Defense Against Adversarial Attacks on No-Reference Image Quality Models with Gradient Norm Regularization

Yujia Liu[1,2]*, Chenxi Yang[3,1]*, Dingquan Li[4], Jianhao Ding[1,2], Tingting Jiang[1,2]✉
[1]NERCVT, School of Computer Science, Peking University, China
[2]National Key Laboratory for Multimedia Information Processing, Peking University, China
[3]School of Mathematical Sciences, Peking University, China [4]Peng Cheng Laboratory, China
{yujia_liu,dingquanli,ttjiang}@pku.edu.cn; {yangchenxi,djh01998}@stu.pku.edu.cn

Abstract

The task of No-Reference Image Quality Assessment (NR-IQA) is to estimate the quality score of an input image without additional information. NR-IQA models play a crucial role in the media industry, aiding in performance evaluation and optimization guidance. However, these models are found to be vulnerable to adversarial attacks, which introduce imperceptible perturbations to input images, resulting in significant changes in predicted scores. In this paper, we propose a defense method to improve the stability in predicted scores when attacked by small perturbations, thus enhancing the adversarial robustness of NR-IQA models. To be specific, we present theoretical evidence showing that the magnitude of score changes is related to the ℓ_1 norm of the model's gradient with respect to the input image. Building upon this theoretical foundation, we propose a norm regularization training strategy aimed at reducing the ℓ_1 norm of the gradient, thereby boosting the robustness of NR-IQA models. Experiments conducted on four NR-IQA baseline models demonstrate the effectiveness of our strategy in reducing score changes in the presence of adversarial attacks. To the best of our knowledge, this work marks the first attempt to defend against adversarial attacks on NR-IQA models. Our study offers valuable insights into the adversarial robustness of NR-IQA models and provides a foundation for future research in this area.

1. Introduction

Deep Neural Networks (DNNs) have demonstrated remarkable performance across various domains [11, 16, 22], and Image Quality Assessment (IQA) is one of them. IQA aims to predict the quality of images consistent with human perception. And it could be categorized as Full-Reference (FR) and No-Reference (NR) according to the access to the reference images. While FR-IQA models specialize in assessing the perceptual disparities between two images, NR-IQA models focus on estimating a quality score for a single input image. The importance of IQA extends to many applications such as image transport systems [8], image inpainting [14] and so on [6, 27, 45]. Leveraging the capabilities of DNNs, recent IQA models have achieved remarkable consistency with human opinion scores [32].

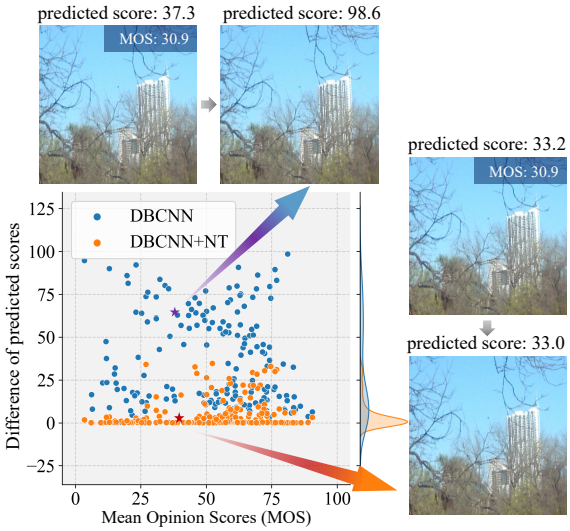

Figure 1. Comparison of DBCNN [46] trained with and without the proposed Norm regularization Training (NT) strategy under the Perceptual Attack [47] using the same setting. The absolute differences between predicted scores before and after the attack ($|s_{\text{after}} - s_{\text{before}}|$) for all test images are presented, with the fitted distribution displayed on the right side of the picture. An example is shown with predicted scores before and after the attack (zoom in for a better view). It is evident that DBCNN+NT exhibits smaller score changes compared to the baseline model.

However, the reliability of DNNs is challenged since they are found to be susceptible to adversarial perturbations. Attackers would mislead DNNs to make decisions inconsistent with human perception by adding carefully designed perturbations to inputs. This manipulation technique

*Equal contribution

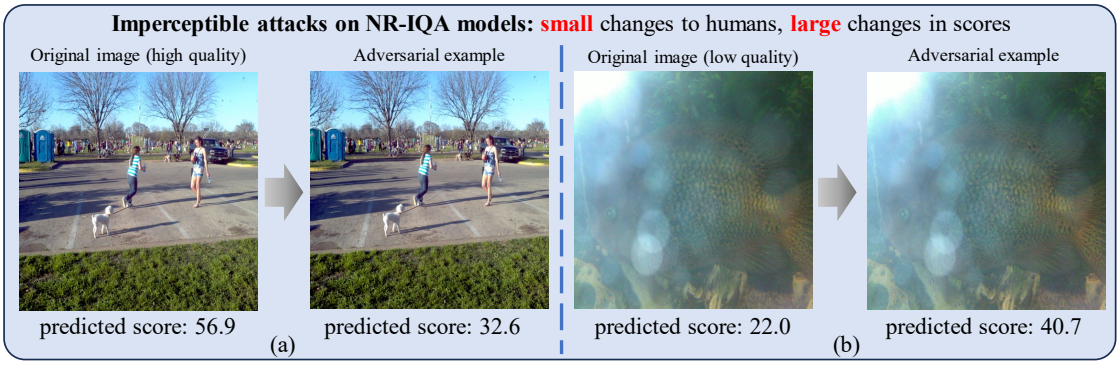

Figure 2. (Zoom in for a better view) Examples of adversarial attacks on the DBCNN [46] model. The range of MOS is $[0, 100]$.

is called the adversarial attack, and the perturbed inputs are called adversarial examples. The initial discovery of DNNs' vulnerability to adversarial attacks was in the context of classification tasks [33]. Subsequently, the threats of adversarial attacks are explored in various tasks, including object detection [34], segmentation [26], natural language processing [49], and many others [1, 19].

Recently, adversarial attacks on IQA models have garnered significant attention. Several attack methods targeting NR-IQA models have been proposed, where attackers aim to significantly change the predicted scores with small adversarial perturbations to input images. For instance, Zhang et al. [47] generated adversarial examples using the Lagrange multiplier method, imposing several constraints on the quality of adversarial examples. Besides, Shumitskaya et al. [30, 31] and Korhonen et al. [17] trained an individual model to generate adversarial examples.

However, despite these proposed attack techniques highlighting vulnerabilities in NR-IQA models, no methods have been put forth to defend against attacks and improve the adversarial robustness of NR-IQA methods. Training robust IQA models is essential for improving the reliability of these models in real-world applications. For instance, in online advertising, the quality of advertising images can significantly impact viewer engagement. Adversarial attacks on NR-IQA metrics could result in low-quality images being rated highly or high-quality images being rated lowly, as cases shown in Figure 2, potentially reducing the effectiveness of online advertising campaigns. Therefore, there is an impending need to train robust NR-IQA models, which is crucial for ensuring both the reliability and security of NR-IQA models in practical applications.

In this paper, we propose a defense method to improve the adversarial robustness of NR-IQA models in terms of reducing the quality score changes before and after adversarial attacks, which is supported by both theoretical foundations and empirical evidence. We analyze existing attacks on NR-IQA models and establish a theoretical foundation demonstrating the strong relationship between the adversarial robustness of an NR-IQA model f and the ℓ_1 norm of its gradient $\nabla_x f(x)$ concerning the input image x. We found that for an NR-IQA model, a smaller $||\nabla_x f(x)||_1$ implies a more robust model. Drawing upon the theoretical analysis, we propose the regularization of the gradient's ℓ_1 norm to enhance the adversarial robustness of NR-IQA models. A direct way to regularize $\|\nabla_x f(x)\|_1$ is adding it to the loss function in the training phase, which needs double backpropagation to compute the gradient of the regularization term with respect to model parameters considering the calculation mechanism of DNNs. However, double backpropagation is not currently scalable for large-scale DNNs [12]. Therefore, we approximate $\|\nabla_x f(x)\|_1$ by finite differences [38] instead of using it directly. The approximation result is utilized as the regularization term, effectively constraining $\|\nabla_x f(x)\|_1$.

To further verify our methodology, we conduct experiments on four baseline NR-IQA models and four attack methods. The results show the effectiveness of the norm regularization strategy in boosting baseline models' robustness against adversarial attacks. To the best of our knowledge, this is the first work to propose a defense method against adversarial attacks on NR-IQA models, which uses the ℓ_1 norm of the gradient as a regularization term. This paper establishes a theoretical connection between the robustness of the NR-IQA model against adversarial attacks and the gradient norm with respect to the input image. To support reproducible scientific research, we release the code at https://meilu.sanago.com/u l- 8 0 . a f f 9

2. Related Work

Adversarial attacks were first studied in classification tasks, so we introduce these attacks along with defense methods in Sec. 2.1. Sec. 2.2 and Sec. 2.3 provide a brief overview of NR-IQA models and attacks on these models.

2.1. Adversarial Attacks and Their Defenses in Classification Tasks

Based on the available knowledge of the target model, adversarial attacks can be divided into white-box attacks and black-box attacks. In white-box scenarios, attackers possess comprehensive knowledge about the target model. Some classic attacks treated the problem of generating adversarial examples as an optimization task [3, 10, 25]. Alternatively, some attacks proposed to train a new model to generate adversarial examples [20, 39, 48]. Conversely, in the case of black-box attacks, attackers are restricted to accessing only the output of the target model. A predominant strategy for executing black-box attacks involves generating adversarial examples on a known source model and subsequently transferring them to the target model [13, 21, 40].

To defend against adversarial attacks in classification tasks, a widely used method is adversarial training [33] and its variants [29, 35, 43, 49]. Adversarial training involves the generation of adversarial examples using specific attacks, which are incorporated into the training dataset so that the model can learn from these adversarial examples in the training phase. It acts as a form of data augmentation and helps to improve the robustness of the network.

2.2. IQA Tasks and Models

IQA tasks aim to predict image quality scores consistent with human perception (*i.e.*, Mean Opinion Score, MOS for short), which could be divided into FR and NR. For FR-IQA, it involves comparing a distorted image and its reference image to predict the quality score of the distorted image. Due to the difficulty of obtaining reference images in some authentic scenes, NR-IQA proposes to predict the quality score with only the distorted image.

NR-IQA methods extract features related to human perception of image quality. Some methods [23, 24] considered the hand-craft feature from Natural Scene Statistics. Further works explored the impact of image semantic information on human perception of image quality. HyperIQA [32] used a hypernetwork to obtain different quality estimators for images with different content. DBCNN [46] extracted distorted information and semantic information of images by two independent neural networks and combined them with a bi-linear pooling. LinearityIQA [18] proposed the normalization of scores in the loss function for faster convergence of the model. Meanwhile, some methods explored the effectiveness of different network architectures. MANIQA [41] and MUSIQ [15] utilized vision transformers [5] and verified their effectiveness in NR-IQA tasks.

2.3. Adversarial Attacks on NR-IQA Models

The issue of adversarial attacks within the context of IQA tasks has garnered some attention, although research in this area remains somewhat limited. Recently, some attack methods have been designed for NR-IQA models.

In white-box scenarios, the Perceptual Attack [47] modeled NR-IQA attacks as an optimization problem, where it employed the Lagrange multiplier method to solve this optimization problem. Perceptual Attack had tried different constraints on the image quality of adversarial examples, including the Chebyshev distance, LPIPS [44], SSIM [37] and DISTS [4]. Shumitskaya et al. [30] proposed to update a universal perturbation through a set of images and added it to clean images to attack NR-IQA models.

In black-box scenarios, the Kor. Attack [17] adapted ideas from attacks in classification tasks, creating adversarial examples through ResNet50 [11] and transferring them to the unknown target model. Likewise, Shumitskaya et al. [31] proposed to train a U-Net [28] to generate different adversarial perturbations for each image.

Although Korhonen et al. [17] adapted basic defense mechanisms from classification models to NR-IQA models, they did not investigate defense methods that are specifically designed for NR-IQA tasks.

3. Preliminary

3.1. Definition of Attacks on NR-IQA Models

Adversarial attacks on NR-IQA models aim to manipulate the predicted score of an input image x by an NR-IQA model f so that the objective score by models is inconsistent with the subjective score by humans. As for a successful attack on an NR-IQA model, there are imperceptible differences to the human eye between original images and adversarial examples, but these subtle perturbations result in large changes in the predicted scores generated by NR-IQA models. Examples are shown in Figure 2 (a) and (b). This attack can be mathematically described as follows:

$$\max \ |f(x+\delta) - f(x)|, \ \text{s.t.} \ D(x+\delta, x) \leqslant \epsilon, \quad (1)$$

where δ symbolizes the perturbation added to x, while the function $D(\cdot, \cdot)$ quantifies the perceptual distance between two images, and ϵ characterizes the tolerance of human eyes for image differences. An assumption is that when $D(x+\delta, x) \leqslant \epsilon$, the subjective score of $x+\delta$ is the same as x. In our methodology, we take $D(\cdot, \cdot)$ as defined by:

$$D(x+\delta, x) = \|\delta\|_\infty, \quad (2)$$

for the convenience of our theoretical analysis. Moreover, it has been used in attacks for IQA tasks [17, 47].

3.2. Robustness Evaluations of NR-IQA Models

The adversarial robustness of NR-IQA measures the stability of the NR-IQA model to imperceptible perturbations of input images generated by attacks. For example, when the

original image and its adversarial example have the same appearance, an NR-IQA model should give both the same quality scores. Figure 2 presents instances where DBCNN fails in this aspect. Researchers tend to assess the adversarial robustness of NR-IQA models by evaluating their IQA performance on adversarial examples [17, 47]. A better performance implies that the model is more robust.

Typically, the performance of an NR-IQA model is measured using four metrics: Root Mean Square Error (RMSE), Pearson's Linear Correlation Coefficient (PLCC), Spearman Rank-Order Correlation Coefficient (SROCC), and Kendall Rank-Order Correlation Coefficient (KROCC).[1] RMSE and PLCC are indicators of prediction accuracy, while SROCC and KROCC assess the prediction monotonicity [36]. When an NR-IQA model is attacked, greater robustness is indicated by smaller RMSE and larger PLCC, SROCC, and KROCC values. In this paper, we provide a theoretical analysis of robustness in terms of RMSE, and test all these metrics in the experimental part.

Recently, some IQA-specific metrics were proposed for evaluating the robustness of NR-IQA models [2, 47]. We will discuss these metrics in the supplementary material.

4. Methodology

In this section, we offer a theoretical exposition on improving NR-IQA models' adversarial robustness in terms of the magnitude of changes in predicted scores. We show that the robustness can be enhanced by regularizing the ℓ_1 norm of the gradient. We also propose a method for training a robust NR-IQA model using the norm regularization method.

4.1. Why to Regularize Gradient Norm?

In this subsection, we will outline the theoretical foundations regarding the relationship between the robustness in terms of score changes and the ℓ_1 norm of the gradient. It raises the necessity of regularizing the ℓ_1 norm of the input gradient of the predicted score. We prove that the magnitude of changes in predicted scores can be effectively approximated by the ℓ_1 norm of $\nabla_x f(x)$, with the assumption that the ℓ_∞ norm of perturbations is bounded.

Theorem 1. *Suppose f represents an NR-IQA model, ϵ is the strength of an attack, and x denotes an input image. The maximum change in predicted scores of x by f against ℓ_∞-bounded attacks is highly correlated to $\|\nabla_x f(x)\|_1$, which can be formulated as*

$$\sup_{\delta:\|\delta\|_\infty \leqslant \epsilon} |f(x+\delta) - f(x)| \approx \epsilon \|\nabla_x f(x)\|_1. \tag{3}$$

Proof. To begin, we apply the first-order Taylor expansion to the function $f(x+\delta)$ in the vicinity of x, yielding:

$$f(x+\delta) \approx f(x) + \delta^T \nabla_x f(x). \tag{4}$$

Then, $|f(x+\delta) - f(x)| \approx |\delta^T \nabla_x f(x)|$. Meanwhile, $|\delta^T \nabla_x f(x)|$ has the maximum value when $\delta = \epsilon \cdot \text{sign}(\nabla_x f(x))$, and this leads to Eq. (3). □

This theorem establishes the connection between changes in predicted scores and the ℓ_1 norm of the gradient. According to Theorem 1, suppose the strength of attacks ϵ is fixed, then the extent of score changes is primarily determined by the ℓ_1 norm of the gradient $\nabla_x f(x)$. In practical terms, this signifies that the regularization of $\|\nabla_x f(x)\|_1$ will lead to smaller fluctuations in predicted scores and thereby improve the adversarial robustness of f against imperceptible attacks.

4.2. How to Regularize Gradient Norm?

To train a robust NR-IQA model incorporating gradient norm regularization, a direct way is to add the ℓ_1 norm of gradients to the loss function, *i.e.*,

$$L(f,x) = L_{\text{IQA}}(f,x) + \lambda \cdot \|\nabla_x f(x)\|_1^2. \tag{5}$$

The loss function $L(f,x)$ comprises two components: the loss L_{IQA} tailored to the specific NR-IQA task, and the norm regularization term with a positive weight λ.

However, directly adding the term $\|\nabla_x f(x)\|_1$ to the loss function leads to the requirement of double backpropagation for computing the gradient of this term with respect to model parameters, which is time-consuming and currently not suitable for large-scale DNNs [7]. Therefore, we employ an approximation technique for the regularization term. Drawing inspiration from the methodology presented in the work [12], we leverage the finite difference [38] technique to estimate $\|\nabla_x f(x)\|_1$, *i.e.*,

$$\|\nabla_x f(x)\|_1 \approx \left| \frac{f(x+h\cdot d) - f(x)}{h} \right|, \tag{6}$$

where $h \in \mathbb{R}^+$ is the step size and $d = \text{sign}(\nabla_x f(x))$.

Finally, the loss function with the regularization of the ℓ_1 norm of the gradient is as follows:

$$L(f,x) = L_{\text{IQA}}(f,x) + \lambda \cdot \left| \frac{f(x+h\cdot d) - f(x)}{h} \right|^2. \tag{7}$$

5. Experiments

In this section, we present extensive experiments conducted on various NR-IQA baseline models to validate the efficacy of our proposed Norm regularization Training (NT) strategy. We briefly overview our experimental setup in Sec. 5.1. Subsequently, in Sec. 5.2, we demonstrate the enhancement in robustness achieved by the NT strategy against a diverse set of attacks. Furthermore, we illustrate the role of the finite difference approximation in reducing the ℓ_1 norm of the gradients (Sec. 5.3), as well as the relationship between attack intensity and robustness (Sec. 5.4). We also perform ablation studies on hyperparameters λ and h in Sec. 5.5.

[1] Formulations of these metrics are in the supplementary material.

5.1. Experimental Settings

Experiments were carried out on the popular LIVEC dataset [9]. We randomly selected 80% of the images for training and the remaining 20% for testing and attacks.

Our experiments to assess the robustness of NR-IQA models are structured along three key dimensions, as depicted in Figure 3. The first dimension revolves around the choice of the baseline models. We evaluate our NT strategy on four prominent NR-IQA baseline models: HyperIQA [32], DBCNN [46], LinearityIQA [18], and MANIQA [41]. Each of these baseline models is referred to as "baseline," while the models trained with NT are denoted as "baseline+NT." The NT strategy is applied to HyperIQA with the weight $\lambda = 0.001$, DBCNN, and LinearityIQA with $\lambda = 0.0005$ and MANIQA with $\lambda = 0.003$. For all models, the step size $h = 0.01$. Further training settings are provided in the supplementary material.

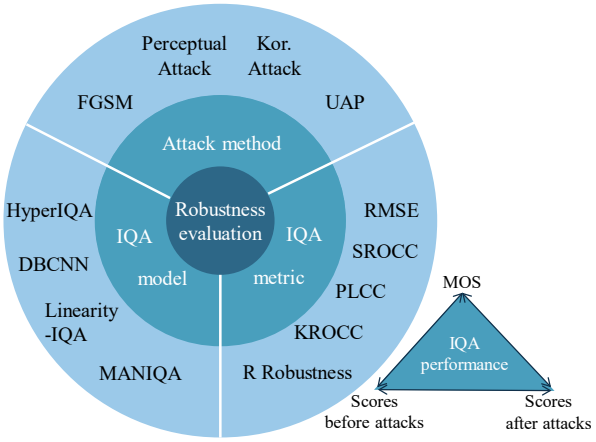

Figure 3. Three dimensions in experimental settings.

The second dimension involves the selection of attack methods. We employ four attack methods designed for NR-IQA tasks. These attacks include two white-box attacks: FGSM[2] [10] and Perceptual Attack [47], as well as two black-box attacks: UAP[3] [30] and Kor. Attack [17]. To ensure fairness in our evaluations, each attack method uses the same setting (*i.e.*, employing the same hyperparameters in attack) when targeting different models. We set different hyperparameters for different attacks and ensure the majority of attacked images' SSIM [37] was above 0.9 to satisfy the assumption that the MOS is the same for both images before and after the attack. Detailed hyperparameter information for these attacks can be found in the supplementary material. Black-box attacks are indicated with an asterisk in the tables of this paper.

[2]FGSM was originally designed for classification tasks, but we modify its loss to for NR-IQA tasks (refer to the supplementary material).

[3]UAP is proposed as a white-box attack. We employ its perturbation generated on PaQ-2-PiQ model [42], and serve UAP as a black-box attack.

The third dimension pertains to the evaluation metrics for NR-IQA models. As detailed in Sec. 3.2, we follow the evaluations in previous works and consider four metrics in this paper, *i.e.*, RMSE, PLCC, SROCC, and KROCC. Additionally, we incorporate the R robustness [47] into our analysis. This metric is proposed to assess the model robustness by measuring the relative score changes before and after attacks. Formulations of these metrics are shown in the supplementary material. Except for R robustness, other metrics are conventionally computed by comparing the predicted scores by models against MOS provided by humans.[4] In our evaluation, we extend the analysis of these metrics to assess predicted scores by a model both before and after attacks, since attackers only possess prediction scores before attacks. Notably, in accordance with the attack definition in Eq. (1), the NT strategy primarily focuses on the magnitude of changes in predicted scores between predicted scores before and after attacks. It could be measured by RMSE between predicted scores before and after attacks.

5.2. Robustness Improvement

In this subsection, we present the performance of NR-IQA models on unattacked images (where R robustness is not applicable), as well as their adversarial robustness against different attack methods. Our experimental results are summarized into four key observations. We provide additional analysis of adversarial robustness improvement in the supplementary material.

Table 1. Performance of NR-IQA models on unattacked images ("baseline / baseline+NT"). **Bold** denotes better value in a cell.

	HyperIQA base / +NT	DBCNN base / +NT	LinearityIQA base / +NT	MANIQA base / +NT
RMSE↓	**9.913** / 12.575	**10.897** / 13.140	**12.730** / 13.173	26.082 / **23.830**
SROCC↑	**0.899** / 0.859	**0.866** / 0.856	**0.832** / 0.820	**0.876** / 0.871
PLCC↑	**0.916** / 0.868	**0.892** / 0.849	**0.840** / 0.827	0.870 / **0.876**
KROCC↑	**0.728** / 0.670	**0.688** / 0.666	**0.641** / 0.627	**0.696** / 0.692

Observation 1. *The NT strategy results in a slight decrease in the performance of NR-IQA models on clean images.*

The performance of both the baseline models and their NT-enhanced versions on unattacked images are shown in Table 1. These metrics are calculated between MOS values and predicted scores on unattacked images. We can see that the NT strategy leads to a slight decrease in RMSE, SROCC, PLCC, and KROCC compared with baseline models. Similar trends were reported in the context of classification models that defense methods would cause a decline in classification accuracy on clean images. [33, 35, 43]. These findings suggest that enhanced robustness is often achieved at the cost of reducing performance on unattacked images.

[4]We normalize MOS to $[0, 100]$ for a straightforward comparison.

Table 2. The **RMSE**↓ metric of NR-IQA models against attacks ("baseline / baseline+NT"). **Bold** denotes better value in a cell.

	MOS & Predicted Score After Attack				Score Before Attack & Score After Attack			
	HyperIQA base / +NT	DBCNN base / +NT	LinearityIQA base / +NT	MANIQA base / +NT	HyperIQA base / +NT	DBCNN base / +NT	LinearityIQA base / +NT	MANIQA base / +NT
FGSM	25.729 / **16.828**	36.758 / **24.711**	50.823 / **40.104**	**24.899** / 25.712	19.174 / **7.885**	32.778 / **19.065**	48.128 / **36.988**	15.549 / **6.562**
Perceptual	13.565 / **12.593**	88.864 / **51.961**	115.395 / **80.949**	22.745 / **21.998**	6.360 / **0.130**	63.991 / **14.524**	115.732 / **80.857**	**0.079** / 0.189
UAP*	17.765 / **16.363**	19.775 / **17.188**	16.997 / **16.847**	**23.109** / 27.832	10.583 / **8.131**	14.833 / **10.922**	20.813 / **19.434**	5.795 / **5.592**
Kor.*	18.564 / **17.667**	**12.617** / 12.707	19.500 / **17.865**	18.423 / **17.395**	13.698 / **10.107**	6.514 / **5.298**	14.807 / **12.407**	7.759 / **6.680**

Table 3. The **SROCC**↑ metric of NR-IQA models against attacks ("baseline / baseline+NT"). **Bold** denotes better value in a cell.

	MOS & Predicted Score After Attack				Score Before Attack & Score After Attack			
	HyperIQA base / +NT	DBCNN base / +NT	LinearityIQA base / +NT	MANIQA base / +NT	HyperIQA base / +NT	DBCNN base / +NT	LinearityIQA base / +NT	MANIQA base / +NT
FGSM	0.021 / **0.810**	-0.318 / **0.200**	-0.375 / **-0.347**	0.417 / **0.772**	0.043 / **0.941**	-0.333 / **0.227**	-0.429 / **-0.426**	0.428 / **0.878**
Perceptual	0.815 / **0.858**	-0.127 / **0.643**	0.477 / **0.567**	**0.876** / 0.871	0.938 / **1.000**	-0.160 / **0.773**	0.542 / **0.685**	**1.000** / **1.000**
UAP*	0.736 / **0.822**	0.705 / **0.760**	0.715 / **0.739**	0.773 / **0.839**	0.825 / **0.941**	0.836 / **0.887**	0.836 / **0.869**	0.923 / **0.976**
Kor.*	**0.808** / 0.802	**0.863** / 0.856	**0.775** / **0.775**	0.828 / **0.847**	0.892 / **0.922**	0.978 / **0.983**	**0.936** / **0.936**	0.942 / **0.969**

Nonetheless, the performance decline induced by the NT strategy is marginal and well within acceptable limits.

Observation 2. *The NT strategy significantly improves the robustness of NR-IQA models in most cases, where the robustness is in terms of RMSE, SROCC, PLCC, KROCC or R robustness.*

Due to space constraints, we only present the robustness results for RMSE (Table 2) and SROCC (Table 3) in this subsection. Comprehensive results for PLCC, KROCC, and R robustness can be found in the supplementary material. In both tables, columns 2-5 display the IQA metric calculated between MOS values of unattacked images and predicted scores on adversarial examples, while columns 6-9 showcase the metric calculated between predicted scores on unattacked images and scores on adversarial examples.

As shown in Table 2, when RMSE is computed between predicted scores before and after attacks, NR-IQA models trained with the NT strategy exhibit smaller score changes under nearly all attack scenarios compared to baseline models. These results confirm the correctness of our theoretical analysis in Sec. 4.1. The robustness improvement is especially significant when models are attacked by FGSM. The only exception is MANIQA when attacked by the Perceptual Attack. In this case, the RMSE of MANIQA is smaller than that of the NT-trained model, where the difference is only 0.11. We think this phenomenon can be attributed to inherent biases among test images. Furthermore, MANIQA and its NT-enhanced version exhibit significant robustness against the Perceptual Attack, as indicated by an SROCC value of 1 between predicted scores before and after the attack. This signifies that the Perceptual Attack has minimal impact on MANIQA and MANIQA+NT, resulting in a reasonably small difference in RMSE between the two models.

When considering RMSE results measured between MOS values and predicted scores after attacks, the robustness of NT-trained models is also improved. For example, the RMSE value of DBCNN under the FGSM attack is about 36.758, whereas that of DBCNN+NT is just 24.711. There are only 3 out of 16 cases where baseline+NT models perform worse than baseline models. Such occurrences are expected because the NT strategy does not leverage MOS information but relies on the original predicted scores.

Results shown in Table 3 demonstrate that the NT strategy can also enhance the robustness in terms of SROCC, although we are not clear about the theoretical connection between the NT strategy and SROCC. The improvement is particularly pronounced in white-box scenarios. Taking the HyperIQA model as an example, the robustness of the baseline model measured by SROCC is notably deficient under the FGSM attack. The SROCC value between MOS values and predicted scores is a mere 0.021, while the SROCC value between predicted scores before and after the FGSM attack is only 0.043. However, with the inclusion of the NT strategy, there is a significant enhancement in SROCC. The SROCC value between MOS values and predicted scores increases to 0.810, while the SROCC between scores before and after the FGSM attack rises to 0.941. This exemplifies the effectiveness of the NT strategy in boosting the SROCC robustness of NR-IQA models.

Observation 3. *In IQA tasks, the robustness in terms of distinct metrics is not completely the same.*

In Table 2 and Table 3, it is evident that a model showing robustness in terms of RMSE when subjected to an attack method may not necessarily exhibit robustness in SROCC.

Take the left part of the two tables as an example, for baseline models, we can see that HyperIQA achieves much better robustness in terms of RMSE than MANIQA against the UAP attack (17.765 vs. 23.109), but it performs worse in SROCC (0.736 vs. 0.773). Similar phenomena also occur with NT-trained models where LinearityIQA+NT shows better RMSE robustness but worse SROCC robustness than DBCNN+NT against the UAP attack. How to make a trade-off between the adversarial robustness from different perspectives brings challenges in IQA tasks, potentially opening up new avenues for further exploration and research.

Observation 4. *The NT strategy exhibits a more effective defense against white-box attacks compared to black-box attacks.*

Table 2 and Table 3 demonstrate that the improvement in RMSE / SROCC from the baseline to baseline+NT models is generally greater under white-box attacks than under black-box attacks. This trend is more clear from Table S7 in the supplementary material by comparing the averaged metrics of improvement.

This happens because the attack capability of existing black-box attacks is generally weaker than that of white-box attacks on NR-IQA models. Hence, the baseline models exhibit better robustness against black-box attacks compared to white-box attacks, making the robustness improvement brought by NT less evident in the black-box scenario. For instance, when attacked by black-box methods, the SROCC values between predicted scores before and after attacks for all baseline models exceed 0.8, while these values for most baseline models under white-box attacks are below 0.6. This observation highlights the importance of exploring effective black-box attacks on IQA models.

5.3. Norm Reduction

To validate the effectiveness of the NT strategy in reducing the norm of the gradient, as well as the accuracy of Eq. (6) in approximating the ℓ_1 norm, we generate distribution plots of $\|\nabla_x f(x)\|_1$. Here, x represents samples from the test set.

Figure 4 compares the norm distribution between baseline and baseline+NT models. We can see that the gradient norms of the baseline+NT models are all shifted towards the left compared to the baseline models. This indicates that models trained with the NT strategy exhibit a smaller ℓ_1 gradient norm concerning the input image compared to the baseline models. These results confirm that Eq. (6) serves as a reliable approximation of the gradient ℓ_1 norm.

To further demonstrate that a smaller $\|\nabla_x f\|_1$ enhances the robustness of an NR-IQA model against adversarial attacks, we draw a scatter plot to show the relationship between the adversarial robustness and the gradient norm. In Figure 5, the horizontal axis represents the logarithm of the average $\|\nabla_x f(x)\|_1$ value across all test images. Points in

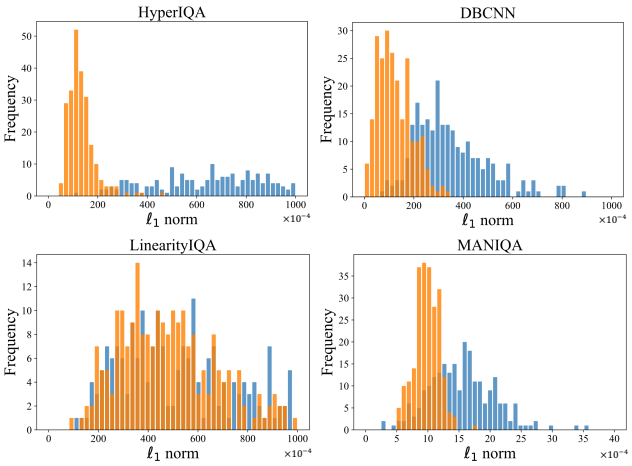

Figure 4. The comparison of ℓ_1 norm distribution of gradient between baseline models (blue) and baseline+NT models (orange).

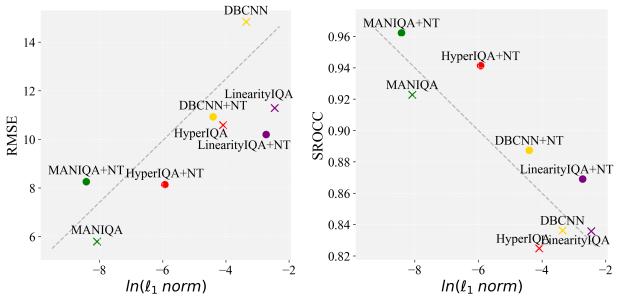

Figure 5. The relationship between the gradient norm and the robustness in terms of RMSE (left) and SROCC (right). The horizontal axis represents the logarithm of the average $\|\nabla_x f(x)\|_1$ value across all test images. All metrics are calculated between predicted scores before and after the UAP attack.

the left part of Figure 5 generally follow a diagonal distribution from bottom left to top right, indicating that models with smaller gradient norms tend to exhibit better robustness in terms of RMSE. Moreover, points in the right part of Figure 5 are generally distributed from the top left to the bottom right. This reflects that models with smaller gradient norms tend to exhibit better robustness in terms of SROCC.

5.4. Attack Intensity and Robustness

To evaluate the robustness of the baseline model and the baseline+NT model under different attack intensities, we adjust the strength of the iterative FGSM attack (illustrated in the supplementary material) with different iterations and the ℓ_∞ norm ϵ of perturbations. Generally, a higher number of iterations and larger ϵ values correspond to more potent attacks. The attack intensity is quantified using SSIM, with smaller SSIM values signifying greater attack intensity.

Figure 6 presents the performance of HyperIQA and its NT-trained versions under attacks with varying intensities. As the attack intensities increase, the RMSE and

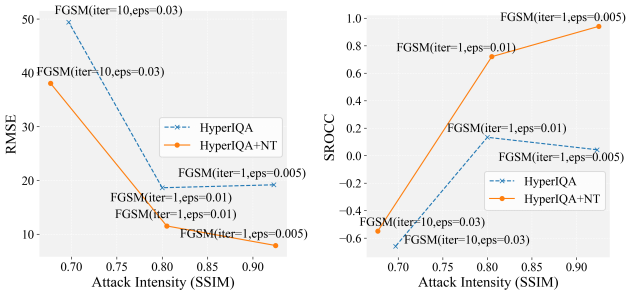

Figure 6. RMSE (left) and SROCC (right) values of HyperIQA and HyperIQA+NT models under different attack intensities. RMSE and SROCC are calculated between predicted scores before and after the FGSM attack.

SROCC values for HyperIQA and its NT version tend to get worse in general. This reflects that stronger attacks lead to decreased performance for both normally-trained and NT-trained models in most cases. Meanwhile, HyperIQA+NT model consistently keeps lower RMSE values than HyperIQA at the same attack intensity, regardless of the intensity levels. This demonstrates the effectiveness of the NT strategy against attacks with varying intensities.

5.5. Ablation Study

We conduct additional experiments to test the impact of hyperparameters in Eq. (7) for the NT strategy: the weight λ of the gradient norm and the step size h in the finite difference. Due to the space limit, we present partial results, and full results are shown in the supplementary material.

In Figure 7, we fix $h = 0.01$ and vary λ in Eq. (7) in the range from 0 to 0.003. Our analysis focuses on two aspects of an NR-IQA model: its performance on unattacked images and its robustness against attacks. For the former, we utilize SROCC on unattacked images across MOS values and predicted scores, and for the latter, we employ the RMSE between predicted scores before and after the FGSM attack. As λ increases, SROCC values on unattacked images tend to decrease on all baseline+NT models, while the RMSE values under the FGSM attack tends to decrease consistently. This implies that increasing λ enhances the robustness of NR-IQA models but leads to a performance decline on unattacked images.

To explore the effect of the step size h in Eq. (7) on the performance of IQA models, we fix $\lambda = 0.0005$ and vary h in $\{0.001, 0.01, 0.1, 1\}$ for DBCNN. In Table 4, we present the SROCC and RMSE values across MOS values and predicted scores for unattacked images, and SROCC and RMSE values between predicted scores before and after the FGSM attack for adversarial examples. In theory, a large h cannot sufficiently represent the neighborhood of x, so the approximation of the ℓ_1 norm is inaccurate. The experimental results also confirm this point where the robustness of the model is worse when $h = 0.1$ and 1. Conversely, an

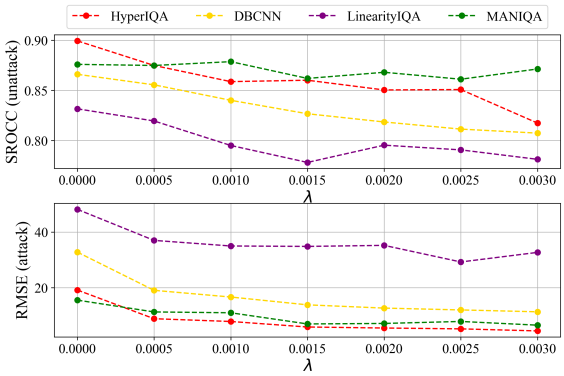

Figure 7. The impact of λ to SROCC on unattacked images and RMSE on FGSM attacked images.

Table 4. The comparison of different h of the DBCNN+NT model with unattacked images and FGSM attacked images.

		h			
		0.001	0.01	0.1	1
Unattacked	SROCC↑	0.788	0.856	0.846	0.844
	RMSE↓	16.099	14.138	12.417	14.809
Attacked	SROCC↑	0.577	0.200	-0.3832	-0.4406
	RMSE↓	7.356	19.065	28.785	18.767

exceedingly small h, such as $h = 0.001$, achieves effective defense performance but leads to a significant performance decline on unattacked images.

6. Conclusion

To the best of our knowledge, this is the first work designing IQA-specific defense methods against adversarial attacks. Our work offers a rigorous theoretical proof that the score changes of NR-IQA models are related to the ℓ_1 norm of the gradient when the perturbation is small. Furthermore, models trained with the proposed NT strategy exhibit significant improvement in adversarial robustness against both white-box and black-box attacks.

Limitations and future work. In this study, our primary theoretical analysis is on reducing the changes in predicted scores when NR-IQA models are exposed to attacks. Nevertheless, an interesting and valuable avenue for future research is the development of NR-IQA models that demonstrate robustness in terms of prediction monotonicity like SROCC. Besides, we intend to explore our method applied to FR-IQA models in future work.

Acknowledgements. This work is partially supported by Sino-German Center (M 0187) and the NSFC under contract 62088102. Thank Zhaofei Yu and Yajing Zheng for their valuable suggestions on the writing and illustrations. Thank Ruohua Shi for the support during rebuttal. Thank High-Performance Computing Platform of Peking University for providing computational resources.

References

[1] Hadi Abdullah, Aditya Karlekar, Vincent Bindschaedler, and Patrick Traynor. Demystifying limited adversarial transferability in automatic speech recognition systems. In *ICLR*, pages 1–17, 2022. 2

[2] Anastasia Antsiferova, Khaled Abud, Aleksandr Gushchin, Sergey Lavrushkin, Ekaterina Shumitskaya, Maksim Velikanov, and Dmitriy Vatolin. Comparing the robustness of modern no-reference image-and video-quality metrics to adversarial attacks. *arXiv preprint arXiv:2310.06958*, 2023. 4

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2016. 3

[4] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 44(5):2567–2581, 2022. 3

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[6] Zhengfang Duanmu, Kede Ma, and Zhou Wang. Quality-of-experience of adaptive video streaming: Exploring the space of adaptations. In *ACM MM*, pages 1752–1760, 2017. 1

[7] Chris Finlay and Adam M. Oberman. Scaleable input gradient regularization for adversarial robustness. *Machine Learning with Applications*, 3:100017, 2021. 4

[8] Haisheng Fu, Feng Liang, Jie Liang, Binglin Li, Guohe Zhang, and Jingning Han. Asymmetric learned image compression with multi-scale residual block, importance scaling, and post-quantization filtering. *IEEE TCSVT*, 33(8):4309–4321, 2023. 1

[9] Deepti Ghadiyaram and Alan C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE TIP*, 25:372–387, 2016. 5

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, pages 1–11, 2015. 3, 5

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 3

[12] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *ICLR*, pages 1–23, 2019. 2, 4

[13] Nathan Inkawhich, Wei Wen, Hai (Helen) Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *CVPR*, pages 7066–7074, 2019. 3

[14] Mariko Isogawa, Dan Mikami, Kosuke Takahashi, Daisuke Iwai, Kosuke Sato, and Hideaki Kimata. Which is the better inpainted image? Training data generation without any manual operations. *IJCV*, 127(11-12):1751–1766, 2019. 1

[15] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021. 3

[16] Manzoor Ahmed Khan, Hesham El-Sayed, Sumbal Malik, Talha Zia, Jalal Khan, Najla Alkaabi, and Henry Ignatious. Level-5 autonomous driving - Are we there yet? A review of research literature. *ACM Computing Surveys*, 55:27:1–27:38, 2023. 1

[17] 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. 2, 3, 4, 5

[18] Dingquan Li, Tingting Jiang, and Ming Jiang. Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *ACM MM*, pages 789–797, 2020. 3, 5

[19] Kaizhao Liang, Jacky Y. Zhang, Boxin Wang, Zhuolin Yang, Sanmi Koyejo, and Bo Li. Uncovering the connections between adversarial transferability and knowledge transferability. In *ICML*, pages 1–11, 2021. 2

[20] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive GAN for generating adversarial patches. In *AAAI*, pages 1028–1035, 2019. 3

[21] Yujia Liu, Ming Jiang, and Tingting Jiang. Transferable adversarial examples based on global smooth perturbations. *Computers & Security*, 121:1–10, 2022. 3

[22] Sushruta Mishra, Anuttam Dash, and Lambodar Jena. Use of deep learning for disease detection and diagnosis. In *Bio-inspired Neurocomputing*, pages 181–201. Springer Singapore, 2021. 1

[23] Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE TIP*, 21(12):4695–4708, 2012. 3

[24] Anush K. Moorthy and Alan C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE TIP*, 20(12):3350–3364, 2011. 3

[25] Seyed Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *CVPR*, pages 2574–2582, 2016. 3

[26] Utku Ozbulak, Arnout Van Messem, and Wesley De Neve. Impact of adversarial examples on deep learning models for biomedical image segmentation. In *MICCAI*, pages 300–308, 2019. 2

[27] Oren Rippel, Sanjay Nair, Carissa Lew, Steve Branson, Alexander G. Anderson, and Lubomir D. Bourdev. Learned video compression. In *ICCV*, pages 3453–3462, 2019. 1

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 3

[29] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NeurIPS*, pages 3353–3364, 2019. 3

[30] Ekaterina Shumitskaya, Anastasia Antsiferova, and Dmitriy S. Vatolin. Universal perturbation attack on differentiable no-reference image- and video-quality metrics. In *BMVC*, pages 1–12, 2022. 2, 3, 5

[31] Ekaterina Shumitskaya, Anastasia Antsiferova, and Dmitriy S. Vatolin. Fast adversarial CNN-based perturbation attack on No-Reference image-and video-quality

metrics. In *The First Tiny Papers Track at ICLR*, pages 1–4, 2023. 2, 3

[32] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *CVPR*, pages 3664–3673, 2020. 1, 3, 5

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, pages 1–10, 2014. 2, 3, 5

[34] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In *CVPR Workshops*, pages 1–7, 2019. 2

[35] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *ICLR*, pages 1–22, 2018. 3, 5

[36] VQEG. Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, 2000. 4

[37] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TIP*, 13(4):600–12, 2004. 3, 5

[38] Paul Wilmott, Sam Howison, and Jeff Dewynne. *The Mathematics of Financial Derivatives: A Student Introduction*. Cambridge University Press, 1995. 2, 4

[39] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *IJCAI*, pages 3905–3911, 2018. 3

[40] 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 *CVPR*, pages 2730–2739, 2019. 3

[41] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *CVPR Workshops*, pages 1190–1199, 2022. 3, 5

[42] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan C. Bovik. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *CVPR*, pages 3575–3585, 2020. 5

[43] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *NeurIPS*, pages 227–238, 2019. 3, 5

[44] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 3

[45] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. RankSRGAN: Generative adversarial networks with ranker for image super-resolution. In *ICCV*, pages 3096–3105, 2019. 1

[46] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE TCSVT*, 30(1):36–47, 2020. 1, 2, 3, 5

[47] 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 *NeurIPS*, pages 2916–2929, 2022. 1, 2, 3, 4, 5

[48] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. In *ICLR*, pages 1–15, 2018. 3

[49] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. FreeLB: Enhanced adversarial training for natural language understanding. In *ICLR*, pages 1–14, 2020. 2, 3