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.

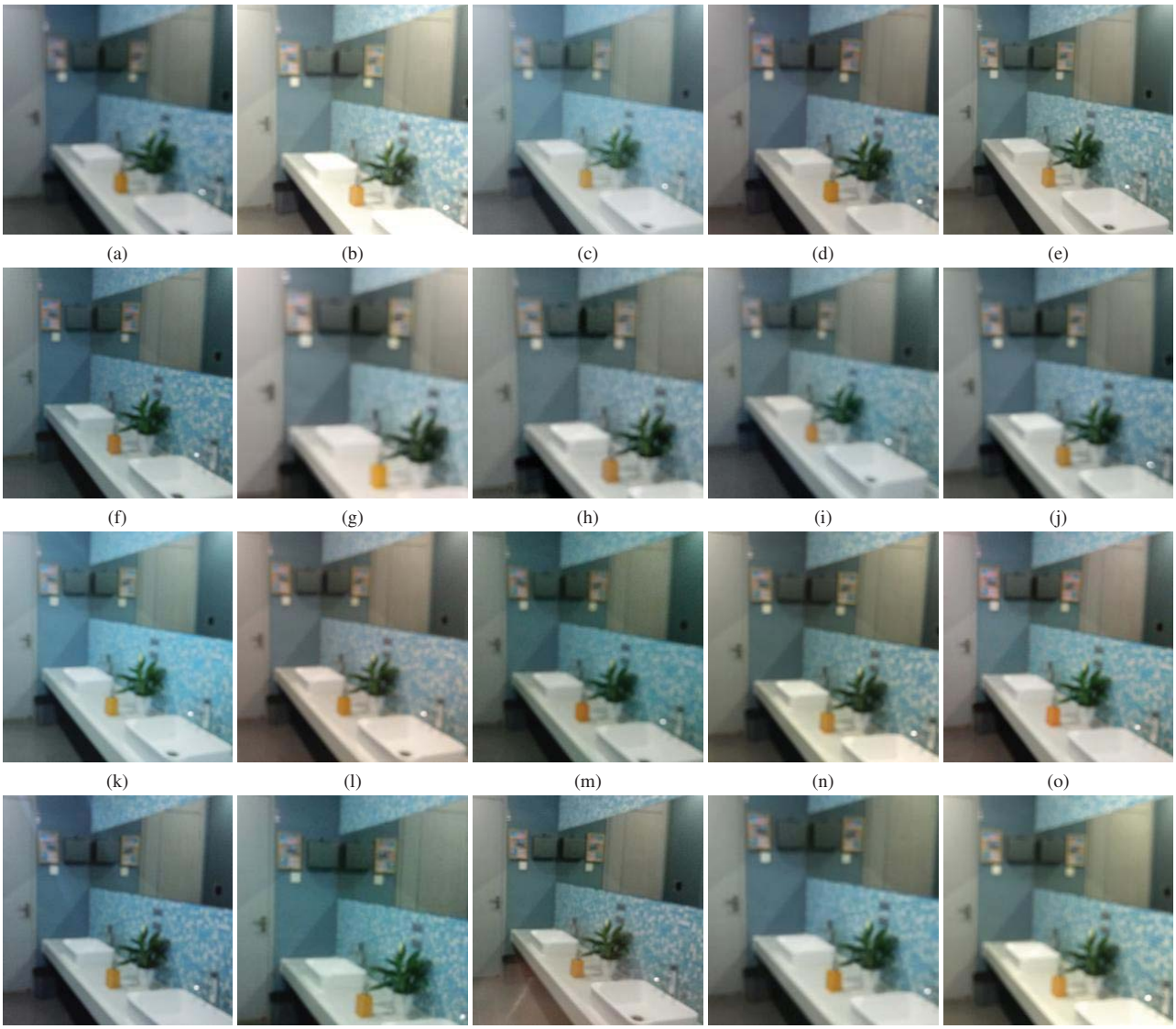
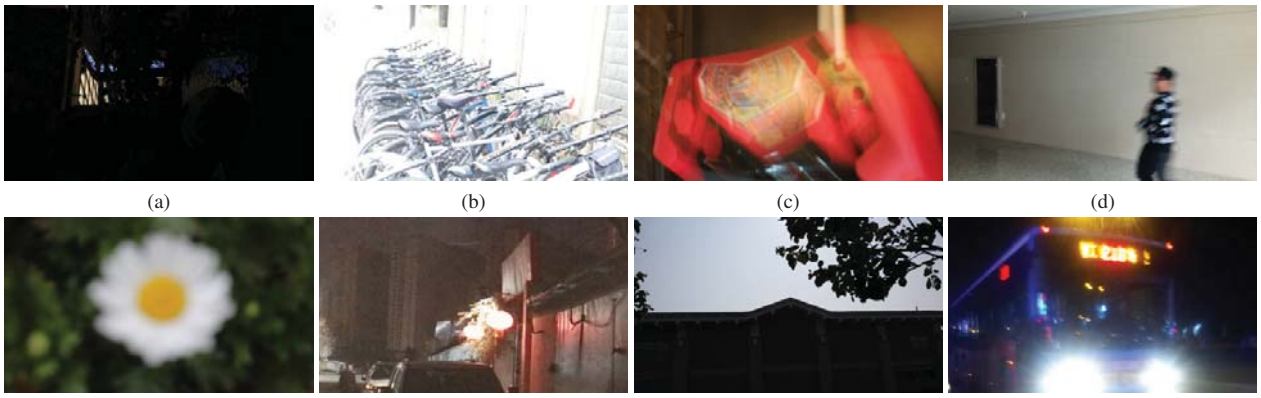


Figure S1. Sample images captured by 20 smartphone cameras with the out-of-focus configuration. (a) Meitu V6, MOS = 34.62. (b) Apple iPhone 6, MOS = 33.09. (c) Meitu M6, MOS = 34.25. (d) Samsung SM-G9200, MOS = 40.67. (e) Xiaomi MIX 2, MOS = 41.67. (f) Oppo A33m, MOS = 39.00. (g) Huawei BLA-AL00, MOS = 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.



(a) (b) (c) (d)

(e) (f) (g) (h)

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.

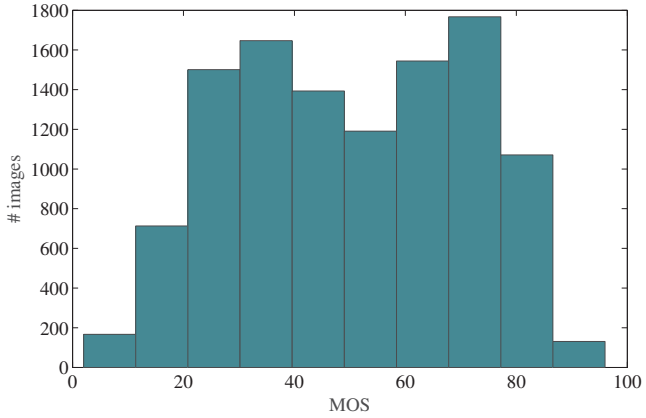


Figure S3. The histogram of MOSs in SPAQ.

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.