Large-scale crowdsourced subjective assessment of picturewise just noticeable difference

Hanhe Lin[a,d], Guangan Chen[a], Mohsen Jenadeleh[a], Vlad Hosu[a], Ulf-Dietrich Reips[b], Raouf Hamzaoui[c], and Dietmar Saupe[a]

***Abstract*—The picturewise just noticeable difference (PJND) for a given image, compression scheme, and subject is the smallest distortion level that the subject can perceive when the image is compressed with this compression scheme. The PJND can be used to determine the compression level at which a given proportion of the population does not notice any distortion in the compressed image. To obtain accurate and diverse results, the PJND must be determined for a large number of subjects and images. This is particularly important when experimental PJND data are used to train deep learning models that can predict a probability distribution model of the PJND for a new image. To date, such subjective studies have been carried out in laboratory environments. However, the number of participants and images in all existing PJND studies is very small because of the challenges involved in setting up laboratory experiments. To address this limitation, we develop a framework to conduct PJND assessments via crowdsourcing. We use a new technique based on slider adjustment and a flicker test to determine the PJND. A pilot study demonstrated that our technique could decrease the study duration by 50% and double the perceptual sensitivity compared to the standard binary search approach that successively compares a test image side by side with its reference image. Our framework includes a robust and systematic scheme to ensure the reliability of the crowdsourced results. Using 1,008 source images and distorted versions obtained with JPEG and BPG compression, we apply our crowdsourcing framework to build the largest PJND dataset, KonJND-1k (Konstanz just noticeable difference 1k dataset). A total of 503 workers participated in the study, yielding 61,030 PJND samples that resulted in an average of 42 samples per source image. The KonJND-1k dataset is available at http://database.mmsp-kn.de/konjnd-1k-database.html**

***Index Terms*—Just noticeable difference (JND), satisfied user ratio (SUR), crowdsourcing, flicker test, JPEG, BPG, dataset**

I. Introduction

IMage compression is essential to meet constraints on transmission bandwidth and storage. With increasing compression levels, an increasing number of users can perceive distortion in the compressed version of an original image. The smallest distortion level that a user can perceive when an image is compressed is called the picturewise just noticeable difference (PJND). Since physiological and visual attention mechanisms vary from one user to another, the PJND varies according to the user. The satisfied user ratio (SUR) is the fraction of users who cannot notice any distortion artifact when comparing an original image with its compressed version for a given distortion level. In mathematical terms, the PJND is a random variable, and the SUR is its complementary cumulative distribution function. Modeling the SUR can help content providers minimize bandwidth costs while guaranteeing user satisfaction for any target proportion of their customers.

To determine a statistical model for the PJND, subjective studies are required. Such studies typically consist of three steps. First, a number of source images are compressed according to a compression scheme at different distortion levels. Next, a group of subjects are asked to identify the PJND of these images (see Section II). Finally, a probability distribution model is fitted to the PJND data.

PJND assessment methods can be categorized according to the search strategy. Two baseline methods are *linear search* and *full search*. In linear search, the reference image is compared with the sequence of compressed images with increasing compression levels (decreasing bit rate) until a difference is noticed. The linear search method is called the "method of limits" in psychophysics. Unlike the linear search method, the full search method conducts randomized comparisons with all compressed images. As these two baseline methods may include many redundant comparisons, more efficient search strategies were developed.

- *Binary search.* The baseline methods can be sped up with a binary search algorithm that quickly narrows down the range of the PJND, resulting in fewer subjective comparisons. The *relaxed binary search* is a more robust version with respect to the nondeterministic outcomes of comparisons. It proceeds by scaling the size of the bracketing interval by 3/4 instead of 1/2 in each iteration.
- *Paired comparisons with scale reconstruction.* Subjects identify the image with higher quality among many sampled two-alternative forced-choice (2AFC) comparisons. A pair for which one of the images collects 75% of the votes is considered to have a perceptual distance equal to 1 JND.

PJND assessment methods can be further grouped according to the way the images are presented for comparison. The reference and the test image can be displayed sequentially or simultaneously for a specified duration. In the latter case,

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 (Project A05) and DFG Project-ID 496858717. Corresponding author: Hanhe Lin, E-mail: h.lin2@rgu.ac.uk

[a]Department of Computer and Information Science, University of Konstanz, 78464 Konstanz, Germany.

[b]Department of Psychology, University of Konstanz, 78464 Konstanz, Germany.

[c]School of Engineering and Sustainable Development, De Montfort University, LE1 9BH Leicester, UK.

[d]National Subsea Centre, Robert Gordon University, AB10 7AQ Aberdeen, UK.

the two images can be viewed side by side or on top of each other, alternating at a certain frequency, a method called the flicker test. For example, the JPEG-XS standard has adopted a flicker test [1] where the reference image and the compressed image alternate at a frequency of 8 Hz.

Note that the PJND depends on how images are assessed and compared. The results depend, for example, on display size, viewing distance, environmental conditions, and whether a single- or dual-stimulus method is used. Therefore, the actual PJND values in a study or application depend on the choice of technique for their measurement.

Conventional subjective PJND assessment studies, however, have two major limitations. First, it is very time-consuming to conduct these studies since they require many comparisons, even when the more efficient binary search strategy is adopted. Second, recruiting participants and setting up the experiment in a laboratory environment is challenging and expensive. In particular, the cost for personnel and participant remuneration can become prohibitive for large-scale studies [2]. As a result, the content of the existing PJND datasets is limited. This makes them unsuitable for developing computational or objective PJND methods, in particular deep learning based approaches [3], [4], [5], which require training on large-scale data with content diversity to ensure their generalizability.

To address the first limitation, we use a method based on slider adjustment. This method allows obtaining measurements that are comparable to those of a binary search while being significantly faster. To address the second limitation, we rely on crowdsourcing, as it has been shown that crowd workers can provide reliable data under an appropriate experimental setup [6] and proper quality control [7], [8].

The main contributions of this paper are summarized as follows.

- We use a new PJND assessment method that significantly differs from traditional visual quality assessment methods used in previous just noticeable difference (JND) [9], [10] and PJND [11], [12] subjective tests. Combining slider-based adjustment with the flicker test, our PJND assessment method can obtain measurements that are comparable to those of binary search while being significantly faster.
- We propose a novel and robust framework for conducting subjective PJND assessment studies via crowdsourcing. A quality control scheme is developed and used to ensure the reliability of the study. While crowdsourcing has been used in various subjective visual quality assessment tests [13], [14], [15], our work is the first study that exploits it to collect PJND samples. Designing a crowdsourced experiment for PJND estimation is different from designing it for a traditional visual quality assessment test because the crowd worker must find a critical level where distortion in a signal becomes visible rather than rate the visual quality of one signal independently or with respect to another signal.
- We create the largest PJND dataset and call it the Konstanz just noticeable difference 1k dataset (KonJND-1k). It contains 1,008 source images, together with distorted versions obtained with two compression schemes: JPEG and BPG. For each image, an average of 42 samples by 503 crowd workers is collected via Amazon Mechanical Turk (AMT). Compared to existing PJND datasets (Table I), KonJND-1k is five times larger than the largest dataset [16] in terms of the number of source stimuli and ten times larger in terms of the number of PJND samples. The KonJND-1k dataset is available at http://database.mmsp-kn.de/konjnd-1k-database.html.

II. Related work

In this section, we provide an overview of the state-of-the-art PJND-based image and video datasets (Table I). We also review other works related to the subjective PJND assessment method proposed in this paper.

A. PJND-based image datasets

Jin *et al.* [11] conducted subjective quality assessment tests to collect PJND samples for JPEG compressed images and built a dataset called MCL-JCI. The tests involved 150 participants and 50 source images with a resolution of 1920×1080. From each source image, 100 compressed versions were generated by varying the JPEG quality factor (QF) from 1 (lowest) to 100 (highest). The image compressed with QF $=$ 100 was used as a reference image. The reference image and a compressed image were displayed side by side on a 65-inch TV with a resolution of 3840×2160 to determine whether they are noticeably different. The viewing distance was 2 m from the center of the monitor. For a given image, PJND samples were collected from 30 subjects. The standard binary search was used to speed up the process. The study found that observers could distinguish only a few distortion levels (five to seven).

Shen *et al.* [16] [12] created a PJND-based image quality dataset using 202 pristine images and their 7,878 encoded versions using versatile video coding (VVC). All reference images were cropped to a uniform aspect ratio of 16:9 and then downsampled to a resolution of 1920×1080. Each reference image was compressed by VTM 5.0 intracoding with QP ranging from 13 to 51. The subjective tests were performed in a controlled laboratory environment, and a Samsung Q7F 55-inch smart UHD TV was used as the display device. Subjective experiments on subjective perception were conducted using pairwise comparisons between a reference image and its encoded versions to assess each JND sample. For each reference image, 20 PJND samples were assessed by 20 subjects. A standard binary search was used to seek the PJND.

Fan *et al.* [21] studied the PJND of symmetrically and asymmetrically compressed stereoscopic images for JPEG 2000 and H.265 intracoding. The study considered 12 stereo images and was conducted by 36 subjects for creating the SIAT-JSSI dataset. Stereo image pairs (one source pair and one distorted pair) were shown side by side on a 65-inch 3D monitor with a native resolution of 3840×2160. The subjects wore polarized glasses and were seated 1.6 m from the monitor. The relaxed binary search was used to collect the PJND sample from a subject. Outlier subjects were detected, and their PJND samples were removed.

TABLE I: Comparison of the state-of-the-art JND-based datasets with KonJND-1k.

Datasets	MCL-JCI	MCL-JCV	Huang *et al.*	VideoSet	JND-Pano	SIAT-JSSI	Shen *et al.*	KonJND-1k
Reference	[11]	[17]	[18]	[19]	[20]	[21]	[12][16]	–
Publication year	2016	2016	2017	2018	2019	2020	2021	2021
Stimuli type	image	video	video	video	panoramic image	stereoscopic image	image	image
Number of references	50	30	40	220	40	12	202	1,008
Resolution of references	1920×1080	1920×1080	1920×1080	1920×1080	5000×2500	1920×1080	1920×1080	640×480
Distortion type	JPEG	H.264/AVC	H.265/HEVC	H.264/AVC	JPEG	JPEG 2000 or H.265/HEVC	H.266/VVC VTM 5.0	JPEG or BPG
Distortion levels per each stimulus	100	51	51	51	100	300 or 51	39	100 or 51
Total number of stimuli	5,050	1,560	2,080	45,760	4,040	3,510	8,080	77,112
Test environment	lab	lab	lab	lab	lab	lab	lab	crowdsourcing
Subjective assessment method	Customized[a]	Customized[b]	Customized[a]	Customized[b]	Customized[c]	Customized[a]	Customized[a]	Customized[d]
Total number of JND samples[e]	1,500	1,500	1,200	6,600	970	442	4,040	41,866
Average JND samples per reference	30	50	30	30	24	36	20	42

[a] Observers were asked to compare two stimuli displayed side by side and determine whether the differences between them are noticeable.
[b] Observers were asked to determine whether the differences between the two video clips displayed one after another are noticeable.
[c] Subjects wore a head-mounted display (HMD) device and were free to control the field of view to compare two panoramic images and determine whether they could see a noticeable difference or not.
[d] A flickering image was displayed, and observers were asked to determine if they could see a flicker effect.
[e] For comparison purposes, here we only consider the samples of the first JND level.

Liu *et al.* [20] created a PJND dataset called JND-Pano for panoramic images viewed using a head-mounted display. The study included 40 source images with a resolution of 5000×2500. JPEG was used to compress each source image with 100 quality factors. The reference image and a compressed image were displayed simultaneously in random order. For each source image, the test included at least 25 observers. A standard binary search was used to identify the PJND. Outliers were removed based on the range and standard deviation.

B. PJND-based video datasets

Wang *et al.* [17] considered 30 source video sequences with a resolution of 1920×1080, a duration of 5 s, and different frame rates. They compressed the video sequences by varying the quantization parameter (QP) of the H.264/AVC video coder from 1 (smallest distortion) to 51 (largest distortion). More than 150 people participated in the study. The viewing distance and display monitor were as in [11]. The video sequence corresponding to QP equal to 1 was used as a reference. The reference and a distorted version were displayed one after another. JND samples were collected from 50 subjects. The standard binary search was used to speed up the process. The resulting PJND dataset was called MCL-JCV.

The study of Lin *et al.* in [22] involved five source images and five video sequences with a resolution of 1920×1080. The images were encoded with JPEG, while the video sequences were encoded with H.264 and H.265. The viewing distance and display monitor were as in [11]. The standard binary search was used to speed up the process.

Wang *et al.* [19] built a large-scale JND video dataset called VideoSet for 220 source videos with durations of five seconds and in four resolutions (1080p, 720p, 540p, 360p). Each source video was compressed with H.264 using QP values from 1 to 51. The viewing distance was set according to the ITU-R BT.2022 recommendation. The source video and a distorted version were displayed one after another. A relaxed binary search was used to collect the PJND sample from each subject. At least 30 observers were involved in the PJND estimation of each video sequence. Unreliable subjects and outlying samples were removed.

Huang *et al.* [18] generated a PJND-based dataset of encoded videos with H.265. The dataset contains 40 high-definition (HD) reference video clips with a frame rate of 30 fps and a duration of five seconds. All reference videos were encoded using the HM 16.0 reference software, with QP values ranging from 0 to 51. To assess the first PJND threshold for each reference video and its 51 encoded versions, a subjective test was conducted with 30 subjects using the pairwise comparison method. The reference video and an encoded version were played side by side time-synchronously on Samsung UN65 F9000 65-inch 4K UHD TV display in a laboratory environment. The standard binary search was used to select the displayed stimuli. Outlying samples were excluded with the three-sigma rule.

C. Other relevant work

Wang *et al.* [23] conducted a subjective experiment to evaluate the just perceptible differences of four important attributes that affect laser projection television: white level, black level, color saturation, and contour rendering. Nine source images with a resolution of 434×434 were used as reference. A subjective test was conducted to evaluate the first four JNDs of the above attributes by a two-alternative forced-choice method with a one-on-two staircase method. A 100-inch laser projection digital light processing (DLP) television using blue and red laser diodes with 4K resolution (3840×2160) was used for the experiment. The maximum and minimum luminance values were $275.93\,\mathrm{cd/m^2}$ and $0.17\,\mathrm{cd/m^2}$, respectively. The experiment was conducted in a dark room, and the TV was set to standard mode. The viewing distance was set to 1.97 m.

Hoffman and Stolitzka [24] proposed tests to determine whether a compressed image differs from a reference image by at most one JND. The testing environment was implemented according to ISO 3664. The monitor used had a 24.3-inch diagonal and a resolution of 1920 × 1200. A reference (uncompressed) image and an image consisting of the alternating reference image and a distorted version were presented side by side. The observer had to identify which of the two images was nonflickering. A dataset of approximately 250,000 responses collected from 35 observers for 18 images was created. The flicker method proposed in this paper was found to significantly increase the visibility of image artifacts and adopted as a standard [1], suitable for testing images compressed and reconstructed in a visually lossless manner.

Zhang *et al.* [25] collected a large-scale dataset of perceptual judgments, which included asking subjects whether one reference patch and one distorted patch are identical. They used 20 types of distortions (e.g., photometric distortions, noise, blurring, and compression artifacts) and sequentially composed pairs of distortions. The two patches had a resolution of 64×64 and were shown for 1 s each, with a 250 ms gap in between.

Redi *et al.* [26] compared the performance of absolute category rating obtained by the single stimulus (SS) method with that of the quality ruler (QR) method. The QR consists of a series of reference images varying in a single attribute (sharpness), with known and fixed quality differences between the samples, given by a certain number of JND units [27]. For the QR method, the quality of an input image is compared to the image qualities on the ruler. The study showed that QR scores have narrower confidence intervals than SS scores.

Visual analog scales, similar to sliders, for assessing perceived quantities, such as length and area, or sensory stimuli, such as loudness or taste, have been studied in psychology and have been shown to be reliable measurement tools [28].

III. Pilot study for PJND assessment methods

In our pilot study [29], we introduced and evaluated two new adjustment methods for subjective PJND assessment, where a subject interactively selects the distortion level at the PJND location by a slider or keystrokes. For both methods, a reference image and its distorted version are compared using the flicker test in which the displayed images alternate at a frequency of 8 Hz, as in the JPEG-XS standard [1]. The images were selected from the MCL-JCI dataset [11]. More than 14 participants took part in the study. Experimental results showed that the PJND samples obtained with our adaptation methods are comparable to those obtained with the relaxed binary search method while being 1.5 to 2 times faster. Moreover, the flicker test provided approximately twice the sensitivity of a side-by-side comparison.

Based on the results of the pilot study, we used the slider-based adjustment method with the flicker test for conducting the first JND-based crowdsourced subjective study.

IV. Image dataset creation

We collected source images from Pixabay.com, a website for sharing royalty-free images, videos, and music. All the images on the website are released under the Pixabay license, which grants the right to edit and redistribute them. Furthermore, the perceptual quality of the uploaded images is generally very good since up to 20 independent Pixabay users cast their votes for accepting or declining an uploaded image based on its perceptual quality. The precise mechanism for deciding which images are considered high quality is not made public. Images on Pixabay are labeled by their authors in categories such as "photo" and "illustration". Thus, we started from the collection of 1,244,635 "photos" and ignored all images that were labeled differently.

Despite the perceptual quality screening provided by Pixabay, not all images are of high quality. For this reason, we selected a subset S_{sampling} of 10,000 images that have characteristics indicative of high quality and are larger than 1024 pixels in both width and height, having an aspect ratio (width by height) between 1.31 and 1.78. In a second step, this subset was reduced to the 1,120 images of our dataset.

Images on Pixabay are annotated by the community, with each assigned a number of favorites, likes, and downloads. While these measures are to some extent correlated with the quality of an image, they are also strongly correlated with how often an image has been viewed. For instance, if an image is viewed more frequently, it is likely to receive more favorites and likes and to be downloaded more often.

Schwarz *et al.* [30] proposed normalizing the favorites $F(I)$ received by an image I, relating them to the number of views $V(I)$ by considering $\log(F(I))/\log(V(I))$. While the ratio accounts for the increase in favorites relative to the number of views, the logarithm accounts for the typical exponential-like initial increase in the number of views. We modified the formula slightly to

$$F_{\text{norm}}(I) = \frac{\ln(F(I) + e)}{\ln(V(I) + e)} \in (0, 1],$$

where e is Euler's constant. The unit interval range of F_{norm} is easier to work with. Similarly, we computed normalized downloads D_{norm} and likes L_{norm}.

A low value for the normalized indicators means that the image has received a low fraction of likes, favorites, and downloads. We considered only images that have at least 100 views so that the normalized indicators are statistically meaningful. Low normalized values of the indicators suggest that the image is not appreciated from the corresponding three points of view. Thus, we defined a joint score S to rank the images. First, we ranked the images by each of the normalized indicators and mapped the rank-order indices to $[0, 1]$. We denote the normalized ranking function of an image I for the popularity measures D, L, F by $R_D(I), R_L(I), R_F(I)$, respectively. The combined ranking score, defined as

$$S(I) = (1 - R_D(I)) \cdot (1 - R_L(I)) \cdot (1 - R_F(I)),$$

assigns a high score to images that do not have any low normalized indicator for any of the three attributes. We took the top 10,000 images ranked by decreasing values of $S(I)$ from the larger sampling subset S_{sampling}.

These images had aspect ratios (width by height) between 1.31 and 1.78 but were then scaled and cropped to a size of 640 × 480 pixels as follows. If the aspect ratio is smaller than 640 × 480, we scaled the image to a width of 640; otherwise, we scaled the image to a height of 480. Finally, we center-cropped the scaled image to a size of 640 × 480. This choice

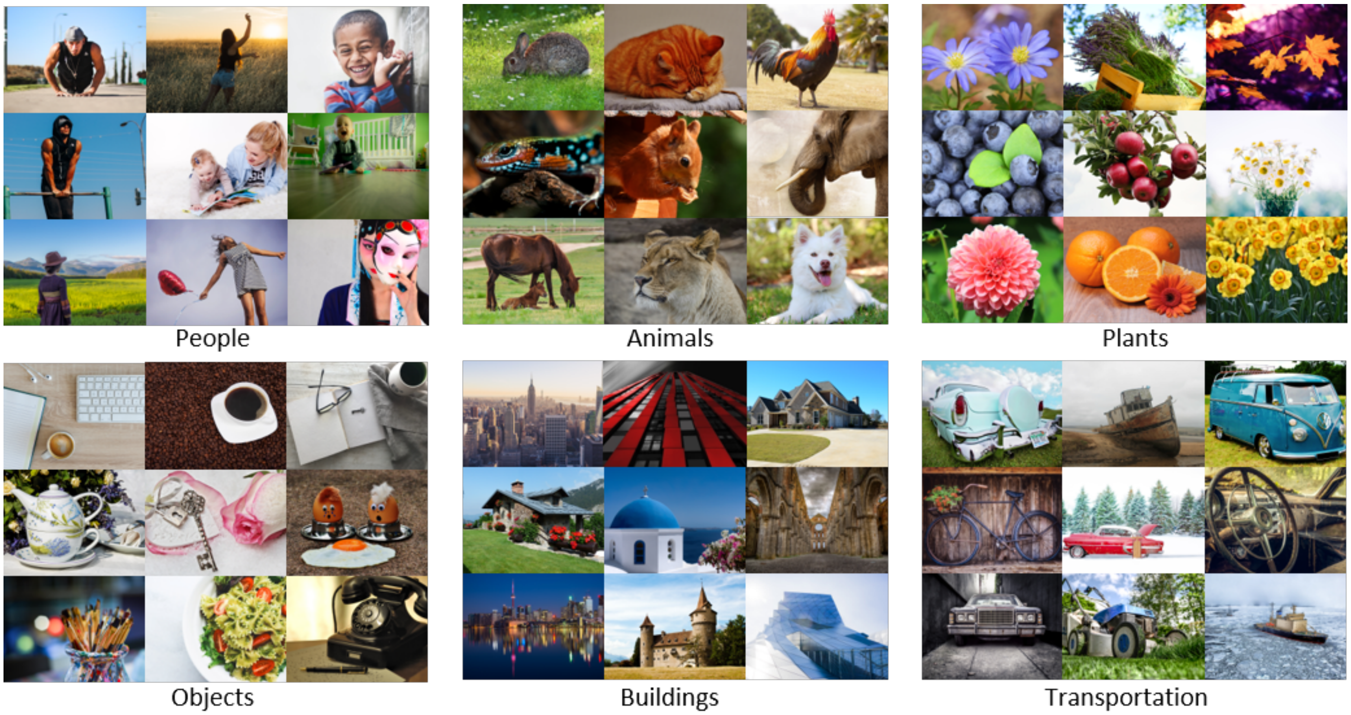


Fig. 1: Samples of source images in the KonJND-1k dataset.

of image resolution was motivated by the layout of the user interface in our experiments and the minimum required screen resolution; see Subsection VI-C.

In the second step, to ensure that the selected images had enough content diversity, we extracted their 2048-dimensional deep features from a ResNet50 model [31] pretrained on ImageNet [32]. We next applied k-means clustering on the deep features to partition the 10,000 images into 1,120 clusters. One image was randomly selected from each cluster as a representative. We randomly split these 1,120 selected images into a *study set* and a *test set*. The study set contained 1,008 images for subjective PJND assessment. The test set contained 112 images for quality control (see Section VI-D). We applied two image compression methods, namely, JPEG and BPG[1], for both sets, where half of the images in each set were compressed with JPEG, while the second half was compressed with BPG.

As shown in Fig. 1, the content of the sampled images is diverse, including categories of people, animals, plants, objects, buildings, and transportation.

V. PJND ASSESSMENT METHOD

For each reference image I, we obtained a sequence $I_d, d = 0, 1, \ldots, 100$ where $I_0 = I$ and I_d is the compressed version at distortion level d. We used two compression types, namely JPEG and BPG. For JPEG compression, we have distortion levels $d = 101 - \text{QF}, d = 1, \ldots, 100$, where QF denotes the JPEG quality factor. For BPG compression, the relation between distortion level d and quantization parameter QP is $\text{QP} = \lceil d/2 \rceil, d = 1, \ldots, 100$. Thus, in both cases, as the distortion level d increases from 0 to 100, the bit rate decreases. For each reference image I, our objective is to search for its PJND among the images I_d.

We used a flicker test, with the reference and the compressed test image being displayed successively at a frequency of 8 Hz. Using this display scheme, we implemented a subjective assessment method for the PJND.

The PJND assessment method is a slider-based adjustment method, as shown in the left part of Fig. 2. The handle of a slider controls the distortion level of a study image. Subjects could either drag the slider or press the left/right arrow keys to move the slider to the position corresponding to the smallest distortion level with noticeable flicker.

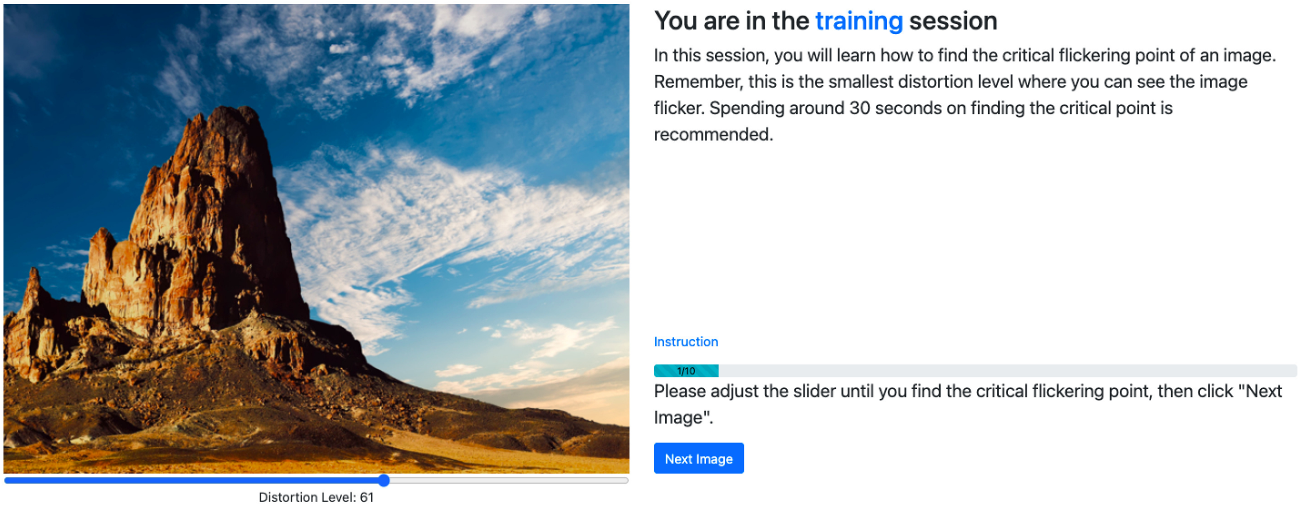
VI. CROWDSOURCED SUBJECTIVE PJND ASSESSMENT

The size of the subjective PJND assessment study for 1,008 images suggests an implementation in a crowdsourcing setting rather than in the laboratory. For this purpose, we chose Amazon Mechanical Turk (AMT). On the AMT platform, *requesters*, who are companies, organizations, or persons, create and submit *human intelligence tasks* (HITs) for *workers*. Workers can work on an HIT, submit the results of their work for it, and collect a reward for completion. Requesters can specify the number of *assignments* for an HIT, i.e., how many workers can submit completed work for the HIT. Each worker can submit only one assignment for each HIT.

Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of the University of Konstanz.

In the following, we explain how we conducted the subjective PJND assessment study in detail.

[1] https://bellard.org/bpg/



You are in the training session

In this session, you will learn how to find the critical flickering point of an image. Remember, this is the smallest distortion level where you can see the image flicker. Spending around 30 seconds on finding the critical point is recommended.

Please adjust the slider until you find the critical flickering point, then click "Next Image".

Fig. 2: Screenshot of the user interface for the PJND subjective study. Workers drag the slider or press left/right arrow keys to move the slider to the position of distortion level, where they start perceiving the flicker. When they finish a PJND assessment, they click ‘Next Image’ for the next assessment. Additional information, such as the progress bar and instructions for the experiment, is provided as well.

A. *Overview*

Fig. 3 shows the workflow of the crowdsourced subjective PJND assessment study. The study contains two types of HITs, a qualification HIT and multiple study HITs. The qualification HIT is used to find and train qualified workers; only qualified workers are allowed to participate in the study HITs.

B. *Instructions*

For both types of HITs, a page with instructions was presented to the workers. The page contains four sections. In the first section, a flicker test image is shown and the critical point, i.e., the earliest (left-most) slider position where an observer starts perceiving the flicker, is introduced. Finally, the purpose of the study, namely finding the critical point by moving the slider, is explained. The detailed steps of the study and the keyboard shortcuts are explained in the second and third sections, respectively. In the final section, a video of an example of how to conduct the study is displayed. Furthermore, a few examples of ‘no flickering’, ‘just noticeable’, and ‘severe flickering’ are presented.

C. *Qualification HIT*

Workers who wanted to join our study had to pass the qualification HIT beforehand. In the following, we describe each step in detail.

Step 1 Eligibility check: To select experienced workers, we required that participants had at least 200 HITs approved by previous requesters and that their approval rate of completed HITs was greater than or equal to 99%. These two requirements were checked by AMT. Workers were able to browse and accept our submitted HITs only if they met the requirements.

In addition, participants were not allowed to continue with the experiment unless the configuration of their device satisfied the following requirements:

- Desktops and laptops are allowed, while mobile phones and tablets are not.
- Use of a Chromium-based browser such as Google Chrome.
- The display screen must have a minimum logical resolution of 1366 × 768.

If any of the requirements was not met, a warning message was displayed, and the experiment was stopped.

Step 2 Calibration: One of the most challenging problems in conducting a PJND study on AMT is that the screens used by workers have different sizes and resolutions. As a result, an image with a specific resolution may have a different physical size when displayed on these screens. Therefore, to align the viewing conditions, we displayed all images at the same physical size and fixed the viewing distance for all workers.

To display the entire graphical user interface (GUI) of our subjective experiment, workers were required to have a minimum display size of 13.3 inches.

After imposing the minimum resolution constraint in Step 1, we checked the display size of workers’ screens by calibration.

Workers were asked to prepare a credit card with a size of 85.60 mm × 53.98 mm or a card of the same size and adjust the size of a frame on the screen until the frame fits the card size (Fig. 4). From the size in pixels of the frame matching the credit card, we calculated the logical pixel density (LPD) of the display in pixels per inch (PPI). From the known screen resolution in pixels, we then estimated the physical size of the display. Workers with a screen size less than 13.3 inches were not allowed to continue.

Displaying an image with a resolution of 640 × 480 on a screen with a logical resolution of 1366 × 768 and a physical size of 13.3 inches requires a physical dimension of 13.797 cm × 10.347 cm. We displayed our test images on all workers’ screens in this physical size.

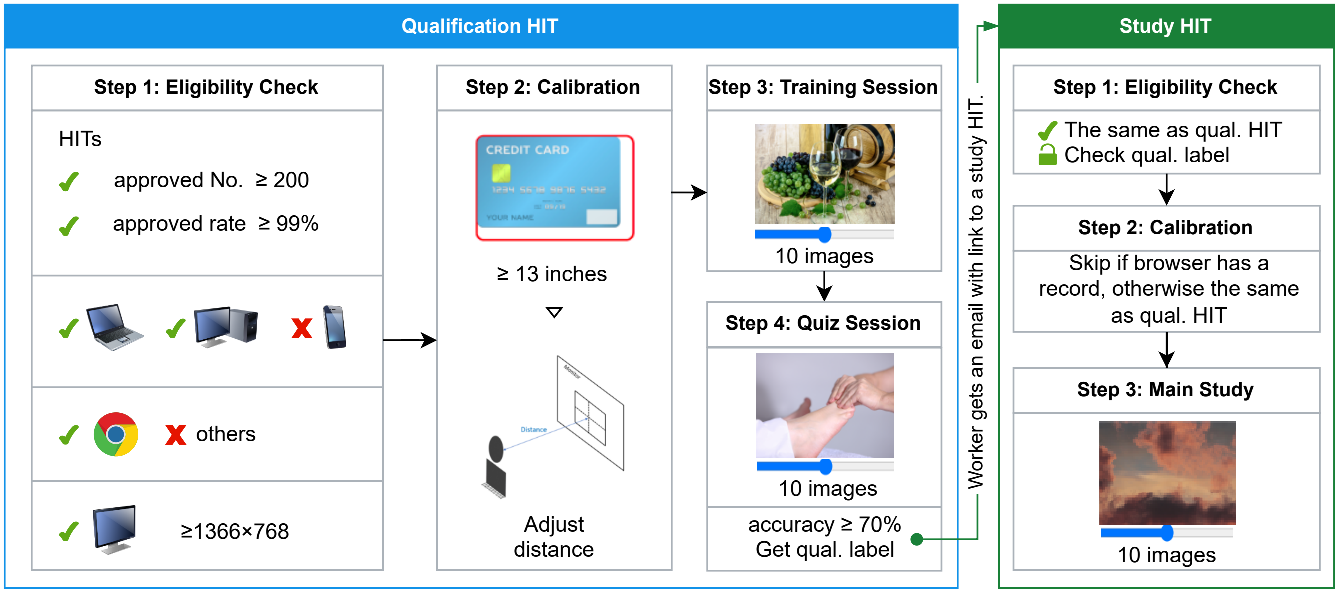
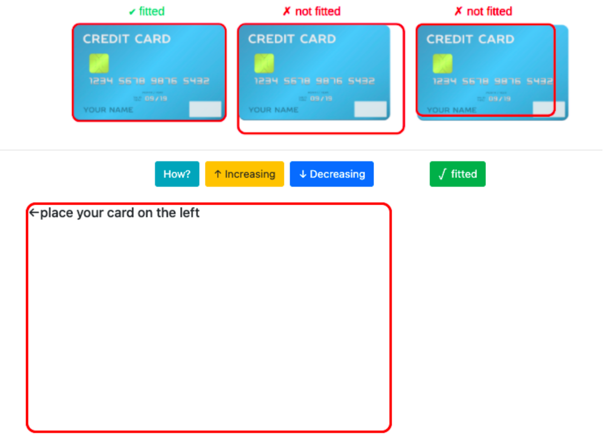


Fig. 3: Workflow of the crowdsourced subjective PJND assessment study. To be eligible for a study HIT, a worker must pass a qualification HIT. Once a study HIT is finished, AMT shows the next one automatically. A qualified worker may participate in up to 30 study HITs.



←place your card on the left

Fig. 4: Screenshot of the user interface for calibration. Workers adjust the frame to match the size of a credit card by using up and down arrow keys on their keyboard or clicking the 'Increasing' and 'Decreasing' buttons, followed by clicking the 'fitted' button to submit calibration results.

The calibration record was stored on the local storage of the browsers of the workers. Workers were not required to repeat the calibration for the following study HITs, provided they used the same computer and display device. Their browsers were blocked whenever they changed the browser zoom level after calibration. In this case, workers had to switch back to the original zoom level or redo the calibration.

After the calibration was finished, we asked workers to adjust their viewing distance to 30 cm. The suggested viewing distance was derived according to trigonometric calculation [33], [34] and ISO standard [35].

Step 3 Training session: In our study, we defined a *question* as asking a worker to identify the PJND of a given image w.r.t. one of the two compression schemes, JPEG or BPG. We provided workers with 10 training questions to guide them on how to use the interface. These 10 training images were selected manually from the downloaded subset S_{sampling}.

For the training, five images were compressed with JPEG, and the other five were compressed with BPG. The order of the questions was randomized at the beginning of each training session. Workers were allowed to work on a question only after all required compressed images had been loaded. While loading images, the slider and button were disabled, and a spinner icon was displayed. While workers were answering a question, the required images of the next questions were already loaded in the background.

The acceptable ranges of the answers for the training questions were determined by an internal study, where 10 subjects were invited to conduct the same study. The range was set to the interval centered at the rounded mean of 10 samples and with the width of two standard deviations.

The user interface to conduct the PJND subjective study is shown in Fig. 2. On the right part of the interface, workers are informed of their current session, and a study time is recommended to them (30 s). Workers are allowed to read the instructions whenever they want. A progress bar is presented to visualize the progression of the session. After conducting a PJND assessment in the left part of the interface, workers could click the 'Next Image' button. If the assessment result is correct, they are allowed to work on the next question. Otherwise, they are informed that their answer is not correct and a range for the slider position is suggested. Workers are

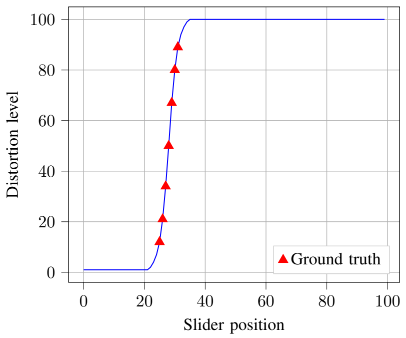


Fig. 5: Example of ground truth generation of quiz questions. The images in the training session are distorted authentically, i.e., the slider position is equal to the distortion level. In contrast, images in the quiz session are distorted nonlinearly such that the distortion level increases from 0 to 100 in a small interval of slider positions. This allows us to place a small ground truth range at arbitrary slider positions. In the example of the figure, the ground truth range was set to $[26, 32]$ (red triangles), where the image shown at slider position 29 corresponds to distortion level 50.

allowed to go to the next question only if they have moved the slider into the correct range.

Step 4 Quiz session: Workers were allowed to take part in a study HIT only if they had passed a quiz. There were 10 quiz questions, where images in five questions were compressed with JPEG, and the images for the other five questions were compressed with BPG. Although the content of the quiz images was different from that of the training images and test images, the steps for answering a quiz question were the same as those for answering a training question, except that no range of valid distortion levels was revealed to the workers.

To make the quiz questions objective and fair, unlike training images where the slider position corresponds to the compression levels, images in the quiz session were compressed nonlinearly such that the distortion level increases from 0 to 100 over an arbitrary range of only approximately 20 consecutive slider positions. This was achieved by letting a sigmoid function of the slider position determine the corresponding distortion levels. The ground truth range was set to those slider positions with a distance of at most 3 to the location whose distortion level was equal to 50. Fig. 5 shows an example of how to generate the ground truth range of a quiz question.

Once a worker finished the quiz, a script on our server downloaded the result and calculated the quiz accuracy immediately. Workers with an accuracy greater than or equal to a given threshold (70% in our experiment) were assigned an AMT qualification label and sent an email with a link to the study HITs. Workers with a lower accuracy received an email to notify them about their failure and were not allowed to take the qualification HIT a second time.

D. Study HITs

Workers who passed the quiz in the qualification HIT were allowed to perform the study HITs. The details are as follows.

Step 1 Eligibility check: Each worker was required to meet the eligibility requirements as explained in Step 1 of Subsection VI-C. In addition, a qualification label issued for the id of the worker was necessary. A new worker who did not have a qualification label but accessed a study HIT would be notified by AMT to first participate in the qualification HIT.

Step 2 Calibration: The local browser storage was checked to determine whether the workers had already performed the calibration. If there was such a record, no action was required; otherwise, they had to perform the calibration again, as in the qualification HIT.

Step 3 Main study: Each study HIT had 10 images to be assessed. Nine of them were from the study set, and one was from the test set. For this purpose, we randomly partitioned the 1,008 images of the study set into nine subsets of 112 images each. Together with the set of 112 test images, we thus had 10 equally sized sets. Each HIT was randomly assembled by sampling one image from each of the subsets without replacement. In this way, we obtained 112 randomly assembled study HITs. Whenever a worker processed an assignment for a study HIT, the order of its 10 images was randomized to avoid any bias due to a fixed sequence in all assignments for an HIT. Each qualified worker was allowed to complete at most 30 study HITs.

The cumulative accuracy of a worker on test questions was calculated by running a script on a server. Workers with an accuracy below 70% after 10 HITs were disqualified and therefore were not allowed to perform more HITs. All data generated by disqualified workers was discarded without replacement. Thus, the ground truth generation of test questions was the same as that for quiz questions.

VII. Results

A. Setup

We recruited workers by posting 600 assignments of the qualification HIT. For each study HIT, we collected 50 assignments, i.e., 50 PJND samples per image. In total, 61,030 samples were collected, where the number of samples on the quiz, test, and study images was 5,030, 5,600, and 50,400, respectively.

For each question, i.e., for each PJND assessment, we recorded the following information for further analysis:

- Slider duration: the time difference between the first and the last slider interaction.
- Number of slider direction changes, i.e., the number of times a slider changed its moving direction.

B. Worker analysis

Of the 503 workers who participated in the qualification HIT, 371 ($\approx$ 74%) passed the quiz. Fig. 6 compares the cumulative distribution function (CDF) of the slider duration and the number of slider direction changes between failed and passed assignments in the qualification HIT. It can be observed that workers who failed the qualification HIT spent less time on it and moved the slider less frequently.

Of the 371 workers who passed the quiz, 317 continued to submit at least one assignment of a study HIT. Fig. 7

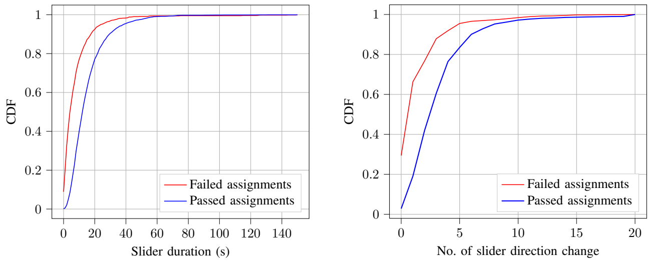


Fig. 6: Cumulative distribution function (CDF) of slider duration (left) and number of slider direction changes (right) in the qualification HIT. Workers passed the HIT when the accuracy of their answers to the 10 quiz questions was at least 70%.

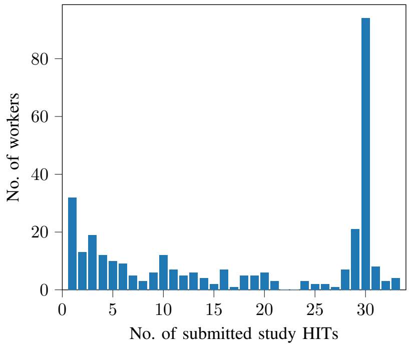


Fig. 7: Histogram of submitted study HITs by workers. In total, 142 workers out of 317 submitted 25 or more HITs.

shows the histogram of submitted study HITs. Although the maximum number of assignments was set to 30 per participant, 15 workers were able to complete a few more due to the communication delay between our lab server and AMT.

Info	Category	Number
Resolution	1366×768	116
	1440×900	22
	1536×864	50
	1600×900	17
	1920×1080	84
	Other×Other	19
OS	Linux	15
	Mac	29
	Windows	274
OS language	English	249
	Portuguese	48
	Spanish	6
	Italian	6
	All others	9

TABLE II: Worker statistics for the study HITs.

Table II presents statistics of the workers who participated in the study HITs. Most of them used screens with resolutions of 1366×768 and 1920×1080. Windows was the most often used OS, and the main language was English.

C. Outlier removal

The qualification HIT cannot ensure that all unreliable workers were disallowed for the study HITs. For example, random clickers might have passed the quiz by chance. Moreover, some of the reliable workers who passed the quiz might have occasionally paid insufficient attention in the following study HITs during their work. Therefore, we detected and removed outliers on three levels as follows.

Worker level: We assumed that workers who completed at least 10 HITs and whose accuracy on the test questions was less than 70% did not pay full attention to the study HITs. There were 10 such workers. Hence, we removed all the 137 assignments that they had submitted. This reduced the number of samples from 50,400 to 49,167.

HIT level: Despite the test questions to control the quality of the experiment, the data might still be noisy due to various reasons. For instance, workers might submit unreliable results after working a very long time without a break. In this case, we should remove these results. Therefore, we propose a robust HIT-level outlier removal method based on worker consensus.

Algorithm 1: HIT-level outlier removal

```
Input:
$\mathcal{H}$: set of all study HITs;
$\mathcal{A}(H)$: set of assignments for study HIT $H \in \mathcal{H}$ (exclude test questions);
$0 < p < 1$: target fraction of retained assignments;
$n_{\max}$: maximum number of iterations;
$r, s$: parameters to be chosen;
$\mathcal{A} \longleftarrow \bigcup_{H \in \mathcal{H}} \mathcal{A}(H)$;
$\mathcal{A}' \longleftarrow \mathcal{A}$;
$n \longleftarrow 0$;
$converged \longleftarrow false$;
while not converged and $n < n_{\max}$ do
    foreach HIT $H \in \mathcal{H}$ do
        foreach question in $H$ do
            Calculate the means and standard deviations of question samples in $\mathcal{A}(H) \cap \mathcal{A}'$;
            Compute z-scores of question samples for all assignments $\mathcal{A}(H)$ using the estimated mean and standard deviation;
        foreach assignment $A \in \mathcal{A}(H)$ do
            $P \leftarrow 1/9$ of the sum of all z-scored samples that are positive;
            $Q \leftarrow 1/9$ of the absolute sum of all z-scored samples that are negative;
            $Z(A) \leftarrow \max(0, rP + sQ - rs) \cdot \max(0, sP + rQ - rs)$;
    Sort all assignments $A \in \mathcal{A}$ according to $Z(A)$ in ascending order;
    $\mathcal{A}'' \longleftarrow$ leading subset of $\mathcal{A}$ of size $p \cdot |\mathcal{A}|$;
    if $\mathcal{A}'' = \mathcal{A}'$ then
        $converged \longleftarrow true$;
    $\mathcal{A}' \longleftarrow \mathcal{A}''$;
    $n \longleftarrow n + 1$;
return $\mathcal{A}'$
```

In an HIT, an assignment can be regarded as an outlier if the answers for the corresponding nine study questions substantially deviate from those of the other assignments of the same HIT. The criterion to accept or reject an assignment is based on the outlier detection given in the ITU-R Recommendation BT.500 [36] for quality assessment methods using the double stimulus presentation and a continuous quality scale. Therein, a set of quality ratings of a study participant are deemed unreliable if two conditions are met: the ratings have large deviations from the mean of the ratings of all participants, and the ratings are inconsistently above and below the corresponding means. The latter condition is intended to ensure that workers with a consistent bias are not classified as outliers. For example, participants with "golden eyes" will detect minute distortions, unlike less critical viewers, and consistently give lower quality ratings, which, however, should not be regarded as outliers.

We propose to make outlier detection robust in the sense that the statistics used to identify outliers do not suffer from the influence of the outliers themselves.

To this end, we apply an iterative procedure structured similarly to the method of [37] for k-means clustering with integrated outlier removal. In previous work, we successfully used such an approach for outlier detection in full-reference image quality assessment using triplet comparisons [38]. Given a target fraction of outlier assignments, for example, 10%, the remaining majority of assignments is used to compute the statistics, namely, the mean and standard deviation for the set of PJND assessments of each study question. These statistics are then used to update the target fractions of outlier assignments and their respective complements, the set of acceptable assignments . This procedure of extracting statistics from the majority and updating the set of outliers is repeated until convergence or a maximal number of iterations is reached.

To compute the statistics in the currently considered set of acceptable assignments (majority), we first calculate the z-score of the assessed PJND samples for each of the nine study images. Then, we apply the same z-score mapping to the remaining assignments in the current set of outliers. For the nine PJNDs collected in any given assignment, we consider their mean absolute error w.r.t. the consensus of the majority, which by construction is equal to 1/9-th of the sum of the absolute values of the z-scores. The mean absolute error is split into a sum $P+Q$, where P corresponds to the contribution of the z-scored samples that are positive, and Q corresponds to those that are negative.

To jointly judge the severity of the overall mean absolute error and the degree by which the signs of the errors differ for the nine judgments in an assignment, we use the product

$$Z(P,Q) = \max(0, rP + sQ - rs) \cdot \max(0, sP + rQ - rs),$$

where $r \leq s$ are two parameters. The target percentage of assignments with the greatest Z-values are marked as outliers, and the remaining assignments are marked as acceptable. See Algorithm 1 for details.

The parameters r, s determine the tradeoff between the two involved criteria. For example, for $r = s = 1$, only the overall mean absolute error $P+Q$ determines the outliers, and for $r = -1, s = 1$, the largest Z-values are obtained on the diagonal $P = Q$ independently of the overall mean absolute error.

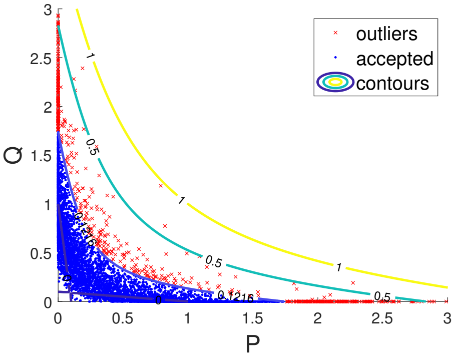


Fig. 8: Example for HIT-level outlier detection. The figure shows the (P,Q)-pairs of 5,459 assignments of 112 HITs. The curves are contour lines of the function $Z(P,Q)$ for the indicated Z-values. The 546 HIT assignments with Z-values larger than 0.1216 are regarded as outliers (red crosses).

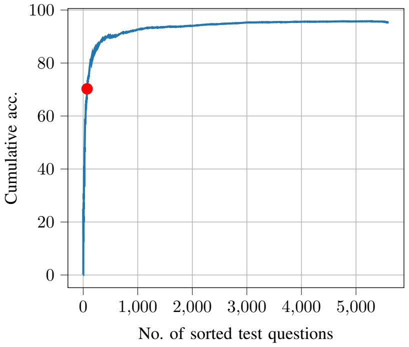


Fig. 9: Setting the time threshold at the question level. The red dot corresponds to the test question used to determine the threshold.

The contour plot in Fig. 8 shows that the Z-value for an assignment with mean absolute error $P + Q$ is positive if and only if the point (P, Q) lies above the two lines with intercepts r and s on the P- and Q-axes, respectively. The contour lines of $Z(P,Q)$ are hyperbolas, and we decrease the Z-level of a contour line until the desired target fraction of assignments has its points (P, Q) above the contour line. In this work, the parameters r and s were empirically set to 0.1 and 1.0, respectively. As a result, 10% of the assignments were removed in all study HITs, which reduced the total number of PJND samples to 44,253.

Question level: It is infeasible to complete a question reliably in a very short time. Therefore, we removed samples for which the slider duration was less than a threshold, determined as follows. Given the results of 5,600 test questions in all study HITs, we sorted them according to slider duration in ascending order. We then calculated the cumulative accuracy w.r.t. the ground truth (Fig. 9). The figure indicates that when workers spent more time on the test questions, they were more likely to provide correct answers. We set the time threshold as the minimum time T such that the accuracy on the set of all test questions whose answers required a slider time of at least T was at least 70%. This corresponds to locating the red dot

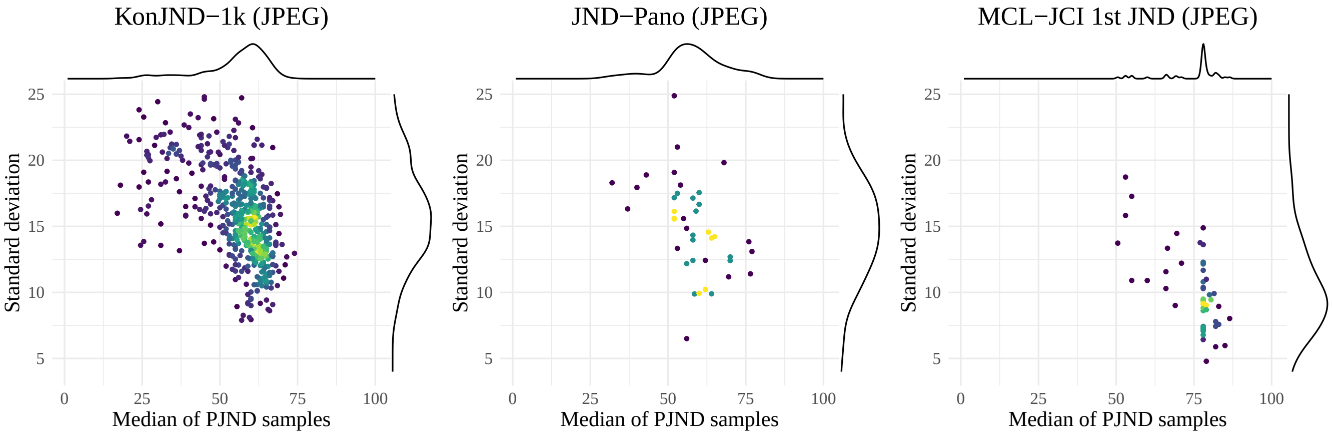


Fig. 10: Scatter plot of the standard deviation vs the median of PJND samples per reference image. The marginal density plots are shown on the sides. The point colors represent the local estimated density (kernel-based density). Dark colors stand for low-density regions, and bright indicate high-density regions.

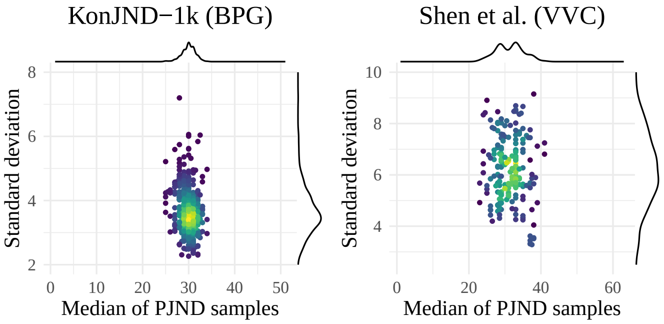


Fig. 11: Scatter plot of the standard deviation vs. the median of PJND samples per reference image. Similar to Fig. 10, the marginal density plots are shown on the sides and point color represents local density.

at the 70% accuracy level in the cumulative accuracy plot in the figure. The result was $T = 2.45\,\text{s}$. This step reduced the number of samples to 42,162.

In addition, we removed extreme values, namely, samples whose chosen distortion levels were less than or equal to 5 and greater than or equal to 95. With this step, the number of samples decreased to 41,974 (20,810 for JPEG and 21,164 for BPG).

D. Comparison with the state-of-the-art JND-based image datasets

We compare KonJND-1k with three existing JND-based image datasets: MCL-JCI, JND-Pano, and Shen *et al.* [16]. We excluded SIAT-JSSI because it has only 12 reference images.

For a first visual overview, we present the basic statistics, which are given by the distributions of the medians of these PJND samples and their standard deviations in Figures 10 and 11 alongside the scatter plots of median PJND versus standard deviation. Each dot in the plots corresponds to a single reference image being subject to a particular codec (JPEG, BPG, or VVC) in each study. For a fair comparison, we must differentiate between codecs that have different numbers of distortion levels (quality factor or quantization parameter). Therefore, we compare KonJND-1k (JPEG) to JND-Pano and MCL-JCI for JPEG, and KonJND-1k (BPG) to Shen et al. (VVC).

The main purpose of the JND-based image datasets is to provide training data for machine learning to predict the JND for any given source image and a particular codec. For this purpose, two aspects stand out as most relevant for the utility of such datasets, namely, the size or diversity of the dataset and the precision of the JND values. The precision of the datasets can be measured with the interrater reliability (IRR), which is the degree of agreement among the independent observers of each stimulus sequence when assessing the JND. Better controlled lab studies generally give better IRR scores.

1) Size and precision: The number of source images for JPEG compression used in KonJND-1k, MCI-JCI, and JND-Pano is 504, 50, and 40, respectively. For KonJND-1k, we ensured diversity by proper sampling of content-related deep features from a ResNet50 model, as described in Sec. IV. For the other codec (BPG), KonJND-1k also provides 504 stimuli, and Shen *et al.* provide 202 sources for VVC. Overall, we find that our dataset KonJND-1k is far superior in terms of size and diversity compared to the others.

To visualize and compare the precision of the median PJNDs in the comparison datasets, we plotted the violin graphs of the 95% confidence interval (CI) of the median of the PJNDs in Fig. 12. In this figure, the gray area indicates the distribution of CI values. The white dot in the center of the graph represents the median, the thick black band above the median is the third quartile area, and the thick black band below the median is the first quartile.

As seen in the left plot of Fig. 12, the KonJND-1k (JPEG) dataset has a lower CI distribution than the other two datasets with JPEG compression. This could be because we used the flicker test method to assess the PJND for the KonJND-1k dataset. The flicker test results in a higher intersubject agreement and thus yields a lower CI. The right part of the figure indicates that KonJND-1k (BPG) has a much lower CI of the median values than the dataset of Shen *et al.*

TABLE III: Ranking of the distribution models according to the negative log-likelihood of MLE and the A-D test for the 1,008 source images of the KonJND-1k dataset. The models are from MATLAB (R2019b) and described in [39], [40]: Half-normal (1), Rayleigh (2), Exponential (3), Generalized Extreme Value (4), Generalized Pareto (5), Stable (6), tLocation Scale (7), Birnbaum-Saunders (8), Extreme Value (9), Gamma (10), Logistic (11), Loglogistic (12), LogNormal (13), Nakagami (14), Normal (15), Poisson (16), Rician (17), and Weibull (18). The results for the two other models available in MATLAB, the Beta distribution and Burr distribution, are not included because fitting the PJND samples with these distributions was not possible.

KonJND-1k	Models	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
JPEG	log-likelihood	16	15	17	1	4	3	7	14	5	11	10	12	13	9	8	18	6	2
	A-D reject	461	370	494	1	380	1	11	89	8	40	1	4	64	23	13	409	16	9
	A-D rank	17	14	18	1	15	1	7	13	5	11	1	4	12	10	8	16	9	6
BPG	log-likelihood	17	16	18	1	15	2	3	13	12	9	6	10	11	7	4	14	5	8
	A-D reject	504	504	504	2	504	0	1	25	15	18	0	0	23	12	7	227	7	6
	A-D rank	15	15	15	5	15	1	4	13	10	11	1	1	12	9	7	14	7	6
Overall	log-likelihood	17	15	18	1	14	2	4	13	7	10	9	11	12	8	6	16	5	3
	A-D reject	965	874	998	3	884	1	12	114	23	58	1	4	87	35	20	636	23	15
	A-D rank	17	15	18	3	16	1	5	13	8	11	1	4	12	10	7	14	8	6

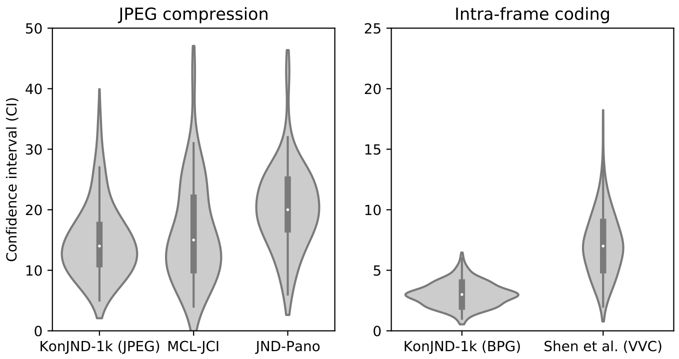


Fig. 12: Violin plots showing the distribution shapes for the 95% confidence interval (CI) of the median PJNDs for five JND-based image datasets, with quartiles indicated.

Dataset	ICC	95% CI	N	k
KonJND-1k (JPEG)	0.200	0.179-0.223	504	41.29
JND-Pano (JPEG)	0.205	0.134-0.316	40	20.25
MCL-JCI, 1st JND (JPEG)	0.342	0.259-0.453	50	30.00

TABLE IV: Intraclass correlation coefficients (ICCs) for several JND-based image datasets for JPEG compression, with confidence intervals (CI). N is the total number of source images used in the analysis, k is the number of measurements per source image calculated using the method of Lessells and Boag [45]. Only the first JND is considered for MCL-JCI.

Therefore, the KonJND-1k datasets have higher precision and lower uncertainty in the median PJNDs than the comparison datasets.

2) Interrater reliability: Evaluating the JND level is a difficult subjective task. It depends on factors including the participant's interpretation, attention, the hardware used, and the environment in which the experiment was conducted. To compare our approach to existing annotated datasets, many of which were obtained in laboratory conditions, we take a closer look at the interrater reliability. We accomplish this via the intraclass correlation coefficient (ICC), which is one of the most prevalent IRR indicators. Shrout and Fleiss [41] outlined six ICC estimators, given different assumptions. Here, we use the ICC(1,1) one-way random effects model [42], which is applicable to crowdsourcing data that has incomplete observations. ICC(1,1) measures the absolute agreement among raters. A high ICC value means that the largest part of the variance between the JND values is explained by differences between individual images and not by differences between participant opinions.

To better interpret the ICC values, putting the values into context is a good approach. Generally, better controlled studies result in higher ICC values for rating experiments. However, a lower ICC can also mean that the task is more difficult, and thus, participants may agree less. The range of the ICC values varies greatly for different tasks. For instance, for absolute category rating (ACR) tasks, such as aesthetics or technical quality assessment, ICCs between 0.3 and 0.7 have been reported in crowdsourcing experiments; the range is higher for lab experiments, between 0.7 and 0.94 [43], [44], [7].

Table IV shows that the ICC values are generally lower for JND-based image datasets. The largest ICC, 0.342, was obtained by the MCL-JCI lab study. This confirms the expected difficulty of the task. Nonetheless, for our crowdsourcing experiments, the ICC value for the JPEG images (0.203) is approximately the same as that for another JND database, namely, JND-Pano (0.205), which was based on a lab study. The tasks that involved assessing the JNDs of BPG and VVC encoded images yield lower ICC values of 0.102 and 0.175, respectively. This is due to the very narrow interval in which the mean of the JND samples per image was distributed relative to the standard deviation within each source image set, as shown by comparing Figures 10 and 11.

In summary, we showed that our crowdsourced dataset has a level of agreement similar to JND-Pano. Compared to MCL-JCI, the reliability metric values are lower, which suggests that our crowdsourced PJND assessment task was more challenging.

VIII. Modeling the SUR function

As in [3], we used maximum likelihood estimation (MLE) and the Anderson-Darling (A-D) test to find the most suitable distribution for the KonJND-1k PJND samples. We used MLE to estimate the parameters of the probabilistic models and ranked them according to increasing negative log-likelihood averaged over the source images in the dataset. The A-D test was applied to the null hypothesis that a set of PJND samples were drawn from a given distribution model at a specified significance level (5% in our work).

As candidate distributions, we used 18 distribution models from MATLAB (R2019b). The results are shown in Table III. For JPEG compression, the generalized extreme value (GEV) distribution ranked first in terms of both the negative log-likelihood and A-D test. For BPG compression, the GEV distribution ranked first in terms of the negative log-likelihood and third in terms of the A-D test.

IX. Objective PJND method validation

Our aim is not only to explore the feasibility of conducting a crowdsourced JND study but also to provide a benchmark PJND dataset to the research community. Such a dataset can support the development and validation of PJND models.

To demonstrate the utility of our dataset, we used it to predict the SUR with a state-of-the-art method (SUR-FeatNet [3]). This method extracts deep features from a pretrained deep model and trains a shallow network to map these features to the ground truth SUR values. The parameters of the GEV distribution are estimated by fitting the predicted SUR values.

k-fold cross-validation ($k = 10$) was used to evaluate the performance. More specifically, the source images were split into 10 nonoverlapping subsets according to compression type. Each subset contained approximately 50 source images and their corresponding distorted images. In each trial, we used one subset as a test set and the remaining subsets as training and validation sets. For each compression type, the overall result was the average of 10 test results.

We used three metrics [3] to evaluate the performance of SUR-FeatNet on the KonJND-1k dataset. These are the mean absolute error (MAE) of the 50% PJNDs, the MAE of the peak signal-to-noise ratio (PSNR) at the 50% PJNDs, and the Bhattacharyya distance [46] between the predicted and ground truth PJND distributions of type GEV.

Table V reports the overall performance of the fine-tuned SUR-FeatNet and a baseline method. As in [3], the baseline method predicts the 50% PJNDs for the test set by the distortion levels corresponding to the average PSNRs at the corresponding 50% PJNDs in the training set. The results show that SUR-FeatNet provided accurate SUR predictions for both JPEG and BPG. Moreover, SUR-FeatNet significantly outperformed the baseline method.

X. Conclusion and future work

We designed a robust framework for conducting crowdsourced subjective PJND assessments. Our framework includes a flicker test together with a slider-based adjustment method to speed up the experiment. It also exploits qualification HITs and test questions to ensure reliability. Our framework can be easily applied to other quality-based crowdsourcing tasks with little modification.

	SUR-FeatNet (Fine-tuning)			Baseline	
Compression	Bhattacharyya	ΔPJND	ΔPSNR (dB)	ΔPJND	ΔPSNR (dB)
JPEG	0.0640	6.95	0.50	30.79	3.13
BPG	0.0588	1.46	0.76	7.89	2.53

TABLE V: Comparison between fine-tuned SUR-FeatNet and the baseline method for KonJND-1k dataset. ΔPJND is the MAE of the 50% PJNDs. ΔPSNR is the MAE of the PSNR at the 50% PJNDs.

Using our framework, we conducted a large-scale subjective PJND study, yielding the largest image PJND dataset, KonJND-1k. The dataset contains 1,008 source images, with an average of 42 samples per image. It provides large content diversity and allows the research community to develop more accurate and more general objective PJND methods than the current state of the art.

In addition to the development of objective PJND models, future investigations of subjective PJND assessment should uncover the flickering locations in an image. Understanding the locations in an image where the just noticeable distortions first appear when the bit rate is reduced will provide another measure, in addition to the PJND, for improving perception-based image compression techniques.

In our study, the resolution of the images was relatively small (640×480). Building a large-scale crowdsourced PJND dataset for high-resolution images is challenging because the number of crowd workers who have access to a full high-definition (FHD) screen is small. To address this issue, one could display only parts (crops) of the high-resolution image at a time or use an interface that allows panning.

References

[1] ISO/IEC 29170-2, "Information technology – Advanced image coding and evaluation – Part 2: Evaluation procedure for visually lossless coding," 2015.

[2] U. Gadiraju, S. Möller, M. Nöllenburg, D. Saupe, S. Egger-Lampl, D. Archambault, and B. Fisher, "Crowdsourcing versus the laboratory: towards human-centered experiments using the crowd," in *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments, Lecture notes in Computer Science 10264*. Springer Cham, 2017.

[3] H. Lin, V. Hosu, C. Fan, Y. Zhang, Y. Mu, R. Hamzaoui, and D. Saupe, "SUR-FeatNet: Predicting the satisfied user ratio curve for image compression with deep feature learning," *Quality and User Experience*, vol. 5, no. 1, pp. 1–23, 2020.

[4] H. Liu, Y. Zhang, H. Zhang, C. Fan, S. Kwong, C. . J. Kuo, and X. Fan, "Deep learning-based picture-wise just noticeable distortion prediction model for image compression," *IEEE Transactions on Image Processing*, vol. 29, pp. 641–656, 2020.

[5] Y. Zhang, H. Liu, Y. Yang, X. Fan, S. Kwong, and C. J. Kuo, "Deep learning based just noticeable difference and perceptual quality prediction models for compressed video," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[6] 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 *IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[7] V. Hosu, H. Lin, and D. Saupe, "Expertise screening in crowdsourcing image quality," in *IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, 2018, pp. 1–6.

[8] R. R. R. Rao, S. Göring, and A. Raake, "Towards high resolution video quality assessment in the crowd," in *IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, 2021, pp. 1–6.

[9] Y. Jia, W. Lin, and A. A. Kassim, "Estimating just-noticeable distortion for video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 7, pp. 820–829, 2006.

[10] Z. Chen and W. Wu, "Asymmetric foveated just-noticeable-difference model for images with visual field inhomogeneities," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 4064–4074, 2019.

[11] 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," *Electronic Imaging*, vol. 2016, no. 13, pp. 1–9, 2016.

[12] X. Shen, Z. Ni, W. Yang, X. Zhang, S. Wang, and S. Kwong, "Just noticeable distortion profile inference: A patch-level structural visibility learning approach," *IEEE Transactions on Image Processing*, vol. 30, pp. 26–38, 2020.

[13] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612–627, 2018.

[14] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[15] R. Furuta, I. Tsubaki, and T. Yamasaki, "Fast volume seam carving with multipass dynamic programming," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1087–1101, 2016.

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

[17] 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 *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 1509–1513.

[18] 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 *IEEE Data Compression Conference (DCC)*, 2017, pp. 42–51.

[19] H. Wang, I. Katsavounidis, J. Zhou, J. Park, S. Lei, X. Zhou, M.-O. Pun, X. Jin, R. Wang, X. Wang *et al.*, "VideoSet: A large-scale compressed video quality dataset based on JND measurement," *Journal of Visual Communication and Image Representation*, vol. 46, pp. 292–302, 2017.

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

[21] C. Fan, Y. Zhang, H. Zhang, R. Hamzaoui, and Q. Jiang, "Picture-level just noticeable difference for symmetrically and asymmetrically compressed stereoscopic images: Subjective quality assessment study and datasets," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 140–151, 2019.

[22] 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," in *Applications of digital image processing XXXVIII*, vol. 9599. International Society for Optics and Photonics, 2015, p. 95990Z.

[23] F. Wang, X. Li, J. Wang, Y. Tu, X. Liu, Y. Gao, Y. Tian, Y. Wang, W. Liu, and Y. Dong, "Study on influential factors of image quality for laser projection television," in *SID Symposium Digest of Technical Papers*, vol. 51. Wiley Online Library, 2020, pp. 82–84.

[24] D. M. Hoffman and D. Stolitzka, "A new standard method of subjective assessment of barely visible image artifacts and a new public database," *Journal of the Society for Information Display*, vol. 22, no. 12, pp. 631–643, 2014.

[25] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[26] J. Redi, H. Liu, H. Alers, R. Zunino, and I. Heynderickx, "Comparing subjective image quality measurement methods for the creation of public databases," in *Image Quality and System Performance VII*, vol. 7529. International Society for Optics and Photonics, 2010, p. 752903.

[27] B. W. Keelan and H. Urabe, "ISO 20462: A psychophysical image quality measurement standard," in *Image Quality and System Performance*, vol. 5294. International Society for Optics and Photonics, 2003, pp. 181–189.

[28] U.-D. Reips and F. Funke, "Interval-level measurement with visual analogue scales in internet-based research: VAS generator," *Behavior Research Methods*, vol. 40, no. 3, pp. 699–704, 2008.

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

[30] K. Schwarz, P. Wieschollek, and H. P. Lensch, "Will people like your image? learning the aesthetic space," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 2048–2057.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[33] K. Rohrschneider, "Determination of the location of the fovea on the fundus," *Investigative Ophthalmology & Visual Science*, vol. 45, no. 9, pp. 3257–3258, 2004.

[34] B. V. Ehinger, K. Häusser, J. P. Ossandon, and P. König, "Humans treat unreliable filled-in percepts as more real than veridical ones," *Elife*, vol. 6, p. e21761, 2017.

[35] ISO 9241-303, "Ergonomics of human-system interaction – part 303: Requirements for electronic visual displays," 2008.

[36] ITU-R Recommendation BT.500, "Methodologies for the subjective assessment of the quality of television images," 2019. [Online]. Available: https://www.itu.int/rec/R-REC-BT.500

[37] S. Chawla and A. Gionis, "k-means–: A unified approach to clustering and outlier detection," in *SIAM International Conference on Data Mining (SDM)*, 2013, pp. 189–197.

[38] H. Men, H. Lin, M. Jenadeleh, and D. Saupe, "Subjective image quality assessment with boosted triplet comparisons," *IEEE Access*, 2021.

[39] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions. Vol.1*. Hoboken, NJ: Wiley-Interscience, 1993.

[40] ——, *Continuous Univariate Distributions. Vol.2*. Hoboken, NJ: Wiley-Interscience, 1994.

[41] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, no. 2, p. 420, 1979.

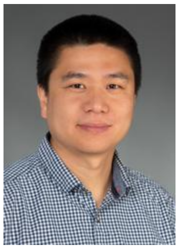
[42] K. A. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," *Tutorials in Quantitative Methods for Psychology*, vol. 8, no. 1, p. 23, 2012.

[43] E. Siahaan, A. Hanjalic, and J. Redi, "A reliable methodology to collect ground truth data of image aesthetic appeal," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1338–1350, 2016.

[44] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Häkkinen, "CID2013: A database for evaluating no-reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 390–402, 2014.

[45] C. Lessells and P. T. Boag, "Unrepeatable repeatabilities: A common mistake," *The Auk*, vol. 104, no. 1, pp. 116–121, 1987.

[46] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.



Hanhe Lin received his Ph.D. at the Department of Information Science, University of Otago, New Zealand in 2016. From 2016 to 2021, he was a postdoc at the Department of Computer and Information Science at the University of Konstanz, Germany, where he was working on project A05 (visual quality assessment) of SFB-TRR 161, funded by the German Research Foundation (DFG). Currently, he is a research fellow at the National Subsea Centre at Robert Gordon University, UK. His research interests include image processing, computer vision, machine learning, deep learning, and visual quality assessment.

Guangan Chen received his M.Sc. in Computer and Information Science at the University of Konstanz, Germany in March 2022. From July 2019 to December 2021, he worked as a student research assistant focusing on visual quality assessment at the University of Konstanz. From May 2022, he will pursue his Ph.D. under the project “multi-camera algorithms and 3D reconstruction for construction site monitoring” in Ghent University, Belgium. His research interests include image processing, 3D reconstruction, deep learning, and visual quality assessment.

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 at the University of Konstanz. He is currently working on his research project titled “JND-based perceptual video quality analysis and modeling” funded by the German Research Foundation (DFG) and collaborating in project A05 (Visual Quality Assessment) of the SFB-TRR 161 funded by the DFG. His research interests include image processing, visual perception, machine learning, deep learning, and crowdsourcing.

Vlad Hosu is a postdoc since 2016 at the Department of Computer and Information Science at the University of Konstanz, Germany. Previously he was a Research Fellow at NUS, Singapore, having received his Ph.D. at the same institution in 2014. His research interests include visual quality assessment, machine learning, image enhancement, crowdsourcing strategies, understanding, and modeling human visual perception.

Ulf-Dietrich Reips is the Full Professor for Research Methods, Assessment and iScience (https://iscience.uni-konstanz.de) at the Department of Psychology, University of Konstanz. He received his PhD in 1996 from the University of Tübingen. His research focuses on internet-based research methodologies and also concerns the psychology of the internet, measurement, assessment, cognition, personality, privacy, social media, and human data science. In 1994, he founded the first laboratory for conducting real experiments on the world wide web. Ulf was a founder of the German Society for Online Research and was elected the first non-North American president of the Society for Computers in Psychology. His over 170 scientific publications include six books. Ulf and his team develop and provide free web tools (available from the iScience Server: http://iscience.eu/) for researchers, teachers, students, and the public. They have received numerous awards for their web applications and methodological work serving the research community.

Raouf Hamzaoui (Senior Member, IEEE) received the M.Sc. degree in mathematics from the University of Montreal, Montreal, QC, Canada, in 1993, the Dr.rer.nat. degree from the University of Freiburg, Freiburg im Breisgau, Germany, in 1997, and the Habilitation degree in computer science from the University of Konstanz, Konstanz, Germany, in 2004. He was an Assistant Professor with the Department of Computer Science, University of Leipzig, Leipzig, Germany, and the Department of Computer and Information Science, University of Konstanz. He joined De Montfort University, Leicester, U.K., in 2006, where he is currently a Professor in media technology. His research interests include image and video coding, multimedia communication systems, error control systems, and machine learning. He served as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2010 to 2016 and IEEE TRANSACTIONS ON MULTIMEDIA from 2017 to 2021.

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 at the University of California, Santa Cruz, USA, and then at the University of Bremen, resulting in his habilitation. From 1993 to 1998, he was a Professor of 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 visualisation, dynamical systems, and sport informatics.