

Exploring Vulnerabilities of No-Reference Image Quality Assessment Models: A Query-Based Black-Box Method

Chenxi Yang^{1,2}, Yujia Liu², Dingquan Li³ and Tingting Jiang²

¹School of Mathematical Sciences, Peking University

²School of Computer Science, Peking University

³Peng Cheng Laboratory

yangchenxi@stu.pku.edu.cn, yujia_liu@pku.edu.cn, dingquanli@pku.edu.cn, ttjiang@pku.edu.cn

Abstract

No-Reference Image Quality Assessment (NR-IQA) aims to predict image quality scores consistent with human perception without relying on pristine reference images, serving as a crucial component in various visual tasks. Ensuring the robustness of NR-IQA methods is vital for reliable comparisons of different image processing techniques and consistent user experiences in recommendations. The attack methods for NR-IQA provide a powerful instrument to test the robustness of NR-IQA. However, current attack methods of NR-IQA heavily rely on the gradient of the NR-IQA model, leading to limitations when the gradient information is unavailable. In this paper, we present a pioneering query-based black box attack against NR-IQA methods. We propose the concept of *score boundary* and leverage an adaptive iterative approach with multiple score boundaries. Meanwhile, the initial attack directions are also designed to leverage the characteristics of the Human Visual System (HVS). Experiments show our attack method outperforms all compared state-of-the-art methods and is far ahead of previous black-box methods. The effective DBCNN model suffers a Spearman rank-order correlation coefficient (SROCC) decline of 0.6972 attacked by our method, revealing the vulnerability of NR-IQA to black-box attacks. The proposed attack method also provides a potent tool for further exploration into NR-IQA robustness.

1 Introduction

Image Quality Assessment (IQA) aims to predict image quality scores consistent with human perception, which can be categorized as Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) according to the access to the undistorted reference images. Among them, NR-IQA has witnessed substantial development recently and has emerged as a suitable method for real-world scenarios [Su *et al.*, 2020; Zhang *et al.*, 2020] because it does not need access to undistorted reference images. NR-IQA models also serve as a crucial component in various visual tasks, such as evaluating im-

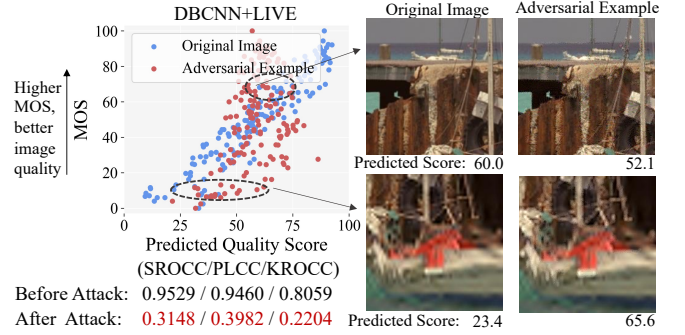


Figure 1: We attack NR-IQA models in a black-box scenario. We successfully attack different NR-IQA models on the individual image level and image set level.

age processing algorithms [Mrak *et al.*, 2003] and optimizing image recommendation systems [Deng and Chen, 2014]. The robustness of NR-IQA methods is vital for providing a stable and dependable basis for comparing different image processing techniques and ensuring consistent user experiences in image recommendation systems.

To scrutinize the robustness of NR-IQA models, recent research has conducted preliminary investigations, shedding light on the susceptibility of IQA models to various attacks [Zhang *et al.*, 2022; Shumitskaya *et al.*, 2022; Korhonen and You, 2022]. These attack methods are designed to generate adversarial examples by causing significant deviations in the prediction quality scores from those of the original samples in two scenarios. In a *white-box scenario* that the entire NR-IQA model under attack is available, generating adversarial examples using gradient-based optimization with the model’s gradients is straightforward [Zhang *et al.*, 2022; Shumitskaya *et al.*, 2022; Sang *et al.*, 2023]. However, this white-box scenario becomes unrealistic when the model parameters are unknown to the attacker. In a more practical *black-box scenario*, attackers possess limited knowledge about the NR-IQA model, often confined to only its output. Korhonen *et al.* utilized a transfer-based method employing a substitute model to generate adversarial examples, which are then transferred to attack unknown target models [Korhonen and You, 2022]. However, the performance of transfer-based black-box methods is limited, highly depending on the choice of substitute model and the constraint condition [Zhang *et al.*,

2022].

However, unlike widely studied black-box attacks for classification problems, attacking NR-IQA presents distinct challenges. Firstly, quantifying the success of attacks on regression-based IQA problems is not straightforward. Different from classification, which naturally defines a classification boundary for determining attack success, regression-based IQA lacks a direct measure of attack “success” due to its continuous output. Secondly, identifying the attack direction becomes particularly challenging when the gradient of an IQA model f is unavailable. Unlike classification tasks, where a small perturbation like Gaussian noise may easily lead to successful attacks, the IQA problem demands a more deliberate design of the attack direction to generate substantial changes in the predicted quality scores. In our preliminary experiment that attacking images for the classification and NR-IQA task with the Gaussian noise, the label misprediction rate achieved 92.6% but the quality score only changes by 2.09 points on average, where the predicted quality image score is in the range of $[0, 100]$. The result shows that the efficiency of this stochastic attack direction dropped dramatically in the context of NR-IQA. This disparity emphasizes the need for a more thoughtful and precise design of attack directions in the context of NR-IQA. Thirdly, NR-IQA tasks are more sensitive to image quality variation compared to classification tasks. So attacking NR-IQA models has a more strict constraint for the perceptual similarity between the adversarial example and its original image, which implies the perturbation is expected to be imperceptible for humans but could cause misjudgments by NR-IQA models. These intricate challenges underscore the significance of developing tailored attack strategies for NR-IQA methods.

We address these three challenges in this paper. Firstly, we introduce the concept of *score boundary* to quantify the success of individual attacks and systematically intensify attacks by setting multiple score boundaries, which enables a more measurable assessment of attack effectiveness. Secondly, we leverage the sensitivity of DNNs to texture information [Ding *et al.*, 2022] and sparse noise [Modas *et al.*, 2019], using texture information and sparse noise extracted from natural images to design the attack direction. We constrain the attack region to the edges and saliency areas of the image to enhance the efficacy of the attack. Thirdly, to ensure perturbation invisibility, we generate adversarial examples with the help of Just Noticeable Difference (JND) [Liu *et al.*, 2010]. JND accounts for the maximum sensory distortion that the Human Visual System (HVS) does not perceive, and it provides a threshold for perturbation for each pixel in an image. When the perturbation of each pixel satisfies the constraint of the JND threshold, the perturbation of the whole image can be considered invisible to human eyes. To optimize the final attack, the Surfree framework [Maho *et al.*, 2021a] is employed, resulting in a more effective and imperceptible attack.

We evaluate the efficacy of our proposed attack methods on four NR-IQA methods: CONTRIQUE [Madhusudana *et al.*, 2022], DBCNN [Zhang *et al.*, 2020], HyperIQA [Su *et al.*, 2020], and SFA [Li *et al.*, 2019] on LIVE [Sheikh *et al.*, 2006] and CLIVE [Ghadiyaram and Bovik, 2016] datasets.

Three correlation metrics and Mean Absolute Error (MAE) are employed to measure the performance of the attack. Perceptual similarity metrics SSIM [Wang *et al.*, 2004] and LPIPS [Zhang *et al.*, 2018] are employed to measure the visibility of perturbations. We compare our approach to three transfer-based attack methods. Our results demonstrate that while maintaining equivalent invisibility of the perturbations, our method achieves superior attack effects. One intuitive case of our attack performance is shown in Figure 1.

Our contributions are as follows:

- We design a novel query-based black box attack against NR-IQA methods featuring adaptive iterative attacks with initial attack direction guidance. To the best of our knowledge, we are the first to design the query-based black-box attack for NR-IQA.
- We propose the concept of *score boundary* of NR-IQA attack and develop adaptive iterative score boundaries to adjust the attack intensity of different images. With prior knowledge of NR-IQA, we design initial attack directions based on the edge and saliency areas of the attacked image. Furthermore, the constraint of JND is conducted, effectively reducing the visibility of the perturbation.
- Extensive experiments show our attack achieves the best black-box performance on different NR-IQA methods, and reveal the vulnerability of NR-IQA under black-box attacks. Our exploration of black-box attacks on NR-IQA provides a convenient tool for further research of NR-IQA robustness.

2 Related Work

2.1 Adversarial Attack for Classification

Adversarial attack is an important problem considering the security and reliability of models. It has been studied extensively in classification, whose goal is to generate adversarial examples misclassified by the model, under the constraint of small perturbations around original images. It can be categorized into white-box attacks and black-box attacks. In white-box scenarios, attackers have access to all details of the target models, including their structures, parameters, and other relevant information [Szegedy *et al.*, 2014; Baluja and Fischer, 2018]. While in black-box scenarios, attackers possess little knowledge about the target model, often limited to just its output [Chen *et al.*, 2017; Papernot *et al.*, 2016]. In practical applications, black-box attacks are more common and challenging [Xie *et al.*, 2019]. Papernot *et al.* [2017] train a local model to substitute for the target model, use the substitute to craft adversarial examples, and then transfer them to target models. On the other hand, without training an additional model, Li *et al.* [2022] only access the decision result of target models, but effectively generate adversarial examples using frequency mixup techniques with limited queries.

2.2 Image Quality Assessment and Its Attack

The quality score of a distorted image can be obtained by the human rating, which is termed the Mean Opinion Score

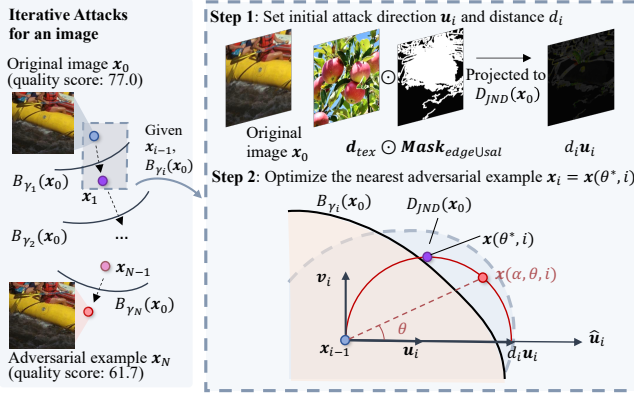


Figure 2: The framework of proposed attack method.

(MOS). An IQA model attempts to predict an image quality score consistent with MOS, which costs less than the human rating. Many researchers have explored image quality-related information inspired by the exploration of the HVS. Among them, the signal-level information like luminance, contrast, and edge in the spatial domain [Wang *et al.*, 2004; Zhang *et al.*, 2011] and natural scene statistics in the frequency domain [Sheikh and Bovik, 2006; Moorthy and Bovik, 2011] are considered. Meanwhile, the image semantic information is also considered with the help of deep neural networks (DNNs). The semantic information extracted from a pre-trained DNN is used for IQA [Li *et al.*, 2019; Su *et al.*, 2020]. On the other hand, JND models the minimum visibility threshold of the HVS, which is considered an important feature related to image quality in some IQA methods [Ferzli and Karam, 2009; Seo *et al.*, 2021]. Additionally, the unsupervised method [Madhusudana *et al.*, 2022], the multi-modality method [Wang *et al.*, 2023], and other advanced methods have been explored facing application scenarios appearing in recent years. For distorted images, their distortion types can be divided into synthetic and authentic distortion, in which the former is artificially created and the latter occurs naturally during the image production process. Authentic distortion has a broader variety and complexity than synthetic distortion.

For attacking IQA, a general goal is to generate the adversarial example within a small perturbation around the original image while the image quality score changes a lot against the original image. In the white-box scenario, Zhang *et al.* [2022] utilize the Lagrangian formula with an FR-IQA as a perceptual constraint and generate adversarial examples with gradient-based optimization; Shumitskaya *et al.* [2022] propose a universal adversarial perturbation to train a single adversarial perturbation for the whole dataset. In the black-box scenario, Korhonen and You [2022] utilize a substitute model to generate adversarial examples and then attack target models. However, the attack performance relies on the choice of the substitute model and training dataset. For example, when Zhang *et al.* [2022] attack CORNIA [Ye *et al.*, 2012], the performance difference with SROCC metric achieves 0.1151 between attacking with the substitute model UNIQUE [Zhang *et al.*, 2021] and BRISQUE [Mittal *et al.*, 2012]. A black-

box method that does not rely on substitute models should be explored.

3 Methodology

In this section, we will introduce our method in a top-down order. We will first illuminate the Global-Local optimization objective of the entire attack process. Then the score boundary for a single-step attack and progressive score boundaries for multistep attacks are defined. Finally, the optimization method with HVS prior for a single-step attack and adaptive score boundaries are introduced in detail. The framework of the attack is shown in Figure 2.

3.1 Global-Local Optimization Objective

To attack IQA models, a primary goal is to make the quality score $f(x_0 + d)$ predicted by attacked IQA model f deviate from the original image x_0 as much as possible. Meanwhile, in the application of IQA, the rank correlation of predicted score with MOS is also important for abundant application, so we also hope to “perceive” rank correlation through adversarial samples. To disturb the rank correlation in an image set, we propose a Global-Local optimization objective for a more reasonable attack. For an image set \mathcal{I} , we split it into higher quality part \mathcal{I}_h and the lower quality part \mathcal{I}_l according to the predicted image score $f(x_0)$. Our objective involves inducing the model to misjudge a high-quality image with a lower quality, and vice versa. The optimization objective is shown in Eq. (1). The higher $f(x)$ corresponds to the higher quality of x .

$$\begin{aligned} \max_d S(x_0) * (f(x_0 + d) - f(x_0)), \\ \text{s.t. } x_0 + d \in \mathcal{D}_{\text{JND}}(x_0), \end{aligned} \quad (1)$$

where

$$S(x_0) = \begin{cases} 1, & x_0 \in \mathcal{I}_l \\ -1, & x_0 \in \mathcal{I}_h \end{cases}. \quad (2)$$

And to restrain the visibility of perturbation, the adversarial sample $x_0 + d$ is restrained to the JND neighborhood $\mathcal{D}_{\text{JND}}(x_0)$ of x_0 . The neighborhood $\mathcal{D}_{\text{JND}}(x_0)$ of an image x_0 could be written as:

$$\begin{aligned} \mathcal{D}_{\text{JND}}(x_0) = \{x \mid |x(l, j, k) - x_0(l, j, k)| < m(l, j, k), \\ 0 \leq l < H, 0 \leq j < W, 0 \leq k < C\}, \end{aligned} \quad (3)$$

where $m(l, j, k)$ is the minimum visibility threshold at pixel $x_0(l, j, k)$ located on (l, j, k) on image x_0 predicted by a JND model. The height, weight and channel of x_0 are H, W and C respectively. The JND model is to estimate the pixel-wise threshold for an image, the perturbed image cannot be visually distinguished from the original image x_0 if the perturbation is under the threshold [Liu *et al.*, 2010].

3.2 Iterative Score Boundaries for Optimization

To qualify the variation of the predicted quality score during attacking, we propose the concept of *score boundary*. For example, for image $x_0 \in \mathcal{I}_l$, and maximum and minimum MOS value MOS_{\max}, MOS_{\min} in the dataset, we set the score boundary $B_{\gamma}^l(x_0)$ as $\{x \mid f(x) = f(x_0) +$

Algorithm 1 Algorithm for Iterative Attacks

Input: Original image x_0 , maximum number of score boundaries N , initial $\gamma_0 = 1/100$, $\gamma_{-1} = 0$

Output: Adversarial point x_N

```

1: for  $i = 1, \dots, N$  do
2:    $\gamma_i \leftarrow \gamma_{i-1} + (\gamma_{i-1} - \gamma_{i-2})$  //To initialize  $\gamma_i$ 
3:    $x_i, \gamma_i \leftarrow \text{SingleAttack}(x_{i-1}, \gamma_i, \dots)$ 
4: end for
5: return  $x_N$ 

```

$\gamma(MOS_{\max} - f(x_0))\}$. This boundary includes samples that have a higher quality score than x_0 . Then the attack is γ -success if $f(x_0 + d) > f(x_0) + \gamma(MOS_{\max} - f(x_0))$. And γ is a scalar to adjust the distance from x_0 to $B_{\gamma}^l(x_0)$, which corresponding to the attack intensity. While for $x_0 \in \mathcal{I}_h$, $B_{\gamma}^h(x_0) = \{x | f(x) = f(x_0) + \gamma(MOS_{\min} - f(x_0))\}$. So we define the adversarial example $x_0 + d$ is γ -success if:

$$f(x_0 + d) \begin{cases} > f(x_0) + \gamma(MOS_{\max} - f(x_0)), x_0 \in \mathcal{I}_l \\ < f(x_0) + \gamma(MOS_{\min} - f(x_0)), x_0 \in \mathcal{I}_h \end{cases} \quad (4)$$

With the criterion in Eq. (4), the success of a single-step attack with intensity γ could be obtained.

Further, to determine the maximum attack intensity of an image, multiple score boundaries are applied. For $x_0 \in \mathcal{I}_l$ and initial γ_0, γ_{-1} , a series of score boundaries $B_{\gamma_1}^l(x_0), \dots, B_{\gamma_N}^l(x_0)$ are set with $\gamma_i = \gamma_{i-1} + (\gamma_{i-1} - \gamma_{i-2})$, $i = 1, \dots, N$. With the multistep attacks, a series adversarial images x_1, \dots, x_N could be generated, which satisfy the property that x_i is γ_i -success ($i = 1, \dots, N$). And x_N is used as the final adversarial example for x_0 . The algorithm for iterative attacks is shown in Algorithm 1. The iterative boundaries guarantee the attack intensity of each iteration is moderate, while multiple boundaries ensure the considerable attack intensity of the whole attack. Further adaptive optimization for iterative score boundaries will be shown in Sec. 3.4.

3.3 Optimization Method for A Single-Step Attack

With the target decomposition in Section 3.2, the attack objective of the i_{th} -step attack of x_0 could be set with:

Solve $x_i \in D_{\text{JND}}(x_0)$, subject to x_i is γ_i -success.

To solve x_i , we leverage a query-based black-box method [Maho *et al.*, 2021a] for classification attack, which reaches low query amounts in attacking classification tasks by utilizing geometrical properties of the classifier decision boundaries. In our attack on NR-IQA, the same analysis could be used. With a start point x_{i-1} , a preset unit attack direction u_i and a distance d_i which satisfies $x_{i-1} + d_i u_i$ is γ_i -success, and a stochastic unit direction v_i orthogonal to u_i , the polar coordinate of a point z near x_{i-1} could be represented as:

$$z(\alpha, \theta, i) = d_i(1 - \alpha)(u_i \cos \theta + v_i \sin \theta) + x_{i-1}, \quad (5)$$

where $\alpha \in [0, 1]$, $\theta \in [-\pi, \pi]$. When α is given, the trajectory of $z(\alpha, \theta, i)$ is an arc, which is shown as the red arc in the lower part of Figure 2 for $\theta \in [0, \pi/2]$. The goal is to choose (α, θ) to raise the probability of $z(\alpha, \theta, i)$ being adversarial. With the theoretical analysis in [Maho *et al.*, 2021a], when

Algorithm 2 *SingleAttack* (Algorithm for A Single-Step Attack)

Input: Start point x_{i-1} , JND neighbourhood $D_{\text{JND}}(x_0)$, score boundary $B_{\gamma_i}(x_0)$, image d_{tex} , mask $\text{Mask}_{\text{edgeUsal}}$ of x_0 , maximum search times for a single-step attack T_{\max}

Output: Adversarial point x_i

```

1: Search times  $T = 0$ 
2: Set initial attack direction
    $\hat{u}_i \leftarrow \tau \cdot d_{\text{tex}} \odot \text{Mask}_{\text{edgeUsal}}, \tau \sim U(-0.1, 0.1)$ 
3: // Project  $\hat{u}_i$  into  $D_{\text{JND}}(x_0)$ 
4:  $d_i \leftarrow \|\text{Proj}_{D_{\text{JND}}(x_0)}(\hat{u}_i)\|$ ,  $u_i \leftarrow \text{Proj}_{D_{\text{JND}}(x_0)}(\hat{u}_i)/d_i$ 
5: if  $x_{i-1} + d_i u_i$  is not  $\gamma_i$ -success then
6:    $T \leftarrow T + 1$ 
7:   if  $T > T_{\max}$  then
8:     // To decrease  $\gamma_i$ 
9:      $\gamma_i \leftarrow \gamma_i - (\gamma_i - \gamma_{i-1})/2$ , go to line 1
10:  else
11:    go to line 2
12:  end if
13: end if
14: if  $x_{i-1} + d_i u_i$  is  $(\gamma_i + 2(\gamma_i - \gamma_{i-1}))$ -success then
15:   // To increase  $\gamma_i$ 
16:    $\gamma_i \leftarrow \gamma_i + (\gamma_i - \gamma_{i-1})$ 
17: end if
18: Set another stochastic attack direction  $\hat{v}_i$ 
19: // Project  $\hat{v}_i$  into  $D_{\text{JND}}(x_0)$ 
20:  $v_i \leftarrow \text{Proj}_{D_{\text{JND}}(x_0)}(\hat{v}_i)$ 
21:  $x(\theta, i) \leftarrow d_i \cos \theta (u_i \cos \theta + v_i \sin \theta) + x_{i-1}$ 
22:  $\theta^* \leftarrow \underset{\theta, x(\theta, i) \text{ is } \gamma_i\text{-success}}{\text{argmin}} \|x(\theta, i) - x_{i-1}\|$ 
23: return  $x_i = x(\theta^*, i), \gamma_i$ 

```

$\alpha = 1 - \cos \theta$, probability of $z(\alpha, \theta, i)$ being adversarial reaches maximum. So we mark $z(1 - \cos \theta, \theta, i)$ as the candidate point $x(\theta, i)$:

$$x(\theta, i) = d_i \cos \theta (u_i \cos \theta + v_i \sin \theta) + x_{i-1}. \quad (6)$$

The adversarial example $x_i := x(\theta, i)$ can be solved by:

$$\begin{aligned} & \min_{x(\theta, i)} \|x(\theta, i) - x_{i-1}\| \\ & \text{s.t. } x(\theta, i) \in D_{\text{JND}}(x_0), x(\theta, i) \text{ is } \gamma_i\text{-success.} \end{aligned} \quad (7)$$

There are two challenges in attacking NR-IQA: 1) The reasonable preset direction u_i should be deliberately designed to guarantee an efficient attack. 2) The generated adversarial example should satisfy the constraint of D_{JND} , which is difficult in solving Eq. (7).

3.3.1 The Design of Attack Direction

In attacking classification tasks, a common approach is to employ stochastic directions, such as Gaussian noise, as the attack direction u . However, this strategy, while effective for classification tasks, often proves inadequate when targeting NR-IQA models. For instance, when applying the attack perturbation δ with preset values $\delta \sim 0.15 * \mathcal{N}(0, 1)$ to the classification task starting point x_0 (normalized to the range $(0, 1)$), we observe a high success rate of 92.6% for inducing misclassification of $x_0 + \delta$ in 500 random trials.

However, when utilizing the same attack direction to target the NR-IQA model HyperIQA, the resulting average change in predicted image quality score between x_0 and $x_0 + \delta$ is merely 2.09 points (within the range of predicted image scores $[0, 100]$). Evidently, the efficiency of this stochastic attack direction dropped dramatically in the context of NR-IQA. In our pursuit of a more effective attack direction, we introduce a method that disrupts image regions that are sensitive to NR-IQA models while ensuring the perturbation remains imperceptible to the human eye.

When designing the attack direction, Our idea is to add disruption to the sensitive regions of the NR-IQA models, while the disruption is not visible to the human eye. Considering the sensitivity of DNNs to image texture [Ding *et al.*, 2022] and sparse noise [Modas *et al.*, 2019; Dong *et al.*, 2020], we leverage the texture information and sparse noise extracted from the high-quality natural images as the foundation for attack direction d_{tex} . The natural images are randomly selected with no content overlap in training/test datasets of attacked NR-IQA. Furthermore, considering the edge region and salient region are often critical to the judgment of IQA models [Zhang *et al.*, 2011; Seo *et al.*, 2021], we introduce a mask $\text{Mask}_{\text{edgeUsal}}$ to confine attacks to these specific regions. By doing so, we amplify the attack’s intensity while maintaining its focus on areas central to NR-IQA assessment.

The designed attack direction could be formulated as $\hat{u}_i = \tau \cdot d_{\text{tex}} \odot \text{Mask}_{\text{edgeUsal}}$, where \odot is the Hadamard product, τ is a stochastic variable drawn from a uniform distribution $U(-0.1, 0.1)$. The introduction of τ imbues different intensities that could be attempted in searching for \hat{u} , enhancing the versatility and adaptability of the proposed method.

3.3.2 The Solution with JND Constraint

The constraint of $D_{\text{JND}}(x_0)$ makes the optimization of Eq. (7) difficult. To solve this problem, we constrain the initial direction u_i and v_i . Eq. (3) indicates D_{JND} is a convex set, so there is a proposition for the JND constraint (the proof is provided in the supplementary material):

Proposition 1 If $x_{i-1} + d_i u_i, x_{i-1} \pm d_i v_i \in D_{\text{JND}}(x_0)$, then $x(\theta, i) \in D_{\text{JND}}(x_0)$ for $\theta \in [0, \pi]$.

So we attempt to constrain u_i, v_i so that they satisfy the conditions of Proposition 1. One feasible method is to project them into a neighborhood in D_{JND} . We obtain u_i, v_i from projection operation:

$$\begin{aligned} \text{Proj}_{D_{\text{JND}}(x_0)}(\hat{u}_i) &:= \underset{u, x_i + d_i u \in D_{\text{JND}}(x_0)}{\text{argmin}} \|u - \hat{u}_i\|, \\ \text{Proj}_{D_{\text{JND}}(x_0)}(\hat{v}_i) &:= \underset{v, x_i \pm d_i v \in D_{\text{JND}}(x_0)}{\text{argmin}} \|v - \hat{v}_i\|, \end{aligned} \quad (8)$$

The obtained u_i, v_i satisfy the conditions of Proposition 1 so that solving a single-step attack progress could be done with the constraint of $D_{\text{JND}}(x_0)$. So the algorithm for a single-step attack is shown in Algorithm 2. Lines 4 and 19 in Algorithm 2 ensure the properties in Proposition 1 are satisfied. Then Eq. (7) is solved with binary search of θ , which can be seen in [Maho *et al.*, 2021b].

3.4 Adaptive Optimization for Score Boundaries

To fine-tune attack intensity for different images, we leverage an adaptive optimization for iterative score boundaries to

set adjustable $\{\gamma_i\}_{i=0}^N$, which means the score boundaries are adaptive for each image and each iteration. When the boundary is too difficult to cross, a closer boundary with a smaller γ is set. When the boundary is too easy to cross, a more distant boundary with a larger γ is set. The benefit of adaptive boundaries is to guarantee a stronger attack, by adjusting the score boundary dynamically.

For two neighboring score boundaries γ_{i-1}, γ_i of an image, there are *Increase* and *Decrease strategies*: a) *Decrease strategy*: when maximum search times for initial attack direction u_i is achieved in a single-step attack, we decrease γ_i to $\gamma_i - (\gamma_i - \gamma_{i-1})/2$. b) *Increase strategy*: when initial attack direction u_i and distance d_i satisfy that $x_{i-1} + d_i u_i$ is $(\gamma_i + 2(\gamma_i - \gamma_{i-1}))$ -success, increase γ_i to $(\gamma_i + (\gamma_i - \gamma_{i-1}))$. These strategies are outlined in lines 9 and 15 of Algorithm 2. When the γ_i is decreased and the difference $\gamma_i - \gamma_{i-1} < 1/400$, the attack will be early stopped. This indicates that the attack intensity is nearing saturation in recent iterations.

4 Experiments

In this section, we first present the setting of attacking, including attacked NR-IQA methods, and the experimental results compared with other methods. Then the effect of different parts of our attack is explored. Finally, the visualization of adversarial examples is presented.

4.1 Experimental Setups

NR-IQA Models and Datasets We choose four NR-IQA models DBCNN [Zhang *et al.*, 2020], HyperIQA [Su *et al.*, 2020], SFA [Li *et al.*, 2019] and CONTRIQUE [Madhusudana *et al.*, 2022]. These methods are based on the various quality features extracted by DNN and are all widely recognized in the NR-IQA field. The LIVE dataset [Sheikh *et al.*, 2006] with synthetic distortions and CLIVE dataset [Ghadiyaram and Bovik, 2016] with authentic distortions are chosen to train and attack NR-IQA models respectively. 80% data of the dataset are split for training and the rest for the attack, and no image content is overlap between the training and the test set. NR-IQA models are retrained on LIVE and CLIVE with their public code. For the attack, we use a random cropping with 224×224 for each image. And cropped images are fixed for all experiments.

Setting of Attacking Experiments We set the number of score boundaries $N = 20$, with $\gamma_0 = 0.01$ for $B_{\gamma_0}(x_0)$. Maximum search times T_{max} is set to 200. And $MOS_{\text{max}} = 100$ and $MOS_{\text{min}} = 0$ are set. The saliency maps of images are predicted with MBS [Zhang *et al.*, 2015]. And edges of images are extracted by Canny operation [Canny, 1986]. For mask $\text{Mask}_{\text{edgeUsal}}(x_0)$, the pixel with a positive value in the saliency map or edge map of x_0 is set to 1 and other pixels are set to 0. The JND model of Liu *et al.* [2010] is used. For the convenience of optimization, the norm in the optimization target of Eq. (8) is set to L_1 norm. For d_{tex} , we randomly choose four high-quality images from the KADID-10k dataset [Lin *et al.*, 2019]. The high-quality images are first been Gaussian blurred, and the difference between the original high-quality image and the blurred image is used as the high-frequency image (shown in the supplementary material). And for each

Attacked NR-IQA	Attack Method	LIVE						CLIVE					
		Attack Performance				Invisibility		Attack Performance				Invisibility	
		SROCC↓	PLCC↓	KROCC↓	MAE↑	SSIM↑	LPIPS↓	SROCC↓	PLCC↓	KROCC↓	MAE↑	SSIM↑	LPIPS↓
DBCNN	Original	0.9529	0.9460	0.8058	9.79	-	-	0.8133	0.8467	0.6292	8.39	-	-
	Korhonen	0.8766	0.8671	0.6928	19.43	0.867	0.186	0.6799	0.6856	0.4986	14.73	0.865	0.113
	UAP	0.8311	0.8145	0.6409	17.38	0.792	0.141	0.7026	0.7083	0.5196	13.10	0.650	0.159
	Zhang	-	-	-	-	-	-	-	-	-	-	-	-
	Ours	0.3148	0.3982	0.2204	21.87	0.896	0.127	0.1161	0.2261	0.084	18.08	0.881	0.114
HyperIQA	Original	0.9756	0.9746	0.8714	4.09	-	-	0.8543	0.8816	0.6762	8.13	-	-
	Korhonen	0.9161	0.9007	0.7417	12.33	0.867	0.186	0.6652	0.663	0.4861	14.38	0.865	0.113
	UAP	0.8685	0.8525	0.6959	10.73	0.792	0.141	0.7847	0.775	0.5909	10.89	0.650	0.159
	Zhang	0.9664	0.9557	0.8421	5.99	0.853	0.084	0.8208	0.8428	0.6347	9.26	0.791	0.113
	Ours	0.7750	0.8208	0.5887	14.95	0.893	0.133	0.4302	0.5121	0.3025	14.45	0.875	0.118
SFA	Original	0.8425	0.8379	0.6546	11.99	-	-	0.805	0.8322	0.6205	9.226	-	-
	Korhonen	0.7479	0.7606	0.5585	21.22	0.867	0.186	0.5967	0.6029	0.4249	22.33	0.865	0.113
	UAP	0.7458	0.7435	0.5469	13.58	0.792	0.141	0.5999	0.6307	0.4257	14.92	0.650	0.159
	Zhang	0.8519	0.8404	0.6612	11.33	0.853	0.084	0.7534	0.7912	0.5677	10.79	0.791	0.113
	Ours	0.5308	0.5894	0.3851	17.04	0.890	0.093	0.3530	0.4169	0.2482	16.17	0.879	0.115
CONTRIQUE	Original	0.8682	0.8386	0.6797	14.73	-	-	0.7143	0.7177	0.5216	18.23	-	-
	Korhonen	0.8042	0.8092	0.6066	16.16	0.867	0.186	0.6748	0.7003	0.4898	15.92	0.865	0.113
	UAP	0.7221	0.7171	0.5203	17.64	0.792	0.141	0.7204	0.7264	0.5322	18.12	0.650	0.159
	Zhang	0.8213	0.8048	0.6259	16.70	0.853	0.084	0.5614	0.5695	0.3983	16.71	0.791	0.113
	Ours	0.4867	0.5455	0.3409	19.72	0.934	0.061	0.0651	0.1318	0.0552	20.85	0.892	0.092

Table 1: Black-Box attack performances on four NR-IQA models. The best attack performances are marked with **bold**. Korhonen, UAP, and Zhang are compared as transfer-based black-box methods. The substitute model used in Korhonen is a variant of ResNet-50. The substitute model in UAP and Zhang are PaQ-2-PiQ and DBCNN, respectively.

single-step attack, one of four high-frequency images is randomly selected as the initial attack direction \mathbf{d}_{tex} . For the whole attack of an image, the overall maximum query times are limited to 8000.

Evaluation of Attack Performance To evaluate the attack performance, we consider the effects of attacks on both individual images and a set of images. For a single image, the absolute error between the predicted score of the adversarial example and MOS is calculated, and it is presented for the whole test set as MAE. For a set of images, we consider the correlation between the predicted quality scores and MOS in the test set. We adopt three correlation indexes: SROCC, Pearson linear correlation coefficient (PLCC), and Kendall rank-order correlation coefficient (KROCC). The R robustness [Zhang *et al.*, 2022] is also presented in the supplementary material. For the invisibility performance, we use SSIM [Wang *et al.*, 2004] and LPIPS [Zhang *et al.*, 2018] to calculate the perceptual similarity between original images and adversarial examples.

Compared Attack Methods To compare with the existing method, we choose the only black-box attack method for NR-IQA that existed from Korhonen and You [2022], marked as Korhonen. For a comprehensive comparison, two white-box attack methods trained with substitute models are compared as transfer-based black-box methods, marked as UAP [Shumitskaya *et al.*, 2022] and Zhang [Zhang *et al.*, 2022]. Details for the three methods are in the supplementary material.

4.2 Attacking Results

We present the prediction performance of NR-IQA models before (marked as Original) and after the attack in Table 1. Our black-box attack method has superior attack effectiveness under the premise of maintaining good invisibil-

ity. It consistently leads to substantial performance degradation across not only correlation metrics but also the MAE. Specifically, the attack on CONTRIQUE within the CLIVE dataset results in an SROCC reduction to 0.0651, indicating a substantial disruption in the order relationship within the image set. Meanwhile, Zhang presents unstable attack performances with failure in attacking SFA. For instance, the SROCC for SFA unexpectedly increases from 0.8425 before the attack to 0.8519 after the attack. Korhonen method performs a better MAE value than our attack when attacking SFA because the substitute model in Korhonen is similar to the model used in SFA. But our method still achieves better SROCC/PLCC/KROCC performance beneficial to the Global-Local optimization objective we set.

For the robustness of NR-IQA, all four attacked NR-IQA present the vulnerability to black-box attacks on both synthetic distortions in LIVE and authentic distortions in CLIVE, which alarms the necessity to explore the security of NR-IQA. Among the four NR-IQA, DBCNN and CONTRIQUE suffer the most performance degradation. Meanwhile, NR-IQA methods are less robust against attacks involving authentic distortions compared to synthetic ones. This could be attributed to the more complex and variable patterns inherent in authentic distortions, presenting a relatively easier target for attacks. It is worthy to be further explored in future work.

4.3 Ablation Study

To examine the effectiveness of different parts of our attack method, we conduct a detailed performance analysis with the DBCNN model for different settings in Table 2.

Effect of Adaptive Iterative Score Boundaries For **iterative boundaries**, we generate adversarial examples with different numbers of adaptive score boundaries. In the first part

	Setting	Attack Performance		Invisibility	
		SROCC↓	MAE↑	SSIM↑	LPIPS↓
# Score Boundaries	N=5	0.7883	16.57	0.935	0.084
	N=10	0.4354	20.55	0.903	0.123
	N=20	0.3148	21.87	0.896	0.127
	N=40	0.3147	21.69	0.895	0.129
Adaptive Boundaries	Fixed	0.8461	15.62	0.962	0.061
	Adaptive	0.3148	21.87	0.896	0.127
Init. Attack Direction	SP Noise	0.5489	18.89	0.941	0.064
	Nat. Image	0.8709	15.21	0.944	0.064
	High Freq.	0.3148	21.87	0.896	0.127
Operation on	Edge	0.4123	20.49	0.931	0.104
Init. Attack	Sal.	0.8531	14.97	0.962	0.072
Direction	Edge+Sal.	0.3148	21.87	0.896	0.127
Constraint of JND	w/o JND	-0.4502	31.58	0.722	0.226
	with JND	0.3148	21.87	0.896	0.127

Table 2: Black-Box attack performance with different settings.

of Table 2, the number of score boundaries N directly affects the intensity of the attack. Totally, a greater N responds to a stronger attack intensity. For **adaptive boundaries**, we compare boundaries with fixed space with $\gamma_i = 0.01i, i = 1, \dots, N$ in the second part of Table 2. The adaptive boundaries guarantee a stronger attack intensity than the fixed boundaries. We also show the change of γ_i in an iterative attack for adaptive boundaries in the supplementary material.

Effect of Design for Initial Attack Direction **1)** Different initial directions: We employed four natural images, and corresponding high-frequency images as attack directions respectively (marked as Nat. Image and High Freq. respectively). The attack direction with salt-and-pepper noise in each image channel (marked as SP noise) is also compared. The third part of Table 2 presents the influence of different initial attack directions. Utilizing direct natural images as the initial attack direction proved to be effective to some degree. The sparse noise in salt-and-pepper noise and high-frequency images makes the attack stronger. Meanwhile, high-frequency images achieve the best attack performance among them. We also show the role of different components in the high-frequency image, and both image texture and sparse noise in it is effective in attacking NR-IQA. **2)** Different image contents of initial attack directions: To verify the effect of different contents of high-frequency images in initial attack directions, we randomly choose ten high-quality images from the KADID-10k dataset, and their high-frequency versions are regarded as the initial attack direction respectively. The mean±standard deviation of the MAE is 16.69 ± 0.31 (more results for each image can be seen in the supplementary material). The effectiveness of the initial attack direction is less related to its image content. **3)** Different operations on the initial attack direction: The fourthly part of Table 2 shows the effect of edge mask and saliency mask operation on the initial attack direction (marked as Edge and Sal. respectively). Both operations lead to an effective attack. Edge mask performs a more important role with an MAE of 20.49. Using both operations makes the best performance with an MAE of 21.87.

Effect of JND Constraint In the last two rows of Table 2, when there is no JND constraint, the invisibility of the attack



Figure 3: (Zoom in for better view) Adversarial examples from different attack methods.

has a dramatic decline, which testifies to the necessity of JND constraint for invisibility.

4.4 Visualization of Adversarial Examples

For an intuitive exhibition, we show the visualization of adversarial examples in Figure 3. The predicted score are normalized to $[0, 100]$, and a higher score means a higher image quality for NR-IQA models. It is noticeable our adversarial examples show good invisibility. The perturbations generated by our method are more concentrated in the high-frequency region, for instance, the balustrade in (a), and the rocks in (c). Other methods tend to have more perturbations in the low-frequency region, for instance, the body of the sky in (a) and (b), which are easier to capture by the human eye. It implies the necessity of our constraint of attack directions.

5 Conclusion

In this paper, we propose the query-based black-box attack for NR-IQA for the first time. We propose the definition of score boundary and leverage an adaptive iterative approach with multiple score boundaries. Meanwhile, the design of attack directions ensures the effectiveness and invisibility of the attack. With the attack, the robustness of four NR-IQA models is examined. It reveals the vulnerability of NR-IQA models to black-box attacks and gives a clue for the exploration of the robustness of NR-IQA models.

References

- [Baluja and Fischer, 2018] Shumeet Baluja and Ian Fischer. Learning to attack: Adversarial transformation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2687–2695, New Orleans, Louisiana, USA, 2018. AAAI Press.
- [Canny, 1986] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, PAMI-8(6):679–698, 1986.
- [Chen et al., 2017] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the ACM Workshop on Artificial Intelligence and Security*, pages 15–26, Dallas, TX, USA, 2017. ACM.
- [Deng and Chen, 2014] Yu Deng and Ke Chen. Image quality analysis for searches, November 25 2014. US Patent 8,997,604.
- [Ding et al., 2022] 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*, 44(5):2567–2581, 2022.
- [Dong et al., 2020] Xiaoyi Dong, Dongdong Chen, Jianmin Bao, Chuan Qin, Lu Yuan, Weiming Zhang, Nenghai Yu, and Dong Chen. Greedyfool: Distortion-aware sparse adversarial attack. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11226–11236. Curran Associates, Inc., 2020.
- [Ferzli and Karam, 2009] Rony Ferzli and Lina J. Karam. A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb). *IEEE Transactions on Image Processing*, 18(4):717–728, 2009.
- [Ghadiyaram and Bovik, 2016] 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, 2016.
- [Korhonen and You, 2022] Jari Korhonen and Junyong You. Adversarial attacks against blind image quality assessment models. In *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, pages 3–11, 2022.
- [Li et al., 2019] Dingquan Li, Tingting Jiang, Weisi Lin, and Ming Jiang. Which has better visual quality: The clear blue sky or a blurry animal? *IEEE Transactions on Multimedia*, 21(5):1221–1234, 2019.
- [Li et al., 2022] Xiu-Chuan Li, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Decision-based adversarial attack with frequency mixup. *IEEE Transactions on Information Forensics and Security*, 17:1038–1052, 2022.
- [Lin et al., 2019] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience*, pages 1–3, 2019.
- [Liu et al., 2010] 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, 2010.
- [Madhusudana et al., 2022] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022.
- [Maho et al., 2021a] Thibault Maho, Teddy Furon, and Erwan Le Merrer. Surfree: A fast surrogate-free black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10430–10439, June 2021.
- [Maho et al., 2021b] Thibault Maho, Teddy Furon, and Erwan Le Merrer. Surfree: A fast surrogate-free black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10430–10439, June 2021.
- [Mittal et al., 2012] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.
- [Modas et al., 2019] Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Sparsefool: A few pixels make a big difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [Moorthy and Bovik, 2011] Anush K. Moorthy and Alan C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, 2011.
- [Mrak et al., 2003] M. Mrak, S. Grgic, and M. Grgic. Picture quality measures in image compression systems. In *The IEEE Region 8 EUROCON 2003. Computer as a Tool.*, volume 1, pages 233–236, 2003.
- [Papernot et al., 2016] Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1312.6199, 2016.
- [Papernot et al., 2017] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the ACM on Asia Conference on Computer and Communications Security*, page 506–519, Abu Dhabi, United Arab Emirates, 2017. ACM.
- [Sang et al., 2023] Qingbing Sang, Hongguo Zhang, Lixiong Liu, Xiaojun Wu, and Alan C Bovik. On the generation of adversarial examples for image quality assessment. *The Visual Computer*, pages 1–16, 2023.
- [Seo et al., 2021] Soomin Seo, Sehwan Ki, and Munchurl Kim. A novel just-noticeable-difference-based saliency-channel attention residual network for full-reference image quality predictions. *TCSVT*, 31(7):2602–2616, 2021.

- [Sheikh and Bovik, 2006] Hamid Rahim Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.
- [Sheikh *et al.*, 2006] Hamid Rahim Sheikh, Muhammad Farooq Sabir, and Alan Conrad Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006.
- [Shumitskaya *et al.*, 2022] Ekaterina Shumitskaya, Anastasia Antsiferova, and Dmitriy S Vatolin. Universal perturbation attack on differentiable no-reference image- and video-quality metrics. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.
- [Su *et al.*, 2020] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqui Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2020.
- [Szegedy *et al.*, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*, Banff, AB, Canada, 2014.
- [Wang *et al.*, 2004] 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, 2004.
- [Wang *et al.*, 2023] Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2555–2563, Jun. 2023.
- [Xie *et al.*, 2019] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, Long Beach, CA, USA, 2019. Computer Vision Foundation / IEEE.
- [Ye *et al.*, 2012] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1098–1105, 2012.
- [Zhang *et al.*, 2011] L. Zhang, L. Zhang, X. Mou, and D. Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.
- [Zhang *et al.*, 2015] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Minimum barrier salient object detection at 80 fps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1404–1412, December 2015.
- [Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.
- [Zhang *et al.*, 2020] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou 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.
- [Zhang *et al.*, 2021] 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, 2021.
- [Zhang *et al.*, 2022] Weixia Zhang, Dingquan Li, Xiongkuo Min, Guangtao Zhai, Guodong Guo, Xiaokang Yang, and Kede Ma. Perceptual attacks of no-reference image quality models with human-in-the-loop. In *Advances in Neural Information Processing Systems*, volume 35, pages 2916–2929. Curran Associates, Inc., 2022.