# SQAD: Automatic Smartphone Camera Quality Assessment and Benchmarking

Zilin Fang[1]*, Andrey Ignatov[2], Eduard Zamfir[3], Radu Timofte[2,3]
[1]National University of Singapore, [2]ETH Zurich, [3]University of Wurzburg

## Abstract

*Smartphone photography is becoming increasingly popular, but fitting high-performing camera systems within the given space limitations remains a challenge for manufacturers. As a result, powerful mobile camera systems are in high demand. Despite recent progress in computer vision, camera system quality assessment remains a tedious and manual process. In this paper, we present the **S**martphone Camera **Q**uality **A**ssessment **D**ataset (SQAD), which includes natural images captured by 29 devices. SQAD defines camera system quality based on six widely accepted criteria: resolution, color accuracy, noise level, dynamic range, Point Spread Function, and aliasing. Built on thorough examinations in a controlled laboratory environment, SQAD provides objective metrics for quality assessment, overcoming previous subjective opinion scores. Moreover, we introduce the task of automatic camera quality assessment and train deep learning-based models on the collected data to perform a precise quality prediction for arbitrary photos. The dataset, codes and pre-trained models are released at* https://github.com/aiff22/SQAD.

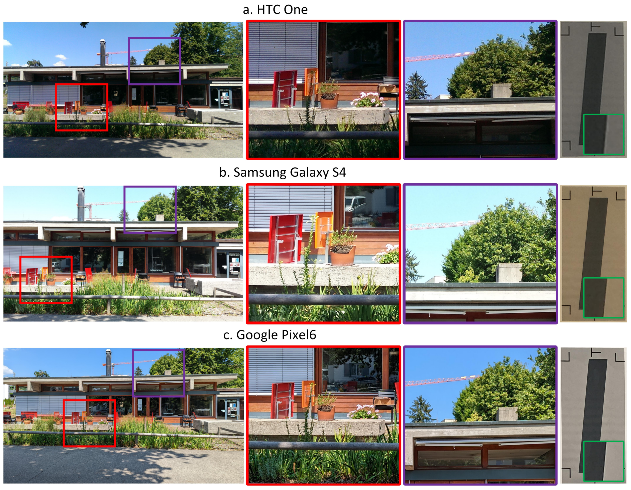


Figure 1: *Sample images of the identical scene from our benchmark captured using different devices.* The training and testing images are captured in hand-held mode in the wild, while the optical patterns, *e.g.* PSF, are obtained under controlled conditions with a fixed tripod 2 meters apart for benchmarking purposes.

## 1. Introduction

Image quality, meaning the representation of details, color, realism, *etc*., has shown to directly influence computer vision algorithms in achieving a more comprehensive understanding of the natural world. Consequently, this evokes the question of precisely quantifying and estimating the perceived image quality by human viewers, as they are usually the ultimate receivers [57]. Besides subjective methods that study human perceptions given physical stimuli [28, 43], objective assessment methods fall into three main categories based on the amount of information available: *no-reference*, *reduced-reference*, or *full-reference*.

No-reference (NF) methods for estimating image quality directly, without requiring undistorted counterparts, are more convenient in realistic vision scenarios. Common quality metrics include image statistical differences, such as BRISQUE [41], BLIINDS-II [48] parameterized by DCT coefficients, NIQE [42] with spatial domain features, and the successor IL-NIQE [62] based on multivariate Gaussian models. Recently, learning-based approaches have led to improved quality assessment methods [36, 13, 38]. However, gaps still exist between perceivable image quality and predicted quantitative results even with neural networks providing better solutions than traditional methods. To address this, new algorithms [9, 29] are being proposed, inspired by GANs [23] and perceptual-oriented optimization [30, 60, 46], that aim to further reduce this gap.

An orthogonal direction in assessing image quality is to consider the device used to take those pictures, which requires not only the recorded imagery but the corresponding camera devices and photometric settings. In this case, manual impairments are not employed to imitate different situations, and quality is measured directly in a more accurate way. Optical patterns are widely applied, making measurements more objective and meaningful [52, 11, 12]. They cover a wide range of distinct quality factors related to

[*] Work done while at ETH Zurich

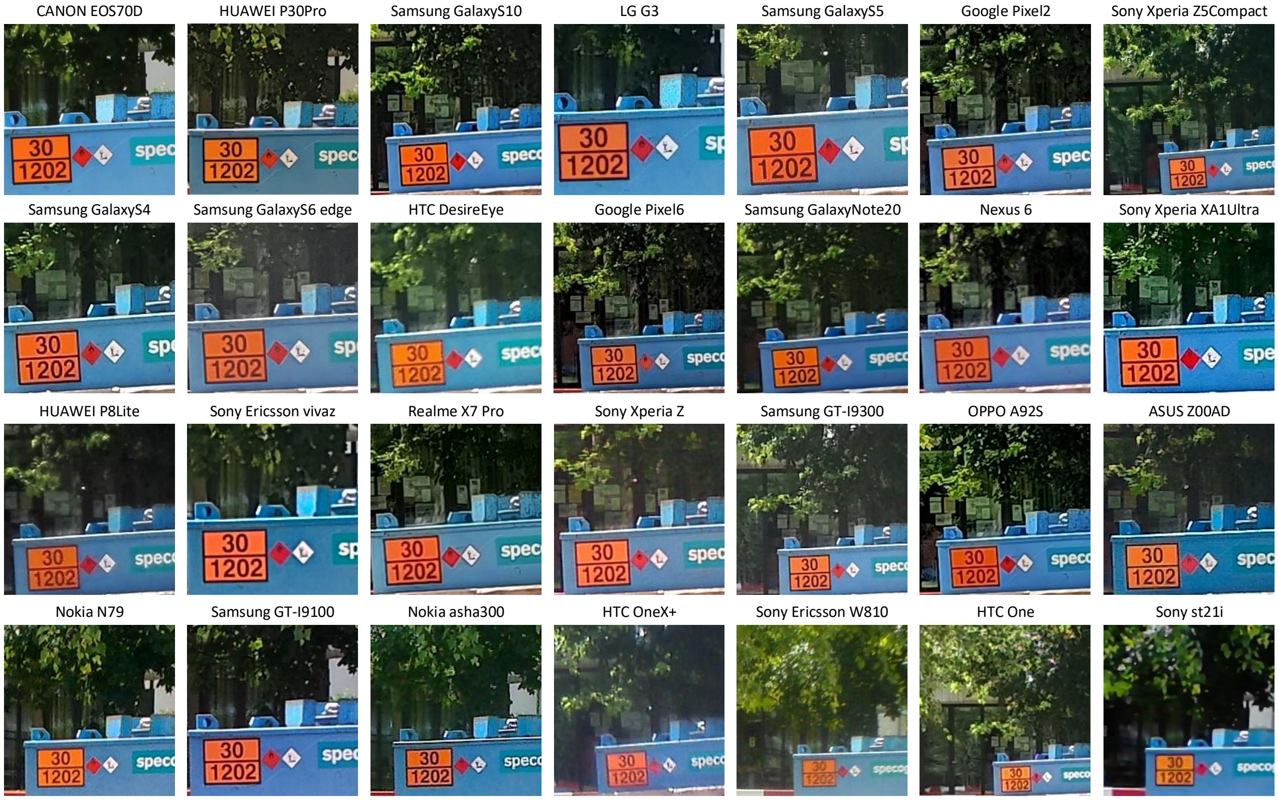


Figure 2: *Smartphone Camera Sensor Quality Assessment Dataset (SQAD).* We present 224 × 224 sized crops extracted from our dataset consisting of images captured by 29 different mobile devices. Images are arranged in descending order of resolution. Orange block scales vary due to differing image dimension settings on smartphones. Perceived sharpness relates more to PSF than resolution. The image taken with Sony Ericsson T630 is in the supplement due to its small size.

the photon nature of light, properties of lenses, signal theory, and more. International standards [5, 6] provide measuring and reporting methods for still-picture cameras based on designated patterns. Besides, several popular commercial software applications [2, 1] provide camera quality tests, including main quality factors such as sharpness, noise level, dynamic range, chromatic aberration, etc. While various methods have been proposed for image quality assessment (IQA), they often have limitations and are not comprehensive. Although IQA methods aim to predict quantitative scores aligning with human perceptions measured by mean opinion scores (MOS) [54], they often lack objectivity, quantization, and physical reasoning due to diverse participant expertise levels and inherent biases. Camera-based methods are accurate and based on physical attributes, but they are complex and time-consuming, and require cameras and their parameters, making them less suitable for web images.

To address previous limitations, we are motivated to provide a standard, measurable, and objective reference, and we make the following contributions: Our proposed SQAD benchmark combines device-based lab measurement approaches with Deep Learning to accurately quantify smartphone camera quality from multiple perspectives using physically meaningful scores. We use a semi-manual evaluation protocol for smartphone camera systems based on ISO standards, covering critical quality factors such as resolution, color accuracy, noise level, dynamic range, aliasing, and Point Spread Function (PSF). Our evaluations are conducted in real-world conditions, reflecting practical scenarios where access to individual camera components, *i.e.* optics and ISP, is limited. As images serve as the ultimate observation of camera quality, we therefore define the camera sensor[1] to encompass all relevant components. Our image collection comprises natural scenes captured by a variety of devices, spanning from the early years of smartphone development to recent high-end phones. Furthermore, we introduce the task of automatically assessing smartphone camera quality, eliminating the need for manual measurements. An overview of our dataset is presented in Fig. 2.

[1] In this work, the term *camera sensor* refers to both hardware and ISP.

## 2. Related Work

We briefly recall closely related research topics in the field of IQA and discuss widely used benchmark databases.

**Image quality assessment.** In full-reference IQA, methods evaluate the differences between a test image and a high-quality reference and provide a score indicating their similarity. Traditional metrics like PSNR and SSIM compute pixel-wise differences or evaluate images by combining luminance, contrast, and structure factors [58]. Advanced Deep Learning methods extract learned features from the reference and test images and compute distances in high dimensional feature space. These deep representations are highly correlated with perceptual image quality and yield superior results, as seen in DeepSIM [21], WaDIQaM [10], DeepQA [31], PieAPP [45], LPIPS [63] and DISTS [18].

In some practical scenarios, high-quality reference images may be unavailable, prompting the development of NF methods that offer solutions. The advent of deep learning has significantly improved the field by introducing databases [35, 29, 8], thereby fostering further developments in this area [59, 64, 39, 16]. Some previous works also focus on quality assessment of distorted images captured in the wild [34, 53], which differ in appearance from synthetic imagery. Su et al. [55] divide the problem into three stages involving semantic feature extraction, perception rule learning, and quality prediction to develop a self-adaptive hypernetwork that performs well on both synthetic and realistic distortions. UNIQUE [65] samples image pairs and employs a fidelity loss to train a unified model for both synthetic and realistic distortions. MetaIQA [66] adopts meta-learning to learn prior knowledge shared by diversified distortions, thereby generalizing better to a vast amount of synthetic and authentic distortions. Chiu et al. [14] link image quality assessment to captioning, vision question answering, and object recognition. More recently, the winner [61] of the NTIRE 2022 challenge [24] (NF-IQA track) employs a transposed attention mechanism to modulate deep features extracted from a ViT [19] backbone, while Swin Transformer [37] layers enhance the interactions of local information before the final quality score prediction.

**Datasets.** To facilitate IQA tasks, numerous databases have been proposed in the past to model various types of synthetic distortions, such as LIVE [50, 51], CSIQ [33], and TID13 [44]. However, in recent years, KADID-10k [35] has emerged as a larger dataset with more diverse contents, where distorted images are categorized into five levels of distortion. With the significant progress achieved in the field of low-level vision through the use of generative techniques, there has been a noticeable gap between the perceptual quality and quantitative evaluations [30, 63]. In light of this, the PIPAL [29] dataset has been developed to include GAN-generated distortions, providing an opportunity for in-depth exploration of perceptual image errors. In addition to hand-crafted and algorithm-based distortions, several image databases contain authentic distortions, such as BID [15], LIVEW [22], and KonIQ-10k [27], which are sampled from multimedia database YFCC100m [56]. SPAQ [20], influenced by the smartphone photography trend, collects a large number of smartphone images with image attributes and scene category labels. Despite the richness of synthetic database distortions and authentic database image contents, quantitative image quality results rely on subjective human opinions. Although KonIQ-10k [27] and SPAQ [20] include several image attributes, they are based on image statistics and subjective ratings, respectively.

Table 1: *IQA databases by size, judgment and distortions.*

| Dataset | # Images | Judgement types | Distortion types |
|---|---|---|---|
| LIVE [50, 51] | 779 | DMOS | synthetic (5) |
| CSIQ [33] | 886 | DMOS | synthetic (6) |
| TID13 [44] | 3,025 | MOS | synthetic (24) |
| KADID-10k [35] | 10,125 | MOS | synthetic (25) |
| PIPAL [29] | 29,250 | MOS | syn. (40) & GAN dist. |
| BID [15] | 585 | MOS | authentic |
| LIVEW [22] | 1,162 | MOS | authentic |
| KonIQ-10k [27] | 10,073 | MOS | authentic |
| SPAQ [20] | 11,125 | MOS | auth. & meta-info |
| **SQAD** (*ours*) | 3,017 | Lab. results | auth. & quality attr. |

In contrast to other image databases that rely on subjective human opinions, our SQAD benchmark uses laboratory measurements to obtain objective scores with greater physical accuracy. Compared to SIDD [7] that only estimates noise ground truth from noisy images with five devices, we measure six various criteria using 29 smartphones. Additionally, we propose the novel task of smartphone camera quality assessment, with a diverse image collection and objective ground truth labeling. Furthermore, while previous works mostly emphasize the *perceptual* aspect, our benchmark provides a technical perspective on image quality, enabling an objective comparison between different smartphone devices. Tab. 1 briefly summarizes the differences between the mentioned IQA databases.

## 3. Smartphone Camera Quality Assessment

Introducing our **S**martphone Camera **Q**uality **A**ssessment **D**ataset (SQAD), we begin by providing detailed information about the dataset and its acquisition process in Section 3.1. Then, in Section 3.2, we delve into the technical background of the quality aspects considered.

### 3.1. Dataset Overview and Benchmarking

Our data collection consists of 29 devices and a total of 3017 images depicting diverse indoor and outdoor scenes captured in the wild without introduced patterns for deep learning purpose. In Tab. 2, we provide an overview of the

Table 2: *Overview of SQAD.* We present a breakdown of SQAD training split and include a Canon EOS70D as a high-quality reference device. More details in supplement Tab. 5.

| Manufacturer | Samsung | Sony | HTC | HUAWEI | Google | LG | Nokia | OPPO | Realme | ASUS | Canon | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # Devices | 7 | 7 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 29 |
| # Images | 786 | 709 | 358 | 169 | 177 | 239 | 238 | 88 | 90 | 114 | 49 | 3017 |

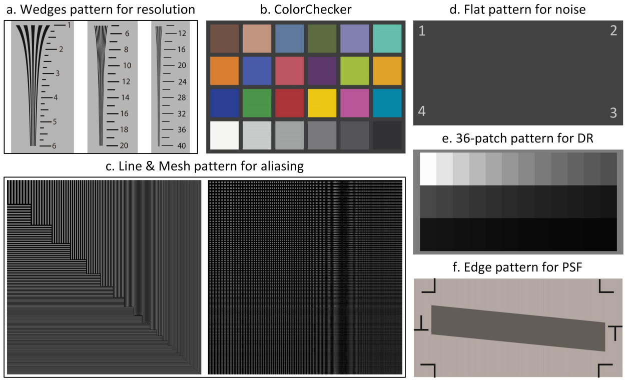

a. Wedges pattern for resolution
1
2
3
4
5
6
6
8
10
12
14
16
18
20
12
16
20
24
28
32
36
40
b. ColorChecker
d. Flat pattern for noise
1
2
4
3
e. 36-patch pattern for DR
c. Line & Mesh pattern for aliasing
f. Edge pattern for PSF

Figure 3: *Patterns used for benchmarking in the laboratory.* (a) Wedge patterns. (b) ColorChecker$^{\text{TM}}$ [4]. (d) Flat pattern. (e) 36-patch pattern with gray background. (f) Edge pattern.

individual devices included in our dataset. The images were taken during different times of the day, including morning, noon, afternoon, and sunset, and under various weather conditions, such as sunny or cloudy. The indoor images include *e.g.* close-ups of multiple objects and various interior settings. In some scenes, we included the ColorChecker$^{\text{TM}}$ [4] pattern to compensate for the lack of vivid colors, such as in lawn scenes. Sample images captured in the morning from the considered smartphones are presented in Fig. 2. We utilize sRGB images for our benchmarking and additional experiments, as RAW image capturing is only available on a small fraction of recent high-end devices. Our aim is to provide a comprehensive overview of smartphone development in recent years. Using RAW images would reduce the diversity of our benchmark, as it would exclude most devices prior to 2015. To ensure consistency, we use the default camera settings, *i.e.* ISO values and aperture, for each device when capturing images.

For the benchmarking process, we mount the devices securely on a tripod to reduce camera shake and ensure a consistent distance between the device and pattern during image acquisition for GT measurements. The patterns are positioned at the center of the frame with similar pattern-frame ratio. In Fig. 3, we present an overview of the patterns used for benchmarking. We sourced the patterns from ISO standards [5] for assessing resolution and Point Spread Function (PSF), reputable commercial companies [2, 4] for evaluating color accuracy and dynamic range, and our self-designed patterns for measuring noise and aliasing. Our experiments are conducted in a dark room to measure noise and dynamic range, and in illuminated environments for the remaining factors. To account for the nonlinear exposure-pixel relationship in sRGB space, we apply gamma correction to each captured image. We transform the images into the color luminance channel for analysis, with the exception of color accuracy.

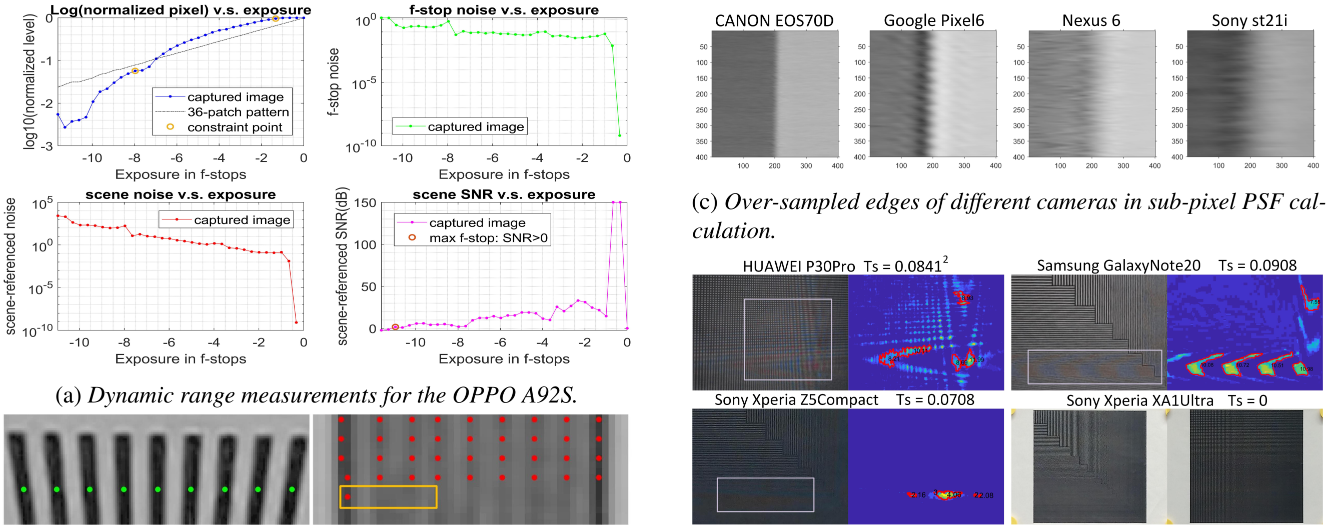
### 3.2. Quality Factors for Benchmarking

We include 6 key image quality factors in our benchmarking protocol, namely resolution, color accuracy, noise, dynamic range, Point Spread Function, and aliasing.

**Resolution.** Resolution is a critical quality factor that describes a camera's ability to render detail with well-defined contrast and texture [5]. It is typically expressed as a single-valued metric and is determined by factors such as the camera lens, addressable photoelements, and electrical circuits. The ISO 12233 standard for still picture imaging offers several test charts for measuring visual resolution, such as the Edge SFR or Siemens star. For our study, we chose the wedge pattern due to its perceived directness in estimating the resolution capability of the devices under investigation. In Fig. 3a, we provide an overview of all the applied wedges, which cover a broad range of resolutions for digital cameras. The wedge chart contains hyperbolic or logarithmic bar patterns [2], and the resolution is determined by the position where the lines become visually indistinguishable from one another. We initialize one peak value point for each line and trace them until the maximum (min. peak prominence $= 2$) cannot be found, as shown in Fig. 4b.

**Color Accuracy.** Color accuracy (CA) refers to a camera's ability to capture colors and shades accurately. However, CA is not always synonymous with color pleasantness as perceived by humans [47], making it a somewhat ambiguous quality factor. To measure CA, we use the classic ColorChecker$^{\text{TM}}$[4] as a reference (see Fig. 3b) and compute the sum of color-differences for all 24 patches (first three rows) within the perceptually uniform CIELAB space. The color-difference formula we use is from CIEDE2000 [49] [2]. Prior to computing the color difference, we apply a different gamma correction using the last row to linearize the image. The estimated gamma value is determined by the slope of the logarithmic pixel response relative to the patch density.

[2] For more details, please refer to the supplementary material.



(a) *Dynamic range measurements for the OPPO A92S.*

(b) *Tracing peak points for wedges in resolution measurement.*

(c) *Over-sampled edges of different cameras in sub-pixel PSF calculation.*

(d) *Comparison of aliasing detection.*

Figure 4: (a) The constraint point in *left top* sub-figure represents upper and lower limits: brightness $< 98\%$ max. RGB and slope $> 7.5\%$ max. response slope on the dark side. Note, scene SNR in *right bottom* sub-figure is truncated to 150 dB if too high on the bright side. (b) Peak points tracing for Samsung Galaxy S10, green dots are initial points and intermediate results are red. The orange rectangle highlights the termination. (c) FWHM values for these four cameras are 17.8, 39.5, 64.8 and 87.8, respectively. (d) We provide picture-heatmap pairs for detected aliasing regions and an example for undetected case.

**Noise Level.** Image noise is an unwanted variation of an imaging system's response caused by the photon nature of light and the thermal energy of heat. This randomness makes noise level a critical factor in overall camera sensor quality. The noise level can be influenced by pixel size, ISO speed setting, exposure time, signal processing components, and more [25]. Measuring the noise level is standardized by ISO 15739 [6], which specifies measuring and reporting the noise versus signal level. Typically, with commonly-used flat patch patterns, noise is approximated as the root mean square (RMS) value measured in pixels, which is equivalent to the standard deviation of the flat patch signal $S$: RMS noise $= \sigma(S)$, since Noise Power $=$ (RMS Noise)$^2$. In our study, we estimate the noise level using a self-designed flat pattern, shown in Fig. 3d, and directly report RMS values over the center region of the captured image as the final noise level estimate. This approach simplifies the calculation of signal-to-noise ratio (SNR) *vs.* input density with multi-patch patterns used in ISO 15739 [6]. The numbers added to the image corners aim to limit autofocus errors since smooth regions lead to blurred measurements. To ensure color uniformity while reducing systematic RMS errors, we display the designed pattern using an OLED screen with a peak and minimal brightness of 605 and 0 nits. Additionally, we solely utilize the center region of the recorded pattern to avoid erroneous estimates due to uneven edge areas.

**Dynamic Range.** Dynamic range (DR) is a critical parameter that measures the luminance range where a camera can produce acceptable SNR and contrast. It is typically expressed as the ratio between the maximum and minimum measurable signal values. An ideal digital camera should capture images with both dark shadows and bright highlights, similar to what the human eye perceives. However, while the DR of real-world scenes can surpass 1,000,000:1 (120 dB), the human eye can detect up to 46.5 f-stops (280 dB), vastly outperforming any consumer camera with a limit of around 100 dB for most devices. To quantify the DR, we first introduce the definitions for SNR and noise, as defined by [6]. The SNR is given by $SNR = \frac{gL}{\sigma}$, where $g = \frac{dpixel}{dL}$, and $\sigma$ is the noise measured in pixels, and $L$ is the luminance derived from patch density $d$, with $L = 10^{-d}$. To express the DR in terms of scene-referenced SNR, as inspired by [2], we modify the transmissive 36-patch chart, which is more accurate than the reflective ISO chart, and adjust its arrangement to fit a rectangular screen better while keeping the density unchanged. Please refer to a more detailed view of the computation in the supplement.

To replicate the transmissive chart illuminated by a backlight, an OLED screen is used to display the custom pattern, ensuring that appropriate dark regions are present. Ideally, the measured range for DR should be from the sensor-saturated patch (less than 98% of maximum RGB) to where $SNR = 1$ (0dB). However, due to flare light severely impacting DR measurement and the lack of a gap between each patch in our pattern, slope-based SNR [3] is applied to limit the measured SNR$_{scene}$ from dark regions. An exemplary output for the OPPO A92S smartphone is shown in Fig. 4a.

**Point Spread Function.** In terms of image quality, the Point Spread Function (PSF) is a crucial factor that characterizes the impulse response of an imaging system. The Fourier transform of PSF is the Modulation Transfer Function (MTF) or Spatial Frequency Response (SFR), which describes the contrast change by the camera as a function of spatial frequency. Although SFR is related to PSF, the latter provides a more comprehensive description of image quality. The International Organization for Standardization (ISO) provides a slanted edge-based method for measuring SFR, as specified in [5]. However, for our purposes, we only need to obtain PSF. After capturing an image, we detect and fit the long slanted edge within a chosen height of 400 pixels. Along two sides of the edge, we oversample the pixel intensities at a small distance of 20 pixels and project them onto a one-dimensional line perpendicular to the edge to create the intensity profile called the Edge Spread Function (ESF). The derivative of the ESF is the Line Spread Function (LSF), which is the line integral of PSF. In practical applications, the difference between LSF and PSF is negligible, so they can be used interchangeably [40]. We report the quantitative result in terms of full-width-at-half-maximum (FWHM), which is commonly used to describe the width of a function. A smaller FWHM indicates a better PSF and a sharper edge in the processed image, as shown in Fig. 4c.

**Aliasing.** Aliasing occurs when signals contain frequencies above the Nyquist criteria and are sampled, resulting in errors. This behavior is typically induced by Bayer color filtering characteristics and often manifests as a color moiré pattern in images. However, unlike other quality factors, there is no established standard for measuring color moiré. To address this, we use the (R-B) parameter [2], which has proven useful in previous experiments. Our measurements rely on self-designed line and mesh patterns, as shown in Fig. 3 (c). The picture is divided into 30 blocks, each containing 8 stripes with logarithmically increasing spatial frequency. By calculating (R-B), we create a heatmap for both patterns. We then detect the largest five bright regions after binarizing the heatmap and determine the position of color moiré based on the region centroid with the maximum mean value. The smaller the period $T_s$, the later color moiré appears (i.e., the higher the frequency $f$), indicating better performance in the aliasing aspect. Four examples of the aliasing measurement are shown in Fig. 4d, and complete detection results are provided in the supplementary material.

# 4. Exploration of the Dataset

## 4.1. Preliminaries

We start our experiments with an exploration of the collected data. The first question we would like to answer is whether deep learning-based models can pick up the general image quality signal from the SQAD dataset and learn to differentiate images captured with different camera devices. For this, we consider the task of smartphone camera classification and analyze several CNN/Transformer-based methods. We mainly conduct our experimentation using ResNet50 [26] and the method proposed by Yang et al. [61], the winner of the NTIRE 2022 NF-IQA challenge [24]. To further take advantage of our dataset, we introduce the task of *automatic* camera quality assessment for smartphone devices, where the models are trained to predict the 6 key image quality factors introduced in the previous sections.

**Training details.** Besides established neural architectures, *e.g.* ResNet50 [26], we select MANIQA [61] as baseline due to its design for NR-IQA. In our experiments, we follow the training settings described in [61] and use Adam optimizer to minimize our loss functions, cross-entropy loss for classification and the MSE loss for quality factor regression. We deviate from described training settings when using the ResNet [26] baseline and train for 200 epochs with a batch size of 64 and a learning rate of $1 \times 10^{-3}$, along with a step scheduler having step size and decay factor set to 50 and 0.5. Furthermore, we randomly extract crops of size $224 \times 224$ and add vertical and horizontal flipping as augmentations. Cropping ultimately results in an increased dataset size, with the number of training samples exceeding 300k. We neglect data augmentations which greatly impact perceptual image quality, such as blurring, color jittering, or inversion. The training-validation-test ratio is approximately 10:1:1.

**Evaluation protocol.** For evaluating our trained networks, we apply additional ensemble techniques. As in other previous works [32, 16, 61], we crop the test image multiple times (16×) randomly. To filter out outliers, such as patches that only contain flat regions like the sky, the final quality factors and class predictions are computed by taking the medians of all patch predictions. Following prior works, we measure the predicted performance of the baseline models trained on SQAD by Spearman's rank-order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC). Both metrics are in range $[-1, 1]$ where positive values denote positive correlation.

## 4.2. Experiments

**Smartphone Camera Classification.** The task of device classification involves predicting the device/sensor used to capture a photo. Although this task may initially seem redundant due to the availability of camera information in EXIF data, EXIF data is usually removed during post-processing or online sharing, which highlights the importance of device classification. Additionally, we aim to investigate the inherent quality signal in our data collection, which makes device classification a valuable task to explore. We adopt the ResNet-50 [26] and MANIQA [61] networks as our baseline

Table 3: *Smartphone Classification and Quality Factor Regression.* We compare ResNet50 [26] and MANIQA [61] on the SQAD test set. (a) We train for camera device classification and report Top-1 and Top-5 accuracy on the test split. (b) We train for quality factor regression and report SROCC/PLCC (↑). We use single-crop and multi-crop (16×) predictions.

(a) *Smartphone Classification.*

| Method | Top-1 (%) | Top-5 (%) |
|---|---|---|
| ResNet50 [26] | 73.33 | 93.95 |
| MANIQA [61] | 93.14 | 99.23 |

(b) *Smartphone Camera Quality Factor Regression.*

| Resolution | | Color Accuracy | | Noise | | Dynamic Range | | PSF | | Aliasing | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SROCC | PLCC | SROCC | PLCC | SROCC | PLCC | SROCC | PLCC | SROCC | PLCC | SROCC | PLCC |
| .8959 | .8802 | .8241 | .8409 | .7758 | .8953 | .6771 | .6373 | .6652 | .8197 | .5987 | .8121 |
| .9431 | .9426 | .9578 | .9613 | .9023 | .9727 | .8441 | .8398 | .8465 | .9256 | .7398 | .8473 |
| .9292 | .9462 | .9474 | .9593 | .9337 | .9785 | .8781 | .9187 | .9430 | .9829 | .9175 | .9812 |
| .9716 | .9842 | .9857 | .9903 | .9809 | .9981 | .9777 | .9862 | .9902 | .9966 | .9826 | .9970 |

methods by changing the final layer to output a probability distribution over the 29 camera devices. Tab. 3a presents classification results of our baseline models that are pre-trained on the ImageNet [17] dataset and then trained on our dataset. Our evaluation results reveal that MANIQA [61] consistently achieves high Top-1 and Top-5 accuracies on our test split, outperforming the ResNet50 [26] baseline. Furthermore, this confirms that image quality is strongly dependent on the sensor used, and each sensor has a distinctive quality signature that sets it apart. As a result, device classification is crucial, and our findings demonstrate this clearly.

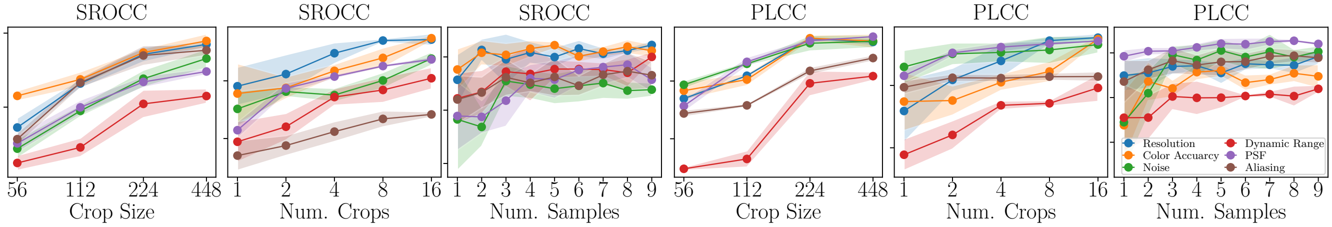
**Smartphone Camera Quality Factor Regression.** Our objective for the task of quality factor regression is to automate the process of determining considered factors from images of real-world scenes captured with respective devices, thereby eliminating the need for manual measurement. In Tab. 3b, we report the results of training ResNet50 [26] and MANIQA [61] instances to regress a single quality factor. We denote the input mini-batch by $\{x_i, s_i\}_{i=0}^{N}$ where $s_i$ is the normalized target quality factor. The model output is a scalar $\hat{s}_i$ representing predicted quality score. Thus, we formulate the MSE loss as $\mathcal{L}_{MSE} = \frac{1}{N}\sum_i^N (s_i - f_\theta(x_i))^2$. Both baselines are successful at regressing quality factors, showing consistently high correlation scores. Still, some quality factors, *e.g.*, dynamic range, PSF, and aliasing, emerge as more difficult to predict. Furthermore, using multiple crops during inference offers enhanced performance, underlying the importance of multiple patch evaluations. Normalizing the quality factors to a range of $s \in [0, 1]$ instead of directly regressing them to their ground truth values (as shown in Tab. 5 in the supplement) provides several advantages. Firstly, it helps the model converge more effectively. Secondly, it highlights the ranking aspect of our dataset and accounts for any potential errors that may have arisen during benchmarking.

**Ablations.** Inadequate crop size can result in inferior performance while the image content is highly influenced by its size. Therefore, we explore training with larger crop sizes and extracting patches from various image locations to obtain better and more robust predictions. Furthermore, we investigated the relationship between the number of crops extracted from test samples and the number of test samples available per camera device. This investigation is motivated by a potential deployment scenario where practitioners may need to provide multiple images for quality factor regression. More precisely, we study whether we not only improve the correlation scores, but also the trustworthiness of our predictions. Fig. 5 visualizes conducted ablation studies using the ResNet50 [26]. We perform our analysis 5 times with varying random seeds and report mean and standard deviation. As anticipated, using small image crops leads to suboptimal performance. However, we found that increasing the number of extracted patches from the test samples not only improves performance but also reduces the deviations in correlation metrics. Additionally, we found that utilizing multiple test samples captured from the same device also exhibits similar behavior. Certain quality aspects, such as DR, are challenging to predict due to various factors, *e.g.* exposure settings during image capture or discontinuities of the test pattern intensity leading to different cameras producing identical DR values. This results in potential errors and inaccuracies not accounted for in our current dataset.

Due to the large variety of smartphones available in the market, it is impractical to measure the quality of each one individually. Therefore, we conduct a more comprehensive study on cross-camera quality prediction, which involves a more rigorous analysis of the relationship between camera quality and various factors. In this experiment, we exclude 5 cameras from the training set and attempt to predict their resolution targets in the test split. To ensure a diverse range of resolutions in the testing phase, we manually select 5 set combinations. In Tab. 4, ResNet50 trained from scratch and Support Vector Regression (SVR) using trained ResNet50 features before the last activation layer are compared to examine the behavior of different regressors. The overall performance, obtained by concatenating results from all sets, achieves SROCC/PLCC scores $> 0.60$, demonstrating the dataset's usefulness and potential for further improvement.

Table 4: *Results of cross-camera regression for resolution.* We report multi-crop correlation metrics (↑) for resolution when excluding different sets of 5 devices from training. Metrics relate each set. The overall score is a concatenation of 25 cameras.

| Method | Set: [IDs] | 26, 28, 08, 18, 11 | | 06, 03, 13, 16, 29 | | 25, 20, 10, 01, 05 | | 23, 09, 04, 19, 27 | | 07, 12, 15, 17, 21 | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | SROCC | PLCC | SROCC | PLCC | SROCC | PLCC | SROCC | PLCC | SROCC | PLCC | SROCC | PLCC |
| ResNet50 [26] | | .5747 | .8161 | .5590 | .4244 | .6700 | .6837 | .4017 | .6539 | .8481 | .6160 | .6212 | .6075 |
| SVR | | .5929 | .8004 | .5541 | .6655 | .7036 | .6781 | .3726 | .6502 | .7913 | .6110 | .6415 | .5141 |



(a)

(b)

Figure 5: *Ablation studies on the crop size, number of crops and number of test samples.* We report *mean* and *std* of (a) SROCC and (b) PLCC results using ResNet50 [26]. We investigate how crop size affects quality factor regression, examine the importance of using multiple crops, and evaluate prediction reliability by increasing the number of test samples per device.

Table 5: *Regression results on web images for MANIQA [61].* We report SROCC and PLCC on multi-crop predictions.

| Resolution | | CA | | Noise | | DR | | PSF | | Aliasing | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| .7794 | .7518 | .6531 | .6726 | .5545 | .5242 | .4333 | .5328 | .6308 | .6699 | .6086 | .5985 |

**Generalization.** Given that most images accessed from web undergo degradation through various methods, we conduct experiments on two degradation approaches: compression and resizing, which aims to evaluate the generalization ability of learned models. With the ResNet50 backbone, resizing all test images yields a top-1 accuracy of 49.4% and 15.7% for ratios $r \in [0.7, 0.5]$, respectively. The expected drop in performance can be limited via a more robust model, *i.e*. stronger data augmentations. More detailed evaluations regarding regression tasks can be found in the supplementary material. Besides, we also include the regression baselines for out-of-distribution data, *i.e*. web images, in Tab. 5. The additional set consists of around 9 images per device obtained from online sources. Despite heavily resized and compressed images, the model exhibits promising generalization with a top-1 classification accuracy of 48.7%.

## 5. Conclusion

This paper introduces the first dataset for evaluating camera sensor quality of mobile devices. Our SQAD benchmark includes 3017 photos captured by 29 distinct devices and validated for quality. Our dataset provides precise measurements of physically-based quality factors obtained in a laboratory setting. The images depict natural scenes and are captured by a diverse range of smartphones, from low-end to high-end devices. We take the next logical step from ordinary assessment of perceptual image quality and introduce a novel task of camera sensor quality estimation, which, to the best of our knowledge, has not yet been explored in the literature. Next, we conduct an extensive study employing our dataset to diverse problems indicating its potential, including camera classification and quality aspects prediction.

The experiments demonstrated that deep learning-based models can easily identify the origin of each unseen photo, reaching a Top-1 accuracy of more than 93%. This shows that the photos captured by each camera sensor have a unique "image quality" imprint that can be used to detect the sensor model used to produce a particular image. When increasing the complexity of the task and trying to predict distinct image quality aspects separately, the obtained results are also quite convincing SROCC/PLCC scores being consistently higher than 0.74 and 0.83, respectively. Besides, the numbers obtained using (sensor-based) cross-fold validation indicate that one can potentially use deep learning-based solutions for image quality assessment of arbitrary and unseen camera sensors by applying models pre-trained on the SQAD dataset to a relatively small number of photos captured with them. One can also expect to see improved numbers when the number of mobile devices/sensor models used for training increases, thus we plan to add more devices to this dataset in the future. Finally, it should be mentioned that the direct correlation between human-perceived quality and sensor quality factors poses a challenging limitation, which is out of the scope of this work. We will also continue exploring this aspect in future research. We hope that the presented dataset and analyses pave the way for future development of effective AI-based camera quality assessment approaches.

## References

[1] Dxomark: What we test and score in camera. https://www.dxomark.com/what-we-test-camera/. 2

[2] Imatest: Solutions on image quality factors. https://www.imatest.com/solutions/iqfactors/. 2, 4, 5, 6

[3] Solutions: Dynamic range. https://www.imatest.com/solutions/dynamic-range/. 5

[4] X-Rite. https://www.xrite.com/. 4

[5] ISO 12233:2017. Photography — Electronic still picture imaging — Resolution and spatial frequency responses. Standard, International Organization for Standardization, 2017. 2, 4, 6

[6] ISO 15739:2017. Photography — Electronic still-picture imaging — Noise measurements. Standard, International Organization for Standardization, 2017. 2, 5

[7] A. Abdelhamed, S. Lin, and M.S. Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, 2018. 3

[8] Anastasia Antsiferova, Alexander Yakovenko, Nickolay Safonov, Dmitriy Kulikov, Alexander Gushchin, and Dmitriy Vatolin. Applying objective quality metrics to video-codec comparisons: Choosing the best metric for subjective quality estimation. pages 199–210. GraphiCon, 11 2021. 3

[9] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 1

[10] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017. 3

[11] 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

[12] Frédéric Cao, Frédéric Guichard, and Hervé Hornung. Dead leaves model for measuring texture quality on a digital camera. In *Digital Photography VI*, volume 7537, pages 126–133. SPIE, 2010. 1

[13] Diqi Chen, Yizhou Wang, and Wen Gao. No-reference image quality assessment: An attention driven approach. *IEEE Transactions on Image Processing*, 29:6496–6506, 2020. 1

[14] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3646–3656, 2020. 3

[15] Alexandre Ciancio, Eduardo AB da Silva, Amir Said, Ramin Samadani, Pere Obrador, et al. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on image processing*, 20(1):64–75, 2010. 3

[16] Marcos V. Conde, Maxime Burchi, and Radu Timofte. Conformer and blind noisy students for improved image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 940–950, June 2022. 3, 6

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[18] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*, 2020. 3

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 3

[20] 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. 3

[21] Fei Gao, Yi Wang, Panpeng Li, Min Tan, Jun Yu, and Yani Zhu. Deepsim: Deep similarity for image quality assessment. *Neurocomputing*, 257:104–114, 2017. 3

[22] 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. 3

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[24] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S Ren, Radu Timofte, Yuan Gong, Shanshan Lao, Shuwei Shi, Jiahao Wang, Sidi Yang, et al. Ntire 2022 challenge on perceptual image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 951–967, 2022. 3, 6

[25] Samuel W Hasinoff, Frédo Durand, and William T Freeman. Noise-optimal capture for high dynamic range photography. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 553–560. IEEE, 2010. 5

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7, 8

[27] 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. 3

[28] P ITU-T RECOMMENDATION. Subjective video quality assessment methods for multimedia applications. 1999. 1

[29] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, pages 633–651. Springer, 2020. 1, 3

[30] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 1, 3

[31] Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. In

*Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1676–1684, 2017. 3

[32] Shanshan Lao, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang. Attentions help cnns see better: Attention-based hybrid image quality assessment network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1140–1149, 2022. 6

[33] 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. 3

[34] Leida Li, Tianshu Song, Jinjian Wu, Weisheng Dong, Jiansheng Qian, and Guangming Shi. Blind image quality index for authentic distortions with local and global deep feature aggregation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 3

[35] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 3

[36] Kwan-Yee Lin and Guanxiang Wang. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 732–741, 2018. 1

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[38] Jupo Ma, Jinjian Wu, Leida Li, Weisheng Dong, Xuemei Xie, Guangming Shi, and Weisi Lin. Blind image quality assessment with active inference. *IEEE Transactions on Image Processing*, 30:3650–3663, 2021. 1

[39] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022. 3

[40] E. W. Marchand. Derivation of the point spread function from the line spread function. *J. Opt. Soc. Am.*, 54(7):915–919, Jul 1964. 6

[41] 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

[42] 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

[43] Margaret H Pinson and Stephen Wolf. Comparing subjective video quality testing methodologies. In *Visual Communications and Image Processing 2003*, volume 5150, pages 573–582. SPIE, 2003. 1

[44] Nikolay Ponomarenko, Oleg Ieremeiev, Vladimir Lukin, Lina Jin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. A new color image database tid2013: Innovations and results. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 402–413. Springer, 2013. 3

[45] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018. 3

[46] Mohammad Saeed Rad, Behzad Bozorgtabar, Urs-Viktor Marti, Max Basler, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Srobb: Targeted perceptual loss for single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2710–2719, 2019. 1

[47] Rajeev Ramanath, Wesley E Snyder, Youngjun Yoo, and Mark S Drew. Color image processing pipeline. *IEEE Signal Processing Magazine*, 22(1):34–43, 2005. 4

[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

[49] Gaurav Sharma, Wencheng Wu, and Edul N Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 30(1):21–30, 2005. 4

[50] HR Sheikh. Live image quality assessment database release 2. *http://live. ece. utexas. edu/research/quality*, 2005. 3

[51] 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. 3

[52] Mukul V Shirvaikar. An optimal measure for camera focus and exposure. In *Thirty-Sixth Southeastern Symposium on System Theory, 2004. Proceedings of the*, pages 472–475. IEEE, 2004. 1

[53] Tianshu Song, Leida Li, Pengfei Chen, Hantao Liu, and Jiansheng Qian. Blind image quality assessment for authentic distortions by intermediary enhancement and iterative training. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 3

[54] Robert C Streijl, Stefan Winkler, and David S Hands. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227, 2016. 2

[55] 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. 3

[56] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 3

[57] Zhou Wang, Alan C Bovik, and Ligang Lu. Why is image quality assessment so difficult? In *2002 IEEE International*

*conference on acoustics, speech, and signal processing*, volume 4, pages IV–3313. IEEE, 2002. 1

[58] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 3

[59] Bo Yan, Bahetiyaer Bare, and Weimin Tan. Naturalness-aware deep no-reference image quality assessment. *IEEE Transactions on Multimedia*, 21(10):2603–2615, 2019. 3

[60] Qingsong Yang, Pingkun Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xuanqin Mou, Mannudeep K Kalra, Yi Zhang, Ling Sun, and Ge Wang. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*, 37(6):1348–1357, 2018. 1

[61] 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. 3, 6, 7, 8

[62] 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

[63] 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. 3

[64] 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

[65] 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. 3

[66] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Metaiqa: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14143–14152, 2020. 3