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

Yujia Liu, Chenxi Yang, Dingquan Li, Jianhao Ding, Tingting Jiang

In the supplementary material, we offer the formulation of NR-IQA metrics (Sec. 3.1), detailed proofs of the finite difference in Eq. (6) (Sec. 4.2), additional implementation details (Sec. 5.1), further robustness analysis (Sec. 5.2), and supplementary ablation study results (Sec. 5.5). Additionally, we present more visualization results.

S1. Formulations of RMSE, SROCC, KROCC, PLCC and R Robustness

In this section, we will introduce IQA-specific metrics RMSE, SROCC, KROCC, PLCC, and R robustness mentioned in Sec 3.2.

RMSE measures the difference between MOS values and predicted scores, which is represented as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - f_i)^2}. \tag{S1}$$

In this equation, N is the number of images. y_i and f_i represent the MOS and predicted score of the i^{th} image, respectively. The smaller the RMSE value is, the smaller the differences between the two groups of scores.

SROCC measures the correlation between MOS values and predicted scores to what extent the correlation can be described by a monotone function. The specific formulation is as follows:

$$\text{SROCC} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \tag{S2}$$

where d_i denotes the difference between orders of the i^{th} image in subjective and objective quality scores. The closer the SROCC value is to 1, the more consistent the ordering is between two groups of scores.

KROCC measures the degree of concordance in the ranking of MOS values and predicted scores. The formulation is:

$$\text{KROCC} = \frac{2(N_{\text{con}} - N_{\text{dis}})}{N(N-1)}. \tag{S3}$$

In this equation, N_{con} and N_{dis} represent the number of image pairs in the test dataset with consistent and inconsistent ranking of subjective and objective quality scores, respectively. The closer the KROCC value is to 1, the more consistent the ordering is between two groups of scores.

PLCC measures the linear correlation between MOS values and predicted scores, which is formulated as

$$\begin{aligned} \text{PLCC} &= \frac{\sum_{i=1}^{N}(y_i-\bar{y})(f_i-\bar{f})}{\sum_{i=1}^{N}(y_i-\bar{y})^2(f_i-\bar{f})^2}, \\ \bar{y} &= \frac{1}{N}\sum_{i=1}^{N} y_i, \bar{f} = \frac{1}{N}\sum_{i=1}^{N} f_i. \end{aligned} \tag{S4}$$

The closer the PLCC value is to 1, the higher the positive correlation between the two groups of scores.

R **robustness** was recently proposed by Zhang et al. [8]. It takes the maximum allowable change in quality prediction into consideration:

$$R = \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\max\{\beta_1 - f(x_i), f(x_i) - \beta_2\}}{|f(x_i) - f(x_i')|}\right), \tag{S5}$$

where N is the number of images, x_i is the i_{th} image to be attack, x_i' is the attacked version of x_i. $f(\cdot)$ is the IQA model for quality prediction. β_1 and β_2 are the maximum MOS and minimum MOS among all MOS values. A larger R value means better robustness.

S2. Proof of Eq. (6)

Eq. (6) in the main context illustrates how to approximate the ℓ_1 norm of $\nabla_x f(x)$ by the finite difference, which is expressed as

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

In this formula, h is a small step size and $d = \text{sign}(\nabla_x f(x))$. We provide a proof of this approximation in this section.

Proof. We expand $f(x+h\cdot d)$ at point x by the first order Taylor estimation, *i.e.*,

$$f(x+h\cdot d) \approx f(x) + h\cdot\nabla_x f(x)^T d. \tag{S6}$$

Since $d = \text{sign}(\nabla_x f(x))$, we have

$$\nabla_x f(x) d = \|\nabla_x f(x)\|_1. \tag{S7}$$

Therefore,

$$f(x + h \cdot d) \approx f(x) + h\|\nabla_x f(x)\|_1, \tag{S8}$$

and $\|\nabla_x f(x)\|_1$ can be approximated by

$$\left| \frac{f(x + h \cdot d) - f(x)}{h} \right|. \tag{S9}$$

□

S3. Experimental Settings

In Sec. 5.1 in the main manuscript, part of the experimental settings are reported. In this section, we report the experimental environment and detailed experimental settings in our experiments.

S3.1. Experimental Environment

We conducted all the training, test, and attack on an NVIDIA GeForce RTX 2080 GPU with 11GB of memory.

S3.2. Training Settings

For the four NR-IQA models considered in our study, namely, HyperIQA [5], DBCNN [7], LinearityIQA [3], and MANIQA [6], we used publicly available code provided by their respective authors to train these models on the same training dataset.

Due to the memory requirements associated with approximating the ℓ_1 norm, the batch size used for training the NR-IQA models had to be adjusted to prevent memory overflow. Other training settings are shown in Table S1. To ensure consistency and fairness in our comparisons, the same setting is utilized when training both the baseline and NT versions of each NR-IQA model.

Table S1. Detailed training settings for NR-IQA models and their NT versions. "Patches per Image" is marked as "-" if the input of the model is the whole image

Model	Architecture	Input Size	Patches per Image	Batch Size	Training Epochs
HyperIQA / HyperIQA+NT	ResNet50	224×224	25	16	16
DBCNN / DBCNN+NT	VGG and Its Variant	500×500	-	6	50
LinearityIQA / LinearityIQA+NT	ResNet34	498×664	-	4	30
MANIQA / MANIQA+NT	ViT-B/8	224×224	20	1	30

S3.3. Normalization of MOS

In our selected 4 NR-IQA models, MANIQA [6] is a special NR-IQA model in which MOS is scaled to the range of $[0, 1]$, and it leads to the predicted score in the range of $[0, 1]$. Furthermore, the different scales of MOS in different NR-IQA models result in a difference in the RMSE metric.

For a fair comparison across different NR-IQA models, we normalize the MOS into the range $[0, 100]$. The normalization formula is depressed as follows:

$$\text{MOS}_n = \frac{\text{MOS} - S_{\text{min}}}{S_{\text{max}} - S_{\text{min}}} \times 100. \tag{S10}$$

In this formula, S_{min} and S_{max} represent the minimal and maximal MOS of the training data, respectively. In this paper, $S_{\text{min}} = 3.42$ and $S_{\text{max}} = 92.43$.

S4. (I-)FGSM for NR-IQA Tasks

We mention the FGSM attack in Sec. 5.1 in the main manuscript. We will introduce the details of the setting of the FGSM attack in this section. FGSM [1] is first proposed for classification tasks, which is concise and efficient in attacking classification models. In our paper, we perceive FGSM as a white-box attack for NR-IQA models with a redesigned loss function. We will first introduce the FGSM attack in classification tasks and then the FGSM attack adapted to NR-IQA models below.

In the context of classification, FGSM is a straightforward non-iterative attack method, which is expressed as follows:

$$x_{\text{adv}} = x + \epsilon \, \text{sign}(\nabla_x \mathcal{L}(f(x), y)). \tag{S11}$$

In this equation, x_{adv} represents the adversarial example, x is the original image, ϵ denotes the ℓ_∞ norm bound of perturbations, $\mathcal{L}$ is the loss function, $f(\cdot)$ signifies the neural network function, and y represents the true label of x. A common use of $\mathcal{L}$ is cross-entropy loss.

I-FGSM is an iterative extension of FGSM, which is described as follows:

$$x^k_{\text{adv}} = \Pi_\epsilon \left\{ x^{k-1}_{\text{adv}} + \alpha \, \text{sign}(\nabla_x \mathcal{L}(f(x), y)) \right\}, \tag{S12}$$

where k is the current iteration step and α is the step size, the total number of iteration steps is K. The operator Π_ϵ projects the adversarial examples onto the space of the ϵ neighborhood in the ℓ_∞-ball around x.

In the NR-IQA task, we take y as the predicted score of the clean image x. In this paper, we choose the optimization object according to the predicted score of the image and define the loss function $\mathcal{L}$ as follows:

$$\mathcal{L}(f(x), y) \triangleq \mathcal{L}_{\text{mid}} = \begin{cases} f(x), & y \leqslant 50, \\ -f(x), & y > 50, \end{cases} \tag{S13}$$

where $f(x)$ represents the predicted score of the attacked image. The object is to maximize the predicted score for a low-quality image, thereby misleading the IQA model into assigning a high score to the adversarial example. Conversely, for a high-quality image, the goal is to minimize the predicted score to generate effective adversarial examples.

There is an interesting observation emerged from the experiment. We find that the choice of the loss function $\mathcal{L}$ has a significant impact on the efficacy of the FGSM attack. Specifically, we also try the mean absolute error loss:

$$\mathcal{L}(f(x), y) \triangleq \mathcal{L}_{\text{mae}} = |f(x) - y|, \tag{S14}$$

and the mean squared error loss:

$$\mathcal{L}(f(x), y) \triangleq \mathcal{L}_{\text{mse}} = (f(x) - y)^2. \tag{S15}$$

Taking the DBCNN as an example, we report the RMSE, SROCC, PLCC, and KROCC after the FGSM attack with different loss functions in Table S2[1] where $\epsilon = 0.005$. It is obvious that the effect of the FGSM attack is notably diminished when the loss function is $\mathcal{L}_{\text{mae}}$ or $\mathcal{L}_{\text{mse}}$. Especially when the mean absolute loss $\mathcal{L}_{\text{mae}}$ is used, the changes of RMSE, SROCC, PLCC, and KROCC for all models are very minimal.

Investigating the relationship between the loss function and the ability of attacks is an interesting domain of research.

Table S2. The attack ability of the FGSM attack with different loss functions. **Bold** denotes better value in a column

	MOS & Predicted Score After Attack			
	RMSE↑	SROCC↓	PLCC↓	KROCC↓
$\mathcal{L}_{\text{mae}}$	10.0734	0.8994	0.8844	0.7177
$\mathcal{L}_{\text{mse}}$	24.354	0.2795	0.2092	0.2096
$\mathcal{L}_{\text{mid}}$	**36.758**	**-0.318**	**-0.383**	**-0.146**
	Predicted Scores Before & After Attack			
	RMSE↑	SROCC↓	PLCC↓	KROCC↓
$\mathcal{L}_{\text{mae}}$	13.0829	0.754	0.705	0.6065
$\mathcal{L}_{\text{mse}}$	14.5819	0.6351	0.5886	0.4689
$\mathcal{L}_{\text{mid}}$	**32.778**	**-0.333**	**-0.418**	**-0.071**

S5. Hyperparameters of Attacks

In Sec. 5 in the main manuscript, 4 attack methods are utilized. For each attack method, there are hyperparameters which affect the strength of the attack. Table S3 summarizes the chosen hyperparameters in tested attack methods in the main experiment, *i.e.*, experiments in Sec. 5.

Table S3. Hyperparameters of attacks

Method	Hyperparameters
FGSM	one step, $\epsilon = 0.005, \alpha = 0.01$
Perceptual Attack	constraint: SSIM, weight =1,000,000
UAP*	scale $= 0.04$
Kor.* Attack	learning rate: 0.2

The meaning of these hyperparameters is explained in the original papers of attacks: FGSM [1], Perceptual attack [8], UAP [4] and Kor. attack [2].

S6. Further Robustness Analysis

In this section, we will further analyze the effectiveness of the NT strategy in improving the robustness of NR-IQA models. In Sec. S6.1, we present the robustness of baseline models and their NT-enhanced versions measured by KROCC, PLCC, and R robustness [8]. In Sec. S6.2, we report the average metrics of RMSE and SROCC improvement for both baseline models and their NT-enhanced versions. For each model, we provide the scatter plots of predicted scores before and after the perceptual attack in Sec. S6.3, which intuitively show the effectiveness of the NT strategy.

S6.1. Robustness in Terms of KROCC, PLCC and R Robustness

In Sec. 5.2 in the main manuscript, the robustness performances in terms of RMSE and SROCC are reported. Table S4, Table S5 and Table S6 show the robustness performances of NR-IQA models in terms of KROCC, PLCC, and R robustness against different attack methods, respectively. Specifically, We evaluate R robustness on four baseline methods as well as their NT-trained models with $\beta_1 = 100, \beta_2 = 0$.

In Table S4, NR-IQA models with the NT strategy outperform their baseline models under all attacks when KROCC is measured between predicted scores before and after attacks. Among them, HyperIQA+NT witnesses a larger improvement in KROCC compared to its baseline under the FGSM attack, with KROCC increasing from 0.043 of HyperIQA to 0.806 using the NT strategy. Meanwhile, MANIQA demonstrates strong robustness against the Perceptual Attack, achieving a KROCC (scores before and after the attack) value of 1. This means Perceptual Attack could not change the rank order of predicted scores before and after the attack on MANIQA. This phenomenon is also observed in the results of SROCC robustness.

In Table S5, NR-IQA models with the NT strategy perform better than their baseline models in most cases when PLCC is measured between predicted scores before and after attacks. For example, when the attack method is the

[1] As for attack methods, larger RMSE and smaller SROCC, KROCC, PLCC represents stronger attack ability.

Table S4. The **KROCC**↑ metric of NR-IQA models against attacks (with "baseline / baseline+NT"). **Bold** denotes better value in a cell

	MOS & Predicted Score After Attack				Score Before Attack & Score After Attack			
	HyperIQA	DBCNN	LinearityIQA	MANIQA	HyperIQA	DBCNN	LinearityIQA	MANIQA
	base / +NT	base / +NT	base / +NT	base / +NT	base / +NT	base / +NT	base / +NT	base / +NT
FGSM	0.020 / **0.610**	-0.146 / **0.136**	-0.197 / **-0.184**	0.296 / **0.584**	0.043 / **0.806**	-0.071 / **0.217**	**-0.156** / -0.171	0.332 / **0.749**
Perceptual	0.627 / **0.669**	-0.079 / **0.471**	0.350 / **0.415**	0.870 / **0.876**	0.837 / **0.997**	-0.091 / **0.628**	0.440 / **0.566**	**1.000** / **1.000**
UAP*	0.548 / **0.628**	0.510 /**0.568**	0.526 / **0.543**	0.578 / **0.651**	0.634 /**0.797**	0.643 / **0.708**	0.664 / **0.694**	0.766 / **0.871**
Kor.*	0.614 /**0.615**	**0.678**/0.669	0.585 /**0.587**	0.637 /**0.658**	0.724 /**0.777**	0.874 /**0.895**	0.777 /**0.786**	0.790 / **0.850**

Table S5. The **PLCC**↑ metric of NR-IQA models against attacks (with "baseline / baseline+NT"). **Bold** denotes better value in a cell

	MOS & Predicted Score After Attack				Score Before Attack & Score After Attack			
	HyperIQA	DBCNN	LinearityIQA	MANIQA	HyperIQA	DBCNN	LinearityIQA	MANIQA
	base / +NT	base / +NT	base / +NT	base / +NT	base / +NT	base / +NT	base / +NT	base / +NT
FGSM	-0.009 / **0.801**	-0.383 / **0.251**	-0.497 / **-0.387**	0.599 / **0.861**	0.042 / **0.926**	-0.418 / **0.196**	-0.569 / **-0.439**	0.535 / **0.929**
Perceptual	0.830 / **0.868**	-0.030 / **0.585**	0.487/ **0.528**	**0.696** / 0.691	0.937 / **1.000**	-0.005/ **0.719**	0.522 / **0.582**	**0.998** / 0.995
UAP*	0.733 / **0.817**	0.701 / **0.776**	0.694 / **0.729**	0.766 / **0.837**	0.826 / **0.943**	0.811 / **0.884**	0.805 / **0.876**	0.928 / **0.978**
Kor.*	0.801 / **0.806**	**0.875** / 0.868	**0.774** / **0.774**	0.838 / **0.856**	0.875 / **0.933**	0.972 / **0.980**	0.914 / **0.933**	0.942 / **0.969**

Percepural Attack, the PLCC of DBCNN is -0.005 while the PLCC of DBCNN+NT is 0.719. The only exception is MANIQA where MANIQA+NT performs worse than MANIQA when attacked by the Perceptual Attack. This trend is consistent with the results reported in RMSE robustness.

From Table S4 and Table S5, we can conclude that the robustness of NR-IQA models in terms of RMSE and PLCC have similar trends, while robustness in terms of SROCC and KROCC show similar patterns. Additionally, the robustness improvement caused by the NT strategy is more obvious when NR-IQA models are attacked in white-box scenarios than in black-box scenarios.

From Table S6, we can see that NR-IQA methods with our NT strategy generally perform better than their baselines. However, it's essential to note that the definition of R robustness assigns a higher weight to images with extremely large scores (close to β_1) or extremely small scores (close to β_2), whereas RMSE treats each image equally. Consequently, in scenarios where the DBCNN model is attacked by Kor. attack, the NT model shows improvement in the RMSE metric but a decrease in the R robustness compared to its baseline. Similar trends are observed when the LinearityIQA model is attacked by UAP. Although different metrics focus on different aspects, the proposed NT strategy improves all robustness metrics in most cases.

S6.2. Averaged Metrics of RMSE and SROCC Improvement

For an NR-IQA model subjected to an attack method, we calculate the difference in RMSE (or SROCC) between its NT version and the original model, denoted as ΔRMSE (or ΔSROCC). We then average ΔRMSE and ΔSROCC for both white-box and black-box attacks, as shown in Table S7, to corroborate *Observation 4* presented in the main manuscript. These results further confirm the effectiveness of NT in mitigating both white-box and black-box attacks.

Table S6. The R ↑ robustness of NR-IQA models against attacks (with "baseline / baseline+NT")

	HyperIQA	DBCNN	LinearityIQA	MANIQA
	base / +NT	base / +NT	base / +NT	base / +NT
FGSM	0.659 / **1.099**	0.328 / **0.671**	1.011 / **1.389**	2.957 / **3.864**
Perceptual	2.249/ **3.047**	0.938 / **2.076**	1.011 / **1.398**	**5.492**/ 4.784
UAP*	1.180 / **1.285**	1.054 / **1.067**	**1.161** / 1.096	3.459 / **3.464**
Kor.*	0.980 / **1.092**	**1.333** / 1.323	0.883 / **1.000**	3.299 / **3.319**

Table S7. Averaged ΔRMSE ↓/ averaged ΔSROCC ↑.

	HyperIQA	DBCNN	LinearityIQA	MANIQA
White	-8.7595 / 0.4800	-31.5900 / 0.7465	-23.0075 / 0.0730	-4.4385 / 0.2250
Black	-3.0215 / 0.0730	-2.5635 / 0.0280	-1.8895 / 0.0165	-0.6410 / 0.0400
Overall	-5.8905 / 0.2765	-17.0768 / 0.3873	-12.4485 / 0.0448	-2.5398 / 0.1325

S6.3. Distributions of Predicted Scores

Figure S1–S4 illustrate the absolute differences between predicted scores before and after various attacks for all test images (from the first row to the last row: FGSM, Perceptual Attack, UAP, and Kor. attack). The fitted distribution is presented on the right side of each image.

It is evident that all models trained with the NT strategy exhibit smaller score changes compared to their corresponding baseline models. Additionally, we observe an interesting trend: the NT strategy enhances robustness for different

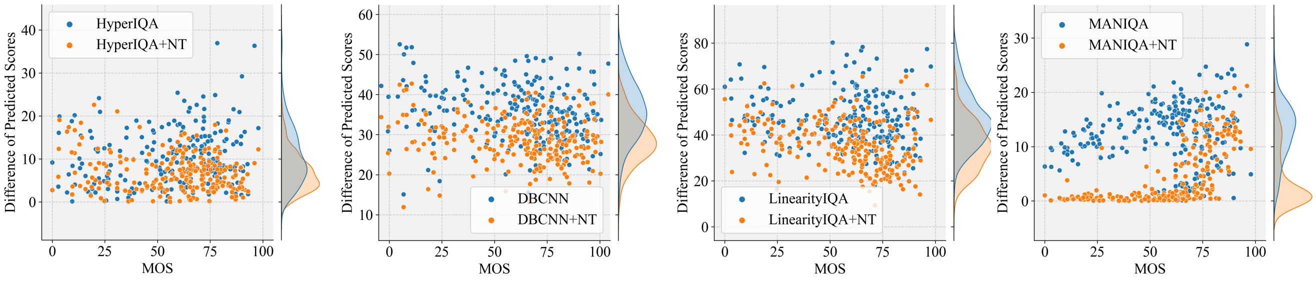

Figure S1. Comparison of four NR-IQA models with/without the NT strategy under the FGSM attack [1]. The absolute differences between predicted scores before and after attack for all test images are presented, with the fitted distribution displayed on the right side.

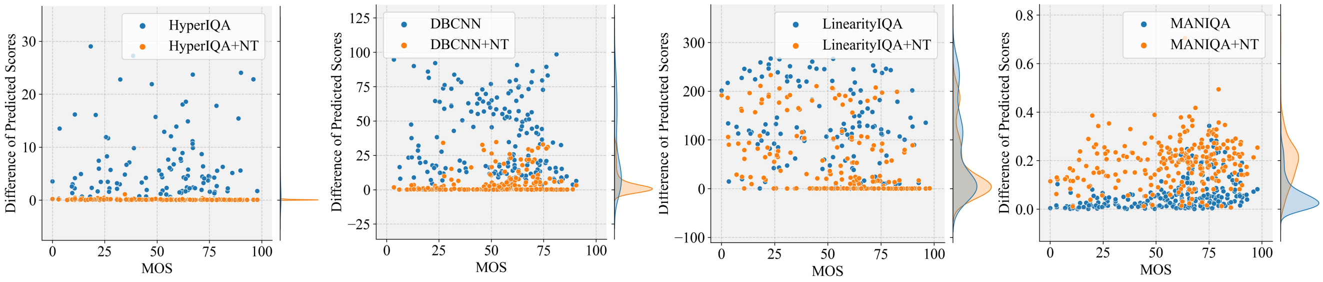

Figure S2. Comparison of four NR-IQA models with/without the NT strategy under the Perceptual attack [8]. The absolute differences between predicted scores before and after attack for all test images are presented, with the fitted distribution displayed on the right side.

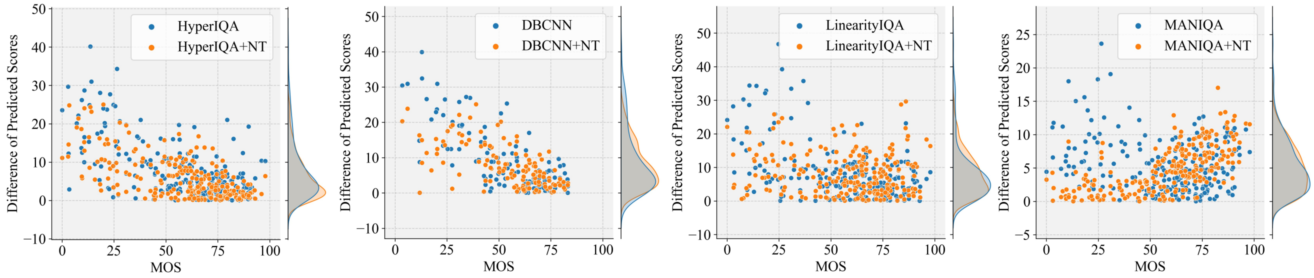

Figure S3. Comparison of four NR-IQA models with/without the NT strategy under the UAP attack [4]. The absolute differences between predicted scores before and after attack for all test images are presented, with the fitted distribution displayed on the right side.

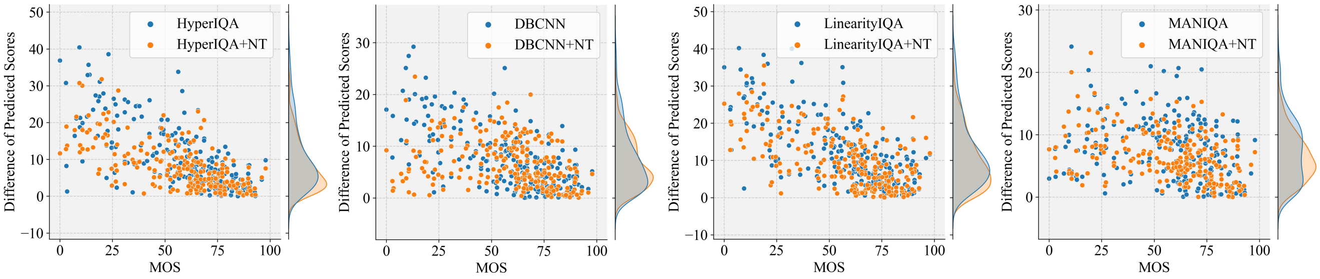

Figure S4. Comparison of four NR-IQA models with/without the NT strategy under the Kor. attack [2]. The absolute differences between predicted scores before and after attack for all test images are presented, with the fitted distribution displayed on the right side.

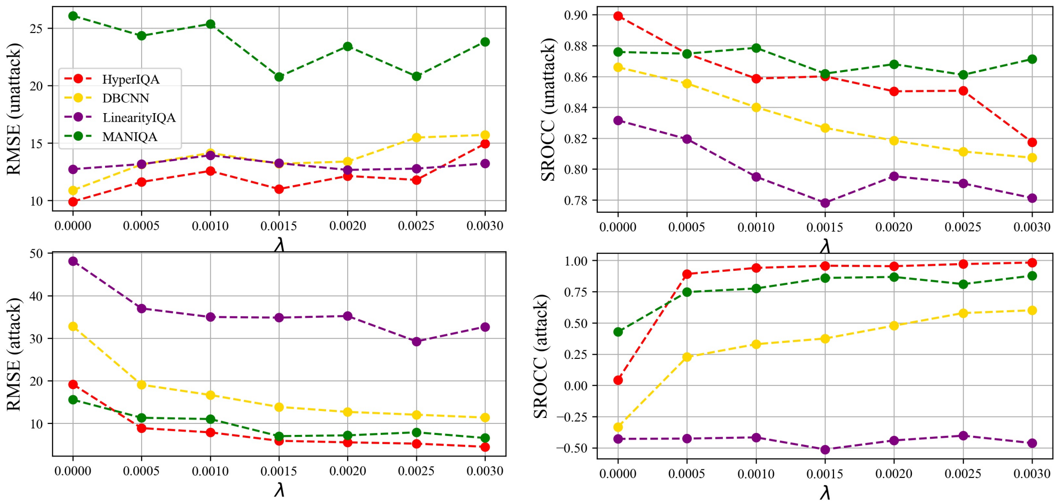

Figure S5. The impact of λ to SROCC and RMSE on both unattacked images and adversarial examples.

NR-IQA models at various image quality levels.

For example, considering the Perceptual Attack, all points for HyperIQA+NT closely align with the line "difference of predicted scores = 0". This highlights the significant effectiveness of the NT strategy in minimizing score changes with small perturbations for HyperIQA. For the DBCNN model, it is clear that the NT strategy brings about more reduction in score changes for images with MOS between $[0, 50)$ and $[75, 100]$.

Conversely, in the case of LinearityIQA, the effectiveness of the NT strategy is more obvious on high-quality images with MOS in $[75, 100]$, while it proves more effective on low-quality images with MOS in $[0, 50]$ for MANIQA. This discovery reflects that the NT strategy has varying impacts on images with different quality levels, and these impacts are closely tied to the NR-IQA models. Exploring the enhancement of adversarial robustness in NR-IQA models across different image quality levels represents a valuable avenue for research. Such investigations can shed light on the properties of NR-IQA models in predicting scores for images of differing quality.

S7. Full Results of Ablation Studies

In Figure S5, we show the full results of the ablation study of λ (mentioned in Sec. 5.5 in the main manuscript). 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, the performance of baseline+NT models has the following trend on unattacked images. SROCC values generally decrease with the rising λ, while RMSE values exhibit an upward trend (except for MANIQA). It is an interesting observation that the RMSE value of MANIQA fluctuates as λ changes, and the RMSE tends to decrease with larger λ. When attacked by the FGSM attack, the RMSE values of all baseline+NT models decrease consistently with the increase of λ. Except for LinearityIQA, the SROCC values of other models increase as λ becomes larger. This implies that increasing λ tends to enhance the robustness of NR-IQA models but leads to a performance decline on unattacked images.

S8. Visualization Results

In this section, we present visualization results to illustrate the effectiveness of the NT strategy under FGSM attack with different attack intensities and UAP. Under FGSM attack with different attack intensities, for each pair of baseline and baseline+NT models, we provide two sets of visualization results: one for high-quality images and the other for low-quality images. We show the normalized MOS of the original image. Under the UAP attack, we provide one adversarial sample for an NR-IQA model. We display adversarial examples for both the baseline model and the baseline+NT model, along with the corresponding score changes (predicted score before attack $\rightarrow$ predicted score after attack).

Figure S6 shows visualization results of FGSM attack for HyperIQA, DBCNN, and their NT versions. Figure S7 displays visualization results of FGSM attack for LinearityIQA, MANIQA, and their NT versions. Figure S8 shows visualization results of UAP attack for HyperIQA, DBCNN, LinearityIQA, MANIQA, and their NT versions.

From Figure S6 and Figure S7, we can see that the imperceptibility of adversarial perturbations for images gets worse as the attack intensity increases. However, de-

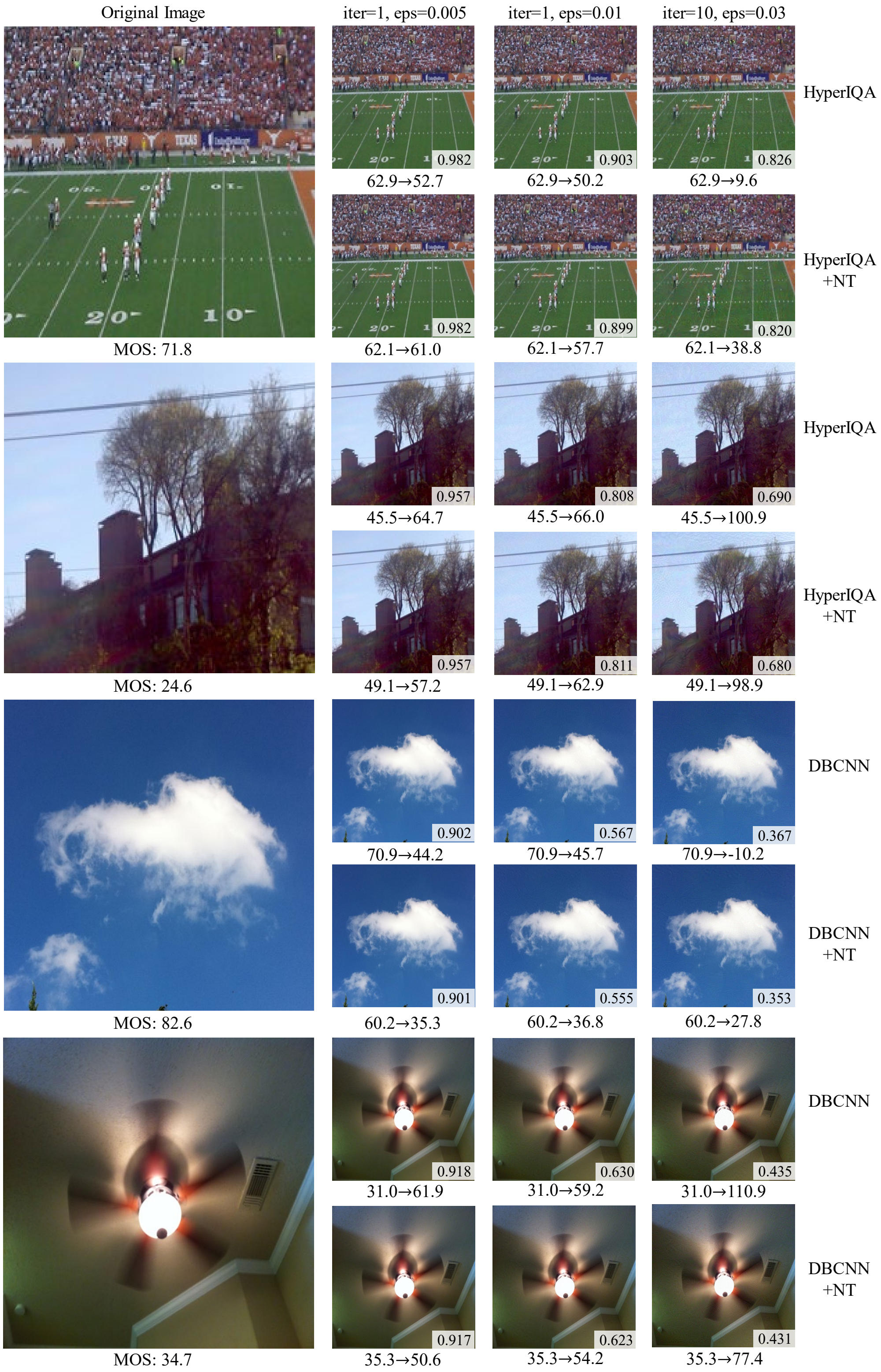

Figure S6. (Zoom in for a better view) Visualization results of adversarial examples generated using the FGSM with different intensities. The normalized MOS is presented. FGSM attack settings are indicated above the figures, and each adversarial example for a model is presented with the format: "predicted score before attack → predicted score after attack" below the respective images. The SSIM between the adversarial example and the original image is displayed at the bottom right corner of each adversarial image.

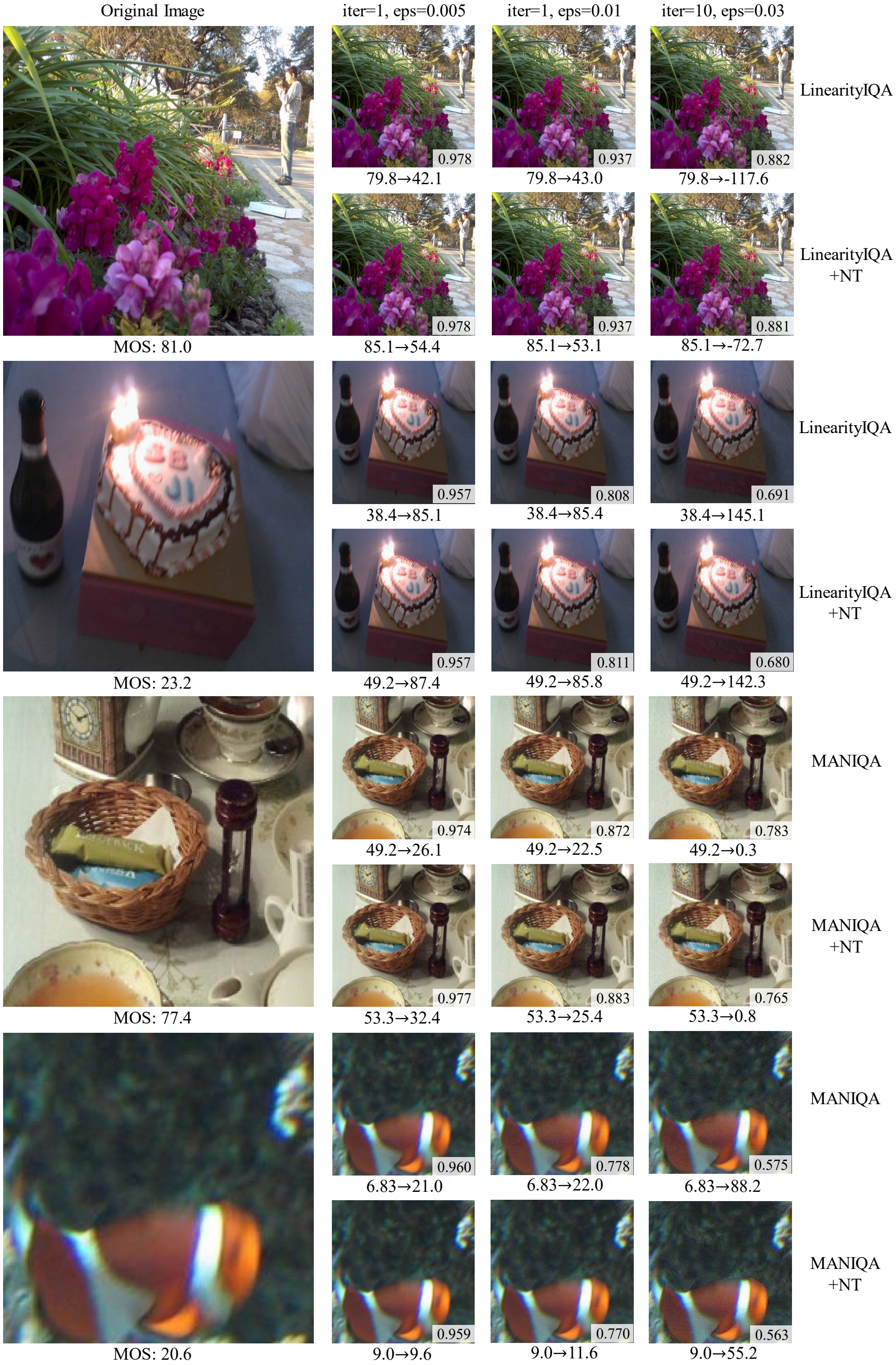

Figure S7. (Zoom in for a better view) Visualization results of adversarial examples generated using the FGSM with different intensities. The normalized MOS is presented. FGSM attack settings are indicated above the figures, and each adversarial example for a model is presented with the format: “predicted score before attack → predicted score after attack” below the respective images. The SSIM between the adversarial example and the original image is displayed at the bottom right corner of each adversarial image.

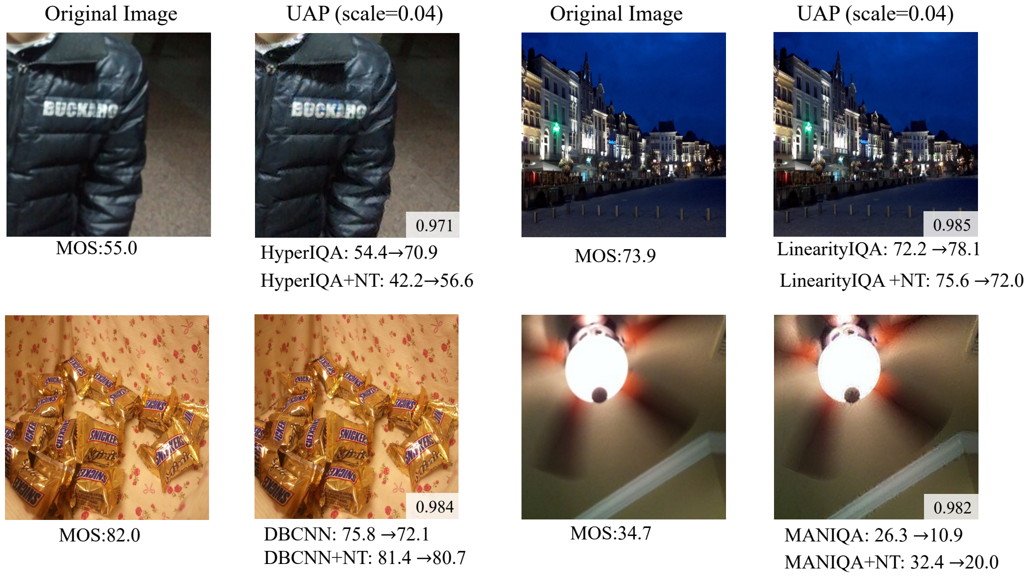

Figure S8. (Zoom in for a better view) Visualization results of adversarial examples generated using the UAP attack. The normalized MOS is presented. Each adversarial example for a model is presented with the format: "predicted score before attack → predicted score after attack" below the respective images. The SSIM between the adversarial example and the original image is displayed at the bottom right corner of each adversarial image.

spite this, the NT models consistently exhibit smaller score changes than the baseline models in most cases. Consider HyperIQA and its NT version as an example. When attacked by the strongest FGSM attack (iter=10, eps=0.01), the score change for HyperIQA+NT on the high-quality image is $62.1 - 38.8 = 23.3$, whereas the score change for HyperIQA is $62.9 - 9.6 = 53.3$. Similarly, for the low-quality image, the score change for the NT version is 49.8, while the change for the baseline model is 55.4. From Figure S8, we can see that with the same adversarial sample, baseline and their NT versions have different defense performances. For example, the score change for HyperIQA+NT on its adversarial sample is $56.6 - 42.2 = 14.4$, whereas HyperIQA on the same adversarial sample is $70.9 - 54.4 = 16.5$.

References

[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, pages 1–11, 2015. 2, 3, 5

[2] 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. 3, 5

[3] 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. 2

[4] 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. 3, 5

[5] 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. 2

[6] 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. 2

[7] 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. 2

[8] 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, 3, 5