

An Image Quality Assessment Dataset for Portraits

Nicolas Chahine^{1,2} Ana-Stefania Calarasanu¹ Davide Garcia-Civiero¹
Théo Cayla¹ Sira Ferradans¹ Jean Ponce^{2,3}

¹DXOMARK ²Département d'informatique de l'Ecole normale supérieure (ENS-PSL, CNRS, Inria)

³Institute of Mathematical Sciences and Center for Data Science, New York University

Abstract

Year after year, the demand for ever-better smartphone photos continues to grow, in particular in the domain of portrait photography. Manufacturers thus use perceptual quality criteria throughout the development of smartphone cameras. This costly procedure can be partially replaced by automated learning-based methods for image quality assessment (IQA). Due to its subjective nature, it is necessary to estimate and guarantee the consistency of the IQA process, a characteristic lacking in the mean opinion scores (MOS) widely used for crowdsourcing IQA. In addition, existing blind IQA (BIQA) datasets pay little attention to the difficulty of cross-content assessment, which may degrade the quality of annotations. This paper introduces PIQ23, a portrait-specific IQA dataset of 5116 images of 50 predefined scenarios acquired by 100 smartphones, covering a high variety of brands, models, and use cases. The dataset includes individuals of various genders and ethnicities who have given explicit and informed consent for their photographs to be used in public research. It is annotated by pairwise comparisons (PWC) collected from over 30 image quality experts for three image attributes: face detail preservation, face target exposure, and overall image quality. An in-depth statistical analysis of these annotations allows us to evaluate their consistency over PIQ23. Finally, we show through an extensive comparison with existing baselines that semantic information (image context) can be used to improve IQA predictions. The dataset along with the proposed statistical analysis and BIQA algorithms are available: <https://github.com/DXOMARK-Research/PIQ2023>

1. Introduction

Social media has made smartphones a vital tool for connecting with people worldwide. Visual media, particularly portrait photography, has become a crucial aspect of sharing content on these platforms. Portrait photography serves numerous applications (e.g., advertisements, social media)

and use cases (e.g., anniversaries, weddings). Capturing a high-quality portrait is a complex exercise that demands careful consideration of multiple factors, such as scene semantics, compositional rules, image quality, and other subjective properties [46].

Smartphone manufacturers strive to deliver the best visual quality while minimizing production costs to rival professional photography. Achieving this requires implementing complex tuning and optimization protocols to calibrate image quality in smartphone cameras. These cameras introduce sophisticated non-linear processing techniques such as multi-image fusion or deep learning-based image enhancement [55], resulting in a combination of authentic (realistic) camera distortions. This makes traditional objective quality assessment [4, 16, 32, 40] that models digital cameras as linear systems unreliable [9]. Therefore, in addition to objective measurements, the tuning process also includes perceptual evaluations where cameras are assessed by image quality experts. This procedure requires shooting and evaluating thousands of use cases, which can be costly, time-consuming, and challenging to reproduce. Automatic image quality assessment (IQA) methods that try to mimic human perception of quality have been around for many years, in order to help in the tuning process [14, 36, 37, 39, 48, 57, 60, 64]. Blind IQA (BIQA), in particular, is a branch of IQA where image quality is evaluated without the need for undistorted reference images. Learning-based BIQA methods [15, 24, 25, 27, 52, 59, 62, 67, 69] have shown good performance on authentic camera distortion datasets [9, 13, 21, 56, 61, 70], annotated by subjective assessment of image quality. Annotating these datasets is considered an ill-posed problem, as the subjective opinions are not deterministic, making it challenging to use BIQA methods as accurate quality measures. Therefore, there is a need to develop a quantitative and formal framework to evaluate and compare subjective judgments in an objective manner. In this paper, we rely on pairwise comparisons performed by image quality experts along a fixed and relevant set of attributes.

Multiple attributes, including target exposure, dynamic

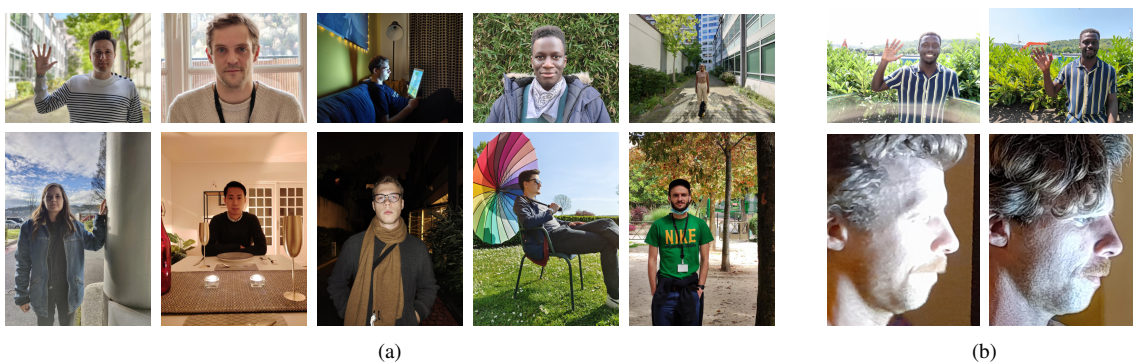


Figure 1. (a) Scenes from the PIQ23 dataset. (b) Examples of the region of interest (ROI) used for different attribute comparisons. *Top*: overall quality; we use a resized version of the full image. *Bottom*: details & target exposure; we use an upscaled face area.

range, color, sharpness, noise, and artifacts, define image quality [3]. Portrait images require additional considerations, such as skin tone, bokeh effect, face detail rendering, and target exposure on the face, which fall under the scope of portrait quality assessment (PQA) [40].

To the best of our knowledge, the problem of assessing the quality of a portrait image has received limited attention. Most of the work on face IQA [49] has been directed towards improving face recognition systems and not as an independent topic. As far as we know, our paper introduces the first-of-its-kind, smartphone portrait quality dataset. We hope to create a new domain of application for IQA and to push forward smartphone portrait photography. Our contributions are the following:

- A new dataset, PIQ23, consisting of 5116 single portrait images, taken using 100 smartphone devices from 14 brands, and distributed across 50 different natural scenes (*scene = fixed visual content*). We have addressed the ethical challenges involved in creating such a dataset, by obtaining from each individual depicted in the dataset a signed and informed agreement, making it the only IQA dataset with such legal and ethical characteristics, as far as we know.
- A large IQA experiment controlled in a laboratory environment with fixed viewing conditions. Using pairwise comparisons (PWC) and following carefully designed guidelines, we gather opinions for each scene, from over 30 image quality experts (professional photographers and image quality experts) on three attributes related to portrait quality: face detail preservation, face target exposure, and overall portrait image quality.
- An in-depth statistical analysis method that allows us to evaluate the precision and consistency of the labels as well as the difficulty of the IQA task. This is particularly important given the fact that image quality la-

bels are heavily affected by subjectivity, disagreement between observers, and the number of annotations.

- An extensive comparison between multiple BIQA models and a simple new method combining scene semantic information with quality features to strengthen image quality prediction on PIQ23.

2. Related work

2.1. BIQA

The main goal of blind IQA (BIQA) is to predict image quality without requiring a pristine reference image. We review the datasets that already exist in this domain as well as the BIQA computational algorithms.

BIQA datasets. Early datasets like LIVE [50], CSIQ [29], TID [44, 45] and BAPPS [66] consist of noise-free images processed with several artificial distortions. These distortions aim to describe image compression or transmission scenarios and most of them fail to capture the complexity of modern smartphone camera systems, with non-linear processing pipelines. Recent “in-the-wild” datasets such as CLIVE [13], KonIQ10k [21] and PaQ-2-PiQ [61] consist of media-gathered images with more complex mixtures of distortions closer to real-world images. However, due to their wild nature and uncontrolled labeling environment, they do not form a strong background to evaluate the quality of digital cameras, which we are most interested in. As an early effort on this topic, Virtanen *et al.* [56] have developed a database (CID2013) that spans 8 visual scenes with 79 digital cameras. In recent work, Zhu *et al.* [70] provide a smartphone IQA dataset (SCPQD2020) of 1800 images shot with 15 devices on 120 visual scenes. They annotate the database in a well-controlled laboratory, by three image quality experts. Fang *et al.* published SPAQ [9], a smartphone IQA dataset with 11125 images shot with 66

devices. Both datasets provide multiple attribute evaluations and scene categories. They include generic visual content and do not deal with PQA. While SCPQD2020 lacks in the number of observers, SPAQ relies on resized images which heavily degrades the quality. All previously mentioned datasets, except TID2013 [44] and BAPPS [66], rely on rating systems (MOS), and do not pay close attention to the difficulty of cross-content observations. In PIQ23, we provide 50 scenes, each annotated independently. We collect opinions from over 30 image quality experts by pairwise comparisons, which has been shown to be more consistent in IQA experiments [34, 43]. We also analyze the uncertainty and consistency of our annotations through a new statistical analysis method.

BIQA methods. BIQA can be separated into classical and deep learning approaches. Early learning-based approaches [14, 36, 39, 48, 60] use a combination of hand-crafted statistical features (natural scene statistics) to train shallow regressors (e.g. SVR). Other approaches try to estimate the quality without the need for training [37, 57, 64]. These methods perform relatively poorly on modern IQA datasets, as they do not fully reflect the human perception of realistic distortions [13, 67]. Consequently, deep BIQA models have been surging in the last decade. Multiple convolutional neural networks (CNN) based methods [24, 27, 67] have demonstrated solid performance on modern datasets. Zhang *et al.* [69] address the problem of uncertainty in IQA and present a method to simultaneously train on multiple datasets using image pairs as training samples. Su *et al.* [52] try to separate semantic features from image quality features by training an adaptive hyper network that captures semantic information. Recent works that adopt transformer architectures [15, 25, 59, 62] to extract global quality information, have shown impressive performances on IQA datasets. Because of the per-scene annotation structure of our dataset, we adopt a semantics-aware multitasking method to adapt the scale and features to the input scene.

2.2. PQA

Despite the lack of portrait quality assessment (PQA) research, solely focusing on evaluating portrait quality, we still recognize the importance of face IQA (FIQA). FIQA aims to assess the quality of face images to boost the performance of face recognition algorithms [1, 17, 20, 31, 41, 47, 49, 58]. The closest FIQA work to PIQ23 is that of Zhang *et al.* [65], where they have developed a dataset to objectively evaluate the illumination quality of a face image. Redi *et al.* [46] define a set of attributes to evaluate the “beauty” of the portrait. Kanafusa *et al.* [23] propose a method to define a standard portrait image, which can be later used to evaluate color rendering and other attributes between cross-media. In this work, we can see a first attempt to use a stan-

dard portrait as a subject for IQA. Chahine *et al.* [40] proposed the first approach to evaluate specific face attributes as a separate metric for PQA on realistic mannequins. Finally, Liang *et al.* [30] have developed a large-scale portrait photo retouching dataset, with multiple use cases and cameras. To the best of our knowledge, PIQ23 is the first smartphone PQA dataset, with a variety of visual scenes, legal validation, and expert annotations.

2.3. Domain shift

The annotation strategies and image content can vary widely between different IQA datasets. Hence, their respective quality scales are usually relative and independent. With this characteristic, we encounter a problem known as domain shift [53, 63, 68, 69]. Since quality scales are relative, similar scores may not indicate the same level of perceptual quality across different datasets. This can lead to confusion when attempting to learn from multiple sources. As a result, understanding image semantics is necessary. Current BIQA models implicitly try to learn semantics and quality simultaneously. However, it is extremely difficult to merge these two problems, as they seem to be contradictory [9, 26]. Some papers try to solve this problem using multi-task learning [9, 22, 53, 63]. Explicitly separating semantic information from quality is not well represented in previous works. Su *et al.* [52] propose HyperIQA, a self-adaptive hyper network that implicitly extracts semantic information and adapts the quality prediction accordingly. The hyper network, however, is not trained to predict categories explicitly. Since PIQ23 consists of multiple relative content-dependent scales, we propose to combine multitasking with HyperIQA in order to adapt the quality scale of each scene based on semantic understanding.

3. PIQ23

3.1. Dataset details

Legal aspects. We believe that unrestricted access to PIQ23 for public research is crucial. Accordingly, we have taken steps to address any potential legal obstacles that may obstruct this access. All individuals in the photos have given explicit permission for image rights via signed transfer and received a privacy notice detailing how their images will be processed. Also, to ensure the effectiveness of people’s rights, we have tagged each photo with a unique identifier assigned to each person by using a face clustering algorithm. This pseudonymization technique prevents access to individuals’ names by dataset users. Finally, we contractually require all dataset users to comply with relevant data protection laws, including the GDPR.

Dataset properties. We have constructed PIQ23 with the intent of reducing annotation biases and covering a variety

of common real-life scenarios. To achieve this goal, we have broken down the factors affecting the quality of a portrait image. We consider lighting to be a primary element influencing the quality. Hence, we have separated lighting conditions into four groups: outdoor, indoor, low light, and night. Also, we have paid attention to lighting homogeneity, which describes the reflection of the light on the subject (*e.g.* front light, side light, backlight). The characteristics of the subject, such as age, skin tone, gender, subject position, framing, face orientation, movement, and subject-to-lens distance play an equally important role. Our skin tone ruler is based on the Fitzpatrick skin type (FST) [12]. We have tried to cover a sizeable chunk of smartphone devices and brands utilized over the past decade. Additionally, we have included diverse smartphone camera lens focal lengths such as zoom, wide, and selfie along with distinct camera modes such as night and bokeh. Furthermore, we have contemplated the possibility of augmenting our dataset with high-quality images sourced from DSLR cameras. Nonetheless, we have found through our experimentation that artificially distorting DSLR images to ensure comparability with photos taken by smartphones is a challenging task. As a result, we have excluded DSLR cameras from our dataset.

To comply with the previous description, we have designed a collection of 50 distinct portrait scenarios (referred to as “scenes”), captured in predetermined locations that encompass a diverse range of factors (see Fig. 1 (a)). The dataset images were taken with about 100 smartphones (2014- 2022) from 14 brands and different price segments. We have collected around 5116 images, averaging 100 images per scene. We note that PIQ23 was subsampled from a larger dataset that was collected over a long period of time (a couple of years) and is a result of cumulative efforts in engineering and photography. We, therefore, believe in its capacity to cover a broad spectrum of smartphone photography. More information about the PIQ23 characteristics can be found in the supplementary material.

3.2. Portrait quality assessment

Portrait quality attributes. In a portrait, most attention is given to the person depicted, which is known as the human region priority (HRP) [30]. Portrait quality may vary significantly depending on the application. For example, Redi *et al.* [46] try to define all the characteristics to capture the ‘beauty’ of a portrait. Quality in this case is strongly correlated with beauty and aesthetics. In other applications, such as FIQA [49], quality assessment is a measure of utility, to filter out poor quality faces from face recognition systems. Neither application totally aligns with PQA [40]. Thus, we intend to broaden the research on PQA by studying a preliminary group of three attributes: face detail preservation, face target exposure, and overall image quality. Additionally, we have conducted a study concern-

ing a fourth attribute, namely global color quality. However, due to the difficulty in annotating this attribute through pairwise comparisons on different content, we have decided to exclude it from our dataset (Fig. 3). The annotation guidelines can be found in the supplementary material.

Annotation strategy. Perception-based IQA experiments present a high degree of difficulty and are usually subjective. Opinions can vary widely depending on multiple factors: viewing conditions, the observer’s cultural and professional backgrounds, image content, etc. The objective of PIQ23 is to deliver image quality annotations obtained (as much as possible) from impartial and unbiased observations. To maximize objectivity and consistency, we propose two elementary steps:

- First, we have chosen to annotate each scene separately using a forced-choice pairwise comparison approach (PWC). Combined with the active sampling technique proposed in [35], we have been able to reach good annotation consistency with a minimal number of comparisons (see Sec. 4).
- Second, we have fixed the region of interest (ROI) for each attribute, as shown in Fig. 1 (b). For details preservation and target exposure, we have extracted the face area using RetinaFace [7]. We have then up-scaled it using standard bicubic upsampling to a reference resolution of about 4.5 megapixels with a fixed aspect ratio. For color and overall attributes, we have resized the images to an approximate Full HD resolution (about 2.5 megapixels) while keeping the original aspect ratio (*i.e.* portrait or landscape).

Experiment details. We have reached out to professional photographers and experts with a solid background in photography and image quality to help us annotate the dataset. The opinions of more than 30 experts were gathered using an internal PWC tool. Observers were asked to select the best out of two images, following the guidelines described in the supplementary material. We have adapted our settings so that the viewing conditions are aligned with that of a human eye, with a cutoff frequency $\nu_{cut} = 30\text{cpd}$. Hence, we have used a BenQ 32” 4k monitor with a pixel pitch of 0.185, and we have fixed the eye-to-screen distance at 65cm. We have calibrated the display to standard sRGB settings (D65 white point with luminance $\geq 75\text{cd}/\text{m}^2$ with no direct illumination of the screen and a background illumination with a lighting panel set to D65 / 15% for reducing eye stress). We have also converted all images in DCI-P3 color space to sRGB. We have kept the sessions short, around five minutes per attribute, in order to reduce fatigue and stress on the observers. The annotation procedure took around eight

months. For each scene and each attribute, we have collected around 4k pairwise comparisons, a total of 600k data points. Though, for a limited number of comparisons, given the subjectivity of the task, noise and outliers are commonly encountered. In Sec. 4, we present a new statistical analysis method to rectify this noise.

4. Statistical analysis

We present a new approach to quantifying uncertainty in IQA experiments. We recall in Sec. 4.1 how quality scores are extracted from a PWC experiment and how uncertainty can be estimated using bootstrapped confidence intervals. We then introduce in Sec. 4.2 a new statistical analysis strategy to go beyond the calculation of confidence intervals. The complete pipeline is illustrated in Fig. 2

4.1. Psychometric scaling and confidence intervals

Psychometric scaling. Designing a PWC experiment requires modeling the statistical distribution of the image quality. Commonly, the quality of an image is described by the Thurstone Case V observer model [5, 54] as a Gaussian distribution $\mathcal{N}(\mu, \sigma)$. The average μ represents the actual quality and σ^2 is its “perceptual” variance across observations. The latter encompasses the intra-variance and inter-variance of the perceptual quality. The intra-variance represents the uncertainty of one observer when the observation is repeated multiple times. The inter-variance represents the uncertainty across multiple observers. Based on this formulation, psychometric scaling methods [42, 43] transform the comparison matrix M constructed from a PWC experiment, into a continuous scale of image scores representing the average opinions across multiple observers. The results are typically expressed in Just-Objectable-Difference (JOD) units [43]. Two images are 1 JOD apart if 75% of observers choose one as better than the other. In our work, we have adopted the psychometric scaling method proposed by Mikhailiu *et al.* [35]. The authors propose an efficient active pair selection technique via approximate message passing and information gain maximization, combined with the TrueSkill scaling algorithm [19] to minimize the PWC experiment cost. Thus, M is typically a very sparse matrix (so-called incomplete design) with a limited number c of non-zero elements.

IQA limitations. The choice of image and observer samples plays a critical role in the accuracy of the JOD scores. From a statistical point of view, these samples are taken from infinitely large populations of images and observers respectively. When sampling images of similar quality, for example, the comparison becomes harder, requiring a correspondingly larger number of comparisons than a sample of images with distinguishable quality differences. Similarly,

a sample of inexperienced observers generally leads to noisier annotations, requiring more observers than a sample of experts. In addition, psychometric scaling algorithms introduce an estimation error that is inversely proportional to the size of the data. In conclusion, estimating the difficulty of the comparison task (linked to image sampling), the quality of the experiment (linked to observer sampling), and the precision of the psychometric scaling algorithm, contribute to what we call the experiment error, which is the image JOD score estimation error.

Estimating the uncertainty in IQA. One way of quantifying the experiment error is by calculating the confidence interval (CI) of the JOD scores. The original formulation of the CI does not directly apply to PWC experiments, since it does not take into account the error introduced by the image sampling process [38]. A practical alternative for computing CIs is bootstrapping [8]. We follow the approach proposed in [42] and resort to the percentile method of bootstrapping [6]. We repeatedly generate JOD scores by sampling, with replacement, the observer comparison matrices, each of which is a unique opinion on all images. The CI boundaries for each image are then defined as the 2.5th and 97.5th percentiles of the JOD scores (Fig. 2 (a)).

4.2. JOD clustering via confidence intervals

CI limitations. Confidence intervals represent well the sample mean error for independent variables and samples, but in PWC these conditions are not achieved [42]. The psychometric scaling algorithm calculates all the JOD scores at the same time. This means that every change in the comparison matrix will simultaneously affect the scores of all images. This behavior makes the image JOD scores somewhat interdependent, which is not apparent in the CIs. To identify which images have a significant difference in quality, we need to analyze the overlap of their confidence intervals. But how to quantify this overlap? Do we consider the overlap of 20%, 30%, or 60% to be significant? What to do in case of multiple overlaps? We address these issues by combining two techniques. We first cluster the images using their CIs, then, we use variance analysis to identify which images have significant quality differences.

Preliminary clustering. Overlapping CIs may indicate possible quality similarity, so we can use this information to group the images. To define the distance between intervals, we consider each CI as a point $C(x, y)$ in the subspace $\{(x, y) \in \mathbb{R}^2 \mid y - x \geq 0\}$ where x is the lower bound of the CI, and y is the upper bound of the CI. In this way, we can calculate the Euclidian distance between C_1 and C_2 . We then resort to the K-means algorithm [33] to define our preliminary quality groups (Fig. 2 (a-b)). We estimate the

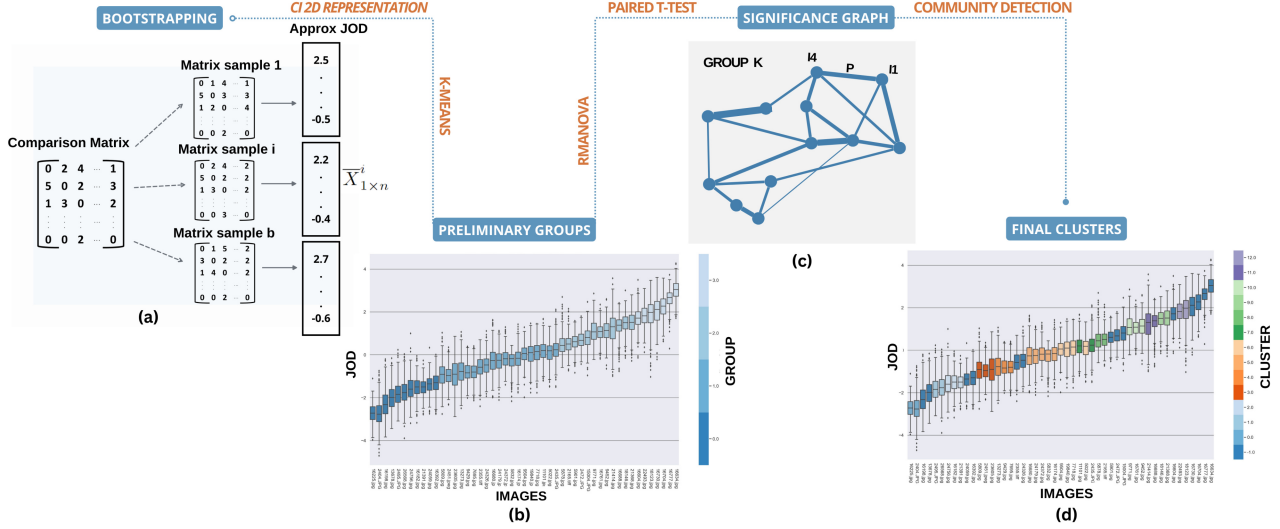


Figure 2. Diagram of the statistical analysis strategy used to estimate the uncertainty in a PWC experiment. (a) Given a PWC matrix, we generate confidence intervals (CIs) using percentile bootstrapping. (a-b) We then apply the K-means algorithm to the “2d representation” of the CIs to cluster the images into preliminary quality groups (b). (b-c) In each group, we apply RMANOVA to detect significant differences between the JOD scores. (c) For groups with such differences, we construct a weighted undirected graph, where the weights consist of the p-value of paired t-tests between the image pairs. (c-d) Finally, we apply Louvain community detection to extract sub-clusters of similar quality inside each group (d). Figures (b) and (d) represent the boxplots of the bootstrapped image scores generated in (a).

number of preliminary quality groups by dividing the total JOD range by the median size of the CIs (Fig. 2 (b)).

4.3. JOD clustering via variance analysis

Variance analysis. To estimate the significance of the CI overlaps, we turn to the analysis of variance (ANOVA) [10, 11] and particularly repeated measures ANOVA (RMANOVA) [18]. RMANOVA is a statistical significance test used to investigate the differences in mean scores of a given continuous variable (called dependent variable), that has been “repeatedly tested”, on the same group of subjects, under three or more different conditions taken from a categorical variable (called the within-subject factor or the independent variable). In a PWC experiment, we interpret the set of images as the independent variable and consider each bootstrapped matrix (Sec. 4.1) as a subject that was tested on different conditions (one matrix \equiv one subject). Finally, since the JOD scores are estimated from the same matrix, we consider them as measurements of the dependent variable, that is, the image quality.

Statistical hypothesis. Let M be the sparse comparison matrix defined in Sec. 4.1. Let $\mathbb{X} = \{\bar{X}_{1 \times n}^i, i = 1, \dots, b\}$, where $\bar{X}_{1 \times n}^i = [\bar{x}_1^i \ \bar{x}_2^i \ \dots \ \bar{x}_n^i]$, be the set of score vectors inferred from b comparison matrices bootstrapped from M , and n the number of images. We define the two

hypotheses:

$$\begin{cases} H_0 : \bar{x}_1 = \bar{x}_2 = \dots = \bar{x}_n; \\ H_1 : \text{At least two means are different.} \end{cases} \quad (1)$$

where \bar{x}_k represents the inferred average score of the image I_k from the PWC experiment. Refusing the null hypothesis H_0 only guarantees that at least two image scores are different (H_1). Accepting H_0 means that all images in the test have indistinguishable quality. We apply the previous hypothesis testing on each of the preliminary groups and deduce whether there is a significant difference between the images or not (Fig. 2 (b-c)). We can identify two cases:

1. **No significant difference has been found:** we consider in this case that all the images of the group have the same average score and variance.
2. **A significant difference exists:** we don’t know how many images are significantly different. In this case, we conduct a post hoc analysis, using paired t-tests [51], at a confidence level of 0.95, on all the possible pairs in the given cluster.

Significance graph. For the groups where the significant difference exists, we create a weighted undirected graph by weighting the connections between pairs with the p-value of their corresponding paired t-tests (Fig. 2 (c)). Then, we apply the Louvain community detection algorithm [2] to group dense regions of nodes into the same “community”

or cluster (Fig. 2 (c-d)). With this method, we separate the graph into sub-clusters, then assign their average score and variance to the corresponding images (Fig. 2 (d)).

4.4. Results and discussion

We show the results of our statistical analysis on 20 scenes for the four attributes in Fig. 3, from which we can make several interesting observations. First, we note that the number of clusters and groups is correlated with the JOD range and the median CI size (rows 1, 2). A wider JOD interval indicates a wider quality coverage, which implies a higher number of quality levels. Similarly, a smaller confidence interval indicates that the images can be more easily separated, which in turn implies a higher number of quality levels. Second, we observe that the median CI size decreases with the JOD range (row 3, left). This confirms our hypothesis in Sec. 4.1 that sampling images of close quality makes the task more difficult. Finally, we note that detail preservation and exposure have smaller CIs, while color has the largest, indicating a greater difficulty in annotating this attribute (row 3, right), thus justifying its omission.

5. Blind image quality assessment with a tweak

Based on the PIQ23 dataset, we introduce a deep BIQA method (SEM-HyperIQA) that adapts to the specific structure of the dataset, where each scene has a separate quality scale. We retrain several existing BIQA methods from the literature and compare them to our proposed approach.

5.1. Semantics aware IQA

The PIQ23 dataset contains individually annotated scenes, each with its own quality scale and unique content. This characteristic introduces a problem known as domain shift (Sec. 2.3), involving both content-dependent and annotation-dependent factors. This emphasizes the need to understand the scene’s semantics and align the predicted quality with its corresponding scale.

In order to address the challenges of domain shift in PIQ23, we propose SEM-HyperIQA, a solution that involves combining the HyperIQA architecture, which integrates semantic information, with multitasking, which allows scene-specific rescaling. Based on the HyperIQA architecture, we concatenate the semantic features of multiple random crops and feed them to a multi-layer perceptron (MLP) that predicts the scene category for the respective image. We then feed the predicted category to a smaller MLP that predicts a multiplier a_i and offset b_i to adapt the predicted quality score of each patch to its respective scene scale, such as $\hat{q}_i = a_i q_i + b_i$, where q_i is the predicted quality score of patch i (Fig. 4). The loss is the sum of the ℓ_1 -norm loss and the cross entropy.

We also propose two other variants, SEM-HyperIQA-SO and SEM-HyperIQA-CO. In the first variant, we omit the

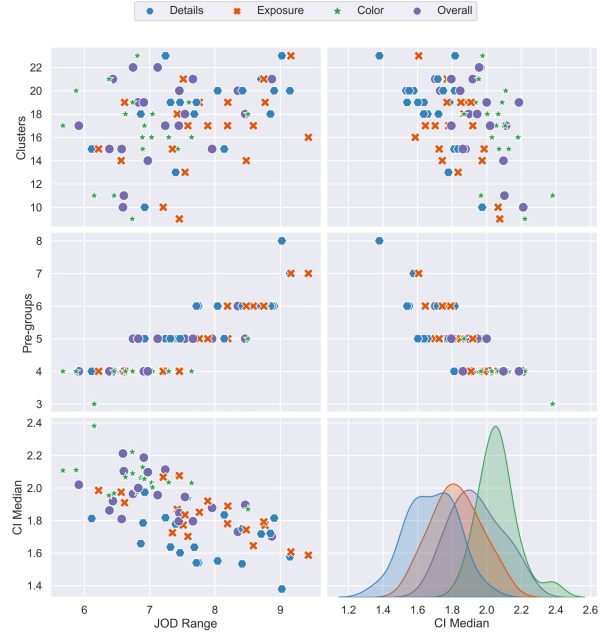


Figure 3. Statistical analysis on 20 scenes for the four attributes. From top to bottom row (shared x-axis): distribution of the number of clusters (row 1) and preliminary groups (row 2) in terms of the JOD range (left) and the median CI size (right). Row 3 displays the distribution of the median CI in terms of the JOD range (left), as well as the median CI distribution per attribute (right).

scene category prediction and instead feed the scene information directly to the MLP that rescales the predicted score. In the second variant, we omit the rescaling part and only keep the scene prediction. The two variants will help us explore the individual importance of scene-specific rescaling and semantic prediction, respectively.

5.2. Performance evaluation

Training strategy. We test different training configurations for all the proposed methods and report the best results. Specifically, we randomly sample 70% of the images in PIQ23 for training and leave the rest for testing. We randomly crop the images to patches of one of the three following sizes: 672, 448, and 224. We use Adam stochastic optimization [28] with different learning rates between 10^{-6} and 10^{-4} . We fix the training for 300 epochs and adopt a learning rate decay factor of 0.05 for every 10 epochs. The final image quality score is computed by averaging the individual patch scores. To evaluate the performance, we compute Spearman’s rank correlation coefficient (SRCC) between the model outputs and the JOD scores. Since each scene is annotated separately, we compute the correlation over the scores for each individual scene and evaluate the performance as $\bar{C} = \frac{1}{s} \sum_{i=1}^s C_i$, where s = number of scenes, C_i = correlation for scene i .

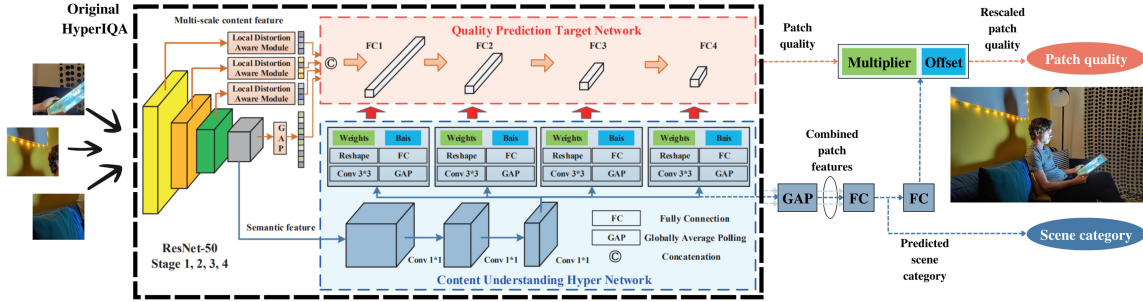


Figure 4. The SEM-HyperIQA architecture. We combine the semantic representation acquired by HyperIQA [52] for multiple patches to predict the scene category. We then use the predicted category to rescale the patch quality. The image score is averaged across patches.

#	Method	Details	Exposure	Overall
1	BRISQUE [36]	0.323	0.307	0.192
2	NIQE [37]	0.378	0.265	0.298
3	ILNIQE [64]	0.353	0.312	0.214
4	DB-CNN [67]	0.628 ± 0.07	0.635 ± 0.06	0.555 ± 0.07
5	HyperIQA [52]	0.649 ± 0.08	0.706 ± 0.04	0.611 ± 0.06
6	MUSIQ [25]	0.671 ± 0.07	0.725 ± 0.04	0.589 ± 0.07
7	SEM-HyperIQA	0.671 ± 0.07	0.71 ± 0.04	0.621 ± 0.06
8	SEM-HyperIQA-SO	0.722 ± 0.06	0.721 ± 0.06	0.642 ± 0.08
9	SEM-HyperIQA-CO	0.664 ± 0.07	0.71 ± 0.06	0.621 ± 0.07

Table 1. Comparison of the baselines according to their average scene Spearman’s rank correlation coefficient with the measured JOD scores and their error margin across the scenes. As shown by the table, the deep learning methods tested perform significantly better than their classical counterparts on PIQ23.

Baseline methods. We compare SEM-HyperIQA with existing BIQA models, including BRISQUE [36], NIQE [37], ILNIQE [64], DB-CNN [67], HyperIQA [52] and MUSIQ [25]. We train these models on PIQ23 using their official implementations. NIQE and ILNIQE do not require any training. DB-CNN and MUSIQ are pre-trained on LIVE Challenge and PaQ-2-PiQ, respectively. HyperIQA is pre-trained on ImageNet. Results are shown in Table 1.

Discussion. From Table 1 we can make the following observations. First, the deep learning methods tested (4-9) perform better than their classical counterparts (1-3), indicating a difficulty to adapt to high-resolution images, scene-specific scales, and attribute-specific annotations. Zhu *et al.* [70] have demonstrated the ineffectiveness of such methods when the annotations do not represent an overall subjective evaluation of the quality. Second, the proposed SEM-HyperIQA method improves upon the original HyperIQA, which indicates the effectiveness of scene semantics and multitasking in quality prediction, especially for separate scene scales. Third, SEM-HyperIQA-SO with scene-specific rescaling achieves the best performance. It notably enhances the detail preservation attribute, possibly due to

the limited information available in face crops for scene analysis. Therefore, semantic information cannot be fully utilized and we are better off using scene-specific rescaling only. Fourth, we note that deep BIQA models perform significantly better for detail preservation and exposure than overall, which directly reflects this task’s difficulty and the uncertainty of the annotations, as discussed in Sec. 4.4.

6. Conclusion

We have presented PIQ23, a new dataset for portrait quality assessment with a wide variety of smartphone cameras and use cases, which has been annotated by image quality experts using pairwise comparisons. We have shown the importance of identifying the uncertainty in the annotations by providing a new statistical analysis method to cluster the quality scale into consistent levels of quality. Finally, we adopt a training strategy and a deep neural network architecture that adapts to the high-resolution images of PIQ23 and profits from semantic information and multitasking, in order to adjust to the per-scene quality scaling of the dataset. Our results have shown the necessity and effectiveness of quality scale quantification and clustering of similar quality images to contain annotation uncertainty, as well as the importance of semantic information in training IQA models. We believe that this work will be the foundation for a new area of application of IQA for portrait images, as well as for a higher caliber of annotations in IQA datasets.

Acknowledgments

This work was funded in part by the French government under the management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), the Louis Vuitton/ENS chair in artificial intelligence and the Inria/NYU collaboration. NC was supported in part by a DXOMARK/PRAIRIE CIFRE Fellowship. We thank DXOMARK engineers and photographers for their time investment and fruitful discussions about image quality.

References

- [1] Lacey Best-Rowden and Anil K Jain. Learning face image quality from human assessments. *IEEE Transactions on Information Forensics and Security*, 13(12):3064–3077, 2018. 3
- [2] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008. 6
- [3] Martin Cadik, Michael Wimmer, Laszlo Neumann, and Alessandro Artusi. Image attributes and quality for evaluation of tone mapping operators. In *National Taiwan University*. Citeseer, 2006. 2
- [4] Frédéric Cao, Frederic Guichard, and Hervé Hornung. Measuring texture sharpness of a digital camera. In *Digital Photography V*, volume 7250, pages 146–153. SPIE, 2009. 1
- [5] Roger R Davidson and Peter H Farquhar. A bibliography on the method of paired comparisons. *Biometrics*, pages 241–252, 1976. 5
- [6] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Number 1. Cambridge university press, 1997. 5
- [7] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotzia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 4
- [8] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992. 5
- [9] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020. 1, 2, 3
- [10] Ronald A Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919. 6
- [11] Ronald Aylmer Fisher et al. 014: On the “probable error” of a coefficient of correlation deduced from a small sample. 1921. 6
- [12] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988. 4
- [13] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015. 1, 2, 3
- [14] Deepti Ghadiyaram and Alan C Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of vision*, 17(1):32–32, 2017. 1, 3
- [15] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1220–1230, 2022. 1, 3
- [16] Yann Gousseau and François Roueff. Modeling occlusion and scaling in natural images. *Multiscale Modeling & Simulation*, 6(1):105–134, 2007. 1
- [17] P Grother, Austin Hom, Mei Ngan, and Kayee Hanaoka. On-going face recognition vendor test (frvt) part 5: Face image quality assessment. *Draft NIST Interagency Report*, 2020. 3
- [18] Ralitza Gueorguieva and John H Krystal. Move over anova: progress in analyzing repeated-measures data and its reflection in papers published in the archives of general psychiatry. *Archives of general psychiatry*, 61(3):310–317, 2004. 6
- [19] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19, 2006. 5
- [20] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019. 3
- [21] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 1, 2
- [22] Chen-Hsiu Huang and Ja-Ling Wu. Multi-task deep cnn model for no-reference image quality assessment on smartphone camera photos. *arXiv preprint arXiv:2008.11961*, 2020. 3
- [23] Kunihiko Kanafusa, Keiichi Miyazaki, Hiroshi Umemoto, Kazuhiko Takemura, and Hitoshi Urabe. A standard portrait image and image quality assessment. In *IS AND TS PICS CONFERENCE*, pages 317–320. SOCIETY FOR IMAGING SCIENCE & TECHNOLOGY, 2000. 3
- [24] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014. 1, 3
- [25] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 1, 3, 8
- [26] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 3
- [27] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11(1):206–220, 2016. 1, 3
- [28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7
- [29] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006, 2010. 2

- [30] Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 653–661, 2021. 3, 4
- [31] Zhang Lijun, Shao Xiaohu, Yang Fei, Deng Pingling, Zhou Xiangdong, and Shi Yu. Multi-branch face quality assessment for face recognition. In *2019 IEEE 19th International Conference on Communication Technology (ICCT)*, pages 1659–1664. IEEE, 2019. 3
- [32] Christian Loebich, Dietmar Wueller, Bruno Kligen, and Anke Jaeger. Digital camera resolution measurements using sinusoidal siemens stars. In *Digital Photography III*, volume 6502, pages 214–224. SPIE, 2007. 1
- [33] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297, 1967. 5
- [34] Rafał K Mantiuk, Anna Tomaszewska, and Radosław Mantiuk. Comparison of four subjective methods for image quality assessment. In *Computer graphics forum*, volume 31, pages 2478–2491. Wiley Online Library, 2012. 3
- [35] Aliaksei Mikhailiuk, Clifford Wilmot, Maria Perez-Ortiz, Dingcheng Yue, and Rafał K Mantiuk. Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2559–2566. IEEE, 2021. 4, 5
- [36] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 1, 3, 8
- [37] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 1, 3, 8
- [38] Ethan D Montag, Louis leon thurstone in monte carlo: creating error bars for the method of paired comparison. In *Image Quality and System Performance*, volume 5294, pages 222–230. SPIE, 2003. 5
- [39] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364, 2011. 1, 3
- [40] Chahine Nicolas and Belkarfa Salim. Portrait quality assessment using multi-scale cnn. In *London Imaging Meeting*, volume 2021, pages 5–10. Society for Imaging Science and Technology, 2021. 1, 2, 3, 4
- [41] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. Sdd-fiq: unsupervised face image quality assessment with similarity distribution distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7670–7679, 2021. 3
- [42] Maria Perez-Ortiz and Rafał K Mantiuk. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint arXiv:1712.03686*, 2017. 5
- [43] Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafał K Mantiuk. From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing*, 29:1139–1151, 2019. 3, 5
- [44] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015. 2, 3
- [45] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelen-sky, Karen Egiazarian, Marco Carli, and Federica Battisti. Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelec-tronics*, 10(4):30–45, 2009. 2
- [46] Miriam Redi, Nikhil Rasiwasia, Gaurav Aggarwal, and Alejandro Jaimes. The beauty of capturing faces: Rating the quality of digital portraits. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8. IEEE, 2015. 1, 3, 4
- [47] Jacob Rose and Thirimachos Bourlai. Deep learning based estimation of facial attributes on challenging mobile phone face datasets. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1120–1127, 2019. 3
- [48] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing*, 21(8):3339–3352, 2012. 1, 3
- [49] Torsten Schlett, Christian Rathgeb, Olaf Henniger, Javier Galbally, Julian Fierrez, and Christoph Busch. Face image quality assessment: A literature survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–49, 2022. 2, 3, 4
- [50] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 2
- [51] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908. 6
- [52] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2020. 1, 3, 8
- [53] Wei Sun, Xiongkuo Min, Guangtao Zhai, and Siwei Ma. Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *arXiv preprint arXiv:2105.14550*, 2021. 3
- [54] Louis L Thurstone. A law of comparative judgment. *Psychological review*, 101(2):266, 1994. 5
- [55] Oliver van Zwanenberg, Sophie Triantaphillidou, Robin Jenkin, and Alexandra Psarrou. Edge detection techniques for quantifying spatial imaging system performance and image quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1
- [56] Toni Virtanen, Mikko Nuutinen, Mikko Vaaherankosa, Pirkko Oittinen, and Jukka Häkkinen. Cid2013: A database

- for evaluating no-reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 24(1):390–402, 2014. 1, 2
- [57] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 995–1002, 2013. 1, 3
- [58] Fei Yang, Xiaohu Shao, Lijun Zhang, Pingling Deng, Xiangdong Zhou, and Yu Shi. Dfqa: Deep face image quality assessment. In *International Conference on Image and Graphics*, pages 655–667. Springer, 2019. 3
- [59] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 1, 3
- [60] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1098–1105. IEEE, 2012. 1, 3
- [61] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3585, 2020. 1, 2
- [62] Junyong You and Jari Korhonen. Transformer for image quality assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1389–1393. IEEE, 2021. 1, 3
- [63] Emin Zeman, Giuseppe Valenzise, and Frederic Dufaux. An extensive performance evaluation of full-reference hdr image quality metrics. *Quality and User Experience*, 2(1):1–16, 2017. 3
- [64] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 1, 3, 8
- [65] Lijun Zhang, Lin Zhang, and Lida Li. Illumination quality assessment for face images: A benchmark and a convolutional neural networks based model. In *International Conference on Neural Information Processing*, pages 583–593. Springer, 2017. 3
- [66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2, 3
- [67] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018. 1, 3, 8
- [68] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Learning to blindly assess image quality in the laboratory and wild. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 111–115. IEEE, 2020. 3
- [69] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021. 1, 3
- [70] Wenhan Zhu, Guangtao Zhai, Zongxi Han, Xionghuo Min, Tao Wang, Zicheng Zhang, and Xiaokang Yang. A multiple attributes image quality database for smartphone camera photo quality assessment. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2990–2994. IEEE, 2020. 1, 2, 8