This CVPR 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Troubleshooting Blind Image Quality Models in the Wild

Zhihua Wang[1], Haotao Wang[2], Tianlong Chen[2], Zhangyang Wang[2], and Kede Ma[1]
[1] City University of Hong Kong, [2] University of Texas at Austin
zhihua.wang@my.cityu.edu.hk, {htwang,tianlong.chen,atlaswang}@utexas.edu, kede.ma@cityu.edu.hk

Abstract

Recently, the group maximum differentiation competition (gMAD) has been used to improve blind image quality assessment (BIQA) models, with the help of full-reference metrics. When applying this type of approach to troubleshoot "best-performing" BIQA models in the wild, we are faced with a practical challenge: it is highly nontrivial to obtain stronger competing models for efficient failure-spotting. Inspired by recent findings that difficult samples of deep models may be exposed through network pruning, we construct a set of "self-competitors," as random ensembles of pruned versions of the target model to be improved. Diverse failures can then be efficiently identified via self-gMAD competition. Next, we fine-tune both the target and its pruned variants on the human-rated gMAD set. This allows all models to learn from their respective failures, preparing themselves for the next round of self-gMAD competition. Experimental results demonstrate that our method efficiently troubleshoots BIQA models in the wild with improved generalizability.

1. Introduction

Over the years, researchers and engineers in the field of image processing and computer vision have realized the importance of blind image quality assessment (BIQA) [42]. Numerous BIQA models [3, 30, 33, 34, 47, 48] have been proposed, focusing mainly on boosting performance on existing IQA datasets of fixed sizes. However, the superior correlation numbers on closed test sets may not translate in a reliable way to generalization in the open visual world [44, 29, 36, 50]. Therefore, computational methods for probing and improving the generalizablity of BIQA models are highly desirable.

In 2008, Wang and Simoncelli [44] described a maximum differentiation (MAD) competition procedure to compare IQA models in the space of all possible images. Ma *et al*. [29] proposed gMAD, a discrete instantiation of the MAD method, by restricting the search space to some specific domain of interest. Both methods are able to automatically and efficiently expose failures of a relatively *weak* IQA model, by letting it compete with a set of *strong* models. Wang and Ma [43] took advantage of gMAD to identify the counterexamples of a BIQA model [52] using a set of stronger full-reference IQA metrics. Furthermore, they demonstrated that harnessing gMAD-selected failures significantly improves the BIQA generalizability.

Figure 1: Failure cases of a "top-performing" BIQA method - UNIQUE [52] spotted by an ensemble of its pruned versions. **(a)** Best/worst-quality images according to the ensemble, with near-identical quality reported by UNIQUE. **(b)** Best/worst-quality images according to UNIQUE with near-identical quality reported by the ensemble.

Despite demonstrated success, the progressive failure identification and model rectification pipeline proposed in [43] have two drawbacks. First, it can only be applied to the synthetic distortion scenario, where full-reference IQA models are computable. For BIQA models in the wild with input images containing realistic camera distortions, it is highly nontrivial to obtain a list of stronger methods to falsify a state-of-the-art model. Second, the competing full-reference models are fixed throughout model development, rendering failure-spotting less effective as the target model becomes stronger [43].

In this paper, we present an innovative extension of the pipeline proposed in [43], to troubleshooting BIQA models in the wild. We start with a "top-performing" BIQA method based on deep neural networks (DNNs) as the target model. Instead of finding strong external methods, the key step in our approach is to compress the target model using network pruning techniques [15, 16, 24, 27], to construct strong "self-competitors" from the target model. The critical underlying rationale takes root in the recent finding [17] in image classification. The authors observed that, network pruning, which usually removes smallest-magnitude weights in a trained network, does not affect all learned classes or samples equally. Rather, it tends to disproportionally hamper the network memorization and generalization on the long-tailed and most difficult images from the training distribution. In other words, those images are not "memorized" well by the current model, and therefore easily "forgotten" when pruning the model. In short, network pruning can effectively spot the samples not yet well learned or represented, hence exposing the weakness of the trained model.

Inspired by this prior wisdom [17], we propose to leverage network pruning in revealing superficial "shortcuts" in (either original or pruned) BIQA models. In order to encourage spotting *diverse* failures of the target model, we create ensembles of subsets of pruned models [53] to compete with the target model in gMAD [29] (see Figure 1). We then jointly fine-tune the target and all pruned variants on the combination of the human-rated gMAD images and previously trained data. This allows all competing models to learn from their respective failures, and prepare themselves for the next round of gMAD competition.

Our method is the first of its kind to troubleshoot BIQA models in the wild. The fine-tuned model shows improved aggressiveness and resistance [29] in gMAD, comparing with itself in previous rounds. In addition, we find that the images in the gMAD sets exhibit increasing transferability to falsify existing BIQA models. Our code is publicly available at https://github.com/wangzhihua520/troubleshooting_BIQA.

2. Related Work

BIQA with DNNs Recently, there has been a surge of interest in developing BIQA models based on DNNs. A major challenge along this direction is to constrain the large set of network parameters using a small set of human-rated images with mean opinion scores (MOSs). Kang *et al*. [20] trained a DNN with one convolution layer on 32×32 patches to compensate for the lack of training data. Bosse *et al*. [3] developed a DNN with more convolution layers using the same patch training strategy. Ma *et al*. [30] leveraged the distortion identification as an auxiliary task to warm up the training. Kim *et al*. [21] used the error map from the Minkowski metric to regularize the training. Ma *et al*. [31] took a step further, and exploited multiple full-reference metrics as noisy annotators for training DNN-based BIQA models without MOSs. These methods are mostly designed to handle synthetic distortions [38, 26], with limited generalizability to realistic distortions [14, 18]. To meet the cross-distortion-scenario challenge, Zhang *et al*. [51] bilinearly pooled two feature representations that are sensitive to synthetic and realistic distortions, respectively. Zhang *et al*. [52] described a simple method to train BIQA models on multiple IQA datasets. The resulting UNIQUE model is capable of assessing image quality in the laboratory and wild, and will be used as the target model to demonstrate the feasibility of the proposed method.

Network Pruning DNNs commonly hinge on over-parameterization [28] and can be effectively compressed [23]. Network pruning [15] has been an effective technique to remove redundant computation at surprisingly little sacrifice of test accuracy. For example, Han *et al*. [15] proposed to prune DNNs by thresholding model *weights* based on their magnitudes. Li *et al*. [24] pruned DNN *filters* with small ℓ_1- or ℓ_2-norms. Liu *et al*. [27] encouraged channel sparsity by adding ℓ_1-constraints on the batch normalization scaling parameters. Molchanov *et al*. [35] estimated filter importance using Taylor expansion. He *et al*. [16] used geometry median to select the most redundant filters. A latest review is referred to [2].

Some researchers have started to rethink pruning beyond just an ad-hoc compression tool, and to explore its in-depth connection with DNN memorization/generalization. Frankle *et al*. [12] pioneered to show that there exist highly sparse "critical subnetworks" from the full DNNs, that can be trained in isolation from scratch. That critical subnetwork could be effectively identified by pruning [49, 13]. The most relevant work is due to Hooker *et al*. [17], who showed that pruning a trained image classifier tends to harm its performance more on the most difficult and long-tailed training images. This implies that pruning might effectively spot samples not well learned by the current model, and provides novel insights to exposing a trained model's potential weakness. We take inspiration from [17], and improve their method to troubleshoot BIQA models, by identifying and leveraging quality-discriminable images between pruned and non-pruned methods.

Active Learning The main idea of active learning is to mine the most valuable samples to label from a large unlabeled dataset [8]. In active learning for regression, query by committee (QBC) selects the most disagreed samples by a committee of models [37, 5]. Expected model change maximization (EMCM) selects samples that can cause the largest change to the current model [6]. Greedy sampling (GS) was originally proposed as a robust clustering method against

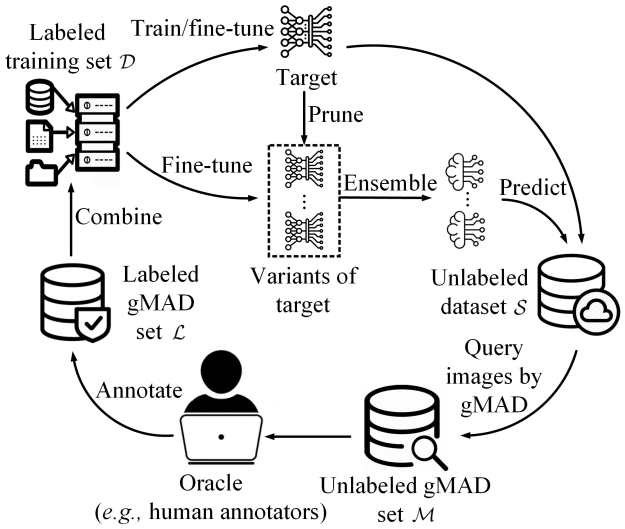


Figure 2: Diagram pf troubleshooting BIQA models in the wild. We start with a differentiable parametric target BIQA model, seek pairs of images by letting it compete with ensembles of pruned variants in gMAD [29], collect human scores for the gMAD set, fine-tune all models on the combination of the previously seen databases and the newly annotated gMAD set. The target and competing models co-evolve for the next round of troubleshooting.

outliers [1]. It was adapted to active learning in [45], with the goal of selecting samples that can increase the diversity of model responses. Residual active learning (RSAL) [9, 10] trains two models to fit the target outputs and the prediction residuals, respectively. The residual model is then used to select samples with the maximum predicted residuals. The procedure of identifying gMAD images [29] in this work can be seen as a form of active sampling with the criterion that the selected images have the greatest potential to falsify the target model. We will compare the error-spotting efficiency of several active learning methods in Section 4.3.

3. Proposed Method

We formulate the general problem of troubleshooting BIQA models in the wild as follows. We assume a strong off-the-shelf BIQA model f that has been trained on the labeled set $\mathcal{D}$ with images captured in the wild. Also assumed is a large-scale unlabeled set $\mathcal{S}$, containing images with much greater scene complexities and realistic distortions. The end goals are to identify diverse failures of f in $\mathcal{S}$ with a limited human labeling budget, and to leverage the exposed failures to further improve the generalizability of f.

The gMAD competition [29] suggests to select images that optimally distinguish f and a stronger competing model because those images are most likely to be its counterexamples. Then the core question is "how to obtain a set of diverse competing models for efficient failure-spotting?" In this paper, we create strong competing models, dubbed "self-competitors," from the target model by network pruning [17]. After obtaining the labeled gMAD set $\mathcal{L}$ through subjective testing, we jointly fine-tune the target and competing models on the combination of $\mathcal{D}$ and $\mathcal{L}$, attempting to learn from spotted failures without forgetting previously seen data [25]. Figure 2 illustrates the proposed diagram of troubleshooting BIQA models in the wild.

3.1. Self-Competitor Construction

The success of failure-spotting of the target model by gMAD [29] depends on the *strength* of the competing models. Here we resort to network pruning for competing model construction due to two reasons. First, DNN-based models are highly overparameterized. Therefore, the performance drop of pruned models on test sets is often insubstantial, and can be easily recovered if fine-tuning is allowed. Second, Hooker *et al*. [17] showed in the context of image classification that images that differentiate between the original and pruned classifiers are the least-memorized and weakest-learned ones by the original models. They appear the most challenging for both models, and sometimes even for humans to classify. In BIQA terms, these are likely to be selected by gMAD as the most informative images to falsify both the target and competing models.

Specifically, we first generate a list of pruned models $\{h_j\}_{j=1}^m$ from the target model f. Thanks to the prosperity of the network pruning field, we are able to leverage a diverse set of state-of-the-art network pruning techniques [15, 16, 24, 27, 35], with different hyperparameter settings, to encourage diversity among pruned models. Furthermore, as ensemble models have been shown to achieve stronger generalizability than individual models in many fields of machine learning [53], we create n ensemble models $\{g_i\}_{i=1}^n$ by randomly combining a subset of s models out of $\{h_j\}_{j=1}^m$:

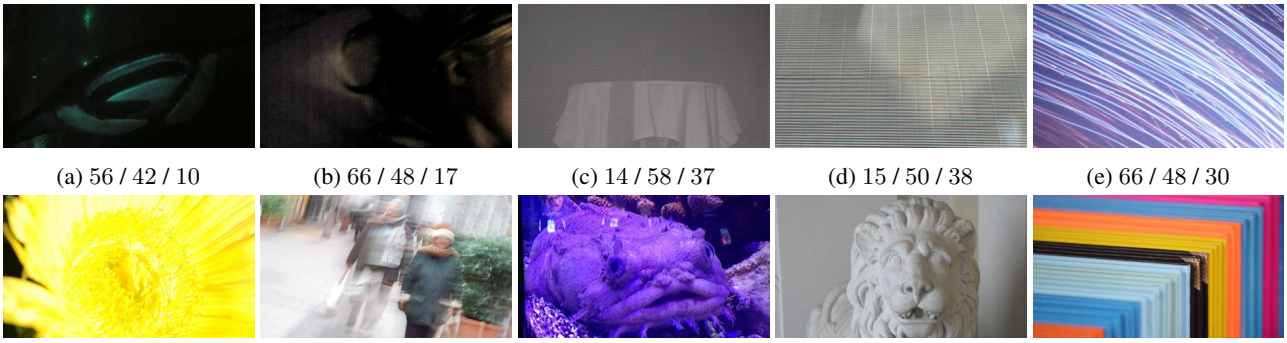
$$g_i(x) = \frac{1}{m}\sum_{j=1}^{m} \alpha_{ij} h_j(x), \quad i = 1, 2, \ldots, n, \tag{1}$$

where $\alpha_{ij} = 1$ if h_j is selected to create g_i, and $\sum_j \alpha_{ij} = s$. In Eq. (1), $\{h_j\}_{j=1}^n$ have been mapped to the same perceptual scale such that the weighted summation is legitimate.

3.2. Failure Identification

Given the large-scale unlabeled dataset $\mathcal{S}$, gMAD selects top-k pairs of images that best discriminate between the target model f and the competing model g_i:

$$(\hat{x}_{ik}, \hat{y}_{ik}) = \arg\max_{x,y} (g_i(x) - g_i(y)) - (f(x) - f(y))$$
$$\text{s.t. } x, y \in \mathcal{S} \setminus \{\hat{x}_{ij}, \hat{y}_{ij}\}_{j=1}^{k-1}, \tag{2}$$



(a) 56 / 42 / 10 (b) 66 / 48 / 17 (c) 14 / 58 / 37 (d) 15 / 50 / 38 (e) 66 / 48 / 30

(f) 58 / 37 / 25 (g) 51 / 29 / 20 (h) 71 / 48 / 42 (i) 26 / 38 / 69 (j) 16 / 35 / 55

Figure 3: Representative failures of the target model in gMAD. **(a)-(e)** Optimization results of Eq. (2) with $k = 10$ and $n = 10$. **(f)-(j)** Optimization results of Eq. (3) with $k = 1$ and $n = 120$, which appear much more diverse in scene content and distortion type. Below each image are scores from the target model / the competing models / humans.

where $\{\hat{x}_{ij}, \hat{y}_{ij}\}_{j=1}^{k-1}$ is the set of $k-1$ pairs of images that have been selected. The roles of f and g_i may be switched by replacing $\arg\max$ with $\arg\min$. However, recursive optimization of Eq. (2) may simply expose different instantiations of failures with the same underlying root causes (as shown in the first row of Figure 3). To encourage spotting diverse failures, a fine-grained version of gMAD can be formulated as

$$(\hat{x}_{ik}, \hat{y}_{ik}) = \arg\max_{x,y} (g_i(x) - g_i(y)) - (f(x) - f(y))$$
$$\text{s.t. } f(x), f(y) \in [a, b], \text{and } x, y \in \mathcal{S} \setminus \{\hat{x}_{ij}, \hat{y}_{ij}\}_{j=1}^{k-1}, \quad (3)$$

where $[a, b]$ defines a quality level, within which f predicts the two images x and y to have similar quality (see the second row of Figure 3). In this case, f is regarded as the *defender*, while g_i is the *attacker*. We may select several (non-overlapping) quality levels to cover the full quality spectrum. All image pairs selected by exhausting the competing models and the quality levels form the gMAD set $\mathcal{M}$, whose size is considerably smaller than the unlabeled set $\mathcal{S}$ and is adjustable to fit the available human labelling budget. In practice, it is possible to further diversify the spotted failures in $\mathcal{M}$ by decreasing k and increasing n provided that the created ensemble models differ to certain degrees (see Figure 3). Finally, we conduct subjective testing to collect human scores for each $(x, y) \in \mathcal{M}$, leading to four possible results:

- **Case I**. Both f and g_i are consistent with humans in ranking the perceived quality of x and y. This happens because f is a top-performing model, while g_i closely resembles f as its pruned version. We may reduce the possibility of this outcome by increasing the size of $\mathcal{S}$ and selecting more suitable network pruning algorithms.
- **Case II**. g_i is consistent with human perception, but f is not. In this case, the selected (x, y) constitutes a failure of f, and is informative for subsequent model rectification.
- **Case III**. f is consistent with human perception, but g_i is not. In this case, f successfully spots a counterexample of g_i, which seems to deviate from the original goal. However, it is worth noting that (x, y) is still useful in improving the performance of f because we are co-evolving f and g_i in the subsequent stage of model rectification. That is, g_i is also given the opportunity to learn from its failures, increasing the possibility of exposing f's weaknesses in the next round of the gMAD competition.
- **Case IV**. Neither f nor g_i is consistent with human perception. In this case, (x, y) manifests itself as a double-failure result, which is the most informative in improving the generalizability of f [17].

3.3. Model Rectification

The labeled gMAD set $\mathcal{L}$ exposes aspects of weaknesses of f, and thus is useful for improving its generalization to the real world. To avoid catastrophic forgetting [32], we choose to combine $\mathcal{L}$ and previously trained dataset $\mathcal{D}$, and jointly fine-tune f and $\{g_i\}_{i=1}^n$ on the combined set. By doing so, all models are able to learn from their respective failures and improve their generalizability for the next round of the gMAD competition. We may iterate this procedure of failure identification and model rectification several rounds, leading to a progressive human-in-the-loop troubleshooting method for BIQA models in the wild. We denote the target and competing models in the first round as $f^{(0)}$ and $\{g_i^{(0)}\}_{i=1}^n$ respectively. In the t-th round, we fine-tune $f^{(t-1)}$ and $\{g_i^{(t-1)}\}_{i=1}^n$ on the combination of $\mathcal{D}$ and $\mathcal{L} = \bigcup_{j=1}^t \mathcal{L}^{(j)}$. We summarize the proposed method in Algorithm 1.

Algorithm 1: Troubleshooting BIQA models in the wild

Input: A training set $\mathcal{D}$, a large-scale unlabeled image set $\mathcal{S}$, a target BIQA model $f^{(0)}$, the number r of fine-tuning rounds

Output: A troubleshot BIQA model $f^{(r)}$

Train/fine-tune $f^{(0)}$ on $\mathcal{D}$
Prune $f^{(0)}$ to generate $\{h_j^{(0)}\}_{j=1}^m$
for $j \leftarrow 1$ **to** m **do**
    Fine-tune $h_j^{(0)}$ on $\mathcal{D}$ to recover the performance
end
$\mathcal{L} \leftarrow \emptyset$
for $t \leftarrow 0$ **to** $r-1$ **do**
    Randomly ensemble subsets of $\{h_j^{(t)}\}_{j=1}^m$ to construct $\{g_i^{(t)}\}_{i=1}^n$
    Compute the responses of $f^{(t)}$ on $\mathcal{S}$
    $\mathcal{M}^{(t+1)} \leftarrow \emptyset$, $\mathcal{L}^{(t+1)} \leftarrow \emptyset$
    for $i \leftarrow 1$ **to** n **do**
        Compute the responses of $g_i^{(t)}$ on $\mathcal{S}$
        Seek pairs of images associated with $f^{(t)}$ and $g_i^{(t)}$ by solving Eq. (3), and include them in $\mathcal{M}^{(t+1)}$
    end
    Collect human scores for $\mathcal{M}^{(t+1)}$ to form $\mathcal{L}^{(t+1)}$
    $\mathcal{L} \leftarrow \mathcal{L} \bigcup \mathcal{L}^{(t+1)}$
    Fine-tune $f^{(t)}$ on $\mathcal{D} \bigcup \mathcal{L}$
    for $j \leftarrow 1$ **to** m **do**
        Fine-tune $h_j^{(t)}$ on $\mathcal{D} \bigcup \mathcal{L}$
    end
end

4. Experiments

In this section, we first describe the real experimental setups, and then provide quantitative and qualitative results to validate the feasibility of the proposed method, followed by an ablation study to test the failure-spotting efficiency of our method.

4.1. Experimental Setups

Target Model f We use UNIQUE [52], a state-of-the-art BIQA model with so far the best cross-distortion-scenario performance to our best knowledge. We retrain it on six IQA datasets, *i.e.*, LIVE [38], CSIQ [22], KADID-10k [26], BID [7], LIVE Challenge [14], and KonIQ-10k [18]. We leave 20% images for monitoring the performance changes of f during troubleshooting.

Unlabeled Dataset $\mathcal{S}$ To construct the large-scale dataset $\mathcal{S}$ for gMAD to seek potential failures of f, we first download $750,000$ images from the Internet followed by automatic pre-screening to remove duplicate and non-photographic images. Afterward, we sample $100,000$ images with marginal distributions nearly uniform with respect to image attributes, including bitrate, JPEG compression ratio, brightness, colorfulness, contrast, and sharpness [40]. Finally, we down-sample the images such that the long edge has $1,024$ pixels as a way of facilitating computational prediction. The constructed $\mathcal{S}$ includes a wide range of realistic camera distortions, such as sensor noise contamination, motion and out-of-focus blurring, under- and over-exposure, contrast reduction, color cast, and a mixture of these.

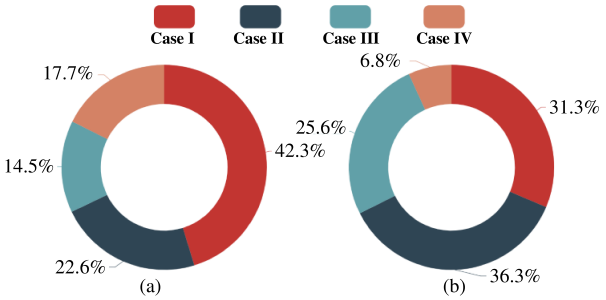


Figure 4: The empirical distributions of the gMAD pairs in **(a)** $\mathcal{L}^{(1)}$ and **(b)** $\mathcal{L}^{(2)}$, respectively.

Competing Models $\{g_i\}_{i=1}^n$ To encourage the diversity of the competing model pool, we adopt six network pruning algorithms: oneshot magnitude pruning (OMP) [15], ℓ_1-filter pruning [24], ℓ_2-filter pruning [24], TaylorFOWeight pruning [35], network slimming [27], and FPGM pruning [16], among which OMP is unstructured weight pruning, and the others are filter pruning. Fine-tuning is conducted after each model pruning method to recover the model performance. In addition, we use three different pruning ratios for each method, resulting in a total of $m = 18$ pruned models $\{h_j\}_{j=1}^{18}$ from f. We randomly combine $s = 8$ out of $m = 18$ pruned models, giving rises to $n = 120$ ensemble models $\{g_i\}_{i=1}^{120}$. Note that ensembling requires all pruned models to use the same perceptual scale. To achieve this, we map all model predictions onto the MOS scale $[0, 100]$ of the LIVE Challenge Database [14], with higher values indicating better perceptual quality. The fitted mapping function can be treated as part of the pruned model. As formulated in Eq. (1), ensembling is implemented by simple averaging, which gives all pruned models equal weights.

Labeled gMAD Set $\mathcal{L}$ To seek gMAD pairs, we set five quality levels, roughly covering the full quality range from "bad", "poor", "fair", "good", to "excellent". Two types of pairs are queried by treating the target model f in Eq. (3) as the defender and the attacker, respectively. We retain a

Table 1: Correlation between model predictions and MOSs on the test set of KonIQ-10k [18], $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$, respectively. Top section lists two knowledge-driven models. Middle section contains three DNN-based models retrained on the training set of KonIQ-10k. Note that the result of f on $\mathcal{L}^{(2)}$ is obtained by fine-tuning it on both $\mathcal{D}$ and $\mathcal{L}^{(1)}$.

SRCC	KonIQ-10k	$\mathcal{L}^{(1)}$	$\mathcal{L}^{(2)}$
NIQE [34]	0.521	0.340	0.293
HOSA [46]	0.520	0.336	0.287
DB-CNN [51]	0.806	0.690	0.641
MetaIQA [54]	0.841	0.801	0.751
HyperIQA [39]	0.902	0.802	0.765
UNIQUE [52] (as f)	0.862	0.570	0.553
PLCC	KonIQ-10k	$\mathcal{L}^{(1)}$	$\mathcal{L}^{(2)}$
NIQE	0.529	0.339	0.293
HOSA	0.519	0.327	0.275
DB-CNN	0.828	0.722	0.639
MetaIQA	0.878	0.799	0.742
HyperIQA	0.921	0.807	0.741
UNIQUE (as f)	0.875	0.565	0.555

single pair that best differentiates between f and g_i at each quality level by setting $k = 1$. We perform two rounds of troubleshooting (*i.e.*, $r = 2$), and the number of gMAD pairs in $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$ are $1,184$ and $1,194$, respectively. We gather human data from 20 subjects in an office environment with a calibrated display [41] using the single stimulus continuous quality rating. We process the raw subjective data using the outlier detection and subject rejection algorithm in [4]. We find that all subjects are valid, and 2.89% and 3.03% of ratings are outliers and subsequently removed.

Fine-Tuning Details For each round of model rectification, we fine-tune all 19 models, including the target model with the same optimization settings. Specifically, the Adam method is used with a learning rate of 10^{-5} and a mini-batch size of 32 - half from the previous training set $\mathcal{D}$ and half from the gMAD set $\mathcal{L}$. The maximum epoch number is set to ten. During fine-tuning, we re-scale and crop the images to 384×384. We test on images of original sizes. It takes about 162 GPU hours for each round of fine-tuning as measured on a machine with a single RTX 2080Ti.

4.2. Main Results

Failure Identification Table 1 lists the Spearman rank correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC) results between model predictions and MOSs on the gMAD sets $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$. We also include the performance on the test set of KonIQ-10k [18] for reference.

Table 2: Global ranking results of $f^{(0)}$, $f^{(1)}$ and $f^{(2)}$ in gMAD. A larger aggressiveness/resistance value indicates better performance [29].

	Aggressiveness	Resistance
$f^{(0)}$	-2.833	-0.053
$f^{(1)}$	1.366	0.019
$f^{(2)}$	1.467	0.034

Several aspects of the results are worth noting. First, the correlation numbers of f on $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$ are much lower than that on KonIQ-10k, indicating the effectiveness of our method in exposing failures of a "top-performing" BIQA model. Second, despite fine-tuned on $\mathcal{L}^{(1)}$, $f^{(1)}$ delivers slightly worse performance on $\mathcal{L}^{(2)}$ compared to $f^{(0)}$ on $\mathcal{L}^{(1)}$. This suggests that the co-evolving ensemble models are able to spot stronger errors of f in the second round of troubleshooting. Third, the identified counterexamples of f in each round show increasing transferability to falsify five existing BIQA models, as evidenced by larger performance drops on $\mathcal{L}^{(2)}$ than $\mathcal{L}^{(1)}$.

In Figure 4, we take a look at the empirical distributions of four possible results of gMAD pairs (see Section 3.2). Generally, ensemble models are more aggressive to falsify the target model, and are more resistant to the target's attacks as well. After the first round of fine-tuning, the failure-spotting capability of all models has been significantly improved.

Model Rectification We progressively fine-tune the target model $f^{(0)}$ on $\mathcal{D} \bigcup \mathcal{L}^{(1)}$ to obtain $f^{(1)}$, which is further fine-tuned on $\mathcal{D} \bigcup \mathcal{L}^{(1)} \bigcup \mathcal{L}^{(2)}$ to obtain $f^{(2)}$. To verify the relative improvements resulting from model rectification, we let $f^{(0)}$, $f^{(1)}$, and $f^{(2)}$ play the gMAD game against one other on $\mathcal{S} \setminus \mathcal{L}$. In each of $\binom{3}{1}$ competitions, we select 100 gMAD pairs at five quality levels for human annotation. As suggested in [29], we aggregate the paired comparison results into two global ranking vectors to indicate how *aggressive* one model is to falsify other models as the attacker and how *resistant* one model is to survive other models' attacks as the defender. Table 2 shows the global ranking results with higher values indicating better performance. It is easy to conclude that f continually evolves to be a better model in terms of both aggressiveness and resistance in gMAD without forgetting previously seen data. This verifies the feasibility of the proposed scheme to troubleshoot BIQA models in the wild.

We next compare $f^{(0)}$, $f^{(1)}$, and $f^{(2)}$ qualitatively. Figure 5 shows four representative gMAD pairs between $f^{(0)}$ and $f^{(1)}$. It is clear that the pairs of images in (a) and (b) exhibit substantially different quality, which is in disagreement with $f^{(0)}$. In contrast, $f^{(1)}$ correctly predicts top images to have

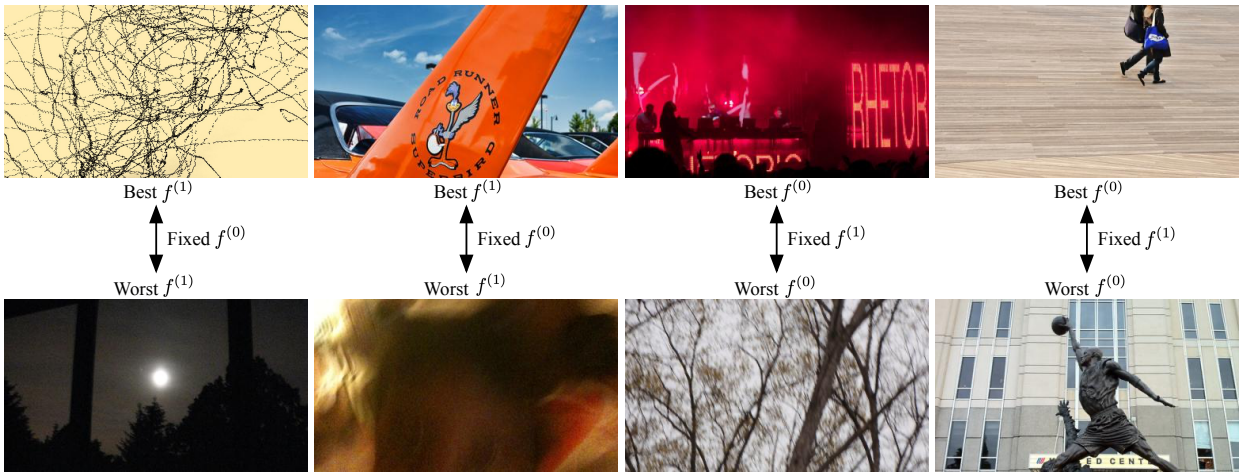


Figure 5: Representative gMAD pairs between $f^{(0)}$ and $f^{(1)}$. **(a)** Fixing $f^{(0)}$ at the low quality level. **(b)** Fixing $f^{(0)}$ at the high quality level. **(c)** Fixing $f^{(1)}$ at the low quality level. **(d)** Fixing $f^{(1)}$ at the high quality level.

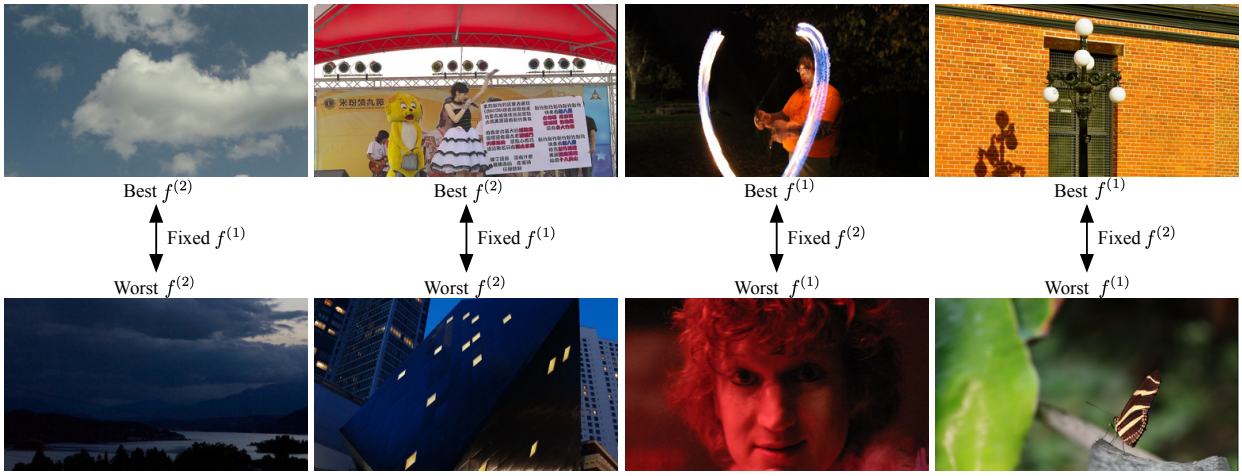


Figure 6: Representative gMAD pairs between $f^{(1)}$ and $f^{(2)}$. **(a)** Fixing $f^{(1)}$ at the low quality level. **(b)** Fixing $f^{(1)}$ at the high quality level. **(c)** Fixing $f^{(2)}$ at the low quality level. **(d)** Fixing $f^{(2)}$ at the high quality level.

much better quality than bottom images. When the roles of $f^{(0)}$ and $f^{(1)}$ are reversed, $f^{(0)}$ still fails to expose failures of $f^{(1)}$ (see (c) and (d)), suggesting that $f^{(1)}$ is significantly improved by learning from the gMAD set in the first round.

Figure 6 depicts four gMAD competition results between $f^{(1)}$ and $f^{(2)}$. In (a) and (b), we observe that $f^{(2)}$ is able to falsify $f^{(1)}$ by finding its counterexamples which appear dark, but the perceptual gaps between best and worst cases are not as large as those when $f^{(1)}$ attacks $f^{(0)}$. In (c) and (d), $f^{(2)}$ successfully survives the attacks from $f^{(1)}$, with pairs of images of similar quality according to human perception. This indicates that perceptual gains from the second round of fine-tuning are not as substantial as those in the first round, which is a common phenomenon in active learning.

We also show four gMAD pairs between $f^{(0)}$ and $f^{(2)}$ in Figure 7 to further demonstrate the improvements of $f^{(0)}$ after two rounds of troubleshooting. $f^{(2)}$ favors the top images in (a) and (b), which is consistent with human judgments, suggesting that $f^{(2)}$ successfully attacks $f^{(0)}$. $f^{(0)}$ fails to penalize the top image in (c) and (d), which are spotted by $f^{(2)}$. This further validates the improved generalizability of f to the real world.

4.3. Ablation Study

In this subsection, we show that gMAD sampling in our method has stronger failure-spotting capability compared

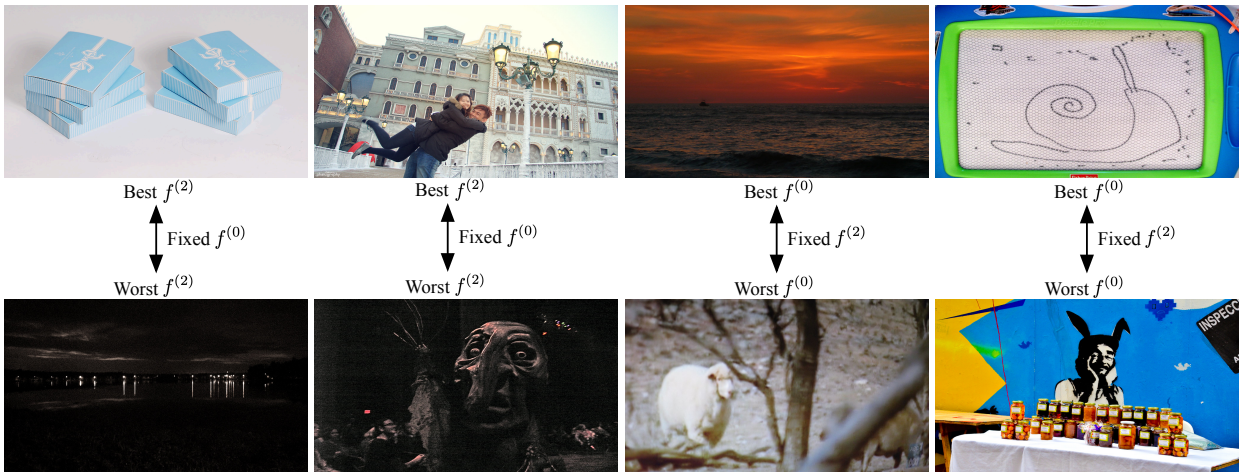


Figure 7: Representative gMAD pairs between $f^{(0)}$ and $f^{(2)}$. **(a)** Fixing $f^{(0)}$ at the low quality level. **(b)** Fixing $f^{(0)}$ at the high quality level. **(c)** Fixing $f^{(2)}$ at the low quality level. **(d)** Fixing $f^{(2)}$ at the high quality level.

Table 3: Comparison of the failure-spotting efficiency between gMAD sampling and five active learning methods on the SPAQ database [11]. A lower correlation coefficient indicates better performance.

Method	SRCC	PLCC
Random	0.729	0.776
QBC	0.180	0.334
RSAL	0.629	0.700
EMCM	0.547	0.584
GS	0.395	0.392
Ours	**0.137**	**0.297**

to five active learning methods for regression: random sampling, QBC [37], EMCM [6], RSAL [9, 10], and GS [1]. We conduct experiments on the smartphone photography attribute and quality (SPAQ) dataset [11], which contains $11,125$ human-rated images captured by 66 smartphones. We sample a subset of 200 images by each method, and compute the SRCC and PLCC between MOSs and predictions by $f^{(0)}$ (*i.e.*, UNIQUE [52]). Table 3 shows the results, with a lower correlation coefficient indicating better performance. As can be seen, the images selected by gMAD [29] lead to the worst performance of $f^{(0)}$, among all methods, which shows the failure-spotting capability of the proposed gMAD sampling.

5. Conclusion

We have introduced a computational method for progressively troubleshooting BIQA models in the wild. The key to success of our method is to construct strong "self-competitors" as random ensembles of pruned versions of a "top-performing" target model. We have demonstrated the effectiveness of the ensemble models in exposing diverse counterexamples of the target model in the gMAD competition. A second advantage of our method is the flexibility to co-evolve the target and competing models, which allows all models to learn from their respective failures, making progressively troubleshooting the target model more effective.

Our work extends a new line of research in BIQA with many important topics to be explored. For example, the current work only performs two rounds of troubleshooting due to the limited human labeling budget. Nevertheless, it is interesting to mathematically analyze the convergence of the proposed method or come up with a practical stop criterion to guide the setting of the fine-tuning round. Moreover, the computational complexity of constructing competing models, *i.e.*, pruning followed by fine-tuning, is relatively high. It is thus worth exploring more computationally efficient methods, *e.g.*, snapshot ensembles [19] for competing model construction. Another future direction is to extend the current work to troubleshoot BIQA models in the laboratory and wild, towards universal and generalizable BIQA.

Acknowledgments

The authors would like to thank all subjects who participated in our subjective study during this period of the coronavirus pandemic. This work was supported in part by the National Natural Science Foundation of China (62071407), and the CityU SRG-Fd and APRC Grants (7005560 and 9610487).

References

[1] A. Bhaskara, S. Vadgama, and H. Xu. Greedy sampling for approximate clustering in the presence of outliers. In *Advances in Neural Information Processing Systems*, pages 11148–11157, 2019. 3, 8

[2] D. Blalock, J. J. G. Ortiz, J. Frankle, and J. Guttag. What is the state of neural network pruning? *arXiv preprint arXiv:2003.03033*, 2020. 2

[3] 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*, 27(1):206–219, 2017. 1, 2

[4] BT.500, ITU-R. Methodology for the subjective assessment of the quality of television pictures. 2002. 6

[5] R. Burbidge, J. J. Rowland, and R. D. King. Active learning for regression based on query by committee. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 209–218, 2007. 2

[6] W. Cai, Y. Zhang, and J. Zhou. Maximizing expected model change for active learning in regression. In *International Conference on Data Mining*, pages 51–60, 2013. 2, 8

[7] A. Ciancio, E. A. da Silva, A. Said, R. Samadani, and P. Obrador. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on Image Processing*, 20(1):64–75, 2011. 5

[8] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994. 2

[9] F. Douak, N. Benoudjit, and F. Melgani. A two-stage regression approach for spectroscopic quantitative analysis. *Chemometrics and Intelligent Laboratory Systems*, 109(1):34–41, 2011. 3, 8

[10] F. Douak, F. Melgani, and N. Benoudjit. Kernel ridge regression with active learning for wind speed prediction. *Applied Energy*, 103:328 – 340, 2013. 3, 8

[11] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang. Perceptual quality assessment of smartphone photography. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020. 8

[12] J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2018. 2

[13] J. Frankle, G. K. Dziugaite, D. M. Roy, and M. Carbin. The lottery ticket hypothesis at scale. *arXiv preprint arXiv:1903.01611*, 8, 2019. 2

[14] D. Ghadiyaram and A. C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015. 2, 5

[15] S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, pages 1135–1143, 2015. 2, 3, 5

[16] Y. He, P. Liu, Z. Wang, Z. Hu, and Y. Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4340–4349, 2019. 2, 3, 5

[17] S. Hooker, A. Courville, G. Clark, Y. Dauphin, and A. Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019. 2, 3, 4

[18] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 2, 5, 6

[19] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger. Snapshot ensembles: Train 1, get M for free. In *International Conference on Learning Representations*, 2017. 8

[20] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014. 2

[21] J. Kim, A. Nguyen, and S. Lee. Deep CNN-based blind image quality predictor. *IEEE Transactions on Neural Networks*, 30(1):11–24, 2019. 2

[22] E. C. Larson and D. M. Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):1–21, 2010. 5

[23] Y. LeCun, J. S. Denker, and S. A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, pages 598–605, 1990. 2

[24] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf. Pruning filters for efficient ConvNets. In *International Conference on Learning Representations*, 2017. 2, 3, 5

[25] Z. Li and D. Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017. 3

[26] H. Lin, V. Hosu, and D. Saupe. KADID-10k: A large-scale artificially distorted IQA database. In *IEEE International Conference on Quality of Multimedia Experience*, pages 1–3, 2019. 2, 5

[27] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang. Learning efficient convolutional networks through network slimming. In *IEEE International Conference on Computer Vision*, pages 2736–2744, 2017. 2, 3, 5

[28] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019. 2

[29] K. Ma, Z. Duanmu, Z. Wang, Q. Wu, W. Liu, H. Yong, H. Li, and L. Zhang. Group maximum differentiation competition: Model comparison with few samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):851–864, 2020. 1, 2, 3, 6, 8

[30] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2018. 1, 2

[31] K. Ma, X. Liu, Y. Fang, and E. P. Simoncelli. Blind image quality assessment by learning from multiple annotators. In *IEEE International Conference on Image Processing*, pages 2344–2348, 2019. 2

[32] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press, 1989. 4

[33] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 1

[34] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 1, 6

[35] P. Molchanov, A. Mallya, S. Tyree, I. Frosio, and J. Kautz. Importance estimation for neural network pruning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11264–11272, 2019. 2, 3, 5

[36] W. J. Scheirer and T. E. Boult. A fusion-based approach to enhancing multi-modal biometric recognition system failure prediction and overall performance. In *IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–7, 2008. 1

[37] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Annual Workshop on Computational Learning Theory*, pages 287–294, 1992. 2, 8

[38] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006. 2, 5

[39] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 6

[40] V. Vonikakis, R. Subramanian, and S. Winkler. Shaping datasets: Optimal data selection for specific target distributions across dimensions. In *IEEE International Conference on Image Processing*, pages 3753–3757, 2016. 5

[41] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment, 2000. 6

[42] Z. Wang and A. C. Bovik. Reduced-and no-reference image quality assessment. *IEEE Signal Processing Magazine*, 28(6):29–40, 2011. 1

[43] Z. Wang and K. Ma. Active fine-tuning from gMAD examples improves blind image quality assessment. *arXiv preprint arXiv:2003.03849*, 2020. 1, 2

[44] Z. Wang and E. P. Simoncelli. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12):1–13, 2008. 1

[45] D. Wu, C.-T. Lin, and J. Huang. Active learning for regression using greedy sampling. *Information Sciences*, 474:90–105, 2019. 3

[46] 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*, 25(9):4444–4457, 2016. 6

[47] P. Ye, J. Kumar, L. Kang, and D. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1098–1105, 2012. 1

[48] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3575–3585, 2020. 1

[49] H. You, C. Li, P. Xu, Y. Fu, Y. Wang, X. Chen, R. G. Baraniuk, Z. Wang, and Y. Lin. Drawing early-bird tickets: Toward more efficient training of deep networks. In *International Conference on Learning Representations*, 2019. 2

[50] P. Zhang, J. Wang, A. Farhadi, M. Hebert, and D. Parikh. Predicting failures of vision systems. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3566–3573, 2014. 1

[51] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2020. 2, 6

[52] W. Zhang, K. Ma, G. Zhai, and X. Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *arXiv preprint arXiv:2005.13983*, 2020. 1, 2, 5, 6, 8

[53] Z.-H. Zhou. *Ensemble Methods: Foundations and Algorithms*. CRC Press, 2012. 2, 3

[54] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi. MetaIQA: Deep meta-learning for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14143–14152. 6