published in *Computer Vision and Pattern Recognition*, Jun. 2020

Perceptual Quality Assessment of Smartphone Photography

Yuming Fang[1*], Hanwei Zhu[1*], Yan Zeng[1], Kede Ma[2†], and Zhou Wang[3]

[1]Jiangxi University of Finance and Economics, [2]City University of Hong Kong,
[3]University of Waterloo

Abstract

As smartphones become people's primary cameras to take photos, the quality of their cameras and the associated computational photography modules has become a de facto standard in evaluating and ranking smartphones in the consumer market. We conduct so far the most comprehensive study of perceptual quality assessment of smartphone photography. We introduce the Smartphone Photography Attribute and Quality (SPAQ) database, consisting of 11,125 pictures taken by 66 smartphones, where each image is attached with so far the richest annotations. Specifically, we collect a series of human opinions for each image, including image quality, image attributes (brightness, colorfulness, contrast, noisiness, and sharpness), and scene category labels (animal, cityscape, human, indoor scene, landscape, night scene, plant, still life, and others) in a well-controlled laboratory environment. The exchangeable image file format (EXIF) data for all images are also recorded to aid deeper analysis. We also make the first attempts using the database to train blind image quality assessment (BIQA) models constructed by baseline and multi-task deep neural networks. The results provide useful insights on how EXIF data, image attributes and high-level semantics interact with image quality, how next-generation BIQA models can be designed, and how better computational photography systems can be optimized on mobile devices. The database along with the proposed BIQA models are available at https://github.com/h4nwei/SPAQ.

1. Introduction

Perceptual image quality assessment (IQA) aims to quantify human perception of image quality. IQA methods can be broadly classified into two categories: subjective and objective IQA [35]. Although time-consuming and expensive, subjective IQA offers the most reliable way of evaluating image quality through psychophysical experiments [30]. Objective IQA, on the other hand, attempts to create computational models that are capable of automatically predicting subjective image quality [3]. In the past decades, there have been a significant number of studies on both directions [1, 12, 17], most of which focus on synthetic distortions, with the assumption that the original undistorted images exist and can be used as reference [37].

In recent years, there has been a fast development of smartphone photography technologies. From a hardware perspective, dual-camera systems prevail, representing major advancements for the unprecedented photography experience. From a software perspective, computational methods play a more and more important role, introducing novel features such as digital zoom, HDR, portrait and panorama modes. It could be argued that the camera system along with the integrated computational photography module has become a crucial part and one of the biggest selling points of smartphones. Nevertheless, the vast majority of pictures are taken by inexperienced users, whose capture processes are largely affected by lighting conditions, sensor limitations, lens imperfections, and unprofessional manipulations. Arguably it is often challenging for professional photographers to acquire high-quality pictures consistently across a variety of natural scenes, especially in low-light and high-dynamic-range (HDR) scenarios [7]. As a result, real-world smartphone photos often contain mixtures of multiple distortions, which we call realistic camera distortions (as opposed to distortions such as JPEG compression that may be synthesized). Therefore, if the visual quality of the captured images could not be quantified in a perceptually meaningful way, it is difficult to develop next-generation smartphone cameras for improved visual experience.

In this paper, we carry out so far the most comprehensive study of perceptual quality assessment of smartphone photography. Our contributions include:

- A large-scale image database, which we name Smartphone Photography Attribute and Quality (SPAQ)

*Equal contribution
†Corresponding author (email:kede.ma@cityu.edu.hk)

Database	# images	# cameras	Type of cameras	Subjective environment	# image attributes	# scene categories	# EXIF tags
BID	585	1	DSLR	Laboratory	N/A	N/A	N/A
CID2013	480	79	DSLR/DSC/Smartphone	Laboratory	4	N/A	N/A
LIVE Challenge	1,162	15*	DSLR/DSC/Smartphone	Crowdsourcing	N/A	N/A	N/A
KonIQ-10k	10,073	N/A	DSLR/DSC/Smartphone	Crowdsourcing	4	N/A	3
SPAQ	11,125	66	Smartphone	Laboratory	5	9	7

Table 1. Comparison of IQA databases of camera distortions. DSLR: Digital single-lens reflex camera. DSC: Digital still camera. N/A: Not applicable. * LIVE Challenge Database provides the number of manufacturers only.

database, consisting of $11,125$ realistic pictures taken by 66 mobile cameras from eleven smartphone manufacturers. To aid comparison among different cameras, a subset of $1,000$ pictures in SPAQ are captured under the same visual scenes by different smartphones [33]. Each image comes with EXIF data, which provide useful information about the scene being captured (*e.g.*, time and brightness) and the camera settings (*e.g.*, ISO and f-number) [32].

- A large subjective experiment conducted in a well-controlled laboratory environment. We carefully design our experimental protocols to collect the mean opinion score (MOS) for each image and verify its reliability. Additionally, each image is annotated with five image attributes that are closely related to perceptual quality [9]. We also classify the images into nine scene categories by content information to facilitate a *first* exploration of the interactions between perceptual quality and high-level semantics.

- An in-depth analysis of the relationship between EXIF tags, image attributes, scene category labels and image quality based on subjective data. Moreover, the cameras of different smartphones are compared and ranked according to our subjective study.

- A family of objective BIQA models for smartphone pictures based on deep multi-task learning [10]. This allows us, for the first time, to investigate how EXIF tags, image attributes and scene labels affect quality prediction from a computational perspective. More importantly, the results shed light on how to create better photography systems for smartphones.

2. Related Work

In this section, we review representative IQA databases and BIQA models, with emphasis on realistic camera distortions.

2.1. Databases for IQA

Databases have played a critical role in scientific research [27]. In IQA, the creation of the LIVE database [30] validates the perceptual advantages of the structural similarity (SSIM) index [37] and the visual information fidelity (VIF) measure [29] over the widely used mean squared error (MSE). The introduction of the CSIQ [16] and TID2013 [25] databases allows objective IQA models to be compared in cross-database and cross-distortion settings, which highlights the difficulties of distortion-aware BIQA methods in handling unseen distortions. The release of the Waterloo Exploration Database [19] along with the group maximum differentiation (gMAD) competition methodology [18] probes the generalizability of BIQA models to novel image content. The above-mentioned databases facilitate IQA research on how humans and machines assess the perceptual quality of images during processing, compression, transmission, and reproduction, where the degradations can be synthesized. However, they become less relevant when we study smartphone captured images, whose distortions are realistic, complex, and hard to simulate.

There has been limited work studying subjective IQA for realistic camera distortions. Ciancio *et al*. [2] made one of the first steps and built a small dataset of 585 realistically blurred pictures taken by a digital single-lens reflex camera. Toni *et al*. [33] constructed a database that spans eight visual scenes with a number of mobile devices. Ghadiyaram and Bovik created the LIVE Challenge Database [5], which contains $1,162$ images by 15 mobile cameras. The MOSs were crowdsourced via a web-based online user study. The current largest IQA database - KonIQ-10k [9] includes $10,073$ images selected from YFCC100M [32]. The perceptual quality of each image was also annotated via crowdsourcing together with four image attributes. By contrast, the proposed SPAQ database is specifically for smartphone photography with stringent hardware constraints on sensors and optics. Along with EXIF data, each image in SPAQ has image quality, attribute annotations, and high-level scene category labels collected in a well-controlled laboratory environment. Table 1 summaries and compares existing databases for realistic camera distortions.

2.2. Objective BIQA Models

Computational models for BIQA do not require an undistorted reference image for quality prediction of a test

image [36]. Early BIQA models [4, 20–23, 39] mainly focus on synthetic distortions, which have been empirically shown to generalize poorly to realistic camera distortions [5, 41]. This is a consequence of domain shift, and is also referred to as the cross-distortion-scenario challenge in IQA [42]. For BIQA of realistic camera photos, Ghadiyaram and Bovik [6] extracted a bag of natural scene statistics (NSS). Zhang *et al*. [41] proposed a deep bilinear model to handle both synthetic and realistic distortions. Later, they introduced a training strategy [42] that is able to learn a *unified* BIQA model for multiple distortion scenarios. The BIQA models proposed in this paper emphasize more on exploiting additional information such as EXIF tags, image attributes, and semantic labels to aid quality prediction.

3. SPAQ Database

In this section, we first describe the construction of the proposed SPAQ database for smartphone photography. Next, we present the subjective assessment environment for collecting human annotations, including MOSs (for image quality), image attribute scores, and scene category labels.

3.1. Database Construction

We collect a total of $11,125$ realistically distorted pictures. To support comparison among smartphones, a subset of $1,000$ images are captured by different cameras under a few challenging scenes, including night, low-light, high-dynamic-range, and moving scenes. SPAQ represents a wide range of realistic camera distortions, including sensor noise contamination, out-of-focus blurring, motion blurring, contrast reduction, under-exposure, over-exposure, color shift, and a mixture of multiple distortions above. Sensor noise often occurs in night scenes or indoor scenes with low-light conditions, where high ISO must be applied. Out-of-focus blur can be created deliberately or unintentionally, and it does not necessarily lead to visual quality degradation [34]. Motion blur appears when camera shakes or object moves rapidly in the scene. Global and local contrast may not be fully reproduced for scenes under poor weather conditions or with high dynamic ranges. Color shift may result from incorrect white balance or other computational methods for post-processing. An important fact of smartphone photography is that mixtures of distortions frequently occur, making the images substantially different from those created by synthetic distortions.

The images are initially saved with high resolution, typically in the order of six megapixels and more. The gigantic image size poses a challenge to existing BIQA models, whose computational complexities are generally high. Therefore, we choose to downsample the raw pictures such that the shorter side is 512, and stored them in PNG format. Sample images in SPAQ can be found in Figure 1.

Each image in SPAQ is associated with

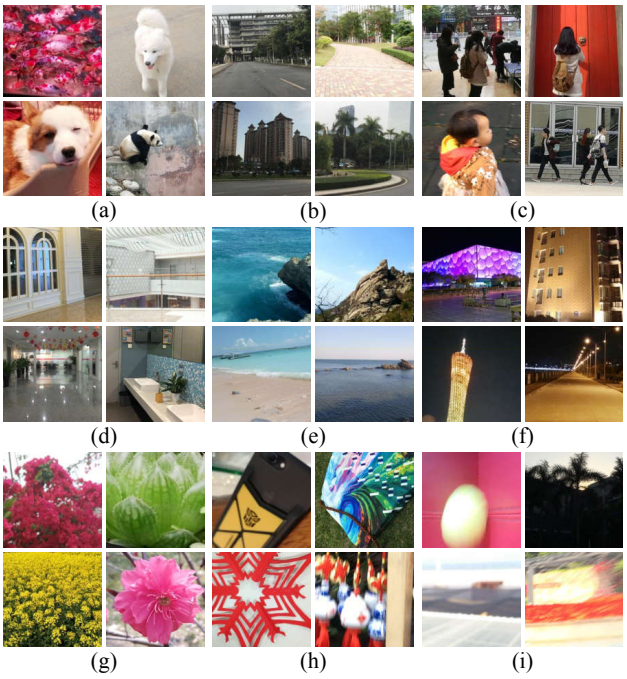
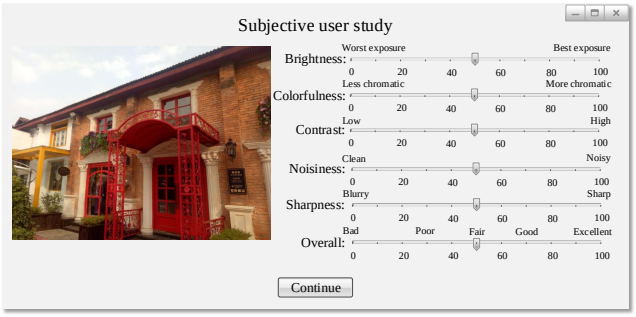


Figure 1. Sample images in SPAQ. (a) Animal. (b) Cityscape. (c) Human. (d) Indoor scene. (e) Landscape. (f) Night scene. (g) Plant (h) Still life. (i) Others. All images are cropped for neat presentation.

- EXIF tags, including 1) focal length, 2) f-number (inversely proportional to aperture size), 3) exposure time, 4) ISO (light sensitivity of sensor), 5) brightness value (brightness of focus point in the scene), 6) flash (flash fired or not), 7) time (when image was recorded). Since the brightness value is not provided by some smartphone manufacturers, we make an educated estimation using the exposure equation [11].
- MOS, a continuous score in $[0, 100]$ to represent the overall quality of the image. A higher score indicates better perceived quality.
- Image attribute scores, including 1) brightness, 2) colorfulness, 3) contrast, 4) noisiness, and 5) sharpness. Similar to MOS, each attribute is represented by a continuous score in $[0, 100]$ (see Figure 2 (a)).
- Scene category labels, including 1) animal, 2) cityscape, 3) human, 4) indoor scene, 5) landscape, 6) night scene, 7) plant, 8) still life, and 9) others. The category of still life refers to images that contain salient static objects (not living things); the category of "others" includes images from which human annotators find difficulty in recognizing the visual content due to abstract nature or extremely poor quality. It is worth noting that one image may have multiple labels (see Figure 2 (b)).



(a)

(b)

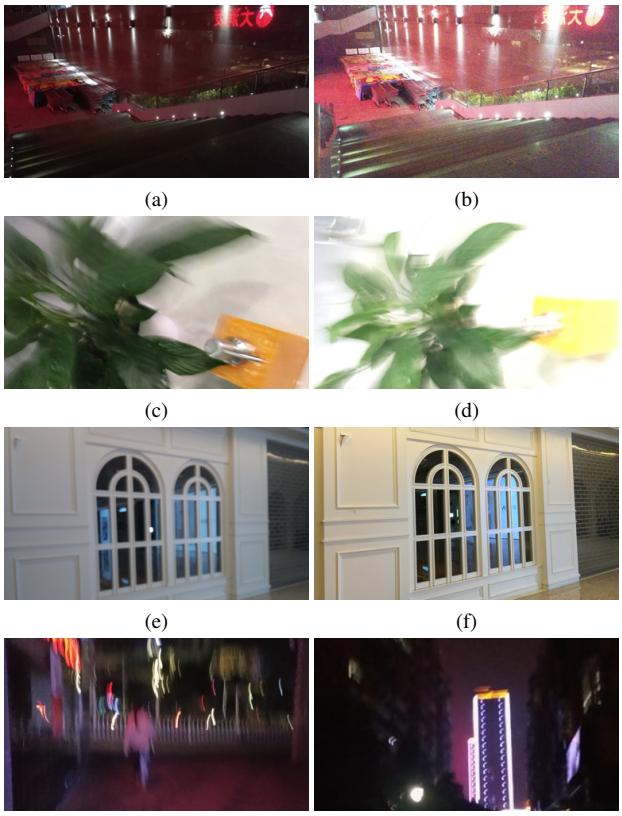
Figure 2. Graphical user interfaces used in our subjective experiments. (a) Quality rating. (b) Scene classification.

3.2. Subjective Testing

MOSs and Image Attribute Scores We invite more than 600 subjects to participate in this subjective test. To obtain consistent and reliable human ratings, the experiment is conducted in a well-controlled laboratory environment. Figure 2 (a) shows the graphical user interface. Subjects are asked to rate the quality of an image on a continuous scale in $[0, 100]$, evenly divided and labeled by five quality levels ("bad", "poor", "fair", "good", and "excellent"). Additionally, we ask the subjects to provide five other continuous scores from 0 to 100, representing the degrees of brightness, colorfulness, contrast, noisiness, and sharpness, respectively.

Scene Category Labels Participants are invited to provide scene category labels for each image in SPAQ using a multi-label method, including animal, cityscape, human, indoor scene, landscape, night scene, plant, still life, and others. The graphical user interface for scene labeling is given in Figure 2 (b), where an image can be labeled by one or more categories.

We refer the interested readers to the supplementary file for a complete description of the subjective experiment regarding the testing environment, the training and testing phases, the outlier removal and the reliability of subjective data.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 3. Sample images in SPAQ. (a) ISO = $1,600$. (b) ISO = $2,000$. (c) exposure time = 0.03s. (d) exposure time = 0.06s. (e) f-number = 2.0. (f) f-number = 2.2. (g) and (h) f-number = 2.2, exposure time = 0.059s, and ISO = $1,757$.

4. Subjective Data Analysis

In this section, we analyze the collected subjective data in SPAQ to reveal the relationships between EXIF tags, image attributes, scene category labels and image quality. We then rank several smartphone cameras based on the subjective results.

4.1. Interactions between Perceptual Image Quality and Various Factors

EXIF Tags To explore the relationship between the EXIF tags and image quality, we present some sample images captured with different camera settings in Figure 3. When playing with ISO, we find that higher ISO numbers yield brighter images but with a significant amount of noise (see Figure 3 (a) and (b)). This shows that ISO is predictable of image quality especially for night scenes. When playing with exposure time, we find that if camera shakes or object is moving fast, motion blurring occurs even for a relatively short exposure (see Figure 3 (c)), and over-exposure also arises if we double the exposure time (see Figure 3 (d)). It

Attribute	Image attribute scores	
	from humans	by MT-A
Brightness	0.784	0.704
Colorfulness	0.844	0.760
Contrast	0.874	0.786
Noisiness	0.893	0.832
Sharpness	0.958	0.904

Table 2. SRCC results between MOSs and image attribute scores from humans and MT-A (our proposed computational model), respectively.

is well-known that different aperture sizes lead to different depths of field. Generally, with a smaller aperture size, the range of distance in focus is larger, and therefore out-of-focus blur is less likely to happen (see Figure 3 (e) and (f)). Finally, the two different visual scenes in Figure 3 (g) and (h) are captured with the same camera setting, where we see that they suffer from a similar combination of distortions, leading to similar perceptual quality. In summary, the EXIF tags convey rich side information that may be helpful for predicting image quality. As will be clear in Section 5.2, computational models that make proper use of EXIF information greatly boost quality prediction performance.

Image Attribute Scores To investigate how each image attribute affects perceived quality, we compute the Spearman's rank correlation coefficient (SRCC) between MOSs and attribute scores, as listed in Table 2. We find that sharpness and noisiness have higher correlations with image quality compared to brightness and colorfulness. This is consistent with the hypothesis that the human eye is highly adapted to extract local structures [37], and is less sensitive to global brightness change.

Scene Category Labels We draw the MOS distribution (discretized to five quality levels) for each scene category in Figure 4, from which we have some interesting observations. First, the MOSs of the "others" category concentrate at low quality levels. This is expected because images in this category are unrecognizable largely due to poor visual quality. Second, images of night scenes generally exhibit poor quality with large under-exposed and noisy regions (see Figure 3 (g) and (h)), emphasizing the challenges for low-light photography. Finally, images with different scene categories have noticeable MOS distributions, suggesting high-level semantic effect on visual perception of image quality.

4.2. Smartphone Camera Comparison

At the beginning, it is important to note that this study is independent of any telecommunication device manufactures or service providers. In SPAQ, a subset of $1,000$ images containing 50 visual scenes are captured by 20 smartphone cameras for performance comparison. Many images are taken under tough conditions to challenge the smartphone photography systems, including low-light, high-dynamic-range, and moving scenes.

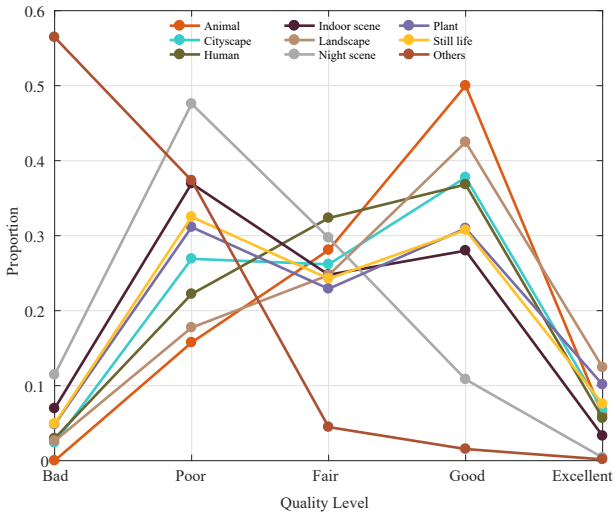


Figure 4. Discretized MOS distributions of images with different scene categories. See Figure 2 (a) for the specification of the five quality levels.

Top-5 cameras		Bottom-5 cameras	
Type	# scenes	Type	# scenes
Apple iPhone 6s Plus	22	Meitu M6	21
Huawei PRA-AL00	19	Vivo X7	17
Oppo A33m	17	Samsung SM-G9006V	16
Oppo R9 Plusm A	16	Xiaomi MI 6	15
Xiaomi MIX 2	15	Apple iPhone SE	14

Table 3. The top and bottom performing smartphone cameras based on image quality.

For the 50 visual scenes, we count the number of top-5 and bottom-5 pictures that belong to each smartphone camera based on image quality. The results are listed in Table 3, where we find that Apple iPhone 6s Plus achieves the best results with 22 scenes at the top-5, while Meitu M6 is at the bottom among 20 smartphone cameras. The success of Apple iPhone 6s plus may result from its stabilization and noise reduction post-processing methods, which help improve the quality of images captured under less ideal conditions. More information about the visual scenes and the smartphone camera ranking samples can be found in the supplementary.

5. Objective Quality Models

Based on the proposed SPAQ database, we train a deep neural network (DNN) to predict the perceptual quality of

smartphone captured images and three variants that make use of EXIF tags, image attributes, and scene category labels, respectively.

5.1. Baseline Model

We adopt ResNet-50 [8] as the backbone to construct our baseline model - BL. We change the final fully connected layer to one output, and drop the softmax function. The parameters of BL are collectively denoted by w_{B}. The training mini-batch consists of $\{x^{(i)}, q^{(i)}\}_{i=1}^{m}$, where $x^{(i)}$ is the i-th input color image and $q^{(i)}$ is the corresponding MOS. We exclude pre-processing that would dramatically alter image quality such as global mean removal and contrast normalization. The output of BL is a scalar $\hat{q}^{(i)}$, representing the predicted quality score of $x^{(i)}$. We replace the cross entropy function for image classification with ℓ_1-norm as the empirical loss

$$\ell_1(w_{\mathrm{B}}) = \|q - \hat{q}\|_1 = \sum_{i=1}^{m} |q^{(i)} - \hat{q}^{(i)}|. \tag{1}$$

In our experiments, we find that fine-tuning ResNet-50 [8] from pre-trained weights performs better than training the network from scratch or starting from other pre-trained network architectures (*e.g.*, AlexNet [15] and VGG16 [31]), which is consistent with the observations in [14].

5.2. Multi-Task Learning from EXIF Tags

We train a variant of the baseline model, namely MT-E, by incorporating EXIF data using multi-task learning. Specifically, each image has seven EXIF tags, among which focal length, f-number and ISO are categorical variables, exposure time and brightness are continuous, flash is binary, and time (image was recorded) is periodic. The input mini-batch samples are formed as $\{x^{(i)}, o^{(i)}, q^{(i)}\}_{i=1}^{m}$, where $o^{(i)}$ is a feature vector containing the encoded EXIF tags of $x^{(i)}$. MT-E consists of two sub-networks. The first sub-network is the same as BL, which takes $x^{(i)}$ as input and regresses a "generic" quality score $\hat{g}^{(i)}$. The second sub-network comprises a simple fully connected layer, which accepts $o^{(i)}$ and produces an offset $\hat{b}^{(i)}$ [26], with parameters denoted by w_{E}. The final quality prediction is computed by

$$\hat{q}^{(i)} = \hat{g}^{(i)} + \hat{b}^{(i)}, \tag{2}$$

where we interpret $\hat{b}^{(i)}$ as a learned bias added to the generic score. When EXIF data are not present as in the case of many Internet images, MT-E reduces gracefully to BL. We train MT-E by optimizing a naïve weighted sum of two ℓ_1-norm losses

$$\ell_2(w_{\mathrm{B}}, w_{\mathrm{E}}) = \alpha_1 \|q - \hat{g}\|_1 + \alpha_2 \|q - \hat{q}\|_1, \tag{3}$$

where α_1 and α_2 are non-negative task weightings, satisfying $\alpha_1 + \alpha_2 = 1$. Since $\hat{g}$ and $\hat{q}$ have the same measurement scale, we simply set $\alpha_1 = \alpha_2 = 0.5$.

5.3. Multi-Task Learning from Image Attributes

Besides subjective quality ratings, we also collect five image attribute scores, including brightness, colorfulness, contrast, noisiness, and sharpness. To explore the influence of image attributes on image quality, we extend BL to MT-A by learning to predict image attributes jointly. Built upon the baseline model, we let the final fully connected layer output six scalars, representing the overall image quality and the degrees of image attributes, respectively. That is, the six tasks share the computation up to the last fully connected layer. The parameters for estimating image attributes are represented by w_{A}. We denote the input mini-batch by $\{x^{(i)}, r^{(i)}, q^{(i)}\}_{i=1}^{m}$, where $r^{(i)}$ is a five-dimensional vector that stores the ground truth image attribute scores of $x^{(i)}$. Similarly, we use a naïve weighted sum of six ℓ_1-norm losses to train MT-A

$$\ell_3(w_{\mathrm{B}}, w_{\mathrm{A}}) = \beta_1 \|q - \hat{q}\|_1 + \frac{\beta_2}{5} \sum_{j=1}^{5} \|r_j - \hat{r}_j\|_1, \tag{4}$$

where $\hat{r}_j$ is an m-dimensional vector that stores the j-th image attribute predictions of the current mini-batch. β_1 and β_2 are non-negative task weightings, satisfying $\beta_1 + \beta_2 = 1$. By default we give the five attribute prediction tasks the same weight. According to our subjective experiment, all tasks are measured in the same scale. We take advantage of this fact and sample β_1 linearly from $[0, 1]$.

5.4. Multi-Task Learning from Scene Labels

To explore the effectiveness of incorporating semantic information into quality prediction, we train another model, namely MT-S, using multi-task learning. Conceptually, scene classification and quality assessment appear to be competing tasks - the former requires feature representations to be insensitive to image quality degradation, while the latter desires the opposite. To address this problem, MT-S splits BL into two sub-networks carefully for the two tasks. The input mini-batch samples are denoted by $\{x^{(i)}, p^{(i)}, q^{(i)}\}_{i=1}^{m}$, where $p^{(i)}$ is a nine-dimensional vector with c non-zero entries, each set to $1/c$ corresponding to the $c \geq 1$ scene labels for $x^{(i)}$. For scene classification, we let the last fully connected layer to produce nine continuous activations $\hat{s}^{(i)}$, followed by softmax nonlinearity to convert them into probabilities $\hat{p}^{(i)}$. The cross entropy function is then used as the loss

$$\ell_4(w_{\mathrm{S}}) = -\sum_{i,j} p_j^{(i)} \log \hat{p}_j^{(i)}, \tag{5}$$

where w_{S} denotes the parameters associated with the scene classification task. For quality regression, we use Eq. (1) as the empirical loss.

It remains to combine the two losses for joint learning, which is nontrivial as they live in substantially different

scales. Grid-based manual tuning for a reasonable weighting is expensive, especially in the context of deep learning. Inspired by [10], we choose to learn the optimal weighting as task-dependent uncertainty. In regression, we define the likelihood function as a Laplace distribution with mean given by the network output and an observation noise scalar σ_1:

$$\hat{p}(q^{(i)}|w_{\mathrm{B}}) \sim \mathrm{Laplace}(\hat{q}^{(i)}, \sigma_1). \tag{6}$$

In classification, we define the likelihood as a scaled version of the model output $\hat{s}^{(i)}$ through a softmax function [10]

$$\hat{p}(y^{(i)}|w_{\mathrm{S}}) \sim \mathrm{Softmax}\left(\frac{1}{\sigma_2}\hat{s}^{(i)}\right), \tag{7}$$

where σ_2 is a positive scalar, governing how uniform the induced discrete distribution is and $y^{(i)} \in \{1, \ldots, 9\}$. It is straightforward to show that the expected negative log likelihood as the joint loss function can be approximated by

$$\ell_5(w_{\mathrm{B}}, w_{\mathrm{S}}) = \frac{\ell_1(w_{\mathrm{B}})}{\sigma_1} + \frac{\ell_4(w_{\mathrm{S}})}{\sigma_2} + m\log\sigma_1 + \frac{m}{2}\log\sigma_2. \tag{8}$$

The above loss discourages high task uncertainty through the two log terms. MT-S can learn to ignore noisy tasks, but is penalized for that [10]. Eq (8) also discourages very low task uncertainty. For example, a low σ_1 will exaggerate the contribution of ℓ_1. σ_1 and σ_2 are estimated along with the model parameters $\{w_{\mathrm{B}}, w_{\mathrm{S}}\}$.

5.5. Performance Evaluation

For the baseline model BL and its variants, we adopt the same training strategy, repeat the training processes for five times, and report the average results to reduce any bias introduced during training. Specifically, we randomly sample 80% of the images in SPAQ for training and leave the rest for testing. The backbone ResNet-50 [8] is initialized with the pre-trained weights for object recognition on ImageNet [28]. We set the mini-batch size to 16 and the epoch number to 30. We use the Adam stochastic optimization package [13] with the initial learning rate of 10^{-3} and a decay factor of 0.1 for every 10 epochs. The input images are randomly cropped to $224 \times 224 \times 3$. For the first 10 epochs, we only train the final fully connected layers by freezing the rest parameters in the networks. For the next 20 epochs, we fine-tune the whole networks by optimizing the respective losses.

During testing, we crop $224 \times 224 \times 3$ patches from a test image with a stride of 112. The final quality and attribute scores are computed by averaging all patch predictions. The dominant scene class is determined by majority vote among all top-1 predictions, which is considered correct if it matches one of multiple ground truth labels.

We compare the proposed methods with seven existing BIQA models, including BRISQUE [22], DIIVINE [24], CORNIA [39], QAC [38], ILNIQE [40], FRIQUEE [6], and DB-CNN [41]. These cover a wide range of design philosophies, including NSS-based [6, 22, 24, 40], codebook-based [38, 39], and DNN-based [41] models. The implementations of the competing models are obtained from the respective authors. We re-train BRISQUE, FRIQUEE, and DB-CNN using the same training set. As for DIIVINE and CORNIA, we directly use the learned models due to the lack of publicly available training codes and the complexity of reproducing the training procedures. Note that QAC [38] and ILNIQE [40] do not require MOSs for training.

Experimental results are shown in Table 4, from which we have several interesting observations. First, BIQA models designed for synthetic distortions (*e.g.*, QAC [38] and DIIVINE [24]) generally do not work well for realistic camera distortions, which is no surprise because there is a significant discrepancy between the two data distributions. Second, verified on the LIVE Challenge Database [5], FRIQUEE [6] delivers superior performance on SPAQ, which verifies the effectiveness of the handcrafted features at capturing the characteristics of realistic distortions. Third, BRISQUE [22] also obtains comparable performance, suggesting that the locally normalized pixel intensities may reveal useful attributes of realistic distortions. Fourth, by bilinearly pooling two sets of features, DB-CNN [41] outperforms all BIQA approaches, including the proposed BL based on ResNet-50. This suggests that DNNs successfully learn hierarchical features sensitive to realistic distortions, and that a more advanced backbone (such as DB-CNN) offers additional performance gains. Finally, the performance of the proposed baseline and its variants are among the best, verifying our training and multi-task learning strategies.

Now, we take a close look at the multi-task learning results. When we add EXIF tags as additional inputs, MT-E achieves a significant improvement compared with BL. This emphasizes the importance of EXIF data to quality prediction of smartphone captured images, which, however, has not been paid much attention by the IQA community. Next, jointly predicting image attributes positively impacts the accuracy of quality prediction, and the results are robust to different task weightings (see Table 5). In addition, the predicted attribute scores by MT-A have high correlations with MOSs, as shown in the second column of Table 2. This indicates that the five image attributes play key roles in determining image quality, and the MT-A model has learned the inherent relationships between image attributes and the overall quality. Lastly and perhaps more interestingly, we observe improved performance of quality prediction in MT-S when jointly trained with scene classification. From Table 6, we find that the approximately op-

Model	QAC [38]	DIIVINE [24]	CORNIA [39]	ILNIQE [40]	BRISQUE [22]	FRIQUEE [6]	DB-CNN [41]	BL	MT-E	MT-A	MT-S
SRCC	0.092	0.599	0.709	0.713	0.809	0.819	0.911	0.908	0.926	0.916	0.917
PLCC	0.497	0.600	0.725	0.721	0.817	0.830	0.915	0.909	0.932	0.916	0.921

Table 4. Average SRCC and PLCC results of our methods across five sessions against seven BIQA models on SPAQ.

Task weights β_1	Task weights β_2	SRCC	PLCC
0.9	0.1	0.917	0.919
0.8	0.2	0.917	0.918
0.7	0.3	0.917	0.918
0.6	0.4	0.916	0.917
0.5	**0.5**	0.916	0.916

Table 5. Average SRCC and PLCC results of MT-A across five sessions as a function of task weighting. The default setting is highlighted in bold.

Splitting position	SRCC	PLCC	Accuracy
Conv1	0.915	0.918	0.702
Conv10	0.916	0.918	0.687
Conv22	0.916	0.919	0.673
Conv40	0.917	0.921	0.673
Conv49	0.915	0.916	0.670

Table 6. Average SRCC, PLCC and accuracy results of MT-S across five sessions as a function of splitting positions. Conv# indicates the splitting happens right after the #-th convolution layer and there are a total of 49 convolution layers in MT-S.

timal splitting for quality regression is at the 40-th convolution layer. However, the scene classification task prefers to split at the first convolution layer (measured by classification accuracy), which suggests two separate networks without sharing weights. This provides indirect evidence that the two tasks compete with each other. Nevertheless, MT-S is able to exploit semantic information to boost the quality prediction performance. These insightful findings inspire further research on how to extract semantic information (*e.g.*, in the form of dense semantic segmentation maps) and how to incorporate it into IQA, with the goal of benefiting both tasks.

6. Conclusion

We build so far the most comprehensive database for perceptual quality assessment of smartphone photography, where each image is attached with rich annotations, including not only quality ratings (in the form of MOSs), but also a series of EXIF, attribute, and semantic information. In SPAQ, $1,000$ images are captured repeatedly by different smartphones of the same scenes, facilitating head-to-head comparisons of smartphone cameras.

We also construct four BIQA models using DNNs to exploit the influence of EXIF tags, image attributes, and high-level semantics on perceived quality of smartphone pictures. Our results suggest that all such side information may be useful in improving prediction accuracy of the BIQA models. We believe that the current database, together with the proposed DNN-based computational models, lay the groundwork for the development of next-generation BIQA methods for smartphone photography, which in turn will impact the future design of smartphone cameras and the integrated computational photography systems.

Acknowledgments

This work was supported in part by the NSFC under Grant 61822109, the National Key R&D Program of China under Grant 2018AAA0100601, and the CityU Start-up Grant (No. 7200630). The authors would like to thank Chenyang Le for fruitful discussions on subjective testing and Weixia Zhang for insightful comments on model implementation. We thank the NVIDIA Corporation for donating a GPU for this research.

References

[1] S. Athar and Z. Wang. A comprehensive performance evaluation of image quality assessment algorithms. *IEEE Access*, 7:140030–140070, Sep. 2019. 1

[2] A. Ciancio, C. A. Da, S. E. Da, A. Said, R. Samadani, and P. Obrador. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on Image Processing*, 20(1):64–75, Jan. 2010. 2

[3] S. J. Daly. Visible differences predictor: An algorithm for the assessment of image fidelity. In *SPIE/IS&T Symposium on Electronic Imaging: Science and Technology*, pages 2–15, 1992. 1

[4] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, and G. Zhai. No-reference quality assessment of contrast-distorted images based on natural scene statistics. *IEEE Signal Processing Letters*, 22(7):838–842, Nov. 2014. 3

[5] D. Ghadiyaram and A. C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, Jan. 2016. 2, 3, 7

[6] D. Ghadiyaram and A. C. Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of vision*, 17(1):32–32, Jan. 2017. 3, 7, 8

[7] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics*, 35(6):192:1–192:12, Nov. 2016. 1

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer vison and Pattern Recognition*, pages 770–778, 2016. 6, 7

[9] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *CoRR*, abs/1910.06180, 2019. 2

[10] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018. 2, 7

[11] D. A. Kerr. APEX - The additive system of photographic exposure. 2017, [Online] Available: http://dougkerr.net/Pumpkin/index.htm#APEX. 3

[12] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Processing Magazine*, 34(6):130–141, Nov. 2017. 1

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, pages 1–15, 2015. 7

[14] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, pages 662–679, 2016. 6

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Nerual Information Processing Systems*, pages 1097–1105, 2012. 6

[16] E. C. Larson and D. M. Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *SPIE Journal of Electronic Imaging*, 19(1):1–21, Jan. 2010. 2

[17] W. Lin and C.-C. J. Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, May 2011. 1

[18] K. Ma, Z. Duanmu, Z. Wang, Q. Wu, W. Liu, H. Yong, H. Li, and L. Zhang. Group maximum differentiation competition: Model comparison with few samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear, 2019. 2

[19] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang. Waterloo Exploration Database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 22(2):1004–1016, Feb. 2017. 2

[20] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao. dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8):3951–3964, Aug. 2017. 3

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

[22] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, Dec. 2012. 3, 7, 8

[23] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letter*, 20(3):209–212, Mar. 2013. 3

[24] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, Dec. 2011. 7, 8

[25] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, Jan. 2015. 2

[26] J. Ren, X. Shen, Z. Lin, R. Mech, and D. J. Foran. Personalized image aesthetics. In *IEEE International Conference on Computer Vision*, pages 638–647, 2017. 6

[27] A. H. Renear, S. Sacchi, and K. M. Wickett. Definitions of dataset in the scientific and technical literature. In *Annual Meeting of the American Society for Information Science and Technology*, pages 1–4, 2010. 2

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F.-F. Li. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec. 2015. 7

[29] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, Feb. 2006. 2

[30] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, Nov. 2006. 1, 2

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 6

[32] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, Nov. 2016. 2

[33] V. Toni, N. Mikko, V. Mikko, O. Pirkko, and H. Jukka. CID2013: A database for evaluating no-reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 24(1):390–402, Jan. 2015. 2

[34] N. Wadhwa, R. Garg, D. E. Jacobs, B. E. Feldman, N. Kanazawa, R. Carroll, Y. Movshovitz-Attias, J. T. Barron, Y. Pritch, and M. Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics*, 37(4):64:1–64:13, Jul. 2018. 3

[35] Z. Wang and A. C. Bovik. *Modern Image Quality Assessment*. San Rafael, CA, USA: Morgan Claypool Publishers, 2006. 1

[36] Z. Wang and A. C. Bovik. Reduced- and no-reference image quality assessment: The natural scene statistic model approach. *IEEE Signal Processing Magazine*, 28(6):29–40, Nov. 2011. 3

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, Apr. 2004. 1, 2, 5

[38] W. Xue, L. Zhang, and X. Mou. Learning without human scores for blind image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 995–1002, 2013. 7, 8

[39] P. Ye, J. Kumar, L. Kang, and D. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1098–1105, 2012. 3, 7, 8

[40] L. Zhang, L. Zhang, and A. C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, Aug. 2015. 7, 8

[41] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. 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, Jan. 2020. 3, 7, 8

[42] W. Zhang, K. Ma, G. Zhai, and X. Yang. Learning to blindly assess image quality in the laboratory and wild. *CoRR*, abs/1907.00516, 2019. 3

Supplementary Materials to "Perceptual Quality Assessment of Smartphone Photography"

Yuming Fang[1]*, Hanwei Zhu[1]*, Yan Zeng[1], Kede Ma[2]†, and Zhou Wang[3]

[1]Jiangxi University of Finance and Economics, [2]City University of Hong Kong,
[3]University of Waterloo

In the supplementary file, we first present in detail the procedures for annotating images in the Smartphone Photography Attribute and Quality (SPAQ) database. Next, we describe the strategies for outlier detection and subject removal, and discuss the reliability of the collected subjective data. Finally, we provide more details and validation results of the proposed blind image quality assessment (BIQA) models based on SPAQ.

1. More about SPAQ

As stated in the manuscript, SPAQ collects so far the richest annotations for each image, including image quality, image attributes, and scene category labels in a well-controlled laboratory environment. In the following, we present more details about SPAQ, including database construction and subjective testing.

1.1. Database Construction

We use 66 smartphones from eleven manufacturers to construct SPAQ of $11,125$ realistically distorted images (see Table S1). A subset of $3,453$ pictures of the same visual scenes are capture with controlled scene configurations and camera settings. As mentioned in Section 4.4, we select $1,000$ images from this subset to rank smartphone cameras. Figure S1 shows 20 images captured by different smartphone cameras with the out-of-focus configuration.

1.2. Subjective Testing

1.2.1 Testing Environment

To obtain reliable human annotations for both quality rating and scene classification, we conduct subjective experiments in a well-controlled laboratory environment using five LCD monitors at a resolution of 1920×1080 pixels, which are calibrated in accordance with ITU-T BT.500 recommendations [S1]. The ambient illumination does not directly reflect off the displays. Each participant requires normal or corrected-to-normal visual acuity with correct color vision. Participants are allowed to move their positions to get closer or further away from the screen for comfortable viewing experience. In our subjective experiment, the male to female ratio is about $3:2$, and their ages are between 18 and 35.

1.2.2 Image Quality

Before the subjective experiment, each participant goes through a training phase, where ten images independent of the testing phase are displayed. Nine of them are provided with reference ratings and detailed instructions. Participants are asked to read the instructions carefully and reminded to focus on image quality rather than image aesthetics. The tenth image without any instruction should be rated by participants as practice. The subjective scores in the training phase are not recorded.

During the testing phase, each participant provides subjective ratings for 80 images in one session, and is involved in at most two sessions with a five-minute break in-between. The 80 images in each session are composed of two parts: the first part includes 75 images selected randomly from $11,125$ images; the second part includes five duplicated images selected randomly from the 75 images in the first part. Eventually, we finish with $2,330$ sessions, and collect $186,400$ subjective ratings in total. Each image is rated at least 15 times.

1.2.3 Image Attributes

Besides image quality, we ask participants to provide five continuous image attribute scores from 0 to 100, representing the degrees of brightness, colorfulness, contrast, noisiness, and sharpness. A low brightness score indicates that the image is poorly exposed. An image with more chro-

*Equal contribution
†Corresponding author (email:kede.ma@cityu.edu.hk)

Manufacturer	Huawei	Apple	Vivo	Oppo	Xiaomi	Nubia	Meitu	Samsung	Meizu	Gionee	Letv
# cameras	22	8	10	9	8	3	5	7	1	1	1
# images	3,086	2,063	1,369	1,127	1,083	882	668	620	149	63	15

Table S1. The number of images by different smartphones from different manufacturers.

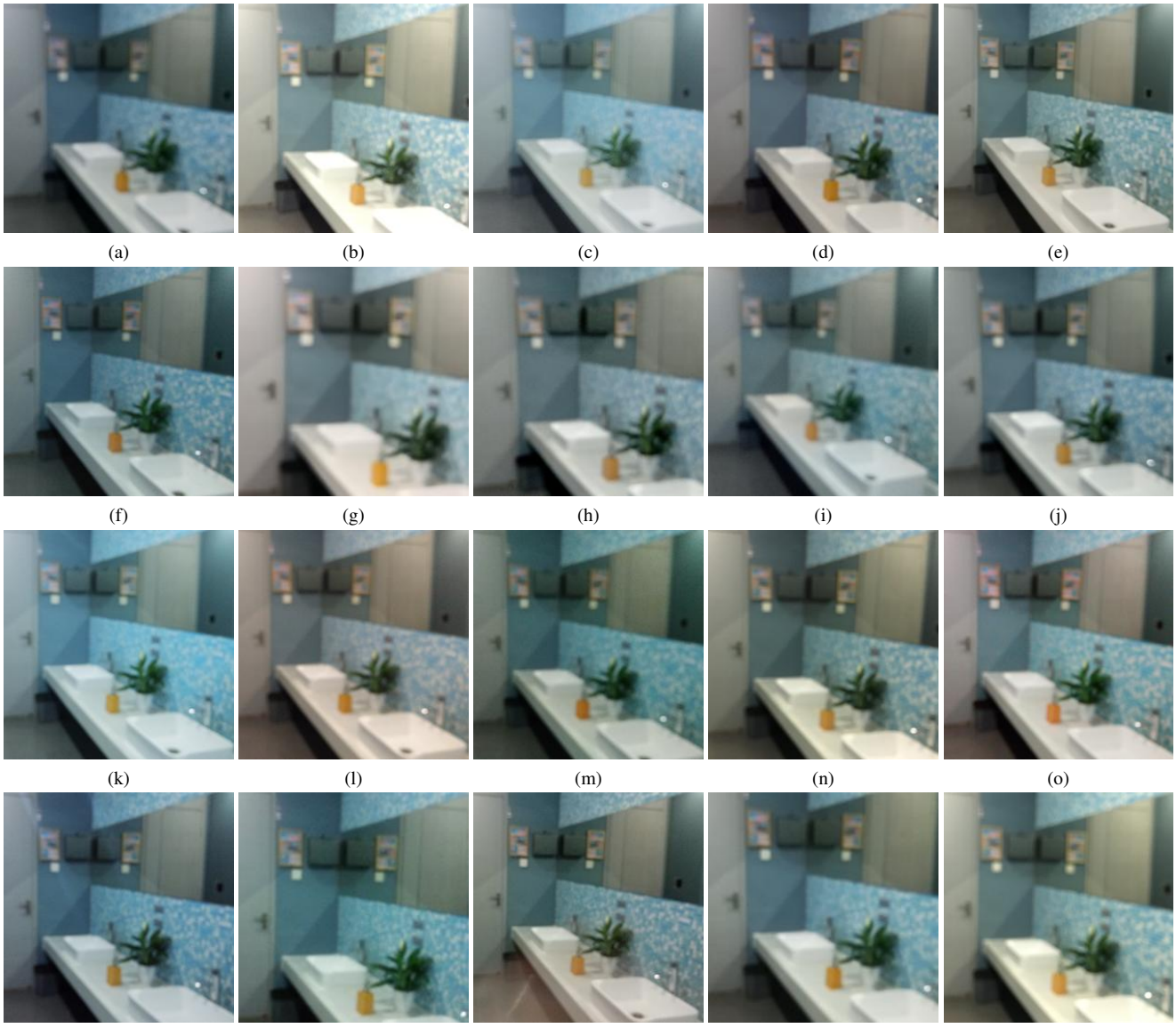


Figure S1. Sample images captured by 20 smartphone cameras with the out-of-focus configuration. (a) Meitu V6, MOS = 34.62. (b) Apple iPhone 6, MOS = 33.09. (c) Meitu M6, MOS = 34.25. (d) Samsung SM-G9200, MOS = 40.67. (e) Xiaomi MIX 2, MOS = 41.67. (f) Oppo A33m, MOS = 39.00. (g) Huawei BLA-AL00, MOS = 40.25. (h) Oppo R9 Plusm A, MOS = 38.00. (i) Meizu M5 Note, MOS = 27.89. (j) Oppo R9s, MOS = 27.44. (k) Meitu T8, MOS = 40.00. (l) Huawei MLA-AL10, MOS = 32.44. (m) Vivo X7, MOS = 24.25. (n) Apple iPhone SE, MOS = 39.43. (o) Huawei TAG-TL00, MOS = 35.33. (p) Meitu M4, MOS = 27.29. (q) Samsung SM-G9006V, MOS = 21.43. (r) Xiaomi MI 6, MOS = 34.71. (s) Huawei PRA-AL00, MOS = 28.00. (t) Apple iPhone 6s Plus, MOS = 33.33.

matic information is given a higher attribute score for colorfulness. An image with reduced contrast is rated with a low contrast score. An image containing a great amount of sensor noise leads to a high noisiness score. The attribute score for sharpness is inversely proportional to the blur level, suggesting that a blurry image should be rated with a low score.

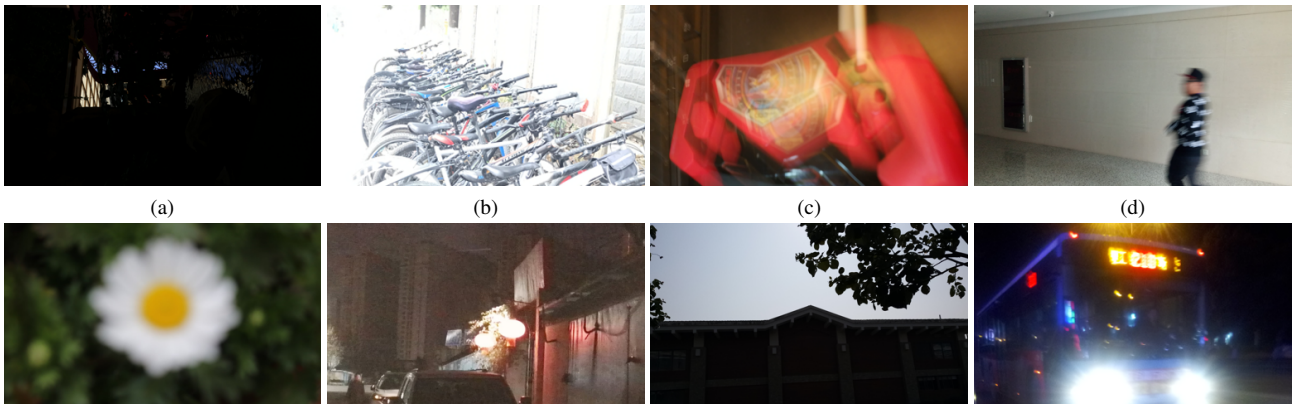


Figure S2. Sample images of typical realistic camera distortions in SPAQ. (a) Under-exposure. (b) Over-exposure. (c) Camera motion blurring. (d) Object motion blurring. (e) Out-of-focus blurring. (f) Sensor noise. (g) Contrast reduction. (h) Mixture of multiple distortions.

1.2.4 Scene Categories

In order to exploit the relationship between scene semantics and image quality, we classify an image into nine scene categories, including animal, cityscape, human, indoor scene, landscape, night scene, plant, still life, and others. Each image may be associate with multiple scene category labels. For example, the image in Figure 2 (b) is labeled with animal and human categories for its content: "humans are playing with a dog". During subjective testing, we remind the subjects to pay attention to foreground objects for scene classification. Five subjects experienced in computer vision annotate the whole $11,125$ images. When there is disagreement between human annotations, majority vote is used to determine the final label.

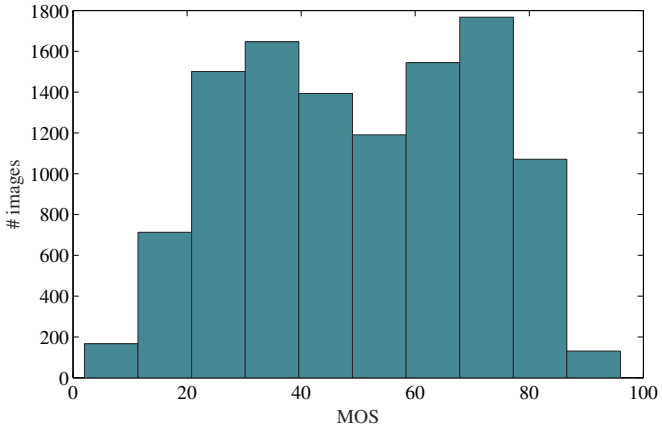


Figure S3. The histogram of MOSs in SPAQ.

2. More about Subjective Data Analysis

2.1. Outlier Detection

We process our raw subjective data by detecting and removing outlier annotations. First, based on the outlier rejection method in [S1], a valid subjective score for each image should be in the range of $[\mu - n\sigma, \mu + n\sigma]$, where μ and σ denote the mean and standard deviation of the subjective scores, respectively. Generally, n is set to 2 if the empirical distribution is Gaussian; otherwise, n is set to $\sqrt{20}$. We use this strategy to check the MOSs and attribute scores from each participant in each session. If there are more than eight subjective scores (*i.e.*, 10%) of the overall quality that are out of the expected range, the subject (in this session) is considered as an outlier, and all subjective scores are subsequently removed.

We also conduct outlier removal based on MOSs of the five duplicated images that are rated twice in each session. We compute the difference between these two MOSs for each duplicated image. If the difference is larger than 20, the subjective scores for this image are invalid. If there are over three images with invalid scores from a participant, we remove all ratings by the participant in this session.

In total, we collect $186,400$ raw human ratings from $2,330$ sessions for all of $11,125$ images, and $18,646$ ratings are detected as outliers. Figure S3 shows the histogram of MOSs of the images in SPAQ.

2.2. Reliability of Subjective Data

Consistency across Sub-Groups We calculate the cross-group consistency using correlations between MOSs from two sub-groups of participants. Specifically, we randomly divide the participants into two equal-size sub-groups, and compute two MOSs for each image in SPAQ from the two sub-groups. We repeat this random splitting 25 times, and report the mean SRCC and PLCC results in Table S2, where we find high cross-group consistency.

Criterion	SRCC	PLCC
Consistency across sub-groups	0.923	0.930
Consistency across subjects	0.841	0.865

Table S2. Subjective data consistency analysis. Consistency across sub-groups: correlations between MOSs from two sub-groups of participants. Consistency across subjects: correlations between ratings from individual participants and MOSs from all participants.

Consistency across Subjects We compute the cross-subject consistency using correlations between ratings from individual participants and MOSs from all participants. The mean SRCC and PLCC results are listed in Table S2, from which we observe that the cross-subject consistency is reasonably high, but not as high as cross-group consistency, suggesting that the variation between individual subjects is larger than that between sub-groups of subjects.

3. More about Proposed BIQA Models

3.1. Model Specification

We use ResNet-50 as the backbone for our BIQA models. Table S3 presents the details of the baseline model (BL) and variants of deep multi-task learning models (MT-A, MT-E, and MT-S).

3.2. Multi-Task Loss for MT-S

In this subsection, we derive the multi-task loss function for both quality regression and scene classification tasks. First, in Eq (6) of the manuscript, the Laplace distribution is given by:

$$\hat{p}(q^{(i)}|w_{\mathrm{B}}) = \frac{1}{2\sigma_1}\exp\left(-\frac{|q^{(i)}-\hat{q}^{(i)}|}{\sigma_1}\right), \tag{1}$$

where $\hat{q}^{(i)}$ is the mean given by the network predication, and σ_1 denotes the observation noise. We then compute the negative log likelihood of a mini-batch containing m training samples to construct the quality regression loss

$$\begin{aligned} L(w_{\mathrm{B}}) &= -\log\prod_{i=1}^{m}\hat{p}(q^{(i)}|w_{\mathrm{B}}) \\ &= -\log\left(\frac{1}{2\sigma_1}\right)^m \exp\left(\frac{-\sum_{i=1}^{m}|q^{(i)}-\hat{q}^{(i)}|}{\sigma_1}\right) \\ &= \frac{1}{\sigma_1}\|q-\hat{q}\|_1 + m\log 2\sigma_1 \\ &\propto \frac{\ell_1(w_{\mathrm{B}})}{\sigma_1} + m\log\sigma_1, \end{aligned} \tag{2}$$

where we drop the constant $m\log 2$. Meanwhile, the log likelihood for the output of scene classification can be written as

$$\begin{aligned} \log\hat{p}(y^{(i)}=j|w_{\mathrm{S}}) =& \mathrm{Softmax}\left(\frac{1}{\sigma_2}\hat{s}_j^{(i)}\right) \\ =& \frac{1}{\sigma_2}\hat{s}_j^{(i)} - \log\sum_k \exp\left(\frac{1}{\sigma_2}\hat{s}_k^{(i)}\right) \\ =& \frac{1}{\sigma_2}\left(\hat{s}_j^{(i)} - \log\sum_k \exp(\hat{s}_k^{(i)})\right) \\ & - \log\frac{\sum_k \exp\left(\frac{1}{\sigma_2}\hat{s}_k^{(i)}\right)}{\left(\sum_k \exp(\hat{s}_k^{(i)})\right)^{\frac{1}{\sigma_2}}} \\ \approx& \frac{1}{\sigma_2}\mathrm{Softmax}(\hat{s}_j^{(i)}) - \frac{1}{2}\log\sigma_2, \end{aligned} \tag{3}$$

where as in [S2] we introduce the assumption that $\frac{1}{\sqrt{\sigma_2}}\sum_k \exp\left(\frac{1}{\sigma_2}\hat{s}_k^{(i)}\right) \approx \left(\sum_k \exp(\hat{s}_k^{(i)})\right)^{\frac{1}{\sigma_2}}$ with $\hat{s}_k^{(i)}$ being the k-th entry of $\hat{s}^{(i)}$. Therefore, the empirical loss for scene classification over a mini-batch of m samples is formulated as

$$\begin{aligned} L(w_{\mathrm{S}}) &= -\sum_{i,j} p_j^{(i)}\log\hat{p}(y^{(i)}=j|w_{\mathrm{S}}) \\ &= -\sum_{i,j} p_j^{(i)}\left(\frac{1}{\sigma_2}\mathrm{Softmax}(\hat{s}_j^{(i)}) - \frac{1}{2}\log\sigma_2\right) \\ &= \frac{\ell_4(w_{\mathrm{S}})}{\sigma_2} + \frac{m}{2}\log\sigma_2. \end{aligned} \tag{4}$$

Final, the complete loss can be computed by the joint negative log likelihood:

$$\begin{aligned} \ell_5(w_{\mathrm{B}}, w_{\mathrm{S}}) &= -\log\prod_i\left(\hat{p}(q^{(i)}|w_{\mathrm{B}})\prod_j \hat{p}(y^{(i)}=j|w_{\mathrm{S}})^{p_j^{(i)}}\right) \\ &= L(w_{\mathrm{B}}) + L(w_{\mathrm{S}}) \\ &= \frac{\ell_1(w_{\mathrm{B}})}{\sigma_1} + \frac{\ell_4(w_{\mathrm{S}})}{\sigma_2} + m\log\sigma_1 + \frac{m}{2}\log\sigma_2, \end{aligned} \tag{5}$$

as desired.

3.3. Cross-Database Validation

In order to verify the robustness of the proposed BIQA model, we evaluate BL in a cross-database setting. We train the BL model on SPAQ and test it on two synthetic distortion databases (LIVE [30] and TID2013 [25]) and three realistic distortion databases (CID2013 [33], LIVE Challenge [5], and KonIQ-10k [9]). We show the SRCC and PLCC results in Table S4, where we find that BL is much easier to generalize to other databases of realistic distortions,

but does not work well on synthetic databases. This suggests a significant domain gap between synthetic and realistic distortions. Therefore, it is necessary to build a realistic database of camera pictures such as SPAQ to lay the groundwork for the next-generation BIQA models for smartphone photography.

References

[S1] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment, 2000. 1, 3

[S2] A. Kendall, Geometry and Uncertainty in Deep Learning for Computer Vision, Ph.D. dissertation, Department of Engineering, University of Cambridge, 2017. 4

Layer name	BL	MT-A	MT-E		MT-S	
Conv1	$7 \times 7, 64$, stride 2				$7 \times 7, 64$, stride 2	
	3×3 MaxPool2d(), stride 2				3×3 MaxPool2d(), stride 2	
Conv2_x	$\begin{bmatrix} 1 \times 1 & 64 \\ 3 \times 3 & 64 \\ 1 \times 1 & 256 \end{bmatrix} \times 3$				$\begin{bmatrix} 1 \times 1 & 64 \\ 3 \times 3 & 64 \\ 1 \times 1 & 256 \end{bmatrix} \times 3$	
Conv3_x	$\begin{bmatrix} 1 \times 1 & 128 \\ 3 \times 3 & 128 \\ 1 \times 1 & 512 \end{bmatrix} \times 4$				$\begin{bmatrix} 1 \times 1 & 128 \\ 3 \times 3 & 128 \\ 1 \times 1 & 512 \end{bmatrix} \times 4$	
Conv4_x	$\begin{bmatrix} 1 \times 1 & 256 \\ 3 \times 3 & 256 \\ 1 \times 1 & 1024 \end{bmatrix} \times 6$				$\begin{bmatrix} 1 \times 1 & 256 \\ 3 \times 3 & 256 \\ 1 \times 1 & 1024 \end{bmatrix} \times 6$	
Conv5_x	$\begin{bmatrix} 1 \times 1 & 512 \\ 3 \times 3 & 512 \\ 1 \times 1 & 2048 \end{bmatrix} \times 3$				$\begin{bmatrix} 1 \times 1 & 512 \\ 3 \times 3 & 512 \\ 1 \times 1 & 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1 & 512 \\ 3 \times 3 & 512 \\ 1 \times 1 & 2048 \end{bmatrix} \times 3$
	AdaptiveAvgPool2d()			8-d (EXIF)	AdaptiveAvgPool2d()	
FC	1-d	6-d	1-d (Generic)	1-d (Bias) + Generic	9-d	1-d
GT	MOS	Image attributes and MOS	MOS	MOS	Scene labels	MOS

Table S3. The network architectures of our BIQA models. We follow the style and convention of ResNet-50 in [8], and the "bottleneck" building blocks are shown in brackets with the number of blocks stacked. FC denotes fully connected layer. GT denotes ground truth annotation.

Training	SPAQ				
Testing	Synthetic database		Realistic database		
	LIVE [30]	TID2013 [25]	CID2013 [33]	LIVE Challenge [5]	KonIQ-10k [9]
SRCC	0.560	0.397	0.754	0.742	0.707
PLCC	0.608	0.570	0.771	0.773	0.745

Table S4. SRCC and PLCC results of the proposed BL model in a cross-database setting.