a list of images $(I_0, I_1, \ldots, I_K)$ where I_0 is a pristine **reference image** and the others are increasingly distorted versions of it. The index k of I_k is called the **distortion level** of I_k.

Triplet. A triplet is a list (i, j, k) of three non-negative indices referring three stimuli (I_i, I_j, I_k) of an image sequence. In a **triplet comparison**, the observer is asked to determine whether the left or the right image $(I_i$ or $I_k)$ is perceptually closer to the so-called **pivot image** I_j in the middle.

Baseline triplet. A triplet of the form $(i, 0, k)$, where the pivot is the pristine reference image I_0.

General triplet. A triplet of the form (i, j, k), where the pivot could be any image (could be the pristine reference image or a distorted one).

Pivot. The stimulus that is placed in the middle of a triplet.

JND. The natural unit on the perceptual scale. The perceptual difference between the two compared images is 1 JND (just noticeable difference) when the difference is perceived by a random observer of the population of observers with a probability of 0.5.

HIT. A human intelligence task is a self-contained, virtual task that a worker can work on. A HIT may contain several questions to be answered.

Assignment. A completed HIT that is submitted by a unique worker.

Response. Equals vote or answer for a pair or triplet comparison.

Rating. Selected image quality or degree of impairment in an ACR or DCR assessment.

Plain. A plain triplet comparison is a conventional one, where the images to be compared are not processed (boosted).

A, Z, F. Abbreviations for triplet comparisons with artefact amplification, zooming, and flicker, respectively.

AZ, AF, ZF, AZF. Abbreviations for triplet comparisons of combinations of A and Z, A and F, Z and F, and A, Z, and F.

II. RELATED WORK

A. BOOSTING BY ARTEFACT AMPLIFICATION AND ZOOMING

Amplification and zooming are common techniques applied in image and video processing, for the purpose of highlighting details in an image or for overall image enhancement. For instance, histogram equalization enhances contrast by enlarging small intensity differences. Image sharpening enhances the appearance of edges, e.g., by adding to the input image a signal that is a scaled high-pass filtered version of the original image. Motion and color magnification reveals subtle variations in video sequences that cannot be perceived by human senses, such as heartbeats showing in faces [40], and can even reconstruct sounds from videos of objects subtly vibrating in response to those sounds [41]. Exaggeration also
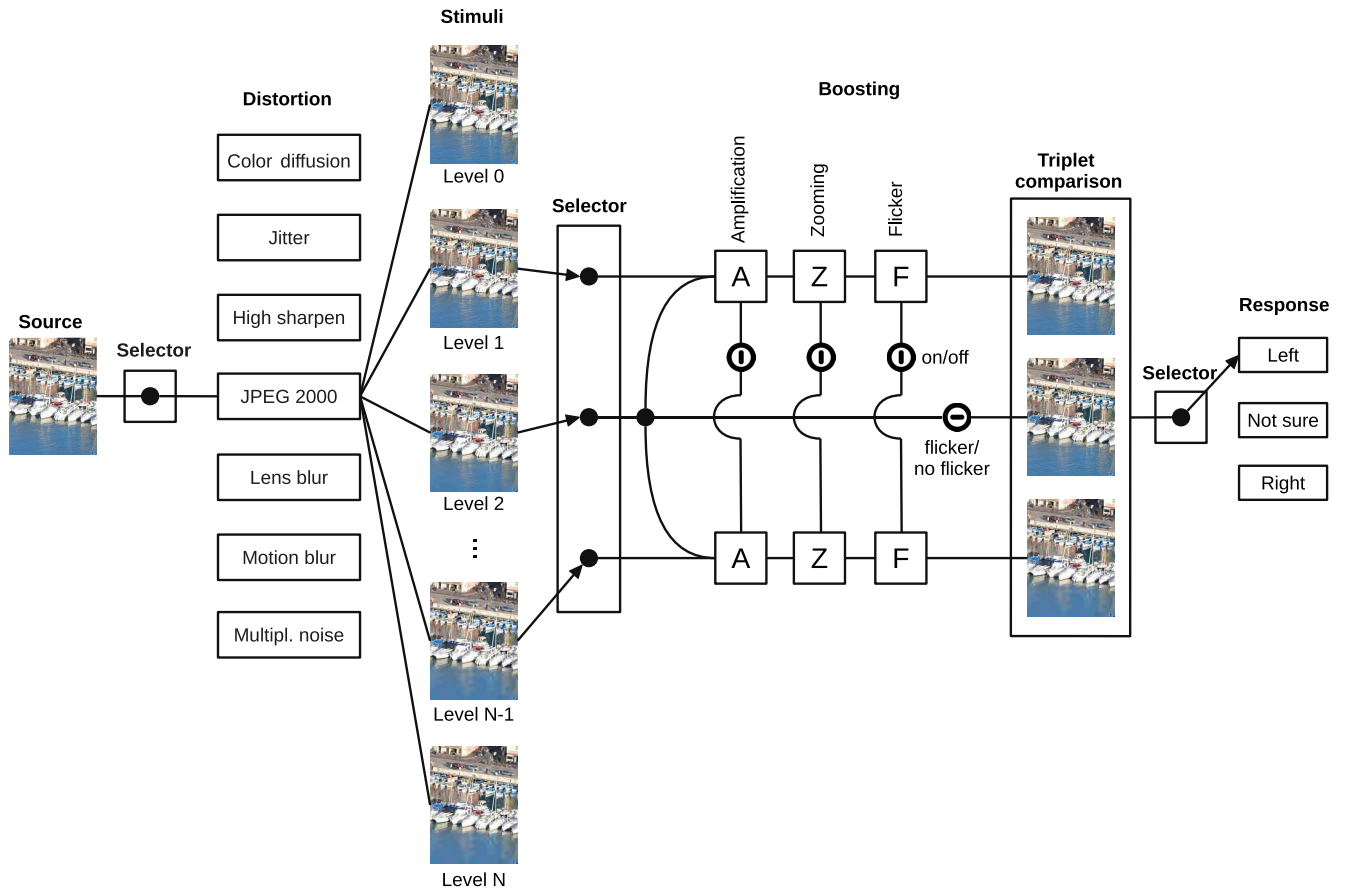
FIGURE 3. Flowchart of the proposed boosted triplet comparison method for subjective IQA. Each pristine source image is distorted by seven types of distortion at *N* levels, respectively. For a given distortion type, samples of three distorted images are drawn, and for each triplet, subjective assessments are made as to whether the left or the right image is perceptually closer to the center, pivot image. Three boosting strategies (amplification, zoom, and flicker) are deployed individually or in combination to enhance the perceptual sensitivity of comparisons between distorted images. With the flicker option turned on, only two (flickering) images are shown side by side. The yielded triplet comparison results are to be used to estimate image quality impairment scales using Thurstonian reconstruction.

is one of Disney's twelve basic principles of animation [42], applied in particular to motion. The cartoon characters were designed to maintain the illusion that they follow the laws of physics, however, in a wilder, more extreme form.

In spite of the widespread applications of amplification and zooming in multimedia applications, we have not become aware of any previous work on adapting, applying, and validating these techniques for the purpose of subjective visual quality assessment.

The only exception is our earlier work [43], in which we applied image distortion amplification and zooming to properly cropped frames to compare interpolated video frames with the corresponding ground-truth frames. However, this technique was a side issue in that contribution, and there was no systematic analysis of the performance and potential of amplification and zooming.

B. BOOSTING BY THE IMAGE FLICKER METHOD

In [39], flicker tests were proposed for the task of determining the JND of distorted images. In order to test for noticeable distortions in compressed images, two images were placed side by side. One of them was a still reference image.

The other was an animation in which a reference image alternated with the distorted image at the same spatial position. The display frame rate was 30 fps, and the alternation occurred every four frames, yielding a flicker frequency of 7.5 Hz. Observers were asked to identify which of the two images was the still (or non-flickering) one (two-alternative forced-choice).

Shortly later, the ISO/IEC standard 29170-2:2015(E) was set up with recommendations for the subjective quality evaluation of single frames from JPEG XS encoded videos [44]. The focus of the standard was on visually lossless, low-latency, and lightweight video compression schemes. Therefore, subjective tests were prescribed for image JND assessment rather than conventional MOS from ACR or DCR procedures. The standard directly follows the ideas in [39], with one notable exception: The flicker rate was recommended to be 10 Hz instead of 7.5 Hz.

The first study, based on the new standard, was published in 2017. In a large lab experiment with 120 subjects, the flicker test was used for video frames, produced by the Video Electronics Standards Association (VESA) Display Stream Compression, which is a lightweight codec designed

for visually lossless compression [45]. An in-depth discussion of this application of the flicker test for the criterion of perceptually lossless compression as prescribed in the ISO/IEC standard was presented in [46]. The authors' conclusion was that ''if the goal is to conservatively evaluate the possibility that a compression artefact might be visible under any situation, then the flicker paradigm is a viable approach as it highlights differences between images regardless of whether they are noticeable in the absence of a reference.'' However, a quantitative comparison of the sensitivity of the flicker test protocol versus the traditional side-by-side presentation was not included.

In 2016, a JPEG XS Call for Proposals with subjective quality evaluations based on the flicker test of the ISO/IEC standard was issued [47]. The results of the submissions to the call were summarized in [48]. Different from the ISO/IEC standard recommendations, a flicker frequency of 8 Hz was applied, and subjects were given a third option for their response, namely to cast a no-decision vote. This was deemed to alleviate subject fatigue.

In other recent work, the flicker test was applied to a palette of different image modalities: high dynamic range (HDR) images [49], foveated images in head-mounted displays (HMD) [50], and stereoscopic imagery [51].

In our pre-study [52], presented at the ICME 2020 Workshop on Data-driven Just Noticeable Difference for Multimedia Communication, we have provided the first experiment to compare the performance of the flicker test with conventional side-by-side comparisons. The purpose of the tests was to assess the JND for JPEG image compression. As a result, we reported that the flicker test was about twice as sensitive as the classical side-by-side comparisons with forced choice. However, this experimental study was small, and the focus was rather on a new adjustment method for JND detection using a slider-based design. Moreover, the flicker tests were done in a lab situation, while the classical 2AFC pair comparison ran on a crowdsourcing platform. So the results of the comparison regarding sensitivity is only a preliminary. Our contribution here provides a much more elaborate study targeted specifically at the validation of the performance of several boosting techniques, with flicker being one of them.

The previous works on the image flicker method mentioned above have applied flicker in a single stimulus or in a double stimulus method where one of the two stimuli was a still image. We extend these procedures by comparing two flickering stimuli in the context of a triplet comparison. Moreover, in past approaches, the flickering was between the undistorted reference image and a test image. We will show the advantages of considering flicker images in comparisons where the flicker is between two test images.

C. RECONSTRUCTION OF SCALE VALUES FROM TRIPLET COMPARISONS

Direct quality assessment proceeds by collecting and averaging quality ratings from a sufficiently large set of observers. Absolute category rating is the most common technique in

visual quality assessment. Scale value reconstruction is an indirect procedure, deriving the scales of latent variables from the pair comparisons of the perceptual quality or from the comparison of quality differences in triplets or quadruplets. Other approaches are possible, like reconstruction from the rankings of images in subsets of stimuli. For the application of boosting methods in subjective visual quality assessment, indirect methods seem more appropriate because boosting enhances the perception of differences between a test stimulus and its corresponding reference.

One of the first indirect approaches, based on the scaling of perceived distances of stimuli, attained from triplet comparisons, was proposed in 1952 by Torgerson [53] and named the *method of triads*. Although the goal was multi-dimensional scaling, it is clear that the method can also be used to derive the scalar values of a latent variable. In a nutshell, for the 1D case, the reconstruction is based on a model of random variables X_i, $i = 0, \ldots, M$, for the latent stimuli qualities with the assumption that their pairwise distances

$$D_{i,j} = |X_i - X_j|$$

are random variables with a normal distribution of unit variance. Then scale values of these distances can be reconstructed from the pairwise comparison of distances arising from the triplet comparisons. This step is analogous to the Thurstonian scale reconstruction from pair comparison of stimulus values, where instead of scales for the random variables X_i, scales for the distances $D_{i,j}$ are considered.

However, since the triplet comparisons (i, j, k) give information only about differences in distances (namely whether $D_{i,j} < D_{k,j}$), the reconstruction of the distances can be determined only up to an additive constant. In [53], the least squares solution to solve the problem of the additive constant was proposed.

At the end, a square matrix of distances $(d_{i,j})$, $i, j = 0, \ldots, M$ is yielded, from which a one-dimensional embedding can be generated. One can solve the optimization problem, where estimates for the latent variables are found as a minimizer of a cost function, for example,

$$\left(\hat{\mu}_0, \ldots, \hat{\mu}_M\right) = \arg \min_{\mu_0, \ldots, \mu_M} \sum_{i<j} \left(|\mu_i - \mu_j| - d_{i,j}\right)^2.$$

The method of triads has been criticized as being ad hoc [54], as it does not directly follow the basic setup of Thurstonian models, where the latent variables of the stimuli themselves are normally distributed. Moreover, distances are non-negative and cannot be modelled accurately by normal distributions.

In our contribution, we propose a complete solution for the reconstruction of scale values from triplet questions that strictly adheres to the considerations of Thurstonian models. Let us assume such a model of a set of normally distributed random variables X_i, $i = 0, \ldots, M$ of equal variance, for the visual qualities of a corresponding set of stimuli. We will make use of a formula for the probabilities $\Pr\left(D_{i,j} < D_{k,j}\right)$

for the outcome of a given triplet comparison (i, j, k), that was derived by Ennis *et al.* [55] in 1988.

Let $Q_{i,j,k}$ denotes the empirical estimation of $\Pr(D_{i,j} < D_{k,j})$ given by the fraction of responses to the comparison for the triplet (i, j, k) that indicate that the left stimulus I_i is closer to the pivot I_j than the right one, I_k. Then the task to be solved is to reconstruct the mean values of the model such that the model predictions for the TC outcomes match the empirical data:

$$\Pr(D_{i,j} < D_{k,j}) \approx Q_{i,j,k}.$$

For this purpose, in [55] the method of least squares was proposed,

$$\min_{\mu_0, \dots, \mu_M} \sum_{i<k, j\neq i,k} \left(Q_{i,j,k} - \Pr\left(D_{i,j} < D_{k,j}\right)\right)^2.$$

This is equivalent to the MLE in which the prediction errors $\Pr(D_{i,j} < D_{k,j}) - Q_{i,j,k}$ are modelled as independent normal random variables with equal variance [56]. This assumption generally cannot hold since for small or large probabilities $\Pr(D_{i,j} < D_{k,j})$ near 0, resp. 1, the error distribution necessarily must be skewed. Therefore we favor the general MLE method, i.e., to maximize the model likelihood of the set of observations $Q_{i,j,k}$.

While this choice follows the common approach taken in psychometrics [57], the most widely used one-dimensional scale reconstruction method for vision science applications is probably maximum likelihood difference scaling (MLDS). It solves the difference scaling problem of quadruplet questions (i, j, k, l), where the perceptual distance of the first pair of stimuli, (I_i, I_j), is compared to that of the second pair, (I_k, I_l), in a 2AFC setting. In MLDS, the decision variable employed by an observer of such quadruplet questions is modelled as

$$Z = |x_j - x_i| - |x_l - x_k| + N_\sigma,$$

where $x_i, i = 0, \dots, M$ are the (crisp) unknown qualities and N_σ is a zero-mean Gaussian noise term with variance σ^2. The unknown variance characterizes the difficulty of the particular set of quadruplet questions together with the uncertainty of the subjects.

It is worth noting that MLDS presents an approach of a fundamentally different type than Thurstonian models. In Thurstonian models, the decision variable is $Z = |X_j - X_i| - |X_l - X_k|$ and deterministic, but the qualities on the perceptual scales are uncertain, given by normally distributed random variables $X_i, i = 0, \dots, M$. In MLDS, it is just the opposite. The quality values are crisp, while there is uncertainty in the decision variable.

The number of free parameters for the $M + 2$ unknowns in MLDS is M, and one may set the range of scale values to $[x_0, x_M] = [0, 1]$ and solve for the variables $x_1, \dots, x_{M-1}$ and σ, using MLE. Alternatively, one can set $x_0 = 0$, the variance σ^2 to a fixed value, and then solve for the scales $x_1, \dots, x_M$.

In this paper, we contribute a method for the selection of the variance σ^2 of the noise term such that the resulting reconstruction yields scale values in approximate JND units.

A recent survey discusses the MLDS method, its variations, and a very large number of applications in different fields [58]. The two contributions, most closely related to our work on visual quality assessment, are [59] and [60] where quadruplet comparisons for image sequences with distortions due to compression were undertaken and analysed by MLDS.

Although MLDS was designed for scale reconstruction from quadruplet comparisons, it is clear that it can also be applied to triplet comparisons (i, j, k), simply by restricting to quadruplets of the form (i, j, j, k). Then the decision variable is $Z = |x_j - x_i| - |x_k - x_j| + N_\sigma$. In practice, it can be expected that its normal distribution is very similar to that for the decision variable $Z = |X_j - X_i| - |X_k - X_j|$ which arises from the Thurstonian model (X_i, X_j, X_k normally distributed with variance 1/2). However, for our particular applications in visual quality assessment (FR-IQA), we prefer a reconstruction method that can produce scale values in JND units. MLDS was not designed for this purpose.

There are several other methods for scale reconstruction from triplet comparisons, some of which have recently originated from the machine learning community [16]. Usually, these methods are for multi-dimensional scaling. Some of them can be restricted to the one-dimensional case. In Section IV, we compare our results with those computed by MLDS and stochastic triplet embedding (STE) [61]. In terms of correlation with ground truth, all methods showed excellent performance. However, as for MLDS, also STE cannot be expected to yield estimates on JND scales.

Let us finally remark that there also is an ISO standard that proposed triplet comparison [37]. Observers rate each image in a triplet using a 5-point ACR scale, and from that, all three pair comparisons in the triplet are deduced. The selections of the triplets in an experiment is prescribed and rather restricted. This setting would not support the boosting strategies discussed in this paper.

III. BOOSTING STRATEGIES
We apply three methods to boost the perceptual sensitivity of comparisons between distorted images: 1) *Artefact amplification* amplifies the artefacts of distorted images relative to their references, 2) *Zooming* enlarges the visual representation of the images, and 3) *Flicker* increases the perceptual sensitivity to the distortions by rapidly alternating between distorted images and their corresponding reference images.

A. ARTEFACT AMPLIFICATION (A)
Many ways can be conceived to amplify artefacts due to distortions in images. For this study, we consider one of the simplest kinds, namely the linear pixel-wise scaling of RGB color differences between the distorted and the reference images. Let $v, \hat{v}$ be the RGB pixel values of a pixel in the reference and a distorted image, respectively. Then we replace $\hat{v}$ by $\hat{v}' = v + \alpha(\hat{v} - v)$.
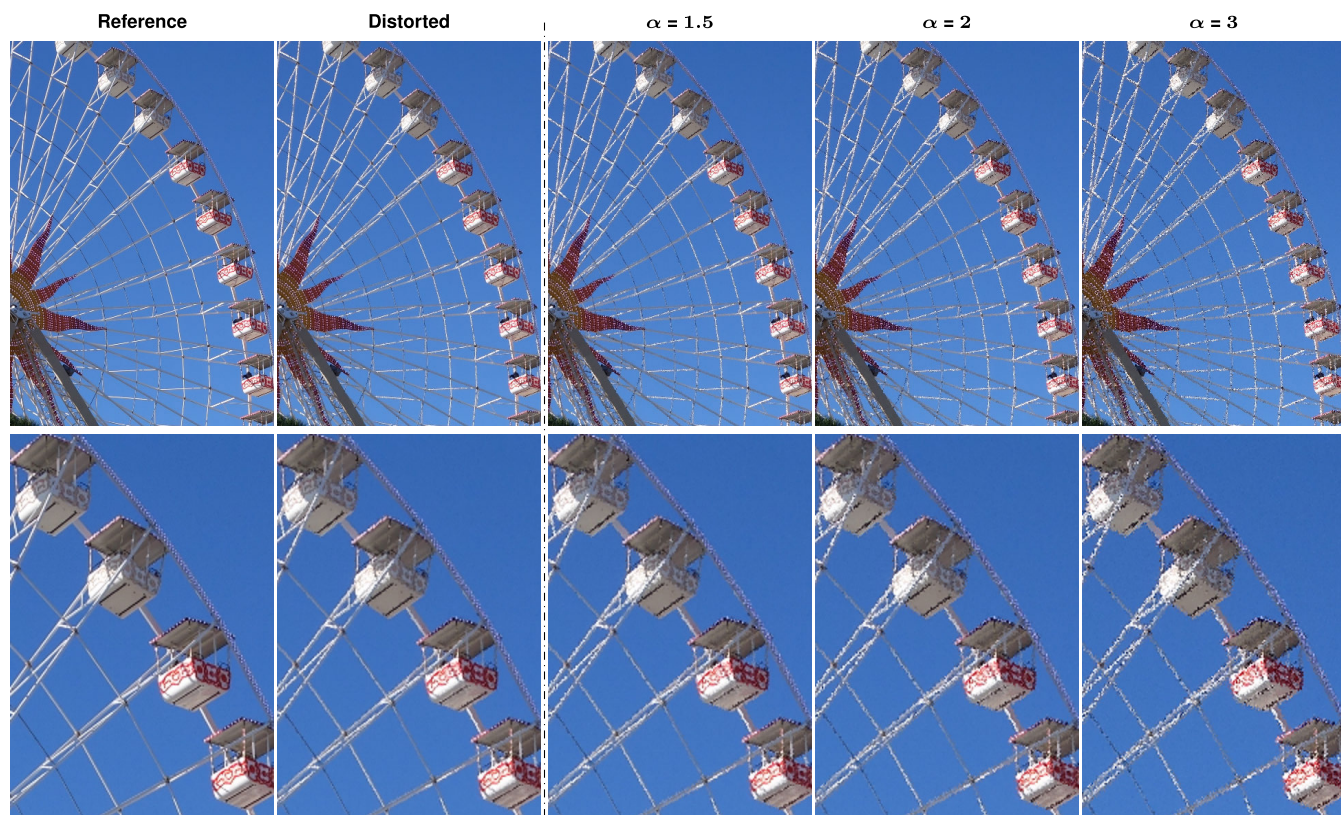
FIGURE 4. Illustration of artefact amplification and zooming. The upper row shows an original source image and its distorted version when jitter is applied, at a level corresponding to 0.5 JND. The artefacts are amplified with factors $\alpha = 1.5$, 2, and 3 in the three top right images. Differences between the reference and the distorted image are barely visible, but with increasing amplification, they become more noticeable. The bottom row presents a zoomed version of the upper row. The visibility of the distortions is further enhanced by zooming.

Algorithm 1 Pixel-Wise Artefact Amplification

1: $\alpha \leftarrow 2$ ▷ default amplification factor
2: $v \leftarrow (v_r, v_g, v_b)$ ▷ ground truth pixel
3: $\hat{v} \leftarrow (\hat{v}_r, \hat{v}_g, \hat{v}_b)$ ▷ distorted pixel
4: **for** c $\in$ {r,g,b} **do**
5:     **if** $\hat{v}_c - v_c > 0$ **then**
6:         $\alpha_{c,max} \leftarrow (255 - v_c)/(\hat{v}_c - v_c)$
7:     **else if** $\hat{v}_c - v_c < 0$ **then**
8:         $\alpha_{c,max} \leftarrow -v_c/(\hat{v}_c - v_c)$
9:     **else**
10:        $\alpha_{c,max} \leftarrow \alpha$
11:     **end if**
12: **end for**
13: $\alpha \leftarrow \min(\alpha, \alpha_{r,max}, \alpha_{g,max}, \alpha_{b,max})$
14: $\hat{v}' \leftarrow v + \alpha(\hat{v} - v)$ ▷ amplified pixel $\hat{v}' \in [0, 255]^3$

The multiplication by the factor $\alpha > 1$ ensures consistency with Fechner's law [62], which states that the subjective sensation is proportional to the logarithm of the stimulus intensity. In our context, this means that equal relative increments of distortion, i.e., the same factor α applied in artefact amplification, should correspond to equal increments of perceived impairment in these images.

However, due to the finite range of RGB color components in digital images, the linear scaling is limited and RGB pixel

TABLE 2. Fraction of pixels clamped in artefact amplification of Figure 4.

Amplification factor	Channel			Pixels overall
	red	green	blue	
$\alpha = 1.5$	0.0012	0.0011	0.0021	0.0022
$\alpha = 2.0$	0.0018	0.0017	0.0038	0.0050
$\alpha = 3.0$	0.0036	0.0039	0.0083	0.0131
$\alpha = 4.0$	0.0068	0.0068	0.0135	0.0242
$\alpha = 5.0$	0.0107	0.0097	0.0184	0.0357

values exceeding the limit must be clamped. Thus, to restrict the RGB components of $\hat{v}'$ to the range [0, 255] for 24-bit color images, we reduce α accordingly for those pixels where clamping is needed. Note that this may cause a local nonlinearity and saturation effect of the artefact amplification. See Algorithm 1 for details.

An example of artefact amplification is shown in Figure 4. Comparing the distorted image with the reference image in the first row, the distortion is hardly visible. In contrast, some distortions are noticeable after artefact amplification and become more and more obvious with the increase of the amplification factor (top right row).

Table 2 shows the fraction of the pixel color components and the overall number of pixels that are clamped in the amplification process, for the example shown in Figure 4. These fractions are monotonically increasing with the amplification factor α. In order to avoid a widespread nonlinearity and the saturation effect due to clamping too many pixels,
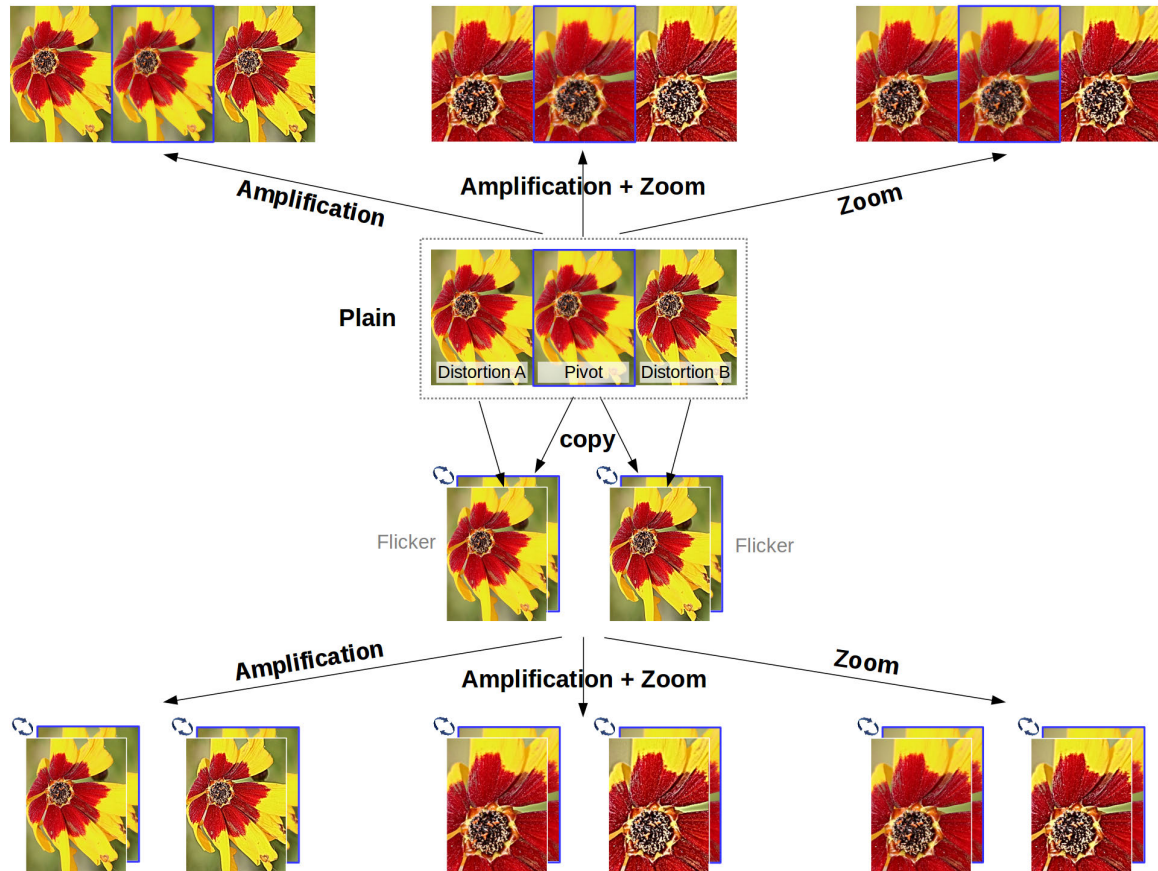
FIGURE 5. Overview of plain triplet comparison and seven types of boosted triplet comparison. In plain triplet, three stimuli, selected from an image sequence with increasing distortion levels, are displayed side by side. The task is to judge which side is perceptually closer to the pivot in the center. The three boosting techniques, artefact amplification, zooming, and flicker, help to improve the accuracy, reliability, and speed of the subjective assessment. When boosting with flicker, the left and right images are displayed side by side, each alternating with the pivot eight times per second. In this case, observers judge which side has the stronger flicker effect.

α should be chosen appropriately. Note that, depending on the application, more or less strong amplification can be used. In triplet comparisons, we applied the amplification relative to the pivot stimulus displayed in the center. Therefore, for baseline triplets of the sort $(i, 0, k)$ the differences that are amplified, are typically larger than for most general triplets (i, j, k) with $i < j < k$ or $i > j > k$. Thus, a more conservative (i.e., smaller) amplification factor α should be used for baseline triplets. In this paper, we set $\alpha = 2$ as the artefact amplification factor.

B. ZOOMING (Z)

Apart from enlarging color differences by amplifying artefacts, artefacts appear more visible when enlarged spatially. In fact, participants in an IQA experiment may be tempted to enlarge the images displayed in their browser or to move closer to the screen to detect fine differences between images. However, to ensure a uniform and controlled quality assessment, participants are asked to refrain from such ad hoc zooming actions. Instead, we propose to deliver the displayed images already in a zoomed and cropped fashion.

Figure 4 (bottom row) shows an example. Images are cropped to half their linear size and zoomed by a factor of two. The cropped regions were manually selected, and bicubic interpolation was adopted for scaling up. Distortions in the zoomed images are more clearly visible, especially as the artefact amplification factor α increases.

Similarly to artefact amplification, larger zoom factors are better for the visual detection of artefacts, but at some point, undesirable side effects like pixelation set a limit on zooming. Due to the required cropping, only a part of the image content is maintained in the zoomed images, possibly masking image areas with more severe local distortions. In this paper, we chose a fixed zoom factor of two.

Regarding the area for cropping, we manually cropped the images based on the content. We cropped to a reasonable and sensible scene. For instance, cutting half of a face was avoided.

C. FLICKER (F)

The visibility of distortions can also be enhanced by making use of the flicker effect as explained in Subsection I-E. In this

scenario, a distorted image and its reference are displayed successively at a frequency of 8 Hz. As already mentioned, for a triplet comparison (i, j, k), using the flicker technique, we show two flickering images side by side, on the left, image I_i alternates with the pivot I_j, and on the right I_k alternates with the pivot. Observers are asked to select the side that has a stronger flickering effect.

D. COMBINATIONS OF BOOSTING TYPES

With the three boosting options on hand, we can apply them individually or in combination, for example, zooming together with artefact amplification. This gives rise to seven cases that we abbreviate with the letters A, Z, and F assigned to the boosting methods artefact amplification, zooming, and flicker, respectively. The combinations are A, Z, F, AZ, AF, ZF, and AZF. Without boosting, we obtain *plain* results, which serve as a reference when assessing the performance of the boosting methods. Figure 5 shows an overview of the different kinds of boosting options as applied for triplet comparison.

IV. THURSTONIAN RECONSTRUCTION FROM TRIPLET COMPARISONS

Let us consider a set of $M + 1$ stimuli. In our experiments, these are a source image I_0 together with derived distorted images $I_1, \ldots, I_M$. The magnitudes of the perceived stimulus qualities are taken to be unknown latent variables, modelled by normally distributed real-valued random variables $X_0, \ldots, X_M$ with variance equal to 1/2. It is the purpose of reconstruction to estimate their means $\mu = (\mu_0, \ldots, \mu_M) \in \mathbb{R}^{M+1}$ from a collection of responses to subjective triplet comparisons of stimuli. This setup corresponds to the assumptions Thurstone established as Case V in his analysis for pair comparisons [7].

In Subsection IV-A, we present formulas for the computation of the probabilities of the responses for the triplet comparisons, followed by MLE of the means in Subsection IV-B. In Subsection IV-C, we compare the reconstruction performances of MLDS, STE, and our method by means of a simulation with available ground truth data. We also compare the probabilistic model for the decision random variable in MLDS with the uncertainty of the means in the Thurstonian model. This gives rise to a choice of the unspecified variance of the MLDS decision variable such that the reconstructions of the means of the stimuli values on the perceptual scale are given in approximate JND units.

A. FORMULAS FOR THE PROBABILITIES OF THE RESPONSES

We define that for a triplet $t = (i, j, k)$ with $i, j, k \in \{0, \ldots, M\}$, a subjective comparison yields a response $R_{ijk} = 1$, if the observer judges the left stimulus, numbered i, closer to the pivot stimulus j than the right one, k. Otherwise, the response is $R_{ijk} = 0$.

Algorithm 2 Probability of a Response $R_{ijk} \in \{0, 1\}$ to a Triplet Comparison (i, j, k)

1: $\boldsymbol{\mu} = (\mu_0, \ldots, \mu_M)$ ▷ stimuli means in model
2: $u_0 \leftarrow \mu_k - \mu_i$
3: $v_0 \leftarrow (\mu_k + \mu_i - 2\mu_j)/\sqrt{3}$
4: $p \leftarrow 1 - \Phi(u_0) - \Phi(v_0) + 2\Phi(u_0)\Phi(v_0)$
5: **if** $R_{ijk} = 1$ **then**   ▷ stimulus i closer to j than k
6:     Return p
7: **else**   ▷ stimulus k closer to j than i
8:     Return $1 - p$
9: **end if**

From the Thurstonian probabilistic model, it follows that observers act according to the sign of the decision variable

$$Z_{ijk} = |X_k - X_j| - |X_i - X_j|, \tag{1}$$

such that

$$R_{ijk} = \begin{cases} 1 & \text{if } Z_{ijk} > 0 \\ 0 & \text{if } Z_{ijk} \leq 0 \end{cases}. \tag{2}$$

Next, we first give an expression, based on a result of [55], for the probability that the decision variable is positive. It will be a function of the unknown means $\boldsymbol{\mu} = (\mu_0, \ldots, \mu_M)$, so we write it as the conditional probability $\Pr(Z_{ijk} > 0 \mid \boldsymbol{\mu})$. Given a triplet comparison $t = (i, j, k)$, the probabilities for the response $R_{ijk} = 1$ (left stimulus i is closer to the pivot stimulus j than stimulus k) and the opposite, $R_{ijk} = 0$, is

$$\Pr(Z_{ijk} > 0 \mid \boldsymbol{\mu}) = 1 - \Phi(\mu_k - \mu_i) - \Phi\left(\frac{\mu_k + \mu_i - 2\mu_j}{\sqrt{3}}\right)$$
$$+ 2\,\Phi(\mu_k - \mu_i)\,\Phi\left(\frac{\mu_k + \mu_i - 2\mu_j}{\sqrt{3}}\right)$$
$$\Pr(Z_{ijk} \leq 0 \mid \boldsymbol{\mu}) = 1 - \Pr(Z_{ijk} > 0 \mid \boldsymbol{\mu}). \tag{3}$$

Algorithm 2 summarizes the computation.

Other probabilistic models differ from the above by specifying a different probability for the triplet responses. In MLDS [63], we have

$$Z_{ijk} = |\mu_k - \mu_j| - |\mu_i - \mu_j| + N_\sigma.$$

In this case,

$$\Pr(Z_{ijk} > 0 \mid \boldsymbol{\mu}) = \Phi\left(\frac{|\mu_k - \mu_j| - |\mu_i - \mu_j|}{\sigma}\right). \tag{4}$$

The default for the parameter is $\sigma = 1$. In stochastic triplet embedding (STE, [61]), the probability for a positive response is given directly as

$$\Pr(Z_{ijk} > 0 \mid \boldsymbol{\mu}) = \frac{e^{-\alpha(\mu_i - \mu_j)^2}}{e^{-\alpha(\mu_i - \mu_j)^2} + e^{-\alpha(\mu_k - \mu_j)^2}}. \tag{5}$$

The parameter $\alpha > 0$ is not contained in the original method. Thus, its default value is $\alpha = 1$. Here, we have introduced it for the purpose of model calibration.

In the special case of baseline triplets of the form $(i, 0, k)$, we have that the response $R_{i0k} = 1$, i.e., that the left stimulus, numbered i, is closer to the pivot 0 than the right stimulus k,

may also be interpreted as the judgement that the impairment in stimulus k is greater than the impairment in stimulus i. In effect, this amounts to a response to a regular pair comparison, and the probabilistic model for the decision random variable simply becomes

$$\Pr(Z_{i0k} > 0 \mid \boldsymbol{\mu}) = \Phi(\mu_k - \mu_i). \qquad (6)$$

The difference from the normal interpretation of a triplet comparison is that here the impairment of the pivot is fixed to be equal to 0. In the general triplet comparison, however, all stimuli are modelled as random variables.

B. MAXIMUM LIKELIHOOD ESTIMATION OF THE MEANS
For the actual reconstruction by the maximum likelihood method we take as input a finite multiset T of annotated triplets, (i, j, k, R_{ijk}), where $R_{ijk} \in \{0, 1\}$ is the response to the triplet comparison (i, j, k) as in the above part. In subjective quality assessments with triplet comparisons, each triplet may be presented multiple times, collecting a response each time. Thus, T may contain multiple copies of both, $(i, j, k, 0)$ and $(i, j, k, 1)$. To keep the notation simple, we trust that it is clear from the context what R_{ijk} refers to in each case. Assuming that the responses are independent, we have that the negative log-likelihood of this data under the model assumptions is given by

$$L(\boldsymbol{\mu}) = - \sum_{(i,j,k,R_{ijk}) \in T} p^{R_{ijk}} (1-p)^{1-R_{ijk}}, \qquad (7)$$
$$p = \Pr(Z_{ijk} > 0 \mid \boldsymbol{\mu}).$$

The MLE estimate of the latent variable then is given by

$$\hat{\boldsymbol{\mu}} = \underset{\boldsymbol{\mu}=(\mu_0,...,\mu_M)}{\arg\min} L(\boldsymbol{\mu}).$$

In our experiments, we allowed a third option for triplet question responses, namely an answer *not sure* (see Subsection V-B). To account for such undecided responses, we simply assign the value $R_{ijk} = 1/2$ to the corresponding triplets (i, j, k) and use Equation (7) as given.

There are many algorithms for such nonlinear optimization problems, and generally, there is no guarantee that the global maximum will be attained. In our computations, we used the "fmincon" function in MATLAB, which is a nonlinear programming solver that finds the minimum of a constrained nonlinear multivariable function.

Solutions to this optimization are unique up to an additive constant. This constant can be chosen arbitrarily, and we have used this option to align all reconstructions such that the reconstructed scale value for the undistorted reference stimulus I_0 is $\mu_0 = 0$. These reconstructions of impairment scales are not yet in JND units because their probabilistic model assumed Gaussian distributions with the variance of 0.5. Since the JND unit corresponds to a value of $\Phi^{-1}(0.75) \approx 0.6745$, we divide the results by that to obtain impairment scales in JND units.
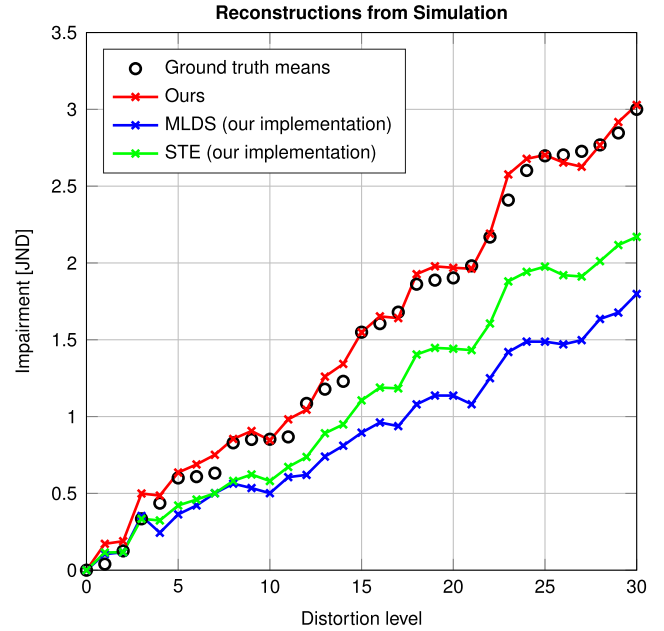
FIGURE 6. Reconstructions from a sample of 20 000 simulated random triplet question responses for 31 stimuli, spread over a range of 3 JND units. All three reconstruction methods yielded excellent correlations of 0.99, see Table 3, but only our reconstruction also reproduced the correct range of the means.

C. COMPARISON OF TRIPLET RECONSTRUCTION ALGORITHMS
Given that there are a number of available methods for the reconstruction of latent scale values from triplet comparisons, the question arises of which of them is the most suitable one to process subjective responses to triplet comparisons for visual quality assessment. For this purpose, we consider in this subsection simulated responses to triplet comparisons based on the Thurstonian model, i.e. the impairment scale of a distorted image is given by a normally distributed random variable with a corresponding mean and variance equal to 1/2.

For the reconstruction by MLE, we compute the likelihoods from one of the equations (3), (4), and (5), corresponding to our proposed reconstruction, MLDS, and STE, respectively. We expect that our proposed method gives the most accurate approximations because Equation (3) is directly derived from the Thurstonian model and the others are not. However, in terms of time complexity, each evaluation of the decision probability $\Pr(Z_{ijk} > 0 \mid \boldsymbol{\mu})$ requires two evaluations of the normal CDF, while MLDS needs only one, and STE none.

For baseline triplets of the form $(i, 0, k)$, where the pivot is given by the undistorted source image I_0 as a reference, we may interpret the response also as a response to a traditional pair comparison (i, k). In this case, we can also apply the usual Thurstonian reconstruction method for pair comparison.

For our simulation, we firstly considered an artificial sequence of 31 stimuli as the ground truth, with impairments on the perceptual scale, ranging from $\mu_0 = 0$ to $\mu_{30} = 2.0235$, which corresponds to $2.0235/\Phi^{-1}(0.75) \approx 3$ JND.

TABLE 3. Simulation for 31 stimuli over a range of 3 JND. Correlation with ground truth, and range of reconstructed means, averaged over 1000 repetitions.

Triplet responses per sample	MLDS ($\sigma = 1$)		STE ($\alpha = 1$)		Ours	
	SROCC	Range (JND)	SROCC	Range (JND)	SROCC	Range (JND)
1000	0.922 ± 0.030	2.055 ± 0.429	0.917 ± 0.053	2.214 ± 0.402	0.913 ± 0.064	3.153 ± 0.652
2500	0.964 ± 0.012	1.885 ± 0.257	0.967 ± 0.010	2.187 ± 0.211	0.967 ± 0.010	3.068 ± 0.326
5000	0.979 ± 0.008	1.840 ± 0.177	0.980 ± 0.006	2.185 ± 0.137	0.981 ± 0.006	3.050 ± 0.215
10 000	0.987 ± 0.005	1.808 ± 0.126	0.988 ± 0.004	2.171 ± 0.097	0.988 ± 0.004	3.024 ± 0.151
20 000	0.992 ± 0.003	1.797 ± 0.090	0.993 ± 0.003	2.168 ± 0.067	0.993 ± 0.003	3.015 ± 0.105
	MLDS ($\sigma = 1.6594$)		STE ($\alpha = 0.5316$)			
20 000	0.992 ± 0.003	2.989 ± 0.150	0.993 ± 0.003	2.974 ± 0.092		

We randomly sampled from the uniform distribution on the interval $[0, \mu_{30}]$ to obtain the remaining 29 stimulus means.

The triplets (i, j, k) were randomly sampled with the constraint $i \neq j \neq k \neq i$. The baseline triplets $(i, 0, k)$ were chosen randomly with $i, k \neq 0$ and $i \neq k$. In five rounds we drew 1000 to 20 000 triplets of each kind and generated one response per triplet according to the Thurstonian probabilistic model. We then applied all triplet reconstruction methods to the responses to triplets of the general type (i, j, k). For the baseline triplets $(i, 0, k)$ we carried out the reconstruction according to pair comparison and our proposed reconstruction. We repeated the procedure 1000 times.

The results of our simulation are presented in Figure 6 and Table 3. As expected, it is confirmed that our reconstruction does faithfully reconstruct the ground truth means from the triplet comparisons that were generated by the same Thurstonian probability model that underlies our reconstruction method. For baseline triplet comparisons, the reconstruction for pair comparisons from 20 000 ratings gave very similar results in terms of correlation as the reconstruction for our triplet comparisons, an SROCC of 0.996 (not shown in the table).

Even more notable are the findings that the other algorithms, MLDS and STE, also produced excellent results in terms of the Pearson linear as well as the Spearman rank-order correlation. However, the reconstruction ranges are around 2 JND, thus, well below the correct 3 JND.

To calibrate the STE and MLDS methods to give results in JND units, one could tune their parameters α and σ such that the range of reconstructed impairment values is equal to 3. For our simulation, we applied the bisection method to determine these parameters and obtained $\alpha = 0.5316$ and $\sigma = 1.6594$. The resulting correlations are excellent again (see Table 3). The RMSE over all 30 reconstructed impairments are 0.1089 for MLDS and 0.0646 for STE, while for our method (without tuning a parameter), we obtained an RMSE of 0.0520.

Of course, the above procedure is not feasible in general because the ground truth range is not known as in our simulation here. One would have to resort to an estimation of a suitable parameter α for MLDS or σ for STE. To this end, we propose two approaches.

1) Estimate the range of the expected scale values. In our simulation, it was 3 JND, for example. Then proceed as in our simulation. Randomly choose a sequence of scale

values in the selected range, and then tune the parameter α, respectively σ, to achieve a reconstruction by MLDS, resp. STE, to match the selected range.

2) Minimize the mean square error of the MLDS probabilities for a response $R_{ijk} = 1$ by selecting $\hat{\sigma}$,

$$\hat{\sigma} = \arg\min_{\sigma > 0} \int_{-\Delta}^{\Delta} \int_{-\Delta}^{\Delta} |e(r, s \mid \sigma)|^2 \, dr \, ds$$

where the error $e(r, s \mid \sigma)$ is given by

$$\Phi\left(\frac{|s| - |r|}{\sigma}\right) - 1 + \Phi(s) + \Phi\left(\frac{r + s}{\sqrt{3}}\right) - 2\,\Phi(s)\,\Phi\left(\frac{r + s}{\sqrt{3}}\right).$$

Here, $r = \mu_i - \mu_j$ and $s = \mu_k - \mu_j$ denote the left and right differences of impairments in the triplet (i, j, k), ranging over the square domain $[-\Delta, \Delta] \times [-\Delta, \Delta]$. Similarly, one can do the same for STE. In Figure 7, we visualize the probability functions according to the Thurstonian model along with their approximation in the MLDS and STE methods and the corresponding errors $e(r, s \mid \sigma)$ resp. $e(r, s \mid \alpha)$. Inspecting this figure, it becomes apparent that the globally optimal parameter σ will be difficult to obtain as it strongly varies locally.

In summary, all three methods produced excellent results. For baseline triplet comparisons, reconstruction by the traditional Thurstonian approach (with MLE) was as good as our method for triplet construction. If one needs to have results on the perceptual scale given in JND units, then our proposed reconstruction should be applied.

V. EXPERIMENTAL SETUP: MATERIALS AND PROCEDURES

The purpose of our experimental studies was to investigate the potential and limitations of the proposed boosting strategies in the application of subjective full-reference image quality assessment. In current FR-IQA datasets, the main approaches have been DCR and PC with the reference image shown additionally, i.e., a case of baseline triplet comparison (Table 1). For both of these, boosting of the underlying image distortions can be applied. Thus, we carried out three main experiments, starting out with baseline triplets, which we then extended to general triplets, finally followed by a smaller study for DCR. In the following, we refer to these as Experiments I, II, and III.
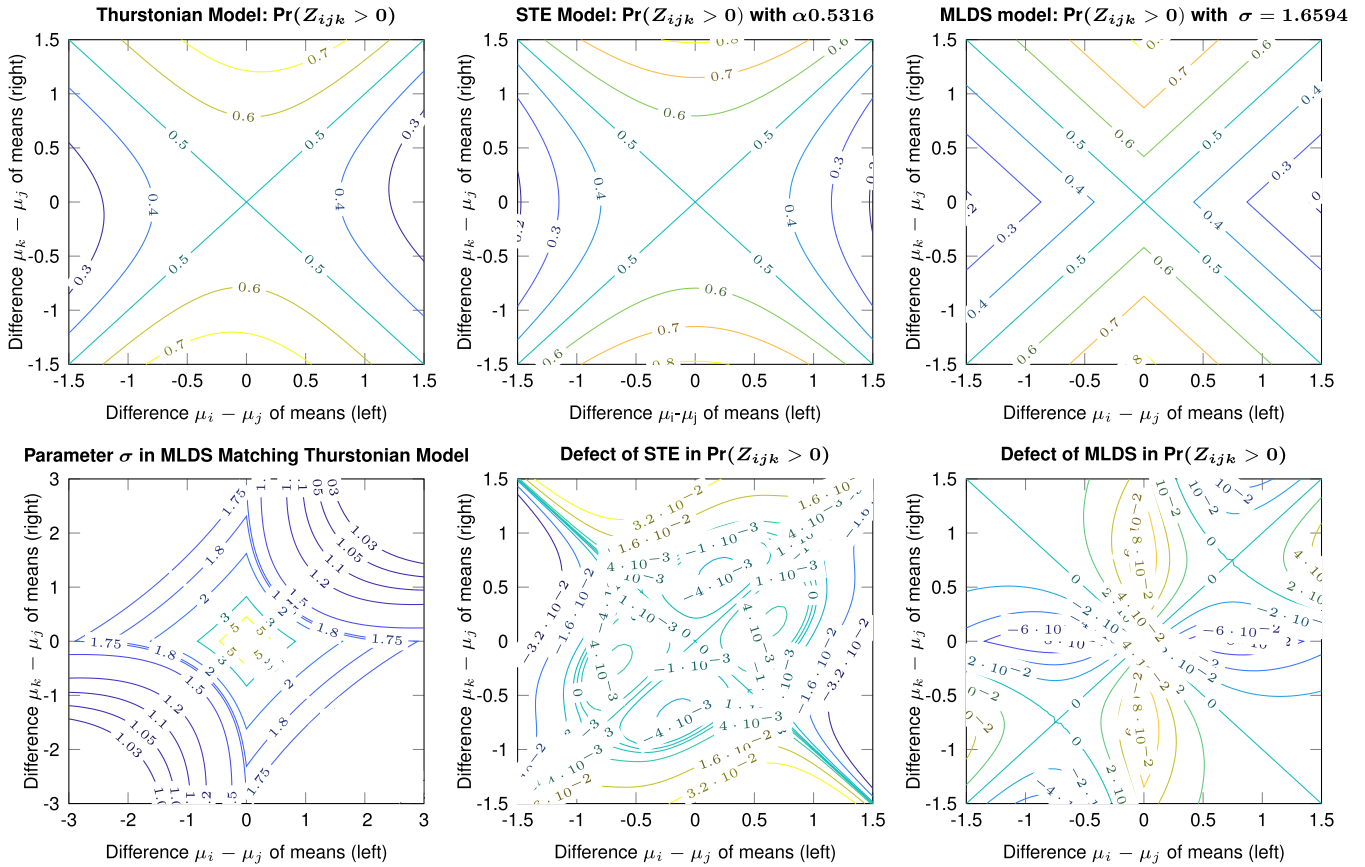
FIGURE 7. The top row shows plots of isolines of the probabilities Pr($Z_{ijk} > 0 \mid \mu$) for the Thurstonian (left), STE (middle), and MLDS (right) models. The parameters $\alpha = 0.5316$ for STE and $\sigma = 1.6594$ for MLDS were obtained by ensuring that the range of the corresponding reconstructed scales is equal to 3 JND. The bottom row (center and right) shows the difference between the resulting probabilities of STE resp. MLDS and those of the Thurstonian model. The model fit for STE is much closer to the Thurstonian reference than that of MLDS. The bottom right plot shows the parameter σ for MLDS that locally yields equality with the Thurstonian probabilities.

In order to evaluate aspects like accuracy, reliability, and convergence, a large number of comparisons is beneficial. Therefore, our subjective IQA experiments were conducted via crowdsourcing. For the study, a set of original pristine source images, each distorted by various types of distortions, was selected. For each source image and each distortion type, a sequence of increasingly distorted images was generated. By means of a pilot study, we took care to calibrate our dataset such that each such image sequence uniformly spans a perceptual quality range of approximately 3 JND. In the following subsections, we briefly describe our setup and procedure to achieve these goals.

A. SUBJECTIVE CROWDSOURCED IQA STUDY

In terms of experimental methodology, lab studies are well established and considered reliable because the experimental environment can be controlled, and the whole procedure can be monitored. On the other hand, the number of images that can be assessed is limited due to the time requirements as well as the cost. Alternatively, crowdsourcing studies are more economical, more efficient, and more scalable, and can have sufficient reliability if the setup, with a quality control mechanism included, is appropriate [64] and a suitable outlier

removal strategy is employed [65]. We have installed several measures of control to ensure the validity of the results from our crowdsourcing campaigns, described in the following.

The experiments were carried out on the Amazon Mechanical Turk [66] platform, in which *requesters* create and submit their *human intelligence tasks* (HITs) for *workers* who carry out the subjective quality assessment. Workers receive a monetary reward for completing a HIT. Requesters specify the number of *assignments* for each HIT to control how many workers can submit work for that HIT.

In our experiments with triplet comparisons, a HIT consisted of 20 questions that gave rise to 20 (ternary) *responses* or *answers* from each crowdworker that completed an assignment for that HIT. For the experiment with degradation category ratings, HITs also had 20 questions each, and workers provided the corresponding *ratings* on a 5-point DCR quality scale.

B. INTERFACE

At the beginning of the experiment, a detailed instruction was shown to the crowd workers, after which they were allowed to start doing assignments.
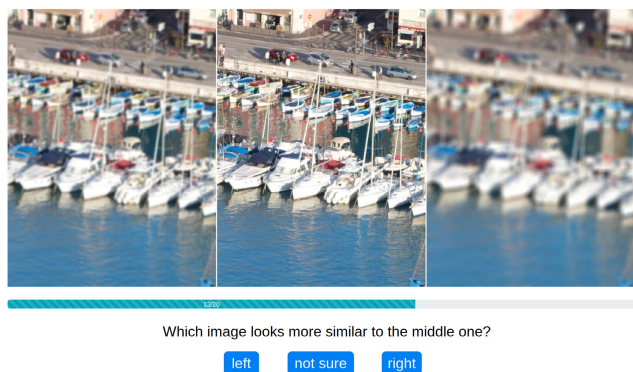
FIGURE 8. The interface of the triplet comparison experiments without a flickering effect. Crowd workers were asked to select the image that they think looks more similar to the pivot. They could choose "not sure" if they could not distinguish the differences. For the experiments with a flickering effect, the interface was the same as the one without a flickering effect shown here, except that three still images were replaced by two flickering images, and the question got changed to "Which image has a stronger flickering effect?".

- In the parts of Experiments I and II that used TC without a flickering effect (Plain, A-, Z-, and AZ-boosted TC), three images were displayed in a row, see Figure 8. Crowd workers selected the image that looked more similar to the pivot image in the middle by clicking "left" or "right". If they could not decide, a third choice, "not sure", was available. This option was introduced in subjective evaluations of the JPEG XS image compression [67] and had been found useful to reduce subject stress and fatigue. In an earlier study on the unforced-choice paradigm in applications in audiology, it was also concluded that the efficiency of pair comparison might be compromised when participants are forced to choose between stimuli [68].
- In the other parts of Experiments I and II that used TC with a flickering effect (F-, AF-, ZF-, and AZF-boosted TC), two flickering images were displayed side by side. Crowd workers selected the image with a stronger perceived flickering effect by clicking "left" or "right", or use the "not sure" option.
- For Experiment III using DCR, two images were displayed side by side, with the reference image on the left and the test image on the right, see Figure 9. Crowd workers rated the distortion of the test image on the 5-scale category ratings ranging from 0 (imperceptible) to 4 (very annoying).

For each of the 20 questions in one assignment, crowd workers had eight seconds to enter their responses. The images were shown only during the first five seconds. In case no answer was given by the crowd worker within the eight seconds, the response was labelled as "skipped". Thus, the total time for an assignment was 2m 40s.

C. HIT-LEVEL QUALITY CONTROL
The subjective assessment of image quality through crowdsourcing may pose some challenges due to the lack of control
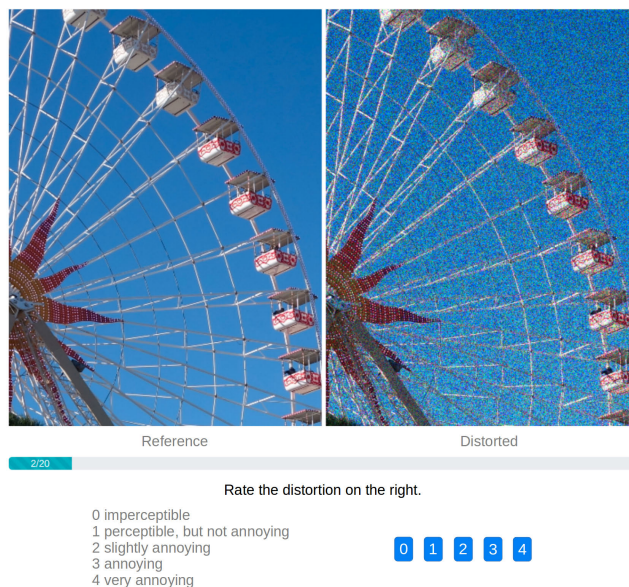
FIGURE 9. DCR Interface. Two images were placed side by side, with the reference image on the left. Crowd workers were asked to rate the distortion of the right image w.r.t. its reference on the left on one of the five categories: 0 (imperceptible), 1 (perceptible but not annoying), 2 (slightly annoying), 3 (annoying), 4 (very annoying).

over the experimental environment, lack of knowledge about the background of the workers, and limited reliability of the experimental results. Therefore, we need to detect and filter out low-quality responses. Unreliable responses may be caused by technical problems with the workers' screens and devices, a misunderstanding of the subjective task, e.g., limited English proficiency of some international workers, and a lack of attention. In addition, some of the workers may try to answer quickly to get the maximum payment in a shorter amount of time, resulting in responses of insufficient quality.

We ensured the quality of workers' answers in the crowdsourcing studies by monitoring the number of questions in each assignment that the workers skipped and including one hidden test question in each assignment. For example, to select suitable test questions for Experiment I, we proceeded as follows. Based on our pilot study (see Subsection V-H2 below), we first chose baseline triplets with distorted images on the left and right, for which the perceptual difference in distortion was relatively large, about 1.75 JND. By a following visual inspection, we then discarded triplet comparisons that did not seem to suggest a straightforward correct response. For Experiments II and III, we proceeded similarly.

If a worker skipped (did not answer) more than three questions in an assignment or the hidden test question was answered incorrectly, the assignment was rejected, and all of the responses were discarded. We did not pay the workers for their rejected assignments.

The responses from rejected assignments were re-collected by making the HITs available again for new workers. This rejection and re-collection procedure was carried out for three

rounds. We did not reject any assignments in the last, smallest re-collection round, regardless of the performance on test questions and the number of provided responses. By this procedure, we filtered out the low-quality responses at the assignment level and ensured that the desired number of assignments was achieved.

Accepted assignments might still contain outlier responses. In the next subsection, we describe the procedure for removing such outliers. The statistics of the rejected assignments and outliers for all experiments are provided in Table 5.

D. ROBUST HIT-LEVEL OUTLIER REMOVAL

During data collection in each experiment, unreliable assignments for HITs were discarded and penalized as described above in Section V-C. However, there may still have been uncovered unreliable data left that should be identified as outliers and removed before the reconstruction of quality scales. It seems inappropriate to classify individual responses to triplet questions as outliers since the answers are not on an interval scale but ternary ("left", "right", and "not sure"). Therefore, we considered outlier removal at the HIT assignment level, requiring a multivariate method. After the quality control during each experiment, each assignment carried 16–19 answers for the 19 triplet comparisons per assignment, not counting the single test question per assignment and allowing for skipping up to three questions.

To this end, we aim at a robust multivariate outlier detection method that flags a prescribed percentage of assignments that markedly differ from the consensus given by the remaining majority of assignments. A robust approach deviates from conventional ones in that the statistics used to identify outliers do not suffer from the influence of the outliers themselves.

The most common and recommended multivariate outlier detection method in this spirit is a fast version of the minimum covariance determinant (Fast-MCD) approach [69]. However, the MCD method operates with Mahalanobis distances that are not suitable in our case since the multivariate data are not only vectors of ternary decisions rather than being from a real Euclidean vector space, but also of variable dimension (16–19).

A similar approach was given by the k-means– algorithm [70]. This modification of the classical k-means clustering algorithm takes a desired number of outliers into account that is farthest from the cluster centers. Convergence of local optima was proven. Cluster centers are given by the means of the cluster data points. For our application, the method would have to be run for a single cluster ($k = 1$). However, this does not work since data from HIT assignments are not simply vectors of some vector space, and these data cannot be averaged.

To remove HIT assignments that are outliers, we propose an adaptation of the above two robust detection methods.

Firstly, we define the consensus of a subset of assignments for a HIT due to the reconstructed impairment scale values for all stimuli involved in the corresponding experimental study. This consensus replaces the cluster means as used in Fast-MCD and k-means–.

Secondly, we need an algorithm to compute the distance of each assignment to the consensus given by the impairment scales reconstructed from the corresponding majority of assignments. A small distance should indicate that the responses collected in a HIT assignment agree well with the reconstructed impairments. Large distances suggest strong disagreement with the consensus and that the corresponding HIT assignments may be regarded as outliers.

1) DISTANCE TO CONSENSUS FOR AN ASSIGNMENT OF TRIPLET COMPARISONS

For the case of triplet comparisons, we define this distance for an assignment as the complementary weighted true positive rate of the corresponding N responses with respect to the given consensus as follows. Considering the n-th response to a triplet question of type (i, j, k) for a particular reference image I_0, a given distortion type, and corresponding distorted images I_i, I_j, I_k that make up the triplet question, we compare it with the corresponding impairment scale values $\hat{\mu}_i, \hat{\mu}_j, \hat{\mu}_k$ from the current consensus. If the answer "left" (resp. "right") for this n-th triplet question is in accordance with the consensus, we assign a score of value $v_n = 1$ to it. The answer "not sure" earns a score of $v_n = 0.5$. The following table completes the definition of the score v_n for the response to the n-th triplet question (i, j, k).

Response	$|\hat{\mu}_k - \hat{\mu}_j| \geq |\hat{\mu}_i - \hat{\mu}_j|$	$|\hat{\mu}_k - \hat{\mu}_j| < |\hat{\mu}_i - \hat{\mu}_j|$
left	1	0
right	0	1
not sure	0.5	0.5

The difficulty of triplet questions varies according to the difference between the left and right differences of impairment scales, $D_l = |\hat{\mu}_i - \hat{\mu}_j|$ and $D_r = |\hat{\mu}_k - \hat{\mu}_j|$. If $D_r \approx D_l$, the decision which is perceptually the smaller one is hard. In this case, an answer that disagrees with the consensus should not be penalized severely. On the other hand, if $|D_r - D_l|$ is large, a wrong decision should be penalized more strongly. Therefore we introduce the weight

$$w_n = |D_r - D_l| = |\hat{\mu}_k - \hat{\mu}_j| - |\hat{\mu}_i - \hat{\mu}_j|$$

for the n-th response and define the distance of the assignment w.r.t. the consensus as

$$d = 1 - \frac{\sum_{n=1}^{N} w_n v_n}{\sum_{n=1}^{N} w_n}. \tag{8}$$

We have that $0 \leq d \leq 1$, and the maximal distance of $d = 1$ implies that all responses in that assignment were against the consensus of the majority.

2) DISTANCE TO CONSENSUS FOR AN ASSIGNMENT OF DCR QUESTIONS

In case of an assignment of DCR questions, the consensus produced by a subset of HIT assignments is given by the

FIGURE 10. The (cropped) source images for our experiment. From upper left to lower right: source images with a resolution of 512 × 384 cropped from the images in MCL-JCI dataset with the following IDs: SRC01, SRC03, SRC06, SRC07, SRC09, SRC17, SRC28, SRC31, SRC45, and SRC50, respectively. The green inset rectangles (256 × 192 pixels) indicate the regions used for the boosting methods with zooming.

corresponding DMOS values for all test images involved in the experiment. In a HIT assignment with ratings r_n for pairs $\left(I_0^n, I_{k(n)}^n\right)$ and corresponding DMOS values $\hat{\mu}\left(I_{k(n)}^n\right)$, $n = 1, \ldots, N$, from the consensus, we define the distance of the given ratings to the consensus as the mean of the absolute differences to the consensus,

$$d = \frac{1}{N} \sum_{n=1}^{N} \left| r_n - \hat{\mu}\left(I_{k(n)}^n\right) \right|. \tag{9}$$

With these definitions made, we now give the iterative, robust outlier removal algorithm.

1) Input: M HIT assignments with responses, a target $L < M$ of assignments to be kept.
2) Start with the subsample of all M HIT assignments.
3) Compute the consensus of the subsample: the reconstruction of all impairment scales.
4) Compute the distances of all M assignments from the consensus with Equation (8), resp. (9).
5) Choose the L smallest distances and create a new subset.
6) Repeat steps 3 to 5 until convergence (the new subset is the same as the old one) or a timeout.

In our experiments, we removed a fraction of 5% of HIT assignments as outliers and observed convergence in just 4–7 iterations.
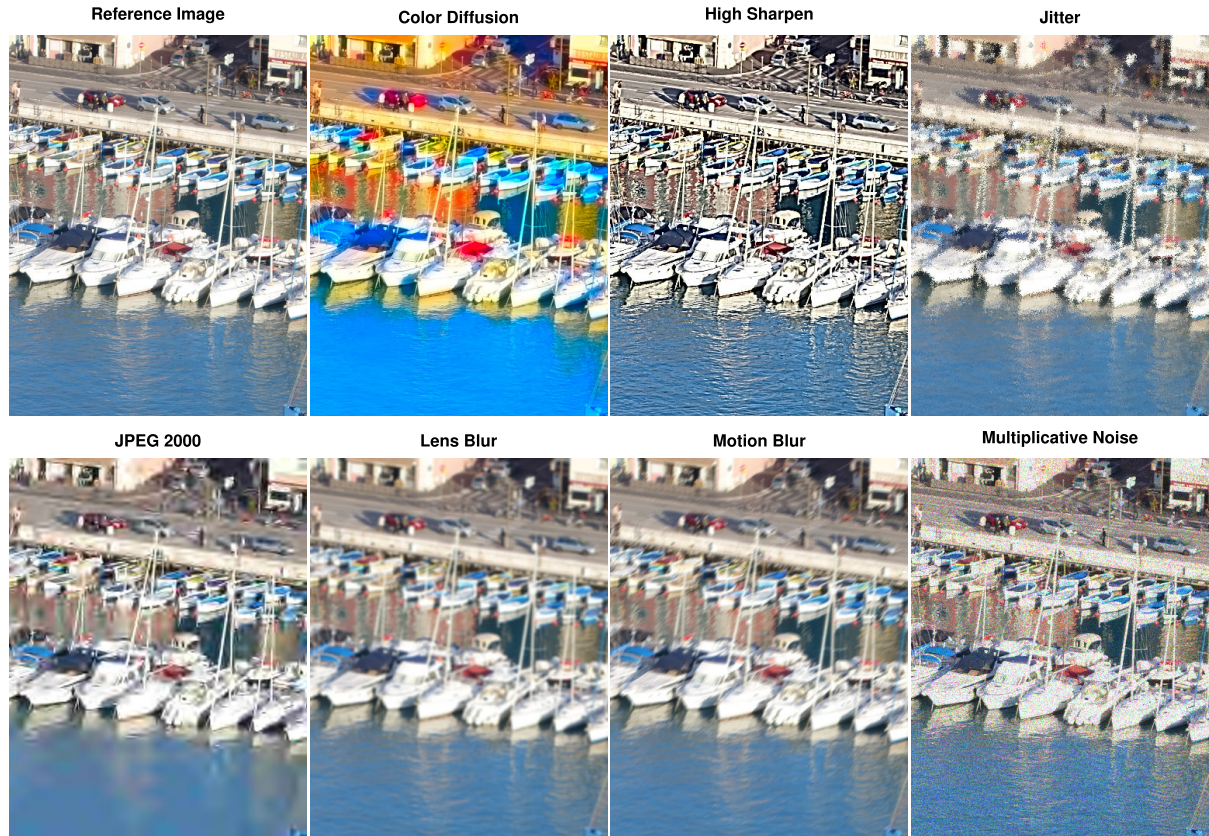
E. SOURCE IMAGES

Ten source images were selected from the MCL-JCI dataset [29], whose original resolution is 1080 × 1920. In our

subjective study, the original resolution is too large to display on the screens of crowd workers. To ensure that a triplet can be displayed without image re-scaling, we manually cropped each image to 512 × 384 pixels. We chose to crop portrait-mode subimages because triplets of such images better utilize screen space. We further cropped the images to 256 × 196 pixels for experiments with boosting by zooming and subsequently upscaled them back to 512 × 384 pixels for display. Figure 10 shows the ten (cropped) source images and their parts used for zooming.

F. DISTORTION TYPES

For our validation experiments, the source images were degraded by seven distortions, selected from [21], [71], [72]. All distortions were implemented in MATLAB, with the source code made available by the authors.

- **Color Diffusion** converts an image from RGB to CIELAB color space, where a 2D Gaussian smoothing kernel is used to blur the a and b channels. Its distortion magnitude is determined by the standard deviation of the kernel.
- **High Sharpen** applies the unsharp masking method to sharpen an image. An image is sharpened by subtracting a blurred unsharp) version of the image from itself. Its distortion magnitude is determined by the parameter of the strength of the sharpening effect.
- **Jitter** warps an image according to two matrices describing random local shifts of each pixel in horizontal

FIGURE 11. One of the source images (upper left) is distorted by each of the seven considered types of distortion. The degree of distortion is the largest used in this study. It corresponds to the third JND w.r.t. the source image.

and vertical directions. The amount of distortion is determined by the magnitude of the shift.

- **JPEG 2000** is an image compression standard with the distortion magnitude determined by the compression ratio.
- **Lens Blur** performs spatial 2D filtering on each color channel of an image with a circular kernel, whose distortion magnitude is determined by the radius of the kernel.
- **Motion Blur** performs spatial 2D filtering on each color channel of an image with a kernel oriented at 45 degrees. The distortion magnitude is determined by the size in the direction of motion.
- **Multiplicative Noise** adds speckle noise to an image I, obtaining $I + n \odot I$, where $n(x, y)$ are i.i.d. random variables, uniformly distributed with zero mean and an adjustable variance. The magnitude of the distortion is determined by the variance.

Figure 11 shows one of our source images together with distorted images for all of the seven distortion types.[2]

G. DISTORTION LEVELS: DESIGN

For each combination of a source image and a distortion type, a sequence of increasingly distorted images is to be defined.

[2]In a pilot study prior to this work (unpublished), we used JPEG compression, Gaussian noise, and Gaussian blur. The results showed that the boosting strategies (A-,Z-, and AZ-Boosting) work well also for these types of distortions.

In previous FR-IQA datasets, only 4–6 levels of distortion were considered (Table 1). The boosting techniques are expected to help differentiate between distorted images that look the same at first glance. Thus, for our first main experiment, we strove to generate image sequences such that the perceptual difference of any two successive images is fixed at only 0.25 JND. This is a very small difference: According to the Thurstonian probabilistic model, in a 2AFC experiment, the fraction of observations that correctly identify the stimulus with better quality is $\Phi(0.25 \, \Phi^{-1}(0.75)) \approx 0.5670$. The corresponding detection rate is only $2 \cdot 0.5670 - 1 \approx 0.1339$. Our second main experiment shrunk the spacing even more, to 0.1 JND, corresponding to an expected detection rate of merely 0.0538. However, in this case, we considered only one particular type of distortion to keep the overall cost within our budget for the crowdsourcing.

Having defined the psychovisual spacing of distortion in each image sequence, the question remained over what range of distortions the sequence should span. In a recent study on just noticeable differences in video sequences, it was found that the perceptual quality at the third JND is between fair and poor on the 5-point ACR scale [32]. The authors concluded that distortions stronger than 3 JND are not acceptable by today's viewers. Thus, we set the range of impairment for each image sequence to 3 JND. For content providers of high-quality media, we believe the first third of this range, from lossless compression up to 1 JND is the most relevant.

Therefore, in our first main experiment, an image sequence consisted of the pristine, original image together with 12 distorted versions at impairments of $0.25\,k$ JND, with $k = 1, \ldots, 12$. In the second experiment, we had even 30 distorted images uniformly ranging over 3 JND. In many of our figures, we show results over these 12, respectively 30, distortion levels. Since all reference source images and all distortion types test images at the same distortion level correspond to nearly the same perceived magnitude of distortion, we can average results over these image sequences for each distortion level.

H. TEST IMAGE GENERATION

In order to generate sequences of distorted images with the desired equal spacing on the perceptual scale, we carried out a pilot study. The overall procedure for each of the 70 image sequences (10 source images, 7 distortion types) was as follows:

1) Generate a sequence of 12 distorted images by increasing the corresponding scalar distortion parameter λ. The parameters are chosen using the method of bisection to yield an image sequence that is equally spaced according to the structural similarity index SSIM [73].

2) Run a crowdsourcing study using pair comparison with the additional display of the corresponding source image, which is equivalent to the baseline triplet comparison strategy.

3) Reconstruct the impairment scales from the comparison data in JND units. Calibrate the scales such that the impairment for the pristine source image is equal to 0. The result is a sequence of impairments, parametrized over the corresponding distortion parameter λ for the given distortion type.

4) Truncate this sequence such that only the last remaining impairment value is outside of the range from 0 to 3 JND. Fit a straight line to these data points, constrained to pass through the point for the source image. Without loss of generality, we may assume that this is the origin, $(0, 0)$. Let the slope of the line fit be $s > 0$. Then the expected impairment at parameter λ according to the line fit is $s\lambda$.

5) Define the sequence of distortion parameters as

$$\lambda_k = \frac{k}{4s}, \quad k = 0, \ldots, 12.$$

6) Generate the sequence of distorted images according to these parameter settings. As a result, the impairment scale of the k-th image will be approximately $0.25\,k$ JND, and the distortions span a total range of about 3 JND.

Figure 12 shows an example of impairment reconstructions for one image sequence together with the line fit and the resulting choices of 12 physical distortion parameters, which we can expect to correspond to distorted images uniformly spaced 0.25 JND apart from each other.

For the second experiment the intended spacing in impairment is 0.1 JND, and the procedure is the same as above,

except for Step 5, where we set $\lambda_k = \frac{k}{10s}$, $k = 0, \ldots, 30$. This was done for 10 image sequences from the 10 source images, but only one distortion type, motion blur.

In the rest of this subsection, we briefly describe the details of the selection of the baseline triplet comparisons in the pilot study, the numbers of collected responses, and the quality control.

1) SAMPLING STRATEGY

Our sampling strategy proceeded in three rounds of data collection. The first round was for initialization. For each of the 70 image sequences, we randomly sampled pairs of the 13 test images (including the source image) by choosing the edges of a random sparse graph with a vertex degree of six and nodes corresponding to images. Thus, each image was randomly compared to six other images of the same sequence. For all sequences together, this resulted in $39 \times 70 = 2730$ triplets of images. We used baseline triplet comparisons, so the pivot image in the center of the triplets was fixed as the corresponding source image.

In the second and third rounds, we applied an active sampling strategy. A minimal spanning tree connecting 13 nodes (i.e., test images) was produced for each sequence. The 12 edges then yielded the test images for the triplet questions and were chosen based on maximizing the expected information gain, following [74]. This resulted in $12 \times 70 = 840$ triplets of images in each round.

Each triplet question was presented to 30 crowd workers. Hence, overall $(2730 + 2 \times 840) \times 30 = 132\,300$ responses were collected.

2) QUALITY CONTROL

The quality of the experiment was controlled as described in Section V-C and by a simplified outlier removal process. There were 1035 assignments that were rejected and recollected because the test question was answered incorrectly, or more than three questions were skipped.

The outlier detection in the pilot study was chosen differently from that in the main experiment. This is because the pilot study's purpose was merely to help generate test image sequences with prescribed impairment scales. This required accurate reconstructions of these scales for the test images used in the pilot study. To identify HIT assignments with unreliable responses, we, therefore, could rely on the ground truth, given by the ordering of the test stimuli on the physical distortion scale.

Consider the n-th baseline triplet comparison $(i, 0, k)$ of a HIT assignment, where i and k denote the indices of the test images in a given sequence. Similar to Subsection V-D1, we gave a score $u_n \in \{0, 0.5, 1\}$ to its response as specified in the following table.

Response	$i < k$	$k < i$
left	1	0
right	0	1
not sure	0.5	0.5
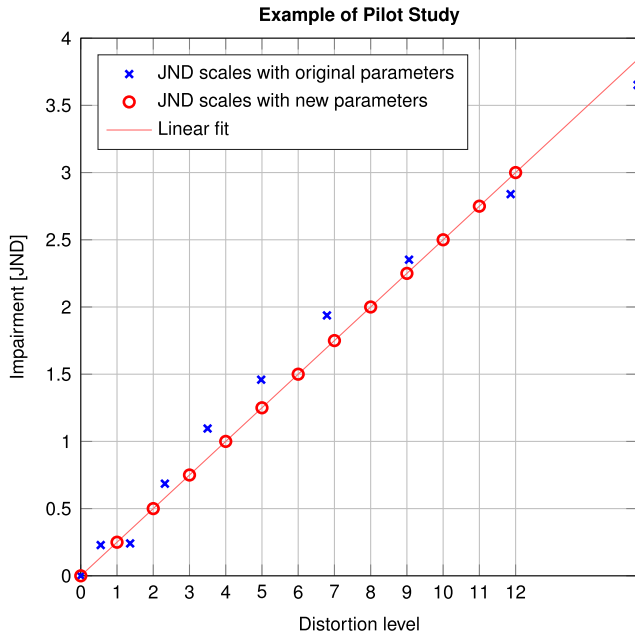
FIGURE 12. Example of the linear fitting procedure in the pilot study for the image sequence with source image SCR03 and distortion type "High Sharpen". On the horizontal axis, the physical distortion parameter for the selected distortion type increases linearly. 13 impairment values (blue crosses) were reconstructed, the last 4 of them being greater than 3 JND. The last 3 of these are disregarded and not shown. From the remaining 10 points, the line fit was generated. Using this result, 13 equally spaced physical distortion parameters, as shown, were obtained, which span the range of 3 JND on the perceptual impairment scale. These parameters define the twelve distortion levels for the image sequence used in Experiments I and III. For this example, a difference of one distortion level corresponds to a difference of 0.322 for the physical distortion parameter of the type "High Sharpen".

Note, that by construction, $i \neq k$. The normalized sum of the scores in an assignment, $\frac{1}{N} \sum_{n=1}^{N} u_n$ can be called the true positive rate. The distance of an assignment w.r.t. the ground truth ordering is defined as

$$d = 1 - \frac{1}{N} \sum_{n=1}^{N} u_n. \tag{10}$$

After computing these distances to ground truth ordering for all assignments, we sorted them according to distance and removed the last 10% of them as outliers.

I. COMPARISON OF RESOLUTIONS OF FR-IQA DATASETS

Our dataset is the first one designed based on perceptual criteria that had been assessed in a pilot study. For each combination of a source image and a distortion type, the goal was to generate a sequence of distorted images with equal increments of perceived impairment. In Part A of our dataset, this increment is 0.25 JND, and in Part B, it is 0.1 JND. In other datasets, the distortion parameters either were selected manually (e.g., LIVE, VCL@FER, CID:IQ, MDID, and KADID-10k) or according to a plan w.r.t. increments of PSNR or bitrate (e.g., TID2008 and TID2013).

For an image sequence in a fine-grained FR-IQA dataset, we expect a large number of distortion levels spread over the
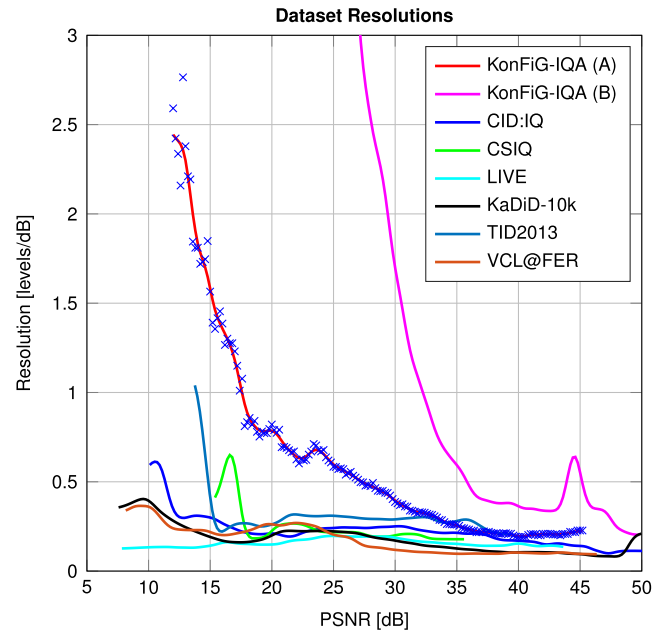

FIGURE 13. The resolution of FR-IQA datasets, defined by the number of distortion levels per dB on the PSNR scale, generally varies between 0.2 and 0.5 levels/dB for conventional datasets. Our new fine-grained datasets, KonFiG-IQA, Parts A and B, have much greater resolution over a large portion of the entire PSNR range. The data points (black crosses) are samples of the resolution function, computed for KonFiG-IQA, Part A. The curves in the figure were obtained by a Gaussian averaging filter of width 2 dB.

respective ranges of distortion, in our case 3 JND. To quantify and compare different FR-IQA datasets in this regard, we introduce the notion of *dataset resolution*. As a reference scale, we adopt the PSNR in dB. The resolution for an FR-IQA dataset then is a function of the PSNR, which indicates how close PSNRs of consecutive images from image sequences are in the neighbourhood of the given PSNR. We chose to define resolution locally this way because of the nonlinear relationship between PSNR and perceived quality which causes the resolution function to vary significantly, particularly for our perceptually guided FR-IQA datasets.

TID2013 adopted a spacing of 3 dB PSNR between consecutive images in a sequence, so its dataset resolution is 1/3 level per dB PSNR, in this case uniformly over the entire range of distortion. A resolution of 0.4 levels/dB at 28 dB PSNR means that the average length of intervals of PSNR values of consecutive images in a sequence including 28 dB, is 2.5 dB, the inverse of the resolution of 0.4 levels/dB.

We computed the dataset resolution functions for our datasets, KonFiG-IQA Parts A and B, and six other datasets, namely LIVE, CSIQ IQA, VCL@FER, TID2013, CID:IQ, and KADID-10k. For each of these, the procedure was as follows. First, we scanned consecutive images in each image sequence of the FR-IQA dataset and collected the corresponding PSNR intervals. Secondly, we sampled the PSNR scale uniformly with a step size of 0.2 dB, and for each sample value, we averaged the lengths of all those intervals that contain the PSNR sample. The inverse of this
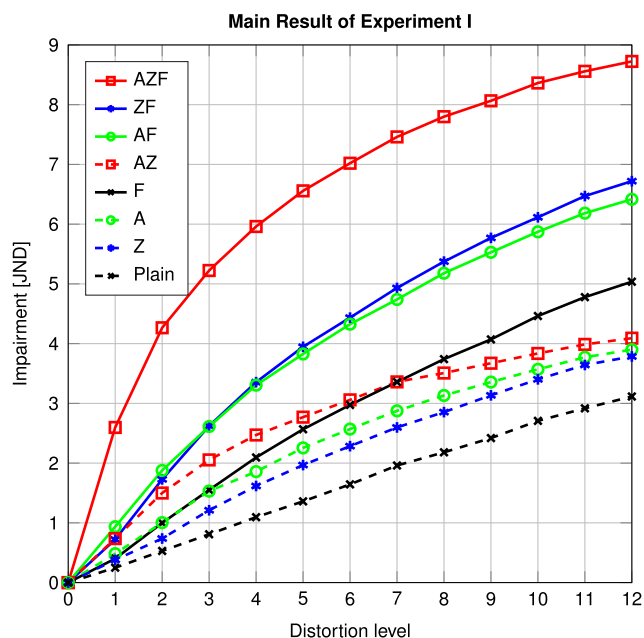
FIGURE 14. Average reconstructed impairment scales over 70 sequences for 8 types of TC with baseline triplets. Scales from AZF-boosted TC have the largest range, almost 3 times as large as the range of Plain TC.

average length is the value of the resolution function at the given PSNR sample value. Due to the discrete nature, the resolution function is only piecewise constant and appears noisy as there are some intervals of very small size. For visualisation, therefore, we show a smooth approximation obtained using a Gaussian averaging filter with a width of 2 dB PSNR.

Figure 13 shows the resulting resolution functions for the selected datasets and also the samples collected for KonFiG-IQA (Part A). The figure confirms that our new fine-grained datasets have much greater resolution over a large portion of the total PSNR range.

We also note that for TID2013, the design goal to have 3 dB PSNR between consecutive images in each sequence (0.33 levels/dB) was not strictly followed. The dataset resolution mainly varies between 0.2 and 0.4 levels/dB and exceeds 1 level/dB at the low end of the PSNR scale.

VI. EXPERIMENT I: BOOSTING FOR BASELINE TRIPLET COMPARISON

The purpose of our first experiment is to apply the proposed boosting methods to our FR-IQA dataset in a crowdsourcing campaign for the analysis of the performance w.r.t. that achieved without boosting as traditionally done. Here we use baseline triplet comparisons, i.e., we present two different distorted test images, I_i and I_k, together with the undistorted reference image I_0 in the middle of the triplet, which is denoted by $(i, 0, k)$. This interface corresponds to that used in TID2008, TID2013, and MDID.

We recall from the previous section that we have 10 source images, each distorted by 7 types of distortion, giving 70 image sequences, each consisting of a reference (source)

image I_0 and 12 increasingly distorted versions of the reference image, $I_1, \ldots, I_{12}$. By design, the distortions span 3 JND units on the perceptual scale, so the perceptual difference between two consecutive images is 0.25 JND. For each of the 70 sequences, we applied 8 types of baseline triplet comparisons, namely the plain TC without boosting and A-, Z-, AZ-, F-, AF-, ZF-, AZF-boosted baseline TC.

In this setup, for each sequence, there are $\binom{13}{2} = 78$ possible triplet comparisons. Since there is less information gain in responses for triplets for which the correct answer is obvious, we only considered the 68 triplets $(i, 0, k)$ with $|k - i| \leq 8$ and $k \neq i$. Thus, triplets $(i, 0, k)$ with a perceptual distance between I_k and I_i greater than 2 JND were omitted. In this way we generated two groups of $68 \times 70 \times 4 = 19\,040$ triplets each. The first group contained those triplets that were used for comparisons without boosting by flicker. For the other group of triplets, to be used with F-boosting, a different interface had to be applied (see Section V-B), so they were collected in separate HITs in the crowdsourcing. The triplets were randomly oriented (either as $(i, 0, k)$ or $(k, 0, i)$) and shuffled in each group. Finally, they were split up and distributed into crowdsourcing HITs. Each HIT consisted of 19 triplet comparisons and 1 additional triplet comparison from our pool of test questions.

We spawned 20 assignments per HIT, i.e., we collected 20 responses for each triplet comparison. We controlled the quality of the experiment and removed outliers as described in Sections V-C and V-D. In the end, $37\,206$ assignments of HITs with $706\,914$ TC responses remained for the reconstruction of impairment scales for 70 image sequences and analysis. For details see Table 4 and 5. Using the method proposed in Section IV, the perceived impairments of the images were reconstructed from the set of TC responses. We evaluated and compared the performances of the eight types of TC by several criteria as follows.

A. RESULTS: RECONSTRUCTED IMPAIRMENT SCALES
Figure 14 summarizes the reconstructed impairment scales for stimuli at the 12 distortion levels. Each curve represents the average taken over all 70 image sequences of seven distortion types. The main findings from this global view over the range of 3 JND are as follows:

- The result from plain triplet comparisons is given by the black dashed curve. It shows a linearly increasing impairment, reaching 3.1 JND for distortion level 12. Thus, Experiment I confirms our expectations from the pilot study quite well, a perceptually linearly increasing impairment over 3 JND.
- The three colored dashed curves are for the results without boosting by flicker. They are well above the baseline given by plain TC and extend the range of perceived distortion from 3 to about 4 JND, with the red curve for combined AZ-boosting yielding the largest increase. Thus, for boosting with artefact amplification, zooming,

TABLE 4. Overview of all subjective studies.

	Sources	Distortion types	Distortion levels	Boosting types	Responses/triplet Ratings/DCR	Reponses/ratings per sequence	Total number of reponses/ratings	Reponses/ratings after outlier removal
Pilot study	10	7	13	Plain	30	1890 $((39 + 12 + 12) \times 30)$	132 300 $((39 + 12 + 12) \times 30 \times 70)$	119 092
Experiment I	10	7	13	Plain, A, Z, AZ F, AF, ZF, AZF	20	1360 (68×20)	761 600 $(68 \times 70 \times 8 \times 20)$	706 914
Experiment II Plain TC	10	1	31	Plain	9	29 070 (3230×9)	290 700 $(3230 \times 10 \times 9)$	271 510
Experiment II Boosted TC	10	1	31	AZF	9	9585 (1065×9)	95 850 $(1065 \times 10 \times 9)$	89 034
Experiment III DCR	10	7	13	Plain AZ	50	650 (13×50)	91 000 $(13 \times 70 \times 2 \times 50)$	76 646

TABLE 5. Number of HIT assignments in all experiments.

Assign-ments*	Pilot study	I Baseline TC	II General TC	III DCR	Alto-gether
Total	8000	48 469	24 234	5387	86 090
Rejected*	1035	8385	3889	597	13 906
Discarded†	–	826	366	539	1708
Outliers	697	2052	1443	217	4419
Kept‡	6268	37 206	18 536	4034	66 057

* Each assignment has 20 questions.
* Assignments were rejected during the experiment because of failing the test questions or skipping more than three questions. Assignments were re-collected after rejection.
† Assignments that were discarded before outlier removal because of line clicking (all 20 responses were the same).
‡ Assignments are remaining for analysis.

or their combination, we have gained 1 JND over the range of the first 3 JND.

- The four solid curves are for the results with boosting by flicker. These provide an additional large increase in performance. Just the boosting by flicker alone extends the range of perceived distortion to 5 JND. The combinations with either zooming or artefact amplification provide about 6.5 JND in place of only 3 JND, given by traditional plain comparisons. However, the top performing boosting is the combination of all three methods, AZF-boosting, giving close to 9 JND and providing an overall increase by a factor of almost 3 over plain comparison.

These findings hold for the averages taken over all distortion types and source images. The impairment curves for the different distortion types, averaged over the 10 source images for the sequences, show more detailed views of the results and can be found in Figure 15.

B. RESULTS: SENSITIVITY GAIN

The strongest effect of the boosting strategies can be observed for smaller distortion levels, up to 1 JND. To quantify the performance of the boosting strategies also locally at different distortion levels, we introduce the concept of sensitivity and sensitivity gain. In general, a sensitivity analysis determines how changes in an independent variable affect a particular dependent variable. If a differentiable function gives their functional dependence, we can use its derivative as a measure of sensitivity.

In our case, we think of the distortion level as the independent variable and the corresponding reconstructed impairment scales from the eight types of TC as dependent variables. Although the distortion levels are discrete, they correspond to equally spaced, physical, and real-valued distortion parameters. It may be assumed that the resulting impairments of image quality can be modelled by a continuous and differentiable function. Given our data, we applied the 5-parameter logistic fitting [17] to the curves in Figure 14 using

$$\beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp\left(\beta_2(x - \beta_3)\right)} \right) + \beta_4 x + \beta_5, \quad (11)$$

where x denotes the distortion level, and β_1 to β_5 are the parameters for the fit. In the numerical optimization procedure for the curve fitting, local optima will be obtained, depending on the choice of initial parameters. Therefore, we ran the optimization multiple times with different initial conditions and visually checked the fitting quality.

The derivatives of the fitted functions then model the observed *sensitivities* of the plain and boosting techniques to assess the impairment. The sensitivity is also a function of the physical distortion magnitude, respectively the distortion level, allowing a local analysis.

We define the *sensitivity gain*, provided by a particular boosting method as the quotient of the sensitivity for a boosting method and the sensitivity of the method using plain triplet comparisons. A gain larger than 1 indicates an increase in the sensitivity of that factor due to boosting. Figure 16 (left) illustrates the procedure for the case of AZ-boosting. Firstly, it clearly shows that the curve fitting yielded visually convincing smooth functional approximations of the empirical data. Secondly, we see that AZ-boosting yielded a sensitivity gain greater than 1 for distortion levels up to 5 (corresponding to about 1.25 JND), but the method using plain baseline TC was more sensitive for larger levels than the one with AZ-boosted TC.

The right part of Figure 16 shows the sensitivity gain as a function of the distortion level for all seven types of boosting. The curves demonstrate that the boosting methods
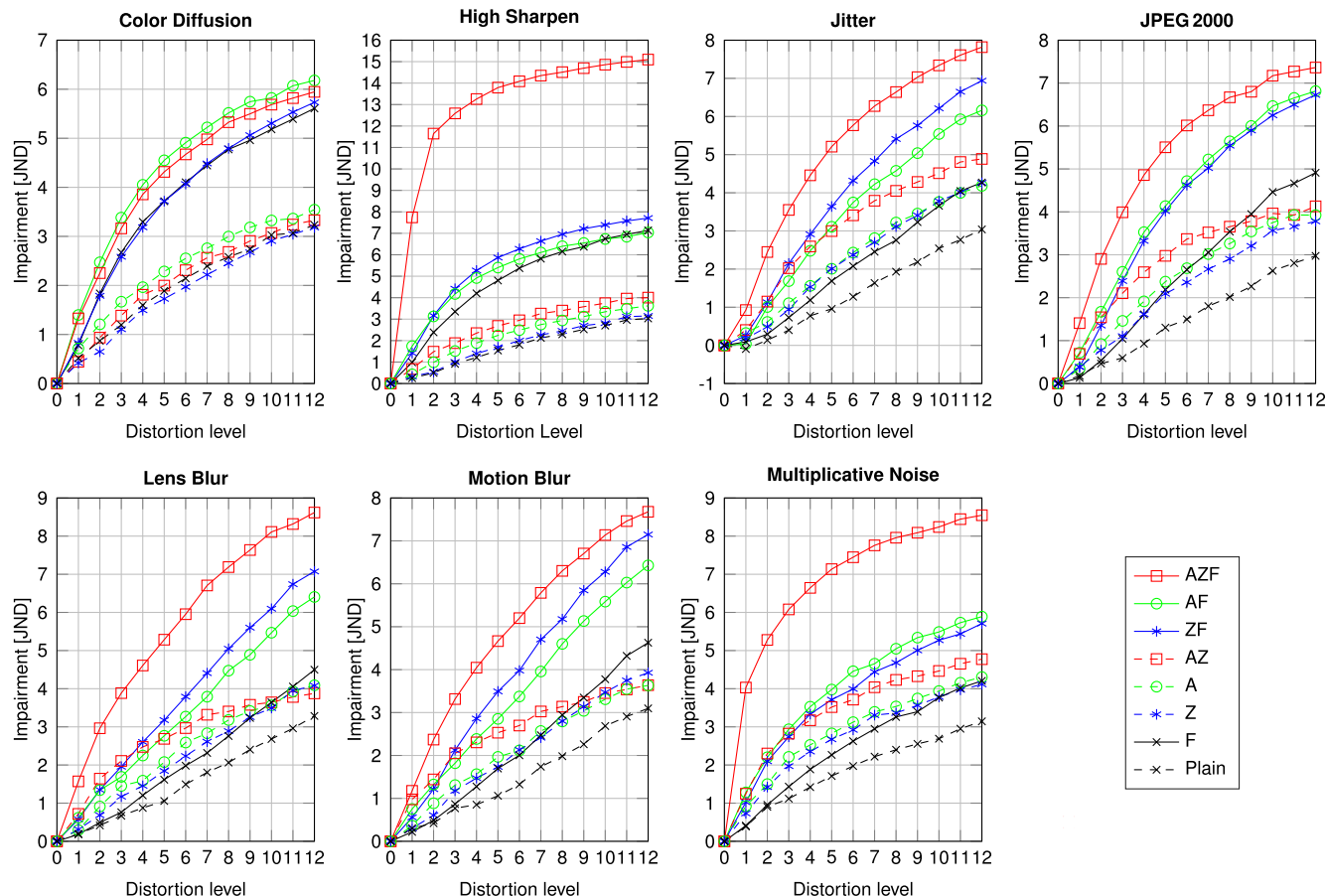
FIGURE 15. Average reconstructed JND scales as in Figure 14 for each type of distortion. Each point on any of these graphs corresponds to the mean impairment scale in JND units of the 10 source images, distorted with the respective type and at the given level.

for baseline triplet comparison are most effective for smaller distortions up to about 1 JND, which corresponds to distortion level 4. Especially for the boosting with flicker, sensitivity gains larger than 2 were achieved.

For larger distortion levels near 2 JND, a kind of saturation effect can be noticed; the gain dropped below 2 but still is larger than 1, except for A- and AZ-boosting. This indicates that perceptually, boosting small distortions to a reference image makes their difference more apparent than boosting large distortions. This effect is particularly strong for artefact amplification among the basic A-, Z- and F-boosting techniques. The sensitivity gain almost linearly drops from 2 at distortion level 0 to 0.8 at distortion level 12. This may, in part, be explained by the nonlinearity of the boosting due to pixel value clamping, see Table 2. Naturally, for small distortions in the test images, there is less clamping to be expected than in test images with large distortions, and such a deficiency does not hold for the other two basic boosting types.

C. RESULTS: TRUE POSITIVE RATE

Recall from Section V-H2 that the true positive rate (TPR) for a set of baseline triplet responses is the ratio of correct answers w.r.t. the ground truth given by the ordering of the

TABLE 6. True positive rate and average response times for triplet comparisons.

Method	TPR	Rank	Response Time (s)	Rank
Plain TC	0.7703	8	2.314 ± 0.412	8
A-boosted TC	0.8001	6	2.229 ± 0.410	7
Z-boosted TC	0.8002	5	2.166 ± 0.408	5
F-boosted TC	0.8470	4	2.225 ± 0.482	6
AZ-boosted TC	0.7791	7	2.146 ± 0.409	4
AF-boosted TC	0.8707	2	2.060 ± 0.438	3
ZF-boosted TC	0.8803	1	2.021 ± 0.440	2
AZF-boosted TC	0.8627	3	1.998 ± 0.434	1

stimuli of an image sequence. In this regard, a response of type "not sure" scores 1/2 point. Therefore, the TPR can also be considered a criterion for validating the performance improvement by the boosting methods. A TPR of 0.5 can be achieved by guessing alone. The TPR for boosted TC is expected to be larger than for plain TC, as boosting should enable observers to make more correct decisions regarding image quality differences.

Table 6 shows the first overall average results for the TPRs in Experiment I (with increasing order) and confirms our expectation. The TPRs show a similar ordering as observed in the average reconstructed impairment scales at larger distortion levels, as shown in Figure 14:

Plain, Z, A, AZ, F, AF, ZF, AZF.
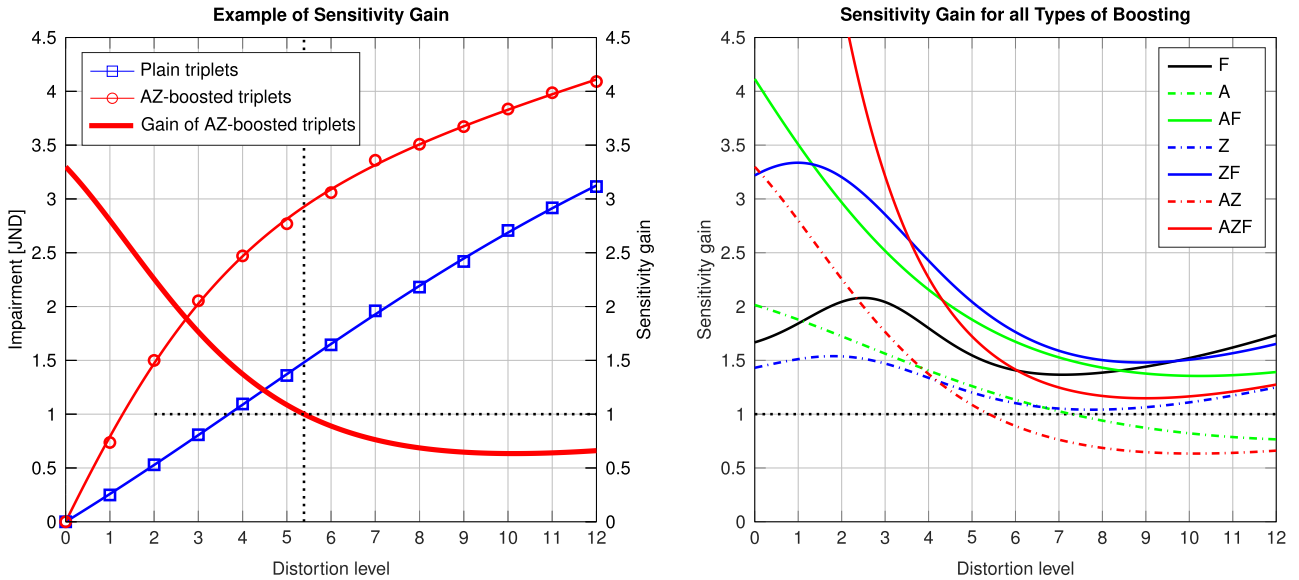
FIGURE 16. Left: The sensitivity gain of AZ-boosted baseline triplet comparison is illustrated. The gain is the ratio of the derivatives of the fitted impairment scale functions. Thus, the gain is the factor by which an increase of perceived distortion is multiplied when AZ-boosting is applied. Here, the sensitivity gain is greater than 1 for distortion levels up to 5 but less than 1 at larger levels. This indicates that AZ-boosting is more sensitive than plain baseline TC for small distortions up to around 1.25 JND and less sensitive for larger levels. Right: Sensitivity gain for all seven types of boosted baseline triplet comparisons. It shows that boosting in baseline triplet comparisons is most effective for smaller distortions.
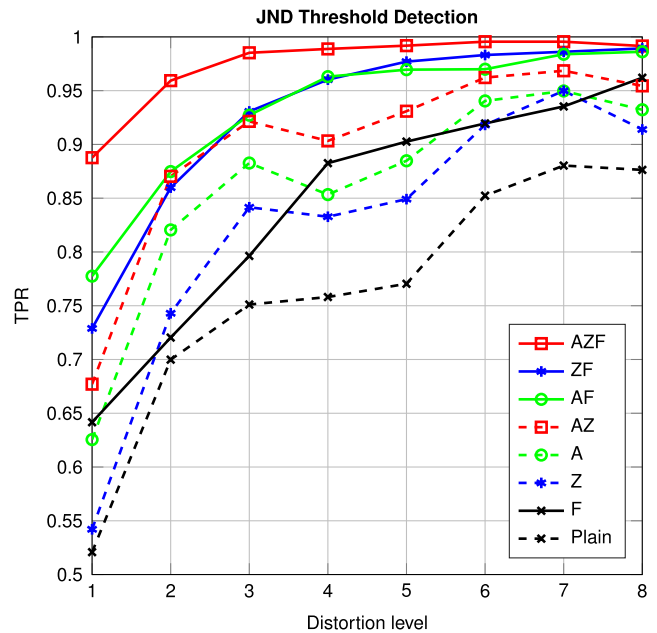
For a more detailed perspective, we consider comparisons as typically done to assess the JND threshold in image sequences with increasing distortion. This amounts to comparisons of distorted images with the corresponding source reference images. In terms of triplet comparisons, we therefore look at triplets $(0, 0, k)$ (or $(i, 0, 0)$). We expect that the TPR increases with the distortion level k (resp. i) and that boosted TCs give rise to larger TPRs.

Figure 17 shows the TPRs for this analysis, averaged over the 70 sequences in our dataset. Corresponding detection rates linearly increase from 0% for a TPR of 0.5 to 100% for a TPR of 1, and the JND threshold on one of the curves is reached at a TPR of 75%. For the case of plain TC, we see that the JND threshold was reached at distortion level 3. The dataset was designed to have the threshold at distortion level 4, and there the TPR is 0.76, still close to 0.75 as expected.

All seven types of boosting increase the TPR compared to plain TC. To give an example, consider test images having just one distortion level difference (0.25 JND). Plain comparison yielded a TPR of only 0.52, which is not much better than guessing.[3] On the other hand, with combined AZF-boosting, the TPR is 0.88, a very strong improvement.

On the far end of the scale at levels 6 to 8, corresponding to distortions 1.5 to 2 JND, the gains in TPR achieved by boosting are smaller. This is due to the saturation effect described earlier, i.e., consistent with our findings for the sensitivity

FIGURE 17. For assessment of the JND threshold, distorted images are compared with the source image. The figure shows the corresponding TPRs from our corresponding triplet comparisons, averaged over all 70 image sequences. Corresponding detection rates linearly increase from 0% for a TPR of 0.5 to 100% for a TPR of 1. Boosting increases the TPR and reduces the JND threshold, which can be read off the graphs at the TPR of 0.75.

gain by boosting. These gains are much more pronounced for small distortions levels.

D. RESULTS: RESPONSE TIME

Boosting not only helps observers to find the correct answers to comparisons but also reduces their response time. Table 6

---

[3]In Section V-G we showed that the expected TPR of the corresponding 2AFC pair comparison for a perceptual difference of 0.25 JND is 0.567, which is a bit larger than the empirical TPR of 0.52 observed here. This may be due to the basic assumption in Thurstonian models that the distributions of the latent perceptual image quality scales are perfectly Gaussian. It is unknown to what extent this assumption holds true.

shows the response times in seconds for all methods with and without boosting, averaged over all baseline triplet comparisons and with corresponding standard deviations. A response for a plain TC required 2.3 seconds on average. All types of boosted TC were faster, with AZF-boosting requiring slightly less than 2 seconds on average. The two-sample *t*-test revealed that these time savings are statistically significant, with very small p-values less than 10^{-11}.

## VII. EXPERIMENT II: BOOSTING FOR GENERAL TRIPLET COMPARISON

In the previous section, we noticed that there is a saturation effect for larger distortion levels, and the reason could be that boosting small distortions of two test images makes their difference stand out better than boosting large distortions with the same difference in distortion levels.

To ameliorate this drop in effectiveness of boosting for larger distortion levels, we consider general triplet comparisons (i, j, k), where the pivot image I_j is not fixed to be the undistorted reference image I_0 of a sequence. Instead, we may allow arbitrary triplets with different pairwise stimuli and select the stimulus with the median distortion magnitude as the pivot.

The artefact amplification then linearly increases the differences with respect to the pivot as before and not the distortions concerning the pristine reference images. Typically, these image differences to be enlarged will be smaller than in Experiment I with baseline triplets. Likewise, in the flicker technique, the flicker is between distortion levels that typically are closer together than with baseline triplets. In Experiment I, we have seen larger sensitivity gains for such smaller quality differences. Thus, we expect to arrive at a sensitivity gain that is enlarged over the whole range of distortion levels, thereby reducing or even eliminating the saturation effect observed for baseline triplet comparisons.

The second motivation to conduct Experiment II is to assess and compare the performance of boosted versus plain TC in terms of the convergence as the number of TCs increases. On the one hand, the reconstructed impairment scales converge, and for the analysis, we consider their confidence intervals from samples of TCs as estimates for the precision of the computed impairment scales. On the other hand, for a given set of distorted images derived from a fixed reference image, the ordering of the reconstructed impairment values should converge, and the Spearman rank-order correlation (SROCC) is the appropriate measure for this analysis.

To study such convergence aspects, a very large pool of TC responses and a challenging set of image sequences are desirable. Therefore, we increased the number of distortion levels in each image sequence from 12 to 30, so that the spacing between consecutive test images is only 0.1 JND. The number of TCs per image sequence was increased from 1360 to 9585 and 29 070 for boosted and plain TC, respectively. To limit the cost for such an enlarged study, we chose to restrict Experiment II to only one type of distortion,
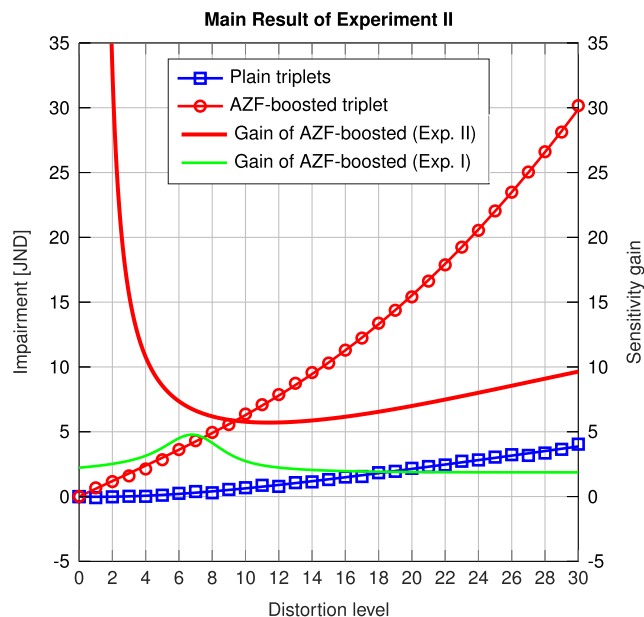
FIGURE 18. Experiment II (general triplet comparisons, motion blur distortion): The reconstruction of impairment scales from plain and AZF-boosted TC are shown by the curves with square and circular markers, respectively. On average, the red curve increases with a slope 7 times as large as that for the blue curve. Thus, the overall gain in sensitivity by AZF-boosting is given by a factor of about 7. The sensitivity gain function for AZF-boosting is also shown (solid red curve). It is globally larger than 5, while the corresponding sensitivity gain function from Experiment I (baseline triplet comparisons, motion blur distortion), shown by the solid green curve, is everywhere below 5.

motion blur, and to the most promising type of boosting, AZF-boosting. All 10 source images were used, resulting in 10 sequences of increasing motion blur distortion.

For 31 images $I_0, I_1, \ldots, I_{30}$ in each of the 10 image sequences, we considered triplets (i, j, k) with $i < j < k$. We further limited the span of these triplets. The span of a triplet (i, j, k) is $S = \max(i, j, k) - \min(i, j, k)$ which in this case is $S = k - i$. For boosted TC, we set the maximal span to 10, and for plain TC to 20, corresponding to 1 and 2 JND, respectively. We anticipated that small differences were harder to detect with plain TC, so larger spans for plain TC than for boosted TC would be advisable for a fair comparison. The number of triplets for each image sequence thus was $\sum_{n=2}^{S}(31 - n)(n - 1)$, which is 3230 for plain and 1065 for boosted TC.

For each of the triplet comparisons, we collected 9 responses from the crowd workers. The presentation of the triplets was randomized in sequence and in orientation, showing either (i, j, k) or (k, j, i). The quality control and outlier removal were carried out as in Experiment I. For details regarding the numbers of HITs that were rejected or classified as outliers, see Table 5.

A. RESULTS: RECONSTRUCTION AND SENSITIVITY GAIN

We reconstructed the impairment scales for the 10 sequences from the collected responses to the plain and the AZF-boosted triplet comparisons. The results, averaged over

Average TPR for Experiment II
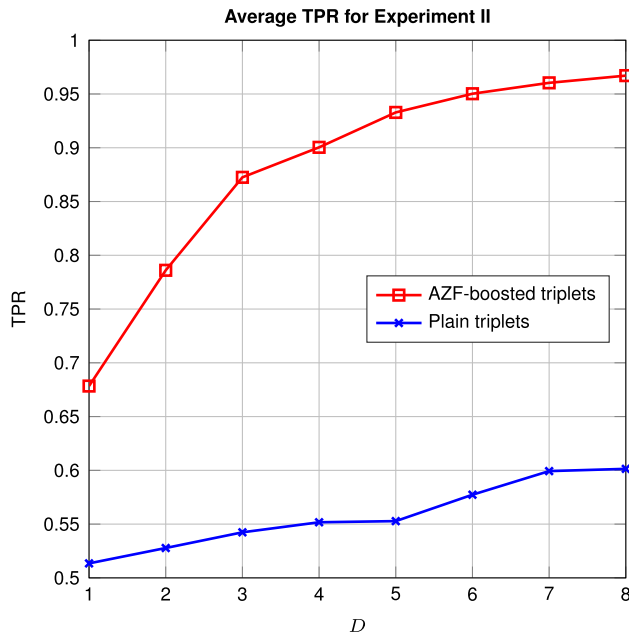
FIGURE 19. Average TPR for all triplets (i, j, k) and distance $D = ||i - j| - |k - j||$ for all 10 sources. The advantage of boosting by artefact amplification, zooming, and flickering is obvious: The easiest plain triplet comparisons are for $D = 8$, and yet they are harder than the hardest AZF-boosted triplet comparisons at $D = 1$.

the 10 sequences, are shown in Figure 18. For plain TC we obtained a very slowly increasing impairment, reaching 4.0 JND at distortion level 30. For AZF-boosted TC, however, we obtained a range of 30 JND. Thus, over the range of 3 JND of motion blur distortion, we recorded an average sensitivity gain of 7.5 for AZF-boosting. The same boosting for baseline triplets gave an average gain of only 1.8.

Figure 18 also displays the achieved local sensitivity gain of AZF-boosted TC over plain TC as functions of the distortion levels. The gain function is shown for baseline triplets used in Experiment I (green curve) and for the general triplets used here (red curve). Over the whole range of distortion levels, the gain achieved in Experiment II is larger than 5. Moreover, the gain is decreasing up until level 12 and then increases again, reaching 9.6 at level 30. In contrast, for baseline TC of Experiment I, the gain is limited below 5 and slowly decreases down to 1.9.

Altogether, our first conjecture, to improve on the sensitivity gain by using general triplets in place of baseline triplets, was clearly confirmed with an impressive tripling of performance in sensitivity.

B. RESULTS: TRUE POSITIVE RATE
The true positive rate is an indicator of the ease of the corresponding set of triplet comparisons. As for Experiment I, we computed the true positive rate for plain and boosted triplet comparison, in this case, averaged over all triplets i, j, k with $i < j < k$ and span $S = k - i \leq 10$. The average TPR for the 10 sequences of motion-blurred images is 0.5583 for plain TC and 0.8810 for AZF-boosted TC.

Compared to Experiment I, we see that AZF-boosting brings about an even larger increase of overall TPR. We obtained an improvement of 0.3227 using AZF-boosted TC over plain TC with general triplets. With baseline triplets, the corresponding improvement was smaller, 0.1313, for the case of motion blur, and 0.0924 for all types of distortion on average.

Figure 19 shows a more detailed view, breaking up the averages into eight parts, based on the absolute differences $D = ||i - j| - |k - j|| = 1, \ldots, 8$. Triplet comparisons with larger values of D can be expected to be easier to judge correctly, and this is reflected in the monotonic increase of the TPR. It is remarkable that the TPR for AZF-boosting, even for the smallest difference of just 1 level between perceptual distances of the left and right image to the pivot image, is larger than any of the detailed TPRs for the plain triplet comparison.

C. RESULTS: CONVERGENCE IN PRECISION
Assuming that the ratios of responses "left", "right", and "not sure" for each TC (i, j, k) converge as the number of responses tends to infinity, we may expect that the reconstructed impairment scales for the corresponding image sequence also converge. In this subsection, we aim to assess the precision of the reconstructions for given budgets of TCs. For this purpose, we consider the 95% confidence intervals (CI) for all of the reconstructed impairment scales.

For each of the 10 image sequences with motion blur, we had collected approximately 8900 responses for plain and also for AZF-boosted TC with a span of at most 10 distortion levels. For a given budget of responses, one may create a sample of that size from each of these pools of 8900 responses for triplets, using random resampling with replacement. From each such sample, a reconstruction can follow. For each image sequence and each budget of responses, we carried out this procedure and computed the 95% confidence interval for the resulting impairment scale of each image and recorded their lengths. We used budgets from 50 up to 10 000 responses per sequence and collected 500 resamplings each time.

For a fair comparison of the lengths of the confidence intervals derived from plain and boosted TCs, we must take the sensitivity gain of boosted TC into account. The impairment scales reconstructed from boosted TC are approximately 7.5 times larger (Section VII-A), and therefore the corresponding confidence intervals should be scaled by $1/7.5 \approx 0.13$.

Figure 20 shows the resulting (scaled) lengths of the CI for the images at distortion levels 5, 15, and 25, averaged over the 10 sources, on a doubly logarithmic grid. For both methods, plain and boosted TC, the sizes of the confidence intervals shrink by about two orders of magnitude when the budget of responses is increased from 50 to 10 000. The precision given by reconstructions from 10 000 responses for plain TC can be achieved by only 300–400 responses with the AZF-boosted TCs. In other words, in this experiment, a single response for a boosted TC gave as much benefit in terms of resulting precision as 25 to 33 responses for plain TCs.
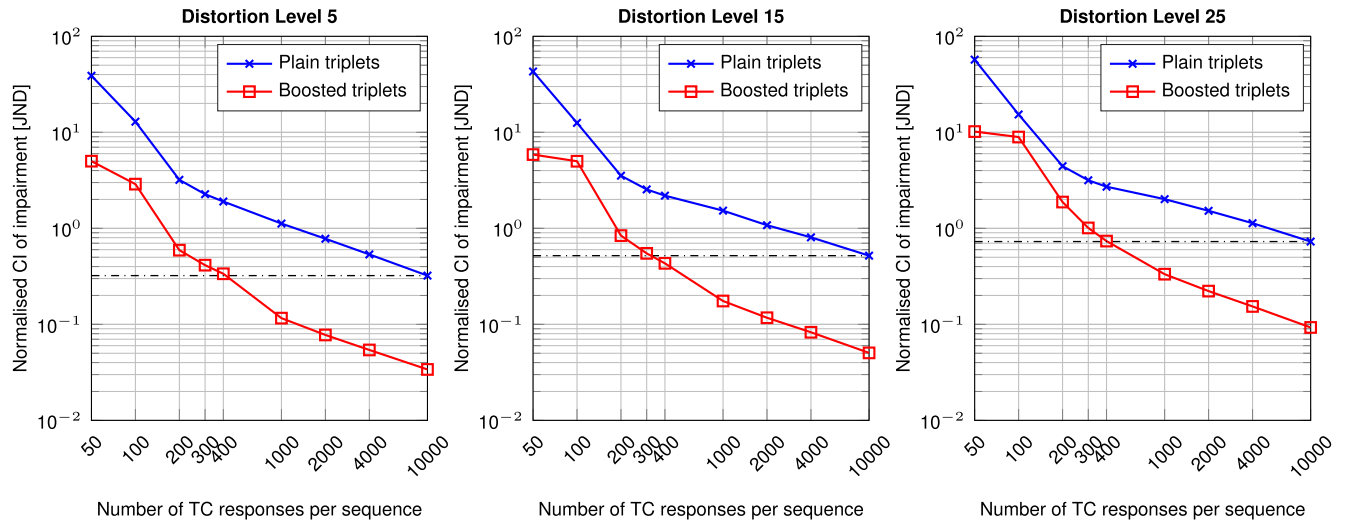
FIGURE 20. Comparative study on the precision of reconstructions from TCs. Impairment scales were reconstructed for each of the 10 image sequences from sets of 500 random samples of a variable number of responses for plain and AZF-boosted TCs. The figure shows the average lengths of the corresponding 10 confidence intervals for the stimuli at distortion levels 5, 15, and 25 as indicators of precision. The best achievable performance for plain TC, obtained from 10 000 responses per sequence, was surpassed by reconstructions derived from as few as 400 responses to AZF-boosted TCs.

**TABLE 7.** Convergence in SROCC for AZF-boosted and plain TC.

TC responses	AZF-boosted TC		Plain TC	
	SROCC	CI length	SROCC	CI length
50	0.9372	0.2267	0.7340	0.7600
100	0.9682	0.2106	0.7409	0.7553
200	0.9915	0.1060	0.7420	0.6689
500	0.9982	0.0273	0.7897	0.5987
1000	0.9993	0.0036	0.8044	0.5687
2000	0.9998	0.0013	0.8633	0.4804
5000	1	0.0006	0.9223	0.3513
10 000	1	0.0002	0.9412	0.2452

D. RESULTS: CONVERGENCE IN ORDERING

To compare the convergence of the quality assessment using boosted TCs with that for plain TCs, we followed the same approach as in the previous Section VII-C. Using resampled data for given budgets of TC responses per sequence of 31 images, we computed the impairment scale reconstructions and their corresponding SROCC w.r.t. the ground truth determined by the distortion levels of the images. For each size of budget, we produced 500 resamplings and recorded the median SROCC and the corresponding 95% CI (using the percentile method). Finally, we averaged the median SROCC and lengths of the CIs over the 10 image sequences.

These results are listed in Table 7 and visualised in Figure 21. The SROCC for plain TC is smaller than 0.9 for all sample sizes up to 2000 responses per sequence, while for AZF-boosted TC, the SROCC is above 0.9, even for samples of only 50 responses per sequence. The SROCC of 0.9412 for plain TC using 10 000 responses is surpassed by the SROCC for boosted TC with as few as 100 responses.

We demonstrate the advantage of boosted TC over plain TC by means of an example. We pick a source image and its 30 distorted versions, labelled by motion blur distortion levels 0 to 30. Then, for each plain and boosted TC, we choose samples of 100, 1000, and 10 000 randomly
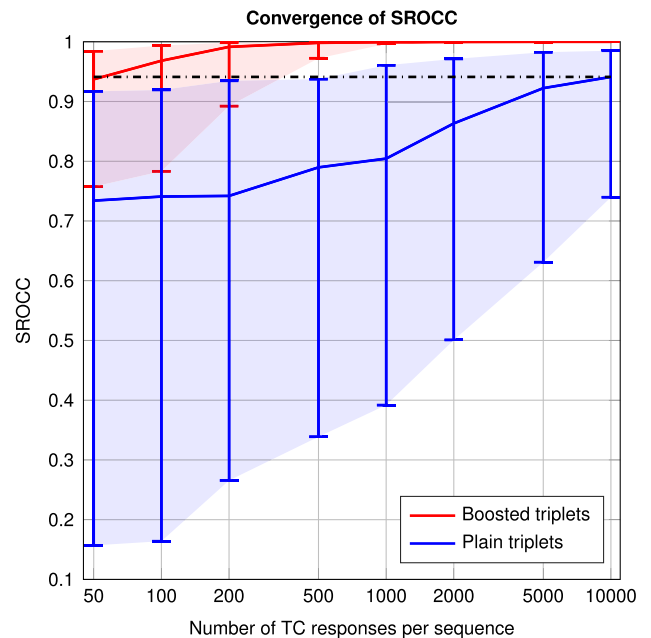
FIGURE 21. This figure illustrates the data from Table 7. Impairment scales were reconstructed from random samples of a variable number of responses for plain and AZF-boosted TCs. We show their (median) rank-order correlation (SROCC) with the ground truth ordering, averaged over image sequences from 10 sources. The upper and lower bound of CI (95%) was averaged over 10 sequences. The best achievable performance for plain TC required 10 000 responses but was beaten by reconstructions derived from as few as 100 responses to AZF-boosted TCs.

selected responses (with replacement), followed by reconstruction of the 31 impairment scales. Sorting the image labels for each reconstruction according to increasing impairment scales yields permutations of (0, 1, . . . , 30). See Table 8 for the results. The table also shows the corresponding SROCC and the number of inversions in the permutations, which express the quality of the orderings (lower is better).

Main Result for Experiment III (DCR)
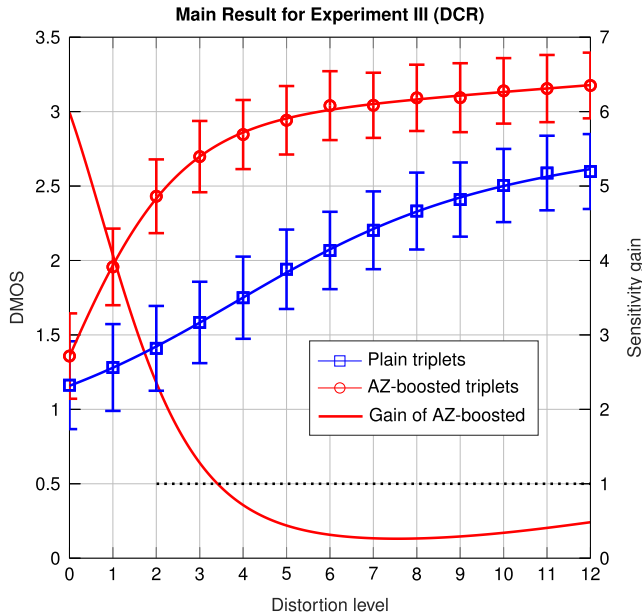
FIGURE 22. The DMOS of the DCR study of Experiment III show a sensitivity gain when assessing the perceptual image quality with small distortions up to about 0.75 JND (corresponding to level 3). The DMOS values are averaged over 70 sequences. 95% confidence intervals were computed for the DMOS of each distorted image. For each of the distortion levels, the average length of the corresponding 70 CIs was computed. The figure shows the centerd CIs with these average lengths. The solid red curve is the corresponding sensitivity gain function.

The number of inversions is equal to the number of swaps required for sorting a permutation by the Bubble Sort algorithm.

The best result for plain TC is from a sample of 10 000 TC responses and has 40 inversions and an SROCC of 0.9331. It has about the same quality as a result obtained from a sample of only 100 AZF-boosted TC responses which has 42 inversions and an SROCC of 0.9484. With 1000 responses for boosted TCs, the ordering of the reconstruction is almost perfect, with only a single inversion left.

To summarise, in this experiment, each of the first 100 responses for AZF-boosted TC was worth more than 100 responses for plain TCs in terms of the resulting SROCC of the reconstruction. To obtain an SROCC of 0.95, our boosting method was 100 times as efficient as plain TC.

VIII. EXPERIMENT III: BOOSTING FOR DEGRADATION CATEGORY RATING

In our third and last experiment, we briefly tested the potential of boosting together with degradation category rating (DCR), which is one of the standard methods for subjective FR-IQA. Since in the conventional DCR approach, the distorted and reference images are displayed side by side, it is not appropriate to show a flickering image in such a scenario. Hence, only still images can be considered. In Experiment I, AZ-boosted TC provided the largest sensitivity gain for TC among the non-flickering boosting techniques. Therefore, here we repeated Experiment I, investigating the performance of plain and AZ-boosted comparisons applied

in the DCR setting, denoted as Plain-DCR and AZ-DCR, respectively.

The experimental setup, the quality control, and the outlier removal for this study were as described earlier in Section V and shown in Figure 9. As in Experiment I, we used all 70 image sequences from 10 sources and 7 distortion types, each containing a source reference image and 12 increasingly distorted images. The set of all Plain-DCR and AZ-DCR questions was shuffled and distributed into a sufficient number of HITs, each one also containing one test question for quality control. For each image, we collected 50 ratings. The statistics of the collected data is detailed in Tables 5 and 4.

A. RESULTS: RECONSTRUCTION AND SENSITIVITY GAIN

Figure 22 shows the DMOS for the two types of DCR, averaged over the 70 sequences. Both methods worked well. The DMOS curve for AZ-DCR is above that for Plain-DCR and spans over a larger interval. Thus, on average, AZ-boosting provided an increased sensitivity as anticipated. However, the gain in sensitivity is restricted to small distortions, up to level 3 (0.75 JND). For larger distortions, the sensitivity gain is less than 1, showing a saturation effect similar to A- and AZ-boosted baseline triplet comparison, compare Figure 16.

Note that the DMOS should ideally be equal to 0 at distortion level 0, since when the reference image is compared to itself, there is no difference between the two, and the distortion should be rated "imperceptible" with a score of 0. However, from Figure 22, the DMOS at level 0 is not equal to 0 but even larger than 1 for both of the DCR tests. We think the reason for this outcome is that in this experiment, observers were instructed to expect to see distorted images on the right side, and during the work on the assignments, this expectation was fulfilled, almost always. Therefore, the participants in the study may have been hesitant to declare that they could not detect any difference. So many rated the distortion for a displayed test image as "perceptible, but not annoying" or even worse, although it actually was identical to the reference image.

B. RESULTS: PRECISION

We evaluated the precision of the acquired DMOS results by computing their 95% confidence intervals. Figure 22 shows the results for Plain- and AZ-DCR, where we have averaged the full width of the confidence intervals over the 70 image sequences. These average confidence intervals, based on up to 50 collected ratings per test image, range from ±0.220 to ±0.295 on the 5-point DCR impairment scale. The confidence intervals for boosted AZ-DCR are slightly smaller than for Plain-DCR.

IX. RESCALING BOOSTED IMPAIRMENTS BY HYBRID TRIPLET COMPARISONS

Our boosting techniques of artefact amplification, zooming, and flickering were designed to perceptually magnify differences between compared stimuli so that human observers are enabled to distinguish fine-grained distortion levels better.

TABLE 8. Orderings from the reconstructions for source image SRC07 and motion blur distortion. Last column: Number of inversions.

| Method	Responses	Ordering	SROCC	Inv.
Plain TC	100	1 4 18 8 2 14 6 5 12 17 11 0 26 20 15 19 7 16 13 21 10 23 9 3 24 22 29 25 27 30 28	0.6661	117
Plain TC	1000	6 5 11 1 8 7 0 18 2 3 15 12 4 27 19 26 9 10 13 14 21 16 23 20 17 30 22 25 24 28 29	0.7754	95
Plain TC	10 000	0 1 3 2 4 5 8 6 7 9 10 11 12 15 13 14 26 17 16 28 23 19 21 18 27 20 24 29 22 25 30	0.9331	40
AZF-boosted TC	100	0 2 1 3 13 8 5 6 4 14 11 7 12 9 10 15 16 18 17 22 21 23 24 20 19 26 29 25 28 27 30	0.9484	42
AZF-boosted TC	1000	0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 29 28 30	0.9996	1
AZF-boosted TC	10 000	0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30	1	0

The experiments of the previous sections have confirmed this intended effect.

However, boosting amounts to a nonlinear scaling of perceptual distortion, as already apparent from Figure 14. Moreover, this nonlinearity may depend on the distortion type and the content of the source images. For example, using boosting by zooming and flicker, the impairment range of 3 JND units for plain TC was stretched to 7 JND for the jitter distortion, but only to 5.5 JND for color diffusion, see Figure 15.

This nonlinear scaling of perceptual distortion could be disregarded when building FR-IQA datasets. After all, it is a commonly accepted practice to use DCR or reconstructions from pair comparison for subjective quality assessment, although also the DMOS scale is not perceptually linear and is also not proportional to the reconstruction from pair comparisons. Moreover, as shown in Table 1, the creators of FR-IQA and FR-VQA datasets have applied various other methods for the assessment of impairment scales, but there is no agreed upon standard that provides any particular scale as a common ground that other scales can be related to. Only recently, procedures were proposed to merge impairment scale values from different methodologies (pair comparison and category rating) and from different datasets, see [75]–[77].

In this section, we propose a similar approach to transform impairment scale values from boosted triplet comparisons back to scales obtained in the traditional way without boosting perceptual discrimination power. For this purpose, we construct a (nonlinear) monotonic transformation of boosted scales for each image sequence. Thereby, we preserve the discrimination of fine-grained distortion differences achieved by the boosting approach while simultaneously ensuring that the ranges of transformed absolute impairment scales match those that are obtained when comparing distorted images without boosting. In other words, the transformed impairment values reflect the perceived qualities of the original distorted images rather than the qualities of the images with perceptually boosted distortions.

Towards this end, we propose a hybrid method by allowing for a fraction of triplet comparisons without boosting. The scale reconstruction from these plain comparisons provides a rough estimate of the desired impairment scales. We then fit a smooth scalar transformation for each sequence of distorted images that maps the boosted scales to the target scales in the least-squares sense. This transformation should be smooth and monotonic so that the high-quality relative differences of close scale values and the overall ordering of the image sequences obtained from boosted comparisons
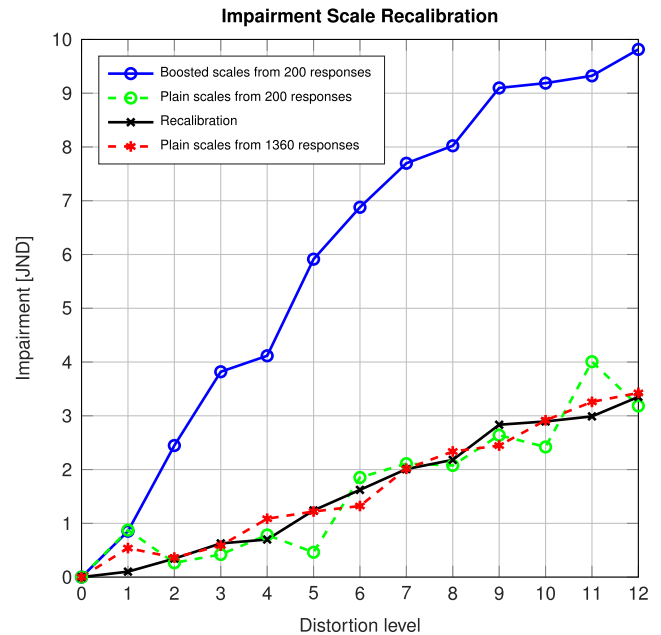
FIGURE 23. Example of the impairment scale recalibration by the hybrid method with parameters $K = 400$ and $\alpha = 0.5$. The data stems from the image sequence for source image SRC06 and distortion type jitter, assessed by plain and AZF-boosted triplet comparisons in Experiment I. The hybrid method fits the reconstruction from $(1 - \alpha)K = 200$ boosted TCs (blue line with circles) to that from $\alpha K = 200$ plain TCs (green dashed line), with the result shown by the black line with crosses. For comparison, the reconstruction from all 1360 plain TC responses is also shown (red dashed line).

are maintained. In this way, we recalibrate the scales from boosted comparisons to follow those from standard comparisons without boosting. By construction, this recalibration is adaptive w.r.t. the source image contents and the type of distortion.

In Algorithm 3, we outline the hybrid method in the form of a pseudo code.[4] Then we provide the results of it when applied to the comparisons that we had collected in Experiment I.

For our implementation of the hybrid method we chose the 5-parameter logistic function of Equation (11) as f_γ. We forced it to be monotonically increasing by the constraint $\beta_1, \beta_2, \beta_4 \geq 0$.

Figure 23 shows the results for the example of one image sequence. For each of the 70 image sequences of Experiments I, we sampled $K = 400$ random triplet

[4]In lines 8 and 9, the functions f_γ, resp. $f_{\hat{\gamma}}$, are applied to each component of their arguments.

Algorithm 3 Hybrid Method: Re-Calibration of Scales From Boosted Triplet Comparisons

1: Input: $I_0, \ldots, I_N$ ▷ sequence with N distorted images
2: Parameters: K, α ▷ budget of comparisons, $0 < \alpha < 1$
3: Do $(1 - \alpha)K$ boosted triplet comparisons
4: Result: $\mu_0^{\text{boost}}, \ldots, \mu_N^{\text{boost}}$ ▷ reconstructed scales
5: Do αK plain triplet comparisons
6: Result: $\mu_0^{\text{plain}}, \ldots, \mu_N^{\text{plain}}$ ▷ reconstructed scales
7: $f_\gamma : \mathbb{R} \to \mathbb{R}$ ▷ select family of monotonic functions
8: $\hat{\gamma} = \arg\min_\gamma \|f_\gamma(\mu^{\text{boost}}) - \mu^{\text{plain}}\|^2$ ▷ function fitting
9: Output: $f_{\hat{\gamma}}(\mu^{\text{boost}})$ ▷ transformed impairment scales

FIGURE 24. Recalibration of the scale reconstruction from boosted triplet comparison of Experiment I by the hybrid method. The right point cloud (blue circles) is the scatter plot of the reconstruction from the AZF-boosted triplet comparison versus those from plain triplet comparisons. The recalibration by the hybrid method produced the left part of the scatters plot along the diagonal as intended for the recalibration. Note that for visual clarity, the figure truncates the reconstructions from boosted comparison to a maximum of 12 JND. There are additional points (blue circles) further to the right, which are not shown. See the main text for more details.

comparisons, half of them as plain triplet comparisons in Step 4 ($\alpha = 0.5$). Since the result of the hybrid method depends on the chosen samples for the TCs, we repeated the hybrid procedure (random sampling and reconstruction, followed by recalibration) 100 times and kept only the median result w.r.t. the mean-square difference between the recalibrated reconstructed scales from boosted TC and the scales from plain TC.

Figure 24 illustrates the resulting rescaled impairment scales for all distorted images from Part A of our dataset KonFiG-IQA. In Experiment I, we had obtained 1360 responses for each type of triplet comparison and each of the 70 sequences with 12 distorted images. The figure shows two scatter plots together. The first one (blue circles) is for the scales, reconstructed from a random sample of only 200 AZF-boosted TCs ($K = 400$, $1 - \alpha = 0.5$) versus those reconstructed from all 1360 responses per sequence from plain TCs. Due to the boosting, the resulting impairment scales are much larger than those from plain comparison.

These raw scales from $(1 - \alpha)K = 200$ boosted TCs were adaptively recalibrated by the hybrid method, using an additional random sample of $\alpha K = 200$ responses

TABLE 9. Recalibration for 200 AZF-boosted TCs per sequence.

Recalibration	RMSE	MAE	PLCC	SROCC
before	11.1439	8.9748	0.3274	0.3300
after	0.4836	0.3916	0.8995	0.9051

(per sequence) from plain TC. As in the previous figure, we kept only the (mean-square difference) median result from recalibrated reconstructions of 101 random samples of the budget of $K = 400$ TC responses. The corresponding scatters plot for the recalibrated results is also included in the graph (black crosses). The recalibration translates each blue circle horizontally to the corresponding black cross. Thus, the recalibration from boosted TCs conforms to the range of scales obtained for plain TCs, i.e., represent the perceptual qualities corresponding to the original distorted images without any boosting. Table 9 gives the corresponding numerical results for the fitting procedure in the hybrid method. RMSE denotes the root-mean-square difference, MAE stands for the mean absolute difference.

We also examined the extent to which the performance of the recalibration by the hybrid method depends on the type of distortion and the choice of the source image. For this purpose, we computed the RMSE between the corresponding recalibrated scales and those derived from all 1360 responses per sequence, using plain comparison. The results are shown in Table 10. The medians of these mean-square differences indeed do not vary strongly between source images or distortion types. However, with only 200 boosted and plain TC responses per sequence of 13 images, the variation between different random samples is larger, as shown by the standard deviations listed in the table.

To judge the usefulness of the hybrid method, it is important to keep in mind that the purpose of the recalibration is not to achieve a ''perfect'' result of zero RMSE or an SROCC equal to 1, but only to match the range of the scales reconstructed from plain comparisons: Due to the increased sensitivity of the boosted image quality assessment, we expect the details of the reconstructed scales from boosted TCs to be more accurate than those reconstructed from plain TCs without boosting.
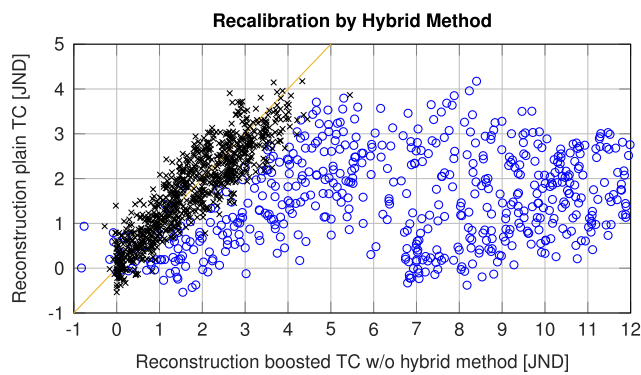
X. DISCUSSION, LIMITATIONS, AND FUTURE WORK
A. OTHER OPTIONS FOR BOOSTING IN FR-IQA
To exploit the full potential of boosting in FR-IQA, other options for boosting can be investigated.

1) TYPE OF ARTEFACT AMPLIFICATION
For the artefact amplification, we have worked with the RGB color space. The RGB color space corresponds well to how images are technically displayed on devices and how colors are processed in the human visual system. However, the RGB space is not a perceptually uniform color space and is not well suited for human interaction. The HSV color space (hue, saturation, value) corresponds better to how people experience color [78]. It separates the chromatic (hue, saturation) from the achromatic (value) color components. In the context

TABLE 10. Performance of the recalibration in the hybrid method per source image and distortion Type: Median RMSE (and STDDEV) w.r.t. Scales from 1360 plain TC per sequence.

	Motion Blur	Lens Blur	Color Diffusion	Jitter	Multiplicative Noise	High Sharpen	JPEG 2000	*Average*
SRC09	0.43 ± 0.91	0.42 ± 0.24	0.42 ± 0.82	0.39 ± 0.29	0.40 ± 0.26	0.43 ± 0.28	0.47 ± 0.24	*0.42*
SRC17	0.34 ± 0.61	0.45 ± 0.45	0.44 ± 0.31	0.44 ± 0.25	0.45 ± 0.30	0.48 ± 0.24	0.45 ± 0.37	*0.44*
SRC28	0.43 ± 0.27	0.36 ± 0.18	0.39 ± 0.27	0.47 ± 0.82	0.41 ± 0.30	0.46 ± 0.24	0.54 ± 0.29	*0.44*
SRC31	0.39 ± 0.20	0.49 ± 0.23	0.40 ± 0.22	0.47 ± 0.18	0.44 ± 0.29	0.50 ± 0.22	0.41 ± 0.32	*0.44*
SRC07	0.45 ± 0.23	0.44 ± 0.26	0.38 ± 0.28	0.48 ± 0.29	0.51 ± 0.33	0.45 ± 0.31	0.47 ± 0.30	*0.45*
SRC01	0.45 ± 0.21	0.40 ± 0.26	0.47 ± 0.26	0.53 ± 0.35	0.49 ± 0.21	0.46 ± 0.35	0.50 ± 0.25	*0.47*
SRC50	0.39 ± 0.25	0.43 ± 0.28	0.51 ± 0.61	0.47 ± 0.25	0.50 ± 0.28	0.44 ± 0.24	0.57 ± 0.31	*0.47*
SRC03	0.46 ± 0.35	0.49 ± 0.25	0.48 ± 0.35	0.42 ± 0.21	0.39 ± 0.50	0.47 ± 0.22	0.58 ± 0.38	*0.47*
SRC45	0.51 ± 0.29	0.44 ± 0.28	0.50 ± 0.32	0.44 ± 0.26	0.50 ± 0.32	0.53 ± 0.26	0.48 ± 0.26	*0.49*
SRC06	0.45 ± 0.23	0.50 ± 0.41	0.59 ± 0.65	0.46 ± 0.20	0.52 ± 0.29	0.46 ± 0.27	0.67 ± 0.41	*0.52*
Average	*0.43*	*0.44*	*0.46*	*0.46*	*0.46*	*0.47*	*0.51*	

of artefact amplification, this implies that the clamping of the value component in HSV space does not affect the color appearance, while when working in RGB space, the clamping of an R, G, or B component does. Another option is to extend the technique to a context-dependent one, which takes into account the local JND (e.g., [79]) when amplifying the image distortion at each pixel.

2) AMPLIFICATION AND ZOOM FACTORS
The benefits of our boosting strategies depend on the choices for the amplification and zoom factors together with the zoom regions. We had made the choices heuristically, based on some manual exploration. However, other amplification and zoom factors may provide improved results, depending on the content, distortion type, and test environment.

3) FLICKER FREQUENCY
In our flickering study, the displayed image buffer was swapped between the reference and the distorted image 8 times per second. In other words, the frequency of the visual signal was 4 Hz. This is different from [38], in which the temporal TCSF suggests a contrast threshold of a flickering frequency of 8 Hz. However, it should be noted that images are different from the test stimuli used in the experiments regarding the TCSF. An interesting future work, therefore, is to characterize the TCSF for our IQA application. Furthermore, buffer swap rates, different from 8 times per second, may increase the sensitivity for subjective IQA even more.

B. LIMITATION OF GENERAL TRIPLET COMPARISON
In Experiment II, we introduced general triplet comparisons that provided an improved sensitivity gain compared to baseline triplet comparisons, especially for larger distortions (2 to 3 JND). However, there is an important limitation of this method. Such triplet comparisons aim at capturing relations between perceptual distances of stimuli. The reconstruction of the corresponding quality scales w.r.t. a reference then relies on the assumption that the given sequence of stimuli can be modelled as a subset of a one-dimensional Euclidean space such that distances properly add up. Thus, if I_a, I_b, I_c denote three stimuli with increasing impairment scales, then we expect for the perceptual distances that $d(I_a, I_b) + d(I_b, I_c) = d(I_a, I_c)$ holds. This assumption appears natural for each

image sequence derived for a single type of distortion, and therefore general triplet comparisons, with or without boosting, may be expected to provide meaningful results.

However, general triplet comparisons are no longer applicable if we mix several distortion types in one image sequence as was done in MDID [13], for example. In this case, consider two images with equal impairment scale, derived from the same reference image, but with two different types of distortion like JPEG compression and color diffusion. Then, clearly, these two distorted images are perceptually noticeably different from each other, yet their difference in impairment is equal to 0. In [76] it was therefore suggested to rename the JND measurement units of impairment scales to "just objectionable differences".

The reason for this discrepancy is that a set of distorted images, derived from different types of distortions or with mixed distortions together, cannot be expected to lie on a one-dimensional continuum in image space. In our future work, we will study to what extent multidimensional scaling (MDS) methods can facilitate the application of general triplet comparisons to impairment scale reconstruction also for image sets derived from a reference image by multiple distortions. For example, one can consider maximum likelihood difference scaling (MLDS, [63]) or stochastic triplet embedding (STE, [61]) for MDS, select a suitable embedding dimension, and finally derive impairment scales from the distances in the multidimensional embedding space. The results can be compared and validated with the corresponding DMOS values.

C. OPTIMAL ALLOCATION OF TRIPLET COMPARISONS IN THE HYBRID METHOD
In Section IX we introduced a hybrid method that combined boosted and plain triplet comparisons in order to recalibrate reconstructed impairment scales to the traditional impairment ranges achieved without boosting. A fraction α of a fixed budget of K comparisons was devoted to plain comparisons. For our empirical analysis, we have used $\alpha = 0.5$. The more responses we collect from boosted TCs, the better the accuracy and precision of the resulting reconstruction. But increasing α reduces the number of auxiliary plain TCs and worsens the alignment with the scale range valid for impairment without boosting.

To investigate the tradeoff between accuracy, respectively precision, on the one side and the alignment with the ground truth result achievable without boosting on the other side, we will carry out suitable simulations and experiments. As a first step we will rerun the computations as provided in Figures 23, 24, and Table 10 with variable parameters K and α. After defining a suitable target functional that combines accuracy, respectively precision, with alignment quality, we can estimate an optimal fraction α of plain TCs for any given budget of K TCs.

D. ADAPTIVE SAMPLING STRATEGIES

The confidence intervals of the reconstructed impairment of quality scales generally will shrink as more and more responses to triplet comparison are collected. In our experiments, we used TCs (i, j, k) with an equal number of target responses per TC. The triplets were moderately restricted, e.g., to spans of 10 or 20 distortion levels in Experiment II. However, an adaptive sampling strategy based on information theoretic considerations may reduce the number of responses required to achieve the same quality of the result. In such a procedure, the expected information gain is considered that can be derived for the next response for a TC or the responses for a batch of several TCs. This expected information gain may be maximized, thereby determining the next TCs to be posted to observers.

Such adaptive sampling approaches have already been proven useful for pair comparisons, see, e.g., [75] and [74]. A similar general adaptive framework for the assessment of psychometric functions is QUEST+ [80]. We propose to tailor such adaptive sampling strategies for the case of triplet comparisons which would further improve the performance of impairment scale assessment using boosted triplet comparisons.

E. APPLICATION SCENARIO

In general, boosting strategies help to evaluate more conservative JND thresholds and increase the discrimination power of subjective image quality assessment. In the existing FR-IQA datasets, the distorted images were typically generated by applying only a few sparse distortion levels to a reference stimulus. In such cases, it may be expected that subjective FR-IQA by comparisons without boosting already provides reliable results. However, in many applications, assessing slight quality differences is desirable. For example, in video compression, fine-grained quality assessment for small impairment scales up to 1 JND would be desirable for content providers that strive to satisfy most of their consumer clients. Our boosting strategies enable faster and less expensive reconstruction for such small distortion scales. If needed, the reconstructed scores of the stimuli with boosted distortions subsequently can be mapped back to the JND scale for the stimuli without boosted distortions, as shown in Section IX. We are currently applying the flicker technique for subjective assessment of the JND and the satisfied user ratio (SUR) in image and video compression methods.

Another application of boosting strategies is robust watermarking. Recently, several methods using JND assessment have been proposed for robust image watermarking [81]–[84]. Our boosting strategies would increase the visibility of the distortions caused by a watermarking algorithm. Therefore, the watermarking algorithm can be optimized so that the watermark distortions would not be visible even when the distortion of the stimulus is boosted. Thus, this procedure would result in more robust watermarking.

A further future research direction is the use of boosting strategies for the subjective evaluation of other imaging modalities, such as stereoscopic imaging and screen content.

F. KonFiG-IQA: THE KONSTANZ FINE-GRAINED IQA DATASET

Our Konstanz Fine-Grained IQA dataset (KonFiG-IQA, Parts A and B) has been publically available online in our dataset repository `http://database.mmsp-kn.de`. Part A contains the 10 source images and corresponding distorted versions (7 distortion types at 12 levels, spaced at 0.25 JND), resulting in 840 distorted images in total. We supply the (MATLAB) code to boost the distortions with respect to a reference image using artefact amplification, zooming, and flicker. Part B provides the distorted images for motion blur only, however, with 30 levels of distortion, spaced at 0.1 JND.

KonFiG-IQA also includes a large number of subjective responses to triplet comparison and DCR ratings. Only those responses and ratings that remained after the data cleaning process and outlier removal are provided. For each TC response, we give a record

```
[source_id, distortion_type, (i,j,k),
  response, time_stamp, time_used,
  worker_id],
```

where `response` denotes the response `left`, `not sure`, `right`, and `worker_id` is an anonymous identifier of the corresponding observer. In total, there are 70 6914 and 36 0544 responses to triplet comparisons for Parts A and B, respectively.

For Part A we also conducted a DCR study, yielding 36 0554 ratings. For each one we provide a record

```
[source_id, distortion_type,
  distortion_level, rating, time_stamp,
  time_used, worker_id].
```

The last part of the dataset consists of the impairment scales, reconstructed from the TC responses and DCR ratings.

The source code for generating the distorted images and for reconstructing impairment scales from triplet responses will be provided on `GitHub`.

XI. CONCLUSION

For many applications of full-reference image quality assessment, reliable and precise subjective image quality

measurement for small distortions near and also below the just noticeable difference on the perceptual quality scale is important. In this contribution, we showed that the sensitivity of conventional assessment methods w.r.t. small changes in perceptual quality can be strongly enlarged by several boosting methods. For this purpose, we proposed artefact amplification, zooming, the flicker test, and their combinations.

We proposed triplet comparisons for the subjective quality assessment and provided the details required to reconstruct impairment scales for distorted images. This reconstruction is obtained by maximum likelihood estimation, based on the Thurstonian probabilistic model of image quality, thus compatible with the state-of-the-art method applied for classical pair comparison.

To assess the potential of the approach, we have created the first FR-IQA dataset, KonFiG-IQA, which was designed based on perceptual criteria. Sequences of distorted images were generated with a fine-grained, perceptually equidistant spacing of distortion levels, only 0.25 JND (or 0.1 JND) apart from each other.

In a large crowdsourced field study, we collected over 1.7 million responses to triplet comparison questions. We gave a detailed analysis of this data in terms of scale reconstructions, their accuracy, convergence, detection rates, the sensitivity gain achieved by the different boosting methods, and more. Boosting methods proved to increase the discriminatory power for the fine-grained dataset, allowing to reduce the number of subjective quality comparisons while improving the accuracy of the resulting relative image impairment scales in terms of SROCC w.r.t. the available ground truth.

Increasing perceptual sensitivity by the boosting strategies necessarily implies that the obtained impairment scales are larger than for conventional methods such as degradation category rating. However, these larger ranges also depend on the respective image contents and the type of distortion. In order to map the high precision boosted impairment scales back to the ranges corresponding to the original distorted images without boosting, we proposed a hybrid method that applies a monotonic numerical transformation of scale values based on a few auxiliary triplet comparisons without boosting.

Our boosting techniques pave the way to fine-grained image quality datasets, allowing for an increased number of distortion levels, yet with high-quality subjective ground truth annotations facilitated by an amplified perceptual discrimination power.

REFERENCES

[1] *Methodologies for the Subjective Assessment of the Quality of Television Images*, document ITU-R Recommendation BT.500, 2019. [Online]. Available: https://www.itu.int/rec/R-REC-BT.500

[2] *Methods for Subjective Determination of Transmission Quality*, document ITU-T Recommendation P.800, 1996. [Online]. Available: https://www.itu.int/rec/T-REC-P.800-199608-I

[3] *Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment*, document1 ITU-T Recommendation P.913, 2008. [Online]. Available: https://www.itu.int/rec/T-REC-P.913-201603-I

[4] B. L. Jones and P. R. McManus, "Graphic scaling of qualitative terms," *SMPTE J.*, vol. 95, no. 11, pp. 1166–1171, Nov. 1986.

[5] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Proc. 17th ACM Int. Conf. Multimedia (MM)*, 2009, pp. 491–500.

[6] M. Seufert, "Statistical methods and models based on quality of experience distributions," *Qual. User Exper.*, vol. 6, no. 1, pp. 1–27, Dec. 2021.

[7] L. L. Thurstone, "A law of comparative judgment," *Psychol. Rev.*, vol. 34, no. 4, p. 273, 1927.

[8] M. Perez-Ortiz and R. K. Mantiuk, "A practical guide and software for analysing pairwise comparison experiments," 2017, *arXiv:1712.03686*. [Online]. Available: http://arxiv.org/abs/1712.03686

[9] L. L. Thurstone, "Equally often noticed differences," *J. Educ. Psychol.*, vol. 18, no. 5, p. 289, 1927.

[10] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Comput. Graph. Forum*, vol. 31, no. 8, pp. 2478–2491, 2012.

[11] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," *Adv. Modern Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.

[12] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, and F. Battisti, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.

[13] W. Sun, F. Zhou, and Q. Liao, "MDID: A multiply distorted image database for image quality assessment," *Pattern Recognit.*, vol. 61, pp. 153–168, Jan. 2017.

[14] H. Men, V. Hosu, H. Lin, A. Bruhn, and D. Saupe, "Subjective annotation for a frame interpolation benchmark using artefact amplification," *Qual. User Exper.*, vol. 5, no. 1, pp. 1–18, Dec. 2020.

[15] W. S. Torgerson, *Theory and Methods of Scaling*. New York, NY, USA: Wiley, 1958.

[16] S. Haghiri, F. A. Wichmann, and U. von Luxburg, "Estimation of perceptual scales using ordinal embedding," *J. Vis.*, vol. 20, no. 9, p. 14, Sep. 2020.

[17] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[18] D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, Jan. 2010, Art. no. 011006.

[19] A. Zarić, N. Tatalović, N. Brajković, H. Hlevnjak, M. Lončarić, E. Dumić, and S. Grgić, "VCL@FER image quality assessment database," *Automatika*, vol. 53, no. 4, pp. 344–354, 2012.

[20] X. Liu, M. Pedersen, and J. Y. Hardeberg, "CID:IQ—A new image quality database," in *Proc. Int. Conf. Image Signal Process. (ICSIP)*. Cham, Switzerland: Springer, 2014, pp. 193–202.

[21] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Proc. 11th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–3.

[22] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H.264/AVC video database for the evaluation of quality metrics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 2430–2433.

[23] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[24] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan, *IVP Subjective Quality Video Database*. The Chinese Univ. Hong Kong, Hong Kong, 2011. [Online]. Available: http://ivp.ee.cuhk.edu.hk/research/database/subjective/

[25] P. V. Vu and D. M. Chandler, "ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imag.*, vol. 23, no. 1, p. 013016, Feb. 2014.

[26] J. Y. Lin, R. Song, T. Liu, H. Wang, and C.-C. J. Kuo, "MCL-V: A streaming video quality assessment database," *J. Vis. Commun. Image Represent.*, vol. 30, pp. 1–9, Jul. 2015.

[27] Netflix. *NFLX Public Dataset*. Accessed: Oct. 7, 2021. [Online]. Available: https://github.com/Netflix/vmaf/blob/master/resource/doc/datasets.md

[28] N. T. Blog. (2016). *Toward a Practical Perceptual Video Quality Metric*. [Online]. Available: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[29] L. Jin, J. Y. Lin, S. Hu, H. Wang, P. Wang, I. Katsavounidis, A. Aaron, and C.-C. J. Kuo, "Statistical study on perceived JPEG image quality via MCL-JCI dataset construction and analysis," *Electron. Imag.*, vol. 2016, no. 13, pp. 1–9, Feb. 2016.

[30] X. Shen, Z. Ni, W. Yang, X. Zhang, S. Wang, and S. Kwong, "A JND dataset based on VVC compressed images," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2020, pp. 1–6.

[31] H. Wang, W. Gan, S. Hu, J. Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C.-C.-J. Kuo, "MCL-JCV: A JND-based H.264/AVC video quality assessment dataset," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1509–1513.

[32] H. Wang, I. Katsavounidis, J. Zhou, J. Park, S. Lei, X. Zhou, M.-O. Pun, X. Jin, R. Wang, X. Wang, Y. Zhang, J. Huang, S. Kwong, and C.-C. J. Kuo, "VideoSet: A large-scale compressed video quality dataset based on JND measurement," *J. Vis. Commun. Image Represent.*, vol. 46, pp. 292–302, Jul. 2017.

[33] C. Fan, Y. Zhang, R. Hamzaoui, and Q. Jiang, "Interactive subjective study on picture-level just noticeable difference of compressed stereoscopic images," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8548–8552.

[34] X. Liu, Z. Chen, X. Wang, J. Jiang, and S. Kowng, "JND-Pano: Database for just noticeable difference of JPEG compressed panoramic images," in *Proc. Pacific Rim Conf. Multimedia (PCM)*. Cham, Switzerland: Springer, 2018, pp. 458–468.

[35] Q. Huang, H. Wang, S. C. Lim, H. Y. Kim, S. Y. Jeong, and C.-C.-J. Kuo, "Measure and prediction of HEVC perceptually Lossy/Lossless boundary QP values," in *Proc. Data Compress. Conf. (DCC)*, Apr. 2017, pp. 42–51.

[36] J. Y. Lin, L. Jin, S. Hu, I. Katsavounidis, Z. Li, A. Aaron, and C.-C. J. Kuo, "Experimental design and analysis of JND test on coded image/video," *Proc. SPIE*, vol. 9599, Sep. 2015, Art. no. 95990Z.

[37] B. W. Keelan and H. Urabe, "ISO 20462: A psychophysical image quality measurement standard," in *Proc. Image Qual. Syst. Perform.*, Dec. 2003, pp. 181–189.

[38] A. B. Watson, *Handbook of Perception and Human Performance*. New York, NY, USA: Wiley, 1986, ch. Temporal Sensitivity, pp. 6.1–6.43.

[39] D. M. Hoffman and D. Stolitzka, "A new standard method of subjective assessment of barely visible image artifacts and a new public database," *J. Soc. Inf. Display*, vol. 22, no. 12, pp. 631–643, Dec. 2014.

[40] M. A. Elgharib, M. Hefeeda, F. Durand, and W. T. Freeman, "Video magnification in presence of large motions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4119–4127.

[41] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–10, Jul. 2014.

[42] O. Johnston and F. Thomas, *The Illusion Life: Disney Animation*. New York, NY, USA: Disney Editions, 1981.

[43] H. Men, H. Lin, V. Hosu, D. Maurer, A. Bruhn, and D. Saupe, "Visual quality assessment for motion compensated frame interpolation," in *Proc. 11th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–6.

[44] *Information Technology—Advanced Image Coding and Evaluation—Part 2: Evaluation Procedure for Visually Lossless Coding*, document ISO/IEC 29170-2, 2015.

[45] R. S. Allison, L. M. Wilcox, W. Wang, D. M. Hoffman, Y. Hou, J. Goel, L. Deas, and D. Stolitzka, "75-2: Invited paper: Large scale subjective evaluation of display stream compression," in *SID Symp. Dig. Tech. Papers*, vol. 48, no. 1, pp. 1101–1104.

[46] R. S. Allison, K. Brunnström, D. M. Chandler, H. R. Colett, P. J. Corriveau, S. Daly, J. Goel, J. Y. Long, L. M. Wilcox, Y. M. Yaacob, S.-N. Yang, and Y. Zhang, "Perspectives on the definition of visually lossless quality for mobile and large format displays," *J. Electron. Imag.*, vol. 27, no. 5, Oct. 2018, Art. no. 053035.

[47] *Call for Proposals for a Low-Latency Lightweight Image Coding System*, document J. ISO/IEC JTC1/SC29/WG1 (ITU-T SG16), JPEG, 2016.

[48] A. Willème, S. Mahmoudpour, I. Viola, K. Fliegel, J. Pospíšil, T. Ebrahimi, P. Schelkens, A. Descampe, and B. Macq, "Overview of the JPEG XS core coding system subjective evaluations," *Proc. SPIE*, vol. 10752, Sep. 2018, Art. no. 107521M.

[49] A. Sudhama, M. D. Cutone, Y. Hou, J. Goel, D. Stolitzka, N. Jacobson, R. S. Allison, and L. M. Wilcox, "85-1: Visually lossless compression of high dynamic range images: A large-scale evaluation," in *SID Symp. Dig. Tech. Papers*, vol. 49, no. 1, May 2018, pp. 1151–1154.

[50] V. Thirumalai, J. Ribera, J. Xiang, J. Zhang, M. Azimi, J. Kamali, and P. Nasiopoulos, "P-23: A subjective method for evaluating foveated image quality in HMDs," in *SID Symp. Dig. Tech. Papers*, vol. 51, no. 1, 2020, pp. 1415–1418.

[51] S. S. Mohona, L. M. Wilcox, and R. S. Allison, "Subjective assessment of display stream compression for stereoscopic imagery," *J. Soc. Inf. Display*, vol. 29, no. 8, pp. 592–607, Mar. 2021.

[52] H. Lin, M. Jenadeleh, G. Chen, U.-D. Reips, R. Hamzaoui, and D. Saupe, "Subjective assessment of global picture-wise just noticeable difference," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2020, pp. 1–6.

[53] W. S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, Dec. 1952.

[54] R. D. Luce and E. Galanter, *Handbook of Mathematical Psychology*. New York, NY, USA: Wiley, 1963, ch. Psychophysical Scaling, pp. 245–307.

[55] D. M. Ennis, K. Mullen, and J. E. R. Frijters, "Variants of the method of triads: Utridimensional thurstonian models," *Brit. J. Math. Stat. Psychol.*, vol. 41, no. 1, pp. 25–36, May 1988.

[56] W. H. Press, B. Flannery, S. Teukolsky, and W. Vetterling, "Least squares as a maximum likelihood estimator," *Numer. Recipes Fortran, Art Sci. Comput.*, vol. 2, pp. 651–655, 1992.

[57] F. A. Wichmann and N. J. Hill, "The psychometric function: I. Fitting, sampling, and goodness of fit," *Perception Psychophys.*, vol. 63, no. 8, pp. 1293–1313, Nov. 2001.

[58] L. T. Maloney and K. Knoblauch, "Measuring and modeling visual appearance," *Annu. Rev. Vis. Sci.*, vol. 6, no. 1, pp. 519–537, Sep. 2020.

[59] C. Charrier, L. T. Maloney, H. Cherifi, and K. Knoblauch, "Maximum likelihood difference scaling of image quality in compression-degraded images," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 24, no. 11, pp. 3418–3426, 2007.

[60] C. Charrier, K. Knoblauch, L. T. Maloney, A. C. Bovik, and A. K. Moorthy, "Optimizing multiscale SSIM for compression via MLDS," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4682–4694, Dec. 2012.

[61] L. van der Maaten and K. Weinberger, "Stochastic triplet embedding," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Sep. 2012, pp. 1–6.

[62] G. T. Fechner, D. H. Howes, and E. G. Boring, *Elements of Psychophysics*, vol. 1. New York, NY, USA: Holt, Rinehart and Winston, 1996.

[63] L. T. Maloney and J. N. Yang, "Maximum likelihood difference scaling," *J. Vis.*, vol. 3, no. 8, p. 5, Oct. 2003.

[64] D. Saupe, F. Hahn, V. Hosu, I. Zingman, M. Rana, and S. Li, "Crowd workers proven useful: A comparative study of subjective video quality assessment," in *Proc. IEEE Int. Conf. Qual. Multimedia Exper. (QoMEX)*, 2016, pp. 1–2.

[65] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 3097–3100.

[66] *Amazon Mechanical Turk (AMTurk)*. Accessed: Oct. 7, 2021. [Online]. Available: https://www.mturk.com/

[67] D. McNally, T. Bruylants, A. Willème, T. Ebrahimi, P. Schelkens, and B. Macq, "JPEG XS call for proposals subjective evaluations," *Proc. SPIE*, vol. 10396, Sep. 2017, Art. no. 103960P.

[68] J. L. Punch, B. Rakerd, and A. M. Amlani, "Paired-comparison hearing aid preferences: Evaluation of an unforced-choice paradigm," *J. Amer. Acad. Audiol.*, vol. 12, no. 4, pp. 190–201, 2001.

[69] C. Leys, O. Klein, Y. Dominicy, and C. Ley, "Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance," *J. Experim. Social Psychol.*, vol. 74, pp. 150–156, Jan. 2018.

[70] S. Chawla and A. Gionis, "K-means−: A unified approach to clustering and outlier detection," in *Proc. SIAM Int. Conf. Data Mining*, May 2013, pp. 189–197.

[71] V. Hosu, H. Lin, and D. Saupe. (2018). *IQA-Experts-300: A Professional Photographer Annotated IQA Database*. [Online]. Available: http://database.mmsp-kn.de

[72] V. Hosu, H. Lin, and D. Saupe, "Expertise screening in crowdsourcing image quality," in *Proc. 10th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May 2018, pp. 1–6.

[73] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[74] J. Li, R. Mantiuk, J. Wang, S. Ling, and P. Le Callet, "Hybrid-MST: A hybrid active sampling strategy for pairwise preference aggregation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 3475–3485.

[75] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1098–1105.

[76] M. Perez-Ortiz, A. Mikhailiuk, E. Zerman, V. Hulusic, G. Valenzise, and R. K. Mantiuk, "From pairwise comparisons and rating to a unified quality scale," *IEEE Trans. Image Process.*, vol. 29, pp. 1139–1151, 2020.

[77] A. Kaipio, M. Ponomarenko, and K. Egiazarian, "Merging of MOS of large image databases for no-reference image visual quality assessment," in *Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2020, pp. 1–6.

[78] K. N. Plataniotis and A. N. Venetsanopoulos, *Color Image Processing and Applications*. Springer, 2013.

[79] J. Wu, L. Li, W. Dong, G. Shi, W. Lin, and C.-C. J. Kuo, "Enhanced just noticeable difference model for images with pattern complexity," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2682–2693, Jun. 2017.

[80] A. B. Watson, "QUEST+: A general multidimensional Bayesian adaptive psychometric method," *J. Vis.*, vol. 17, no. 3, p. 10, Apr. 2017.

[81] W. Wan, K. Zhou, K. Zhang, Y. Zhan, and J. Li, "JND-guided perceptually color image watermarking in spatial domain," *IEEE Access*, vol. 8, pp. 164504–164520, 2020.

[82] J. Wang and W. Wan, "A novel attention-guided JND model for improving robust image watermarking," *Multimedia Tools Appl.*, vol. 79, nos. 33–34, pp. 24057–24073, Sep. 2020.

[83] K. Zhou, Y. Zhang, J. Li, Y. Zhan, and W. Wan, "Spatial-perceptual embedding with robust just noticeable difference model for color image watermarking," *Mathematics*, vol. 8, no. 9, p. 1506, Sep. 2020.

[84] L. Qin, X. Li, and Y. Zhao, "A new JPEG image watermarking method exploiting spatial JND model," in *Proc. Int. Workshop Digit. Watermarking (IWDW)*. Cham, Switzerland: Springer, 2019, pp. 161–170.
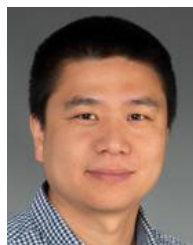
HANHE LIN received the Ph.D. degree from the Department of Information Science, University of Otago, New Zealand, in 2016. He is currently a Postdoctoral Researcher with the Department of Computer and Information Science, University of Konstanz, Germany. His research interests include machine learning and deep learning-based application, visual quality assessment, and crowdsourcing.

MOHSEN JENADELEH (Member, IEEE) received his Dr.rer.nat. degree from the Department of Computer and Information Science, University of Konstanz, Germany, in 2019. He is currently a Postdoctoral researcher with the University of Konstanz. His research interests include image processing, visual perception, machine learning, and crowdsourcing.
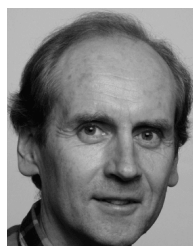
DIETMAR SAUPE was born in Bremen, Germany, in 1954. He received the Dr.rer.nat. degree in mathematics from the University of Bremen, Germany, in 1982.

From 1985 to 1993, he was an Assistant Professor with the Departments of Mathematics, first with the University of California at Santa Cruz, Santa Cruz, USA, and then with the University of Bremen, resulting in his habilitation. From 1993 to 1998, he was a Professor in computer science with the University of Freiburg, Germany, the University of Leipzig, Germany, until 2002, and since then, the University of Konstanz, Germany. He is the coauthor of the book *Chaos and Fractals* (Springer-Verlag, 1992), which won the Association of American Publishers Award for Best Mathematics Book of the Year, the book *The Science of Fractal Images* (Springer-Verlag, 1988), and well over 100 research articles. His research interests include image and video processing, computer graphics, scientific visualization, dynamical systems, and sport informatics.

• • •

HUI MEN received the master's degree from the Department of Computer Science, University of Saarland, Germany, in 2016, and the Ph.D. degree from the Department of Computer and Information Science, University of Konstanz, Germany, in 2021. She is currently a Postdoctoral Researcher in the Department of Psychology, University of Marburg, Germany. Her research interests in include visual quality assessment, crowdsourcing, frame interpolation, and visual perception.