arXiv:2401.05217v3 [cs.CV] 26 Apr 2024

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

Chenxi Yang, Yujia Liu, Dingquan Li, Tingting Jiang

***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 method outperforms all compared state-of-the-art attack methods and is far ahead of previous black-box methods. The effective NR-IQA model DBCNN suffers a Spearman's rank-order correlation coefficient (SROCC) decline of 0.6381 attacked by our method, revealing the vulnerability of NR-IQA models to black-box attacks. The proposed attack method also provides a potent tool for further exploration into NR-IQA robustness.**

***Index Terms*—No-reference image quality assessment, black-box attack, query-based attack, robustness.**

I. 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 pristine reference images. Among them, NR-IQA has witnessed substantial development recently and has emerged as a suitable method for real-world scenarios [1], [2] because it does not rely on reference images. NR-IQA models also serve as a crucial component in various visual tasks, such as evaluating image processing algorithms [3] and optimizing image recommendation systems [4]. 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.

To scrutinize the robustness of NR-IQA models, recent research has conducted preliminary investigations, shedding light on the vulnerability of IQA models to various attacks [5]–[7]. These attack methods are designed to generate adversarial examples by causing significant deviations in the predicted 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 [5], [6], [8]. However, this white-box scenario becomes unrealistic when the model parameters are unknown to the attacker. In a *black-box scenario*, attackers possess limited knowledge of 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 [7]. However, the performance of transfer-based black-box methods is limited, highly depending on the choice of substitute models and constraints [5].

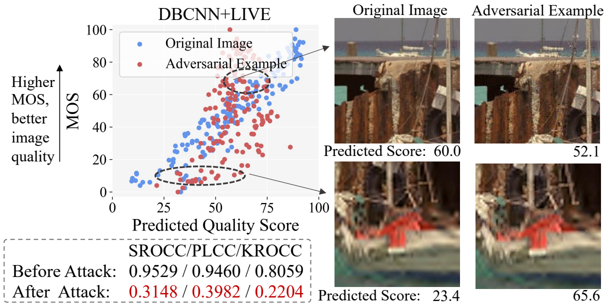

Fig. 1. An NR-IQA model (DBCNN) is attacked in a black-box scenario. The top left shows the predicted quality scores by DBCNN on the LIVE dataset for original images and adversarial examples generated by our attack method. The bottom left shows the SROCC/PLCC/KROCC before and after the attack. The right figures show two sample images before and after the attack as well as the predicted scores.

This situation raises a challenging and intricate question: How can we attack IQA models in the black-box scenario, and without using substitute models? A potential resolution is to leverage query-based black-box attack methods, which are extensively explored in classification tasks [9], [10]. These approaches aim to design attack direction with stochasticity and prior information to generate adversarial examples that cross the classification boundary (*i.e.* the prediction of the attacked model changed).

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 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 where we attacked images for the classification and NR-IQA task with the Gaussian noise, the misclassification rate achieved 92.6% but the quality score only changed by 2.09 on average, with the predicted image quality score within 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 delicate design of attack directions in the context of NR-IQA. Thirdly, NR-IQA tasks are more sensitive to image quality variation than 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 that 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 black-box attack strategies for NR-IQA methods.

We address these three challenges in this paper. Firstly, the concept of *score boundary* is introduced 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, leveraging the sensitivity of deep neural networks (DNNs) to texture information [11] and sparse noise [12], we extract the texture and sparse noise from natural images and use them to design the attack direction. We constrain the attack region to the edges and salient areas of an 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) [13]. 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, we employ the SurFree framework [14]. This framework capitalizes the geometric properties of score boundaries and provides an effective query-based attack.

The efficacy of our attack method is evaluated on four NR-IQA methods across two datasets. The assessment employs three correlation metrics and Mean Absolute Error (MAE) to quantify the performance of the attack. Additionally, two perceptual similarity metrics SSIM [15] and LPIPS [16] are employed to measure the visibility of perturbations. We compare our approach to three transfer-based attack methods. The results demonstrate that while maintaining comparable invisibility of the perturbations, our method achieves superior attack effects. One intuitive case of our attack performance is shown in Fig. 1.

Our contributions are as follows:

- A novel query-based black box attack method against NR-IQA methods is proposed, featuring adaptive iterative attacks with initial attack direction guidance. To the best of our knowledge, this is the first work to design the query-based black-box attack for NR-IQA.
- We propose the concept of *score boundary* for NR-IQA attacks 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 salient areas of the attacked image. Besides, the constraint of JND is introduced, effectively reducing the visibility of the perturbation.
- Extensive experiments show our attack achieves the best black-box performance on different NR-IQA methods, which reveals 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.

II. Related Work

A. Adversarial Attack in Classification Tasks

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 [17], [18]. Most white-box attacks generate adversarial examples by solving a constrained optimization problem, where the constraint ensures the similarity between the original images and the generated examples. Commonly used conditions for this constraint include ℓ_∞ norm [19], ℓ_2 norm [20] and others [12], [21].

While in black-box scenarios, attackers possess little knowledge about the target model, often limited to just its output [22], [23]. In practical applications, black-box attacks are more common and challenging [24]. There are two primary approaches for designing black-box attacks: transfer-based methods and query-based methods. Transfer-based methods first leverage known substitution models to generate adversarial examples, which are then transferred to attack unknown target models. Papernot *et al.* [25] train a 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, query-based methods directly approximate the gradient by querying the target model and obtaining its output, allowing them to design adversarial perturbations based on the gradient. These methods do not require training a substitute model, focusing instead on direct interactions with the target model. For instance, Guo *et al.* [9] propose a strategy where adversarial examples are generated by iteratively adding or subtracting vectors from a predefined orthonormal basis, although this method requires a significant number of queries to ensure attack success. To address the inefficiency of high query demands, Thibault *et al.* [14] introduce a method that capitalizes on the geometric properties of classifier decision boundaries to reduce the number of required queries. Meanwhile, using frequency mixup techniques, Li *et al*. [10] effectively generate adversarial examples with limited queries.

B. Image Quality Assessment

IQA plays an important role in evaluating the perceptual quality of images, aiming to align closely with human visual judgment, commonly quantified as the Mean Opinion Score (MOS). IQA models strive to predict image quality in a cost-efficient manner compared to extensive human rating processes. Among them, FR-IQA measures the perceptual difference between the distorted image and its undistorted version, while NR-IQA predicts the image quality of a distorted image with no reference image. Some IQA methods consider the signal-level information like luminance and edge in the spatial domain [15], [26] and natural scene statistics in the frequency domain [27], [28]. FSIM index [26] is an exemplary method that quantifies image similarity, taking into account factors like chrominance and phase congruency between a pristine reference and a distorted image. Furthermore, the image semantic information is also considered in the IQA task [1], [29]–[31]. The SFA method [29] designs statistics derived from features extracted via neural networks trained on classification tasks. HyperIQA [1] utilizes similar features to predict parameters of a quality prediction network. DDNet [31] used a dynamic filtering module to extract content-adaptive features. An additional dimension in IQA research considers the JND. It models the minimum visibility threshold of the HVS, as a critical component in several IQA methods [32], [33]. Recent years have also seen the exploration of unsupervised methods [34], [35], multi-modality method [36], and other innovative techniques, addressing the application scenarios appearing in recent years.

Distortions in images are typically categorized into synthetic and authentic distortion. 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.

C. Adversarial Challenges in Quality Assessment Tasks

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*. [5] employ a gradient-based optimization strategy, incorporating the Lagrangian method with a Full-Reference IQA (FR-IQA) as a perceptual constraint, to generate adversarial examples; Shumitskaya *et al*. [6] propose a universal adversarial perturbation to train a single adversarial perturbation applicable across an entire dataset. They further propose four different attack methods with universal adversarial perturbation to verify the adversarial robustness of IQA models in [37].

In the black-box scenario, Korhonen and You [7] utilize a substitute model to generate adversarial examples and then attack target models. The efficacy of such attacks is significantly influenced by the choice of the substitute model and the dataset used for training. For example, when Zhang *et al*. [5] demonstrate a notable variance in attack performance against the CORNIA model [38] depending on the substitute model employed, illustrating a disparity of 0.1151 in the Spearman's Rank-Order Correlation Coefficient (SROCC) when using different substitute models (UNIQUE [39] vs. BRISQUE [40]).

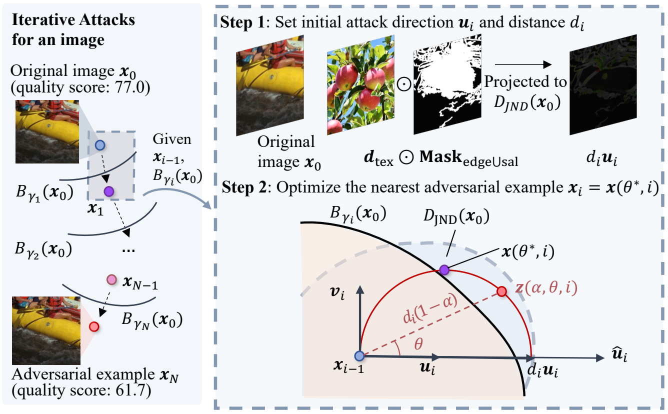
Iterative Attacks for an image
Original image $\boldsymbol{x}_0$ (quality score: 77.0)
Given $\boldsymbol{x}_{i-1}$, $B_{\gamma_i}(\boldsymbol{x}_0)$
$B_{\gamma_1}(\boldsymbol{x}_0)$
$\boldsymbol{x}_1$
$B_{\gamma_2}(\boldsymbol{x}_0)$
...
$\boldsymbol{x}_{N-1}$
$B_{\gamma_N}(\boldsymbol{x}_0)$
Adversarial example $\boldsymbol{x}_N$ (quality score: 61.7)
Step 1: Set initial attack direction $\boldsymbol{u}_i$ and distance d_i
Original image $\boldsymbol{x}_0$
$\boldsymbol{d}_{\text{tex}} \odot \textbf{Mask}_{\text{edgeUsal}}$
Projected to $D_{JND}(\boldsymbol{x}_0)$
$d_i\boldsymbol{u}_i$
Step 2: Optimize the nearest adversarial example $\boldsymbol{x}_i = \boldsymbol{x}(\theta^*, i)$
$B_{\gamma_i}(\boldsymbol{x}_0)$
$D_{\text{JND}}(\boldsymbol{x}_0)$
$\boldsymbol{x}(\theta^*, i)$
$\boldsymbol{z}(\alpha, \theta, i)$
$\boldsymbol{v}_i$
$d_i(1-\alpha)$
θ
$\boldsymbol{x}_{i-1}$
$\boldsymbol{u}_i$
$d_i\boldsymbol{u}_i$
$\hat{\boldsymbol{u}}_i$

Fig. 2. The framework of the proposed attack method.

This scenario underlines the necessity for black-box approaches that operate independently of substitute models, aiming to mitigate the reliance on specific training datasets and models. In the domain of Video Quality Assessment (VQA), Zhang *et al*. [41] propose a patch-based random search technique coupled with a score-reversed boundary loss for executing query-based black-box attacks on videos. While its score-reversed boundary loss provides effective guidance to the attack, there is still room for improving the efficiency due to its patch-based random search.

III. 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 multi-step attacks are defined. Finally, the optimization method with HVS prior for a single-step attack and adaptive score boundaries are described in detail. The framework of the attack is shown in Fig. 2.

A. Global-Local Optimization Objective

To attack an NR-IQA model f, a primary goal is to make the predicted quality score $f(\boldsymbol{x}_0 + \boldsymbol{\delta})$ of the attacked image deviate from the original score $f(\boldsymbol{x}_0)$ as much as possible, where $\boldsymbol{\delta}$ is the perturbation on the image $\boldsymbol{x}_0$. Meanwhile, the rank correlation of the predicted score with MOS is also important for NR-IQA, 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 (GL) optimization objective for a more reasonable attack. For an image set $\mathcal{I}$, we split it into the higher quality part $\mathcal{I}_h$ and the lower quality part $\mathcal{I}_l$ according to the original image score $f(\boldsymbol{x}_0)$. Our goal involves inducing the model to misjudge a high-quality image with a lower quality, and vice versa.

A larger $f(\boldsymbol{x}_0)$ corresponds to a higher quality of $\boldsymbol{x}_0$. The optimization objective is

$$\begin{aligned} \max_{\boldsymbol{\delta}} S(\boldsymbol{x}_0) * (f(\boldsymbol{x}_0+\boldsymbol{\delta}) - f(\boldsymbol{x}_0)), \\ \text{s.t. } \boldsymbol{x}_0+\boldsymbol{\delta} \in \mathcal{D}_{\text{JND}}(\boldsymbol{x}_0), \end{aligned} \tag{1}$$

where

$$S(\boldsymbol{x}_0) = \begin{cases} 1, & \boldsymbol{x}_0 \in \mathcal{I}_l \\ -1, & \boldsymbol{x}_0 \in \mathcal{I}_h \end{cases}. \tag{2}$$

And to restrain the visibility of perturbation, the adversarial sample $\boldsymbol{x}_0+\boldsymbol{\delta}$ is restrained to the JND neighborhood $\mathcal{D}_{\text{JND}}(\boldsymbol{x}_0)$ of $\boldsymbol{x}_0$. The neighborhood $D_{\text{JND}}(\boldsymbol{x}_0)$ of an image $\boldsymbol{x}_0$ could be written as:

$$\begin{aligned} D_{\text{JND}}(\boldsymbol{x}_0) = \{\boldsymbol{x} \big| |\boldsymbol{x}(l,j,k) - \boldsymbol{x}_0(l,j,k)| < m(l,j,k), \\ 0 \le l < H, 0 \le j < W, 0 \le k < C\}, \end{aligned} \tag{3}$$

where $m(l,j,k)$ is the minimum visibility threshold at pixel $\boldsymbol{x}_0(l,j,k)$ located on (l,j,k) on image $\boldsymbol{x}_0$ predicted by a JND model. The height, weight, and channel of $\boldsymbol{x}_0$ 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 $\boldsymbol{x}_0$ if the perturbation is under the threshold [13].

B. 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 an image $\boldsymbol{x}_0 \in \mathcal{I}_l$, and maximum and minimum MOS value $\text{MOS}_{\text{max}}, \text{MOS}_{\text{min}}$ in the dataset, we set the score boundary $B^l_\gamma(\boldsymbol{x}_0)$ as

$$B^l_\gamma(\boldsymbol{x}_0) = \{\boldsymbol{x} | f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \gamma(\text{MOS}_{\text{max}} - f(\boldsymbol{x}_0))\}. \tag{4}$$

This boundary includes samples with a higher quality score than $\boldsymbol{x}_0$. Then the attack is γ-success if $f(\boldsymbol{x}_0+\boldsymbol{\delta}) > f(\boldsymbol{x}_0) + \gamma(\text{MOS}_{\text{max}} - f(\boldsymbol{x}_0))$. And γ is a scalar to adjust the distance from $\boldsymbol{x}_0$ to $B^l_\gamma(\boldsymbol{x}_0)$, which corresponds to the attack intensity. While for $\boldsymbol{x}_0 \in \mathcal{I}_h$, $B^h_\gamma(\boldsymbol{x}_0) = \{\boldsymbol{x} | f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \gamma(\text{MOS}_{\text{min}} - f(\boldsymbol{x}_0))\}$. So we define the adversarial example $\boldsymbol{x}_0+\boldsymbol{\delta}$ is **$\boldsymbol{\gamma}$-success** if:

$$f(\boldsymbol{x}_0+\boldsymbol{\delta}) \begin{cases} > f(\boldsymbol{x}_0) + \gamma(\text{MOS}_{\text{max}} - f(\boldsymbol{x}_0)), \boldsymbol{x}_0 \in \mathcal{I}_l \\ < f(\boldsymbol{x}_0) + \gamma(\text{MOS}_{\text{min}} - f(\boldsymbol{x}_0)), \boldsymbol{x}_0 \in \mathcal{I}_h \end{cases}. \tag{5}$$

With the criterion in Eq. (5), 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 $\boldsymbol{x}_0 \in \mathcal{I}_l$ and initial γ_0, γ_{-1}, a series of score boundaries $B^l_{\gamma_1}(\boldsymbol{x}_0), ..., B^l_{\gamma_N}(\boldsymbol{x}_0)$ are set with $\gamma_i = \gamma_{i-1} + (\gamma_{i-1} - \gamma_{i-2}), i = 1, ..., N$. With the multi-step attacks, a series of adversarial images $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$ could be generated, which satisfy the property that $\boldsymbol{x}_i$ is γ_i-success $(i = 1, ...N)$. And $\boldsymbol{x}_N$ is used as the final adversarial example for $\boldsymbol{x}_0$. 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. III-D.

Algorithm 1 Algorithm for Iterative Attacks

Input: Original image $\boldsymbol{x}_0$, maximum number of score boundaries N, initial $\gamma_0 = 1/100$, $\gamma_{-1} = 0$

Output: Adversarial point $\boldsymbol{x}_N$

```
for $i \leftarrow 1, ..., N$ do
    $\gamma_i \leftarrow \gamma_{i-1} + (\gamma_{i-1} - \gamma_{i-2})$
    $\boldsymbol{x}_i, \gamma_i \leftarrow SingleAttack(\boldsymbol{x}_{i-1}, \gamma_i, ...)$ // Algorithm 2
end for
return $\boldsymbol{x}_N$
```

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

Input: Start point $\boldsymbol{x}_{i-1}$, original image $\boldsymbol{x}_0$, JND neighbourhood $D_{\text{JND}}(\boldsymbol{x}_0)$, score boundary $B_{\gamma_i}(\boldsymbol{x}_0)$, image $\boldsymbol{d}_{\text{tex}}$, maximum search times for a single-step attack T_{max}

Output: Adversarial point $\boldsymbol{x}_i$

```
 1: Search times $T \leftarrow 0$
 2: Generate $\mathbf{Mask}_{\text{edge}\cup\text{sal}}(\boldsymbol{x}_0)$
 3: Set initial attack direction
    $\hat{\boldsymbol{u}}_{\boldsymbol{i}} \leftarrow \tau \cdot \boldsymbol{d}_{\text{tex}} \odot \mathbf{Mask}_{\text{edge}\cup\text{sal}}(\boldsymbol{x}_0), \tau \sim U(-0.1, 0.1)$.
 4: $d_i \leftarrow ||\text{Proj}_{D_{\text{JND}}(\boldsymbol{x}_0)}(\hat{\boldsymbol{u}}_{\boldsymbol{i}})||, \boldsymbol{u}_i \leftarrow \text{Proj}_{D_{\text{JND}}(\boldsymbol{x}_0)}(\hat{\boldsymbol{u}}_{\boldsymbol{i}})/d_i$
 5: if $\boldsymbol{x}_{i-1} + d_i\boldsymbol{u}_i$ is not $\gamma_i$-success or $d_i < 1$ then
 6:     $T \leftarrow T + 1$
 7:     if $T > T_{\text{max}}$ 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 $\boldsymbol{x}_{i-1} + d_i\boldsymbol{u}_i$ 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 $\boldsymbol{v}_i$
19: $\boldsymbol{x}(\theta, i) \leftarrow d_i \cos\theta\,(\boldsymbol{u}_i \cos\theta + \boldsymbol{v}_i \sin\theta) + \boldsymbol{x}_{i-1}$
20: $\theta^* \leftarrow \operatorname*{argmin}_{\theta, \boldsymbol{x}(\theta,i) \text{ is } \gamma_i\text{-success}} \|\boldsymbol{x}(\theta, i) - \boldsymbol{x}_{i-1}\|$
21: $\boldsymbol{x}_i \leftarrow \text{Proj}_{D_{\text{JND}}(\boldsymbol{x}_0)}(\boldsymbol{x}(\theta^*, i))$
22: return $\boldsymbol{x}_i, \gamma_i$
```

C. Optimization Method for A Single-Step Attack

With the target decomposition in Sec. III-B, the attack objective of the ith-step attack of $\boldsymbol{x}_0$ could be set with:

Find $\boldsymbol{x}_i \in \mathcal{D}_{\text{JND}}(\boldsymbol{x}_0)$, subject to $\boldsymbol{x}_i$ is γ_i-success.

To solve this problem, we leverage a query-based black-box method [14] 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 $\boldsymbol{x}_{i-1}$, a preset unit attack direction $\boldsymbol{u}_i$ and a distance d_i which satisfies $\boldsymbol{x}_{i-1} + d_i\boldsymbol{u}_i$ is γ_i-success, and a stochastic unit direction $\boldsymbol{v}_i$ orthogonal to $\boldsymbol{u}_i$, the polar coordinate of a point $\boldsymbol{z}$ near $\boldsymbol{x}_{i-1}$ could be represented as

$$\boldsymbol{z}(\alpha, \theta, i) = d_i(1-\alpha)(\boldsymbol{u}_i \cos\theta + \boldsymbol{v}_i \sin\theta) + \boldsymbol{x}_{i-1}, \tag{6}$$

where $\alpha \in [0,1], \theta \in [-\pi, \pi]$. Given α, the trajectory of $\boldsymbol{z}(\alpha, \theta, i)$ is an arc, which is shown as the red arc in the lower part of Fig. 2 for $\theta \in [0, \pi/2]$. The goal is to choose (α, θ) to raise the probability of $\boldsymbol{z}(\alpha, \theta, i)$ being adversarial. With the theoretical analysis in [14], when $\alpha = 1 - \cos\theta$, probability of $\boldsymbol{z}(\alpha, \theta, i)$ being adversarial reaches maximum. So we mark $\boldsymbol{z}(1 - \cos\theta, \theta, i)$ as the candidate point $\boldsymbol{x}(\theta, i)$:

$$\boldsymbol{x}(\theta, i) = d_i \cos\theta \left(\boldsymbol{u}_i \cos\theta + \boldsymbol{v}_i \sin\theta\right) + \boldsymbol{x}_{i-1}. \tag{7}$$

The adversarial example $\boldsymbol{x}_i := \boldsymbol{x}(\theta, i)$ can be solved by

$$\begin{aligned} &\min_{\boldsymbol{x}(\theta, i)} \left\|\boldsymbol{x}(\theta, i) - \boldsymbol{x}_{i-1}\right\|, \\ &\text{s.t. } \boldsymbol{x}(\theta, i) \in D_{\text{JND}}\left(\boldsymbol{x}_0\right), \boldsymbol{x}(\theta, i) \text{ is } \gamma_i\text{-success.} \end{aligned} \tag{8}$$

There are three questions in attacking NR-IQA: 1) The reasonable preset direction $\boldsymbol{u}_i$ should be deliberately designed to guarantee an efficient attack. 2) The preset direction $\boldsymbol{v}_i$ should be designed to guarantee the orthogonality with $\boldsymbol{u}_i$. 3) The generated adversarial example should satisfy the constraint of D_{JND}, which is difficult in solving Eq. (8).

For the design of attack direction $\boldsymbol{u}_i$, in attacking classification tasks, a common approach is to employ stochastic perturbations, such as Gaussian noise, as the attack direction $\boldsymbol{u}_i$ in Eq. (7). However, this strategy, while effective for classification tasks, often proves inadequate when targeting NR-IQA models. For instance, when applying the attack perturbation $\boldsymbol{\delta}$ with preset values $\boldsymbol{\delta} \sim 0.15 \cdot \mathcal{N}(0,1)$ to the starting point $\boldsymbol{x}_0$ (normalized to the range $[0,1]$) of the classification task, we observe a high success rate of 92.6% for inducing misclassification of $\boldsymbol{x}_0 + \boldsymbol{\delta}$ in 500 random trials. However, when utilizing the same attack direction to target the NR-IQA model HyperIQA, the resulting average change between $f(\boldsymbol{x}_0)$ and $f(\boldsymbol{x}_0 + \boldsymbol{\delta})$ is merely 2.09 (within the range of predicted image scores $[0, 100]$). The efficiency of this stochastic attack direction dropped dramatically in the context of NR-IQA. In our pursuit of a more effective attack perturbation, we introduce a method that disrupts image regions that are sensitive to NR-IQA models while ensuring the perturbation remains invisible to the human eye.

1) Designing the preset direction $\boldsymbol{u}_i$: We generate $\boldsymbol{u}_i$ with three steps. Firstly, we design an initial perturbation $\boldsymbol{d}_{\text{tex}}$, which contains image texture and sparse noise. Secondly, the perturbation $\hat{\boldsymbol{u}}_i$ is obtained by confining $\boldsymbol{d}_{\text{tex}}$ to special image regions. Finally, the attack direction $\boldsymbol{u}_i$ is obtained by applying the projection and normalization operation on $\hat{\boldsymbol{u}}_i$.

In the first step, regarding the initial perturbation $\boldsymbol{d}_{\text{tex}}$, we are inspired by existing work exploring the sensitivity of DNNs to both image texture [11] and sparse noise [12], [42], and leverage the texture information and sparse noise extracted from the high-quality natural images $\boldsymbol{I}_{\text{nat}}$. The extracted information is denoted as high-frequency information $\boldsymbol{I}_{\text{hreq}} = g(\boldsymbol{I}_{\text{nat}})$, with an extraction function $g(\cdot)$. Meanwhile, $\boldsymbol{d}_{\text{tex}}$ is crafted to match the dimensions of the attacked image $\boldsymbol{x}_{i-1}$. More options for $\boldsymbol{d}_{\text{tex}}$ are explored in Sec. IV-D1.

In the second step, for designing $\hat{\boldsymbol{u}}_i$, our idea is to add disruption to the sensitive image regions for NR-IQA models, while the disruption is not visible to the human eye. Noting that the edge region and salient region are often critical to the judgment of IQA models [26], [33], we introduce a mask $\textbf{Mask}_{\text{edge}\cup\text{sal}}(\boldsymbol{x}_0)$ to confine attacks to these specific regions, whose role is to preserve perturbations in the edge and salient regions of $\boldsymbol{x}_0$, and remove perturbations in other regions. The designed perturbation could be formulated as $\hat{\boldsymbol{u}}_i = \tau \cdot \boldsymbol{d}_{\text{tex}} \odot \textbf{Mask}_{\text{edge}\cup\text{sal}}(\boldsymbol{x}_0)$, where $\odot$ is the Hadamard product, τ is a stochastic scalar drawn from a uniform distribution $U(-0.1, 0.1)$. The τ introduces different intensities in searching for $\hat{\boldsymbol{u}}_i$, enhancing the versatility and adaptability of the proposed method.

In the third step, to obtain the initial attack direction $\boldsymbol{u}_i$, $\hat{\boldsymbol{u}}_i$ is firstly modified by a projection to D_{JND}. The projection operation is defined as:

$$\text{Proj}_{D_{\text{JND}}(\boldsymbol{x}_0)}(\hat{\boldsymbol{u}}_i) := \underset{\tilde{\boldsymbol{u}}, \boldsymbol{x}_{i-1} + \tilde{\boldsymbol{u}} \in D_{\text{JND}}(\boldsymbol{x}_0)}{\operatorname{argmin}} \|\tilde{\boldsymbol{u}} - \hat{\boldsymbol{u}}_i\|. \tag{9}$$

Then the resulting projected vector is then normalized to obtain $\boldsymbol{u}_i$:

$$\begin{aligned} d_i &= ||\text{Proj}_{D_{\text{JND}}(\boldsymbol{x}_0)}(\hat{\boldsymbol{u}}_i)||, \\ \boldsymbol{u}_i &= \frac{\text{Proj}_{D_{\text{JND}}(\boldsymbol{x}_0)}(\hat{\boldsymbol{u}}_i)}{d_i}. \end{aligned} \tag{10}$$

To ensure $\boldsymbol{u}_i \in D_{\text{JND}}$, any $\hat{\boldsymbol{u}}_i$ with $d_i < 1$ is discarded and a new $\hat{\boldsymbol{u}}_i$ is regenerated.

2) Designing the preset direction $\boldsymbol{v}_i$: For $\boldsymbol{v}_i$, we follow the practice in [14] and generate $\boldsymbol{v}_i$ with the stochastic sample on the low-frequency subband of the original image. Firstly the image is transformed to the frequency domain with the full Discrete Cosine Transform (DCT) as in [43]. Then a fraction ρ of the transform coefficients is selected in the low-frequency subband. These selected transform coefficients are reassigned values uniformly distributed over $\{-1, 0, 1\}$, while the remaining coefficients are set to 0. The inverse DCT transform yields the direction $\boldsymbol{v}_i$. Then, to guarantee the orthogonality between $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$, the Gram-Schmidt process [44] is employed.

3) Generation of adversarial examples with D_{JND}: With $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$, Eq. (8) could be solved with a binary search of θ to

$$\boldsymbol{x}_i = \text{Proj}_{D_{\text{JND}}(\boldsymbol{x}_0)}(\boldsymbol{x}(\theta^*, i)) := \underset{\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}} \in D_{\text{JND}}(\boldsymbol{x}_0)}{\operatorname{argmin}} \|\tilde{\boldsymbol{x}} - \boldsymbol{x}(\theta^*, i)\|. \tag{11}$$

The algorithm for a single-step attack is shown in Algorithm 2.

D. 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 *Decreasing* and *Increasing strategies*: a) *Decreasing strategy*: when maximum search times for initial attack direction $\boldsymbol{u}_i$ is achieved in a single-step attack, we decrease γ_i to $\gamma_i - (\gamma_i - \gamma_{i-1})/2$. b) *Increasing strategy*: when initial attack direction $\boldsymbol{u}_i$ and distance d_i satisfy that $\boldsymbol{x}_{i-1} + d_i \boldsymbol{u}_i$ 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 16 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.

IV. Experiments

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

A. Experimental Setups

1) NR-IQA Models and Datasets: We choose four NR-IQA models DBCNN [2], HyperIQA [1], SFA [29], and CONTRIQUE [34], which are based on the various quality features extracted by DNN and are all widely recognized in the NR-IQA field. The LIVE dataset [45] with synthetic distortions and CLIVE dataset [46] 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 testing and the attack. No image content overlaps between the training and the test set. NR-IQA models are retrained on LIVE and CLIVE with their public code. The predicted scores are normalized to $[0, 100]$. For the attack, we use a random cropping with 224×224 for each image. And cropped images are fixed for all experiments.

2) Setting of Attacking Experiments: We set the number of score boundaries $N = 20$, with $\gamma_0 = 0.01$ for $B_{\gamma_0}(\boldsymbol{x}_0)$. Maximum search times T_{max} is set to 200, $\text{MOS}_{\text{max}} = 100$ and $\text{MOS}_{\text{min}} = 0$. The $\mathcal{I}_h$ and $\mathcal{I}_l$ are split by whether $f(\boldsymbol{x}_0)$ exceeds 50. The saliency maps of $\boldsymbol{x}_0$ is predicted with MBS [47], and edges of $\boldsymbol{x}_0$ are extracted by Canny operation [48]. In $\textbf{Mask}_{\text{edge}\cup\text{sal}}(\boldsymbol{x}_0)$, the pixel with a positive value in the salient map or edge map of $\boldsymbol{x}_0$ is set to 1 and other pixels are set to 0. The JND model of Liu *et al.* [13] is used, which predicts a single-channel JND map of an image $\boldsymbol{x}_0$. We subsequently apply this JND map on each color channel of the image, as the $D_{\text{JND}}(\boldsymbol{x}_0)$. For the convenience of optimization, the norm in the optimization target of Eq. (9) and (11) is set to L_2 norm. For $\boldsymbol{I}_{\text{hfre}}$, two high-quality images I60 and I71 are selected from the KADID-10k dataset [49] as $\boldsymbol{I}_{\text{nat}}$. The high-frequency information $\boldsymbol{I}_{\text{hfre}}$ is obtained by:

$$\boldsymbol{I}_{\text{hfre}} = g(\boldsymbol{I}_{\text{nat}}) = \boldsymbol{I}_{\text{nat}} - g_{\text{blur}}(\boldsymbol{I}_{\text{nat}}), \tag{12}$$

where $g_{\text{blur}}(\cdot)$ is a Gaussian blur operation with a 3×3 kernel. The select $\boldsymbol{I}_{\text{nat}}$ and their extracted $\boldsymbol{I}_{\text{hfre}}$ are shown in the first two images in the first row and second rows of Fig. 3. For each single-step attack, one of two $\boldsymbol{I}_{\text{hfre}}$ is randomly selected as the initial attack direction $\boldsymbol{d}_{\text{tex}}$. The ρ for the generation of $\boldsymbol{v}_i$ is set to 0.5 in our experiments. For the whole attack of an image, the maximum number of queries is limited to 8000.

3) 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 analyze the correlation between the predicted quality scores and MOS in the test set, employing three correlation indices: SROCC, Pearson linear correlation coefficient (PLCC), and Kendall rank-order correlation coefficient (KROCC). SROCC measures the monotonicity of the relation between MOS and predicted quality scores. PLCC measures the linear correlation between MOS and predicted quality scores, which accounts for the prediction accuracy. And KROCC measures the rank correlation with pairwise comparison. For M images within an image set, with MOS values represented as $l_1, ..., l_M$ and predicted scores as $f_1, ..., f_M$, the correlation indices are calculated as follows:

$$\text{SROCC} = 1 - \frac{6\sum_{i=1}^{M} d_i^2}{M(M^2-1)}, \tag{13}$$

$$\text{PLCC} = \frac{\sum_{i=1}^{M}(l_i - \mu_l)(f_i - \mu_f)}{\sqrt{\sum_{i=1}^{M}(l_i - \mu_l)^2 (f_i - \mu_f)^2}}, \tag{14}$$

$$\text{KROCC} = \frac{2(F_{\text{con}} - F_{\text{dis}})}{M(M-1)}, \tag{15}$$

where d_i is the rank difference between the MOS and predicted score for ith image. μ_l and μ_f denote the mean values of MOS and predicted scores, respectively. And F_{con} and F_{dis} indicate the count of concordant and discordant pairs. in the image set. Furthermore, we introduce an analysis of robustness, R, as initially proposed by Zhang *et al.* [5], to evaluate the variation in predicted quality scores before and after the attack:

$$R = \frac{1}{M}\sum_{i=1}^{M} \log\left(\frac{\max\{f_i - \beta_1, \beta_2 - f_i\}}{|f_i - f_i^\star|}\right), \tag{16}$$

where f_i and f_i^* are the predicted scores of the original and attacked version of ith image be attacked. M is the total number of attacked images. And β_1 and β_2 correspond to the minimum and maximum values of MOS in the image set. A smaller R value corresponds to a stronger attack for the attacker. In our experiments, β_1, β_2 are 3.42, 92.43 for the LIVE dataset, and 3.50, 90.55 for the CLIVE dataset. For the invisibility performance, we use SSIM [15] and LPIPS [16] to calculate the perceptual similarity between original images and adversarial examples.

4) Compared Attack Methods: To compare with the existing method, we choose the only black-box attack method for NR-IQA from Korhonen and You [7]. It utilizes a variant of ResNet50 [50] as its substitute model. We use a learning rate of 2 to generate the adversarial examples with its public code from the authors and mark it 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 [6] and Zhang [5]. For UAP, we use the perturbation generated with the substitute model PaQ-2-PiQ [51]. The amplitude for the perturbation is set to 0.024. For Zhang [5], we re-generate adversarial examples with

TABLE I
BLACK-BOX ATTACK PERFORMANCES ON FOUR NR-IQA MODELS. THE BEST AND SECOND ATTACK PERFORMANCES ARE MARKED WITH **BOLD** AND UNDERLINE.

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	**0.3985**	**0.4952**	**0.2802**	**20.93**	0.867	0.082	**0.1148**	**0.1943**	**0.0854**	**18.25**	0.891	0.100
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	0.6630	0.4861	14.38	0.865	0.113
	UAP	0.8685	0.8525	0.6959	10.73	0.792	0.141	0.7847	0.7750	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	**0.8334**	**0.8293**	**0.6525**	**13.01**	0.868	0.065	**0.4055**	**0.5240**	**0.2823**	**14.59**	0.879	0.111
SFA	Original	0.8425	0.8379	0.6546	11.99	-	-	0.8050	0.8322	0.6205	9.23	-	-
	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	**0.5953**	**0.6578**	**0.4335**	15.47	0.867	0.079	**0.3368**	**0.3818**	**0.2329**	16.52	0.882	0.101
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	**0.5705**	**0.6063**	**0.4046**	**19.40**	0.901	0.040	**0.0667**	**0.1234**	**0.0509**	**21.26**	0.896	0.078

substitute model DBCNN [2] with the perceptual constraint of LPIPS [16], and Lagrangian multiplier $\lambda = 9 \times 10^6$.

B. Attacking Results

We present the prediction performance of NR-IQA models before (marked as Original) and after the attack in Table I. Our method has superior attack effectiveness under the premise of maintaining good invisibility. 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 from above 0.7 to under 0.1, indicating a substantial disruption in the order relationship within the image set. Meanwhile, Zhang presents unstable attack performances with failure in attacking SFA on the LIVE dataset. For instance, the SROCC for SFA unexpectedly increases from under 0.85 before the attack to above 0.85 after the attack. The Korhonen method performs a better MAE value than our attack in targeting SFA because the substitute model it used is similar to the model used in SFA. But our method still achieves better SROCC/PLCC/KROCC performance.

Attack performance compared with the R metric is shown in Table II. Our method shows superior results, either ranking the best or second-best in attacking all four NR-IQA models. It achieves an R value of 0.826 attacking DBCNN on the CLIVE dataset, significantly surpassing the second-best method by over 0.14. When considering the SFA, we observe that the Korhonen method attains a lower R value on both datasets. This can be explained by the similarity between the substitute model employed by Korhonen and the model utilized in SFA.

For the robustness of NR-IQA models, all models present the vulnerability to black-box attacks on both synthetic distortions in the LIVE dataset and authentic distortions in the CLIVE dataset, which alarms the necessity to explore the security of NR-IQA. Among the four NR-IQA models, DBCNN and SFA suffer the most performance degradation with low correlation metrics, MAE, and R value. Meanwhile, NR-IQA models are less robust against attacks on images with authentic distortions compared to synthetic ones. This could be attributed to the more complex and variable patterns in authentic distortions, presenting an easier target for attacks. It is worthy to be further explored in future work.

TABLE II
ATTACK PERFORMANCE COMPARISON WITH THE R METRIC. THE BEST AND SECOND PERFORMANCES ARE MARKED WITH **BOLD** AND UNDERLINE.

Attacked NR-IQA	Attack Method	Attack Performance (R↓)	
		LIVE	CLIVE
DBCNN	Korhonen	1.571	0.982
	UAP	0.942	1.805
	Zhang	1.544	1.185
	Ours	**0.869**	**0.826**
HyperIQA	Korhonen	0.987	**0.762**
	UAP	**0.870**	1.153
	Zhang	1.403	1.220
	Ours	0.982	0.890
SFA	Korhonen	**0.654**	**0.495**
	UAP	1.168	0.866
	Zhang	1.510	1.107
	Ours	0.843	0.887
CONTRIQUE	Korhonen	1.516	1.123
	UAP	1.256	1.038
	Zhang	1.537	1.064
	Ours	**1.058**	**0.851**

C. Analysis of Adaptive Iterative Score Boundaries

In this subsection, we explore the impact of proposed adaptive iterative score boundaries, namely "iterative boundaries" and "adaptive optimization". The experiments are conducted on attacking the DBCNN model within the LIVE dataset.

1) Iterative Boundaries Analysis: Adversarial examples are generated with different numbers of adaptive score boundaries. In part B of Table III, the number N of score boundaries directly affects the intensity of the attack. Totally, an increase in N correlates with a heightened attack intensity. This correlation underscores the direct impact of iterative numbers on the intensity of adversarial examples generated.

TABLE III
BLACK-BOX ATTACK PERFORMANCE WITH DIFFERENT SETTINGS OF SCORE BOUNDARIES. EXPERIMENTS ARE CONDUCTED ON ATTACKING THE DBCNN MODEL WITHIN THE LIVE DATASET.

	Setting	Attack Performance		Invisibility	
		SROCC↓	MAE↑	SSIM↑	LPIPS↓
A. Original	-	0.9529	9.79	-	-
B. #Score Boundaries	$N = 5$	0.8147	16.55	0.900	0.055
	$N = 10$	0.5042	20.50	0.864	0.108
	$N = 20$	0.3985	20.93	0.867	0.082
	$N = 40$	**0.3963**	**21.74**	0.855	0.090
C. Adaptive Boundaries	Fixed	0.8849	15.17	0.919	0.061
	Adaptive	**0.3985**	**20.93**	0.867	0.082

2) Adaptive Optimization Analysis: Fixed boundaries with $\gamma_i = 0.01i, i = 1, ..., N$ are attempt comparing with adaptive optimization. N is set to 20. Results are shown in part C of Table III. In contrasting fixed boundaries with adaptive optimization, we observe that fixed boundaries, despite resulting in less perceptible perturbations, generally yield inferior attack performance. Adaptive optimization, where γ_i is dynamically adjusted across images and iterations, distinctly enhances attack intensity. This adaptability ensures that each attack is optimally tailored to a different target image, facilitating a stronger attack intensity.

D. Analysis of Different Settings for Initial Perturbation $\boldsymbol{d}_{tex}$

For the initial perturbation $\boldsymbol{d}_{\text{tex}}$ in Sec. III-C, the high-frequency information $\boldsymbol{I}_{\text{freq}}$ is employed. Different settings for $\boldsymbol{d}_{\text{tex}}$ are explored by attacking the DBCNN model within the LIVE dataset.

1) Different Options for $\boldsymbol{d}_{tex}$: There are different options for $\boldsymbol{d}_{\text{tex}}$, like natural images $\boldsymbol{I}_{\text{nat}}$ and high-frequency information $\boldsymbol{I}_{\text{hfre}}$. We verify the effect of different options for $\boldsymbol{d}_{\text{tex}}$: utilizing $\boldsymbol{I}_{\text{nat}}$ or $\boldsymbol{I}_{\text{hfre}}$ as $\boldsymbol{d}_{\text{tex}}$. For $\boldsymbol{I}_{\text{nat}}$, four high-quality images are randomly selected from the KADID-10k dataset, as shown in the first two images in the first row of Fig. 3, and one of them is randomly chosen as the $\boldsymbol{d}_{\text{tex}}$ in a single-step attack. For $\boldsymbol{I}_{\text{hfre}}$, the high-frequency information extracted from the first two images in Fig. 3 are used, and one of them is randomly chosen as the $\boldsymbol{d}_{\text{tex}}$ in a single-step attack. The attack performance is shown in the first part of Table IV. Utilizing $\boldsymbol{I}_{\text{nat}}$ as the $\boldsymbol{d}_{\text{tex}}$ proves to be effective to some degree, which makes SROCC decrease by around 0.15. Using $\boldsymbol{I}_{\text{hfre}}$ extracted from the same natural images achieves much better attack performance. It is attributed to the role of image texture and sparse noise in high-frequency images.

TABLE IV
BLACK-BOX ATTACK PERFORMANCE WITH DIFFERENT SETTINGS FOR $\boldsymbol{d}_{\text{TEX}}$. EXPERIMENTS ARE CONDUCTED ON ATTACKING THE DBCNN MODEL WITHIN THE LIVE DATASET.

	Setting	Attack Performance		Invisibility	
		SROCC↓	MAE↑	SSIM↑	LPIPS↓
A. Original	-	0.9529	9.79	-	-
B. Options of $\boldsymbol{d}_{\text{tex}}$	$\boldsymbol{I}_{\text{nat}}$	0.7925	16.95	0.907	0.039
	$\boldsymbol{I}_{\text{hfre}}$	**0.3985**	**20.93**	0.867	0.082
C. Components of $\boldsymbol{d}_{\text{tex}}$	Sparse Noise	0.7930	15.62	0.928	0.030
	Image Texture	0.4086	20.67	0.869	0.102
	Both	**0.3985**	**20.93**	0.867	0.082

TABLE V
THE MEAN AND STANDARD DEVIATION OF ATTACK PERFORMANCE MATRICES AND INVISIBILITY MATRICES WITH 10 $\boldsymbol{I}_{\text{HFRE}}$ RELATED TO DIFFERENT CONTENTS. THE EXPERIMENTS ARE CONDUCTED ON ATTACKING THE DBCNN MODEL WITHIN THE LIVE DATASET.

High-Frequency Image Name	Attack Performance		Invisibility	
	SROCC↓	MAE↑	SSIM↑	LPIPS↓
Original	0.9529	9.79	-	-
I60	0.5848	19.89	0.890	0.070
I71	0.5263	19.40	0.876	0.070
I03	0.5456	20.04	0.885	0.068
I52	0.6494	18.46	0.898	0.055
I66	0.5339	20.18	0.887	0.069
I35	0.6461	19.35	0.899	0.065
I36	0.5964	19.44	0.883	0.068
I31	0.5757	20.23	0.894	0.066
I77	0.4843	20.61	0.877	0.070
I22	0.4121	21.00	0.894	0.067
Mean	0.5555	19.86	0.888	0.067
Std	0.0723	0.73	0.008	0.004

2) Different Contents Related to $\boldsymbol{I}_{hfre}$: When using $\boldsymbol{I}_{\text{hfre}} = g(\boldsymbol{I}_{\text{nat}})$ as $\boldsymbol{d}_{\text{tex}}$, whether the image content of $\boldsymbol{I}_{\text{nat}}$ influences attack performance and invisibility of adversarial attacks? We randomly select 10 high-quality images $\boldsymbol{I}_{\text{nat}}$ with different image contents from the KADID-10K dataset, as shown in Fig. 3, which vary from scenarios and contents. Their corresponding $\boldsymbol{I}_{\text{hfre}}$ are regarded as ten different $\boldsymbol{d}_{\text{tex}}$. The same test set within the LIVE dataset is attacked ten times with these different $\boldsymbol{d}_{\text{tex}}$ respectively. As shown in Table V, the standard deviations of SROCC and MAE are merely under 0.1 and 1, which implies the variations among different $\boldsymbol{d}_{\text{tex}}$ are remarkably low. Meanwhile, the variation of SSIM and LPIPS are also minimal. It implies the image content of high-quality images has little effect on the performance and invisibility of the attack.

3) Image Texture vs. Sparse Noise in $\boldsymbol{I}_{hfre}$: As shown in Fig. 3, $\boldsymbol{I}_{\text{hfre}}$ is composed of two components: image texture and sparse noise. When using $\boldsymbol{I}_{\text{hfre}}$ as $\boldsymbol{d}_{\text{tex}}$, what are the effectiveness of different components of $\boldsymbol{I}_{\text{hfre}}$? To examine it, we segment[1] the high-freq image $\boldsymbol{I}_{\text{hfre}}$ into image texture and sparse noise, and regard them as the initial perturbation $\boldsymbol{d}_{\text{tex}}$ respectively. The high-quality images I60 and I71 are selected from the KADID-10k dataset [49] as $\boldsymbol{I}_{\text{nat}}$. The result of attacking DBCNN on the LIVE dataset is shown in Table IV. Both image texture and sparse noise contained in $\boldsymbol{I}_{\text{hfre}}$ are effective for attacking. Image texture witnesses a small SROCC and

[1] The segmentation was done manually on https://segment-anything.com.

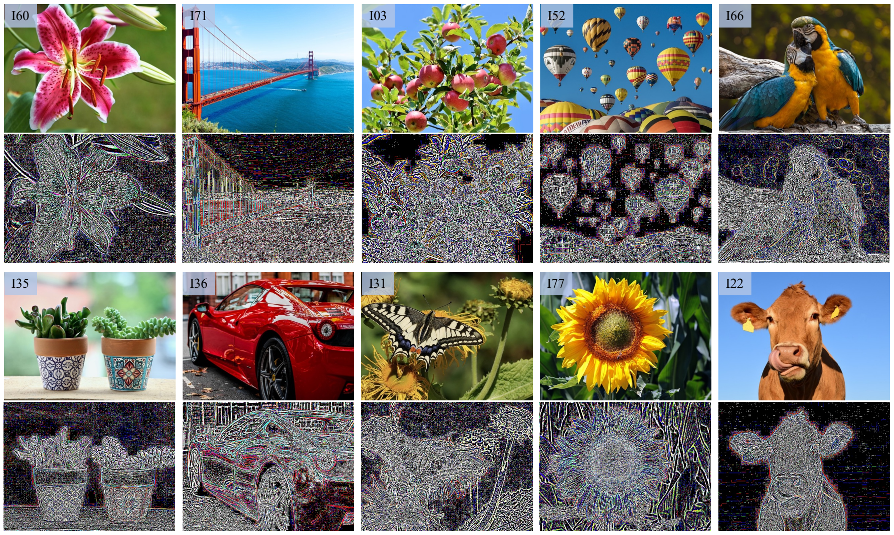

Fig. 3. High-quality images randomly selected from the KADID-10k dataset in the first/third row, their names are labeled in the top left corner. The corresponding extracted texture and noised images are in the second/fourth row.

larger MAE value, which implies it plays a more important role. Meanwhile, when both image texture and sparse noise are utilized (i.e. the whole $\boldsymbol{I}_{\text{hfre}}$ is used), the attack performance achieves the best among them. It confirms the role of image texture and sparse noise in high-frequency images when used as the initial attack direction.

E. Ablation Study

To examine the effectiveness of different parts of our attack method, we conduct a detailed performance analysis by attacking the DBCNN model within the LIVE dataset for different settings in Table VI. The original performance on unattacked images is shown in part A of Table VI.

TABLE VI
BLACK-BOX ATTACK PERFORMANCE WITH DIFFERENT SETTINGS. THE EXPERIMENTS ARE CONDUCTED ON ATTACKING THE DBCNN MODEL WITHIN THE LIVE DATASET.

	Setting	Attack Performance		Invisibility	
		SROCC↓	MAE↑	SSIM↑	LPIPS↓
A. Original	-	0.9529	9.79	-	-
B. Operation on Initial Attack Direction	Edge	0.4650	20.81	0.878	0.093
	Sal.	0.7207	17.14	0.900	0.076
	Edge+Sal.	0.3985	20.93	0.867	0.082
C. Strategy for Optimization	Incr.	0.8801	20.54	0.893	0.098
	GL	0.3985	20.93	0.867	0.082
D. Constraint of JND	w/o JND	-0.2989	29.91	0.729	0.212
	with JND	0.3985	20.93	0.867	0.082

1) Effect of $\boldsymbol{Mask}_{edge\cup sal}$ in Sec. III-C: The employment of edge mask and saliency mask operation confine perturbations to specific regions, as detailed in sec. III-C. The effect of these operations (marked as Edge and Sal. respectively) are explored in part B of Table VI. Compared to the original performance, both operations lead to an effective attack. Only using the edge mask performs a more important role with an MAE above 20. Utilizing both masks together yields the most potent attack.

2) Effect of GL Optimization in Sec. III-A: The GL optimization strategy, formulated in Eq. (1), aims to attack high-quality images to obtain lower quality scores and vice versa. To verify the effect of GL optimization, we compare it with an increasing strategy, as recommended in [6], [7]. The increasing strategy aims to obtain higher quality scores for all attacked images (marked as Incr.). The results are shown in part C of Table VI. Both Incr. and GL strategies provide effective attacks compared to the original performance, with a larger MAE value compared to the original performance. Meanwhile, the employment of GL optimization provides a dramatic decline in SROCC. It is primarily attributed to, in GL optimization, the different strategies for higher/lower quality images significantly altering the ranking of predicted scores after the attack.

3) Effect of JND Constraint in Sec. III-C: The JND constraint in Sec. III-C is designed to preserve the quality of adversarial examples, ensuring their perceptual invisibility. As shown in part D of Table VI, the absence of this constraint significantly compromises the attack's invisibility, highlighting the JND constraint's critical role in balancing effectiveness with imperceptibility.

F. Visualization of Adversarial Examples

The imperceptibility of perturbation is an important part of the adversarial attack. We guarantee it with the JND constraint. Meanwhile, to verify the effectiveness of the JND constraint, we calculated the perceptual similarity between the adversarial example and its original image with two metrics.

For an intuitive exhibition, we show the visualization of adversarial examples generated on the LIVE and CLIVE datasets in Fig. 4 and Fig. 5, respectively. It is noticeable our adversarial examples show good similarity with the original images. The perturbations generated by our method are more concentrated in the high-frequency region, like the rocks in Fig. 4 (c), and the black wires in Fig. 5 (b). Other methods

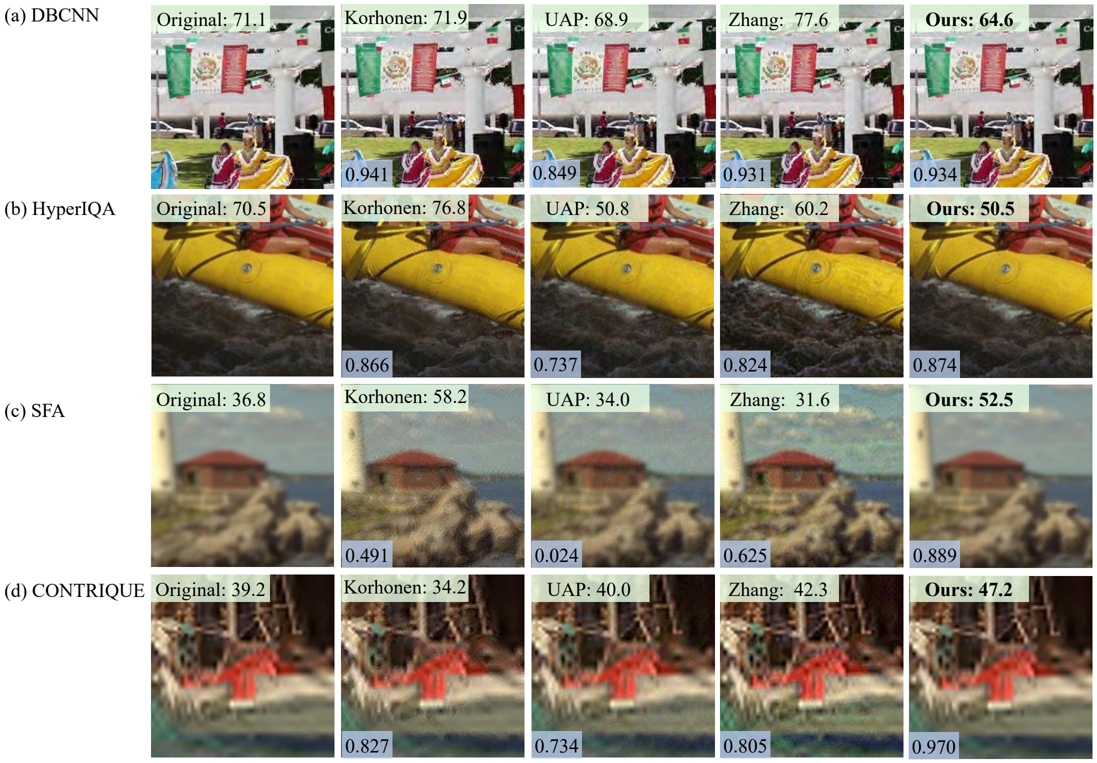

Fig. 4. Adversarial examples from different attack methods on the LIVE dataset. Each row shows one attacked NR-IQA model, and each column corresponds to one attack method. The predicted quality score is shown on the top of each image. The SSIM value between the adversarial example and the original image is shown at the bottom of each image. Our method is marked with **bold**.

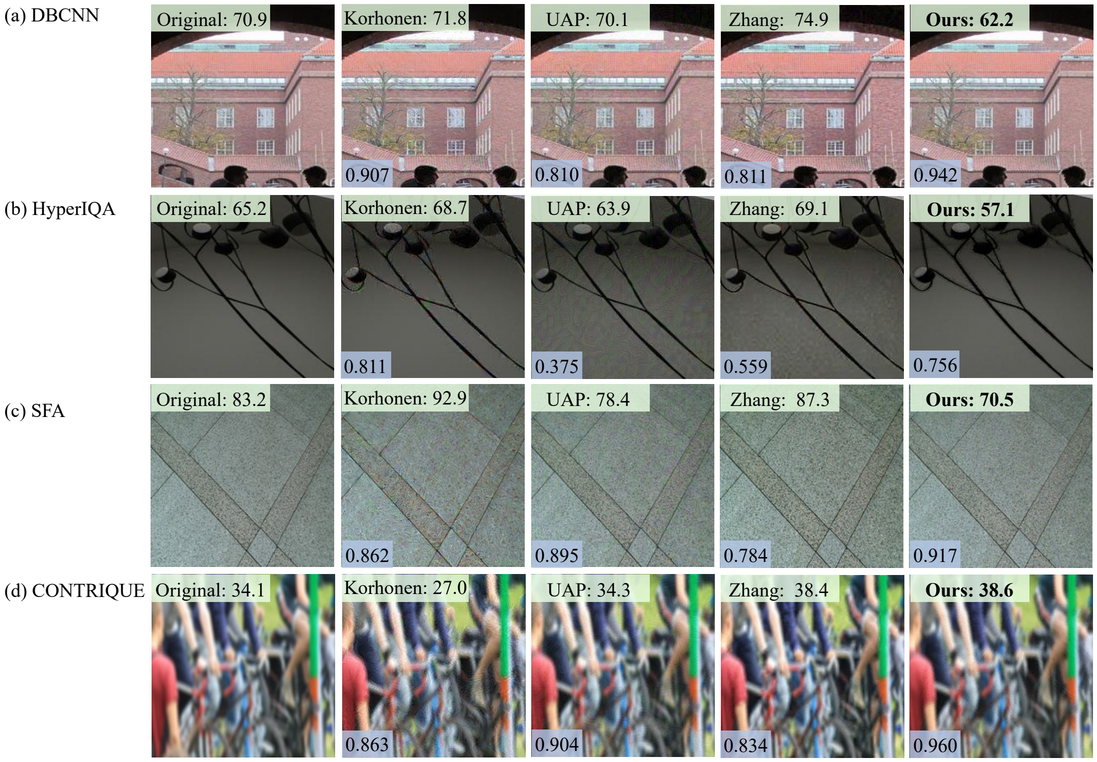

Fig. 5. Adversarial examples from different attack methods on the CLIVE dataset. Each row shows one attacked NR-IQA model, and each column corresponds to one attack method. The predicted quality score is shown on the top of each image. The SSIM value between the adversarial example and the original image is shown at the bottom of each image. Our method is marked with **bold**.

tend to have more perturbations in the low-frequency region, like the sky in Fig. 4 (c), and the wall in Fig. 5 (b), which are easier to be captured by the human eye. It implies the necessity of our constraint with $\mathbf{Mask}_{\text{edge}\cup\text{sal}}$. On the other hand, our attack performs a good invisibility on the blurred images like Fig. 4 (c) and Fig. 5 (d). In contrast, the other three attack methods generate textures that appear unnatural in these images, resulting in poor invisibility.

G. Comparison with Other Attack Methods

1) Comparison with Black-Box Attack in Image Classification Task: Although there are many existing query-based black-box attack methods proposed for the image classification task. They face the problem that classification boundaries can not be transferred to NR-IQA tasks and the querying without any prior is inefficient.

To compare our attack method with existing black-box attacks in classification tasks, we adopt a classical query-based black-box attack method SimBA [9] to the attack of NR-IQA. SimBA uses pixel-wise search to decrease the probability of correct classification predicted by the attacked model. To adopt SimBA for attacking NR-IQA, we utilize the optimization objective in Eq. (1). The loss function is:

$$J(\boldsymbol{x}_0, \boldsymbol{d}) = S(\boldsymbol{x}_0) * (f(\boldsymbol{x}_0 + \boldsymbol{d}) - f(\boldsymbol{x}_0)), \\ \text{s.t. } \boldsymbol{x}_0 + \boldsymbol{d} \in \mathcal{D}_{\text{JND}}(\mathbf{x_0}), \tag{17}$$

where

$$S(\boldsymbol{x}) = \begin{cases} -1, & \boldsymbol{x} \in \mathcal{I}_l \\ 1, & \boldsymbol{x} \in \mathcal{I}_h \end{cases}. \tag{18}$$

The adopted algorithm SimBA-IQA is shown in Algorithm 3. We use step size $\epsilon = 20/255$ for SimBA-IQA. The Cartesian basis of orthogonal search vectors Q is selected, which corresponds to each iteration we are increasing or decreasing one color of a single randomly chosen pixel.

The performance when attacking DBCNN on the LIVE dataset is shown in Table VII. Within the same number of queries, SimBA-IQA obtains substantially weaker attack intensity than our attack. With increasing the number of queries to $20,000$, the attack intensity of SimBA-IQA is improved, but it is still weaker than our attack, which is only $8,000$ query times used. Though SimBA-IQA guarantees a better SSIM value than ours, visual results (Fig. 6) indicate that human observers can not notice the difference between ours and SimBA-IQA. SimBA-IQA performs a low efficiency in attacking NR-IQA, cause it does not consider the prior information for attacking NR-IQA, and it uses the pixel-wise attack mechanism. Meanwhile, the set of initial attack directions and the optimization with SurFree in our attack method guarantee both effective attack intensity and a low number of queries.

2) Comparison with Black-Box Attack in VQA Task: In the field of VQA, Zhang *et al.* [41] propose a black-box attack method for videos. It employs a patch-based random search method, to assign a universal perturbation $\mathbf{m}$ to randomly selected patches across video frames. Each element in $\mathbf{m}$ is independently sampled from a discrete set $\{+\gamma, -\gamma\}$. Additionally, a score-reversed boundary loss is used to mislead the VQA model to inaccurately predict quality scores: lower scores for high-quality videos and vice versa. The difference between [41] and our attack is the perturbation generation strategy. Unlike a universal perturbation, our strategy leverages specific characteristics of the image content, such as texture or saliency information, to craft perturbations, which are more efficient and effective.

Algorithm 3 Algorithm for SimBA-IQA

Input: Original image $\boldsymbol{x}_0$, the set of orthogonal search vectors Q, and the step size ϵ

Output: Adversarial example $\boldsymbol{x}$

```
$\boldsymbol{d} \leftarrow 0$
$\hat{J} \leftarrow J(\boldsymbol{x}_0, \mathbf{0})$
while $T < T_{\text{max}}$ do
    Pick randomly without replacement: $\boldsymbol{q} \in Q$
    for $\alpha \in \{\epsilon, -\epsilon\}$ do
        $J' = J(\boldsymbol{x}_0, \alpha\boldsymbol{q})$
        if $J' < \hat{J}$ then
            $\boldsymbol{d} \leftarrow \boldsymbol{d} + \alpha\boldsymbol{q}$
            $\hat{J} \leftarrow J'$
        end if
    end for
end while
return $\boldsymbol{x}_0 + \boldsymbol{d}$
```

TABLE VII
PERFORMANCE COMPARISON WITH BLACK-BOX METHOD SIMBA-IQA. EXPERIMENTS ARE CONDUCTED ON ATTACKING THE DBCNN MODEL WITHIN THE LIVE DATASET.

Attack Method	#Queries	Attack Performance		Invisibility	
		SROCC↓	MAE↑	SSIM↑	LPIPS↓
Original	-	0.9157	12.07	-	-
SimBA-IQA	8,000	0.7713	16.23	0.956	0.006
SimBA-IQA	20,000	0.7127	17.77	0.946	0.007
Ours	8,000	**0.3985**	**20.93**	0.867	0.082

To evaluate our proposed attack against this method, we adopt it with the NR-IQA task, denoted as PatchAttack-IQA. In this adaptation, the height and width of the universal

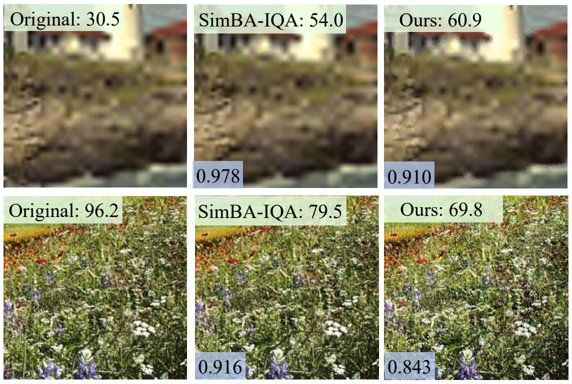

Fig. 6. Adversarial examples generated from SimBA-IQA and our attack against DBCNN on the LIVE dataset. The quality score predicted by the DBCNN/SSIM value is on the top/bottom of the images. Although the SSIM value of SimBA-IQA is a little higher, we can hardly distinguish the difference between adversarial samples generated by ours and SimBA-IQA.

perturbation $\mathbf{m}$ are set to 14×14, with γ set to $12/255$. The performance when attacking DBCNN on the LIVE dataset is shown in Table VIII. PatchAttack-IQA shows effective attack performance with a decline of SROCC with near 0.08 when the number of queries is $8,000$. Meanwhile, under the same number of queries, our attack shows a better attack performance with a decline of SROCC exceeding 0.6. With increasing the number of queries to $40,000$, the attack intensity of PatchAttack-IQA increased but was still weaker than our method. Though PatchAttack-IQA utilizes a score-reversed boundary loss similar to our GL optimization, its performance in the NR-IQA context is notably limited. This limitation can be attributed to its simplistic perturbation generation strategy, where no prior information on images is taken into consideration. Conversely, our method shows a more efficient attack with better invisibility.

TABLE VIII
PERFORMANCE COMPARISON WITH THE BLACK-BOX METHOD PATCHATTACK-IQA. EXPERIMENTS ARE CONDUCTED ON ATTACKING THE DBCNN MODEL WITHIN THE LIVE DATASET.

Attack Method	#Queries	Attack Performance		Invisibility	
		SROCC↓	MAE↑	SSIM↑	LPIPS↓
Original	-	0.9529	9.79	-	-
PatchAttack-IQA	$8,000$	0.8739	12.06	0.786	0.136
PatchAttack-IQA	$40,000$	0.8579	11.18	0.784	0.140
Ours	$8,000$	**0.3985**	**20.93**	0.867	0.082

V. 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

[1] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, “Blindly assess image quality in the wild guided by a self-adaptive hyper network,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3667–3676.

[2] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, “Blind image quality assessment using a deep bilinear convolutional neural network,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020.

[3] J. Gu, H. Cai, H. Chen, X. Ye, J. Ren, and C. Dong, “Image quality assessment for perceptual image restoration: A new dataset, benchmark and metric,” *arXiv preprint arXiv:2011.15002*, 2020.

[4] Y. Deng and K. Chen, “Image quality analysis for searches,” Nov. 25 2014, uS Patent 8,897,604.

[5] W. Zhang, D. Li, X. Min, G. Zhai, G. Guo, X. Yang, and K. Ma, “Perceptual attacks of no-reference image quality models with human-in-the-loop,” in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 2916–2929.

[6] E. Shumitskaya, A. Antsiferova, and D. S. Vatolin, “Universal perturbation attack on differentiable no-reference image- and video-quality metrics,” in *British Machine Vision Conference*, 2022, pp. 1–12.

[7] J. Korhonen and J. You, “Adversarial attacks against blind image quality assessment models,” in *Proceedings of the 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, 2022, pp. 3–11.

[8] Q. Sang, H. Zhang, L. Liu, X. Wu, and A. C. Bovik, “On the generation of adversarial examples for image quality assessment,” *The Visual Computer*, pp. 1–16, 2023.

[9] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, “Simple black-box adversarial attacks,” in *International Conference on Machine Learning*, 2019, pp. 2484–2493.

[10] X. Li, X. Zhang, F. Yin, and C. Liu, “Decision-based adversarial attack with frequency mixup,” *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1038–1052, 2022.

[11] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, “Image quality assessment: Unifying structure and texture similarity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022.

[12] A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard, “SparseFool: A few pixels make a big difference,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9087–9096.

[13] A. Liu, W. Lin, M. Paul, C. Deng, and F. Zhang, “Just noticeable difference for images with decomposition model for separating edge and textured regions,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1648–1652, 2010.

[14] T. Maho, T. Furon, and E. Le Merrer, “SurFree: A fast surrogate-free black-box attack,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 430–10 439.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[17] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in *International Conference on Learning Representations*, 2014, pp. 1–10.

[18] S. Baluja and I. Fischer, “Learning to attack: Adversarial transformation networks,” in *AAAI Conference on Artificial Intelligence*, vol. 32, 2018, pp. 2687–2695.

[19] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *International Conference on Learning Representations*, 2015, pp. 1–11.

[20] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “DeepFool: A simple and accurate method to fool deep neural networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2574–2582.

[21] Z. Zhao, Z. Liu, and M. A. Larson, “Towards large yet imperceptible adversarial image perturbations with perceptual color distance,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1036–1045.

[22] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, “ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models,” in *ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 15–26.

[23] N. Papernot, P. McDaniel, and I. Goodfellow, “Transferability in machine learning: From phenomena to black-box attacks using adversarial samples,” *arXiv preprint arXiv:1605.07277*, 2016.

[24] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, “Improving transferability of adversarial examples with input diversity,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2730–2739.

[25] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” in *ACM on Asia Conference on Computer and Communications Security*, 2017, pp. 506–519.

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

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

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

[29] D. Li, T. Jiang, W. Lin, and M. Jiang, “Which has better visual quality: The clear blue sky or a blurry animal?” *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1221–1234, 2019.

[30] T. Song, L. Li, D. Cheng, P. Chen, and J. Wu, “Active learning-based sample selection for label-efficient blind image quality assessment,”

IEEE Transactions on Circuits and Systems for Video Technology, 2023, early access.

[31] Z. Zhou, J. Li, D. Zhong, Y. Xu, and P. Le Callet, “Deep blind image quality assessment using dynamic neural model with dual-order statistics,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, early access.

[32] R. Ferzli and L. J. Karam, “A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB),” *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 717–728, 2009.

[33] S. Seo, S. Ki, and M. Kim, “A novel just-noticeable-difference-based saliency-channel attention residual network for full-reference image quality predictions,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 7, pp. 2602–2616, 2021.

[34] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, “Image quality assessment using contrastive learning,” *IEEE Transactions on Image Processing*, vol. 31, pp. 4149–4161, 2022.

[35] Z. Zhou, F. Zhou, and G. Qiu, “Blind image quality assessment based on separate representations and adaptive interaction of content and distortion,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, early access.

[36] J. Wang, K. C. Chan, and C. C. Loy, “Exploring CLIP for assessing the look and feel of images,” in *AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 2555–2563.

[37] E. Shumitskaya, A. Antsiferova, and D. Vatolin, “Towards adversarial robustness verification of no-reference image- and video-quality metrics,” *Computer Vision and Image Understanding*, vol. 240, p. 103913, 2024.

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

[39] W. Zhang, K. Ma, G. Zhai, and X. Yang, “Uncertainty-aware blind image quality assessment in the laboratory and wild,” *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, 2021.

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

[41] A. Zhang, Y. Ran, W. Tang, and Y.-G. Wang, “Vulnerabilities in video quality assessment models: The challenge of adversarial attacks,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 51 477–51 490.

[42] X. Dong, D. Chen, J. Bao, C. Qin, L. Yuan, W. Zhang, N. Yu, and D. Chen, “Greedyfool: Distortion-aware sparse adversarial attack,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 11 226–11 236.

[43] H. Li, X. Xu, X. Zhang, S. Yang, and B. Li, “QEBA: Query-efficient boundary-based blackbox attack,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1221–1230.

[44] E. Cheney and D. Kincaid, *Linear Algebra: Theory and Applications*. Jones and Bartlett Publishers, 2009.

[45] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms,” *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[46] D. Ghadiyaram and A. C. Bovik, “Massive online crowdsourced study of subjective and objective picture quality,” *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016.

[47] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, “Minimum barrier salient object detection at 80 FPS,” in *IEEE International Conference on Computer Vision*, 2015, pp. 1404–1412.

[48] J. Canny, “A computational approach to edge detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[49] H. Lin, V. Hosu, and D. Saupe, “KADID-10k: A large-scale artificially distorted IQA database,” in *International Conference on Quality of Multimedia Experience*, 2019, pp. 1–3.

[50] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[51] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. C. Bovik, “From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3575–3585.