Group Maximum Differentiation Competition: Model Comparison with Few Samples

Kede Ma, Zhengfang Duanmu, Zhou Wang, *Fellow, IEEE,* Qingbo Wu, Wentao Liu, Hongwei Yong, Hongliang Li, *Senior Member, IEEE,* and Lei Zhang, *Fellow, IEEE*

Abstract—In many science and engineering fields that require computational models to predict certain physical quantities, we are often faced with the selection of the best model under the constraint that only a small sample set can be physically measured. One such example is the prediction of human perception of visual quality, where sample images live in a high dimensional space with enormous content variations. We propose a new methodology for model comparison named group maximum differentiation (gMAD) competition. Given multiple computational models, gMAD maximizes the chances of falsifying a "defender" model using the rest models as "attackers". It exploits the sample space to find sample pairs that maximally differentiate the attackers while holding the defender fixed. Based on the results of the attacking-defending game, we introduce two measures, *aggressiveness* and *resistance*, to summarize the performance of each model at attacking other models and defending attacks from other models, respectively. We demonstrate the gMAD competition using three examples—image quality, image aesthetics, and streaming video quality-of-experience. Although these examples focus on visually discriminable quantities, the gMAD methodology can be extended to many other fields, and is especially useful when the sample space is large, the physical measurement is expensive and the cost of computational prediction is low.

Index Terms—Model comparison, gMAD competition, image quality, image aesthetics, streaming video quality-of-experience.

1 Introduction

In many science and engineering fields, we desire to construct computational models that can predict certain measurable physical quantities. A common constraint we are often faced with is that the physical measurement process is costly. As a result, only a small number of samples can be measured, relative to the large sample space within which the computational models attempt to make predictions. This casts major challenges to the validation, comparison, and improvement of the computational models. One such example is the prediction of perceptually discriminable quantities such as image quality [1], where multiple computational models for image quality prediction are available and we are asked which one performs the best.

Model comparison has been a long-standing problem [2]. A common theme of conventional direct model comparison methods is to prepare a number of samples from the sample space, collect physical measurements as the ground-truth, and select the model that best fits the ground-truth measurements in terms of certain statistical criteria. Such criteria include statistics on (1) prediction accuracy, *e.g.*, the mean squared error (MSE) [3] and the Pearson's linear correlation coefficient [4] between model predictions and ground truths; (2) prediction monotonicity, *e.g.*, the Spearman's rank correlation coefficient (SRCC) [4] between model predictions and ground truths; and (3) prediction consistency, *e.g.*, the outlier ratio that accounts for the percentage of unreasonable predictions. Assume the evaluation is independent, meaning that the models have never seen the test samples, then such statistics provide a basis for comparing computational models on a given set of samples.

There are two interrelated problems of direct model comparison methods. First, it is commonly believed unfair to compare models with different complexities solely by their goodness of fit [5]. The principle of Occam's razor [6] suggests that for equal goodness of fit, a simpler model is better. Many model comparison approaches that incorporate a simplicity measure have been proposed, including the Akaike information criterion [5], the shortest data description [7], and the Bayesian information criterion [8]. Regardless of whether the heuristic of picking the simpler model or the Occam's razor is well justified, measuring the complexity of a model is a difficult problem by itself. The most intuitive idea is to count the number of parameters, but even a single continuous-valued parameter contains an infinite amount of information. An alternative measure is the description length, which depends on the description method, and the absolute shortest description, *i.e.*, the Kolmogorov complexity [9], is not computable. Thus a question that follows is whether the complexity in computing the description length should be counted as part of the model complexity. Moreover, how to strike a right balance between goodness of fit and simplicity is a difficult question to answer and could be application dependent. In addition, recent studies of deep neural networks suggest that deeper networks with a gigantic number of parameters could generalize better than shallower networks with a smaller number of parameters [10], adding more complications to model comparison

- *K. Ma, Z. Duanmu, W. Liu, and Z. Wang are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: {k29ma, zduanmu, w238liu, zhou.wang}@uwaterloo.ca).*
- *Q. Wu and H. Li are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China (e-mail: wqb.uestc@gmail.com; hlli@uestc.edu.cn).*
- *H. Yong and L. Zhang are with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: {cshyong, cslzhang}@comp.polyu.edu.hk).*

methods that incorporate complexity measures.

Second, there is often a major conflict between the large scale (and possibly high dimensionality) of the sample space and the limited scale of the affordable physical measurement. For example, consider the space of all visual images. This sample space is of the same dimension as the pixel number in the image, which is often in the order of millions. Collecting ground-truth data via subjective testing is expensive and time-consuming. Therefore, a typical "large-scale" subjective experiment allows for a maximum of a few thousand sample images to be examined, which are deemed to be extremely sparsely distributed in the sample space. Model comparison methods based on limited samples assume that the samples are *sufficiently representative*, an assumption that is often doubtful. The verification of such representativeness is by itself a challenging problem without enough ground-truth samples.

Conventional direct model comparison methods have two features in common. First, they provide *absolute* assessments, meaning that the evaluation of one model is independent of other models. Model comparison occurs after such absolute assessments have been performed on all competing models. Second, all assessments attempt to *prove* a model to be correct by measuring its goodness of fit. An exception is the model falsification methodology [11], where a model is rejected when certain statistical criteria between the ground-truth measurements and the model predictions are outside some prescribed bounds. A significant departure from conventional direct model comparison approaches started from the MAximum Differentiation (MAD) competition method [12]. Given two computational models, MAD works by *falsifying* a model using the second model in the most efficient way and a model that is more difficult to be falsified is considered better. To select samples that maximally discriminate between the two models, MAD employs a gradient-based iterative algorithm to synthesize a pair of samples that maximize/minimize the responses of one model while holding the other fixed. The procedure is repeated with the roles of the two models switched. Only such extreme samples are subject to physical measurement. MAD gives us an opportunity to largely reduce the number of samples for testing because theoretically only one counterexample is sufficient to falsify a model.

Nevertheless, several limitations of MAD impede its wide usage in practical applications. First, MAD relies on gradient information of the two models to solve a constrained and possibly nonconvex optimization problem. This is not plausible for sophisticated computational models, whose gradients are difficult to compute, if not impossible. Second, MAD-synthesized samples may be highly unnatural [12], whose practical implications on how to improve existing models in real-world applications may be limited. Third, it is difficult to control the synthesized samples to fall in any specific *domain of interest*, which may be a subset of the sample space. Fourth, it applies to two models only and the extension to account for multiple models is nontrivial.

We aim to develop the principle behind MAD [12] towards an efficient and practical methodology for comparing multiple computational models of measurable physical quantities. We name our method the group MAximum Differentiation (gMAD) competition. When attempting to falsify a model (denoted as the *defender*), we work with a large-scale sample set without performing physical measurements. We search for sample pairs that maximize/minimize the responses of a group of other models (denoted by *attackers*), while fix the responses of the defender. The attacks are optimal in the sense that the defender is most likely to be falsified by the attackers. gMAD runs this game among all models until each and every of them has played the defender role once. Psychophysical experiments on generated sample pairs are then conducted. Moreover, we introduce the *aggressiveness* and *resistance* measures to quantify how aggressive an attacker is at falsifying a defender and how resistant a defender is at defending itself against an attacker, respectively. The pairwise aggressiveness and resistance statistics are aggregated into a global ranking. The gMAD competition is readily extensible, allowing future models to be added with minimal additional work.

To demonstrate gMAD in a practical setting, we apply it to the field of image quality assessment (IQA) [1], [3] and report the competition results on 16 IQA models. Careful inspections of selected gMAD image pairs shed light on how to improve existing IQA models and develop next-generation models. We also explore gMAD in two more applications—image aesthetics evaluation [13] and streaming video quality-of-experience (QoE) prediction [14].

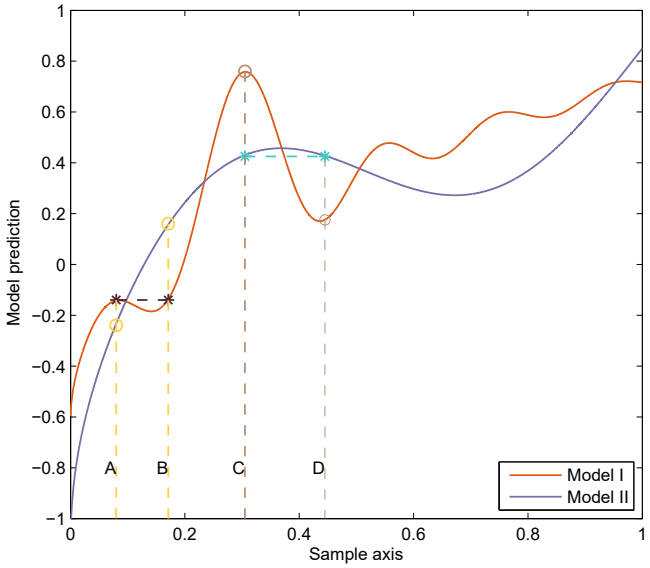
2 The gMAD Competition Methodology

2.1 Problem Formulation and A Toy Example

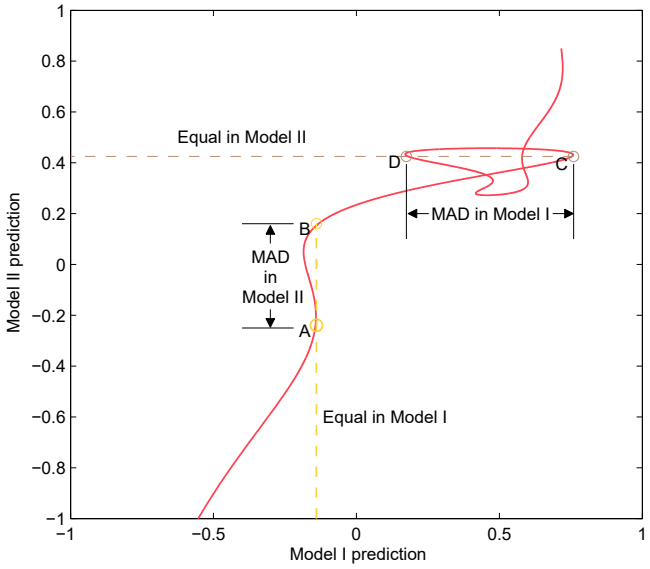
We assume a sample space $\mathcal{S}$, upon which a physical quantity $q \in \mathbb{R}$ is measurable for any sample $s \in \mathcal{S}$. A group of computational models $\{C_i\}_{i=1}^{M}$ are also assumed, each of which takes a sample s as input and makes a prediction of $q(s)$. The goal is to compare the relative prediction performance among all models with a limited number of physical measurements.

The gMAD competition method works with a sample set $O = \{s_i\}_{i=1}^{N} \subset \mathcal{S}$. gMAD selects the sample set O that covers the domain of interest within the sample space $\mathcal{S}$. A good example is the natural image subset in the set of all possible digital images. Only a very small number of samples are selected by gMAD for physical measurements and thus the size of O is not a major concern. As such, O can be selected to densely cover the domain of interest.

We first illustrate the idea of the gMAD competition using a toy example as shown in Fig. 1. We assume two models, Model I and Model II, to predict a continuous quantity q, which varies over an one-dimensional sample space. The physical measurement of q is expensive at any sample point, but the computation of model predictions is cheap. The predictions by Model I and Model II are shown in Fig. 1(a), where we observe that the models generally agree with each other but may make very different predictions at certain sample points. The question is how to determine the better model with a minimal number of samples being physically measured. gMAD aims to maximize the efficiency of falsifying the models by letting them compete. The process is better explained in a scatter plot (Fig. 1(b)) of Model I versus Model II. The samples that have the same Model I response may expect different Model II responses, among



(a)



(b)

Fig. 1. A toy example of the gMAD competition. (a) Predictions by Model I and Model II. (b) Plot of Model II against Model I. (A, B) and (C, D) are the gMAD sample pairs subject to physical measurement.

which we are interested in two samples corresponding to the minimal and maximal Model II responses, respectively. When we go through the Model I axis, the pair with the maximum response difference of Model II are selected, as Points A and B in Fig. 1(b). Similarly, the sample pair that maximize the response difference of Model I for equal Model II response are selected, as Points C and D in Fig. 1(b). The two sample pairs (A, B) and (C, D) are subject to physical measurement. (A, B) is the best counterexample that Model II finds to falsify Model I. If the ground-truth $q(A)$ and $q(B)$ are substantially different, it provides a strong case to falsify Model I. Similarly, (C, D) could falsify Model II. The outcome of the test falls in one of the three cases. In the first case, one model is falsified and the other is not; a clear winner is obtained. In the second case, no model is falsified, indicating that the two models have strong agreement with each other and cannot be differentiated. In the third case, both models are falsified. Although there is no clear winner, the results may help us identify model problems and suggest ways to combine them into a single better model.

2.2 gMAD Competition Method

We summarize the gMAD competition procedure below.

- **Step 1**. Apply all M models to all samples in O to create a model prediction matrix $\mathbf{P} \in \mathbb{R}^{M \times N}$, where the entry p_{ij} is the prediction of $q(s_j)$ given by C_i.
- **Step 2**. Choose C_1 as the defender ($i = 1$). The rest $M - 1$ models are the attackers.
- **Step 3**. Divide the samples into K bins based on $\mathbf{p}_i$ (the i-th row), indexed by $k \in \{1, 2, \cdots, K\}$. Initialize k to 1.
- **Step 4**. Group the samples in the k-th bin to a subset O_{ik}, which are considered to have similar responses of the defender C_i.
- **Step 5**. Choose C_j ($j \neq i$) as the current attacker.
- **Step 6**. Within O_{ik}, find a pair of samples (s^l_{ijk}, s^u_{ijk}) that correspond to the minimal and maximal responses of C_j. This extremal pair is referred to as the gMAD counterexample suggested by the attacker C_j, attempting to falsify the defender C_i at the level k.
- **Step 7**. Choose another model C_j as the attacker and repeat Step 6 until all $M - 1$ attackers are exhausted.
- **Step 8**. Set $k = k + 1$ and repeat Steps 4-7 until all levels are exhausted ($k = K$).
- **Step 9**. Choose the next model C_i as the defender by setting $i = i + 1$ and repeat Steps 3-8 until all models are exhausted ($i = M$).
- **Step 10**. Perform physical measurements on the selected gMAD sample pairs.
  - **Case 1**. Record $2M(M-1)K$ physical measurements for all $M(M-1)K$ sample pairs.
  - **Case 2**. For the defender C_i and the attacker C_j, find the pair $(s^l_{ij}, s^u_{ij}) = (s^l_{ijk^\star}, s^u_{ijk^\star})$, where

$$k^\star = \underset{k \in \{1,2,\cdots,K\}}{\arg\max} \left| q(s^u_{ijk}) - q(s^l_{ijk}) \right|. \quad (1)$$

    Record $2M(M-1)$ physical measurements for $M(M-1)$ extremal pairs.

- **Step 11**. Conduct statistical analysis (Section 2.3) on the physical measurements for model comparison.

In Step 10, two physical measurement processes are presented. In the first case, the defender is attacked by every attacker at every response level, while in the second case, the defender is attacked once by each attacker at the most differentiable response level. The specific usage of the two cases is application dependent.

2.3 Data Analysis Method

Each gMAD sample pair is associated with two models. We first compare the models in pairs and aggregate the pairwise statistics into a global ranking via rank aggregation tools [15]. We introduce the notions of *aggressiveness* and *resistance*. The aggressiveness a_{ij} measures how aggressive the attacker model C_i is at falsifying the defender model C_j and is computed by

$$a_{ij} = \frac{\sum_{k=1}^{K} w_{jk}\bar{q}_{ijk}}{\sum_{k=1}^{K} w_{jk}}, \tag{2}$$

where $\bar{q}_{ijk}$ describes the preference to the sample s^u_{ijk} over s^l_{ijk} selected from the k-th subset with C_i and C_j being the attacker and the defender, respectively. $\bar{q}_{ijk}$ is obtained through the physical measurements and the specific approach could be application dependent (examples given in Sections 3 and 4). A higher $\bar{q}_{ijk}$ suggests $q(s^u_{ijk})$ is clearly larger than $q(s^l_{ijk})$ and vice versa. When $\bar{q}_{ijk}$ is close to 0, $q(s^u_{ijk})$ and $q(s^l_{ijk})$ are difficult to differentiate. w_{jk} is the number of samples in the k-th subset, acting as a weight factor. a_{ij} is expected to be non-negative with a larger value indicating stronger aggressiveness of C_i over C_j. However, it may be negative in theory, meaning that the order of the sample pair selected by C_i contradicts the physical measurements. In other words, a negative a_{ij} reveals a strong failure case of C_i. The pairwise aggressiveness statistics of all models form an aggressiveness matrix $\mathbf{A}$.

The resistance r_{ij} measures how resistant the defender model C_i is to be defeated by the attacker model C_j and is computed by

$$r_{ij} = \frac{\sum_{k=1}^{K} w_{ik}(1 - |\bar{q}_{jik}|)}{\sum_{k=1}^{K} w_{ik}}. \tag{3}$$

A higher r_{ij} indicates stronger resistance of C_i against C_j. The pairwise resistance statistics of all models form a resistance matrix $\mathbf{R}$.

The pairwise comparison results may be aggregated into a global ranking via the maximum likelihood method for multiple options [15]. Let $\boldsymbol{\mu}^x = [\mu^x_1, \mu^x_2, \cdots, \mu^x_M] \in \mathbb{R}^M$ be the global ranking score vector, where $x \in \{a, r\}$. We maximize the log-likelihood of $\boldsymbol{\mu}^x$

$$\begin{aligned} \arg\max_{\boldsymbol{\mu}^x} \quad & \sum_{ij} x_{ij} \log\left(\Phi(\mu^x_i - \mu^x_j)\right) \\ \text{subject to} \quad & \sum_i \mu^x_i = 0, \end{aligned} \tag{4}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF). The constraint $\sum_i \mu^x_i = 0$ is added to resolve the translation ambiguity. Other constraints such as setting the first score to zero $\mu^x_1 = 0$ are also applicable. The optimization problem in (4) is a convex one and enjoys efficient solvers. When $M = 2$, the maximum likelihood estimate reduces to the Thurstone's law [16] and has a closed form solution (assuming $\mu^x_1 + \mu^x_2 = 0$)

$$\mu^x_1 = -\mu^x_2 = \Phi^{-1}\left(\frac{x_{12}}{x_{12} + x_{21}}\right), \tag{5}$$

where $\Phi^{-1}(\cdot)$ is the inverse CDF of the standard normal. The pairwise resistance statistics can be aggregated in a similar fashion. Other ranking aggregation algorithms such as hodgeRank [17] and ranking by eigenvectors [18] may also be applied.

The aggregated aggressiveness and resistance measures μ^a_i and μ^r_i represent two different aspects of the model competitiveness. μ^a_i summarizes the success of a model as an attacker. A larger μ^a_i means the model is better at finding test samples to falsify other models. μ^r_i describes the success of a model as a defender. A larger μ^r_i means the model is more difficult for other models to find failure cases. μ^a_i or μ^r_i does not have theoretical advantage over one another and both measures are useful. In practice, a model of stronger aggressiveness presumably should also have stronger resistance, but in theory, they are not necessarily correlated. It would be interesting to observe the cases when μ^a_i and μ^r_i disagree. One such example is when a model offers highly accurate predictions (consistent rankings) on most samples in the sample space, but performs poorly on a small percentage of corner cases, where the competing models perform well. In this case, the model has strong aggressiveness to attack other models (using the samples it predicts accurately), but is vulnerable as a defender (being easily defeated by other models using the corner cases). Therefore, the disparity between μ^a_i and μ^r_i of a model is highly insightful to reveal the defects of a generally good model.

2.4 Discussion

The toy example in Section 2.1 is a simplified demonstration of the gMAD competition. In real-world applications, the samples could live in a much higher dimensional space, the number of models under competition could be much larger, and the domain of interest in the sample space could vary according to specific applications. In summary, we mention several useful features of the gMAD competition. First, it is straightforward and flexible to apply gMAD to sample sets tuned for specific applications. Second, the number of sample pairs for physical measurements depends on the model number only and is independent of the sample size. As a result, gMAD is an ideal fit in the applications, where the physical measurement is expensive but the computational prediction is cheap. In such scenarios, gMAD encourages to expand the sample set to cover as many cases as possible. Third, each gMAD sample pair is associated with two models. The defender believes that the pair would produce the same response q while the attacker suggests that they are very different. Fourth, it is cost-effective to add new models to the competition. No change is necessary for the current gMAD pairs. The only additional work is to select a total of $2MK$ (Case 1 of Step 10) or $2M$ (Case 2 of Step 10) new sample pairs for physical measurements.

3 Application to IQA Models

In this section, we apply the gMAD competition to computational models of perceived image quality [20].

3.1 Background

Digital images undergo many transformations in their lifetime [21], any of which may introduce distortions, resulting

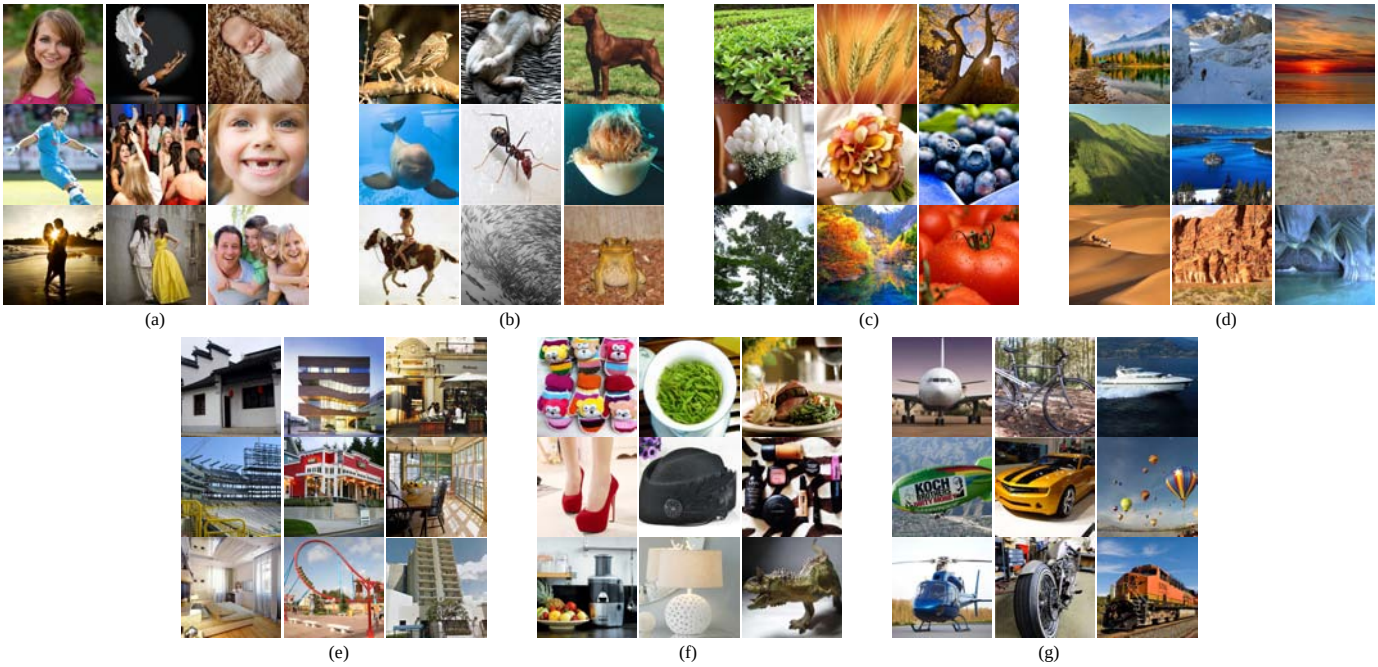


Fig. 2. Sample images in [19]. (a) Human. (b) Animal. (c) Plant. (d) Landscape. (e) Cityscape. (f) Still-life. (g) Transportation.

in visual quality degradation [3]. Being able to automatically predict the perceived image quality by humans is of fundamental importance in image processing and computer vision. Depending on the availability of a distortion-free reference image, computational IQA models may be categorized into full-reference (FR), reduced-reference (RR) and no-reference (NR) methods, where the reference image is fully, partially, and completely not accessible, respectively.

Depending on how test images are presented to human subjects, subjective testing for collecting ground-truth image quality measurements may be roughly classified into three categories: the single-stimulus method, the paired comparison method, and the multiple-stimulus method [22]. In a single-stimulus experiment, one test image is shown at a time and is given ratings of image quality independently. In a paired comparison experiment, a pair of images are shown simultaneously and the subjects are asked which image has better quality. In a multiple-stimulus experiment, multiple images are shown and the subjects rate them based on their perceptual quality. Given n test images, $\mathcal{O}(n)$ evaluations are needed for single-stimulus and multiple-stimulus methods, and $\mathcal{O}(n^2)$ for paired comparison. Although the paired comparison method is often preferred to collect reliable subjective measurements, an exhaustive paired comparison is impractical when n is large. Many methods have been proposed to improve its efficiency. Four types of balanced subset designs were developed in the 1950's [23], among which the square design became popular. An alternative method was to randomly select a small subset of image pairs. It was shown that at least $\mathcal{O}(n \log n)$ distinct pairs are necessary for large random graphs to guarantee the graph connectivity and to achieve a robust global ranking [17]. In [24], a Swiss competition principle was adopted with a decreased complexity of $\mathcal{O}(n \log n)$.

Computational IQA models are typically tested using conventional direct model comparison methods on existing small-scale image quality databases (*e.g.*, LIVE [25] and TID2013 [24]). The goodness of fit is usually measured by the correlation between subjective mean opinion scores (MOS) and objective model predictions. As previously discussed, only a few thousand images can be evaluated by humans due to the limited scale of affordable subjective testing. Moreover, given the combination of reference images, distortion types, and distortion levels, only a few dozen reference images may be included. It is difficult to justify how the few reference images can provide a sufficient representation of real-world content variations. In addition, state-of-the-art IQA models often involve supervised learning or manual parameter adjustments to boost the performance on existing databases. Therefore, it is questionable whether the reported competitive performance can be generalized to the real-world images with much richer content variations and quality degradations.

Wang and Simoncelli adopted MAD [12], [26] to compare two FR-IQA models—the MSE and the structural similarity (SSIM) index [27], and showed that MSE is more easily falsified by SSIM [12]. MAD relies on gradient computation in a constrained optimization process to synthesize test images and is not applicable to advanced IQA models, which are often non-differentiable. Recently, Ma *et al.* introduced three evaluation criteria [19], namely the pristine/distorted image discriminability test (D-Test), the listwise ranking consistency test (L-Test), and the pairwise preference consistency test (P-Test) for IQA models, which do not call for subjective testing. However, the preparations of these tests require reference images and degradation specifications [19].

3.2 Experimental Setup

3.2.1 Database

We choose the Waterloo Exploration Database [19] to constitute the test sample set. It contains $4,744$ high-quality natural images and spans a great deal of image content, including human, animal, plant, landscape, cityscape, still-life, and transportation. Sample images are shown in Fig. 2. Four distortion types—JPEG and JPEG2000 compression, white Gaussian noise contamination, and Gaussian blur—each with five distortion levels are used to generate $94,880$ distorted images. As a result, the Exploration database contains a total of $99,624$ images, which is currently the largest one used by the IQA community. The database focuses on the four aforementioned distortion types because many state-of-the-art IQA models declare themselves for successfully handling them [28]–[32] on small-scale IQA databases. Whether these models survive from the gMAD competition on the Exploration database provides strong evidence of their generalizability in the real world.

3.2.2 Computational IQA Models

A total of 16 computational IQA models are selected to participate in the gMAD competition to cover a wide variety of IQA methodologies with emphasis on NR models. These include FR models 1) PSNR, 2) SSIM [27], 3) MS-SSIM [33], 4) FSIM [34], and NR models 5) BIQI [28], 6) BLINDS II [35], 7) BRISQUE [36], 8) CORNIA [29], 9) DIIVINE [37], 10) IL-NIQE [38], 11) LPSI [39], 12) M3 [40], 13) NFERM [41], 14) NIQE [30], 15) QAC [42] and 16) TCLT [43]. The gradients of most models are extremely difficult to compute or approximate, therefore limiting the pairwise comparison using MAD [12]. The implementations of all models are obtained from the original authors. For IQA models that involve training, we use all images in the LIVE database [25]. To make a consistent comparison, we adopt a logistic nonlinear function to map all model predictions into the same perceptual scale $[0, 100]$ with a higher value indicating better perceptual quality.

We define six quality levels ($K = 6$) evenly spaced on the quality scale with a good coverage from low- to high-quality. The quality range for each level is one standard deviation (std) of MOSs in LIVE [25] so as to guarantee that the images in the same level have similar quality by the defender model. The attacker models then search for gMAD image pairs from the six levels, as described in Section 2. On the scatter plot, finding a gMAD image pair corresponds to selecting points that have the longest distance in a given row or column, as exemplified in Fig. 3, where SSIM [27] competes with MS-SSIM [33]. The corresponding image pairs are shown in Fig. 4, from which we may obtain a first impression on their relative performance in gMAD. Specifically, the images in the first row of Fig. 4 exhibit approximately the same perceptual quality (in agreement with MS-SSIM [33]) and those in the second row have drastically different perceptual quality (in disagreement with SSIM [27]). This suggests that MS-SSIM is a solid improvement over SSIM. In the end, a total of $16 \times (16-1) \times 6 = 1,440$ gMAD image pairs are chosen for the subsequent subjective experiment.

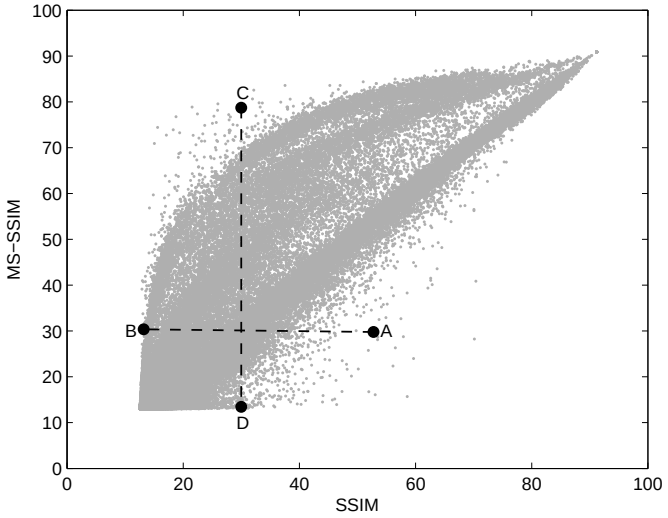


Fig. 3. gMAD image pairs from the Waterloo Exploration Database [19]. The image pair (A, B) is selected by maximizing/minimizing SSIM while holding MS-SSIM fixed. Similarly, (C, D) is selected by maximizing/minimizing MS-SSIM while holding SSIM fixed.

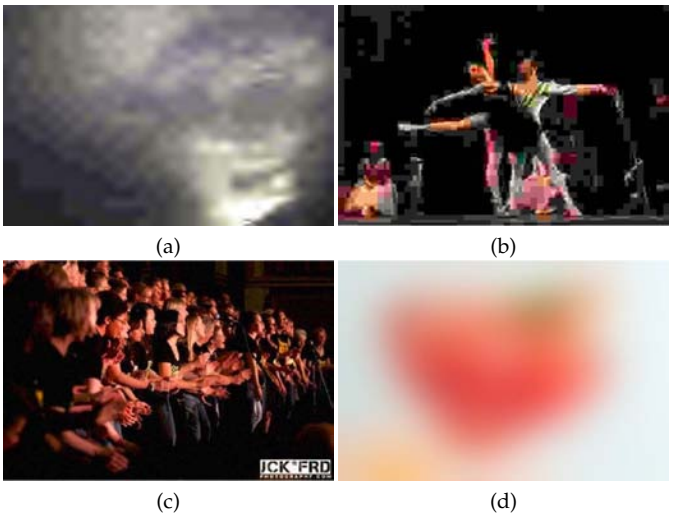


Fig. 4. gMAD image pairs between SSIM and MS-SSIM. Images (a)-(d) correspond to points A-D in Fig. 3. (a) MS-SSIM = 30, SSIM = 53. (b) MS-SSIM = 30, SSIM = 13. (c) SSIM = 30, MS-SSIM = 78. (d) SSIM = 30, MS-SSIM = 13.

3.3 Subjective Testing

A subjective user study is conducted in an office environment with a normal indoor illumination level. The display is a true-color LCD monitor at a resolution of $2,560 \times 1,600$ pixels and is calibrated in accordance with the recommendations of ITU-R BT.500 [22]. A customized MATLAB interface is created to render an image pair at their exact pixel resolutions but in random spatial order. A scale-and-slider applet is used for assigning a quality score, as shown in Fig. 5. A total of 31 naïve human subjects (16 males and 15 females) of age 22 to 30, participate in the subjective experiment. All subjects have a normal or correct-to-normal visual acuity. Sample image pairs (independent of the test pairs) are shown to the subjects in a training session to familiarize them with image distortions and the experimental procedure. For each gMAD image pair, the

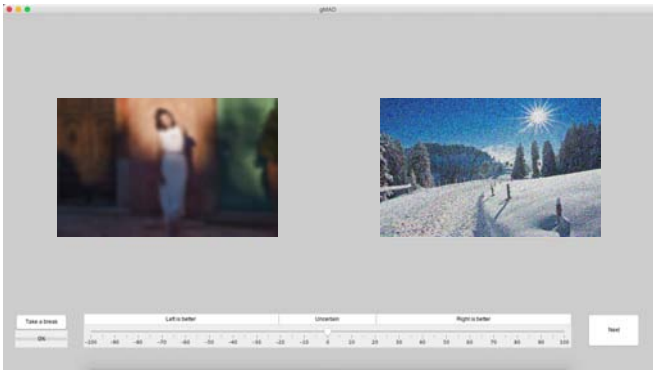


Fig. 5. User interface for subjective testing.

subjects can assign a score between -100 and 100 to indicate their preference to either the left image $[-100, -20]$ (labeled as "left is better") or the right image $[20, 100]$ (labeled as "right is better"). When the subjects are uncertain about their decision, they can also assign a score between $[-20, 20]$ (labeled as "uncertain"), where a score of zero indicates completely neutral. The proposed soft version of the paired comparison method better captures the subjects' confidence when expressing their preference. We divide the experiment into four sessions, each of which is limited to a maximum of 30 minutes. The subjects are asked to take a five-minute break to minimize the influence of the fatigue effect. All subjects participate in all sessions.

3.4 Data Analysis

We adopt the outlier detection and subject rejection algorithm suggested in [22] to screen the raw subjective data. Specifically, a score for an image pair is considered to be an outlier if it is outside two stds for the Gaussian case or outside $\sqrt{20}$ stds for the non-Gaussian case. A subject is removed if more than 5% of his/her evaluations are outliers. As a result, one subject is rejected. Among all scores given by the valid subjects, about 1.4% of them are identified as outliers and are removed subsequently.

We average the subjective measurements of each gMAD image pair and compute the pairwise aggressiveness and resistance statistics for every pair of 16 IQA models. Fig. 6 shows the aggressiveness matrix $\mathbf{A}$ and the resistance matrix $\mathbf{R}$, where the higher value of an entry (warmer color), the stronger aggressiveness or resistance of the corresponding row model against the column model.

We aggregate the pairwise comparison results into a global ranking via the maximum likelihood method for multiple options. Fig. 7 shows the results, from which we have several interesting observations. First, an IQA model with stronger aggressiveness generally exhibits stronger resistance. Second, FR-IQA models are generally better than NR-IQA ones, which is not surprising because FR models make use of reference images. Third, the best performance is obtained by MS-SSIM [33], which is a multi-scale version of SSIM [27] and a significant improvement upon it. This suggests that multi-scale analysis is beneficial to IQA. Fourth, CORNIA [29], NIQE [30], and ILNIQE [38] perform the best among all NR-IQA models. They are derived from

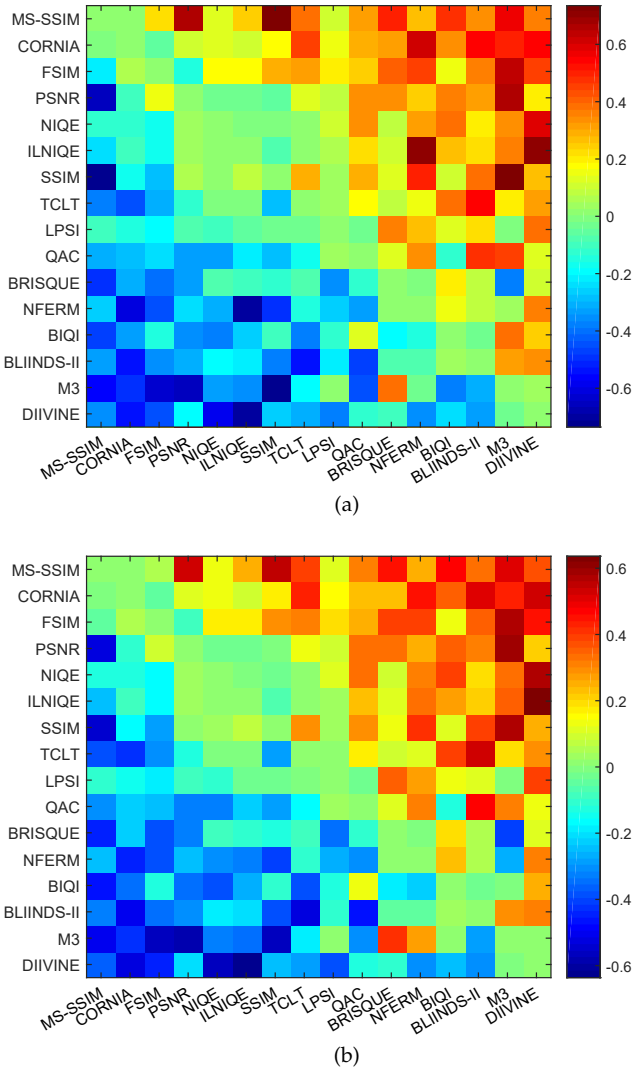


Fig. 6. Pairwise comparison results of the 16 IQA models. (a) Aggressiveness matrix. (b) Resistance matrix. Each entry indicates the aggressiveness/resistance of the row model against the column model. $\mathbf{A} - \mathbf{A}^T$ and $\mathbf{R} - \mathbf{R}^T$ are drawn here for better visibility.

perception- and distortion-relevant natural scene statistics, which map raw images into a perceptually meaningful space for comparison. Finally, machine learning-based IQA models, though outstanding on small-scale IQA databases, generally do not perform well in the current gMAD competition. This may be because the training samples are not sufficient to represent the population of real-world natural images and thus the risk of overfitting is high.

3.5 Further Testing

The conventional direct model comparison method and the proposed gMAD competition test different aspects of computational models. The former evaluates the overall goodness of fit, while the latter focuses on falsifying a model in the most efficient way. They are complementary and are not intended to replace one with the other. It is interesting to observe how much they align with each other in terms of model ranking performance in real-world applications. Presumably, a model that is outstanding in

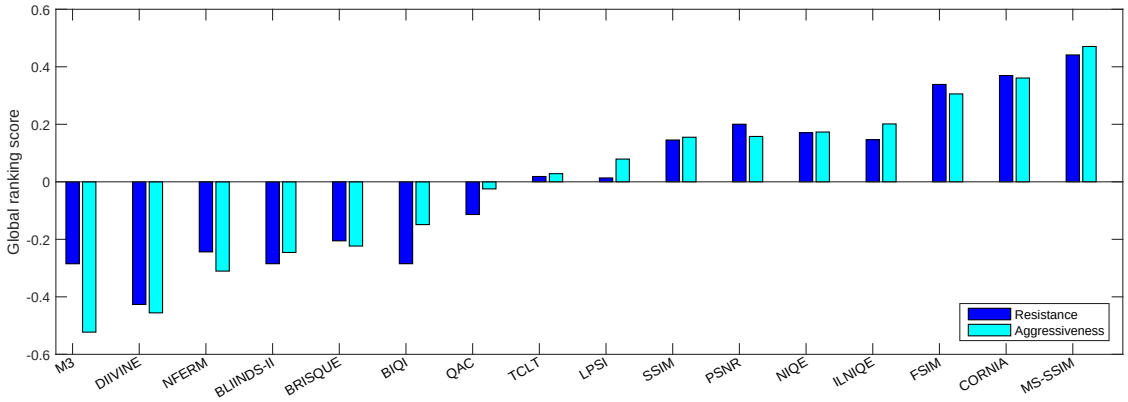


Fig. 7. Global ranking results of the 16 IQA models.

TABLE 1
SRCC results between $K = 6$ as the reference and other K values

SRCC	Aggressiveness	Resistance
$K = 1$	0.930	0.885
$K = 2$	0.929	0.906
$K = 3$	0.965	0.968
$K = 4$	0.982	0.985
$K = 5$	0.997	0.985

one testing methodology is likely to do well in another. Here, we test IQA models on a small subject-rated database, to which both model comparison methods are applicable, and compare their ranking results. Specifically, we choose the CSIQ [44] database, which contains 30 pristine-quality and 866 distorted images with six distortion types and five distortion levels. Each image is associated with a MOS, which spans the range $[0, 1]$ with 1 indicating the worst perceptual quality. We test the 16 IQA models using gMAD and the MSE between model predictions and MOSs (a conventional direct model comparison method). The subjective measurement $\bar{q}_{ijk}$ of a gMAD image pair is computed by the MOS difference between the two images.

We compare the global ranking results by the aggressiveness/resistance of gMAD and MSE, and find that they are well correlated (an SRCC of $0.953/0.941$), suggesting general agreement between gMAD and the direct model comparison method on CSIQ [44]. We also compare the aggressiveness/resistance ranking on CSIQ with that on the Waterloo Exploration Database [19] and observe an SRCC of $0.618/0.726$. The performance discrepancy may be explained by the large difference between the two databases [5] in terms of their sizes and content variations. In general, gMAD benefits from larger image sets of greater diversity, which makes it easier to falsify an IQA model and to differentiate similar models.

To investigate the impact of the number of quality levels K on the global ranking, we experiment with $K \in \{1, 2, 3, 4, 5, 6\}$. For small K values, we are only able to compare IQA models on certain quality levels. For example, for $K = 1$, we choose the low-quality level to search for gMAD image pairs. We adopt the global ranking with $K = 6$ as the reference (Fig. 7) and compute the SRCC with other K values. The results are shown in Table 1, where we see that K has little impact on the global ranking. The robustness of the gMAD ranking with respect to K in our experiment may be because a less competitive IQA model tends to perform poorly at all quality levels. As a result, the results at limited quality levels are representative for the full quality range.

4 More Applications

The application scope of gMAD is broad in the sense that it can be used to compare any group of computational models that predict certain physical quantities. In this section, we demonstrate the gMAD competition methodology with two more examples of perceptually discriminable quantities—image aesthetics and streaming video QoE.

4.1 Comparison of Image Aesthetics Models

Image aesthetics refers to the experience of beauty for subjects when viewing an image [50]. It is generally believed that image aesthetics is determined by a combination of low-level features such as composition, lighting, color arrangement and camera settings, and high-level semantics such as simplicity, realism, content type and topic emphasis [51]. A successful computational image aesthetics model plays an important role in many fields such as image editing, image retrieval, and personal photo management.

Computational image aesthetics assessment is not an easy task. Most existing image aesthetics models only make a binary decision on whether an image is a high-quality professional photo or a low-quality snapshot [13], [50]. Consequently, those models can only be tested on subject-rated image aesthetics databases with binary annotations [51]. In practice, the perceived aesthetics of real-world images is much more diverse than just two levels, and thus continuous-valued models are highly desirable.

Here we aim to apply gMAD to compare continuous-valued aesthetics models. We first randomly select more than $170,000$ images from ImageNet [45] as the test sample set, whose content and aesthetics levels are diverse. Sample images are shown in Fig. 8. We select four image aesthetics models, including GIST+SVR [46], aesthetics-aware features with SVR (AAF+SVR) [47], Jin16 [49], and Kong16 [48]. We

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

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

Fig. 8. Sample images from ImageNet [45] used for the gMAD competition of image aesthetics models. (a)-(h) Images with increasing degrees of perceived aesthetics according to our subjective testing.

TABLE 2
Pairwise comparison results of image aesthetics models in the gMAD competition. Row model: attacker. Column model: defender

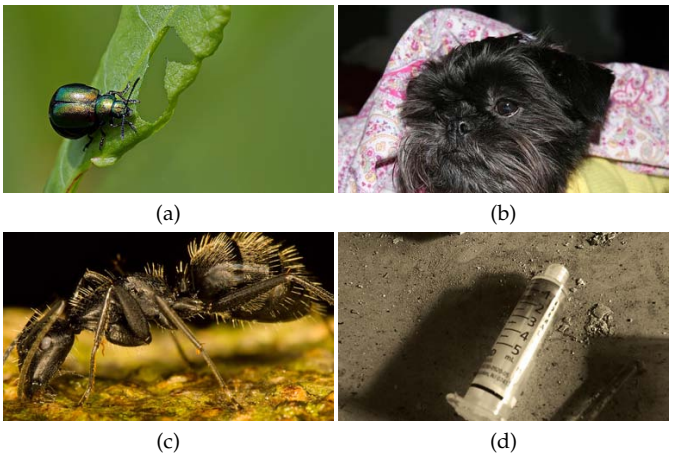
Aesthetics model	Aggressiveness				Resistance			
	GIST+SVR [46]	AAF+SVR [47]	Kong16 [48]	Jin16 [49]	GIST+SVR [46]	AAF+SVR [47]	Kong16 [47]	Jin16 [49]
GIST+SVR	—	0.216	0.103	0.031	—	0.686	0.713	0.541
AAF+SVR	0.314	—	0.182	0.160	0.662	—	0.708	0.534
Kong16	0.287	0.292	—	0.299	0.741	0.648	—	0.422
Jin16	0.459	0.466	0.578	—	0.934	0.810	0.701	—

implement GIST+SVR and AAF+SVR algorithms by ourselves, and the program packages of the other two models are obtained from the original authors. For GIST [46], we work with five scales, four orientations and 16 blocks, and process RGB channels separately, resulting in a total of $5 \times 4 \times 16 \times 3 = 960$ features per image. Linear SVR [52] is adopted with hyperparameters optimized for the best prediction. For AAF, we choose the $1,323$-dimensional features proposed by Mavridaki and Mezaris [47], who implement a set of generally accepted photographic rules such as simplicity, colorfulness, sharpness, image pattern, and composition. Linear SVR with the same hyperparameter optimization strategy is adopted. Jin16 [49] is a convolutional neural network (CNN)-based algorithm that inherits the VGG16 [53] architecture and fine-tunes the weights using a weighted MSE loss. Kong16 [48] is another CNN-based model that fine-tunes the weights from AlexNet [54] using a weighted sum of a regression loss, a pairwise ranking loss, and an attribute loss. We train and validate GIST+SVR and AFF+SVR on AVA [55]. The weights of Jin16 [49] and Kong16 [48] fine-tuned on AVA [55] and AADB [48], respectively, are used for testing. Finally, we use the Waterloo IAA Database [56] to map all model predictions into the same perceptual space for comparison.

We choose three aesthetics levels and generate $4 \times 3 \times 3 = 36$ gMAD image pairs. The subjective testing procedure is similar to that described in Section 3.3 and we highlight the differences here. 30 subjects, 18 males and 12 females, participate in the experiment. Each subject takes about ten minutes to finish rating all the pairs. After running the outlier detection and subject rejection algorithm, all subjects are valid and 2.1% of the subjective measurements are outliers.

TABLE 3
Global ranking results of image aesthetics models in gMAD

Aesthetics model	Aggressiveness	Resistance
GIST+SVR [46]	-0.577	-0.097
AAF+SVR [47]	-0.189	-0.064
Kong16 [48]	0.145	-0.098
Jin16 [49]	**0.621**	**0.260**



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

Fig. 9. gMAD competition between Jin16 [49] and Kong16 [48] at the high-aesthetics level. (a) Best Jin16 for fixed Kong16. (b) Worst Jin16 for fixed Kong16. (c) Best Kong16 for fixed Jin16. (d) Worst Kong16 for fixed Jin16.

We list the pairwise and global ranking results of the four image aesthetics models in terms of aggressiveness and

Fig. 10. Sample frames from the proposed streaming video database for the gMAD competition of QoE models. (a) YellowStone: natural, high motion. (b) StreetDance: outdoor, high motion. (c) SplitTrailer: human, high motion. (d) CSGO: animation, high motion. (e) UCLY: indoor, slow motion. (f) WildAnimal: animal, slow motion. (g) Rose: plant, slow motion. (h) Food: still-life, slow motion.

TABLE 4
Encoding ladder of video clips. kbps: kB per second

Representation	Bitrate (kbps)	Resolution
Bad	235	320×240
Poor	560	512×384
Fair	$1,050$	640×480
Good	$2,350$	$1,280 \times 720$
Excellent	$5,800$	$1,920 \times 1,080$

resistance in Tables 2 and 3, respectively. It can be observed that Jin16 [49], a CNN-based model, exhibits the strongest aggressiveness and resistance. To take a closer look, we show two gMAD image pairs, where Jin16 competes with Kong16 [48] at the high-aesthetics level in Fig. 9. It is clear that Jin16 successfully falsifies Kong16 by finding the image pair in the first row, where image (a) looks more beautiful than image (b) for most subjects. Meanwhile, Jin16 survives from the attack by Kong16 as evidenced by similar aesthetics of the image pair in the second row according to our subjective testing. We conjecture that the superiority of Jin16 over Kong16 arises because 1) the backbone of Jin16—VGG16 [53]—might be easier to generalize to novel tasks than AlexNet [54] used in Kong16; 2) the weighted loss that offsets the aesthetics level imbalance in Jin16 has more potentials to improve the performance than adding the pairwise ranking and attribute losses in Kong16. Moreover, it is not surprising that the general-purpose feature representation GIST [46] for holistic scene modeling is defeated by AAF [47] under the same training configuration. After all, the AAF representation is motivated by years of practices of professional photographers and thus is more relevant to image aesthetics. Finally, the hand-crafted AAF representation is slightly better in terms of resistance than the end-to-end optimized Kong16 [48]. This suggests that more training data, novel network architectures, and advanced optimization techniques are needed to learn more robust end-to-end aesthetics models.

4.2 Comparison of Streaming Video QoE Models

Video streaming services have gained increasing popularity due to the fast deployment of network infrastructures and the proliferation of smart mobile devices. Being able to predict the QoE of end users is of great importance because it plays a critical role in the user choices of video streaming services [57]. Three major factors affect the QoE for HTTP adaptive streaming (HAS) [58], [59]. The first is the presentation quality of video segments encoded in different bitrates, spatial resolutions, and frame rates. The second is the stalling events due to poor or unstable network conditions, characterized by their frequencies and time durations. The third is the switchings of video segments of different bitrates, spatial resolutions, and frame rates from one time segment to another, adapting to varying network conditions. Developing computational QoE models that jointly consider these factors and their interactions is a challenging task. In recent years, many QoE models have been developed [14], [60], but most of them have not been calibrated against subjective data with sufficient video content variations and distortion types. Note that the largest subject-rated streaming video database so far only contains hundreds of videos [61].

We build a large-scale streaming video database as the playground for the gMAD competition of computational QoE models. Specifically, we first download 50 high-quality 4K videos with 24-30 frames per second (fps) from the Internet, which carry a Creative Commons license. We downsample all videos to $1,920 \times 1,080$ to further damp possible compression artifacts. They are selected to cover sufficient content variations and motion patterns. Sample frames of representative videos are shown in Fig. 10. From each video we extract a ten-second video clip, which is further divided into five non-overlapping two-second segments. Each segment is encoded using H.264 into five representations selected from the Netflix's encoding ladder [62], representing "bad", "poor", "fair", "good", and "excellent" presentation quality, respectively. The details of the encoding ladder are given in Table 4. After that, we prepend a stalling event to each encoded segment with a time duration of zero, two, or four seconds, representing "no", "short", and

"long" stalling, respectively. We concatenate all possible combinations of two-second segments from the same source content along with the stalling events, resulting in a total of $3^5 \times 5^5 \times 50 = 37,968,750$ test video clips.

We let three computational QoE models play the gMAD game. These are Liu12 [63], Yin15 [64], and SQI [58]. Liu12 [63] adopts the bitrate and the stalling percentage as two features. On top of Liu12, Yin15 [64] adds two more features—the switching magnitude and the initial buffering duration. Linear regression is used for the two models. Instead of using the bitrate as the indication of presentation quality, SQI [58] resorts to advanced video quality models such as SSIMplus [65] to predict presentation quality and considers the interactions between video presentation quality and playback stalling experiences. We make use of the Waterloo QoE Database [58] and map all model responses to the same perceptual scale.

We choose three QoE levels and generate $3 \times 2 \times 3 = 18$ gMAD video pairs. 30 subjects participate in the subjective experiment. Two video clips in the same pair are played consecutively but in random order. Subjects are allowed to replay them until they are confident about their judgment on the relative QoE for the two video clips. Each subject takes about 20 minutes to finish the experiment. After subjective data screening, no subject is rejected and 3.0% of the subjective measurements are identified as outliers.

The pairwise and global ranking results of Liu12 [63], Yin15 [64], and SQI [58] are listed in Tables 5 and 6, respectively. It appears that SQI outperforms the other two QoE models in terms of both aggressiveness and resistance. We also show the gMAD video pairs between SQI and Yin15 in Fig. 11, where we find that SQI defeats Yin15 at all QoE levels. For example, at the mid-QoE level in Fig. 11(a), Yin15 regards a video sequence of smooth playout (small quality oscillation between excellent and good quality without any stalling) to have similar QoE to a video sequence significantly interrupted by multiple stalling events. The better performance of SQI may arise from that SQI replaces the bitrate with SSIMplus [65], which is a human visual system inspired model and is in close agreement with human perception of presentation quality. Taking into account the interactions between presentation quality and stalling events may be another important ingredient for SQI to win the competition. However, SQI does not consider the quality switching effect to the overall QoE. We believe a joint modeling of video presentation quality, stalling events, and switchings is a potential direction to further improve SQI. Compared to Liu12, Yin15 adds two more features, attempting to model the switching and initial buffering effects. Unfortunately, we observe a performance degradation through the gMAD competition. This may be because Yin15 captures the switching effect with an oversimplified measure—the absolute difference between the bitrates of two consecutive video segments, which may in turn hamper the overall performance. Specifically, the bitrate and its difference exhibit a strong nonlinearity and (possibly non-monotonicity) to the overall QoE. Incorporating it into the model linearly may not be an appropriate choice. In addition, the results in [59] show that users have clearly different behaviors when experiencing positive and negative adaptations. In other words, the switching direction matters, but the absolute operation in Yin15 ignores such information. In summary, modeling user experience when viewing streaming videos is challenging and the current models only work to some degrees. A complete treatment of the aforementioned three factors is desirable to better predict streaming video QoE.

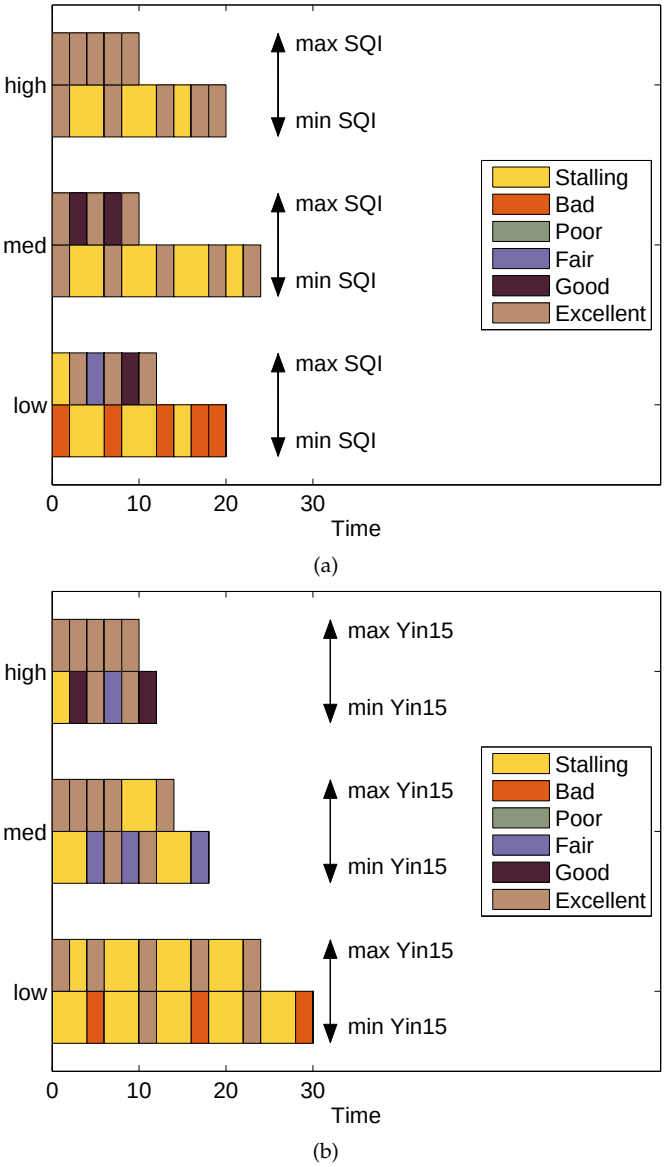


Fig. 11. gMAD competition between Yin15 [64] and SQI [58]. (a) Yin15 as defender. (b) SQI as defender. Video playout sequences at different video quality levels (as listed in Table 4) and the stalling events are represented using different color bars. The gMAD competitions are performed at high-, mid-, and low-QoE levels, respectively.

TABLE 6
Global ranking results of QoE models in the gMAD competition

QoE model	Aggressiveness	Resistance
Liu12 [63]	−0.106	0.010
Yin15 [64]	−0.161	−0.112
SQI [58]	**0.267**	**0.102**

TABLE 5
Pairwise comparison results of QoE models in the gMAD competition. Row model: attacker. Column model: defender

QoE model	Aggressiveness			Resistance		
	Liu12 [63]	Yin15 [64]	SQI [58]	Liu12 [63]	Yin15 [64]	SQI [58]
Liu12 [63]	—	0.000	0.687	—	0.570	0.434
Yin15 [64]	0.430	—	0.077	0.636	—	0.223
SQI [58]	0.566	0.777	—	0.313	0.499	—

5 Conclusion and Discussion

We propose a new methodology, namely the gMAD competition, for efficient comparison of computational predictive models. Aiming for maximizing the speed of falsifying models, gMAD automatically searches from a large-scale sample set for a small number of model-dependent sample pairs. gMAD is particularly useful when the sample space is large, the physical quantity being predicted is expensive to measure, and the model prediction is cheap to compute. Unlike conventional direct model comparison approaches [3]–[5], the number of physical measurements required by the gMAD competition does not scale with the size of the sample space and only depends on the number of competing models. This feature allows gMAD to exploit a sample set of arbitrary size with a low and manageable cost. gMAD also provides two well-defined measures (aggressiveness and resistance) to indicate the relative performance of computational models, through which useful insights may be gained to design better models.

Although the current work demonstrates gMAD using three perceptually discriminable quantities—image quality, image aesthetics, and video QoE—there are a much wider variety of scenarios that gMAD can come into play. To give a few examples, these include comparisons of image/video emotion predictors in the field of cognitive vision [50], the relative attributes (sportiness and furriness) estimators in the field of semantic image search [66], machine translation quality estimators in the field of computational linguistics [67], and thermal comfort models in the field of thermal environment of buildings [68].

The current gMAD requires computational models to produce continuous-valued responses. How to adapt gMAD to account for discrete-valued models has great potentials to impact other computer vision and machine learning applications. For example, instead of building a database larger than ImageNet [45] for testing, it is of great interest to see how existing image classification algorithms behave in a discrete version of gMAD setting. In addition, the current gMAD requires computational models to be scalar-valued, manifesting themselves in predicting a measurable quantity. It is interesting to extend gMAD to work with vector-valued models, for example, to compare different feature representations in a computational vision task [69].

Acknowledgements

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada, the China NSFC Grants (No. 61672446, No. 61831005, No. 61601102, and No. 61525102). The authors would like to thank Dr. Eero Simoncelli for fruitful discussions and insights.

References

[1] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. Morgan & Claypool, 2006.
[2] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media, 2003.
[3] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, Jan. 2009.
[4] G. U. Yule, *An Introduction to the Theory of Statistics*. C. Griffin, Limited, 1919.
[5] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.
[6] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Occam's razor," *Information Processing Letters*, vol. 24, no. 6, pp. 377–380, Apr. 1987.
[7] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, Sep. 1978.
[8] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.
[9] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, 1993.
[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[11] S. De, P. T. Brewick, E. A. Johnson, and S. F. Wojtkiewicz, "Investigation of model falsification using error and likelihood bounds with application to a structural system," *Journal of Engineering Mechanics*, vol. 144, no. 9, pp. 1–15, Jul. 2018.
[12] Z. Wang and E. P. Simoncelli, "Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities," *Journal of Vision*, vol. 8, no. 12, pp. 8.1–8.13, Sep. 2008.
[13] R. Datta, D. Joshi, J. Li, and J. Wang, "Studying aesthetics in photographic images using a computational approach," in *European Conference on Computer Vision*, 2006, pp. 288–301.
[14] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming services," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 20–29, Apr. 2012.
[15] K. Tsukida and M. R. Gupta, "How to analyze paired comparison data," , University of Washington, Tech. Rep. UWEETR-2011-0004, May 2011.
[16] L. L. Thurstone, "A law of comparative judgment," *Psychological Review*, vol. 34, no. 4, pp. 273–286, Jul. 1927.
[17] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye, "Statistical ranking and combinatorial hodge theory," *Mathematical Programming*, vol. 127, no. 1, pp. 203–244, Mar. 2011.
[18] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. SIAM, 2000.
[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*, vol. 22, no. 2, pp. 1004–1016, Feb. 2017.
[20] K. Ma, Q. Wu, Z. Wang, Z. Duanmu, H. Yong, H. Li, and L. Zhang, "Group MAD competition – A new methodology to compare objective image quality models," in *IEEE Conference on Computer Vsion and Pattern Recognition*, 2016, pp. 1664–1673.
[21] A. Rosenfeld, "Picture processing by computer," *ACM Computing Surveys*, vol. 1, no. 3, pp. 147–176, Sep. 1969.
[22] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," 2000. [Online]. Available: http://www.vqeg.org

[23] W. H. Clatworthy, "Partially balanced incomplete block designs with two associate classes and two treatments per block," *Journal of Research of the National Bureau of Standards*, vol. 54, no. 4, pp. 177–190, Apr. 1955.

[24] 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*, vol. 30, pp. 57–77, Jan. 2015.

[25] H. R. Sheikh, Z. Wang, A. C. Bovik, and L. K. Cormack, "Image and video quality assessment research at LIVE." [Online]. Available: http://live.ece.utexas.edu/research/quality/

[26] Z. Wang and E. P. Simoncelli, "Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics," in *Human Vision and Electronic Imaging IX*, 2004, pp. 99–109.

[27] 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*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[28] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513–516, May 2010.

[29] 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*, 2012, pp. 1098–1105.

[30] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[31] 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*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.

[32] 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*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[33] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems and Computers*, 2003, pp. 1398–1402.

[34] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[35] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[36] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[37] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[38] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[39] Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," in *IEEE International Conference on Image Processing*, 2015, pp. 339–343.

[40] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.

[41] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 50–63, Jan. 2015.

[42] 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*, 2013, pp. 995–1002.

[43] Q. Wu, H. Li, F. Meng, K. N. Ngan, B. Luo, C. Huang, and B. Zeng, "Blind image quality assessment based on multi-channel features fusion and label transfer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 425–440, Mar. 2016.

[44] 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*, vol. 19, no. 1, pp. 1–21, Jan. 2010.

[45] 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*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[46] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, May 2001.

[47] E. Mavridaki and V. Mezaris, "A comprehensive aesthetic quality assessment method for natural images using basic rules of photography," in *IEEE International Conference on Image Processing*, 2015, pp. 887–891.

[48] 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*, 2016, pp. 662–679.

[49] B. Jin, M. V. O. Segovia, and S. Süsstrunk, "Image aesthetic predictors based on weighted CNNs," in *IEEE International Conference on Image Processing*, 2016, pp. 2291–2295.

[50] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, Sep. 2011.

[51] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1930–1943, Dec. 2013.

[52] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004.

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

[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[55] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2408–2415.

[56] W. Liu and Z. Wang, "A database for perceptual evaluation of image aesthetics," in *IEEE International Conference on Image Processing*, 2017, pp. 1317–1321.

[57] Cisco IBSG Youth Focus Group, "Cisco IBSG youth survey," Nov. 2010. [Online]. Available: http://www.cisco.com/c/dam/en_us/about/ac79/docs/ppt/Video_Disruption_SP_Strategies_IBSG.pdf

[58] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang, "A quality-of-experience index for streaming video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 154–166, Feb. 2017.

[59] Z. Duanmu, K. Ma, and Z. Wang, "Quality-of-experience of adaptive video streaming: Exploring the space of adaptations," in *ACM Multimedia*, 2017, pp. 1752–1760.

[60] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[61] Z. Duanmu, A. Rehman, and Z. Wang, "A quality-of-experience database for adaptive video streaming," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 474–487, Jun. 2018.

[62] Netflix Inc., "Per-title encode optimization," 2015. [Online]. Available: http://techblog.netflix.com/2015/12/per-title-encode-optimization.html

[63] X. Liu, F. Dobrian, H. Milner, J. Jiang, V. Sekar, I. Stoica, and H. Zhang, "A case for a coordinated Internet video control plane," in *ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, 2012, pp. 359–370.

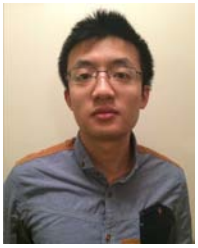
[64] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, "A control-theoretic approach for dynamic adaptive video streaming over HTTP," *ACM SIGCOMM Computer Communication Review*, vol. 45, no. 4, pp. 325–338, Sep. 2015.

[65] A. Rehman, K. Zeng, and Z. Wang, "Display device-adapted video quality-of-experience assessment." in *SPIE Human Vision and Electronic Imaging*, 2015, pp. 1–11.

[66] A. Kovashka, D. Parikh, and K. Grauman, "WhittleSearch: Image search with relative attribute feedback," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2973–2980.

[67] Y. Graham, "Improving evaluation of machine translation quality estimation," in *Association for Computational Linguistics*, 2015, pp. 1804–1813.

[68] J. F. Nicol and M. A. Humphreys, "Adaptive thermal comfort and sustainable thermal standards for buildings," *Energy and Buildings*, vol. 34, no. 6, pp. 563–572, Jul. 2002.
[69] A. Berardino, V. Laparra, J. Ballé, and E. P. Simoncelli, "Eigen-distortions of hierarchical representations," in *Advances in Neural Information Processing Systems*, 2017, pp. 3530–3539.



Kede Ma (S'13-M'18) received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2012, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2014 and 2017, respectively. He is currently a Research Associate with Howard Hughes Medical Institute and New York University, New York, NY, USA. His research interests include computational vision, and computational photography.

Zhengfang Duanmu (S'15) received the B.A.Sc. and M.A.Sc. degrees in electrical and computer engineering from the University of Waterloo in 2015 and 2017, respectively, where he is currently working toward the Ph.D. degree in electrical and computer engineering. His research interests include perceptual image processing and quality-of-experience.

Zhou Wang (S'99-M'02-SM'12-F'14) received the Ph.D. degree from The University of Texas at Austin in 2001. He is currently a Canada Research Chair and Professor in the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include image and video processing and coding; visual quality assessment and optimization; computational vision and pattern analysis; multimedia communications; and biomedical signal processing. He has more than 200 publications in these fields with over 40,000 citations (Google Scholar).

Dr. Wang serves as a Senior Area Editor of *IEEE Transactions on Image Processing* (2015-present). Previously, he served as a member of IEEE Multimedia Signal Processing Technical Committee (2013-2015), an Associate Editor of *IEEE Transactions on Circuits and Systems for Video Technology* (2016-2018), *IEEE Transactions on Image Processing* (2009-2014), *Pattern Recognition* (2006-present) and *IEEE Signal Processing Letters* (2006-2010), and a Guest Editor of *IEEE Journal of Selected Topics in Signal Processing* (2013-2014 and 2007-2009). He is a Fellow of Royal Society of Canada and Canadian Academy of Engineering, and a recipient of 2017 Faculty of Engineering Research Excellence Award at University of Waterloo, 2016 IEEE Signal Processing Society Sustained Impact Paper Award, 2015 Primetime Engineering Emmy Award, 2014 NSERC E.W.R. Steacie Memorial Fellowship Award, 2013 IEEE Signal Processing Magazine Best Paper Award, and 2009 IEEE Signal Processing Society Best Paper Award.

Qingbo Wu (S'12-M'13) received the B.E. degree from the Hebei Normal University in 2009, and the Ph.D. degree from the University of Electronic Science and Technology of China in 2015. He is currently an Associate Professor with the School of Information and Communication Engineering, University of Electronic Science and Technology of China. His research interests include image/video coding, quality evaluation, and perceptual modeling and processing.

Wentao Liu (S'15) received the B.E. and M.E. degrees from Tsinghua University, Beijing, China in 2011 and 2014, respectively. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His research interests include perceptual quality assessment of images and videos.

Hongwei Yong received the B.Sc. and M.Sc. degrees from Xi'an Jiaotong University, Xi'an, China, in 2013 and 2016, respectively. His current research interests include low-rank modeling, background subtraction, and video analysis.

Hongliang Li (SM'12) received the Ph.D. degree in electronics and information engineering from Xi'an Jiaotong University, China, in 2005. He is currently a Professor with the School of Information and Communication Engineering, University of Electronic Science and Technology of China. His research interests include image segmentation, object detection, and visual attention. Dr. Li has authored or co-authored numerous technical articles in international journals and conferences. He is an Associate Editor of *IEEE Transactions on Circuits and Systems for Video Technology* and *Journal on Visual Communications and Image Representation*, and the Area Editor of *Signal Processing: Image Communication*. He served as a Technical Program Chair of IEEE VCIP 2016 and PCM 2017, the General Chair of the ISPACS 2017, and the Local Chair of the IEEE ICME 2014.

Lei Zhang (M'04-SM'14-F'18) received the Ph.D. degree from the Northwestern Polytechnical University, Xi'an, China, in 2001. He is currently a Chair Professor in the Department of Computing, The Hong Kong Polytechnic University. His research interests include computer vision, pattern recognition, image and video processing, and biometrics. He has more than 200 publications in these fields with over 37,000 citations (Google Scholar). Dr. Zhang is a Senior Area Editor of *IEEE Transactions on Image Processing*, an Associate Editor of *SIAM Journal of Imaging Sciences*, and *Image and Vision Computing*. He is a Clarivate Analytics Highly Cited Researcher from 2015 to 2018.