Image Quality Assessment – Subjective Methods and Applications

Yuming Fang (方玉明)

School of Information Technology
Jiangxi University of Finance and Economics, Nanchang, China
江西财经大学 信息管理学院

Part I: Subjective Methods

- Introduction about image quality assessment
- Subjective datasets

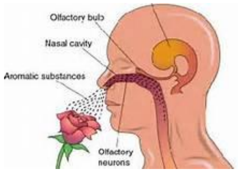
The Quality of Multimedia Content

- Audition (hearing)

- Vision (seeing)

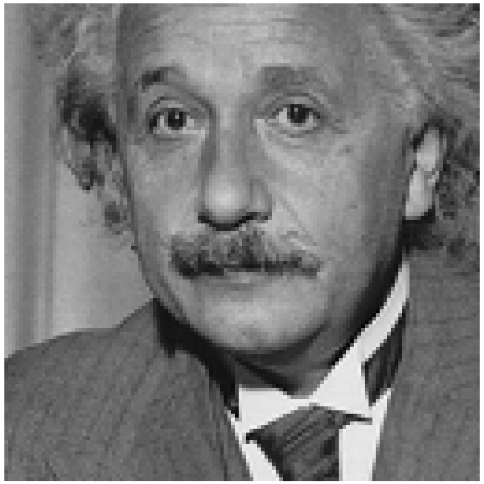
- Touch (taction)

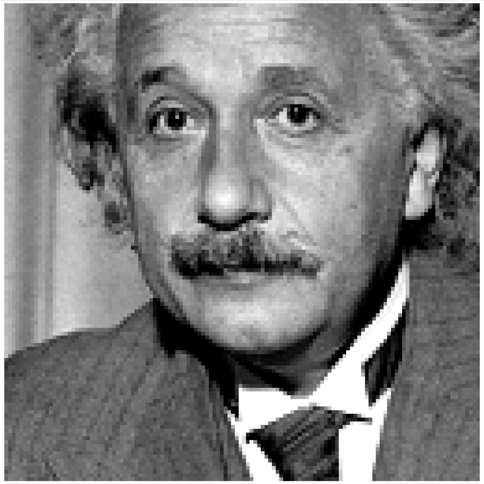
- Olfaction (smell)

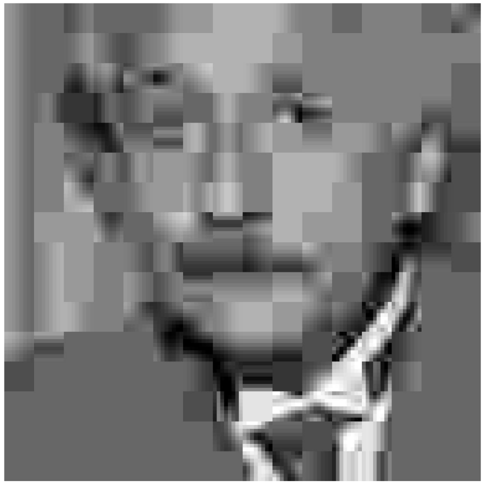


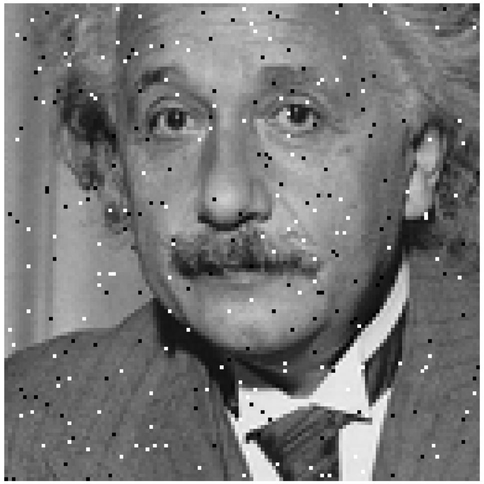
- Gustation (tasting)

What is Image Quality Assessment?









Synthetic and Authentic Distortions

Synthetic Distortions: Simulated by Pristine Image

Pristine image

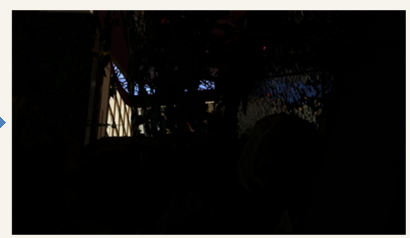
BLUR: level 4

JPEG: level 4

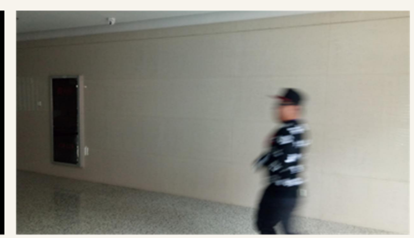
JP2K: level 4

Realistic Distortion: Captured from Mobiles Devices

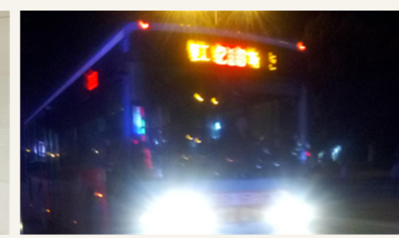
Smartphone Photography



Under-exposure



Motion blurring



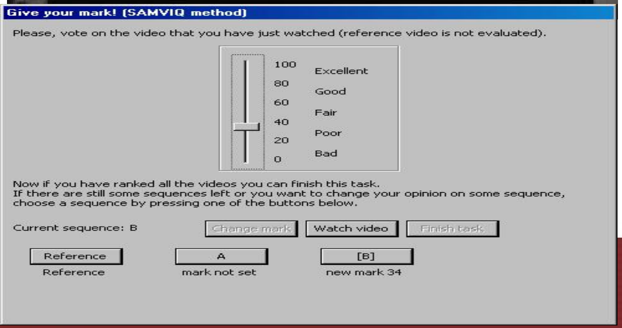
Mixture distortions

Visual Quality Assessment

Subjective Quality Assessment
Getting human quality evaluation

Observers

Environment



Methodology

Ground truth for training and/or validation →

Objective Quality Assessment
Output of a computational model

Input visual data

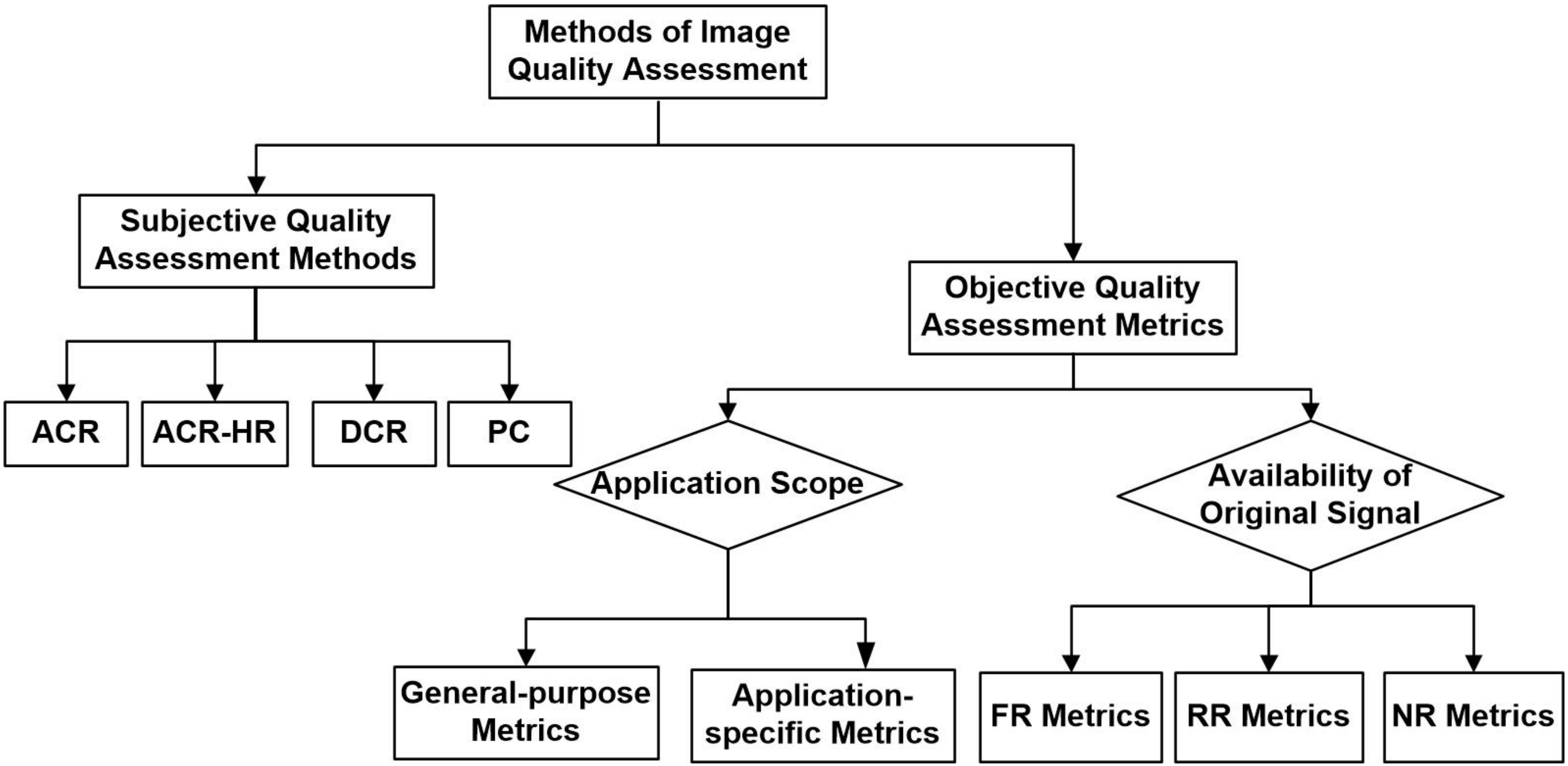
↓

Objective model

↓

Prediction Results

Visual Quality Assessment: Taxonomy



Yuming Fang, and W. Lin, 'Methods for image quality assessment', *Encyclopedia of Electrical and Electronics Engineering*, 2015.
Yuming Fang, W. Lin and S. Winkler, 'Review of Existing QoE Methodologies', Invited Chapter in *Multimedia Quality of Experience (QoE): Current Status and Future Requirements*, 2015.

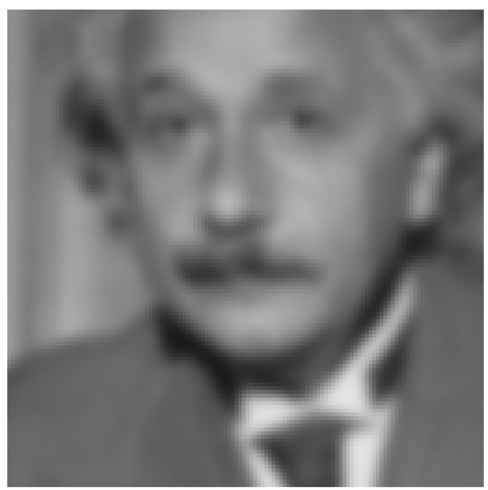
Visual Quality Assessment

- Subjective quality assessment
  - Reliable and accurate quality prediction of visual content
  - Time-consuming, laborious and expensive
  - Not applicable in practical applications

- Objective quality assessment
  - Predict perceived visual quality automatically
  - Applicable in practical applications

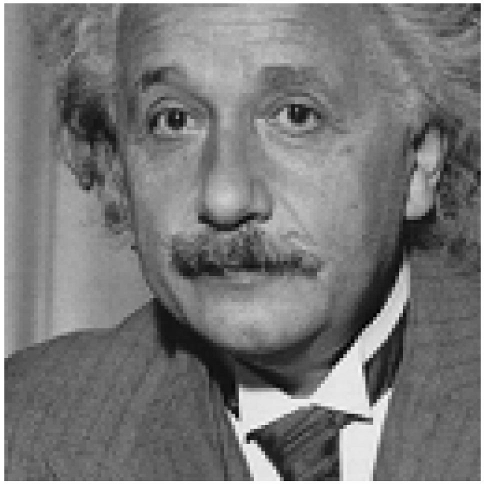
Subjective Image Quality Assessment

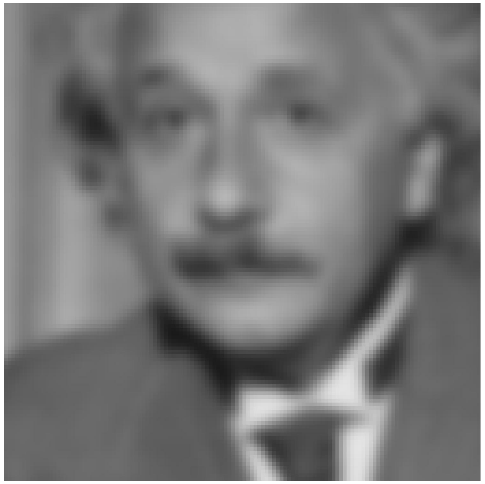
- Absolute category rating (ACR)
  - Single stimulus method
  - Test images are presented one at a time without reference information
  - Voting time: less or equal to 10 seconds depending on the voting method
  - Presentation time: 10 seconds depending on the test image content
  - Five-level or nine-level scale overall rating
- Absolute category rating with hidden reference (ACR-HR)
  - The only difference from the ACR method: a reference version of each test image must be included as the test stimulus, which is termed as a hidden reference condition



Subjective Image Quality Assessment

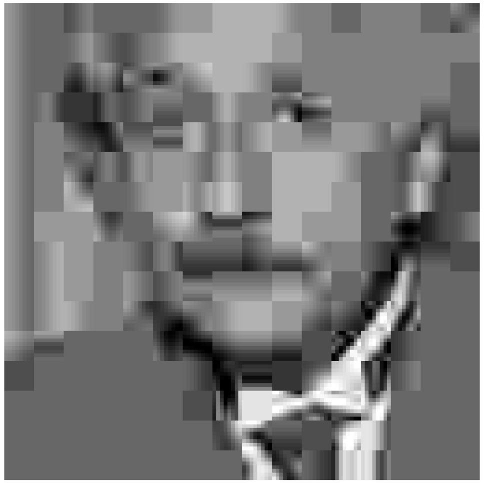
- Degradation category rating (DCR)
  - Double stimulus method
  - Test images are presented in pairs: one is reference image, while the other is distorted image
  - Voting time: less or equal 10 seconds depending on voting method
  - Presentation time: 10 seconds depending on the image content
  - Five-level scale overall rating

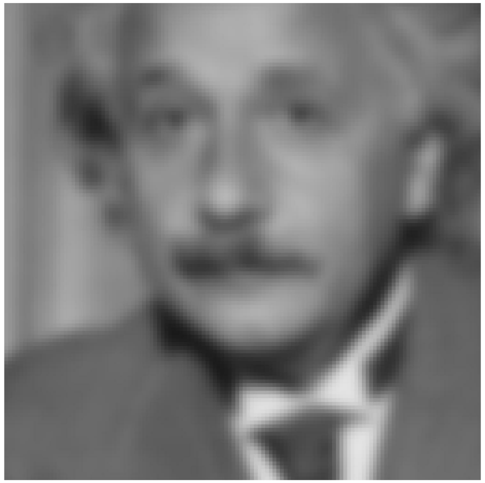




Subjective Image Quality Assessment

- Pair comparison (PC)
  - Double stimulus method
  - Two test images from two different systems are presented in pair from the same reference image
  - Participants are asked to provide the judgment on which one is preferred in the test pair
  - All possible pairs are compared
    - N stimuli -> N(n-1)/2 pairs





A B

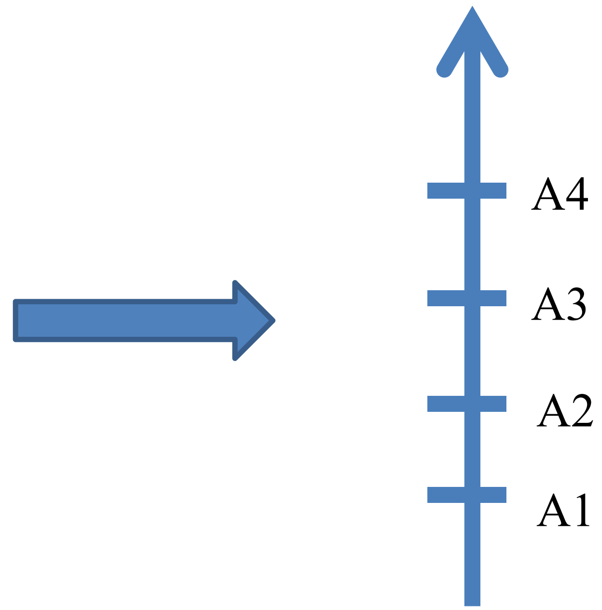
- Which one do you prefer?

Subjective Image Quality Assessment

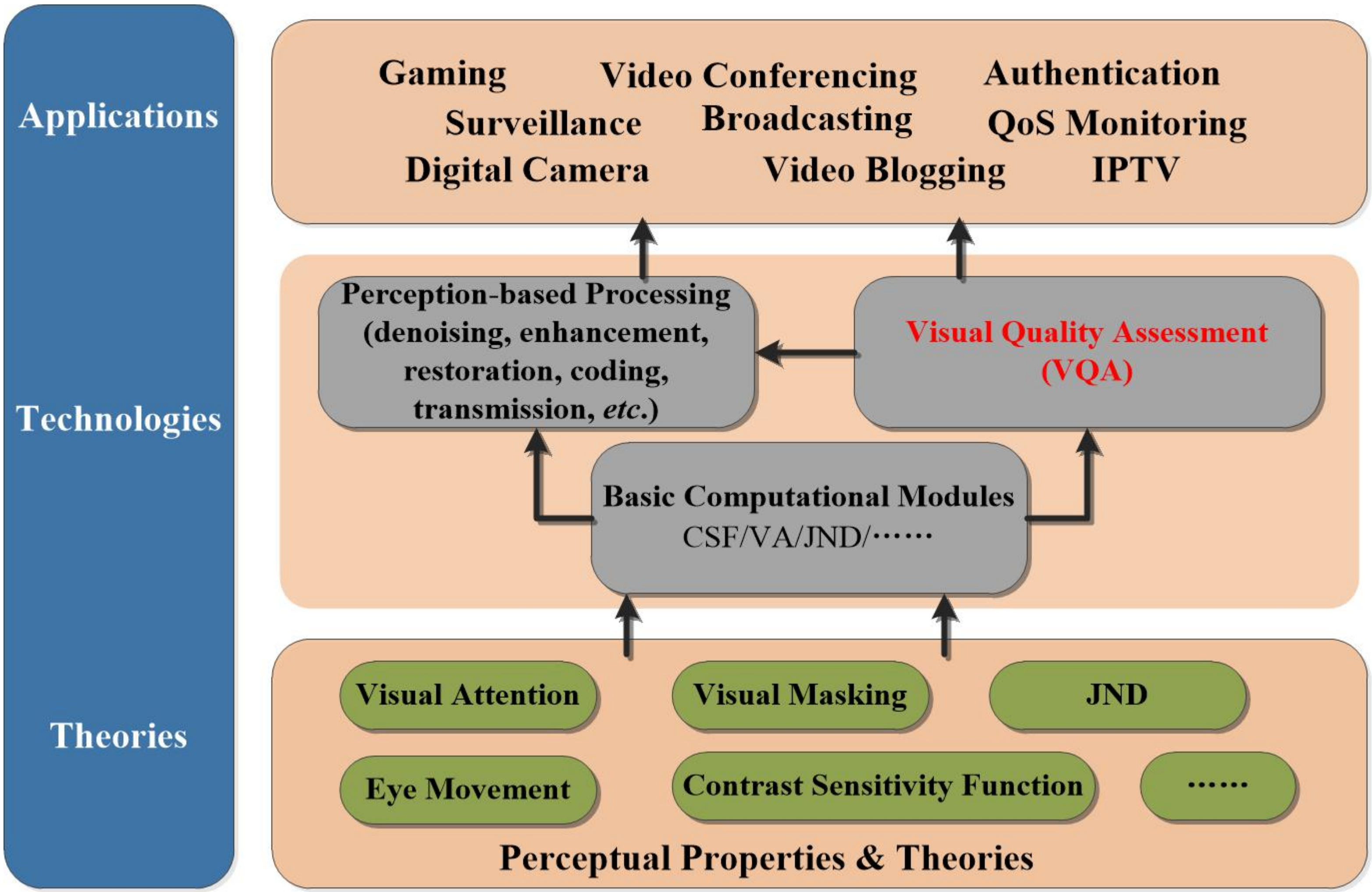
- Paired comparison model: converting paired comparison data to scale values

Pair comparison matrix

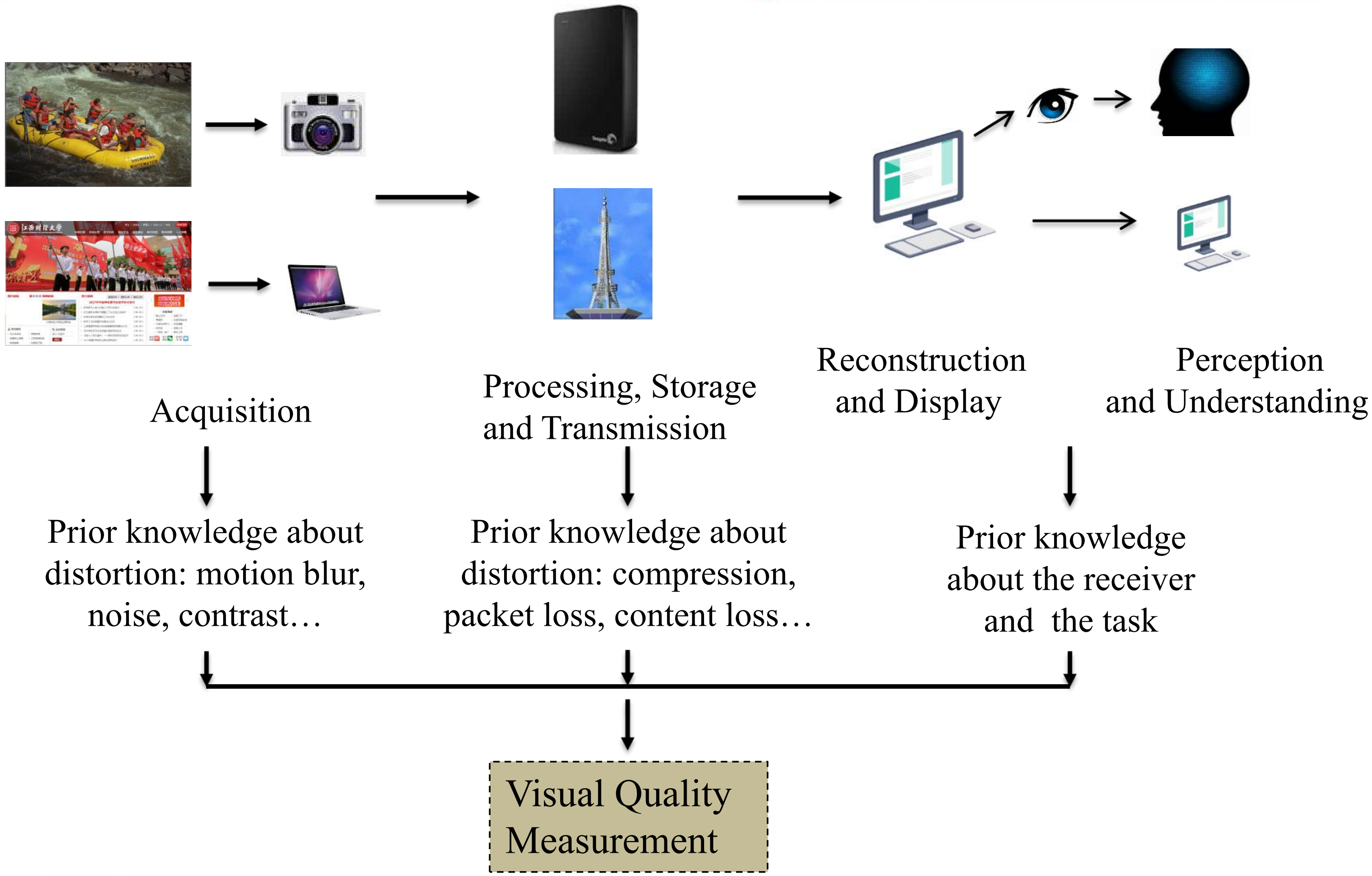
	A1	A2	A3	A4
A1	-	4	3	2
A2	6	-	4	1
A3	7	6	-	4
A4	8	9	6	-



Perceptual Visual Processing

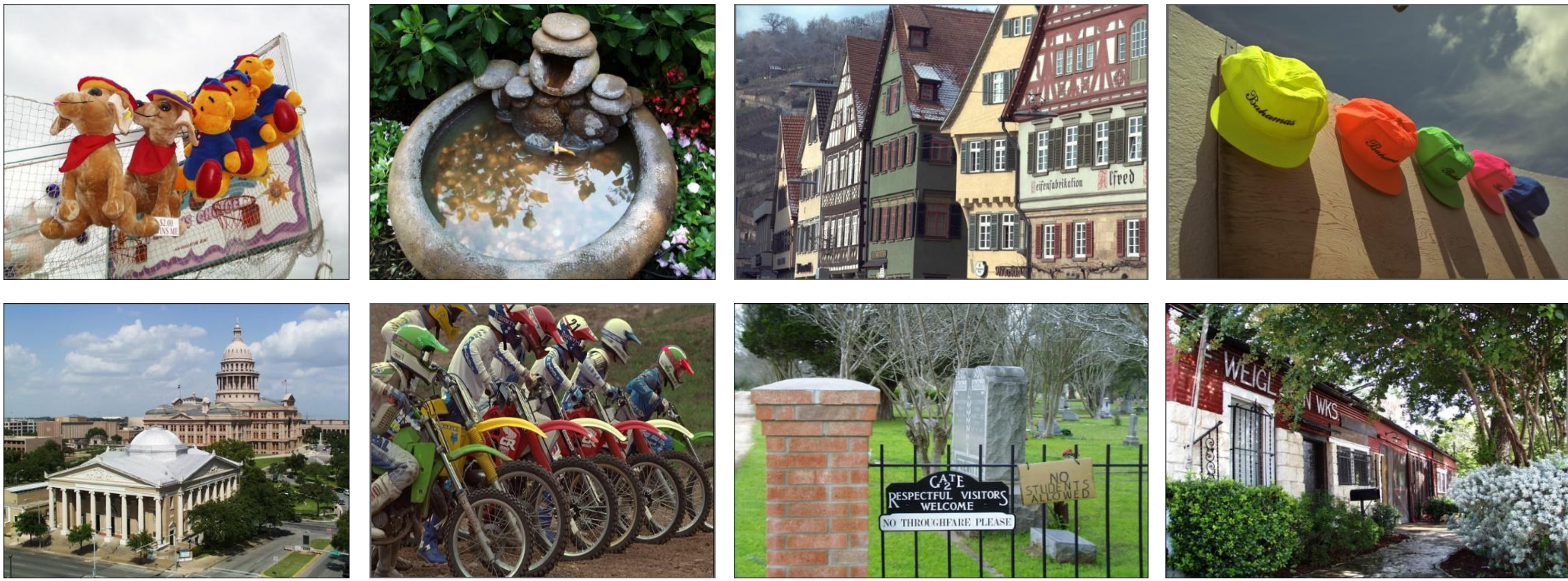


Visual Quality Assessment: A Communication Framework



LIVE Dataset

Some Reference Images in LIVE

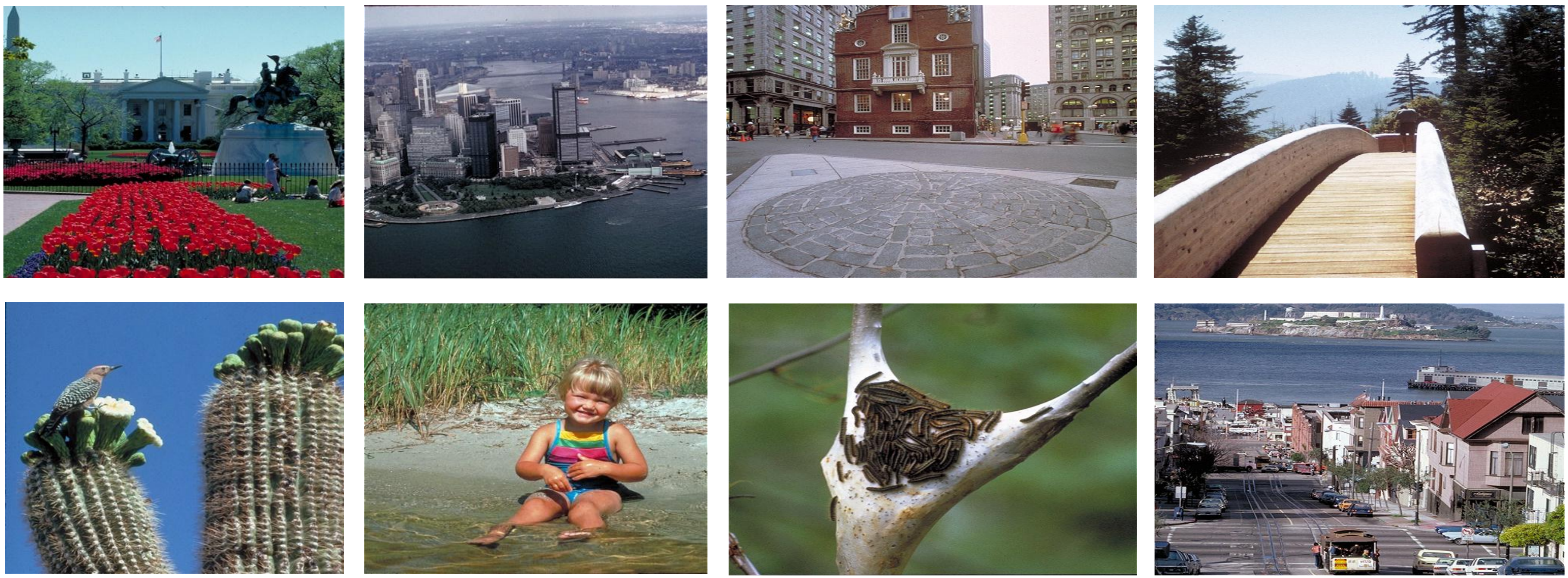


- Reference images: 29. Distorted images: 779.
- Distortion types: 5 (fast fading, Gaussian blur, JP2K, JPEG, white noise)

H. R. Sheikh, M. F. Sabir and A. C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, *IEEE T-IP*, 2006.

CSIQ Dataset

Some Reference Images in CSIQ



- Reference images: 30. Distorted images: 866.
- Distortion types: 6 (JPEG, JP2K, Gaussian blur, white noise, contrast change, pink noise)

E. C. Larson and D, M. Chandler, Most apparent distortion: Full-reference image quality assessment and the role of strategy, *Journal of Electronic Imaging*, 2010.

TID2013 Dataset

Some Reference Images in TID2013

- Reference images: 25. Distorted images: 3000.
- Distortion types: 24 (fast fading, Gaussian blur, JP2K, JPEG, white noise, etc.)

N. Ponomarenko, O. Ieremeiev, *et al.*, Color image database TID2013: Peculiarities and preliminary results, in *European Workshop on Visual Information Processing*, 2013.

KADID-10K Dataset

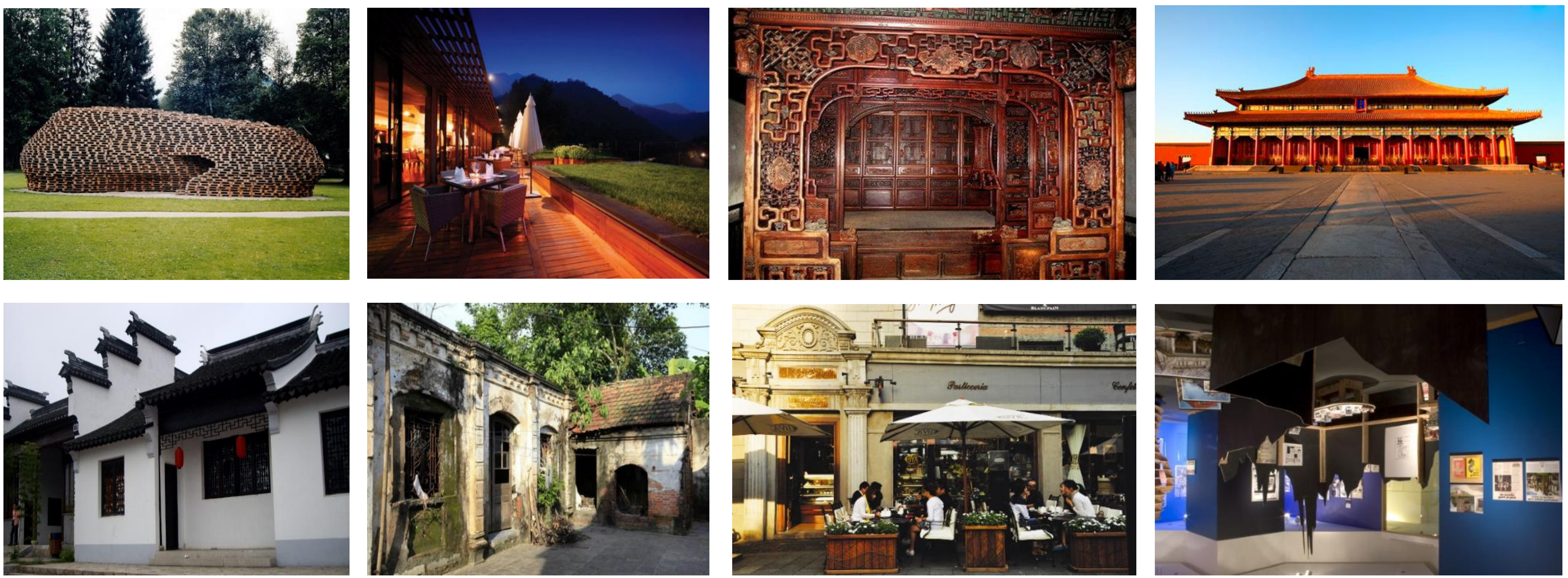
Some Reference Images in KADID-10K

- Reference images: 81. Distorted images: 10125.
- Distortion types: 25 (Gaussian blur, JP2K, JPEG, white noise, motion blur, etc.)

H. Lin, V. Hosu and D. Saupe, KADID-10K: A large-scale artificially distorted IQA database, in 2019 *Eleventh* International Conference on Quality of Multimedia Experience, 2019.

Waterloo Exploration Dataset

Some Reference Images in Waterloo Exploration



- Reference images: 4744. Distorted images: 94880.
- Distortion types: 4 (Gaussian blur, JP2K, JPEG, White noise.)

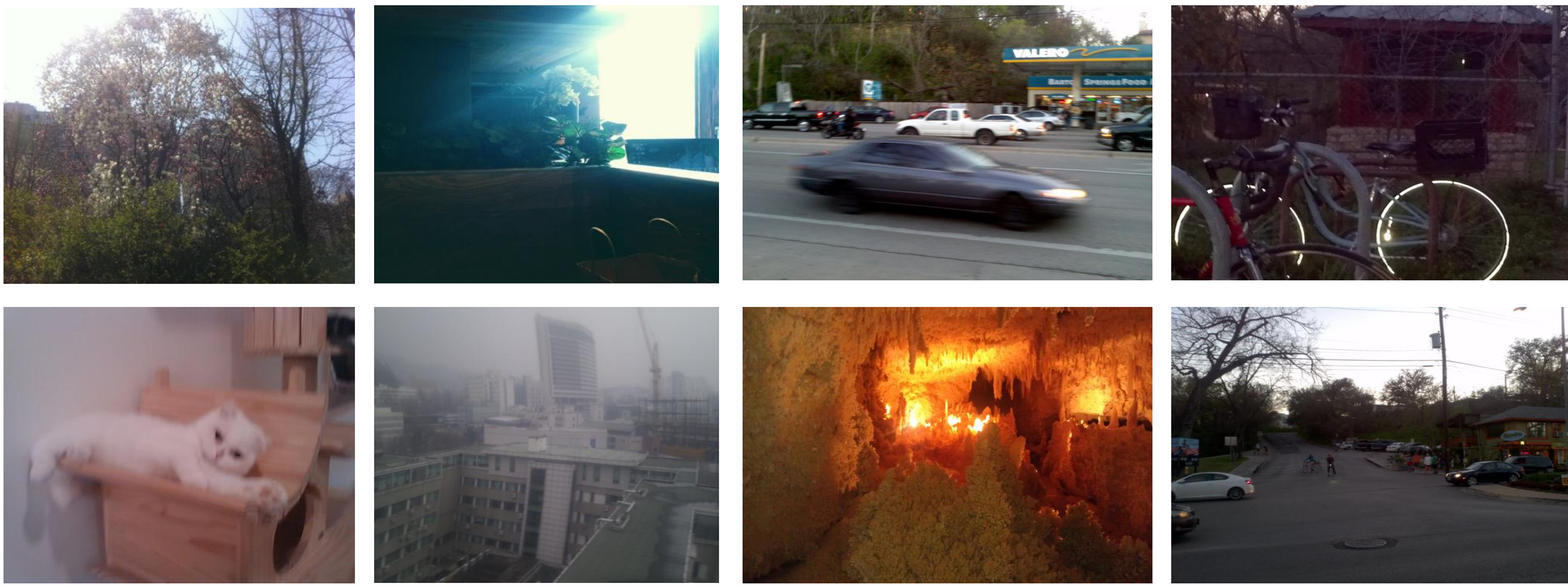
Kede Ma, *et al*., Waterloo exploration database: New challenges for image quality assessment models, *IEEE T-IP*, 2017.

Summary

- **Shortcoming**: we are often faced with the realistic distortions in real world, rather than the synthesized distortions.
- **Advantage**: it is very convenient/easy to build a large-scale database with diverse content, acting as the new independent test bed for IQA models or providing sufficient samples.

LIVE Challenge Dataset-Authentic Distortion

Some Samples in LIVE Challenge



- Distorted images: 1162.
- Distortion types: Complex.

D. Ghadiyaram and A. C. Bovik, Massive online crowdsourced study of subjective and objective picture quality, *IEEE T-IP*, 2015.

KonIQ-10K Dataset-Authentic Distortion

Some Samples in KonIQ-10K







- Distorted images: 10073.
- Distortion types: Complex.

V. Hosu, H. Lin, T. Sziranyi and D. Saupe, KonIQ-10K: An ecologically valid database for deep learning of blind image quality assessment, *IEEE T-IP*, 2020.

Smartphone Photography

Smartphone manufactures

Fast development of smartphone photography technologies:

- Hardware: Dual-camera system, wide-angle lens
- Software: HDR, portrait, panorama

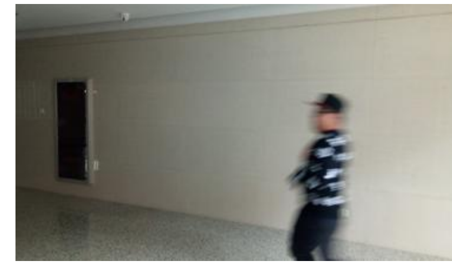
Smartphone Photography Attribute and Quality (SPAQ) Database

Sample Images in SPAQ

Under-exposure

Over-exposure

Contrast reduction



Moving object blurring

Sensor noise

Out-of-focus

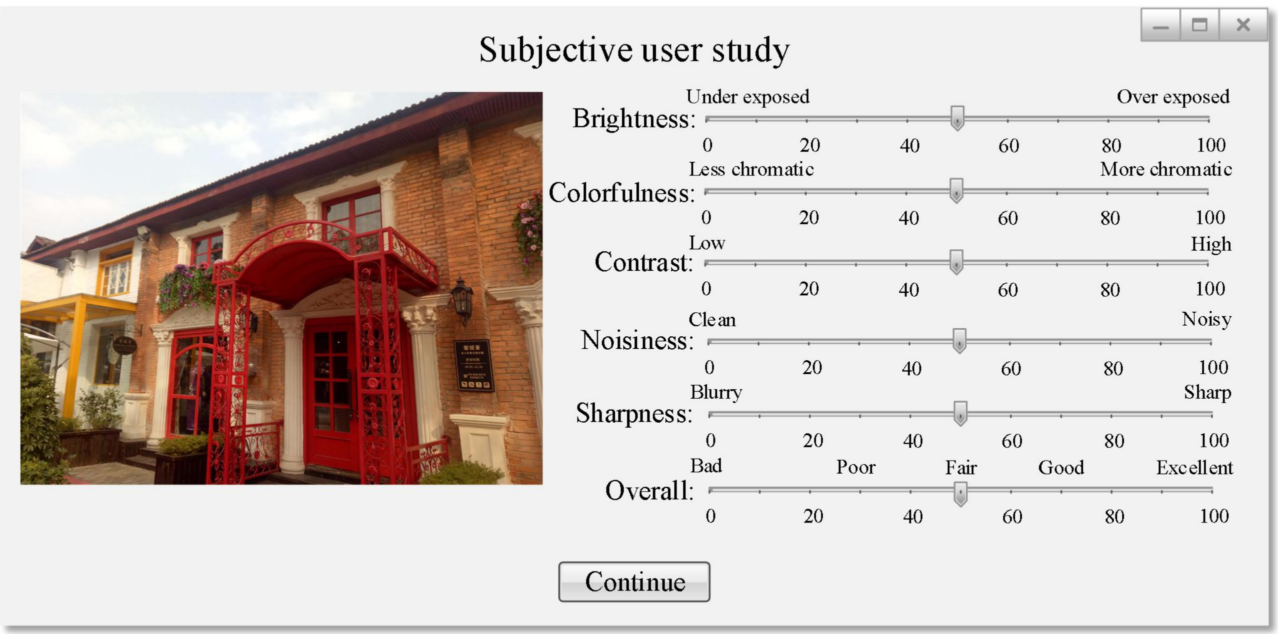
Camera motion blurring

Mixture distortions

We introduce a new image database, consisting of **11,125** pictures taken by 66 smartphones with 11 manufacturers.

Subjective Experiments

We conduct so far the most comprehensive study of perceptual quality assessment of smartphone photography, 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.



Subjective user study for image quality and image attributes

- First stage: 1,125 images was rated by 104 sessions, and each session rated 80 images.
- Second stage: More than 600 subjects were invited to involved in the experiment, and each session rated 80 images (70 random selected from 10,000 images + 5 duplicated images + 5 images from first stage).

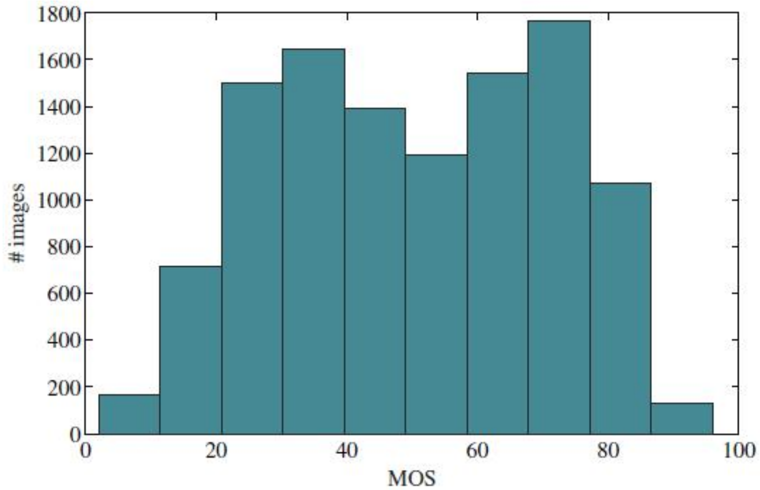
Subjective Experiments

Subjective user study for scene category labels

Exchangeable image file format (EXIF) tags: 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 scene), 6) flash (flash fired or not), 7) time (when image was recorded).

The EXIF data carry useful information about the scene being captured and the camera settings, which may help to predict the quality of smartphone photography.

Subjective Data Analysis



The histogram of MOSs of the images in our database.

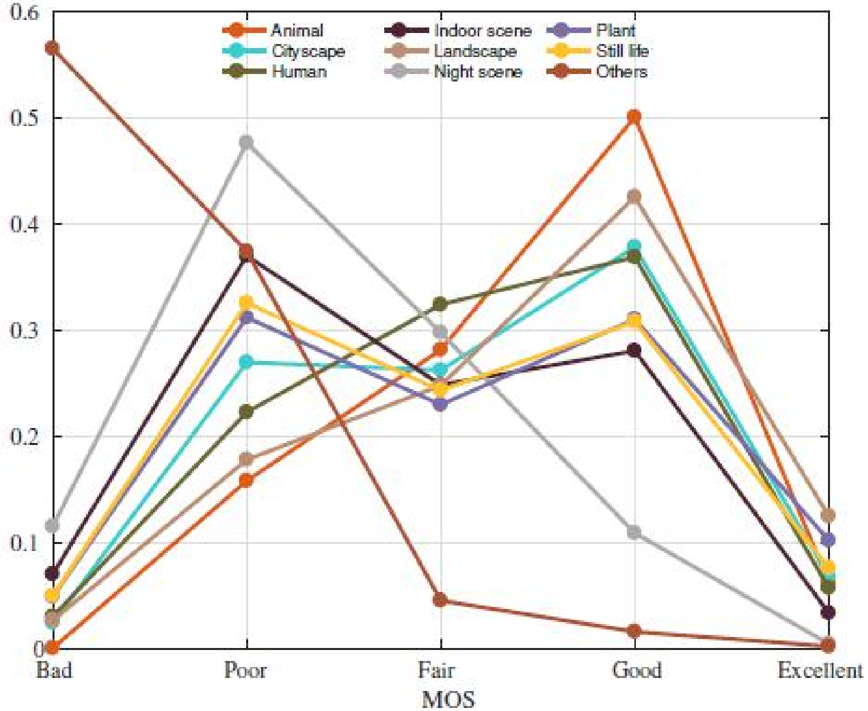
Attribute	Image attribute scores from subjects	Image attribute scores by MT-A
Brightness	0.784	0.704
Colorfulness	0.844	0.760
Contrast	0.874	0.786
Noisiness	0.893	0.832
Sharpness	0.958	0.904

Criteria	SRCC	PLCC
Consistency across duplicated images	0.893	0.903
Consistency across sub-groups	0.923	0.930
Consistency across subjects	0.841	0.865

- Consistency across duplicated images: correlation between MOSs of duplicated images at two stages.
- Consistency across sub-groups: correlation between MOSs from two sub-groups of participants.
- Consistency across subjects: correlation between MOSs of individual articipant and all participants.

- First column: SRCC values between MOSs and attribute scores from participants.
- Second column: SRCC values between MOSs and predicted attribute scores by MT-A.

Subjective Data Analysis

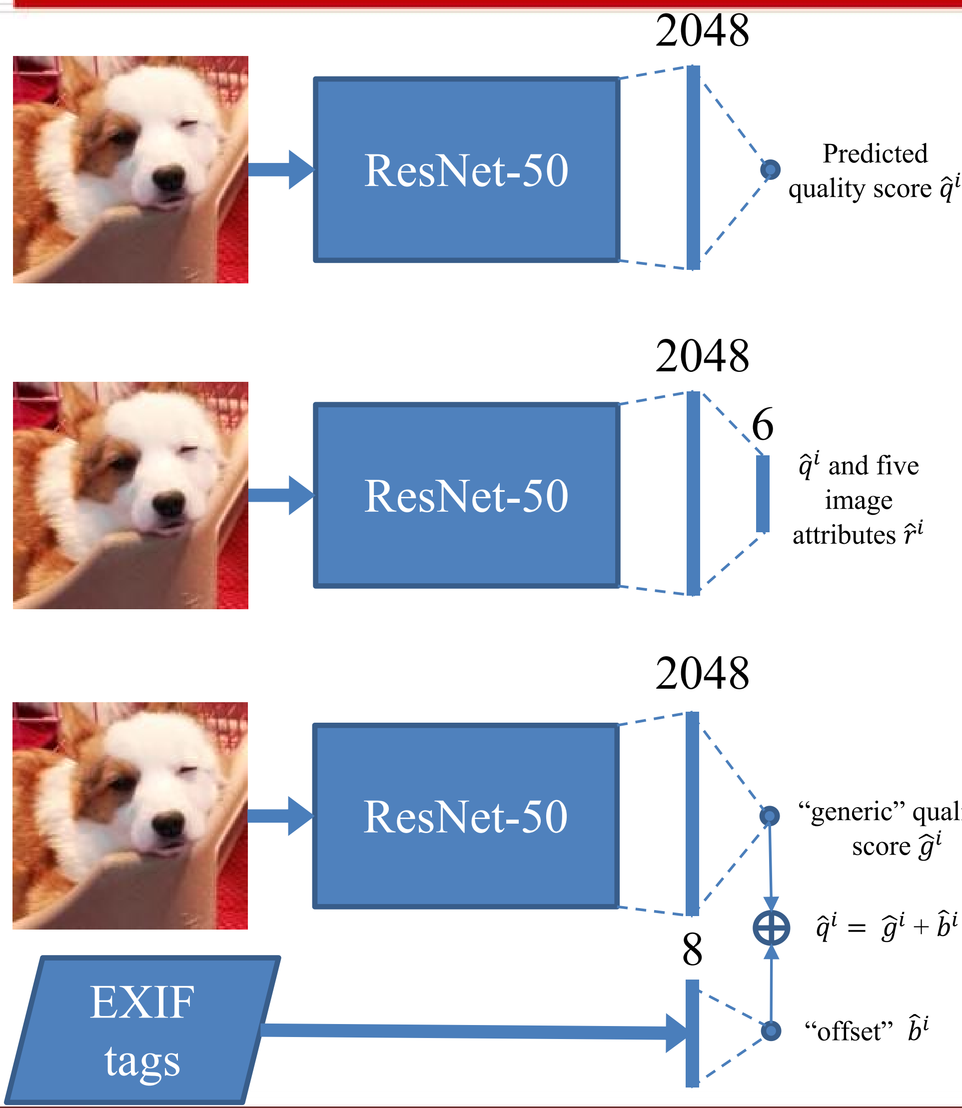


The MOS distribution of images in five quality levels for each scene category. Bad: MOS ∈ [0, 19], Poor: MOS∈ [20, 39], Fair: MOS ∈ [40, 59], Good: MOS ∈ [60, 79], and Excellent: MOS ∈ [80, 100].

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

The top five and bottom five smartphone cameras based on image quality.

Objective Quality Models



$$\mathcal{L}_1(\mathcal{W}_B) = \|q - \hat{q}\|_1 = \sum_{i=1}^{m} |q^i - \hat{q}^i|$$

$\mathcal{W}_B$ indicates the parameters in the baseline model (BL), and q^i is the MOS of *i-th* image.

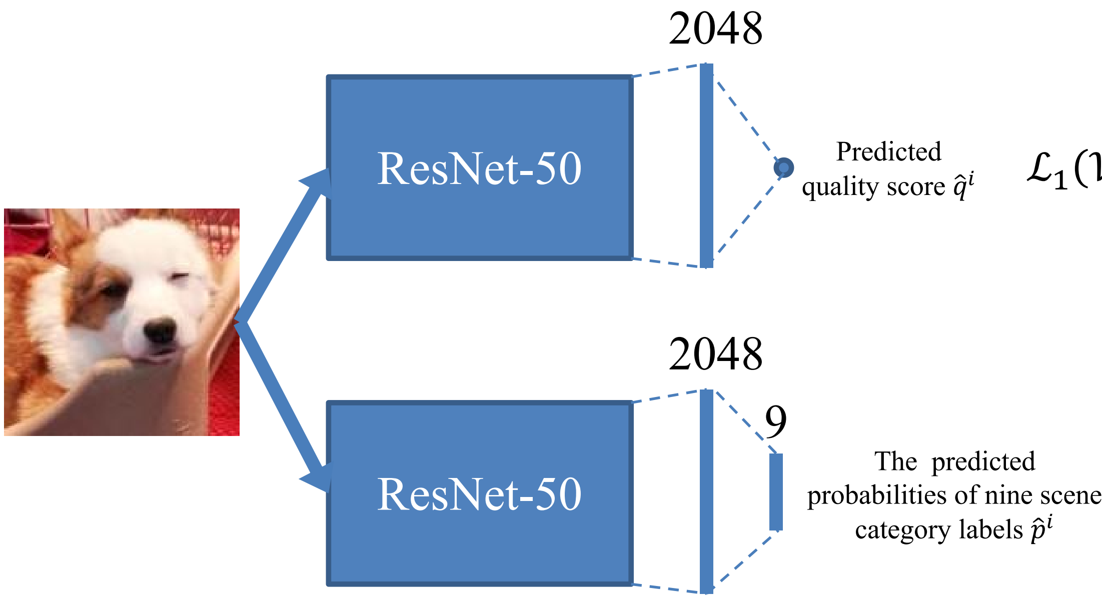
$$\mathcal{L}_2(\mathcal{W}_B, \mathcal{W}_A) = \beta_1 \|q - \hat{q}\|_1 + \frac{\beta_2}{5} \|r_j - \hat{r}_j\|_1$$

$\mathcal{W}_A$ indicates the parameters in the multi-task attribute (MT-A) model, and r^i is the attribute scores of *i-th* image. $\beta_1 + \beta_2 = 1$.

$$\mathcal{L}_3(\mathcal{W}_B, \mathcal{W}_E) = \alpha_1 \|q - \hat{g}\|_1 + \alpha_2 \|q - \hat{q}\|_1$$

$\mathcal{W}_E$ indicates the parameters in the multi-task EXIF (MT-E) model. $\alpha_1 = \alpha_2 = 0.5$.

Objective Quality Models



$$\mathcal{L}_1(\mathcal{W}_B) = \|q - \hat{q}\|_1 = \sum_{i=1}^{m} \left|q^i - \hat{q}^i\right|$$

$$\mathcal{L}_4(\mathcal{W}_S) = -\sum_{i,j} p_j^{(i)} \hat{p}_j^{(i)}$$

$$\mathcal{L}_5(\mathcal{W}_B, \mathcal{W}_S) = \frac{\mathcal{L}_1(\mathcal{W}_B)}{\sigma_1} + \frac{\mathcal{L}_4(\mathcal{W}_S)}{\sigma_2} + m\log\sigma_1 + \frac{m}{2}\log\sigma_2$$

$\mathcal{W}_S$ indicates the parameters in the multi-task scene semantic (MT-S) model, and p^i is the scene category label of *i-th* image. σ_1 *and* α_2 are two jointly learning parameters, which help to balance the image quality regression task and image classification task.

Performance Evaluation

Average PLCC and SRCC Results of Our Methods against Seven BIQA Models on SPAQ

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

Average PLCC and SRCC Results of the Proposed BL Model in a Cross-Database Setting

Training	SPAQ				
Testing	Synthetic database		Realistic database		
	LIVE	TID2013	CID2013	LIVE Challenge	KonIQ-10k
PLCC	0.608	0.570	0.771	0.773	0.745
SRCC	0.560	0.397	0.754	0.742	0.707

Summary

- BIQA models designed for synthetic distortions (e.g., QAC and DIIVINE) generally do not work well for realistic camera distortions.
- Verified on the LIVE Challenge Database, FRIQUEE delivers superior performance on the proposed database.
- DBCNN outperforms all BIQA approaches, including the proposed BL based on ResNet-50, which suggests that DNNs successfully learn hierarchical sensitive features to realistic distortions.
- Image attributes (MT-A) positively impacts the accuracy of quality prediction.
- 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.
- 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

Database & Models:

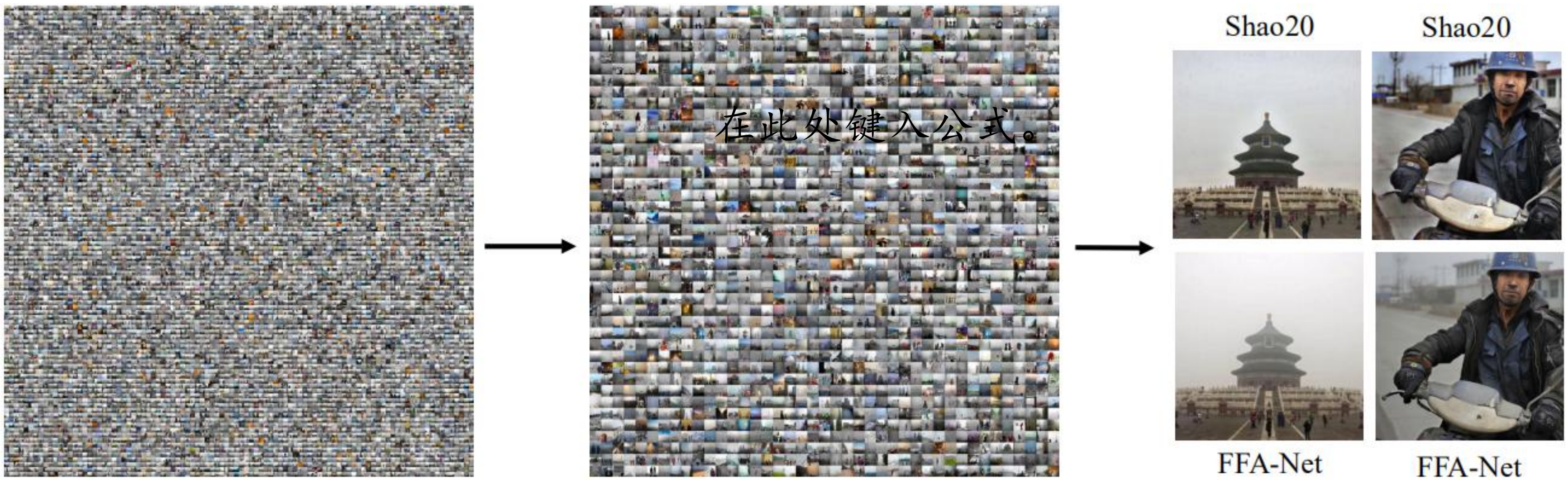
https://github.com/h4nwei/SPAQ

Yuming Fang, H. Zhu, Y. Zeng, **Kede Ma**, Z. Wang, Perceptual quality assessment of smartphone photography, in ***IEEE CVPR***, 2020.

How to Create debiased IQA Databases

- Conventional subjective testing requires manually pre-selecting a small set of visual examples, which may suffer from three sources of biases:
  - *Sampling bias* due to the extremely sparse distribution of the selected samples in the image space;
  - *Algorithm bias* due to potential overfitting the selected samples;
  - *Subjective bias* due to further potential cherry-picking test results.
- Target: debiased subjective assessment.

Debiased Subjective Assessment of Real-World Image Enhancement



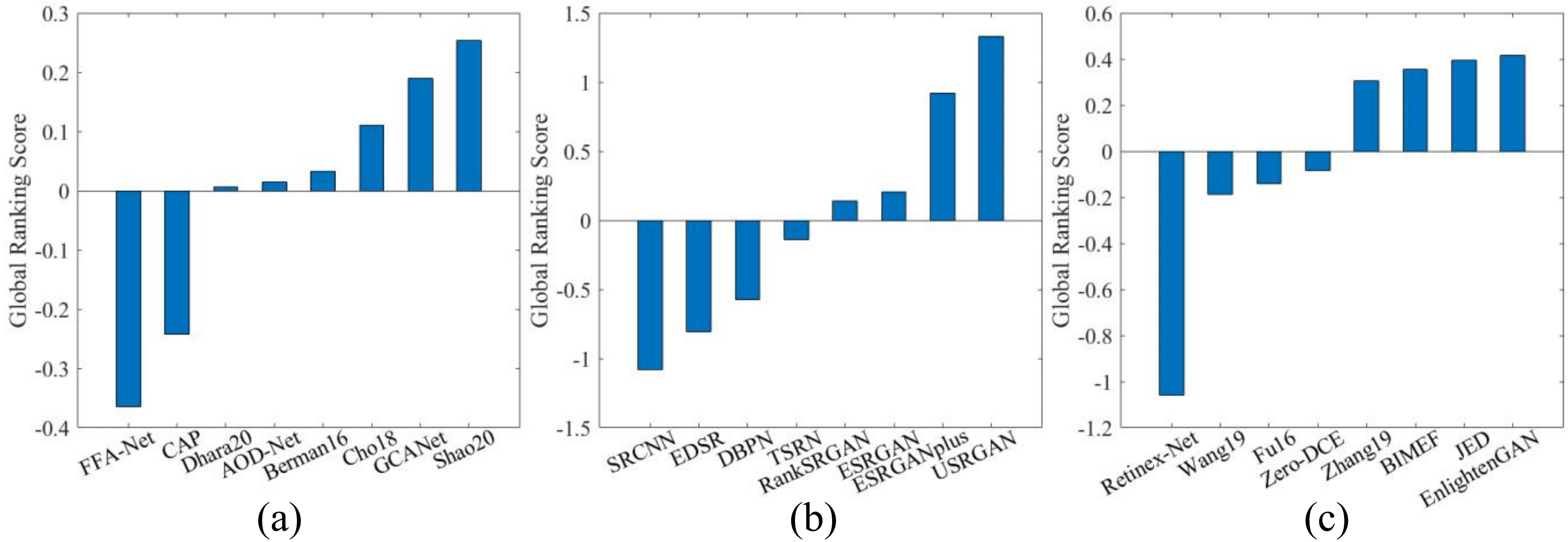
(a) (b) (c)

Overview of our debiased subjective assessment in the context of single image dehazing (a) A large set of hazy images. (b) Top-*K* hazy images selected from (a) to best discriminate between Shao20 and FFANet by optimizing Eq. (1). (c) Pairs of dehazed images corresponding to representative hazy images in (b).

$$\hat{x}^{(k)} = \arg\max_{x \in \chi \setminus S} D_1\big(f_1(x), f_2(x)\big) + \lambda_1 D_2(x, S) \quad (1)$$

P. Cao, Z. Wang, **Kede Ma**, Debiased subjective assessment of real-world image enhancement, in ***IEEE CVPR***, 2021.

Debiased Subjective Assessment of Real-World Image Enhancement



Global ranking results of (a) single image dehazing, (b) single image super-resolution, and (c) low-light image enhancement.

P. Cao, Z. Wang, **Kede Ma**, Debiased subjective assessment of real-world image enhancement, in ***IEEE CVPR***, 2021.