Towards Adversarial Robustness in Blind Image Quality Assessment with Soft Thresholding Norm

Desen Yuan[1] and Lei Wang[2]

Abstract— **In this study, we address the issue of adversarial robustness within the context of Blind Image Quality Assessment (BIQA), an area of heightened importance due to the inherent susceptibility of Deep Neural Networks (DNNs) to adversarial assaults. Current approaches primarily rely on adversarial training, which, despite its efficacy, imposes a significant computational burden. Our research proposes an alternative strategy known as the Soft Thresholding Norm (ST Norm). This approach counters the 'feature shift' phenomenon, identified by a substantial Euclidean Distance Statistics (EDS) between original and adversarial features, through the imposition of sparse constraints on potential features following batch normalization. This novel method offers several advantages: it reduces the Lipschitz constant yielding smoother models, seamlessly integrates with existing models, and boasts inherent denoising capabilities, thereby effectively mitigating the impact of adversarial perturbations. Results suggest that our approach achieves robustness comparable to adversarial training but with significantly less computational overhead. Moreover, it consistently outperforms other adversarial defense strategies on BIQA datasets, highlighting its practical effectiveness in enhancing adversarial robustness. This study underscores the potential of the Soft Thresholding Norm within the realm of IQA tasks, positioning it as a resource-efficient alternative to traditional adversarial training methodologies.**

I. INTRODUCTION

The task of blind image quality assessment aims to enable computers to automatically evaluate the quality of images, imbuing them with human-like image perception capabilities. However, current neural networks are highly susceptible to adversarial attacks, where the model can be deceived into producing incorrect predictions after small perturbations are added to the samples. This poor robustness against minor disturbances severely hampers the practical application of image quality assessment. Furthermore, this lack of robustness to tiny disturbances exposes a significant deficiency in the current image quality assessment models in simulating human perception, as human perception has the Just Noticeable Distortion (JND) effect [36]. The JND model simulates the threshold level at which the human eye can detect distortions in the underlying features of an image. Only when disturbances exceed a certain threshold can they be perceived.

The JND characteristics inherent in image quality assessment tasks dictate that the upper limit of perturbations cannot be too large, as this would objectively lead to a decrease in image quality [34]. It can be found in Figure 1 that the real MOS is confusing after adding too much perturbation.

Unlike IQA tasks, researchers often subject the model to significant perturbations in classification tasks, and mainstream adversarial defense research focuses on experimental verification of methods in the field of image classification, where robustness against perturbations is essential.

To improve the adversarial robustness of DNNs, researchers have made significant efforts, among which adversarial training is the most effective and intuitive defense method. However, the additional computational cost that adversarial training incurs during model training is enormous [28], [13], severely slowing down the training speed of the model. Subsequent research has continued to enhance the effectiveness of adversarial training by introducing more challenging adversarial examples, increasing network margins, optimizing regularized surrogate loss functions, and other methods.

The Lipschitz constant of constrained models is also a method to improve model robustness [15], [2]. The constant characterizes the variation in model output following input perturbations. Without constraints on the Lipschitz constant, as the network deepens, the Lipschitz constant increases with the depth of layers and dimensionality, leading to a decrease in network robustness. Current studies [33] have imposed a constraint of 1 on the Lipschitz constant for the models. Although this enhances model robustness in these circumstances, such global constraints are excessively strong, resulting in a significant decline in the model's expressive capacity.

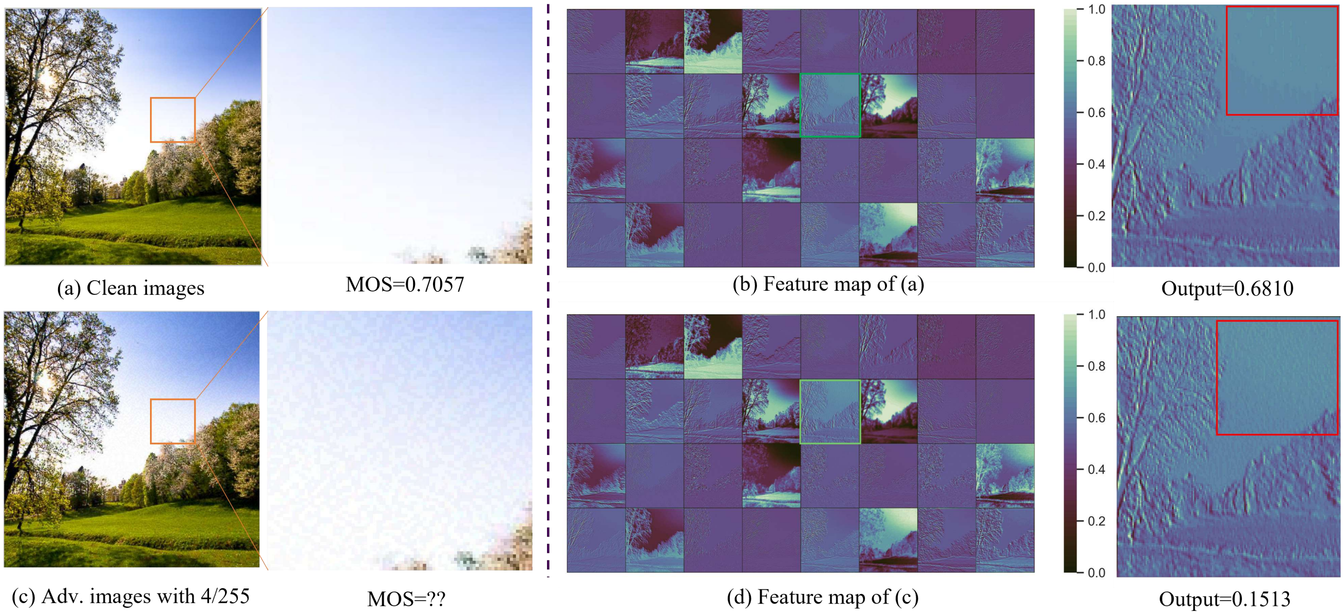
Given the differences between classification tasks and blind IQA tasks discussed in the previous paragraph, this paper attempts to propose a model defense approach for no-reference image quality assessment that does not require adversarial training. Through the changes in the feature spectrum of the adversarial examples in Figure 1, we found that the feature map of the image quality assessment task is sparse and repetitive. The changes in the feature map caused by adversarial examples are not obvious. Therefore, we further analyze the statistics of the characteristic spectrum. The existing literature [30] investigates on batch normalized features in adversarial scenarios. There are also differences in the feature distribution statistics of adversarial samples and clean samples after batch normalization on IQA task, as shown in Figure 2 (a).

We hypothesize that this difference affects the final network output values in a predictable manner. We use Eu-

*The contributions of the authors of this paper are equal.

[1]Desen Yuan is with the School of Communication and Information Engineering, University of Electronic Science and Technology of China, China desenyuan@gmail.com

[2]Lei Wang is with the School of Communication and Information Engineering, University of Electronic Science and Technology of China, China wangleiuestc@outlook.com


(a) Clean images

MOS=0.7057

(b) Feature map of (a)

Output=0.6810

(c) Adv. images with 4/255

MOS=??

(d) Feature map of (c)

Output=0.1513

Fig. 1. The characteristics of image quality assessment tasks in adversarial scenes. The image quality of clean images will change after being perturbed by adversarial scenes. There are subtle changes in the feature map after perturbation. Noise appeared in the original smooth feature map. The ?? marks after the 4/255 attack indicate that the MOS has changed. The ?? indicates that under conditions of significant perturbation, the Mean Opinion Score (MOS) becomes unpredictable. Our empirical observation shows that there will be no significant disturbance changes when the attack rates are 1/255 and 2/255.

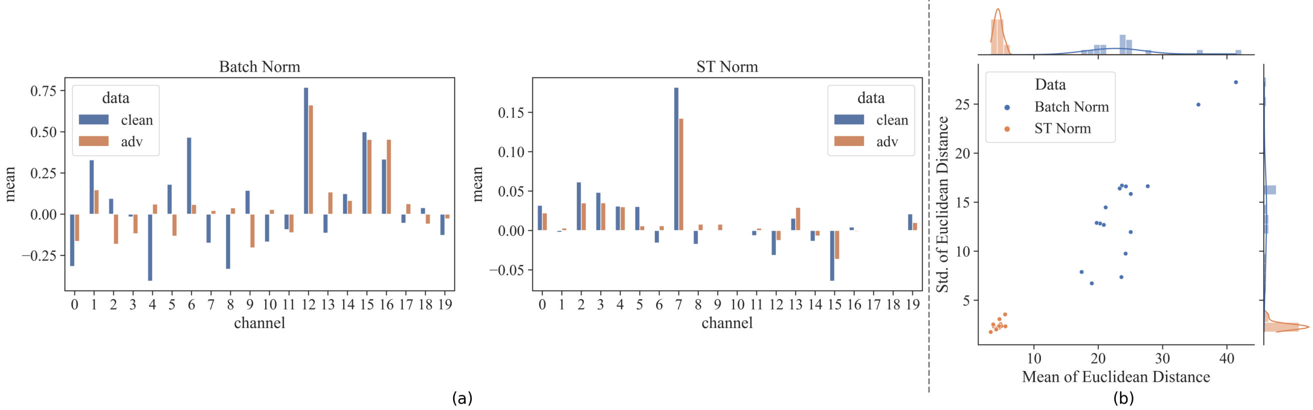


Fig. 2. (a) Statistics of the mean of Batch Norm and ST Norm on randomly sampled 20 channels in a ResNet-50's res4 block. This suggests that clean and adversarial images induce significantly different statistics. (b) EDS between randomly sampled 16 clean samples and paired adversarial samples with Batch Norm or ST Norm on a clean training ResNet-50's res4 block. This suggests that clean and adversarial images induce significantly different statistics.

clidean distance to measure the feature shift value of batch normalized features relative to the feature vector of clean samples. Specifically, we calculate the Euclidean distance in the channel domain, and then calculate the mean and standard deviation of the values obtained in the spatial domain as the final EDS. We hypothesized that the smaller EDS, the more robust the network. To achieve this, we add a post-processing to the features after batch norm to alleviate the feature shift. See Figure 2 (b), our method significantly reduces the feature shift statistics. Specifically, we add sparse constraints to the features after batch norm to alleviate the dimensionality disaster problem [1] in the feature offset. By solving the sparse constraint, we finally get the form of soft thresholding normalization. In addition, the soft thresholding normalization can provide the network with a local 1-Lipschitz constant constraint. The created band-flat area enables the model to have numerical quantification characteristics against input transformations, thereby moderating the output fluctuations caused by input perturbations.

Our experimental results demonstrate that the proposed Soft Thresholding Norm method can achieve results comparable to adversarial training methods without employing adversarial training strategies, and it requires minimal computational resources. Compared to other adversarial defense methods in the image recognition domain, our proposed approach has consistently outperformed on relevant image quality assessment datasets. For instance, on the LIVE dataset, utilizing ResNet as the backbone architecture, the model accuracy under adversarial attacks for the proposed method has seen an enhancement in SROCC from 0.1597 to 0.7904. Concurrently, the accuracy of the adversarial training approach stands at 0.7279, requiring a significant amount of computational resources. Additionally, we have also explored the limitations of the Soft Thresholding Norm method and potential directions for future improvements.

II. Realted Works

A. Adversarial Robustness in Classification

In an effort to counteract the erroneous guidance provided by adversarial examples to models, much work has been dedicated to enhancing the adversarial robustness of models, with existing studies primarily focused on the field of image classification. Adversarial training [6] is one of the most accessible and widely used methods to improve model robustness. However, relevant studies have shown that while this approach significantly increases additional computational costs, it sacrifices the accuracy for clean inputs and only provides empirical robustness [28]. Additionally, from a certified robustness perspective, research has mainly focused on convex relaxation [25], random smoothing [4], and Lipschitz constant constraints [15], [2]. Convex relaxation methods are often complex, carry high computational costs, and pose challenges when applied to deeper neural networks and large-scale models. Random smoothing can provide (probabilistic) certified robustness guarantees for general models. Research indicates that the robustness radius depends on input dimension and struggles with larger infinity perturbations.

While the Lipschitz constant can bolster model robustness, it might degrade performance for clean inputs. Although some suggest altering neural networks or introducing Lipschitz networks to address this, these diverge significantly from traditional neural network structures, impacting their versatility [32]. Therefore, we use flat transformations between network layers to impose local Lipschitz constraints, harmonizing with existing networks and ensuring robustness.

B. Blind Image Quality Assessment

Prior to the rise of deep learning [19], [20], [18], [27], [26], the field of Blind Image Quality Assessment (BIQA) was dominated by the theory of Natural Scene Statistics (NSS) [17]. Recently, deep learning-driven Blind Image Quality Assessment (BIQA) [11], [35], [10], [31], [21], [22], [24], [23] methods have emerged, notably enhancing model performance. CNN-based frameworks, especially using ResNet, effectively merge feature learning and regression. There's also a surge in large-scale IQA datasets. Methods often use image patches or features from different CNN layers for multi-scale representation.

III. Methods

A. Background

Given datasets $D = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in X$ and $y_i \in Y \in \mathbb{R}$ denote images and its corrsponding MOSs. IQA models which is parameterized by DNN can be defined as a function $f_\theta(\cdot) : X \to \mathbb{R}$.

Adversarial robustness is the model's performance on test data with adversarial attacks. Adversarial attacks are intentional perturbations of the model's inputs that aim to fool the model into making incorrect predictions or results. Adversarial robustness can be expressed by the following formula:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{B}} L(f_\theta(x+\delta), y) \right], \tag{1}$$

where L is the MSE loss $L(f_\theta(x), y) = \|f_\theta(x) - y\|_2^2$, θ are the model's parameters, $\mathcal{D}$ is the data distribution, f_θ is the model's prediction function, and perturbation $\delta \in$ collection of neighborhoods $\mathcal{B}$ which are ℓ_∞-norm bounded. From Eq. 1, we find that the robustness of the model examines not only the accuracy of the sample point but also the accuracy of the neighborhood around the sample point.

Generally speaking, the input space of the image $x_i \in \mathbb{R}^{C\times H\times W}$ is huge relative to the number of samples n, and vanilla deep IQA model training will not have enough supervised training on the input space. The training objective function of an IQA task can be expressed as:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[L(f_\theta(x), y) \right]. \tag{2}$$

It can be observed that the sample points used in our supervised training included only contain $(x, y) \sim \mathcal{D}$. However, the input space of the image $x_i \in \mathbb{R}^{C\times H\times W}$ is huge relative to the number of samples n.

Therefore, vanilla deep IQA model training will not have sufficient supervised training for the input space. There are numerous gaps where supervised training has not been conducted. The existing literature proves that the output value of the IQA model will vary greatly in a small neighborhood of ℓ_∞-norm around the input of vanilla supervised training [34]. Examples that cause drastic changes in output are called adversarial samples, and the most traditional method of obtaining adversarial examples is FGSM:

$$\boldsymbol{\eta} = \epsilon \operatorname{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right), \tag{3}$$

where θ be the parameters of a model, and $J(\theta, x, y)$ be the cost used to train the neural network. We can linearize the cost function around the current value of θ, obtaining an optimal max-norm constrained pertubation $\boldsymbol{\eta}$. Naturally, supervised training of the voids around normal by taking adversarial samples is called adversarial training. Specifically, a regular term related to adversarial samples is added to the training loss:

$$\tilde{J}(\boldsymbol{\theta}, \boldsymbol{x}, y) = \alpha J(\boldsymbol{\theta}, \boldsymbol{x}, y) + (1-\alpha) J\left(\boldsymbol{\theta}, \boldsymbol{x}+\boldsymbol{\eta}, y\right). \tag{4}$$

However, adversarial training has a huge training cost, because each update must be backpropagated to obtain the adversarial samples corresponding to normal samples, not only that, different attack methods will obtain different adversarial samples. Compared to the huge space of infinite norm spheres around normal samples, adversarial samples have only one point for supervised training. Therefore, some literature suggests that models trained on adversarial samples can only gain empirical robustness and are only effective against specific attacks [7].

B. Soft Thresholding Norm

Unlike classification tasks, which rely heavily on adversarial training for robustness models for large-scale Lbounded neighborhoods, IQA tasks typically only examine the robustness of small-scale bounded neighborhoods [34]. The robustness goal of the IQA task is easier to achieve than the classification task. So we ask the question: Is it possible to obtain robust IQA models without relying on adversarial training? The problem we need to solve is: without adversarial training, there is no additional supervision target, how to ensure that there will be no drastic changes in a small area of the neighborhood?

Literature indicates [30], [1] that adversarial samples and normal samples exhibit differences in their feature distributions. This discrepancy can be discerned from the variations in the mean and variance within the Batch Normalization (BN) layer. Our approach is inspired by the distributional differences between adversarial and normal samples. Such differences lead to biases in the layer-by-layer outputs of the network, resulting in a significant discrepancy between the outputs of adversarial samples and those of original samples, thereby deceiving the model.

Our paper calls this feature shift phenomenon. As shown in Figure 2, we find that we can measure the offset of features in terms of Euclidean distance $\|(x+\delta)-x\|_2$. And the Euclidean distance between the confrontational features and the original features is very large. The conventional dot product operation of existing networks is based on Euclidean distance, resulting in drastic changes in output under perturbation [1].

Based on the aforementioned observations, we introduce additional sparse constraints to the potential features after batch normalization. This alleviates the phenomenon of feature shift, thereby further enhancing the model's robustness against adversarial samples.

Specifically, for a bounded signal, if the signal is sparse, then it will have a smaller Euclidean distance. The formulaic formulation is:

$$\|\delta\|_2 > \|S(x+\delta) - S(x)\|_2, \tag{5}$$

where S represents the post-processing function, which makes the 0 norm of the input larger, but minimizes the difference between the input and output.

$$\min_S \|S(x) - x\|_2 \quad \text{s.t.} \max_S \|S(x)\|_0 \tag{6}$$

Suppose that S can implement the function of thinning, such that the channel C_s is 0 and the remaining channel C_r values are unchanged, that is : $S(x_c) = x_c, S(x_c+\delta) = x_c + \delta \quad \text{s.t.} \quad c \in C_r$ and $S(x_c) = S(x_c+\delta) = 0 \quad \text{s.t.} \quad c \in C_s$. Then the Euclidean distance can be expressed as:

$$\begin{aligned}
&\|S(x+\delta) - S(x)\|_2 \\
&= \sqrt{\sum_{c\in C_r}(S(x_c+\delta) - S(x_c))^2 + \sum_{c\in C_s}(S(x_c+\delta) - S(x_c))^2} \\
&= \sqrt{\sum_{c\in C_r}(x_c+\delta-x_c)^2 + \sum_{c\in C_s}(0-0)^2} \\
&< \sqrt{\sum_{c\in C}(x_c+\delta-x_c)^2} = \|\delta\|_2.
\end{aligned} \tag{7}$$

Relax the 0-norm constraint to a 1-norm constraint and use the Lagrange multiplier method to transform the optimization objective of S into:

$$\arg\min_S \|S(x) - x\|_2 + \lambda\|S(x)\|_1. \tag{8}$$

To solve the appeal objective function, we get the soft thresholding form, which serves as post-processing of BatchNorm:

$$S(x) = \operatorname{sign}(x)\max\left(|x| - \lambda, 0\right). \tag{9}$$

To make the threshold setting more reasonable, inspired by the robust work that normalization contributes to, and layer normalization proved to be useful for model training, we added layer normalization so that the data were distributed in the range of -1 to 1 before soft thresholding processing. The threshold value of our empirical value is 0.2. We propose a soft thresholding norm (ST Norm) that is merged with BatchNorm as a new robust norm module.

$$S(x) = \operatorname{sign}\left(\frac{x}{\|x\|_\infty}\right)\max\left(|\frac{x}{\|x\|_\infty}| - \lambda, 0\right), \tag{10}$$

where $\mathbf{x}$ is the output of BatchNorm.

C. Properties of ST Norm

Soft thresholding possess two properties that benefit robust models. First, soft thresholding allow the model to have smaller lip constants. The Lipschitz constant measures how much the network output changes with respect to the input. It is defined as

$$Lip(f) = \sup_{\delta\neq 0}\frac{\|f(x) - f(x+\delta)\|_p}{\|\delta\|_p}. \tag{11}$$

A lower Lipschitz constant implies a smoother and more robust network, since it means that small perturbations in the input will not cause large changes in the output. Since the soft thresholding function can be divided into intra-threshold and out-of-threshold functions, where the Lipschitz constant of out-of-threshold is equal to 1, and Lipschitz constant outside of the intra-threshold is equal to 0. That means

$$Lip(f) \geq Lip(S \circ f). \tag{12}$$

Second, ST Norm can then guarantee that the training is stable.

ST Norm can be plugged directly into existing models without losing too much performance. Because the soft thresholding function is an identity map when the threshold is equal to 0 :

$$S(x, \lambda) = x \quad \text{s.t.} \quad \lambda = 0. \tag{13}$$

We can empirically incrementally increase the threshold to get a suitable model.

IV. Experiments

A. Experimental setting

In this study, all comparisons of Blind Image Quality Assessment (BIQA) methods and ablation studies adhere to this setting.

1) Datasets.: We tested the proposed method on 3 prominent Image Quality Assessment (IQA) datasets, including the synthetic distortion dataset LIVE [16]. We also employed two real-world distortion datasets: LIVEC [5] and Koniq-10k [9].

2) Implementation Details.: The model's batch size was set to 16, with an initial learning rate of 1.0e-4, utilizing an ADAM optimizer with a weight decay of 0.01. We selected ResNet-50 as the network for our experiments. In the experiments, the adversarial examples for each method were generated individually using either FGSM [6] or PGD [12] targeted at each respective method. In the following experiment's tables, "-C" indicates the results under clean, original data; "-A" represents the results of the corresponding model when under attack; "-AT" denotes the outcomes of adversarial training [6]; and "-A-STE" refers to the results when using STE combined with an attack.

B. Performance of ST Norm

As the results in Table I, we investigate the robustness of the proposed ST Norm method across multiple datasets, specifically, on ResNet-50 [8] and LIVE and LIVEC datasets.

The comparison methods include, LWTA [14], and Denoise [29] - these represent prevalent strategies employed in the literature for improving the adversarial robustness of neural networks. Our results indicate that across all considered scenarios, the adversarial robustness of ST Norm is either superior to or offers comparable performance. ST Norm provides similar performance without the need for extensive computational resources. For clean sample inputs, the performance of ST Norm slightly decreases; however, the performance gap between clean inputs and under-attack conditions is minimal, underscoring the robustness of the proposed method. Existing research suggests that adversarial defense techniques can sometimes compromise the performance on clean inputs [32], [33].

C. Ablation Study

1) λ ablation: As the results in Table II, we applied ST Norm to the ResNet model and investigated the hyperparameter lambda's impact. Empirical data shows peak robustness at lambda=0.2, consistent across datasets, suggesting it as a potential starting point for other models.

TABLE I
The SRCC, PLCC, and MSE results of the ResNet methods under 1/255 attack rates with clean training and adversarial training. Adversarial samples of Adversarial training are all got by FGSM with 1/255 attack rate.

Datasets	LIVE SROCC	 PLCC	 MSE	LIVEC SROCC	 PLCC	 MSE
ResNet-C	0.9382	0.8712	0.0272	0.8237	0.8609	0.0118
ResNet-A	0.1597	0.2141	0.0888	-0.0010	-0.0200	0.0611
LWTA-C	0.3735	0.2268	0.1736	0.4776	0.4965	0.0816
LWTA-A	0.3878	0.2300	0.1733	0.4948	0.5227	0.0814
Denoise-C	0.9071	0.8797	0.0320	0.8194	0.8550	0.0126
Denoise-A	0.1141	0.1460	0.0900	-0.0395	-0.0633	0.0639
ST Norm-C	0.8547	0.7105	0.0401	0.6867	0.7018	0.0218
ST Norm-A	**0.7904**	**0.6346**	**0.0462**	**0.6435**	**0.6607**	**0.0242**

2) STE or not: Recent research [3] has revealed that many adversarial defense techniques, such as defensive distillation, operate primarily due to gradient masking. Methods reliant on gradient masking are essentially deceptive. When attackers are aware of the defense mechanisms in place, they can effortlessly bypass them, rendering the deployed defenses ineffective against adversarial samples. To verify whether our proposed method, ST Norm, exhibits gradient masking phenomena, we referenced the backward differentiable approximation [3] and specifically employed the Straight Through Estimator (STE) for attack evaluation. As shown in Table III, following the application of STE-based attacks, the adversarial robustness of ST Norm did not deteriorate. This confirms that the efficacy of our proposed method does not stem from gradient masking but rather from its inherent robustness.

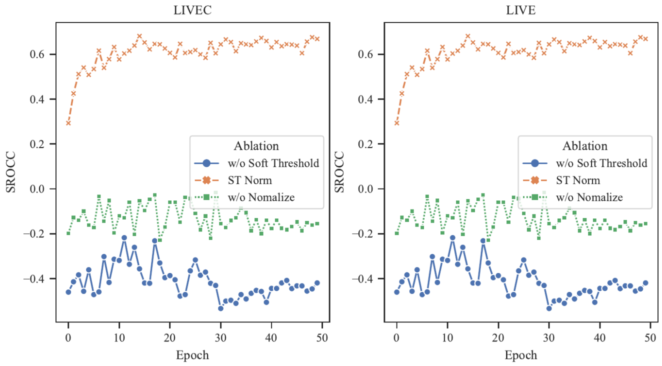


Fig. 3. Ablation study of the ST Norm on LIVE and LIVEC datasets with attack rate 1/255 by FGSM.

3) l_∞ norm & Soft Thresholding or not: To conduct an ablation study on the infinity norm prior to soft thresholding and on the soft thresholding itself, we experimented with ST Norm on the LIVE and LIVEC datasets, assessing configurations both with and without the infinity norm or soft thresholding. As depicted in Figure 3, the results of soft thresholding without norm normalization significantly underperform compared to those after norm normalization followed by soft thresholding. This is consistent with the

TABLE II

λ ABLATION ON LIVE DATASET WITH ATTACK RATE 1/255 BY FGSM.

λ	clean			1/255			2/255		
	SRCC	PLCC	MSE	SRCC	PLCC	MSE	SRCC	PLCC	MSE
0.1	0.8225	**0.7533**	0.0484	0.5914	0.5355	0.0677	0.6019	0.5713	0.0613
0.2	**0.8547**	0.7105	**0.0401**	**0.7904**	**0.6346**	0.0462	**0.6956**	**0.6531**	**0.0415**
0.3	0.5164	0.4930	0.0458	0.5035	0.4872	**0.0448**	0.5244	0.5047	0.0424

TABLE III

ATTACK USE STE OR NOT ON LIVE, LIVEC AND KONIQ DATASETS WITH ATTACK RATE 1/255 BY FGSM.

Datasets	LIVE			LIVEC			KONIQ		
	SROCC	PLCC	MSE	SROCC	PLCC	MSE	SROCC	PLCC	MSE
ST Norm-A	0.7904	0.6346	0.0462	0.6435	0.6607	0.0242	0.6661	0.6925	0.0101
ST Norm-A-STE	0.8009	0.6504	0.0428	0.6758	0.6917	0.0228	0.6888	0.7135	0.0095

TABLE IV

CROSS DATASET EVALUATION OF ST NORM TRAINED ON THE LIVE DATASET WITH ATTACK RATE 1/255 BY FGSM.

Datasets	LIVEC		KONIQ	
	SROCC	PLCC	SROCC	PLCC
ResNet-A	-0.5582	-0.5004	-0.3597	-0.3318
LWTA-A	0.0161	-0.0053	0.0417	0.0146
Denoise-A	-0.5582	-0.5004	-0.3874	-0.3684
ResNet-AT-A	-0.2553	-0.2161	0.0373	0.0279
ST Norm-A	**0.2210**	**0.2338**	**0.3487**	**0.3352**

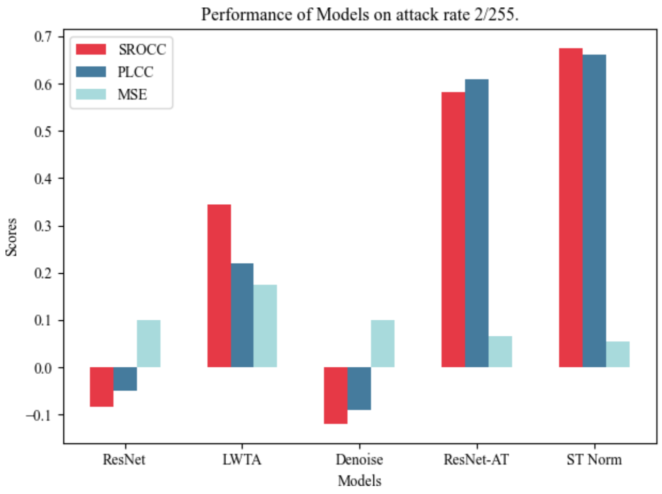


Fig. 4. Performance of ResNet, LWTA, Denoise, ResNet-AT, and ST Norm under attack rate 2/255 on LIVE dataset by FGSM.

rationale behind our proposed method, as normalizing with the infinity norm allows for a more appropriate selection of the threshold value for soft thresholding, whereas neglecting to normalize may result in values not falling within the desired threshold range. Moreover, the results without soft thresholding are also suboptimal, since mere normalization cannot effectively enforce sparsity constraints or generate flat regions.

4) Different attack rates: To evaluate the proposed method's efficacy under different attack rates, we tested it on the LIVE dataset at a 2/255 attack rate and contrasted it with benchmark techniques. Figure 4 shows that ST Norm outperforms benchmarks in adversarial robustness, especially under heightened attack rates. Compared to resource-intensive adversarial training, ST Norm excels in the SRCC and PLCC metrics and records a lower MSE.

5) Cross dataset evalution: As shown in Table IV, to validate the generalizability of the proposed ST Norm, we trained on the LIVE dataset and tested on the KONIQ and LIVEC datasets.

6) Different attack methods: To assess the effectiveness of the proposed method, ST Norm, against various attack techniques, we conducted cross-dataset evaluations. Our trained model on the LIVEC dataset was tested on the LIVE dataset. The PGD attack method was employed with an attack rate of 2/255. The experimental results, as depicted in Figure 5, reveal that the proposed ST Norm method demonstrates robust performance against PGD attacks even across different datasets, outperforming alternative approaches.

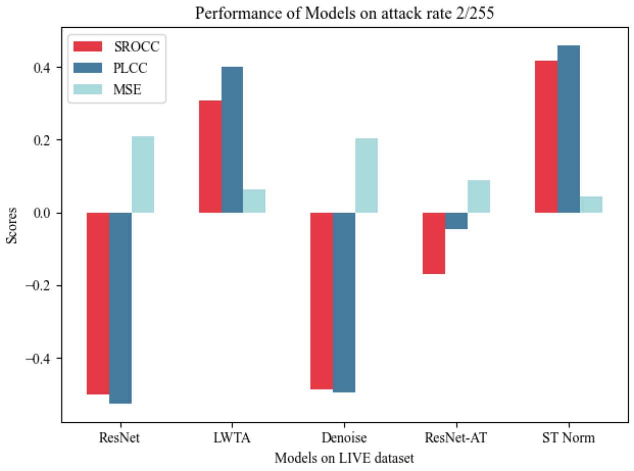


Fig. 5. Models trained on the LIVEC dataset were tested on the LIVE dataset with an attack rate of 2/255 using the PGD attack method.

V. Conclusion

In this paper, we assess the vulnerability of image quality assessment models under adversarial attacks and identify evidence of feature displacement in image quality evaluation. To address this, we propose a simple yet effective method named Soft Thresholding Norm to enhance the model's adversarial robustness, requiring only minimal additional computational cost. Compared to the adversarial training approaches, the proposed method consumes significantly less computation yet achieves comparable adversarial robustness. Moving forward, we aim to explore ways to boost adversarial resilience without compromising the performance on clean inputs.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 977–988. IEEE, 2022.

[2] Cem Anil, James Lucas, and Roger Grosse. Sorting out lipschitz function approximation. In *International Conference on Machine Learning*, pages 291–301. PMLR, 2019.

[3] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

[4] Avrim Blum, Travis Dick, Naren Manoj, and Hongyang Zhang. Random smoothing might be unable to certify l robustness for high-dimensional images. *The Journal of Machine Learning Research*, 21(1):8726–8746, 2020.

[5] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[7] Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4842–4851, 2019.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[10] Kede Ma, Huan Fu, Tongliang Liu, Zhou Wang, and Dacheng Tao. Deep blur mapping: Exploiting high-level semantics by deep neural networks. *IEEE Transactions on Image Processing*, 27(10):5155–5166, 2018.

[11] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2017.

[12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[13] Preetum Nakkiran. Adversarial robustness may be at odds with simplicity. *arXiv preprint arXiv:1901.00532*, 2019.

[14] Konstantinos Panousis, Sotirios Chatzis, Antonios Alexos, and Sergios Theodoridis. Local competition and stochasticity for adversarial robustness in deep learning. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3862–3870. PMLR, 13–15 Apr 2021.

[15] Haifeng Qian and Mark N Wegman. L2-nonexpansive neural networks. In *International Conference on Learning Representations*. International Conference on Learning Representations, ICLR, 2019.

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

[17] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001. PMID: 11520932.

[18] Cong Wang, Jinshan Pan, Wanyu Lin, Jiangxin Dong, Wei Wang, and Xiao-Ming Wu. Selfpromer: Self-prompt dehazing transformers with depth-consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5327–5335, 2024.

[19] Cong Wang, Jinshan Pan, Wei Wang, Jiangxin Dong, Mengzhu Wang, Yakun Ju, and Junyang Chen. Promptrestorer: A prompting image restoration method with degradation perception. *Advances in Neural Information Processing Systems*, 36:8898–8912, 2023.

[20] Cong Wang, Jinshan Pan, Wei Wang, Gang Fu, Siyuan Liang, Mengzhu Wang, Xiao-Ming Wu, and Jun Liu. Correlation matching transformation transformers for uhd image restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5336–5344, 2024.

[21] Lei Wang, Qingbo Wu, Desen Yuan, King Ngi Ngan, Hongliang Li, Fanman Meng, and Linfeng Xu. Learning with noisy low-cost mos for image quality assessment via dual-bias calibration. *arXiv preprint arXiv:2311.15846*, 2023.

[22] Lei Wang and Desen Yuan. Beyond mos: Subjective image quality score preprocessing method based on perceptual similarity. *arXiv preprint arXiv:2404.19666*, 2024.

[23] Lei Wang and Desen Yuan. Causal perception inspired representation learning for trustworthy image quality assessment. *arXiv preprint arXiv:2404.19567*, 2024.

[24] Lei Wang and Desen Yuan. Perceptual constancy constrained single opinion score calibration for image quality assessment. *arXiv preprint arXiv:2404.19595*, 2024.

[25] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. *Advances in Neural Information Processing Systems*, 31, 2018.

[26] Wangyu Wu, Tianhong Dai, Xiaowei Huang, Fei Ma, and Jimin Xiao. Top-k pooling with patch contrastive learning for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2310.09828*, 2023.

[27] Wangyu Wu, Tianhong Dai, Xiaowei Huang, Fei Ma, and Jimin Xiao. Image augmentation with controlled diffusion for weakly-supervised semantic segmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6175–6179. IEEE, 2024.

[28] Cihang Xie, Mingxing Tan, Boqing Gong, Alan Yuille, and Quoc V Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020.

[29] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 501–509, 2019.

[30] Cihang Xie and Alan Yuille. Intriguing properties of adversarial training at scale. *arXiv preprint arXiv:1906.03787*, 2019.

[31] Desen Yuan. Balancing easy and hard distortions: A multi-rate knowledge distillation strategy for blind image quality assessment. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8265–8269. IEEE, 2024.

[32] Bohang Zhang, Tianle Cai, Zhou Lu, Di He, and Liwei Wang. Towards certifying l-infinity robustness using neural networks with l-inf-dist neurons. In *International Conference on Machine Learning*, pages 12368–12379. PMLR, 2021.

[33] Bohang Zhang, Du Jiang, Di He, and Liwei Wang. Rethinking lipschitz neural networks and certified robustness: A boolean function perspective. *Advances in neural information processing systems*, 35:19398–19413, 2022.

[34] 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. *Advances in Neural Information Processing Systems*, 35:2916–2929, 2022.

[35] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018.

[36] XH Zhang, WS Lin, and Ping Xue. Improved estimation for just-noticeable visual distortion. *Signal Processing*, 85(4):795–808, 2005.