Perceptual Attacks of No-Reference Image Quality Models with Human-in-the-Loop

Anonymous Author(s)
Affiliation
Address
email

Abstract

No-reference image quality assessment (NR-IQA) aims to quantify how humans perceive visual distortions of digital images without access to their undistorted references. NR-IQA models are extensively studied in computational vision, and are widely used for performance evaluation and perceptual optimization of man-made vision systems. Here we make one of the first attempts to examine the perceptual robustness of NR-IQA models. Under a Lagrangian formulation, we identify insightful connections of the proposed perceptual attack to previous beautiful ideas in computer vision and machine learning, including the Carlini-Wagner attack, the maximum a posterior estimation and more. We test one knowledge-driven and three data-driven NR-IQA methods under four full-reference IQA models (as approximations to human perception of just-noticeable difference). Through carefully designed psychophysical experiments, we find that all four NR-IQA models are vulnerable to the proposed perceptual attack. More interestingly, we observe that the generated counterexamples are not transferable, manifesting themselves as distinct design flows of respective NR-IQA methods.

1 Introduction

Over the past decade, deep neural networks (DNNs) have revolutionized a wide range of computer vision and machine learning applications. For example, ResNet [1] reduced the object recognition error rate on ImageNet [2] to 3.56%, surpassing the performance of informed humans. A similar story is also written in the field of image quality assessment (IQA), where DNN-based IQA models [3, 4, 5, 6] exhibit very high correlations with human opinions of perceptual quality on various subject-rated datasets [7, 8, 9, 5]. IQA models can be roughly classified into full-reference IQA (FR-IQA) and no-reference IQA (NR-IQA) ones depending on the availability of the pristine undistorted image as a reference. Representative FR-IQA methods include the Minkowski distance (*i.e.*, the ℓ_p-norm induced metric), the structural similarity (SSIM) index [10], the learned perceptual image patch similarity (LPIPS) method [3], and the deep image structure and texture similarity (DISTS) model [4], which are widely used for measuring signal fidelity and quality in various vision applications. NR-IQA models [11] are extensively studied in computational vision, which mimics the human ability to judge the perceptual quality of a test image without comparison to any reference image. NR-IQA plays an indispensable role in the design and optimization of real-world image processing algorithms.

Despite the remarkable achievements of DNN-based models, recent work has identified their vulnerability to adversarial perturbations [13, 14]. For example, in natural image classification, a visually indistinguishable perturbation added to a natural image would mislead "top-performance" classifiers. This imperceptible perturbation is called below the just-noticeable difference (JND) of human perception in psychophysics, and is less relevant in image classification [15]. This is because for the majority of practical adversarial attacks [16] considered in image classification (*e.g.*, the ℓ_∞-norm

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

(a) (b)

Figure 1: Visual comparison of **(a)** a natural photographic image and **(b)** a computer-generated example from (a) by attacking BRISQUE [12] using the projected gradient method, under an ℓ_∞-norm constraint with a radius of $8/255$. It is clear that the semantic information of the image remains intact, while its perceptual quality degrades due to the introduced "mosquito" noise.

constrained attack), the allowable perturbations, even if above JNDs, do not alter the image semantics: they are label-preserving. In the context of IQA, it becomes highly nontrivial to craft label-preserving attacks for two main reasons. First, for adversarial perturbations that are above JNDs, they are highly likely to be perceived as some form of visual distortions that lead to quality degradation (see Fig. 1). Second, computational prediction of JNDs for natural images [17] remains an open research problem as it depends on the combination of the image content and the perturbation type, constrained by the psychophysical experimental conditions (*e.g.*, the maximum time of visual inspection).

In this work, we take initial steps to examine the perceptual robustness of NR-IQA models. Our main contributions are threefold.

- We propose a two-step perceptual attack for NR-IQA methods with human-in-the-loop. First, we define the objective of the perceptual attack as a Lagrangian function of an FR-IQA model (as the "perceptual" constraint) and the NR-IQA model (to be examined). By varying the Lagrange multiplier, we generate a series of perturbed images of different distortion visibility. Second, we ask human observers in a well-controlled psychophysical experiment to determine whether the perturbation of each image is discriminable. The output of our perceptual attack is a computer-generated counterexample, that is below the JND while leading to the most significant change in quality prediction.
- We draw connections between the proposed perceptual attack and conventional wisdom, including the Carlini-Wagner attack [18] in classification, maximum a posterior (MAP) estimation, maximum differentiation (MAD) competition [19], and eigen-distortion analysis [20].
- We conduct an extensive experiment to examine four NR-IQA models, the knowledge-driven BRISQUE [12], the shallow learning-based CORNIA [21], and the deep learning-based Ma19 [22] and UNIQUE [6] under four FR-IQA models, the Chebyshev distance (*i.e.*, the ℓ_∞-norm induced metric), SSIM, LPIPS, and DISTS (as approximations to human perception of JND). We arrive at several important observations, among which the most interesting one is that the proposed perceptual attack succeeds in fooling all four NR-IQA models, but the generated counterexamples are not transferable, manifesting themselves as distinct design flows of respective NR-IQA methods.

2 Related Work

In this section, we give a review of NR-IQA models, and summarize representative adversarial attacks in classification and discuss them in a broader context of "analysis by synthesis" [23, 24].

2.1 NR-IQA Models

Early NR-IQA models were designed to deal with specific distortion types, *e.g.*, JPEG compression [25] and JPEG2000 compression [26]. In the past decade, research on general-purpose NR-IQA become popular with the proposal of a variety of quality-aware features based on natural scene statistics (NSS) [12, 11, 27]. Unsupervised feature learning such as codebook construction [21] was also explored. Since the work of [28], DNN-based methods began to revolutionize the field of NR-IQA.

Towards developing more accurate NR-IQA models, recent research includes joint learning from multiple databases [6], active learning for improved generalizability [29], patch-to-picture learning for local quality prediction [30], and continual learning for handling streaming visual data [31].

2.2 Adversarial Attacks in Classification

Machine learning models are, for a long time, known to be vulnerable to adversarial examples [13, 32, 33, 34]: data samples that have been modified very slightly but enough to falsify a machine learning model. Most frequently, adversarial examples are automatically generated by projected gradient-based methods. Szegedy *et al*. [14] used a box-constrained L-BFGS method to discover the prevalence of adversarial examples in DNN-based classifiers. Goodfellow *et al*. [35] introduced a fast gradient sign method (FGSM), which modifies samples along the steepest descent direction under the ℓ_∞-norm constraint. Kurakin *et al*. [36] provided an extension to FGSM, which iteratively generates adversarial examples with a small step size. Madry *et al*. [37] formulated adversarial training as robust optimization [38], where adversarial example generation corresponds to solving the inner maximization problem by projected gradient descent (PGD). The authors recommended a practical trick to construct more transferable adversarial examples: adding a small amount of uniform noise to each pixel of the initial image as a form of dequantization. Moosavi *et al*. [39] defined the adversarial attack as the process to find the minimal perturbation that leads to erroneous model behavior. Close to the present work, Carlini and Wagner [18] considered a Lagrangian relaxation of the ℓ_p-norm constrained optimization problem to search for adversarial examples. Laidlaw *et al*. [40] switched to a more perceptual FR-IQA model, LPIPS [3], under which adversarial training gives improved robustness to unseen attacks. Here, we consider the Lagrangian formulation in a different context (*i.e*., NR-IQA), for a different purpose (*i.e*., encouraging generating samples below JNDs), and under different FR-IQA models (*i.e*., the Chebyshev distance, SSIM, LPIPS, and DISTS).

We conclude this section by putting adversarial attacks in a broader context of "analysis by synthesis", which is a core idea in the pattern theory by [24]. Analysis by synthesis suggests to test a machine learning model in generative rather than discriminative ways, which is well demonstrated in the field of texture modeling [41]. Adversarial attacks can be seen as a form of analysis by synthesis, where the machine learning model is tested on the automatically generated worst-case data samples, which are less likely to be included in the fixed and excessively reused test sets [42]. In the context of computational vision, the MAD competition [19], the eigen-distortion analysis [20], and the controversial stimuli synthesis [43] all fall into the category of analysis by synthesis. As one of our main contributions, we will make insightful connections of the proposed perceptual attack to these methods.

3 Perceptual Attacks on NR-IQA Models

We first formulate the perceptual attack in NR-IQA, and then put it in proper context by making connections to previously related techniques from different engineering fields.

3.1 Problem Formulation

An adversarial attack aims to modify a benign data sample x_0 to $x^\star$ such that the prediction of a machine learning model $f_w(\cdot)$, parameterized by a vector w, undesirably deviates from the ground-truth label $c = f(x_0)$, where $f(\cdot)$ denotes the underlying true hypothesis. This problem can be generally formulated as

$$x^\star = \arg\max_x L(f_w(x), f(x_0)), \text{ s.t. } D(x, x_0) \leq T, \tag{1}$$

where $L(\cdot,\cdot)$ denotes the loss function of a particular machine learning task, $D(\cdot,\cdot)$ is a signal fidelity metric to define the feasible set of perturbations, and T constrains the maximum magnitude of all possible perturbations. In image classification, one may adopt the cross-entropy function as L (or simply the negative logit of the ground-truth category for untargeted attacks and the logit difference of the specified and the ground-truth categories for targeted attacks). ℓ_p-norm induced metrics are frequently used to implement D. We refer readers to [44, 45, 46, 47, 48, 49] for instantiations of Eq. (1) for attacking other vision tasks.

It is tempting to reuse Eq. (1) to instantiate the adversarial attack on NR-IQA models, where $f_w : \mathbb{R}^M \mapsto \mathbb{R}$ takes an image $x \in \mathbb{R}^M$ as input, and computes a real number as the quality estimate.

Without loss of generality, we assume a larger $f_w(x)$ indicates higher predicted quality. The true perceptual quality, $f(x)$, referred to as the mean opinion score (MOS) of x, can be collected via a standard psychophysical experiment. As a regression task, we may specify L as some discrepancy function between the quality prediction of the perturbed image, $f_w(x)$, and the MOS of the initial image $f(x_0)$, *e.g.*, the mean squared error (MSE), $(f_w(x) - f(x_0))^2$. D can be implemented by any "perceptual" FR-IQA model.

We point out three caveats of this constrained optimization. First, existing FR-IQA models only provide a rough account for human perception of image quality. For example, the α-level set of D w.r.t. x_0, *i.e.*, $\{x|D(x, x_0) = \alpha\}$, may contain two images of drastically different quality. Computational models that are specifically designed for measuring JNDs [17, 50, 51, 52, 53, 54, 55], are only tested under simple perturbations (*e.g.*, noise shaping and compression), and may have their own adversarial examples. As a consequence, the feasible set $D(x, x_0) \leq T$ is little likely to have the quality-preserving property. Second, working with an imperfect FR-IQA model requires setting a content-dependent threshold $T(x_0)$ to ensure a quality-preserving feasible set. However, this is a highly non-trivial task, and may result in a trivial solution (*e.g.*, $T(x_0) \to 0$ due to the inaccuracy of D). Third, solving the constrained optimization problem using PGD (even with careful step size scheduling) appears to converge very slowly (due to the highly nonlinearity and nonconvexity of D). In addition, PGD tends to make the constraint tight with equality and thus is less likely to promote images that are below JNDs. All these motivate us to consider a Lagrangian relaxation of Problem (1):

$$x^\star = \arg\max_x - D(x, x_0) + \lambda(f_w(x) - f(x_0))^2, \tag{2}$$

where λ is the Lagrange multiplier. An immediate advantage of Eq. (2) is that it allows simultaneous maximization of the discrepancy term measured by the MSE and minimization of the FR-IQA term as the fidelity constraint, which encourages perturbed images to be below JNDs. An alternative way of viewing Eq. (2) is that we perturb x_0 in the quality increase and decrease directions specified by the NR-IQA model f_w, respectively, under the constraint of D, and we choose to optimize along the direction that leads to the best discrepancy-fidelity trade-off. Empirically, we find that, if x_0 is of poor quality, it is easier to expose counterexamples in the direction of quality improvement, and vice versa. Meanwhile, it is noteworthy that our discrepancy term relies on the MOS of the initial image, $f(x_0)$, which permits naturally occurring failure examples. This is the case when f_w makes a poor quality prediction of x_0, leading to large discrepancy and high fidelity. Nevertheless, $f(x_0)$ can be replaced by $f_w(x_0)$ if they are close.

3.2 Perceptually Imperceptible Counterexample Generation

We describe a variant of the steepest ascent method to solve Problem (2). Given the initial image x_0, we first add to each pixel some noise sampled uniformly from a discrete set $\{-1/255, 0, 1/255\}$ to trigger the optimization [37]. We then compute the steepest ascent direction g of the objective w.r.t. x, and take a step of γ along the direction of ℓ_p-normalized g for faster convergence. The pixel values of the intermediate image are clamped into the valid range, *i.e.*, $[0, 1]$, whenever necessary. After the maximum number of iterations is reached, we quantize it to obtain a 24-bit color image as the output.

By varying the Lagrange multiplier λ, we are able to generate a sequence of images with different distortion visibility. We next design a psychophysical experiment to identify the perceptually imperceptible counterexample, giving careful treatment to two subtleties. First, the distortion visibility generally reduces with the increase of λ and the decrease of D (see Eq. (2)), but there are exceptions due to the high dimensionality and non-convexity of the optimization problem. Therefore, standard psychophysical staircase methods [56] that assume monotonicity w.r.t. the stimulus intensity may not be straightforwardly applied. Second, ideally, we would like to measure the just-noticeable distortion[1], but practically, it is difficult to collect such human thresholds in a reliable way unless excessive instructions are given. This is especially true when x_0 has already been severely distorted (*e.g.*, blurred or compressed), and it is no easy task to determine whether the added perturbation further degrades the image quality.

We thus choose the yes-no task for screening perturbed images below JNDs [57]. Specifically, human participants undergo a series of trials, each consisting of a pair of the perturbed image (corresponding

[1] In some situations, the visible differences may not be perceived as distortions by the human visual system (*e.g.*, image enhancement).

Algorithm 1 Perceptually Imperceptible Counterexample Generation

Require: An NR-IQA model $f_w(\cdot)$, an FR-IQA model $D(\cdot,\cdot)$ as the fidelity measure, a set of K hyperparameter values $\{\lambda_i\}_{i=1}^K$, and the step size γ
Input: An initial image x_0
Output: A perceptually imperceptible counterexample $x^\star$

for $i = 1 \rightarrow K$ **do**
  $x \leftarrow x_0 + \epsilon$, where ϵ is randomly sampled from $\{-1/255, 0, 1/255\}$
  while the maximum iteration is not reached **do**
    $\lambda \leftarrow \lambda_i$
    Compute the objective value $J \leftarrow -D(x, x_0) + \lambda(f_w(x) - f(x_0))^2$
    Compute the steepest ascent direction $g \leftarrow \arg\max_v \{\nabla_x J^T v \mid \|v\|_p = 1\}$
    Update the image $x \leftarrow x + \gamma \cdot g / \|g\|_p$
    Clamp x into $[0, 1]$
  $y_i \leftarrow x$
  Quantize y_i to a 24-bit color image
Identify the perceptually imperceptible counterexample $x^\star$ from $\{y_i\}_{i=1}^K$ in a psychophysical experiment

to a certain λ) and the initial image. For each trial, they must judge whether the two images look identical. A perturbed image is said to be below the JND if participants fail to distinguish it from the initial image 75% of the time. *We identify the perceptually imperceptible counterexample as the perturbed image that 1) is below the JND and 2) causes the largest change in quality prediction.* More details about the psychophysical experiment can be found in Sec. 4.1, and the procedure of counterexample generation of NR-IQA models is summarized in Algorithm 1.

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

Figure 2: "Enhanced" versions from **(a)** the initial image by maximizing **(b)** BRISQUE, **(c)** CORNIA, **(d)** Ma19, and **(e)** UNIQUE, respectively.

3.3 Connections to Previous Work

In this subsection, we make insightful connections to several beautiful ideas in signal processing, computational vision, and machine learning.

Carlini-Wagner attack [18] has the closest relationship to the proposed perceptual attack on NR-IQA models with three differences. First, Carlini and Wagner chose to work with the Lagrangian formulation mainly from the perspective of *computational convenience*, while we arrive at the same formulation from a different *perceptual* perspective: no computational models exist to reliably compute JNDs. Second, for classification, there are $C-1$ directions for the Carlini-Wagner attack to look for adversarial examples, where C is the number of categories. For NR-IQA as regression, there are two such directions. Third, the Carlini-Wagner attack suffices to employ the binary search to find the desired Lagrange multiplier due to the label-preserving property of the majority perturbations in image classification. In contrast, the proposed perceptual attack may have to rely on carefully designed psychophysical experiments to determine the value of λ in NR-IQA, which is common practice in computational vision.

MAP estimation [58, 59, 60, 61] seeks to estimate a latent image $x^\star$ given a corrupted observation x_0 by maximizing the posterior probability $p(x|x_0)$:

$$x^\star = \arg\max_x p(x|x_0) = \arg\max_x \log p(x_0|x) + \lambda \log p(x), \tag{3}$$

where $\log p(x_0|x)$ is the data fidelity term implemented by an FR-IQA model (*e.g.*, the MSE) and $\log p(x)$ is the image prior term to measure the naturalness of the image. A better prior leads to improved estimation performance. The added hyperparameter λ is often manually tuned, where $\lambda = 1$ makes Eq. (3) hold. It is widely acknowledged that an ideal NR-IQA model must rely solely on knowledge of the appearance of natural undistorted images. This suggests that the NR-IQA method should embody an image prior model, and perhaps even that the quality predictions should be monotonically related to probability densities [22]. With such conceptual equivalence between NR-IQA methods and natural image priors, we see that the proposed perceptual attack differs only from MAP estimation in the choice of λ. Specifically, our attack sets a λ to produce the image below the JND relative to x_0 while causing the most significant change of f_w. In contrast, MAP estimation tests $p(x)$ using a λ that leads to the best estimation performance. More importantly, we shall come to a conclusion: in order for NR-IQA models to work in MAP estimation, they must first survive the proposed perceptual attack as an easier task. Taking a step further, we may set $\lambda = +\infty$, which corresponds to the maximization of f_w alone as a way of performing image enhancement. Not surprisingly, all state-of-the-art NR-IQA models fail this task, generating images with annoying distortions but higher predicted quality scores (see Fig. 2).

MAD competition [19] is an efficient methodology for comparing computational methods of perceptually discriminable quantities (*e.g.*, image quality). In the context of comparing two IQA models, f_1 and f_2, MAD automatically synthesizes an image pair $(x^\star, y^\star)$ that are likely to falsify at least one IQA model in competition by solving

$$(x^\star, y^\star) = \arg\max_{x,y} \; f_1(x) - f_1(y), \text{ s.t. } f_2(x) = f_2(y) = f_2(x_0), \tag{4}$$

where x_0 is the initial image and $f_2(x_0)$ specifies a quality level. By doing so, we obtain a pair of counterexamples, on which f_1 and f_2 hold contradictory opinions: f_1 predicts $x^\star$ to have much better quality than $y^\star$, while f_2 treats them as images of identical quality. Similar as the proposed perceptual attack, psychophysical testing on $(x^\star, y^\star)$ is required to declare which model is the winner. In contrast to our attack that is constrained in the set of images that look identical to x_0 according to human perception (approximated by an FR-IQA model), MAD operates at the level set of another IQA model f_2 that contains images of the same predicted quality as x_0. Moreover, MAD generates the best-quality and worst-quality images in terms of f_1 to compare the relative performance of two IQA models, while we only tests the perceptual robustness of a single NR-IQA model.

Eigen-distortion analysis [20] is a computational method for comparing perceptual image representations, $f : \mathbb{R}^M \mapsto \mathbb{R}^N$. It computes the eigenvectors of the Fisher information matrix[2] with the largest and smallest eigenvalues to account for the model-predicted most- and least-noticeable distortion directions, respectively. A psychophysical experiment based on a standard staircase method is necessary to measure the ratio of the visibility thresholds induced by the two directions. When $N = 1$ (*i.e.*, f can be an NR-IQA method), the most-noticeable direction reduces to the gradient of f w.r.t. the input image x, $\partial f / \partial x$. Correspondingly, the least-noticeable direction (with an eigenvalue of zero) can be any M-dim unit vector living in the $M-1$-dim subspace orthogonal to $\partial f / \partial x$. Taking consecutive small steps along least-noticeable directions creates a predicted imperceptible perturbation, which is subject to human verification. In this sense, the proposed perceptual attack aims to tackle a "dual" problem, generating a true imperceptible perturbation to maximize the prediction discrepancy.

4 Experiments

In this section, we first set up the experiments, including descriptions of NR-IQA and FR-IQA models, details of the psychophysical experiment, and new evaluation metrics for measuring the perceptual robustness of NR-IQA models. We then present the quantitative and qualitative results accompanied by an in-depth analysis.

4.1 Experimental Setups

Choice of NR-IQA Models. We test four NR-IQA models that are believed to be representative in the field: **BRISQUE** [12], **CORNIA** [21], **Ma19** [22], and **UNIQUE** [6]. BRISQUE is a knowledge-

[2] Assuming the additive white Gaussian noise in the response space, we can compute the Fisher information matrix by $\frac{\partial f}{\partial x}^T \frac{\partial f}{\partial x}$, where $\frac{\partial f}{\partial x}$ is the Jacobian matrix.

Table 1: SRCC and average ratio results of NR-IQA models under intra-model attacks.

Metric	SRCC				R defined in Eq. (6)			
FR-IQA	Chebyshev	SSIM	LPIPS	DISTS	Chebyshev	SSIM	LPIPS	DISTS
BRISQUE	0.0959	0.2023	0.1221	0.1308	0.4116	0.3824	0.4412	0.7392
CORNIA	0.1998	0.0663	0.2092	0.0139	0.9141	0.4181	0.3682	0.2197
Ma19	0.2441	0.1447	0.0959	0.2005	0.6188	0.3914	0.3943	0.3828
UNIQUE	-0.0994	-0.1064	-0.1813	-0.1726	0.2059	0.1539	0.0927	0.1229

Table 2: SRCC and average ratio results of NR-IQA models under inter-model attacks. The number in the parentheses shows the original performance of the NR-IQA model.

	Attack	BRISQUE (0.9231)			CORNIA(0.9650)		
	Under Attack	CORNIA	Ma19	UNIQUE	BRISQUE	Ma19	UNIQUE
SRCC	Chebyshev	0.8230	0.6103	0.7777	0.8596	0.9730	0.9660
	SSIM	0.8684	0.9091	0.9276	0.9538	0.9790	0.9555
	LPIPS	0.7725	0.8806	0.8806	0.9154	0.9451	0.9451
	DISTS	0.7254	0.8806	0.8806	0.8997	0.9660	0.9625
	Attack	Ma19 (0.9091)			UNIQUE(0.9301)		
	Under Attack	BRISQUE	CORNIA	UNIQUE	BRISQUE	CORNIA	Ma19
	Chebyshev	0.7725	0.8387	0.8422	0.9381	0.9224	0.9242
	SSIM	0.8771	0.9102	0.9137	0.9346	0.9137	0.9276
	LPIPS	0.8718	0.8387	0.8893	0.9276	0.9346	0.9346
	DISTS	0.8387	0.8422	0.8858	0.9276	0.8945	0.9259
	Attack	BRISQUE ($+\infty$)			CORNIA ($+\infty$)		
	Under Attack	CORNIA	Ma19	UNIQUE	BRISQUE	Ma19	UNIQUE
R defined in Eq. (6)	Chebyshev	2.1064	2.7651	2.4192	3.2428	4.3930	4.2005
	SSIM	2.4097	3.1472	3.2371	3.3378	4.1283	4.2050
	LPIPS	1.7400	3.1070	3.0490	3.5316	4.2963	4.6841
	DISTS	1.8017	3.1368	2.9193	3.8798	3.8258	4.0432
	Attack	Ma19 ($+\infty$)			UNIQUE ($+\infty$)		
	Under Attack	BRISQUE	CORNIA	UNIQUE	BRISQUE	CORNIA	Ma19
	Chebyshev	2.4957	3.0544	3.1594	4.0762	3.6509	3.9139
	SSIM	3.6393	4.3902	4.4728	4.6850	2.9966	4.2564
	LPIPS	2.9033	3.0309	3.5103	4.4247	3.2835	4.9144
	DISTS	3.6089	3.4737	3.5533	4.8580	2.6109	4.1454

driven model, extracting NSS from mean-subtracted contrast-normalized pixel values. CORNIA [21] relies on unsupervised learning of a visual codebook from image patches, followed by soft-assignment coding and max pooling, to obtain image representations. As feature engineering is not involved, CORNIA can be seen as a shallow learning-based method. Ma19 [22] is a four-layer DNN with generalized divisive normalization [62] as nonlinear activation, which is trained from a large number of corrupted image pairs without reliance on MOSs. UNIQUE [6] trains a variant of ResNet-34 [1] on multiple IQA datasets to handle both synthetic and realistic camera distortions. We use the training codes provided by the original authors to re-train BRISQUE and CORNIA on LIVE [7], Ma19 [22] on our own collected dataset[3], and UNIQUE on six human-rated IQA databases [7, 63, 64, 9, 5, 65]. There is no overlap between all training sets and the initial images used to generate counterexamples.

As suggested by the Video Quality Experts Group [66], a four-parameter logistic function can be adopted to compensate for the prediction nonlinearity, making different NR-IQA more comparable:

$$q \circ f_w(x) = \frac{\beta_1 - \beta_2}{1 + \exp^{-\frac{f_w(x) - \beta_3}{|\beta_4|}}} + \beta_2. \tag{5}$$

β_1 to β_4 are fitting parameters, where β_1 and β_2 determine the maximum and minimum mapping values. Empirically, we find that, for different NR-IQA models, the estimated β_1 and β_2 can be quite different. Thus, we choose to manually enforce $\beta_1 = 10$ and $\beta_2 = 0$. We consider the learned q as part of the NR-IQA model.

[3]We use the same pristine-quality images in [22], but generate the training set with only four types of distortions overlapping with LIVE, *i.e.*, white Gaussian noise, Gaussian blur, JPEG compression, and JPEG2000 compression.

Choice of FR-IQA Models. We consider four FR-IQA models to approximate the perceptual distance between the initial and perturbed images: the **Chebyshev** distance, **SSIM**[4] [10], **LPIPS** [3], and **DISTS** [4]. The Chebyshev distance constrains the maximum pixel difference within an ℓ_∞-ball. SSIM [10] is arguably the most successful "perceptual" metric that compares luminance, contrast, and structure, separately. LPIPS [3] computes the Euclidean distance between deep representations of two images, which shows reasonable effectiveness in explaining image quality. DISTS [4] is the first FR-IQA method that unifies structure and texture similarity, and is competitive in perceptual optimization of various image processing tasks [67]. We apply the Chebyshev distance to each color channel, and take the maximum distance across three channels. We enable SSIM to be aware of color information by treating the color-to-grayscale conversion as a fixed differentiable front-end, which allows the gradient to be back-propagated to the input color image. Both LPIPS and DISTS are based on variants of VGG-16 [68].

Details of the Psychophysical Experiment. We collect twelve images as initializations from the publicly available LIVE IQA database [7] with different perceptual quality and different distortion types (see Fig. 5 in the Appendix). For each of sixteen combinations of NR-IQA and FR-IQA models, and each of the twelve initial images, we set λ to 32 values, and optimize the objective in Eq. (2) to generate 32 perturbed images. For the Chebyshev distance, we use the steepest ascent direction in ℓ_∞-norm, and for the remaining three FR-IQA models, we choose the steepest ascent direction in ℓ_2-norm, namely the gradient direction. We set the step size γ to 10^{-3} and the maximum number of iterations to 200, respectively.

We recruit fifteen human subjects (with normal or corrected-to-normal vision) to participate in the psychophysical experiment, viewing the image pairs from a fixed distance of twice the screen height. A training session is performed to familiarize each subject with the study. For each yes-no trial, subjects are shown (for one second each with a half-second gray screen between images, and in randomized order) a perturbed image and the corresponding initial image, and then asked to determine whether the two images look visually different. All image pairs are displayed on a $24''$ LCD monitor at a resolution of 1920×1080 (in an office environment with normal illumination). To avoid the fatigue effect, subjects are required to take a break during the experiment. In total, we generate $4 \times 4 \times 12 = 192$ counterexamples to test the robustness of NR-IQA models.

Evaluation Metrics. Unlike image classification, it is generally difficult (or even pointless) for a regression task to define the success of an adversarial attack on a per-image basis. As image quality is a *relative* perceptual quantity, we adopt the Spearman rank-order correlation coefficient (SRCC) to measure the prediction monotonicity between model predictions and MOSs of 24 images (12 initial images and 12 corresponding counterexamples). In addition, we define the average ratio of maximum allowable change in quality prediction to actual change over counterexamples in a logarithmic scale:

$$R = \frac{1}{S} \sum_{i=1}^{S} \log \left(\frac{\max \{\beta_1 - f_w(x_i), f_w(x_i) - \beta_2\}}{|f_w(x_i) - f_w(x_i^\star)|} \right), \tag{6}$$

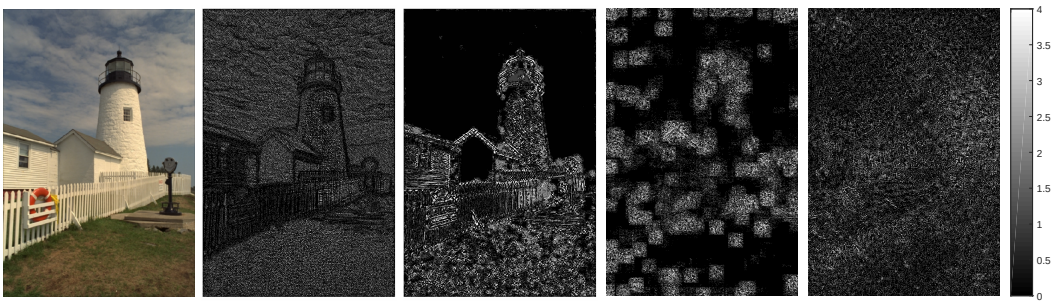
where S is the number of initial images and x_i denotes the i-th one. A larger R means better stability (*i.e.*, robustness), but not necessarily better quality prediction performance as the MOS is not involved in the computation.

4.2 Main Results

We summarize the SRCC and average ratio results of NR-IQA models under intra-model attacks (*i.e.*, using counterexamples generated by the proposed attack with different FR-IQA models) and inter-model attacks (*i.e.*, using counterexamples originally spotted to falsify another NR-IQA model) in Table 1 and Table 2, respectively. The results without perceptual attacks are also presented as references. From the tables, we have a number of insightful observations.

First, the intra-attacks are capable of falsifying the corresponding NR-IQA models, no matter which FR-IQA model is adopted. This is evidenced by a catastrophic SRCC drop relative to the original performance. We then conclude that none of the evaluated NR-IQA design philosophies (*i.e.*, NSS-based, codebook-based, and DNN-based approaches), are inherently perceptually robust. Second, compared with intra-model attacks, NR-IQA methods appear much more robust to inter-model attacks.

[4] In our implementation, we use $1-$ SSIM as the $D(\cdot)$ because a larger SSIM value indicates a higher image fidelity level.



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

Figure 3: Perturbations added to **(a)** the initial image by **(b)** BRISQUE, **(c)** CORNIA, **(d)** Ma19, and **(e)** UNIQUE, respectively, where DISTS is the image fidelity measure.

This manifests the poor transferability of perceptually imperceptible counterexamples, which fail to craft black-box attacks in NR-IQA. On the positive side, the generated examples by the proposed perceptual attack are informative to reveal distinct design flows of different NR-IQA models, and may point out promising ways to improve a model or to combine the best aspects of multiple models. Third, when switching to the average ratio (Eq. (6)), we find that UNIQUE is the least stable model consistently for all FR-IQA methods, suggesting that the overparameterization of UNIQUE creates an abundant space for overfitting the current IQA datasets. Although the observations have been obtained from the expensive psychophysical experiment on only twelve images, we empirically find these to be consistent across a wide range of images with different content and distortion complexities.

To compare the perturbations generated for different NR-IQA models in a more intuitive way, we visualize the absolute residual images (*i.e.*, $(|x_0 - x^\star|)$ in Fig. 3, where DISTS is used as the image fidelity measure. The primary observation is that the difference in perturbations provides useful clues on how they extract quality-aware features. Specifically, perturbations for BRISQUE [12] mainly emerge in smooth and textured regions, where manipulation of individual and product of locally normalized luminances is much easier, as a way of cheating the built NSS model. The learned codebook in CORNIA [21] contains many Dirac delta functions of different locations and edge filters of different orientations, for which the selected texture and edge patches from a test image give the maximum response and thus are used for quality computation. This may explain the perturbations in Fig. 3 (c) are concentrated along the edges and in strong textures. The blocking perturbations in Fig. 3 (d) appear primarily on the objects (*e.g.*, the lighthouse and the fence) and along strong edges (*e.g.*, the image borders due to zero padding in DNNs). We believe this arises from the spatial pyramid pooling [69] layer used in Ma19 [22]. When the pyramid pooling is replaced with global average pooling, the perturbations by Ma19 resemble those in Fig. 3 (e) by UNIQUE. In addition, compare to BRISQUE and CORNIA that only accept grayscale images, the perturbations by Ma19 and UNIQUE occur in all color channels (not shown). Finally, nearly all pixel perturbations are less than $4/255$, justifying the effectiveness of our psychophysical experiment to identify counterexamples that are below JNDs.

5 Conclusion

We have described a perceptual attack on NR-IQA models, based on which their perceptual robustness has been thoroughly evaluated. The proposed attack has close connections to previous techniques. We found that neither the conventional knowledge-driven NR-IQA models nor the modern DNN-based methods are inherently robust to perceptually imperceptible perturbations. Moreover, the generated adversarial examples by one NR-IQA model do not transfer in an efficient way to falsify other models, which contain valuable information to expose the design flows of respective models.

In the future, we plan to develop perceptually robust NR-IQA methods. From the model construction perspective, we will identify and combine robust building blocks of different NR-IQA methods. From the optimization perspective, we may directly employ adversarial training as suggested by [37]. Meanwhile, we may draw inspiration from the connections between the proposed perceptual attack and previous analysis by synthesis techniques, and develop robust regularizers [70] to facilitate training robust NR-IQA methods.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[3] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[4] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear, 2020.

[5] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, Jan. 2020.

[6] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, Mar. 2021.

[7] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, Nov. 2006.

[8] Ponomarenko Nikolay, Jin Lina, Ieremeiev Oleg, Lukin Vladimir, Egiazarian Karen, Astola Jaakko, Vozel Benoit, Chehdi Kacem, Carli Marco, Battisti Federica, and C.-C. Jay Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, Jan. 2015.

[9] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, Jan. 2016.

[10] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, Apr. 2004.

[11] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, Mar. 2013.

[12] Anish Mittal, Anush Krishna Moorthy, and Alan C Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, Dec. 2012.

[13] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 99–108, 2004.

[14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[15] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems*, pages 8312–8323, 2018.

[16] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *CoRR*, abs/1902.06705, 2019.

[17] Chun-Hsien Chou and Yun-Chin Li. A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. *IEEE Transactions on Circuits and Systems for Video Technology*, 5(6):467–476, Dec. 1995.

[18] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017.

[19] Zhou Wang and Eero P Simoncelli. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12):8.1–8.13, Sep. 2008.

[20] Alexander Berardino, Johannes Ballé, Valero Laparra, and Eero Simoncelli. Eigen-distortions of hierarchical representations. In *Advances in Neural Information Processing Systems*, pages 3531–3540, 2017.

[21] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1098–1105, 2012.

[22] Kede Ma, Xuelin Liu, Yuming Fang, and Eero. P. Simoncelli. Blind image quality assessment by learning from multiple annotators. In *IEEE International Conference on Imaging Processing*, pages 2344–2348, 2019.

[23] David Mumford. Pattern theory: A unifying perspective. In *First European Congress of Mathematics*, pages 187–224, 1994.

[24] Ulf Grenander and Michael I Miller. *Pattern theory: From representation to inference*. Oxford University Press, 2007.

[25] Zhou Wang, Hamid R Sheikh, and Alan C Bovik. No-reference perceptual quality assessment of JPEG compressed images. In *IEEE International Conference on Image Processing*, pages 477–480, 2002.

[26] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. Perceptual blur and ringing metrics: Application to JPEG2000. *Signal Processing: Image Communication*, 19(2):163–172, Feb. 2004.

[27] Deepti Ghadiyaram and Alan C Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of Vision*, 17(1):32–32, Jan. 2017.

[28] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014.

[29] Zhihua Wang, Haotao Wang, Tianlong Chen, Zhangyang Wang, and Kede Ma. Troubleshooting blind image quality models in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16256–16265, 2021.

[30] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan C 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 3572–3582, 2020.

[31] Weixia Zhang, Dingquan Li, Chao Ma, Guangtao Zhai, Xiaokang Yang, and Kede Ma. Continual learning for blind image quality assessment. *CoRR*, abs/2102.09717, 2021.

[32] Daniel Lowd and Christopher Meek. Adversarial learning. In *ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 641–647, 2005.

[33] Battista Biggio, Giorgio Fumera, and Fabio Roli. Multiple classifier systems for robust classifier design in adversarial environments. *International Journal of Machine Learning and Cybernetics*, 1:27–41, Oct. 2010.

[34] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 387–402. Springer, 2013.

[35] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[36] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations Workshop*, 2017.

[37] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[38] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

[39] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[40] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations*, 2021.

[41] Bela Julesz. Visual pattern discrimination. *IRE Transactions on Information Theory*, 8(2):84–92, Feb. 1962.

[42] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pages 5389–5400, 2019.

[43] Tal Golan, Prashant C Raju, and Nikolaus Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47):29330–29337, 2020.

[44] Chaowei Xiao, Ruizhi Deng, Bo Li, Fisher Yu, Mingyan Liu, and Dawn Song. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In *European Conference on Computer Vision*, pages 217–234, 2018.

[45] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *IEEE International Conference on Computer Vision*, pages 1369–1378, 2017.

[46] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019.

[47] Jiahang Wang, Sheng Jin, Wentao Liu, Weizhong Liu, Chen Qian, and Ping Luo. When human pose estimation meets robustness: Adversarial algorithms and benchmarks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11855–11864, 2021.

[48] Shuai Jia, Chao Ma, Yibing Song, and Xiaokang Yang. Robust tracking against adversarial attacks. In *European Conference on Computer Vision*, pages 69–84, 2020.

[49] He Wang, Feixiang He, Zhexi Peng, Tianjia Shao, Yong-Liang Yang, Kun Zhou, and David Hogg. Understanding the robustness of skeleton-based action recognition under adversarial attack. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14656–14665, 2021.

[50] XK Yang, WS Ling, ZK Lu, Ee Ping Ong, and SS Yao. Just noticeable distortion model and its applications in video coding. *Signal Processing: Image Communication*, 20(7):662–680, Aug. 2005.

[51] Anmin Liu, Weisi Lin, Manoranjan Paul, Chenwei Deng, and Fan Zhang. Just noticeable difference for images with decomposition model for separating edge and textured regions. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(11):1648–1652, Nov. 2010.

[52] Jinjian Wu, Guangming Shi, Weisi Lin, Anmin Liu, and Fei Qi. Just noticeable difference estimation for images with free-energy principle. *IEEE Transactions on Multimedia*, 15(7):1705–1710, Nov. 2013.

[53] Haiqiang Wang, Ioannis Katsavounidis, Jiantong Zhou, Jeonghoon Park, Shawmin Lei, Xin Zhou, Man-On Pun, Xin Jin, Ronggang Wang, Xu Wang, et al. VideoSet: A large-scale compressed video quality dataset based on JND measurement. *Journal of Visual Communication and Image Representation*, 46:292–302, Jul. 2017.

[54] Huanhua Liu, Yun Zhang, Huan Zhang, Chunling Fan, Sam Kwong, C-C Jay Kuo, and Xiaoping Fan. Deep learning-based picture-wise just noticeable distortion prediction model for image compression. *IEEE Transactions on Image Processing*, 29:641–656, Aug. 2019.

[55] Xuelin Shen, Zhangkai Ni, Wenhan Yang, Xinfeng Zhang, Shiqi Wang, and Sam Kwong. Just noticeable distortion profile inference: A patch-level structural visibility learning approach. *IEEE Transactions on Image Processing*, 30:26–38, Nov. 2020.

[56] Tom N Cornsweet. The staircase-method in psychophysics. *The American Journal of Psychology*, 75(3):485–491, 1962.

[57] Wilson P Tanner Jr and John A Swets. A decision-making theory of visual detection. *Psychological Review*, 61(6):401–409, 1954.

[58] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, Nov. 1992.

[59] Stefan Roth and Michael J Black. Fields of Experts: A framework for learning image priors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 860–867, 2005.

[60] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *International Conference on Computer Vision*, pages 479–486, 2011.

[61] Lei Zhang and Wangmeng Zuo. Image restoration: From sparse and low-rank priors to deep priors. *IEEE Signal Processing Magazine*, 34(5):172–179, Sept. 2017.

[62] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations*, 2017.

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

[64] Alexandre Ciancio, A. L. N. T. Targino da Costa, E. A. B. da Silva, Amir Said, Ramin Samadani, and Pere Obrador. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on Image Processing*, 20(1):64–75, Jan. 2011.

[65] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10k: A large-scale artificially distorted IQA database. In *International Conference on Quality of Multimedia Experience*, pages 1–3, 2019.

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

[67] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, 129(4):1258–1281, Jan. 2021.

[68] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015.

[70] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sec. 5. We focus on evaluation of perceptual robustness on NR-IQA models in this paper, leaving the development of perceptually robust models to future work.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] But we will release all the source code and data upon acceptance.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See both Sec. 4.1 and the Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] As we described in Sec. 4.1, we use the publicly available LIVE IQA database [7].
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] LIVE IQA database is freely available online.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See the descriptions of the details of the psychophysical experiment in Sec. 4.1 and a screenshot of the user interface in the Appendix.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Implementation Details

We evaluate all four FR-IQA models (*i.e.*, Minkowski distance, SSIM, LPIPS, and DISTS) for making predictions of JNDs for natural images. Specifically, we collect four undistorted images from the LIVE dataset [7], on which small perturbations are applied. We present each pair of a perturbed image and the associated source image along with the quality estimated by an FR-IQA model in Fig. 6, from which we find that although Fig. 6 (b), (d), (f), and (h) are assessed as being of high quality by the FR-IQA models, all the added perturbations are above JND. Therefore, we have to conduct well-controlled psychophysical experiments to verify that the perceptually imperceptible counterexamples are below JND. The subjective experiment was carried out in an indoor office environment with a normal lighting condition (with approximately 200 lux) and without reflecting ceiling walls and floors. The peak luminance of the displayed images is mapped to 200 cd/m^2. These are suggested in the BT. 500 recommendations.

To generate candidate perceptually imperceptible counterexamples, we use 32 different lambda values: [0.2, 0.4, 0.6, 0.8, 1, 3, 5, 7, 9, 20, 40, 60, 80, 100, 300, 500, 700, 900, 2,000, 4,000, 6,000, 8,000, 10,000, 30,000, 50,000, 70,000, 90,000, 200,000, 600,000, 1,000,000, 5,000,000, 9,000,000].

All the experiments are conducted on a machine equipped with an Intel(R) Core(TM) i9-9900K CPU @ 3.60GHZ, 48G RAM, and an NVIDIA Geforce GTX 2080Ti 11G GPU.

A.2 One perturbation to fail multiple NR-IQA models

Under the current circumstance that the counterexamples are not transferable (see Table 2), it is worth exploring one perceptually imperceptible perturbation to simultaneously fail multiple NR-IQA models. Conceptually, it is straightforward to extend the current computational method to look for such universal perturbation by replacing the second term in Eq. 2 with the average of discrepancy over N NR-IQA models:

$$x^\star = \arg\max_x - D(x, x_0) + \frac{\lambda}{N}\sum_{i=1}^{N}(f_w^{(i)}(x) - f(x_0))^2. \tag{7}$$

We have made a preliminary experiment to search for such a perturbation on an initial image as shown in Fig. 8 (a). By using the extended attack in Eq. 7 to simultaneously attack BRISQUE, CORNIA, Ma19, and UNIQUE with DISTS as the image fidelity measure, we find an image (Fig. 8 (b)) with perturbation below JNDs with human-in-the-loop. We also visualize the perturbation with an amplified magnitude in Fig. 8 (c), from which we observe a perturbation with mixed patterns, indicating that the resulting perturbation originates from multiple different NR-IQA models. Finally, we compare the quality predictions of Fig. 8 (a) and Fig. 8 (b) by the four NR-IQA models in Table 3, reaching an important observation that the generated single perturbation can simultaneously change the quality predictions of the four NR-IQA models on the counterexample (note that we have limited the range of quality predictions of all four NR-IQA models within $[0, 10]$ by manually enforcing $\beta_1 = 10$ and $\beta_2 = 0$ in Eq. 5).

Table 3: Quality predictions of an initial image and a counterexample by four NR-IQA models.

NR-IQA	BRISQUE	CORNIA	Ma19	UNIQUE
Initial Image	6.9573	6.9410	6.8126	7.7884
Counterexample	9.0697	9.0678	2.3151	9.1327

A.3 Defending against the perceptual attack

One of the key observations is that the counterexamples generated from one NR-IQA model are not transferrable to falsify other models. Intuitively, we may directly leverage model ensemble tricks to mitigate the vulnerability of perceptual quality metrics, but it turns out not a viable solution as indicated by the preliminary results in Sec. A.2. Following the common practice of using the method of Madry et al. [37], we have also conducted some preliminary experiments to adversarially

training the UNIQUE model. Counterexamples during training are generated under an ℓ_∞-constraint with a radius of $\frac{4}{255}$, and the number of attacking iterations and the step size are set to 5 and $\frac{2}{255}$, respectively. After training, we test the adversarially trained model using the proposed perceptual attack, and conduct an informal psychophysical experiment to verify the perceptual imperceptibility of the found counterexamples (by two of the authors). As shown in Table 4, vanilla adversarial training significantly improves the robustness of the UNIQUE model. In the future, we will delve deep into the design and training of perceptually robust NR-IQA models with improved effectiveness and efficiency.

Table 4: SRCC and average ratio results of the UNIQUE model after the standard and adversarial training.

Standard Training		Adversarial Training	
SRCC	R defined in Eq. (6)	SRCC	R defined in Eq. (6)
-0.0994	0.2059	0.6382	1.6395

A.4 Visualization of perturbations

To visualize what the perceptual attacks look like, we include perceptual counterexamples with amplified perturbations in Fig. 7. It deserves to note that the perturbations are imperceptible (verified by the psychophysical experiments) if we do not scale their intensities.

We also show pairs of initial images and their associated counterexamples **without** amplification of the perturbations in Fig. 9. It is clear that the perturbations are invisible, verifying the effectiveness of the psychophysical experiments to make sure that the perturbations are below the JNDs.

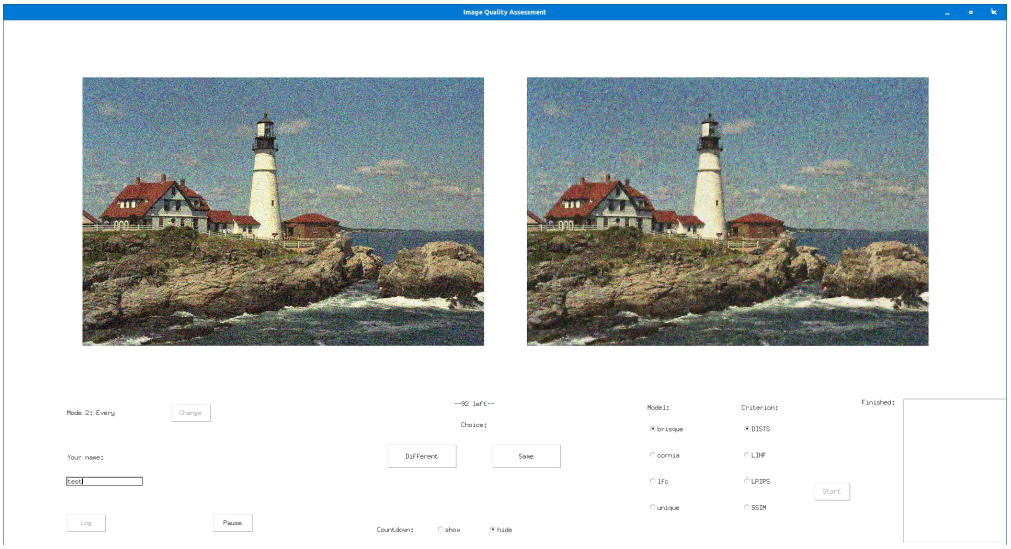


Figure 4: Screenshot of the user interface used in our psychophysical experiment.

Figure 5: Twelve initial images collected from the LIVE IQA database [7]. Images are cropped for improved visibility.

(a) (b) Chebyshev distance = $\frac{3}{255}$

(c) (d) SSIM = 0.9903

(e) (f) DISTS = 0.0057

(g) (h) LPIPS = 0.0075

Figure 6: (a), (c), (e), and (g) are undistorted images collected from LIVE dataset [7]. (b), (d), (f), and (h) are perturbed versions of (a), (c), (e), and (g), respectively. The quality of each perturbed image is estimated by a specific FR-IQA model. A higher SSIM value indicates higher quality, and a lower value of each of the remaining three FR-IQA models indicates higher quality.

Figure 7: Perturbed images derived from **(a), (f), (k)** the initial images by **(b), (g), (l)** BRISQUE, **(c), (h), (m)** CORNIA, **(d), (i), (n)** Ma19, and **(e), (j), (o)** UNIQUE, respectively, where DISTS is the image fidelity measure. Perturbations are amplified by a same factor ($\times 16$), otherwise they are invisible.

Figure 8: (a) An initial image. (b) A counterexample that falsifies four NR-IQA models. (c) Perturbation (amplified) added to (a). DISTS is used as the image fidelity measure.

(a) (b)

(c) (d)

(e) (f)

Figure 9: (a), (c), and (e) are initial images. (b), (d), and (f) are counterexamples of (a), (c), and (e), respectively.