Active Fine-Tuning from gMAD Examples Improves Blind Image Quality Assessment

Zhihua Wang, Student Member, IEEE, and Kede Ma, Member, IEEE

Abstract—The research in image quality assessment (IQA) has a long history, and significant progress has been made by leveraging recent advances in deep neural networks (DNNs). Despite high correlation numbers on existing IQA datasets, DNN-based models may be easily falsified in the group maximum differentiation (gMAD) competition with strong counterexamples being identified. Here we show that gMAD examples can be used to improve blind IQA (BIQA) methods. Specifically, we first pre-train a DNN-based BIQA model using multiple noisy annotators, and fine-tune it on multiple subject-rated databases of synthetically distorted images, resulting in a top-performing baseline model. We then seek pairs of images by comparing the baseline model with a set of full-reference IQA methods in gMAD. The resulting gMAD examples are most likely to reveal the relative weaknesses of the baseline, and suggest potential ways for refinement. We query ground truth quality annotations for the selected images in a well controlled laboratory environment, and further fine-tune the baseline on the combination of human-rated images from gMAD and existing databases. This process may be iterated, enabling active and progressive fine-tuning from gMAD examples for BIQA. We demonstrate the feasibility of our active learning scheme on a large-scale unlabeled image set, and show that the fine-tuned method achieves improved generalizability in gMAD, without destroying performance on previously trained databases.

Index Terms—Blind image quality assessment, deep neural networks, gMAD competition, active learning, subjective quality assessment.

1 INTRODUCTION

A s a fundamental problem in computational vision, objective image quality assessment (IQA) involves matching how humans perceive image distortions [1], and has been studied since 1970s [2]. High quality prediction performance can be achieved by comparing a test image to its original counterpart, a setting known as full-reference IQA [3]. Humans are able to perform quality evaluation without any reference at amazing speed and efficiency, and therefore it is reasonable to build computational models to accomplish a similar goal [4]. The resulting blind IQA (BIQA) methods are applicable to a variety of image processing and computer vision tasks [5], [6], where reference images may not exist. Moreover, the problem of BIQA itself provides an important test bed for our understanding of natural photographic images.

Early attempts to BIQA are distortion specific [7], [8]. For example, if JPEG compression is assumed, it is straightforward to make measurements to detect 8 × 8 blocking artifacts. Later, general purpose solutions were developed based on models of natural scene statistics (NSS) [9]–[11]. The underlying assumption is that sensory neurons are highly adapted to the statistical properties of the natural environment through both evolutionary and developmental processes [12]. It follows that a measure of the destruction of “naturalness” can provide a good approximation to perceived image quality. NSS-based BIQA models often transform raw images to more compact and sparser representations [13]–[15], so that the statistical regularities can be more easily revealed and summarized using common probability density functions, e.g., generalized Gaussian distributions.

This general methodology is widely practiced by state-of-the-art BIQA models before 2015, some of which add a data-driven component, learning dictionaries [16], and quality-aware centroids [17] directly from distorted patches.

In the past five years, data-driven BIQA models [18], [19] based on deep neural networks (DNNs) come to outperform knowledge-driven models based on NSS, in terms of correlation with human data on existing IQA databases [20], [21]. These methods are built upon successive stages of convolution, nonlinear activation, and downsampling, which can approximate a diversity of interesting functions. Training such architectures with millions of parameters would require massive quality annotations in the form of mean opinion scores (MOSs), which are, however, largely lacking due to significant costs of performing large-scale subjective experiments. Several strategies have been proposed to compensate for the lack of human-rated data, including fine-tuning pre-trained networks [22], [23], training on image patches [18], exploiting degradation processes [19], [24], leveraging multiple noisy annotators [25], and combining IQA databases [26].

However, the impressive correlation numbers of DNN-based BIQA models are questionable for two main reasons. First, model comparison has been performed using a small set of images, which are not sufficiently representative of the whole image population. Second, the same test images have been used to evaluate the models for many years. This raises the risk of overfitting by extensive adaptation to existing IQA datasets. In fact, even for the best-performing BIQA models, dramatic failures can be found automatically via the group maximum differentiation (gMAD) competition [27], a computational method of efficiently falsifying the models by selecting pairs of the most discriminative images (see Fig. 1).

In addition to testing the model generalizability using
2.1 DNNs for BIQA

The main challenge to train DNNs for BIQA is that the small number of human-rated images may not be sufficient to constrain the large number of model parameters, typically in the order of millions. Directly fine-tuning pre-trained DNNs on image classification for BIQA seems a straightforward approach \[22, 26\]. However, it is unclear whether such network architectures and functional units are optimal for the BIQA task. Another strategy is to pre-train DNNs using quality-relevant data that can be generated efficiently. For example, Kang \et al.\ [25], Liu \et al.\ [24], and Zhang \et al.\ [23] exploited the distortion type (and level) information to learn perceptually meaningful initializations. Kim \et al.\ [31] and Ma \et al.\ [25] made use of quality predictions from full-reference IQA models as pseudo ground truths. Methods of this kind hold much promise in handling synthetic distortions, on which they have been trained. It remains a challenge to develop distortion-unaware BIQA methods with good generalizability to unseen distortion types. We choose to predict relative quality differences, and combine the methods in \[25\] and \[26\] to create a top-performing BIQA model (see Table 2), as the starting point of our active and progressive learning of BIQA from gMAD examples.

2.2 gMAD Competition

gMAD \[27\] is a discrete instantiation of the maximum differentiation (MAD) competition \[32\], a general methodology for accelerating the comparison of perceptual models. Specifically, given two IQA models, MAD first synthesizes a pair of images by solving the following constrained optimization problem

\[
(x^*, y^*) = \arg\max_{x,y} f_1(x) - f_1(y)
\]

s.t. \[f_2(x) = f_2(y) = \alpha, x, y \in \mathcal{I},\]

where \(f_j\) for \(j = 1, 2\) are two objective quality models with larger predicted values indicating better perceived quality.
and \( I \) denotes the set of all possible images. The pair of images \((x, y)\) are confined in the \( \alpha \)-level set of \( f_2 \). By varying \( \alpha \), we are able to compare \( f_1 \) and \( f_2 \) at different quality levels. In Problem 1, \( f_1 \) serves as an “attacker”, whose difference of the responses to the pair of images \((x^*, y^*)\) are maximized, while \( f_2 \) works as a “defender”, whose responses to \((x^*, y^*)\) are indistinguishable. MAD repeats this optimization, but with the roles of the two models reversed \([32]\). The resulting small set of synthesized images constitutes the strongest possible examples to falsify competing models, reveal their relative weaknesses, and suggest potential ways for improvements \([32]\).

However, MAD requires a projected gradient descent solver to synthesize images, which is computationally expensive, and is not friendly to non-differentiable IQA models. Moreover, the MAD-synthesized images may be highly unnatural, offering little insight into the relative performance of competing models in real-world applications. gMAD overcomes the above limitations by restricting the search space to a fixed set of images \( S \), i.e., a particular domain of interest. Efficient discrete optimizers can be adopted to solve Problem 1 to global optima. Based on subjective data, gMAD introduces two quantitative measures, aggressiveness and resistance, to summarize the performance of a model at attacking and defending against other models, respectively \([27]\). Several researchers \([23, 33, 34]\) have adopted gMAD to test the generalizability of their proposed models. However, little work has been dedicated to exploiting gMAD examples to improve the generalizability of BIQA models.

### 2.3 Machine Learning from Hard Examples

There is a rich body of literature on learning from hard examples, and the definition of “hardness” depends on the formulation and the goal of the machine learning task at hand. In the case of hard negative mining \([35]\) (also referred to as bootstrapping \([36]\)), training is prioritized for samples with high loss at each iteration, with the goal of making training more effective and efficient. In the case of continual learning \([37]\), also called lifelong learning, the model tries to transfer knowledge learned from previous tasks to new ones with resistance to catastrophic forgetting. The hard examples are mainly from new tasks that may cause performance degradation of previously seen data. In the case of active learning \([38]\), the hard examples are generally informative samples that the model is least certain or expects most change. Active learners aim to train on as few labeled instances as possible to achieve high performance, thereby minimizing the cost of labeling \([38]\). Our training paradigm can be seen as a form of active fine-tuning, where we actively seek informative samples for visual inspection by means of the gMAD competition. The selected examples are most likely to be the strongest possible counterexamples, which may lead to the greatest change to the model. However, the goal here is different: instead of minimizing the effort of subjective testing in IQA \([39]\), we aim to improve the generalizability of the BIQA model by learning from the selected gMAD examples.

![Fig. 2.](image-url) The active fine-tuning cycle for improving BIQA models. We start with a differentiable parametric BIQA model, seek a small number of image pairs by letting it compete with a set of full-reference IQA methods in gMAD \([27]\), collect human opinions on visual quality of the selected images, fine-tune it from the combination of existing IQA databases and newly annotated gMAD set. The model then relies on its new experience to choose which gMAD instances to annotate next.

### 3 Proposed Method

In this section, we describe the proposed method for BIQA, including constructing the baseline model and fine-tuning it in an active and progressive manner (see Fig. 2).

#### 3.1 Constructing the Baseline Model

We build our baseline model in two steps: 1) per-train a DNN on a large-scale database, with images annotated by a set of full-reference IQA methods \([25]\) and 2) fine-tune it on multiple IQA databases simultaneously \([26]\). The first step is used to supply perceptually meaningful initializations for subsequent fine-tuning in the second step.

Given an image \( x \), let \( f(x) \) represent its true perceptual quality. We utilize \( n \) IQA annotators \( \{f_j\}_{j=1}^n \), which compute \( n \) nonlinear and noisy quality estimates of \( f(x) \), collectively denoted by \( \{f_j(x)\}_{j=1}^n \). To cope with different model nonlinearities, an image pair \((x, y)\) is formed and associated with \( n \) binary labels \( \{q_j\}_{j=1}^n \), where \( q_j = 1 \) if \( f_j(x) \geq f_j(y) \) and \( q_j = 0 \) otherwise. The training set is in the form of \( D_1 = \{(x^{(i)}, y^{(i)}), q_1^{(i)}, \ldots, q_n^{(i)}\}_{i=1}^m \), where \( m \) is the number of training pairs. The reliability of each annotator is explicitly modeled by probabilities of correct answer and rejection rates

\[
\alpha_j = \text{Pr}(q_j = 1 | q = 1) \quad (2)
\]

and

\[
\beta_j = \text{Pr}(q_j = 0 | q = 0), \quad (3)
\]

respectively, where \( q = 1 \) if \( f(x) \geq f(y) \) and \( q = 0 \) otherwise.

Our goal is to learn a differentiable function \( f_w(x) \), parameterized by a vector \( w \), which computes a quality value of \( x \). Assuming the Thurstone’s Case V model \([40]\), the probability that \( x \) is of higher quality than \( y \) can be computed by

\[
p_w(x, y) = \text{Pr}(f(x) \geq f(y); w) = \Phi \left( \frac{f_w(x) - f_w(y)}{\sqrt{2}} \right), \quad (4)
\]
where the standard deviation (std) is fixed to one. $\Phi(\cdot)$ is the standard Normal cumulative distribution function. The model parameters $w$ along with the uncertainty variables $\{\alpha, \beta\}$ are jointly estimated using maximum likelihood \cite{25}

$$\{\hat{w}, \hat{\alpha}, \hat{\beta}\} = \arg\max_w \Pr(D_1; w, \alpha, \beta),$$ \hspace{1cm} (5)

where

$$\Pr(D_1; w, \alpha, \beta) = \prod_{i=1}^{m} \left( p_w(x^{(i)}, y^{(i)}) \prod_{j=1}^{n} \Pr(q_{j}^{(i)}|q = 1) ight. 
+ \left. (1 - p_w(x^{(i)}, y^{(i)})) \prod_{j=1}^{n} \Pr(q_{j}^{(i)}|q = 0) \right).$$ \hspace{1cm} (6)

As shown in \cite{25}, the learned model is capable of handling distortion types that have been pre-specified in the training set $D_1$, but does not generalize well to unseen distortions, especially those with substantially different visual appearances.

To enhance model generalizability, we leverage the training technique proposed in \cite{26}, and fine-tune our BIQA model on multiple subject-rated IQA databases simultaneously. Given $n$ IQA databases, $m_j$ pairs of images $\{(x_j^{(i)}, y_j^{(i)})\}_{i=1}^{m_j}$ are randomly sampled from the $j$-th database, and a total of $m = \sum_{j=1}^{n} m_j$ image pairs are constructed. For each pair $(x, y)$, a continuous quality annotation is computed, indicating the probability of $x$ having higher perceived quality than $y$

$$p(x, y) = \Pr(f(x) \geq f(y)) = \Phi \left( \frac{\mu(x) - \mu(y)}{\sqrt{\sigma^2(x) + \sigma^2(y)}} \right),$$ \hspace{1cm} (7)

where the Thurstone’s model \cite{40} is assumed, and $\mu(x)$ and $\sigma(x)$ are the MOS of $x$ and the corresponding std, respectively. The training set is therefore in the form of $D_2 = \{(x_j^{(i)}, y_j^{(i)}, p_j^{(i)})\}_{i=1}^{n}, \forall j$, where we effectively combine multiple databases without performing additional subjective experiments for perceptual scale realignment. In \cite{26}, the fidelity loss \cite{41} is used to measure the similarity between two discrete probability distributions

$$\ell(x, y; p, w) = 1 - \sqrt{p(x, y)p_w(x, y)} - \sqrt{(1 - p(x, y))(1 - p_w(x, y))}.$$ \hspace{1cm} (8)

Compared to cross entropy and Kullback-Leibler divergence, the fidelity loss has several desired properties. First, it has a clear physical interpretation, and is used to measure the difference between two states of a quantum object. Second, the minimal loss at zero is achievable for all ground truth $p \in [0, 1]$. Third, the fidelity loss is bounded between zero and one (see Fig. 3). In this paper, we will also use the fidelity loss to monitor the progress of our BIQA model and to help pick gMAD pairs for qualitative comparison.

Finally, the model parameters $w$ are fine-tuned by minimizing the mean fidelity loss over the combined database $D_2$

$$\ell(D_2; w) = \frac{1}{|D_2|} \sum_{i,j} \ell(x_j^{(i)}, y_j^{(i)}, p_j^{(i)}, w),$$ \hspace{1cm} (9)

where $|D_2|$ denotes the cardinality of $D_2$.

3.2 Active Fine-Tuning from gMAD Examples

After acquiring the baseline model $f_w$, we are able to actively fine-tune it using a small set of model-dependent images adaptively selected by gMAD. We first build a large-scale unlabeled image set $S$ as the playground for gMAD. As the size of the gMAD set $U$ subject to visual inspection is orthogonal to that of $S$, we may make $S$ arbitrary large such that it spans a great variety of natural scenes, distortion types and levels. We assume a subjective assessment environment, where we can collect the MOS of $x \in S$ and its corresponding std. We also assume a set of full-reference IQA method $\{f_j\}_{j=1}^{n}$, each of which takes a distorted image $x$ and its corresponding reference $x'$ as input, and computes an estimate of the true perceptual quality, $f_j(x)$, where we have omitted $x'$ in the parenthesis to keep the notation uncluttered. Fixing a quality level $\alpha$, we first let our model and the $j$-th full reference IQA method be the defender and the attacker, respectively. The optimal pair of images in terms of discriminating $f_w$ and $f_j$ can be found by solving

$$\begin{align*}
(x^*, y^*) &= \arg\max_{x,y} f_j(x) - f_j(y) \\
&\text{s.t. } f_w(x) = f_w(y) = \alpha, \quad x, y \in S,
\end{align*}$$ \hspace{1cm} (10)

where the $j$-th full-reference method believes that, for the selected image pair, $x^*$ has much better visual quality than $y^*$, while our model suggests that they are of approximately
the same quality. The subjective result of \((x^r, y^r)\) roughly falls into three categories:

- **Case I.** \(p(x^r, y^r) \approx 1: x^r\) is indeed of better quality than \(y^r\). In this case, \(f_j\) makes a successful attack, identifying a counterexample of \(f_w\). The selected pair of images contain constructive information about improving \(f_w\).

- **Case II.** \(p(x^r, y^r) \approx 0.5: x^r\) and \(y^r\) have very similar visual quality. In this case, \(f_w\) survives the attack from \(f_j\), which is in disagreement with human visual inspection. \((x^r, y^r)\) is informative in discriminating the two models, but may contribute less to performance improvement of \(f_w\).

- **Case III.** \(p(x^r, y^r) \approx 0: y^r\) has better quality than \(x^r\). In this case, \((x^r, y^r)\) is able to falsify both models, leading to a double-failure result. The selected pair is useful for the refinement of \(f_w\).

We then switch the roles of the two models, and seek an image pair \((x^a, y^a)\), to which the difference of the responses of \(f_w\) is maximized in the \(\alpha\)-level set of \(f_j\). That is, \(f_w\) thinks \(x^a\) is perceived much better than \(y^a\), while \(f_j\) considers they are indistinguishable in terms of image quality. Subjective testing on \((x^a, y^a)\) leads to another three possible outcomes:

- **Case IV.** \(p(x^a, y^a) \approx 1: x^a\) is of clearly higher quality than \(y^a\). In this case, \(f_w\) successfully spots a counterexample of \(f_j\). However, \((x^a, y^a)\) may be less useful to further enhance \(f_w\).

- **Case V.** \(p(x^a, y^a) \approx 0.5: x^a\) and \(y^a\) are of approximately the same quality. In this case, the attack by \(f_w\) is not successful, which exposes its own weakness when competing with \(f_j\). As a result, \((x^a, y^a)\) can be used to improve \(f_w\).

- **Case VI.** \(p(x^a, y^a) \approx 0: y^a\) has clearly better quality than \(x^a\). In this case, we reach a double-failure conclusion once again. As the responses of \(f_w\) to \((x^a, y^a)\) are opposite to human judgments, harnessing \((x^a, y^a)\) would impart the largest change to \(f_w\).

For a relatively weak BIQA model, when competing with a group of full-reference IQA methods, the selected gMAD pairs are more likely to fall into Case I and Case V, which manifest themselves as strong gMAD counterexamples, and offer various ways for enhancement. For a high-performance BIQA model (as is the case in our paper), we would expect to see gMAD pairs belonging to Case II and Case IV more often (see Fig. 7).

In practice, we assume \(l\) quality levels (i.e., \(\alpha\) can take on \(l\) values), and for each quality level, we choose top-\(k\) gMAD pairs with \(k\) largest response differences computed by the objective in Problem \(P(10)\). We then reverse the roles of the two models, finding another top-\(k\) gMAD pairs. After pairwise comparison with \(n\) full-reference methods, we obtain an unlabeled gMAD set \(U\) that contains \(2 \times k \times l \times n\) pairs. We invite a number of subjects to rate each image \(x \in U\) in a well-controlled laboratory environment (see Section 4.2.3 for details). The MOS \(\mu(x)\) and the associated std \(\sigma(x)\) can be computed accordingly. The ground truth annotation \(p(x, y) \in [0, 1]\) for a gMAD pair \((x, y)\) can also be derived using Eq. (7), leading to a labeled gMAD set \(L\) of the same size. After active fine-tuning on \(L\), we may iterate this process several rounds: leverage new knowledge acquired by \(f_w\) to seek another set of gMAD examples, request human annotations for selected images, and improve \(f_w\) based on the labeled results. This gives us a progressively increased gMAD set \(D_3 = \{L(i):i=1\}\), which is in the form of \(\{((x(i), y(i)), p(i))\}_{i=1}^{m}\), where \(m = 2 \times k \times l \times n \times (T - 1)\) and \(T\) is the maximum rounds of subjective experiments. Note that we reserve \(L(T)\) for only testing purposes.

We now describe the \(t\)-th round of the fine-tuning procedure using the combination of image pairs from \(D_2\) and \(D_3\), where \(D_3 = \{L(i):i=1\}\). The goal is to harness gMAD examples without overfitting, and preserve performance on previously trained IQA databases. In general, the size of \(D_3\) is much smaller compared to that of \(D_2\). We alleviate this data imbalance in two ways. First, instead of directly adapting to the selected gMAD pairs, we randomly pair up gMAD images, which results in an augmented training set \(D_3\) containing \(n \times (2m + 1)\) pairs. Second, we weight the loss functions according to the number of instances in the respective databases:

\[
\ell(D_2, D_3; w) = \frac{1}{|D_2|} \sum_{i,j} \ell(x(i), y(i), p(j); w) + \frac{1}{|D_3|} \sum_{i} \ell(x(i), y(i), p(i); w). \tag{11}
\]

Algorithm 1 summarizes the entire procedure of the proposed method.
TABLE 1
Summary of different IQA databases. MOS stands for mean opinion score. DMOS is inversely proportional to MOS

| Database                  | # of original images | # of distorted images | # of distortion types | Score type | Score range          | Subjective testing methodology         |
|--------------------------|----------------------|-----------------------|-----------------------|------------|----------------------|----------------------------------------|
| LIVE [20]                | 29                   | 779                   | 5                     | DMOS       | [0, 100]             | Single-stimulus continuous scale       |
| CSIQ [43]                | 30                   | 866                   | 6                     | DMOS       | [0, 1]               | Multi-stimulus absolute category       |
| TID2013 [21]            | 25                   | 3,000                 | 24                    | MOS        | [0, 9]               | Two-alternative forced choice          |
| KADID-10k [44]          | 81                   | 10,125                | 25                    | MOS        | [1, 5]               | Double-stimulus absolute category      |
| Waterloo Exploration [29] | 4,744                | 94,800                | 4                     | N.A.       | N.A.                 | Need-based                             |

Fig. 4. The network architecture of our BIQA model. The parameterization of the convolutional layers is denoted as “filter support | input channel × output channel”. The number of parameters for each layer is given at the bottom, summing up to 154,865.

4 EXPERIMENTS
In this section, we demonstrate the feasibility of the proposed method in real settings. We first present in detail the baseline BIQA model for synthetic distortions. We then describe the active fine-tuning cycle, including the construction of the large-scale unlabeled image set S, the implementation of the gMAD competition, the environment of the subjective experiment, the procedure of active fine-tuning. Last, we conduct both quantitative and qualitative analysis of the proposed method with a number of interesting observations.

4.1 Specification of the Baseline Model

4.1.1 Network Architecture
Our BIQA model is adapted from [25] and is specified in Fig. 4. \( f_w \) is a generic four-layer convolutional network. Each layer applies a bank of 3 \( \times 3 \) convolutional filters to its inputs. Following each convolution, we employ generalized divisive normalization (GDN), in which all responses are divided by pooled responses of their rectified and exponentiated neighbors [45]. It implements a form of local gain control, which is useful in explaining nonlinear behaviors of cortical neurons [46]. GDN is defined as

\[
{v_i} = \frac{u_i}{(\omega_i + \sum_j \gamma_{ij} u_j^2)^\gamma}, \tag{12}
\]

where \( u \) and \( v \) are the input to and the output of GDN, respectively, and \( \{\omega, \gamma\} \) are the parameters to be determined. Apart from IQA [19], [47], GDN has also been successfully adopted in density modeling [45] and image compression [48]. The normalization responses are max-pooled by a factor of two along each spatial dimension. The spatial statistics are summarized by a fixed-length representation using spatial pyramid pooling [49] regardless of input image resolution. Last, the quality value is computed by two fully connected layers with a rectified linear unit (ReLU) nonlinearity in between.

4.1.2 Construction of \( D_1 \)
We build the pseudo-labeled image set \( D_1 \) based on the reference images from the Waterloo Exploration Database [29]. We simulate eighteen common distortion types [25] at each five levels. We assemble four types of image pairs [25]: same reference image and distortion type, with different distortion levels; same reference image, but different distortion types and levels; two different reference images, distortion types and levels; two different reference images, with one undistorted. We generate a total of 600,000 training pairs, whose labels are supplied by six full-reference IQA models.

4.1.3 Construction of \( D_2 \)
We build the subject-rated image set \( D_2 \) by combining four synthetically distorted image databases - LIVE [20], CSIQ [43], TID2013 [21], and KADID-10k [44] (see Table 1 for details). We randomly sample 80% of the reference images and their corresponding distorted ones to form the fine-tuning set \( D_2 \), and leave the rest for evaluation. In order to guarantee the full content independence, special treatment is given when we partition overlapping reference images in LIVE and TID2013. In the end, we generate 50,000, 50,000, 100,000, and 200,000 image pairs from LIVE, CSIQ, TID2013, and KADID-10k, respectively, yielding a total of 400,000.

4.1.4 Details of Pre-Training, Fine-Tuning and Testing
Pre-training is performed by maximizing the likelihood of \( D_1 \) (in Eq. (5)), using the Adam optimizer [54] with a minibatch of 16 and a learning rate of \( 10^{-4} \). After each iteration, we project the parameters \( \omega \) and \( \gamma \) in GDN onto the interval \([2^{-10}, \infty] \), and constrain \( \gamma \) to be symmetric. The maximum epoch number is set to eight. Fine-tuning is performed by minimizing the mean fidelity loss on \( D_2 \) (in Eq. (9)). The Adam solver is adopted with a mini-batch size of 16, a learning rate of \( 10^{-4} \), and a maximum epoch number of 1. These include additive white Gaussian noise, multiplicative noise, pink noise, salt and pepper noise, Gaussian blur, JPEG compression, JPEG2000 compression, Gaussian denoising, color quantization, dithering, neighboring patch substitution, flat patch substitution, contrast change, saturation decrease, chromatic aberration, over-exposure, under-exposure, and ghosting.
Matching the two full-reference models. The performance improvements of $D_{\text{lar}}$ are clearly strong, which is not surprising because the distortion BIQA models. Performance on LIVE and CSIQ is particularly strong, which is not surprising because the distortion types in the two test sets have been seen during training. After fine-tuning on $D_2$, we observe significant performance improvements of $f_w$ on TID2013 and KADID-10k, closely matching the two full-reference models. The performance on LIVE and CSIQ drops slightly as a consequence of balancing more unseen distortion types. In summary, by combining the training techniques in [25] and [26], we arrive at a top-performing BIQA model that is capable of handling a number of synthetic distortions.

### 4.2 Specification of the Active Fine-Tuning Cycle

#### 4.2.1 Construction of $S$.

We collect a large-scale unlabeled image set $S$ as the candidate pool to seek gMAD examples for active fine-tuning. Specifically, we first download high-quality and high-definition natural images from Internet that carry Creative Common licenses. They can be loosely grouped into twelve categories: amphibian, bird, fish, flower, fruit, furniture, geological formation, mammal, musical instrument, reptile, tool, and vehicle (see representative images in Fig. 5). We remove near-duplicate images using the command line tool imgdupes([2] and delete those with inappropriate content. This leaves us 10,000 natural photographic images, and the number in each category is approximately the same. We downsample the images to a maximum width or height of 1,024 as a way of further reducing possibly visible artifacts. After data screening, we add 25 types of distortions with five levels of severity, which are the same in KADID-10k [44] and can be roughly classified into seven categories: blurring, color-related distortion, compression, noise-related distortion, intensity change, contrast change and others. Finally, for each reference image, we randomly choose 5 out of 25 distortion types and 2 out of 5 levels, resulting in a total of $5 \times 2 \times 10,000 = 100,000$ distorted images.

#### 4.2.2 Construction of $l^{(i)}$

We let our method compete with nine state-of-the-art full-reference IQA models - SSIM [50], MS-SSIM [28], NLPD [47], VSI [56], MAD [43], VIF [57], MDSI [58], PieAPP [51], and WaDIQaM [18], among which the former seven are knowledge-driven methods based on our understanding of the image source, the image distortion and the HVS, while the latter two are purely data-driven methods based on DNNs. All implementations are obtained from the original authors, expect for WaDIQaM which we use a publicly available re-implementation [3]. gMAD requires all competing models to work in the same perceptual space. Therefore, we map all model predictions using Eq. (13) onto the LIVE MOS scale $[0, 100]$, with higher values indicating better perceptual quality. Five levels ($l = 5$) are specified to roughly cover bad, poor, fair, good, and excellent quality. The quality range ($i.e.$, bin width) is half of the mean std in LIVE, ensuring that the images in the same level have similar quality in terms of the defender model. Two types of gMAD pairs are queried by treating our baseline model as the defender and the attacker, respectively. We take the subjective testing effort into account, and search for a maximum of $k = 12$ pairs at each quality level. During this process, we find that if our model fails in one corner case, more failure examples of the same case may be picked out repeatedly by other competing models. To guarantee content and distortion diversity of the

| Table 2 | Correlation (SRCC and PLCC) between model predictions and MOSs on $T$. Top section lists two representative full-reference models. Second section contains four knowledge-driven and three data-driven DNN-based BIQA models. The results on the databases used to train the respective models are not shown. The top two correlations obtained by BIQA models are highlighted in boldface. |
|---------|-------------------------------------------------|
| SRCC | LIVE | CSIQ | TID2013 | KADID-10k |
| SSIM [50] | 0.951 | 0.871 | 0.719 | 0.747 |
| PieAPP [51] | 0.919 | 0.891 | 0.814 | 0.886 |
| BRISQUE [19] | – | 0.568 | 0.407 | 0.335 |
| NIQE [11] | 0.922 | 0.618 | 0.315 | 0.404 |
| HOSA [52] | – | 0.602 | 0.469 | 0.353 |
| dipIQ [33] | 0.944 | 0.501 | 0.412 | 0.293 |
| MEON [19] | – | 0.741 | 0.379 | 0.214 |
| NIMA [53] | 0.506 | 0.521 | 0.301 | 0.233 |
| deepIQA [18] | 0.807 | 0.752 | – | 0.595 |
| Baseline ($D_1$) | 0.910 | 0.870 | 0.875 | 0.621 |
| Baseline ($D_2$) | 0.896 | 0.859 | 0.822 | 0.861 |
| PLCC | LIVE | CSIQ | TID2013 | KADID-10k |
| SSIM | 0.940 | 0.861 | 0.784 | 0.738 |
| PieAPP | 0.902 | 0.880 | 0.876 | 0.887 |
| BRISQUE | – | 0.677 | 0.544 | 0.394 |
| NIQE | 0.919 | 0.742 | 0.427 | 0.460 |
| HOSA | – | 0.760 | 0.590 | 0.436 |
| dipIQ | 0.945 | 0.758 | 0.454 | 0.400 |
| MEON | – | 0.786 | 0.486 | 0.403 |
| NIMA | 0.511 | 0.601 | 0.476 | 0.348 |
| deepIQA | 0.839 | 0.814 | – | 0.612 |
| Baseline ($D_1$) | 0.910 | 0.902 | 0.711 | 0.628 |
| Baseline ($D_2$) | 0.915 | 0.897 | 0.837 | 0.866 |

eight. During testing, we quantify the performance using the Spearman’s rank correlation coefficient (SRCC) and the Pearson linear correlation coefficient (PLCC). For the latter, a pre-processing step is added to linearize model predictions by fitting a four-parameter monotonically function

$f_w(x) = \frac{(\eta_1 - \eta_2)}{1 + \exp(-f_w(x) - \eta_3)} + \eta_4. \tag{13}$

The test set consists of four subsets of images from LIVE, CSIQ, TID2013, and KADID-10k, respectively, which we collectively denote by $T$.

#### 4.1.5 Preliminary Results

We compare our baseline model with seven BIQA methods, including BRISQUE [10], NIQE [11], HOSA [52], dipIQ [33], MEON [19], NIMA [53], and deepIQA [18]. The former four are knowledge-driven, among which NIQE relies solely on a prior probability model of natural undistorted images and does not need MOSs for training. The latter three are data-driven DNN-based models, among which NIMA is optimized for predicting perceptual image aesthetics using the AVA database [55]. We also include two full-reference IQA methods - SSIM and PieAPP [51] for reference. Table 2 shows the SRCC and PLCC results on $T$ from four IQA databases. Pre-trained on $D_1$, our model outperforms most BIQA models. Performance on LIVE and CSIQ is particularly strong, which is not surprising because the distortion types in the two test sets have been seen during pre-training. After fine-tuning on $D_2$, we observe significant performance improvements of $f_w$ on TID2013 and KADID-10k, closely matching the two full-reference models. The performance on LIVE and CSIQ drops slightly as a consequence of

2. https://github.com/knjcode/imgdupes#against-large-dataset

3. https://github.com/lidq92/WaDIQaM
Fig. 5. Sample images from the large-scale unlabeled set $S$ for gMAD competition. (a) Amphibian. (b) Bird. (c) Fish. (d) Flower. (e) Fruit. (f) Furniture. (g) Geological formation. (h) Mammal. (i) Musical instrument. (j) Reptile. (k) Tool. (l) Vehicle. Images are cropped for improved visibility.

Fig. 6. Graphical user interface for subjective testing.

selected images, we enforce several additional constraints on pair selection for each pairwise model comparison: (1) images of the same content appear at most twice; (2) images of the same distortion type appear at most three times; (3) combinations of the same two distortion types appear at most once.

4.2.3 Subjective Testing

We set up the subjective experiment in an office environment with a normal indoor illumination level. The display we use is a true-color LED monitor with the resolution of $2560 \times 1920$ pixels, and we calibrated it according to the recommendation of ITU-R BT.500 [59]. Fig. 6 illustrates the graphical user interface we customize for this experiment.

A gMAD pair is rendered at full image resolution, but in random spatial order. Two scale-and-slider applets are utilized to collect the quality score of each image, with 0 and 100 indicating the worst and the best quality, respectively. The viewing distance is fixed to 32 pixels per degree of visual angle. For each $U(t)$, we gather data from fifteen subjects with normal or correct-to-normal visual acuity. They have general knowledge of image processing and computer vision, but do not know the detailed purpose of the study. We include a training session to familiarize them with image distortions, using several sample images that are independent of those in $U(t)$. Each subject is asked to give scores to all gMAD images. To minimize the influence of the fatigue effect, the subjects are allowed to take a break for a while after a maximum of 30-minute experiment. We process the raw data using the outlier detection and subject rejection algorithm in [60]. In total, we perform three rounds of subjective experiments ($T = 3$) by repeating the same procedure. $L(1)$ and $L(2)$ are used to evaluate and refine $f_w$ in the active fine-tuning cycle, while $L(3)$ is reserved for testing. After data purification, we find that all subjects are valid, and $2.82\%$, $2.68\%$ and $2.26\%$ of all ratings are identified as outliers and subsequently removed in $L(1)$, $L(2)$ and $L(3)$, respectively.

Fig. 7 shows the empirical distributions of $p(x^r, y^r)$ and $p(x^a, y^a)$ computed by Eq. (7) on $L(1)$. When the baseline model is the defender, it is effortless for the set of full-reference IQA methods to spot its failures, as evidenced by a large percentage of pairs with $p(x^r, y^r) > 0.8$ (belonging to
Fig. 7. The empirical distributions of (a) \( p(x', y') \) and (b) \( p(x^a, y^a) \) on \( \mathcal{L}^{(1)} \). It is clear that full-reference IQA methods (as attackers) can easily falsify our BIQA model, and vice versa.

Fig. 8. The progress of our method in terms of the mean fidelity loss (± standard error) on the gMAD sets, when playing the role of the defender and the attacker, respectively.

Case I). These are strong counterexamples of \( f_w \), shedding light on how to improve it. When our model works as the attacker, it performs surprisingly well in falsifying full-reference models with a large portion of the selected pairs belonging to Case IV. This adds new direct evidence to our claim of the superiority of the baseline model before active fine-tuning.

4.2.4 Details of Active Fine-Tuning

For each round of active fine-tuning, we minimize the weighted mean fidelity loss in Eq. (11). The Adam optimizer is used with a mini-batch size of 16 - half from \( D_2 \) and half from \( D_3 \). This amounts to oversampling \( D_3 \), and provides an equivalent implementation of Eq. (11) in the mini-batch setting. The learning rates for shallow layers (up to the second GDN layer) and deep layers are set to \( 10^{-5} \) and \( 10^{-4} \), respectively. The maximum epoch number is set to eight. SRCC, PLCC, and the mean fidelity loss are used to quantify the performance during testing.

4.3 Main Results

4.3.1 Quantitative Analysis

Table 3 lists the SRCC and PLCC results between model predictions and MOSs on the gMAD image sets \( \mathcal{L}^{(1)}, \mathcal{L}^{(2)}, \) and \( \mathcal{L}^{(3)} \), respectively. Before active fine-tuning, all full-reference IQA models surpass the baseline on \( \mathcal{L}^{(1)} \), except for SSIM \[50\] in terms of SRCC. After the first round of active fine-tuning on \( \mathcal{L}^{(1)} \), our method is able to learn from and combine the best aspects of the competing models, outperforming all of them by a large margin. As expected, the performance of the full-reference models on \( \mathcal{L}^{(2)} \) deteriorates. After the second round of active fine-tuning on both

| Method | LIVE | CSIQ | TID2013 | KADID-10k |
|--------|------|------|---------|-----------|
| Round 1 | 0.918 | 0.863 | 0.805   | 0.850     |
| Round 2 | 0.914 | 0.871 | 0.826   | 0.872     |

| Method | LIVE | CSIQ | TID2013 | KADID-10k |
|--------|------|------|---------|-----------|
| Round 1 | 0.901 | 0.890 | 0.821   | 0.858     |
| Round 2 | 0.901 | 0.911 | 0.846   | 0.881     |

Table 4 Correlation (SRCC and PLCC) of model predictions by \( f_w \) against human ratings on \( T \) after active and progressive fine-tuning

| Method | LIVE | CSIQ | TID2013 | KADID-10k |
|--------|------|------|---------|-----------|
| Baseline | 0.896 | 0.859 | 0.822   | 0.861     |
| Round 1 | 0.918 | 0.863 | 0.805   | 0.850     |
| Round 2 | 0.914 | 0.871 | 0.826   | 0.872     |
| Round 3 | 0.901 | 0.890 | 0.821   | 0.858     |
| Round 4 | 0.901 | 0.911 | 0.846   | 0.881     |

TABLE 3 Correlation (SRCC and PLCC) results on the gMAD image sets. Our results on \( \mathcal{L}^{(1)}, \mathcal{L}^{(2)}, \mathcal{L}^{(3)} \) are obtained by the proposed method before active fine-tuning, after the first round of active fine-tuning on \( \mathcal{L}^{(1)} \), and after the second round of active fine-tuning on both \( \mathcal{L}^{(1)} \) and \( \mathcal{L}^{(2)} \), respectively. See Algorithm 1 for the detailed procedure

| Method | LIVE | CSIQ | TID2013 | KADID-10k |
|--------|------|------|---------|-----------|
| Baseline | 0.743 | 0.659 | 0.484   | 0.514     |
| Round 1 | 0.820 | 0.739 | 0.603   | 0.644     |
| Round 2 | 0.811 | 0.773 | 0.629   | 0.646     |
| Round 3 | 0.877 | 0.774 | 0.674   | 0.688     |
| Round 4 | 0.799 | 0.736 | 0.649   | 0.640     |
| Round 5 | 0.686 | 0.700 | 0.698   | 0.700     |
| Round 6 | 0.887 | 0.776 | 0.669   | 0.689     |
| Round 7 | 0.866 | 0.800 | 0.722   | 0.765     |
| Round 8 | 0.906 | 0.818 | 0.732   | 0.770     |

TABLE 4 Correlation (SRCC and PLCC) of model predictions by \( f_w \) against human ratings on \( T \) after active and progressive fine-tuning

| Method | LIVE | CSIQ | TID2013 | KADID-10k |
|--------|------|------|---------|-----------|
| Baseline | 0.915 | 0.897 | 0.837   | 0.866     |
| Round 1 | 0.930 | 0.900 | 0.821   | 0.858     |
| Round 2 | 0.931 | 0.911 | 0.846   | 0.881     |

WaDIQaM \[18\] is trained on KADID-10k.
Fig. 9. gMAD image pairs with the maximum fidelity losses (i.e., the worst-case scenarios) selected in (a) $L^{(1)}$, (b) $L^{(2)}$, and (c) $L^{(3)}$, respectively, when our model is the defender and VSI [56] is the attacker.

Fig. 10. gMAD image pairs with the maximum fidelity losses selected in (a) $L^{(1)}$, (b) $L^{(2)}$, and (c) $L^{(3)}$, respectively, when our model is the defender and MDSI [58] is the attacker.

$L^{(1)}$ and $L^{(2)}$, we do not observe noticeable improvements of our model on $L^{(3)}$. We speculate that the gMAD examples in $L^{(2)}$ contain less useful information in refining the proposed method. More importantly, our model may begin to overfit $L^{(1)}$ and $L^{(2)}$, as indicated by performance improvements of most full-reference models on $L^{(3)}$ compared to that on $L^{(2)}$. We treat it as a stopping signal of the active fine-tuning cycle, which strikes a good balance between subjective testing budget and model performance. From Table 3, it is interesting to note that the behaviors of the full-reference IQA methods in the gMAD competition are consistent with those on KADID-10k, which shares the same distortion types. When using our method as the anchor in gMAD, we successfully track the progress of full-reference IQA, where the two recent DNN-based models are among the best.

We take a closer look at the performance changes of our method, when it plays the role of the defender and the attacker, respectively. Fig. 8 shows the mean fidelity losses, where we have several interesting observations. First, after the first round of active fine-tuning, both resistance and aggressiveness of $f_w$ (in terms of the mean fidelity loss) improve significantly. This suggests that without increasing model capacity (e.g., adding more convolution and GDN
layers), our model is able to harness hard gMAD examples. Second, we find that the associated standard errors also largely reduce, suggesting that the improvements are consistent across a majority of the selected gMAD image pairs. Third, the second round of active fine-tuning slightly improves the resistance, but degrades the aggressiveness of $f_w$, which confirms our previous analysis of potential overfitting.

Last, we summarize the SRCC and PLCC results of our model on $T$ in Table 4. Noticeable improvements are achieved on all four test sets after two rounds of active fine-tuning. This may be due to two main reasons: 1) more exposure to the training images in $D_2$ and 2) incorporation of the gMAD image pairs. We conduct an ablation experiment, where we only include images in $D_2$ for further fine-tuning, and find that the first reason is the dominant factor. Therefore, we arrive at a conservative conclusion: the proposed active learning cycle can be used to improve the robustness of the BIQA model, without sacrificing the performance on previously seen data.

4.3.2 Qualitative Analysis

We further qualitatively evaluate the progress of our model in the active fine-tuning cycle. Fig. 9 shows three gMAD image pairs with the maximum fidelity losses (i.e., the worst-case scenarios) selected in (a) $L^{(1)}$, (b) $L^{(2)}$, and (c) $L^{(3)}$, respectively, when VIF [57] is the defender and our model is the attacker.

Fig. 11. gMAD image pairs with the maximum fidelity losses selected in (a) $L^{(1)}$, (b) $L^{(2)}$, and (c) $L^{(3)}$, respectively, when VIF [57] is the defender and our model is the attacker.

Fig. 12. gMAD image pairs with the maximum fidelity losses selected in (a) $L^{(1)}$, (b) $L^{(2)}$, and (c) $L^{(3)}$, respectively, when PieAPP [51] is the defender and our model is the attacker.
pairs with the maximum fidelity losses (as the worst-case scenarios) in $L^{(1)}$, $L^{(2)}$, and $L^{(3)}$, respectively, when our model is the defender and VSI [55] is the attacker. The pair of images in (a) exhibit dramatically different perceptual quality (in disagreement with our model), while those in (c) have very similar perceptual quality (in disagreement with VSI). This shows that great progress has been made by our model by correcting predictions for strong color distortions. A similar result is obtained when MDSI [58] attacks our model (see Fig. 10).

We also examine the gMAD image pairs with the maximum fidelity losses, when our model is the attacker. Fig. 11 shows the results of VIF [57] being under attack. The perceptual quality of the images in (a) is close, which is in disagreement with our model. However, the images in (b) are slightly discriminable, indicating that the aggressiveness of our model is improving. Finally, the images in (c) are clearly discriminable, where VIF gives the blurred image less penalty. Fig. 12 shows the results of PieAPP [51] being the defender. Similarly, in $L^{(3)}$, we successfully identify a strong failure case of PieAPP.

Last, we visualize the changes of predictions by our model on images with similar content, as shown in Fig. 13. In the beginning, our baseline model gives high ratings to severely darkened images, while make low quality predictions on images of wood textures. After incorporating images of similar content into the first round of active fine-tuning, our model gives more reasonable predictions to images of similar content not appearing in $L^{(1)}$. More accurate predictions on images of wood textures can be made after the second round of active fine-tuning. In summary, we observe a trend that our model adapts gradually to gMAD examples.

5 CONCLUSION AND DISCUSSION

We have introduced an active fine-tuning cycle for improving BIQA methods. Combining with the training techniques for constructing the baseline, we have presented a complete and practical framework to learn a top-performing BIQA model that 1) relies on only a handful of human-labeled images, 2) delivers superior performance on existing IQA databases of synthetic distortions, 3) exhibits strong aggressiveness and resistance in gMAD, even when competing with a set of full-reference IQA methods.

We used the gMAD competition methodology to seek informative samples for active fine-tuning. It is of interest to examine whether traditional query strategies [38], such as those based on uncertainty sampling, expected model change and expected error reduction, can facilitate the robustness of the BIQA model, and to compare the results with ours under the same human-labeling budget. Recently, Wang et al. [61] extended the idea of gMAD to compare a number of ImageNet classifiers. It is natural to explore the current work in the context of image classification as a way of improving the generalizability of the classifiers to natural image manifold.

Our work presents a new line of research in BIQA. We conclude by listing other research directions that, we believe, are worth exploring. First, it is desirable to adapt BIQA models trained on a fixed set of synthetic distortion types to unseen ones. Xu et al. [52] made one of first attempts by exploiting higher order image statistics. Second, a practical BIQA model should be able to handle both synthetic and realistic camera distortions. It is interesting to extend our work to such a cross-distortion-scenario setting. Third, a universal BIQA method should embody a prior probability model of natural undistorted images. Mittal et al. [11] developed such a model with reasonable generalizability. Fourth,
how to incorporate high-level semantics into the design of BIQA is yet another challenging problem for future research.

**Acknowledgment**

The authors would like to thank Dingquan Li for fruitful discussions throughout the development of this project, and Yiru Yao for coordinating the subjective experiments. This project was supported by the CityU Start-up Grant (No. 7200630).

**References**

[1] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. Morgan & Claypool, 2006.

[2] J. L. Mannos and D. J. Sakrison, “The effects of a visual fidelity criterion of the encoding of images,” *IEEE Transactions on Information Theory*, vol. 20, no. 4, pp. 525–536, 1974.

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

[4] Z. Wang and A. C. Bovik, “Reduced- and no-reference image quality assessment: The natural scene statistic model approach,” *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 29–40, 2011.

[5] X. Kong and Q. Yang, “No-reference image quality assessment for image auto-denoising,” *International Journal of Computer Vision*, vol. 126, no. 5, pp. 537–549, 2018.

[6] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, “Learning a no-reference quality metric for single-image super-resolution,” *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017.

[7] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, “Perceptual blur and ringing metrics: Application to JPEG2000,” *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 163–172, 2004.

[8] Z. Wang, H. R. Sheikh, and A. C. Bovik, “No-reference perceptual quality assessment of JPEG compressed images,” in *IEEE International Conference on Image Processing*, 2002, pp. 477–480.

[9] 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, 2011.

[10] 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, 2012.

[11] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a completely blind image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 200–212, Mar. 2013.

[12] E. P. Simoncelli and B. A. Olshausen, “Natural image statistics and neural representation,” *Annual Review of Neuroscience*, vol. 24, no. 1, pp. 1193–1216, Mar. 2001.

[13] N. Ahmed, T. Natarajan, and K. R. Rao, “Discrete cosine transform,” *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[14] S. Mallat, *A Wavelet Tour of Signal Processing*. Elsevier, 1999.

[15] O. Schwartz and E. P. Simoncelli, “Natural signal statistics and sensory gain control,” *Nature Neuroscience*, vol. 4, no. 8, pp. 819–825, 2001.

[16] 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.

[17] 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.

[18] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, “Deep neural networks for no-reference and full-reference image quality assessment,” *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2017.

[19] 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, 2017.

[20] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms,” *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.
[44] H. Lin, V. Hosu, and D. Saupe, “KADID-10k: A large-scale artificially distorted IQA database,” in *International Conference on Quality of Multimedia Experience*, 2019, pp. 1–3.

[45] J. Ballé, V. Laparra, and E. P. Simoncelli, “Density modeling of images using a generalized normalization transformation,” in *International Conference on Learning Representations*, 2016, pp. 1–14.

[46] D. J. Heeger, “Normalization of cell responses in cat striate cortex,” *Visual Neuroscience*, vol. 9, no. 02, pp. 181–197, 1992.

[47] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelli, “Perceptual image quality assessment using a normalized Laplacian pyramid,” in *International Symposium on Electronic Imaging, Human Vision and Electronic Imaging*, 2016, pp. 1–6.

[48] J. Ballé, V. Laparra, and E. P. Simoncelli, “End-to-end optimized image compression,” in *International Conference on Learning Representations*, 2017, pp. 1–27.

[49] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” in *European Conference on Computer Vision*, 2014, pp. 346–361.

[50] 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, 2004.

[51] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, “PieAPP: Perceptual image-error assessment through pairwise preference,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1808–1817.

[52] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, “Blind image quality assessment based on high order statistics aggregation,” *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016.

[53] H. Talebi and P. Milanfar, “NIMA: Neural image assessment,” *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[54] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[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] L. Zhang, Y. Shen, and H. Li, “VSI: A visual saliency-induced index for perceptual image quality assessment,” *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, 2014.

[57] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[58] H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet, “Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator,” *IEEE Access*, vol. 4, pp. 5579–5590, 2016.

[59] 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](http://www.vqeg.org)

[60] ITU-R BT.500, “Methodology for the subjective assessment of the quality of television pictures,” 2002. [Online]. Available: [https://www.itu.int/rec/R-REC-BT.500](https://www.itu.int/rec/R-REC-BT.500)

[61] H. Wang, T. Chen, Z. Wang, and K. Ma, “I am going MAD: Maximum discrepancy competition for comparing classifiers adaptively,” in *International Conference on Learning Representations*, 2020, pp. 1–12.