

FS-IQA: Certified Feature Smoothing for Robust Image Quality Assessment

Ekaterina Shumitskaya ^{1,2,3}, **Dmitriy Vatolin** ^{1,2,3}, **Anastasia Antsiferova** ^{2,1,4}

¹ ISP RAS ² MSU AI Institute ³ Lomonosov MSU ⁴ Innopolis University

Abstract

We propose a novel certified defense method for Image Quality Assessment (IQA) models based on randomized smoothing with noise applied in the feature space rather than the input space. Unlike prior approaches that inject Gaussian noise directly into input images, often degrading visual quality, our method preserves image fidelity while providing robustness guarantees. To formally connect noise levels in the feature space with corresponding input-space perturbations, we analyze the maximum singular value of the backbone network’s Jacobian. Our approach supports both full-reference (FR) and no-reference (NR) IQA models without requiring any architectural modifications, suitable for various scenarios. It is also computationally efficient, requiring a single backbone forward pass per image. Compared to previous methods, it reduces inference time by **99.5%** without certification and by **20.6%** when certification is applied. We validate our method with extensive experiments on two benchmark datasets, involving six widely-used FR and NR IQA models and comparisons against five state-of-the-art certified defenses. Our results demonstrate consistent improvements in correlation with subjective quality scores by up to **30.9%**. Code is publicly available at [link is hidden for a blind review](#).

Introduction

Image Quality Assessment (IQA) plays a crucial role in numerous applications, from image processing algorithms development to medical imaging and video streaming. Accurate and robust IQA models are essential to reliably measure the perceptual image quality under varying conditions. However, recent studies have shown that IQA models are vulnerable to adversarial perturbations (Zhang et al. 2022; Yang et al. 2024; Antsiferova et al. 2024), which can lead to inaccurate quality scores and compromise their trustworthiness (Deng et al. 2024; Yu et al. 2025; Gotin et al. 2025).

To address this challenge, certified defenses for IQA have emerged as a promising direction, providing robustness guarantees by either constraining the model architecture (Ghazanfari et al. 2023) or injecting Gaussian noise at the input image level (Shumitskaya et al. 2025). Still, these solutions aren’t perfect. Architectural constraints often degrade model accuracy, reducing the correlation between predicted quality scores and true subjective scores. Meanwhile, input-space noise augmentation like randomized smoothing or median smoothing (Cohen, Rosenfeld, and Kolter 2019;

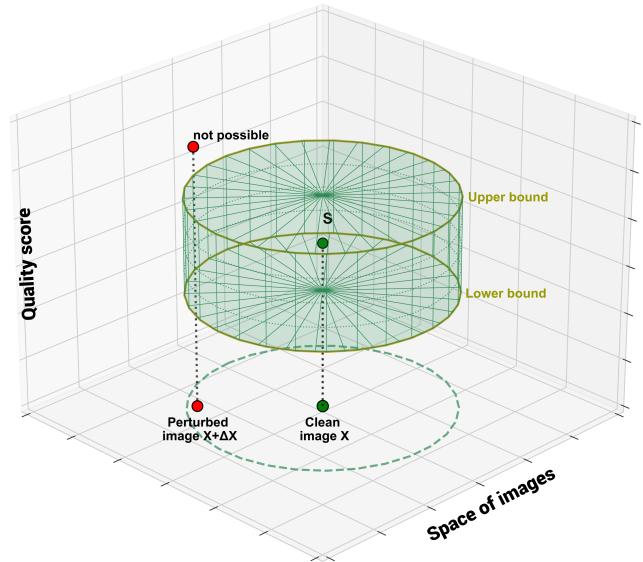


Figure 1: Visualization of a multidimensional cylinder bounding predicted quality score variations under input-space adversarial perturbations. The volume of the cylinder scales with feature-space noise level.

Chiang et al. 2020), though theoretically effective, distorts the original image content. These distortions suppress or affect important visual feature — like fine textures, edges, or compression artifacts — which play a fundamental role in assessing perceptual quality. For instance, if subtle compression artifacts that degrade image quality are smoothed out or masked by added noise, the model may fail to detect these degradations properly, resulting in lower correlation between predicted quality scores and actual human subjective assessments. This trade-off between robustness and accuracy limits the practical usage of such defenses.

To overcome this limitation, we propose to shift the smoothing operation from the input space to a more semantically meaningful feature space. Unlike input-space noise, smoothing features instead of raw images can help preserve critical quality-related information while still providing robustness guarantees. Feature smoothing has been explored

in prior work, primarily for developing defenses in classification tasks (Addepalli et al. 2021; Ma, Dong, and Xu 2023). However, these approaches are empirical and do not establish a formal relationship between noise levels in the feature space and those in the input space. Our research aims to fill this gap by proposing a certified defense applied in the feature space specifically for the IQA task.

The main purpose of our approach is to identify a theoretical multidimensional cylinder that tightly bounds the potential variations of predicted quality scores under adversarial perturbations in the image space (see Figure 1 for visualization). By adjusting the noise level in the feature space, we control the volume of this cylinder — the higher the noise level, the larger the volume.

Our method decomposes the IQA model into two components: a backbone and a scorer module. This design eliminates the need to retrain the backbone network; only the scorer requires fine-tuning. The backbone processes the input image to produce a feature representation, where randomized smoothing is applied. Subsequently, the scorer module generates the final quality score along with robustness certificates. To formally relate the feature space noise to the input perturbation magnitude, we analyze the maximum singular value of the backbone’s Jacobian matrix. More precisely, for a given image, our method outputs a tuple (S, R, S^l, S^u) , where S denotes the predicted quality score, and R represents the input-space radius within which the output of the defended IQA model is guaranteed to lie between S^l (lower bound) and S^u (upper bound). This corresponds to a multidimensional cylinder that bounds quality score variations under allowable input perturbations.

The primary contributions of this work can be summarized as follows:

- To the best of our knowledge, this work presents the first certified defense for IQA, that operates in the feature space instead of the input space, preserving image quality while providing certified robustness. Our method supports both full-reference (FR) and no-reference (NR) IQA models without requiring any architectural modifications.
- We theoretically connect feature-space noise levels to input-space perturbations by analyzing the maximum singular value of the backbone’s Jacobian.
- We conduct extensive experiments on **two datasets** using **six widespread FR and NR IQA models**, comparing our approach against **five** existing state-of-the-art methods. Our method consistently improves the correlation with subjective quality scores by up to **30.9%**.
- Our method is computationally efficient, requiring only one backbone forward pass per image. Compared to existing techniques, it reduces inference time by **99.5%** without certification and by **20.6%** with certification.
- We analyze the method’s usability beyond certified guarantees and demonstrate that it suppresses adversarial gain, improving empirical robustness by **69.9%**.

Our code is publicly available at *link is hidden for a blind review*.

Related Work

Preliminaries. Randomized Smoothing

Randomized Smoothing (**RS**) (Cohen, Rosenfeld, and Kolter 2019) is a widely used certification technique that provides provable robustness guarantees for large-scale models, requiring only black-box access to model evaluations. Originally introduced for classification tasks, RS constructs a smoothed classifier g from a base classifier f by averaging predictions over Gaussian noise perturbations e added to the input x :

$$g(x) = \mathbb{E}\{f(x + e)\}, e \sim \mathcal{N}(0, \sigma^2 I).$$

This approach relies on the robustness of f under Gaussian noise to certify that g is resistant to adversarial perturbations within an l_2 -norm ball of radius ϵ :

$$\epsilon = \frac{\sigma}{2} (\Phi^{-1}(pA) - \Phi^{-1}(pB)),$$

where pA and pB are lower and upper confidence bounds on the top and second-top class probabilities, respectively, and Φ^{-1} denotes the inverse Gaussian cumulative distribution function.

Randomized Smoothing for regression

In this section, we review existing approaches for defending IQA models as well as regression models in general. Since regression models produce continuous outputs for given inputs, developing certified defenses for them is fundamentally more difficult compared to discrete tasks such as classification.

RS-Reg (Rekavandi, Ohrenko, and Rubinstein 2025).

In this paper, the authors generalize randomized smoothing for application to regression models. They define a smoothed function as $g(x) = \text{mean}\{f(x + e)\}, e \sim \mathcal{N}(0, \sigma^2 I)$, where x is an input image, f is the base regression model and e is Gaussian noise sampled from a multivariate normal distribution with covariance matrix $\sigma^2 I$. Using this definition, the authors further derive a probabilistic certified upper bound on input perturbations for the base regression model, assuming the outputs are bounded.

Cert-Reg (Miri Rekavandi et al. 2024). The authors of this paper extend randomized smoothing to regression models using powerful tools from robust statistics, specifically using α -trimming filter as the smoothing function: $g(x) = \frac{1}{N-2[\alpha N]} \sum_{i=[\alpha N]+1}^{N-[\alpha N]} f(x + e), e \sim \mathcal{N}(0, \sigma^2 I)$. Here, the notation is consistent with the previous paper, the values $f(x + e)$ are assumed to be sorted, N denotes the total number of noise samples, and $\alpha \in [0, 0.5]$ is the trimming parameter. In other words, this operator removes the lowest α -fraction and the highest α -fraction of outputs and computes the average of the remaining ones. This approach improves robustness by reducing sensitivity to outliers. Moreover, the authors derive a certification radius for the input perturbation under the assumption that the model outputs are bounded.

MS, DMS (Chiang et al. 2020). In this paper, the authors propose a novel defense approach for regression models using the median as the smoothing function: $g(x) =$

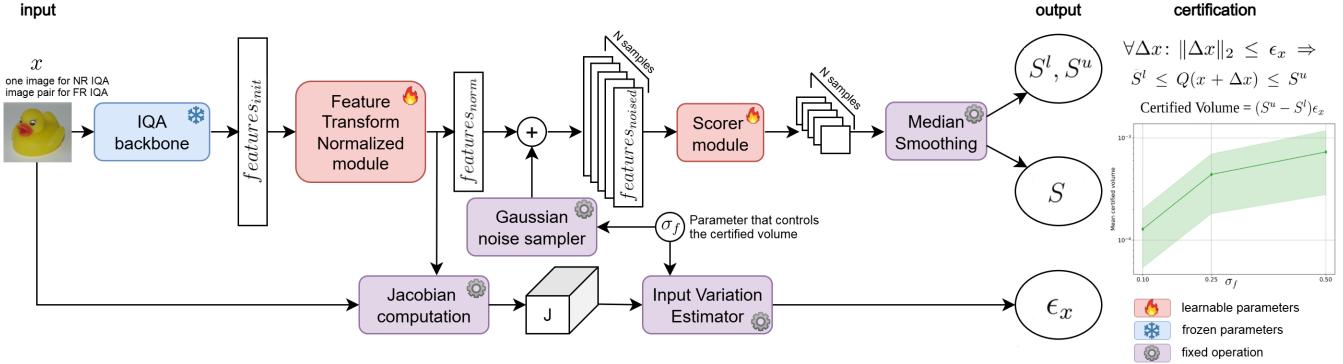


Figure 2: Overview of the proposed FS-IQA defense. The input image or image pairs are first processed by the IQA backbone and the FTN module. Gaussian noise is then added in the feature space. Using Median Smoothing theory (Chiang et al. 2020), we derive the final quality score S along with its lower and upper certified bounds (S^l, S^u). Finally, the proposed Input Variation estimator is used to determine the corresponding input perturbation constraints ϵ_x for the given feature noise level σ_f .

$\text{median}\{f(x + e)\}, \quad e \sim \mathcal{N}(0, \sigma^2 I)$. This operator is significantly more robust to outliers compared to both averaging and the α -trimming filter. To provide certification, the authors prove theorems imposing restrictions on the model’s output within a fixed ℓ_2 -norm ball around the input. Furthermore, they address the potential loss in model performance caused by the added noise and propose an extension of their method that includes an additional denoising step applied after adding noise but before evaluating the regression model’s output. In this paper, we refer to this enhanced method as Denoised Median Smoothing (DMS), while the original approach is called Median Smoothing (MS).

DMS-IQA (Shumitskaya et al. 2025). In this paper, the authors extend the concept of Median Smoothing specifically to defend IQA models. They propose a novel denoiser training scheme based on a composite loss function consisting of three components: pixel-wise MSE, MSE between predicted and true subjective scores, and a differentiable rank loss. Their experimental results demonstrate that this composite loss function enables the training of better denoiser for use within the Median Smoothing framework.

Defenses for IQA

Various empirical defense methods for IQA models have been proposed, including image purification (Gushchin et al. 2024; Liu et al. 2025) and adversarial training (Liu et al. 2024; Chistyakova et al. 2024). However, research on certified defenses specifically designed for IQA remains in its early stages, with only a few existing works addressing this challenge. For example, LipSim (Ghazanfari et al. 2023) proposes using Lipschitz networks to certify a FR IQA model. While this provides robustness guarantees, it is not a defense mechanism in itself, but rather a single robust FR IQA architecture. As a result, this approach lacks scalability to NR IQA models and other FR IQA architectures. Another relevant method, DMS-IQA (Shumitskaya et al. 2025), applies randomized smoothing in the image space as a defense for IQA models. However, this approach perturbs the images

themselves, which degrades image quality and affect the accuracy of quality assessment.

To address these challenges, we introduce a certified defense based on feature-space smoothing, which preserves image fidelity while delivering robust and scalable defense applicable to both FR and NR IQA models.

Proposed Method

Notation and Problem Statement

First, we introduce the notation. We define the defended model $Q(\cdot)$, which maps an input image (or image pairs) to a tuple of outputs:

$$Q : x \in \mathbb{R}^{3 \times H \times W} \rightarrow (S, \epsilon_x, S^l, S^u),$$

where x is an input image for NR IQA and an image pair for FR IQA, S is the predicted quality score of x , S^l and S^u the lower and upper bounds on the quality score under allowable perturbations, and ϵ_x the maximum perturbation magnitude of x .

Our objective is to guarantee that for any input image perturbation Δx satisfying $\|\Delta x\|_2 \leq \epsilon_x$, the predicted quality score remains within the bounds:

$$S_{Q(x)} \leq S_{Q(x+\Delta x)} \leq S_{Q(x)}^u.$$

Note that for FR IQA, the perturbation Δx refers to perturbation of the distorted image only, while the reference image is considered fixed.

Overview of the FS-IQA Architecture

Figure 2 provides an overview of the proposed FS-IQA method. The defended model Q is composed of three CNN-based modules: the IQA backbone $b(\cdot)$, the Feature Transform Normalization Module $FTN(\cdot)$, and the scorer module $S(\cdot)$. Beyond the neural modules, FS-IQA includes two components that provide certified robustness guarantees: the Median Smoothing operator (Chiang et al. 2020) and the proposed Input Variation Estimator. These components will be described in detail in the following subsections.

Feature Preparation Pipeline

First, the input image x is processed by the IQA backbone $b(\cdot)$, producing the initial features:

$$f_{init} = b(x)$$

Next, these features are passed through the Feature Transform Normalize Module to obtain normalized features:

$$f_{norm} = FTN(f_{init})$$

The primary objectives of the Feature Transform Normalize Module are the following:

- Dimensionality reduction — the module reduces the feature dimensionality to 512. This step is important for memory-efficient computation of the Jacobian matrix of the module's output with respect to the input image.
- Normalization — the module constrains the feature values within the $[0, 1]$ range. This normalization prepares the features for the subsequent consistent application of Gaussian noise.

Then, we generate N samples of Gaussian noise and add them to the normalized features to obtain smoothed features:

$$\{f_{noised}^i = f_{norm} + e_i \mid e_i \sim \mathcal{N}(0, \sigma_f^2 I), \quad i = 1, \dots, N\},$$

where σ_f denotes the noise standard deviation, I is the identity matrix, and N is the number of samples.

Input Variation Estimator

Next, given the feature noise standard deviation σ_f , we use the proposed Input Variation Estimator module to determine the maximum allowable perturbation ϵ_x on the input image in terms of its l_2 norm, i.e.,

$$\epsilon_x = IVE(FTN \circ b, x, \sigma_f).$$

The value ϵ_x quantifies how much the input image x can be perturbed without causing feature variations larger than σ_f . To estimate ϵ_x , we rely on a linear approximation of the composite function $B = FTN \circ b$ around x , formalized in the following theorem:

Theorem 1 (Input Variation Estimator). *Let $B : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a differentiable feature map at point x , and let $J_B(x) \in \mathbb{R}^{m \times n}$ be its Jacobian at x . For any input perturbation $u \in \mathbb{R}^n$ satisfying*

$$\|u\|_2 \leq \epsilon_x,$$

the change in features satisfies

$$\max_{\|u\|_2 \leq \epsilon_x} \|B(x + u) - B(x)\|_2 \leq \sigma_f.$$

Using the first-order approximation $B(x + u) \approx B(x) + J_B(x)u$, the maximal allowed input perturbation ϵ_x is given by

$$\epsilon_x = \frac{\sigma_f}{\|J_B(x)\|_2},$$

where $\|J_B(x)\|_2$ denotes the spectral norm (operator norm induced by the Euclidean norm) of the Jacobian matrix.

Proof. By definition, the maximal input perturbation ϵ_x is the largest radius ρ such that for every vector u with $\|u\|_2 \leq \rho$, the corresponding feature change does not exceed σ_f :

$$\epsilon_x = \max_{\rho \geq 0} \{\forall u, \|u\|_2 \leq \rho \implies \|B(x + u) - B(x)\|_2 \leq \sigma_f\}.$$

For sufficiently small perturbations u , we can use the first-order Taylor expansion: $B(x + u) \approx B(x) + J_B(x)u$, where $J_B(x)$ is the Jacobian matrix at the point x . This approximation implies that the change in features is approximately linear with respect to u , so

$$\|B(x + u) - B(x)\|_2 \approx \|J_B(x)u\|_2.$$

$$\|J_B(x)u\|_2 \leq \|J_B(x)\|_2 \|u\|_2.$$

To ensure the feature change never exceeds σ_f for all u with $\|u\|_2 \leq \rho$, it suffices to require

$$\|J_B(x)\|_2 \rho \leq \sigma_f,$$

which rearranges to

$$\rho \leq \frac{\sigma_f}{\|J_B(x)\|_2}.$$

Since ϵ_x is defined as the maximal such ρ , we get:

$$\epsilon_x = \frac{\sigma_f}{\|J_B(x)\|_2}.$$

□

Remark 1. *The spectral norm $\|J_B(x)\|_2$ of the Jacobian matrix is equal to its largest singular value $\sigma_{\max}(J_B(x))$. This equivalence is important: the largest singular value represents the greatest factor by which the feature map's linear approximation can amplify the input perturbation in any direction. Bounding the input perturbation via this norm therefore provides control over the worst-case behavior of the feature change.*

Median Smoothing in the feature space

To obtain the final quality score, as well as the lower and upper bounds on the model output, we use Median Smoothing theory (Chiang et al. 2020). Final quality score calculates as the value of Median Smoothing operator, applied to the set of smoothed features:

$$\bar{v} = \{\text{Scorer}(f_{noised}^i) \mid i = 1, \dots, N\}$$

$$S = \text{median}(\bar{v}) \tag{1}$$

$$(S^l, S^u) = MS_{\text{Cert}}(\bar{v}, \sigma_f, N, \alpha)$$

where α is the confidence level, $MS_{\text{Cert}}(\cdot)$ is an operator that provides certified guarantees for Median Smoothing when $\text{Scorer}(\cdot)$ is a regression function, i.e., a mapping from \mathbb{R}^k to \mathbb{R} . Formally, these guarantees follow from the theorem below:

Theorem 2. (Chiang et al. 2020) If $\text{Scorer}(\cdot)$ is a regression function from the space of features to the space of real numbers, f_{norm} is normalized features, G is the Median Smoothing operator in the form (1), then for all $\|u\|_2 \leq \varepsilon_f$:

$$S^l \leq G(S(f_{\text{norm}} + u)) \leq S^u. \quad (2)$$

Here, $\underline{p} = \Phi(-\frac{\varepsilon_f}{\sigma_f})$, $\bar{p} = \Phi(\frac{\varepsilon_f}{\sigma_f})$ and $\Phi(\cdot)$ is the Gaussian cumulative density function. G is an operator of Median Smoothing, S^l and S^u are defined as the percentiles of the smoothed function:

$$\begin{aligned} S^l &= \sup_{y \in \mathbb{R}} \{ \mathbb{P}_{r \sim \mathcal{N}(0, \sigma_f^2 I)} [S(f_{\text{norm}} + r) \leq y] \leq \underline{p} \}, \\ S^u &= \inf_{y \in \mathbb{R}} \{ \mathbb{P}_{r \sim \mathcal{N}(0, \sigma_f^2 I)} [S(f_{\text{norm}} + r) \leq y] \geq \bar{p} \}. \end{aligned}$$

Training and application

The parameters of the IQA backbone $b(\cdot)$ are frozen, while $FTN(\cdot)$ and $\text{Scorer}(\cdot)$ are trainable. The training objective focuses solely on predicting the quality scores S , since the primary goal of the IQA task is to produce scores that highly correlate with human assessments. Accordingly, we use the mean squared error (MSE) loss between the predicted scores S and the ground-truth subjective quality scores.

Pseudocode for the inference procedure of the FS-IQA model is provided in Algorithm 1.

Algorithm 1: **Pseudocode** for Quality Prediction and Certification for Feature-Smoothed IQA Model Q on x

Input: Image (or image pair) x , feature noise bound σ_f , number of samples N , confidence level α , Jacobian norm threshold $\tau > 0$
Output: Quality score S , certified input bound ϵ_x , and lower and upper output bounds S^l, S^u such that $\forall \Delta x : \|\Delta x\|_2 \leq \epsilon_x \Rightarrow S^l \leq Q(x + \Delta x) \leq S^u$, or ABSTAIN if certification is not reliable

```

1:  $f_{\text{init}} \leftarrow IQA_{\text{backbone}}(x)$  {Extract initial features}
2:  $f_{\text{norm}} \leftarrow FTN(f_{\text{init}})$  {Apply Feature Transform Normalization}
3:  $J \leftarrow \text{Jacobian}(FTN \circ IQA_{\text{backbone}}, x)$  {Compute Jacobian at  $x$ }
4: if  $\|J\|_2 < \tau$  then
5:   return ABSTAIN {Jacobian norm too small — certification not reliable}
6: end if
7:  $\epsilon_x \leftarrow \frac{\sigma_f}{\|J\|_2}$  {Maximal allowed input perturbation}
8: Initialize empty list scores
9: for  $i = 1 \rightarrow N$  do
10:   Sample noise vector  $e \sim \mathcal{N}(0, \sigma_f I)$ 
11:    $f_{\text{noised}} \leftarrow f_{\text{norm}} + e$  {Add noise in feature space}
12:    $S_{\text{cur}} \leftarrow \text{Scorer}(f_{\text{noised}})$ 
13:   Append  $S_{\text{cur}}$  to scores
14: end for
15:  $S \leftarrow \text{median}(\text{scores})$ 
16:  $(S^l, S^u) \leftarrow M_{\text{Cert}}(\text{scores}, \sigma_f, N, \alpha)$ 
17: return  $S, S^l, S^u, \epsilon_x$ 

```

Experimental verification of theoretical constraints can be found in the supplementary material.

Experiments

FS-IQA parameters

The FTN and Scorer modules have a simple architecture consisting of fully connected layers. More details in the supplementary material. In our experiments, we use a sample size of $N = 2000$, a confidence level of $\alpha = 0.999$, and a Jacobian norm threshold $\tau = 0.001$. To benchmark the effectiveness of our FS-IQA approach, we conducted comparisons with five certified defenses on NR and FR IQA models. Experiments were conducted using a GPU server powered by NVIDIA A100 GPUs.

NR and FR IQA models

As NR models we selected **DBCNN** (Zhang et al. 2020), **HyperIQA** (Su et al. 2020) and **KonCept** (Hosu et al. 2020). As FR models we selected **LPIPS** (Zhang et al. 2018), **DISTS** (Ding et al. 2020) and **DreamSim** (Fu et al. 2023). These models are widely recognized in the field of IQA due to their strong performance and their diverse architectural approaches (Antsiferova et al. 2022).

Compared Methods

For comparison, we include all known certified defense methods to date, both those created specifically for IQA and those made for general regression tasks that can be adapted to IQA. These methods are: **RS-Reg** (Rekavandi, Ohrimenko, and Rubinstein 2025) and **Cert-Reg** (Miri Rekavandi et al. 2024), proposed for classification; **MS** and **DMS** (Chiang et al. 2020), proposed for detection; and **DMS-IQA** (Shumitskaya et al. 2025), designed specifically for IQA. Detailed descriptions of these methods are provided in the Related Work section.

Datasets

For experiments on NR IQA models, we use the **KoniQ-10k** image dataset (Hosu et al. 2020), consisting of 10,073 images. For FR models, we use the **Kadid-10k** database (Lin, Hosu, and Saupe 2019), consisting of 10,125 images. These datasets are commonly used for designing IQA models. All images have a resolution of 384×512 pixels. Both datasets were split into training and testing subsets, using an 80% / 20% ratio. For defense methods that require a training stage (DMS, DMS-IQA, and the proposed FS-IQA), the training subset was used. All methods were then evaluated on the testing subset using a sample size of $N = 2000$ and a confidence level of $\alpha = 0.999$. Other parameters are detailed in the supplementary material.

Evaluation metrics

As evaluation metrics for the compared methods, we primarily use Spearman’s rank correlation coefficient (**SRCC**) and Pearson’s linear correlation coefficient (**PLCC**) with human-assessed subjective scores, since the main goal of IQA is to accurately estimate image quality. Additionally, for comparison, we measure the computational **time** required to process one image, both with and without certification. The runtime was averaged over 100 runs.

Method	SRCC/PLCC			time	
	$\sigma = 0.1$	$\sigma = 0.25$	$\sigma = 0.5$	with cert	no cert
RS-Reg	$0.66 \pm 0.02/0.67 \pm 0.04$	$0.38 \pm 0.07/0.38 \pm 0.06$	$-0.04 \pm 0.05/-0.02 \pm 0.06$	$4.2 \pm 1.7 \text{ sec}$	$3.8 \pm 1.7 \text{ sec}$
Cert-Reg	$0.66 \pm 0.02/0.67 \pm 0.04$	$0.38 \pm 0.07/0.38 \pm 0.06$	$-0.04 \pm 0.05/-0.02 \pm 0.06$	$4.2 \pm 1.7 \text{ sec}$	$4.2 \pm 1.7 \text{ sec}$
MS	$0.66 \pm 0.02/0.67 \pm 0.04$	$0.38 \pm 0.07/0.38 \pm 0.06$	$-0.04 \pm 0.05/-0.02 \pm 0.06$	$4.2 \pm 1.7 \text{ sec}$	$3.8 \pm 1.7 \text{ sec}$
DMS	$0.84 \pm 0.04/0.87 \pm 0.03$	$0.74 \pm 0.05/0.78 \pm 0.05$	$0.63 \pm 0.04/0.67 \pm 0.05$	$7.9 \pm 1.8 \text{ sec}$	$7.4 \pm 1.7 \text{ sec}$
DMS-IQA	$0.86 \pm 0.02/0.88 \pm 0.02$	$0.80 \pm 0.03/0.83 \pm 0.02$	$0.70 \pm 0.03/0.73 \pm 0.02$	$7.9 \pm 1.8 \text{ sec}$	$7.4 \pm 1.7 \text{ sec}$
FS-IQA (ours)	$0.91 \pm 0.01/0.93 \pm 0.01$	$0.91 \pm 0.01/0.93 \pm 0.01$	$0.91 \pm 0.02/0.92 \pm 0.01$	$6.9 \pm 2.3 \text{ sec}$	$33 \pm 8 \text{ ms}$

Table 1: Comparison of methods on NR IQA models (DBCNN, HyperIQA and KonCept) using the KonIQ-10k dataset. All metrics are averaged across the IQA models. Data are presented as mean \pm standard error of the mean (SEM). Time shows time to process one image with resolution 384×512 . Detailed results for each IQA model are in the supplementary material.

Method	SRCC/PLCC			time	
	$\sigma = 0.1$	$\sigma = 0.25$	$\sigma = 0.5$	with cert	no cert
RS-Reg	$0.54 \pm 0.08/0.54 \pm 0.05$	$0.33 \pm 0.12/0.33 \pm 0.13$	$0.14 \pm 0.11/0.13 \pm 0.13$	$6.9 \pm 2.9 \text{ sec}$	$6.4 \pm 2.9 \text{ sec}$
Cert-Reg	$0.54 \pm 0.08/0.54 \pm 0.05$	$0.33 \pm 0.12/0.33 \pm 0.13$	$0.14 \pm 0.11/0.13 \pm 0.13$	$6.9 \pm 2.9 \text{ sec}$	$6.9 \pm 2.9 \text{ sec}$
MS	$0.54 \pm 0.08/0.54 \pm 0.05$	$0.33 \pm 0.12/0.33 \pm 0.13$	$0.14 \pm 0.11/0.13 \pm 0.13$	$6.9 \pm 2.9 \text{ sec}$	$6.4 \pm 2.9 \text{ sec}$
DMS	$0.75 \pm 0.02/0.70 \pm 0.04$	$0.62 \pm 0.01/0.62 \pm 0.02$	$0.50 \pm 0.01/0.53 \pm 0.01$	$10.5 \pm 2.9 \text{ sec}$	$10.0 \pm 2.9 \text{ sec}$
DMS-IQA	$0.78 \pm 0.02/0.72 \pm 0.04$	$0.63 \pm 0.01/0.62 \pm 0.02$	$0.49 \pm 0.04/0.51 \pm 0.03$	$10.5 \pm 2.9 \text{ sec}$	$10.0 \pm 2.9 \text{ sec}$
FS-IQA (ours)	$0.89 \pm 0.01/0.89 \pm 0.02$	$0.90 \pm 0.01/0.89 \pm 0.02$	$0.89 \pm 0.01/0.89 \pm 0.01$	$7.5 \pm 3.6 \text{ sec}$	$40 \pm 14 \text{ ms}$

Table 2: Comparison of methods on FR IQA models (LPIPS, DISTs and DreamSim) using the Kadid-10k dataset. All metrics are averaged across the IQA models. Data are presented as mean \pm standard error of the mean (SEM). Time shows time to process one image with resolution 384×512 . Detailed results for each IQA model are in the supplementary material.

Results

Tables 1 and 2 present the comparison results of FS-IQA with prior certified defenses on NR IQA and FR IQA models, respectively. For RS-Reg, Cert-Reg, and MS methods, no significant differences were observed. This suggests that mean, α -trimmed mean, and median data reduction techniques yield similar performance for IQA models. Since IQA models are typically trained on noisy data and are robust to noise outliers, developing complex methods specifically for outlier handling is unnecessary. However, these three methods exhibit poor SRCC and PLCC scores, which dramatically decrease as σ increases. DMS and DMS-IQA methods show significantly better SRCC and PLCC results, though at the cost of increased computation time due to the denoising step. Finally, the proposed FS-IQA method demonstrates the best performance in terms of SRCC and PLCC scores, with only minor decreases as σ rises. This indicates that FS-IQA leverages more semantic features rather than relying solely on input images, making the defense pipeline adaptable and effective even under high noise levels. On average, FS-IQA improves SRCC by approximately **31.3%** and PLCC by **30.5%** compared to the best existing DMS-IQA method across the three σ_f values.

Regarding computational time, FS-IQA with certification mode is slightly slower than classic RS-Reg, Cert-Reg, and MS methods, but approximately **20.6%** faster than the state-of-the-art DMS and DMS-IQA approaches. We also evaluated the methods in a non-certification mode — i.e., when only the certified quality score is computed without calculating restrictions. This mode is useful for real-time applications that use a reliable IQA model proven robust on specific

data types, without spending extra time checking restrictions for each item. As observed, all prior certified methods struggle to deliver fast performance in this setting because they require running the IQA backbone multiple times. In contrast, our approach shifts the smoothing operation to the feature space, overcoming this limitation. This results in a dramatic speed advantage: 33 milliseconds versus the second-best time of 7.4 seconds. The difference is significant, highlighting FS-IQA’s practical benefits for real-time applications.

Figure 3 presents a visualization of the certified guarantees provided by FS-IQA for NR IQA models at different noise levels σ_f . The graph shows the average difference between the upper and lower bounds versus ϵ_x , grouping the data into 10 quantile-based bins according to ϵ_x , with mean values plotted for each group. As observed, increasing σ_f results in wider bounds, meaning that by adjusting the σ_f level, we can control the strength of the certification guarantees. Additionally, as shown in Figure 3, for IQA models it is preferable to be located in the bottom-right corner of the plot, which indicates that the model’s output changes very little within a large neighborhood around the input — thus providing stronger guarantees. This visualization allows us to compare different IQA backbones and select the best option. For example, the KonCept line consistently lies below the HyperIQA line across all σ_f levels, indicating that for the same ϵ_x values, KonCept imposes tighter restrictions on the model output. Therefore, we can conclude that KonCept, combined with FS-IQA, provides stronger certified guarantees than HyperIQA.

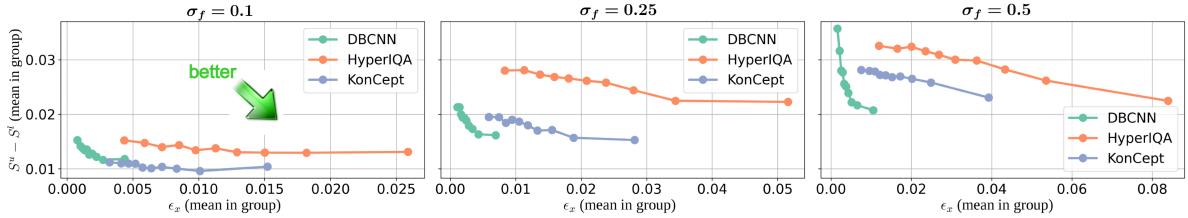


Figure 3: Average difference between upper and lower bounds ($S^u - S^l$) versus input perturbation l_2 norm constraint ϵ_x for three NR IQA models (DBCNN, HyperIQA, KonCept) at noise scales $\sigma \in \{0.1, 0.25, 0.5\}$, produced using FS-IQA. Data are grouped into 10 quantile-based bins by ϵ_x , with mean values plotted for each group.

Discussion

Beyond Certified Guarantees

It is a popular topic of discussion that today’s certified defenses are not practically useful because they provide tight certified guarantees at the cost of increased computation time. Additionally, questions arise about how these methods can defend against much larger perturbations. To explore this issue, we conducted additional experiments to evaluate the empirical robustness of our approach against perturbations with amplitudes much more greater than the certified robustness guarantee. Specifically, we applied I-FGSM (10 iterations) (Kurakin, Goodfellow, and Bengio 2017) to generate adversarial perturbations on 100 images from the KonIQ dataset using the KonCept NR model, testing multiple perturbation norms: $\{0.02, 0.05, 0.1, 0.15, 0.20, 0.25\}$. We then measured the performance degradation of both the original KonCept model and the FS-IQA defended KonCept model under these stronger attacks (see Figure 4). FS-IQA’s scores decreased, matching human perception, while the undefended KonCept’s scores grew by up to 60%. This shows FS-IQA stays robust even beyond its certified guarantees.

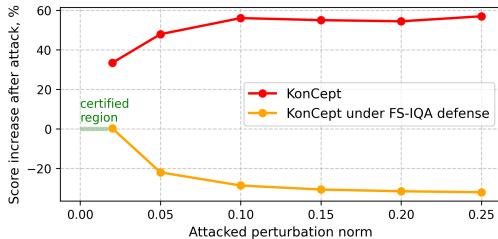


Figure 4: Evaluation of FS-IQA robustness beyond the certified region against adversarial perturbations generated by I-FGSM (10 iterations).

Output stability

FS-IQA is a stochastic method; however, we additionally demonstrated that running the model multiple times on the same image results in only minor variations in the output. We measured the deviation of the model’s output on a single image across multiple runs and found that the change in output is only 0.06%. Details in the supplementary material.

Limitations

One main limitation of the FS-IQA approach is its high computational complexity in certification mode. Currently, the processing time is not suitable for real-time applications. However, with increased computational resources, this may become more practical. Therefore, we suggest using the FS-IQA method in real-time applications without the certification mode, combined with pre-testing on a specific data type using certification to estimate robustness, since similar data typically exhibit similar certified properties. As demonstrated in our experiments, this mode is fast—comparable to the speed of standard IQA models, while providing good empirical robustness. It is worth noting that previous certified methods without certification mode are still relatively slow, making our method more suitable for real-time use.

The second limitation is the number of abstentions. When the certification is not reliable, the method outputs an abstain decision. In all our experiments, the percentage of abstains was 1.5%. Although this rate is not very high, it depends on the IQA backbone and the specific data. Therefore, in practice, users can select a suitable IQA backbone for the given data to minimize the number of abstains.

Conclusion

This paper introduces a novel certified defense method for IQA models based on randomized smoothing with noise applied in the feature space instead of the input space. Through extensive experiments on two benchmark datasets with six popular FR and NR IQA models, we showed that in contrast to prior approaches, our method achieves 30.9% better correlation with subjective scores. Importantly, FS-IQA offers a significant speed advantage, especially in non-certification mode, boosting efficiency by 99.5%, making it practical for real-time applications where certified defenses have historically struggled due to resource-intensive computations. Beyond theoretical guarantees, FS-IQA exhibited strong empirical robustness against perturbations significantly larger than the certified bounds, addressing common criticisms of certified defenses. Overall, FS-IQA presents a promising direction for robust and efficient certified defenses in IQA, balancing theoretical guarantees with practical usability. Code is publicly available at [link is hidden for a blind review](#).

References

- Addepalli, S.; Jain, S.; Sriramanan, G.; and Babu, R. V. 2021. Boosting adversarial robustness using feature level stochastic smoothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 93–102.
- Antsiferova, A.; Abud, K.; Gushchin, A.; Shumitskaya, E.; Lavrushkin, S.; and Vatolin, D. 2024. Comparing the robustness of modern no-reference image-and video-quality metrics to adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 700–708.
- Antsiferova, A.; Lavrushkin, S.; Smirnov, M.; Gushchin, A.; Vatolin, D.; and Kulikov, D. 2022. Video compression dataset and benchmark of learning-based video-quality metrics. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 13814–13825. Curran Associates, Inc.
- Chiang, P.-y.; Curry, M.; Abdelkader, A.; Kumar, A.; Dickerson, J.; and Goldstein, T. 2020. Detection as regression: Certified object detection with median smoothing. *Advances in Neural Information Processing Systems*, 33: 1275–1286.
- Chistyakova, A.; Antsiferova, A.; Khrebtov, M.; Lavrushkin, S.; Arkhipenko, K.; Vatolin, D.; and Turdakov, D. 2024. Increasing the robustness of image quality assessment models through adversarial training. *Technologies*, 12(11): 220.
- Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, 1310–1320. PMLR.
- Deng, W.; Yang, C.; Huang, K.; Liu, Y.; Gui, W.; and Luo, J. 2024. Sparse adversarial video attack based on dual-branch neural network on industrial artificial intelligence of things. *IEEE Transactions on Industrial Informatics*, 20(7): 9385–9392.
- Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2020. Image Quality Assessment: Unifying Structure and Texture Similarity. *CoRR*, abs/2004.07728.
- Fu, S.; Tamir, N.; Sundaram, S.; Chai, L.; Zhang, R.; Dekel, T.; and Isola, P. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. *Advances in Neural Information Processing Systems*, 36: 50742–50768.
- Ghazanfari, S.; Araujo, A.; Krishnamurthy, P.; Khorrami, F.; and Garg, S. 2023. Lipsim: A provably robust perceptual similarity metric. *arXiv preprint arXiv:2310.18274*.
- Gotin, G.; Shumitskaya, E.; Antsiferova, A.; and Vatolin, D. 2025. Cross-Modal Transferable Image-to-Video Attack on Video Quality Metrics. In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 3: VISAPP*, 880–888. INSTICC, SciTePress. ISBN 978-989-758-728-3.
- Gushchin, A.; Abud, K.; Bychkov, G.; Shumitskaya, E.; Chistyakova, A.; Lavrushkin, S.; Rasheed, B.; Malyshev, K.; Vatolin, D.; and Antsiferova, A. 2024. Guardians of image quality: Benchmarking defenses against adversarial attacks on image quality metrics. *arXiv preprint arXiv:2408.01541*.
- Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29: 4041–4056.
- Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2017. Adversarial examples in the physical world. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.
- Lin, H.; Hosu, V.; and Saupe, D. 2019. KADID-10k: A large-scale artificially distorted IQA database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 1–3. IEEE.
- Liu, Y.; Yang, C.; Li, D.; Ding, J.; and Jiang, T. 2024. Defense against adversarial attacks on no-reference image quality models with gradient norm regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25554–25563.
- Liu, Y.; Yang, C.; Yu, Z.; and Huang, T. 2025. Enhancing NR-IQA Model Robustness through Simple Image Compression Techniques. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Ma, Y.; Dong, M.; and Xu, C. 2023. Adversarial robustness through random weight sampling. *Advances in Neural Information Processing Systems*, 36: 37657–37669.
- Miri Rekavandi, A.; Farokhi, F.; Ohrimenko, O.; and Rubinstein, B. I. 2024. Certified Adversarial Robustness via Randomized α -Smoothing for Regression Models. *Advances in Neural Information Processing Systems*.
- Rekavandi, A. M.; Ohrimenko, O.; and Rubinstein, B. I. 2025. RS-Reg: Probabilistic and Robust Certified Regression Through Randomized Smoothing. *TMLR*.
- Shumitskaya, E.; Pautov, M.; Vatolin, D.; and Antsiferova, A. 2025. Stochastic BIQA: Median randomized smoothing for certified blind image quality assessment. *Computer Vision and Image Understanding*, 104447.
- Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 2020. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yang, C.; Liu, Y.; Li, D.; and Jiang, T. 2024. Exploring vulnerabilities of no-reference image quality assessment models: A query-based black-box method. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Yu, Y.; Xia, S.; Lin, X.; Yang, W.; Lu, S.; Tan, Y.-P.; and Kot, A. 2025. Backdoor attacks against no-reference image quality assessment models via a scalable trigger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9698–9706.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.
- Zhang, W.; Li, D.; Min, X.; Zhai, G.; Guo, G.; Yang, X.; and Ma, K. 2022. Perceptual attacks of no-reference image quality models with human-in-the-loop. *Advances in Neural Information Processing Systems*, 35: 2916–2929.

Zhang, W.; Ma, K.; Yan, J.; Deng, D.; and Wang, Z. 2020.
Blind Image Quality Assessment Using A Deep Bilinear
Convolutional Neural Network. *IEEE Transactions on Cir-
cuits and Systems for Video Technology*, 30(1): 36–47.