Записки научных
семинаров ПОМИ
Том 552, 2026 г.

E. A. Shumitskaya, D. S. Vatolin, A. V. Antsiferova

EXPLOITING STOCHASTIC WEAKNESSES OF RANDOMIZED SMOOTHING IN REGRESSION MODELS: ADAPTIVE ATTACK ANALYSIS AND EMPIRICAL LIMITATIONS

Abstract. Randomized smoothing has recently been extended from classification to regression-based tasks such as Image Quality Assessment (IQA), offering formal robustness guarantees against bounded input perturbations. Among these methods, median smoothing has emerged as a robust alternative to traditional mean-based smoothing, promising greater resistance to outliers. However, its practical robustness against adaptive adversaries remains insufficiently explored. In this work, we investigate the vulnerability of median-smoothed IQA models to gradient-based attacks. We demonstrate that while theoretical robustness guarantees hold under idealized conditions of infinite sampling, real-world implementations are susceptible to finite-sample uncertainty and stochastic bias. To expose these weaknesses, we extend Projected Gradient Descent (PGD) with Expectation over Transformation (EOT), enabling adaptive attacks that explicitly exploit the stochastic nature of randomized smoothing. Experiments on three state-of-the-art IQA models – KonCept, PaQ-2-PiQ, and HyperIQA – demonstrate that the proposed attack significantly increases both Adversarial Gain (AdvG) and Certified Attack Success Rate (cASR), outperforming baseline methods such as FGM and PGD. We further show that the robustness of median smoothing strongly depends on the sampling budget and noise level: insufficient sampling or suboptimal noise parameters lead to complete breakdowns of certification. These findings reveal a substantial gap between theoretical and empirical robustness and highlight the need for improved smoothing-based defenses in regression tasks.

Key words and phrases: randomized smoothing, median smoothing, certified defenses, regression models, image quality assessment, IQA robustness, adaptive attacks.

The research was carried out using the MSU-270 supercomputer of Lomonosov Moscow State University and ISP RAS computing cluster.

§1. Introduction

Adversarial robustness has become a central topic in modern computer vision. While numerous studies have explored methods to protect classifiers from small, human-imperceptible perturbations [1-2], regression-based problems – such as Image Quality Assessment (IQA) – have only recently received attention [3-4]. These problems are particularly relevant in perceptual optimization, compression, and image enhancement, where reliable continuous predictions are essential.

Unlike classifiers that output discrete labels, regression models produce continuous-valued predictions, making certified robustness more challenging. In regression, robustness must be defined in terms of bounded changes in the output value itself, rather than crossings of discrete decision boundaries as in classification.

Randomized smoothing (RS) [1] has emerged as one of the few defense mechanisms that can provide formal robustness guarantees. In classification, it aggregates predictions across random perturbations using a majority vote, yielding models that are provably robust within a certified radius [5]. Extending this idea, Median Randomized Smoothing [6] has been proposed to defend regression models by replacing the mean with a more robust aggregation function. This design improves tolerance to outliers, which is particularly relevant for IQA, where score distributions are typically heavy-tailed or skewed due to subjective variations in human perception.

However, in practice, the robustness guarantees of smoothing-based regressors rely on finite-sample Monte Carlo approximations rather than the actual underlying distribution [1]. This difference allows adaptive attackers to exploit the randomness in the sampling process and influence the estimated results. In classification, such attacks have been able to bypass randomized defenses when the number of samples is small [7]. However, similar attacks on regression-based smoothing have not been studied in detail.

This paper addresses this gap by studying median-smoothed IQA models. The approach for RS defense in regression is similar to classical classification solutions: for IQA, we generate N noisy variants of an image using Gaussian noise with a given σ, and take the median of the model outputs as the final prediction. Figure 1 illustrates this process for N samples ($N = [10, 50, 100]$) and noise levels $\sigma = [4/255, 8/255]$ when defending

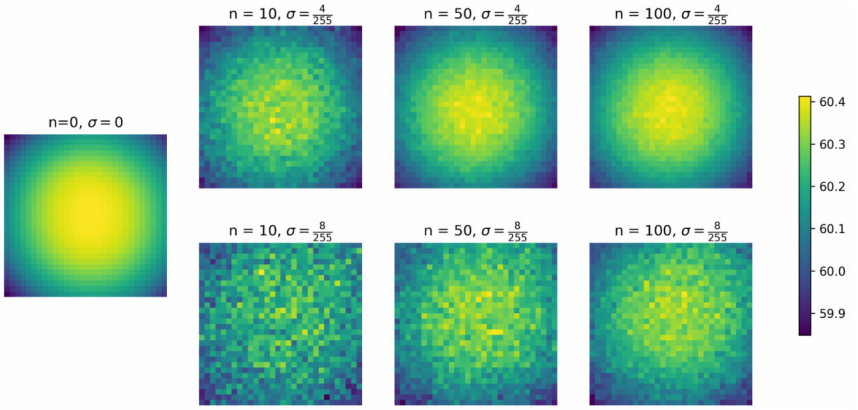

Figure 1. Median-smoothed outputs of the KonCept IQA model for different N and σ. Higher noise reduces output variance, and more samples produce smoother, tighter peaks. The leftmost plot shows the base model without median smoothing.

the KonCept IQA model on a randomly selected image from the KonIQ dataset.

In these plots, we show noisy samples along two random, fixed directions in the input image space, with a step size of 0.002. The leftmost plot shows the outputs of the base KonCept IQA model without median smoothing. We observe that higher noise levels reduce the variance of the outputs, while increasing the number of samples N produces smoother outputs and tighter peaks at higher values compared to the original model.

The goal of this paper is to systematically evaluate adaptive gradient-based attacks on median-smoothed IQA models and uncover critical vulnerabilities that call their real-world security into question. Specifically, we demonstrate that, although formal robustness guarantees hold under idealized assumptions of infinite sampling, in practice they can be violated due to finite-sample estimation errors and stochastic bias in the smoothing process. Here, stochastic bias refers to a systematic mismatch between the true population statistic of the smoothed model (e.g., the median of $f(x + \delta)$ under the noise distribution) and its finite-sample estimate used at inference time. When the output distribution under noise is asymmetric

or unstable, this bias does not vanish with typical sample sizes and can be consistently exploited by an adaptive adversary. As a result, robustness guarantees derived for the infinite-sample setting may fail to hold in practical deployments.

Main contributions of this paper are as follows:

- Adaptive attack formulation for smoothed regressors. We extend the Expectation-over-Transformation (EOT) framework to regression-based randomized smoothing, enabling gradient-based adaptive attacks on median-smoothed IQA models;
- Analysis of certification limitations. We experimentally examine the probabilistic nature of robustness certification under finite sampling and show how percentile estimation uncertainty can be exploited by adaptive adversaries;
- Comprehensive empirical evaluation. Through extensive experiments on standard IQA datasets, we quantify the gap between certified and empirical robustness. Our proposed EOT-PGD attack consistently breaks certified bounds under realistic sampling budgets;
- Robustness–efficiency trade-offs. We study the interplay between noise variance and sample size, highlighting the key trade-off that determine the practical robustness of smoothed regression models.

All code and experimental configurations used in this study is available [8].

The rest of this paper is organized as follows: Section 2 reviews related work on randomized smoothing and certified defenses in regression settings; Section 3 describes the proposed method; Section 4 contains experimental setup; Section 5 presents and discusses the results; and Section 6 concludes the paper.

§2. Background and related work

2.1. Attacks on randomized smoothing. Developed by Lecuyer et al. [5], Randomized Smoothing (RS) is a general method for achieving certified local robustness, regardless of the model architecture. Consider a trained binary classifier $f : R^d \to \{0, 1\}$. RS defines a new classifier g_σ as follows:

$$g(x) = \arg\max_{y\in\{0,1\}} P[f(x+\eta) = y], \;\; \eta \sim N(0, \sigma^2 I) \tag{1}$$

where P denotes the probability, x represents the input image, and η is a sample of additive Gaussian noise with standard deviation σ. Cohen et al. [1] demonstrated that the main advantage of g_σ lies in its certified robustness. Suppose the exact values of the two probabilities are given $\pi_0(x) = P[f(x+\eta) = 0]$ and $\pi_1(x) = 1 - \pi_0(x)$, then we can calculate the certified radius:

$$R(x, \sigma) = \sigma \Phi^{-1}(\pi_{g_\sigma(x)}(x)) \tag{2}$$

where Φ is the cumulative distribution function of the standard normal distribution $N(0, \sigma^2 I)$. In other words, RS guarantees that no adversarial example exists within this radius. The strength of this approach lies in the fact that adversarial attacks no longer need to be considered, as robustness is formally ensured. However, in practice, the actual class probabilities are unknown and cannot be computed exactly due to the high dimensionality of the input space used by modern classifiers. Instead, Monte Carlo sampling methods are used to estimate these probabilities. Consequently, the empirical classifier $g_{\sigma,n}$, behaves like g_σ only in the limit as $n \to \infty$. This issue has recently been discussed in the literature.

Maho et al. [7] examined the gap between the theoretical guarantees of randomized smoothing and its empirical robustness, emphasizing the challenges posed by Monte Carlo estimation in high-dimensional models. Delattre et al. [9] derive a new certified robust radius for RS that accounts for the Lipschitz continuity of the base classifier. However, applying this result in practice presents a key challenge: computing local Lipschitz constants, which are difficult to estimate precisely with certification. To approximate these local Lipschitz constants, the authors adopt the method proposed by Yang et al. [10], which employs a gradient ascent procedure similar to projected gradient descent (PGD) used in adversarial attack generation. Kumari et al. [11] review the theoretical foundations and empirical effectiveness of randomized smoothing, and discuss several challenges, including the curse of dimensionality, high computational cost, and large sample requirements for certification.

2.2. Randomized smoothing for regression. RS has been adapted for regression tasks. Applying RS to regression problems is more challenging because the model outputs are continuous rather than discrete. As a result, RS-based methods for certified defenses in regression typically fix the allowable strength of input perturbations and then determine the corresponding restriction in the model's output space, or vice versa.

The main idea behind RS in regression remains the same: sample N noisy variants of an input and compute the model's predictions for these samples. An aggregation function is then applied to obtain the final prediction, followed by theoretical analysis to provide certification guarantees.

The most straightforward adaptation of RS to regression is to aggregate the model outputs using a mean function. This approach, called RS-Reg, was proposed by Rekavandi et al. [3]. They demonstrate the asymptotic properties of a simple averaging function in settings where the regression model is unconstrained. Furthermore, they derive a certified upper bound on input perturbations for a family of regression models with bounded outputs.

Rekavandi et al. [12] further extend this basic approach by using tools from robust statistics, specifically the α-trimming filter. In this method, the aggregation function first sorts the predicted values corresponding to the noised inputs, removes the lowest and highest $\alpha\%$ of them, and then computes the mean of the remaining values. By adjusting the hyperparameter α, one can achieve a smooth trade-off between certified robustness and model utility.

Chiang et al. [6] propose median smoothing to enable certified regression in cases where standard mean smoothing fails. This aggregation function is more robust to outliers and can provide tighter certification bounds. The authors demonstrate that it yields considerably stronger guarantees for regression networks, such as object detectors. To obtain these certifications, they use a theory based on percentile-smoothed functions.

Recently, median smoothing has been adapted specifically for image quality assessment (IQA) tasks. Shumitskaya et al. [4] developed a method combining median smoothing with an additional convolutional denoiser and a ranking loss to improve the SROCC and PLCC scores of the defended IQA model. Their results show that this pipeline can serve as a robust and accurate quality estimator, though requiring substantial computational resources.

§3. Proposed method

We consider an adversarial setting where the base predictor is a regressor $f : R^d \to R$ that maps an image $x \in [0,1]^d$ to a scalar quality score. The defended model under attack is a randomized-smoothed variant based on median smoothing [6]: given a Gaussian noise distribution $\eta \sim N(0, \sigma^2 I)$,

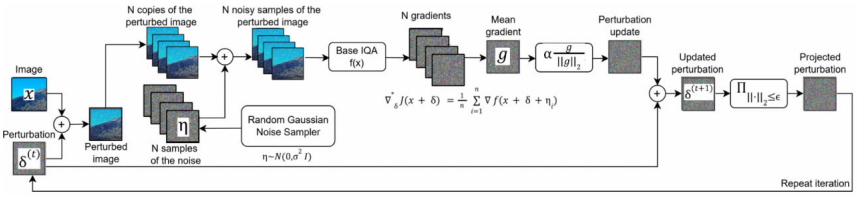

Figure 2. Overview of the proposed adaptive attack on median-smoothed regression models. Gradients are estimated by averaging over multiple noisy samples (EOT) to update the perturbation within a l_2 constraint.

the smoothed predictor $g_\sigma : R^d \to R$ is defined by the sample median of f under additive noise:

$$g_\sigma(x) = \text{median}\{f(x+\eta) \mid \eta \sim N(0, \sigma^2 I)\} \tag{3}$$

Median smoothing also yields certified bounds $Q^l(x)$ and $Q^u(x)$ such that, for a given radius $\epsilon > 0$,

$$Q^l(x) \leqslant g(x+u) \leqslant Q^u(x) \quad \forall u : ||u||_2 \leqslant \epsilon. \tag{4}$$

In practice, the defender estimates $g_\sigma(x)$ and its certified bounds by Monte-Carlo sampling of noisy copies of the input.

Our objective is to construct an adversarial perturbation δ with $||\delta||_2 \leqslant \epsilon$ that causes the smoothed predictor to produce a score outside its certified interval (or to maximize the change in the predicted quality otherwise). Directly attacking g_σ is difficult because g_σ is randomized and, when defined via the median, generally non-differentiable. We therefore propose an adaptive attack that (i) optimizes a differentiable surrogate objective formed by averaging over noisy copies (Expectation over Transformations, EOT) and (ii) uses an empirical gradient estimator obtained by averaging gradients computed on multiple noisy samples per attack step. Figure 2 gives an overview of the attack pipeline. The following subsections describe the method and its implementation details.

3.1. Attacker objective and surrogate. Instead of attempting to differentiate the sample median directly, we use a smooth surrogate objective based on the expectation of the base model under additive noise:

$$J(x+\delta) = E_{\eta\sim N(0,\sigma^2 I)}[f(x+\sigma+\eta)] \tag{5}$$

This expectation is differentiable under mild assumptions on f, such as measurability and finite expectation under Gaussian noise, which are typically satisfied by neural networks. The gradient of J can be estimated using Monte Carlo sampling. Although the median and the expectation are different statistics, perturbing J effectively searches directions that change the distribution of f under noise and therefore often leads to changes in the median as well. In practice, we also evaluate attack success with respect to the median g.

The attacker solves the constrained optimization:

$$\max_{||\delta||_2 \leqslant \epsilon} J(x+\delta). \tag{6}$$

3.2. Monte Carlo gradient estimator (EOT). The gradient of J is calculated as:

$$\nabla_\delta J(x+\delta) = E_{\eta\sim N(0,\sigma^2 I)}[\nabla f(x+\delta+\eta)] \tag{7}$$

We estimate this expectation by drawing n independent noise samples $\{\eta_i\}_{i=1\ldots n}$ and computing the following expression:

$$\nabla^*_\delta J(x+\delta) = \frac{1}{n}\sum_{i=1}^{n} \nabla f(x+\delta+\eta_i) \tag{8}$$

This estimator is unbiased for the gradient of the expectation and its variance decreases as $O(\frac{1}{n})$. In implementation, we compute $\nabla f(\cdot)$ using standard backpropagation through the base model.

3.3. Projected gradient iteration. Given the gradient estimator $\nabla^*_\delta J$, we perform an iterative projected gradient ascent (PGD) in the l_2-norm ball. Let $\delta^{(t)}$ denote the perturbation at iteration t. The update for the next iterate is

$$g = \nabla^*_\delta J(x+\delta^{(t)}), \tag{9}$$

$$\delta^{(t+1)} = \delta^{(t)} + \alpha\frac{g}{||g||_2}, \tag{10}$$

$$\delta^{(t+1)}_\Pi = \Pi_{||\cdot||_2\leqslant\epsilon}(\delta^{(t+1)}), \tag{11}$$

where $\alpha > 0$ is the step size and $\Pi_{||\cdot||_2 \leqslant \epsilon}$ is the projection onto the l_2-norm ball of radius ϵ. Finally, the adversarial image on t iteration is

$$x_{adv}^{(t+1)} = \text{clip}(x + \delta_{\Pi}^{(t+1)}, 0, 1). \tag{12}$$

In practice, the inner gradient estimator $\nabla_{\delta}^{*} J$ is computed by accumulating per-sample gradients on n noisy copies of the current perturbed image $x + \delta^{(t)}$ and averaging them.

§4. Experiments

4.1. Quality metrics. To evaluate the effectiveness of the attacks, we introduce two quantitative metrics: Adversarial Gain (AdvG) and Certified Attack Success Rate (cASR).

Adversarial Gain (AdvG) measures the average change in the IQA model's score caused by the attack. It is defined as

$$AdvG = \frac{100}{nR_Q} \sum_{i=1}^{n} (Q_i^a - Q_i) \tag{13}$$

where n is the number of test images, Q_i denotes the IQA model's score for the clean image i, Q_i^a denotes the score for the corresponding adversarial image, and R_Q is the total range of possible IQA scores. In other words, this metric quantifies the average shift in the IQA model's predictions induced by the attack.

The second metric, **Certified Attack Success Rate (cASR)**, measures the proportion of adversarial examples for which the IQA score deviates beyond the certified bounds obtained via median smoothing. It is defined as

$$cASR = \frac{100}{n} \sum_{i=1}^{n} I[(Q_i^a > Q_i^u) \vee (Q_i^a < Q_i^l)] \tag{14}$$

where Q_i^u and Q_i^l are the certified upper and lower bounds on the clean image score Q_i, such that for any perturbation u satisfying $||u||_2 < \epsilon$, the following holds:

$$Q_i^l \leqslant Q(x + u) \leqslant Q_i^u, \tag{15}$$

$$Q_i^l = \sup_{y \in R} \{P_{\eta \sim N(0, \sigma^2 I)}[f(x + \eta) \leqslant y] \leqslant p^l\} \tag{16}$$

$$Q_i^u = \inf_{y \in R} \{P_{\eta \sim N(0,\sigma^2 I)}[f(x+\eta) \leqslant y] \geqslant p^u\} \tag{17}$$

Here, Q_i^a denotes the IQA score of the attacked image, and the perturbation applied by the attack, u^a, satisfies $||u^a||_2 < \epsilon$. The function $I[\cdot]$ is the indicator function, which equals 1 if the condition inside the brackets is true (i.e., the adversarial score exceeds the certified bounds) and 0 otherwise, $p^l = \Phi(-\frac{\epsilon}{\sigma_{\text{smooth}}})$, $p^u = \Phi(\frac{\epsilon}{\sigma_{\text{smooth}}})$ and $\Phi(\cdot)$ is the Gaussian cumulative density function. Thus, cASR represents the percentage of adversarial examples that cause the IQA model's output to fall outside its certified robustness interval.

4.2. Experiment setup. All experiments use a random subset of 1,000 images from the KonIQ-10k dataset [13]. As IQA models, we select KonCept [13], PaQ-2-PiQ [14], and HyperIQA [15], since these models achieve high Spearman Rank-Order Correlation Coefficients (SRCC) with human subjective quality scores on the KonIQ-10k dataset. We evaluate the proposed attack with three types of experiments: (1) a parameter sweeps to identify when median smoothing certification fails, (2) an attack cost vs. success trade-off study to find efficient attack parameters, and (3) a comparative evaluation against three baseline attacks. For every experiment we report Adversarial Gain (AdvG) and Certified Attack Success Rate (cASR) as defined in Section 3.1.

4.2.1. *Common attack configuration.* Unless stated otherwise, the proposed attack uses the following default parameters:

- mean-gradient sample count $n_{\text{att}} = 16$,
- Gaussian noise standard deviation for attack generation $\sigma_{\text{att}} = 4/255$,
- l_2 budget $\epsilon = 4/255$,
- total iterations $T = 10$,
- step size $\alpha = \sigma_{\text{att}}/T$.

All perturbation norms are measured in the l_2 norm.

4.2.2. *Search for median-smoothing failure modes.* **Goal.** Our goal is to identify combinations of median smoothing parameters for which the certified bounds no longer protect the IQA model from the proposed adversarial examples.

Model. We use the KonCept IQA model [13] because it achieves high performance on KonIQ and was trained specifically on this dataset.

Procedure.

- Generate adversarial examples with the default attack configuration given above.
- For each adversarial example, apply median smoothing with a grid of smoothing parameters: number of smoothing samples $n_{\text{smooth}} \in \{1, 4, 16, 64, 256, 1024, 2048\}$, smoothing noise $\sigma_{\text{smooth}} \in \{2/255, 4/255, 6/255, 8/255, 10/255\}$.
- For each $(n_{\text{smooth}}, \sigma_{\text{smooth}})$ pair measure AdvG and cASR.

Outcome. The result is identified parameter regions where median smoothing fails (high AdvG and/or high cASR), revealing combinations for which the certified bounds are violated by our adversarial examples.

4.2.3. *Attack cost vs. success trade-off.* **Goal.** Quantify how attack success depends on computational cost (number of samples used in gradient estimation and number of iterations), and identify efficient operating points.

Model. We use the PaQ-2-PiQ model [14] for this experiment because it is the fastest IQA model in our set, which reduces runtime for parameter sweeps.

Procedure.

- Fix smoothing parameter to isolate attack-cost effects at: $n_{\text{smooth}} = 1024$ samples and $\sigma_{\text{smooth}} = 4/255$.
- Generate adversarial examples with varying attack parameters: mean-gradient sample count $n_{\text{att}} \in \{8, 16, 32\}$ and total iterations $T \in \{5, 10, 20\}$. All other attack settings are the defaults listed in the common configuration.
- For each (n_{att}, T) pair compute AdvG and cASR.

Outcome. Produce heatmaps showing trade-offs between runtime cost (proportional to $n_{\text{att}} \times T$) and attack effectiveness, and recommend efficient (n_{att}, T) combinations.

4.2.4. *Comparison with baseline methods.* **Goal.** Demonstrate the effectiveness of the proposed method relative to simple and established baselines:

- Random noise: Add zero-mean Gaussian noise with l_2 scaling so that the perturbation norm is $\epsilon = 4/255$. This baseline tests whether random perturbation alone can (a) increase IQA score or (b) break median smoothing certification.

- FGM [16]: Apply a Fast Gradient Sign Method attack on the base IQA model (no smoothing) to produce adversarial examples, under the same l_2 budget. The rationale is to test whether adversarial perturbations crafted for the base model transfer to the median-smoothed model.
- PGD [17]: Apply a Projected Gradient Descent attack on the base IQA model with the same l_2 budget and number of iterations as our attack.

Attack hyperparameters for comparison.

- l_2 budget $\epsilon = 4/255$ for all attack methods,
- number of iterations $= 10$ for both PGD and our method (FGM is single-step),
- all other settings for our method are the same as in Section 4.2.1.

Procedure. For each method, generate adversarial examples and evaluate AdvG and cASR under the median smoothing configurations with $n_{\text{smooth}} = 1024$ samples and $\sigma_{\text{smooth}} = 4/255$.

Outcome. Report comparative AdvG and cASR. Higher values indicate a stronger attack; we use these metrics to argue for the superiority of the proposed approach.

§5. Results

All reported metrics are averaged over the 1,000 images randomly selected from the KonIQ-10k dataset. The evaluation uses the two metrics defined in Section 4.1 – Adversarial Gain (AdvG) and Certified Attack Success Rate (cASR) – which respectively measure the magnitude of score deviation and the fraction of adversarial examples that violate certified robustness bounds.

The first set of experiments investigates the conditions under which median smoothing certification fails. Figure 3 presents the results obtained using the KonCept IQA model with a fixed attack configuration while varying the parameters of median smoothing. When the number of smoothing samples is small (e.g., $n_{\text{smooth}} = 1$ or $n_{\text{smooth}} = 4$), the proposed attack is highly effective, reaching up to 100% cASR and demonstrating a large shift in IQA scores (high AdvG). As the number of samples increases, both metrics gradually decrease, indicating that a larger sample count provides better robustness against adversarial perturbations. A similar trend is observed for the smoothing noise parameter σ_{smooth}: higher noise levels

Attack method	KonCept [13]		PaQ-2-PiQ [14]		HyperIQA [15]	
	AdvG↑	cASR↑	AdvG↑	cASR↑	AdvG↑	cASR↑
Random Noise	0.01%	0.0%	0.02%	0.0%	0.00%	0.0%
FGM [16]	0.53%	25.0%	0.60%	69.4%	0.36%	12.0%
PGD [17]	0.54%	25.4%	0.60%	70.8%	0.36%	12.6%
EOT-PGD (ours)	**0.93%**	**74.8%**	**0.88%**	**94.5%**	**0.82%**	**87.0%**

Table 1. Results of comparison with previous attacks when defending using Median Smoothing with number samples $n_{\text{smooth}} = 1024, \sigma_{\text{smooth}} = 6/255$. All attacks were run to maintain l_2-norm= $4/255$.

generally improve defense effectiveness, although excessive noise can degrade the base model's accuracy on clean images. In practice, it is most efficient to set slightly higher than the attacker's noise level, as this provides a good balance between clean-image accuracy and defense strength. Figure 4 shows image degradation resulting from varying σ_{smooth}.

The second experiment explores the trade-off between attack efficiency and computational cost. In this case, we used the PaQ-2-PiQ model and fixed the median smoothing parameters ($n_{\text{smooth}} = 1024, \sigma_{\text{smooth}} = 4/255$) while varying the attack parameters – the number of samples used for gradient estimation ($n_{\text{att}} = 1, 8, 16, 32$) and the total number of iterations ($T = 1, 5, 10, 20$). The results, summarized in Figure 5, show that increasing either parameter leads to higher AdvG and cASR values, but with diminishing returns. Beyond $n_{\text{att}} = 16$ and $T = 10$, the improvement in attack success becomes negligible, while computational cost grows substantially. Therefore, these values represent an efficient operating point that balances attack success and computational efficiency.

Finally, Table 1 presents a comparison between the proposed adaptive attack and three baseline approaches: random noise, FGM, and PGD. Random noise perturbations produce negligible changes in both metrics, confirming that unstructured noise cannot effectively break median smoothing. FGM and PGD attacks, although designed for the base IQA models without smoothing, exhibit a notable transfer effect and achieve moderate AdvG and cASR values when applied to median-smoothed models. PGD performs slightly better than FGM due to its iterative nature. The proposed adaptive attack, however, consistently achieves the highest AdvG

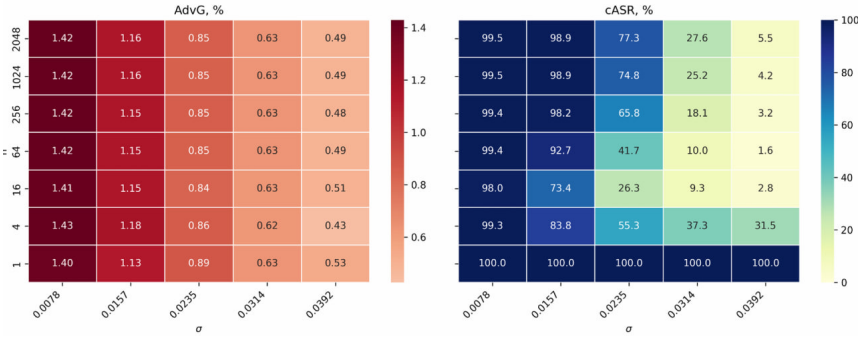

Figure 3. Results of experiment for the search for median-smoothing failure modes. The attack parameters are fixed, while the median-smoothing parameters $(n_{\text{smooth}}, \sigma_{\text{smooth}})$ vary. We can see that in overall the more samples is better, and also higher noise level sigma is better for defense against the proposed adaptive attack.

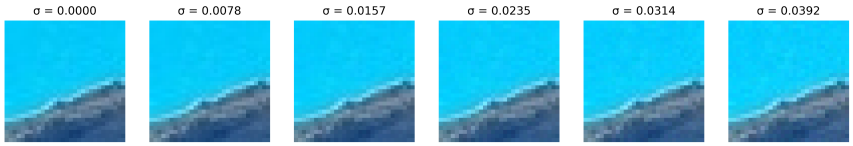

Figure 4. Visualization of image degradation resulting from varying σ_{smooth}. Increasing σ_{smooth} leads to stronger perturbations and a corresponding loss of image detail.

and cASR values across all tested IQA models. This demonstrates that averaging gradients across multiple noisy samples enables the attack to approximate the behavior of the randomized defense better and to find more effective perturbation directions.

FGM and PGD can be viewed as special cases of our adaptive attack: when the number of samples for mean-gradient estimation is set to one and the attack noise is zero, FGM corresponds to a single-step version $T = 1$, while PGD represents a multi-step version $T = 10$. This relationship shows that our method generalizes these classical approaches by averaging gradients across multiple noisy samples, thereby stabilizing optimization and improving transferability under randomized smoothing.

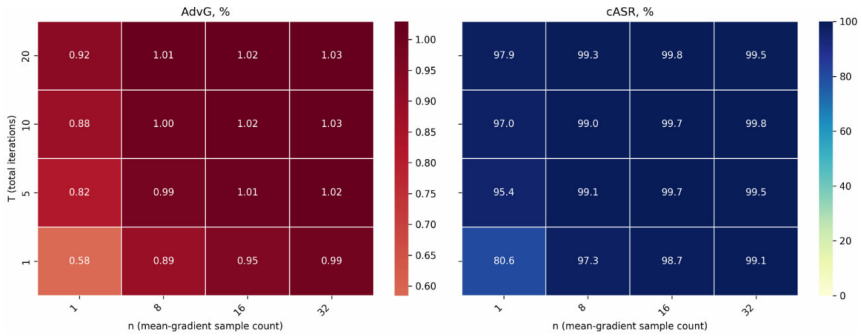

Figure 5. Results of experiment to identify efficient operating points in Attack cost vs. success trade-off. Parameters of median smoothing are fixed, and parameters of attack (n_{att}, T) are varied. We can see that the more n_{att} and T are better; however, after $n_{\text{att}} = 16$ and $T = 10$, increasing parameters does not matter significantly.

In summary, the results demonstrate that the proposed adaptive attack effectively compromises median-smoothed IQA models through noise-averaged gradient estimation. At the same time, the experiments provide practical insights for defense design: increasing the number of smoothing samples and carefully adjusting the smoothing noise level substantially improves robustness, whereas excessive noise can degrade performance on clean images. Overall, the proposed adaptive attack establishes a stronger and more realistic threat model for assessing the robustness of IQA systems employing randomized median smoothing.

§6. Conclusion

This work provides the first systematic analysis of the robustness of median-smoothed IQA models under adaptive adversarial attacks. While randomized smoothing offers formal robustness guarantees, our findings reveal that these guarantees can be compromised in practice due to the finite-sample nature of Monte Carlo estimation. We proposed an adaptive noise-averaged gradient attack, which extends PGD with Expectation over Transformation (EOT) and accounts for stochastic variability during optimization.

Extensive experiments on three state-of-the-art IQA models – KonCept, PaQ-2-PiQ, and HyperIQA – demonstrate that the proposed method consistently achieves higher Adversarial Gain (AdvG) and Certified Attack Success Rate (cASR) compared to existing baselines, including FGM and PGD. We also showed that FGM and PGD can be interpreted as special cases of our adaptive formulation, further underscoring its generality. Importantly, our analysis revealed that median smoothing provides meaningful robustness only when a sufficiently large number of noise samples and an adequately tuned noise level are used; otherwise, its certification guarantees can fail under adaptive attack.

Overall, these results highlight a significant gap between theoretical and practical robustness in median-smoothed regression models. The proposed adaptive attack thus defines a stronger and more realistic threat model for evaluating the security of IQA systems. Future work will focus on developing improved randomized smoothing schemes that maintain robustness under finite-sample constraints and exploring certified defenses specifically tailored to non-differentiable regression tasks.

References

1. J. Cohen, E. Rosenfeld, and Z. Kolter, *Certified Adversarial Robustness via Randomized Smoothing*, in: Proceedings of the International Conference on Machine Learning (ICML), 2019, pp. 1310–1320.
2. G. Yang, T. Duan, J. E. Hu, H. Salman, I. Razenshteyn, and J. Li, *Randomized Smoothing of All Shapes and Sizes*, in: Proceedings of the International Conference on Machine Learning (ICML), 2020, pp. 10693–10705.
3. A. M. Rekavandi, O. Ohrimenko, and B. I. Rubinstein, *RS-Reg: Probabilistic and Robust Certified Regression Through Randomized Smoothing*, Transactions on Machine Learning Research (2023).
4. E. Shumitskaya, M. Pautov, D. Vatolin, and A. Antsiferova, *Stochastic BIQA: Median Randomized Smoothing for Certified Blind Image Quality Assessment*, Computer Vision and Image Understanding **259** (2025), 104447.
5. M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, *Certified Robustness to Adversarial Examples with Differential Privacy*, in: Proceedings of the IEEE Symposium on Security and Privacy (SP), 2019, pp. 656–672.
6. P. Y. Chiang, M. Curry, A. Abdelkader, A. Kumar, J. Dickerson, and T. Goldstein, *Detection as Regression: Certified Object Detection with Median Smoothing*, Advances in Neural Information Processing Systems **33** (2020), 1275–1286.
7. T. Maho, T. Furon, and E. Le Merrer, *Randomized Smoothing Under Attack: How Good Is It in Practice?*, in: ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 3014–3018.

8. *Exploiting Stochastic Weaknesses of Randomized Smoothing in Regression Models: Adaptive Attack Analysis and Empirical Limitations* (2026). Available: `https://github.com/katiashh/adaptive_attacks_rs`.
9. B. Delattre, P. Caillon, Q. Barthélemy, E. Fagnou, and A. Allauzen, *Bridging the Theoretical Gap in Randomized Smoothing*, in: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2025, pp. 3997–4005.
10. Y. Y. Yang, C. Rashtchian, H. Zhang, R. R. Salakhutdinov, and K. Chaudhuri, *A Closer Look at Accuracy vs. Robustness*, Advances in Neural Information Processing Systems **33** (2020), 8588–8601.
11. S. Jindal, D. Bhardwaj, and A. Kumari, *Rethinking Randomized Smoothing from the Perspective of Scalability*, in: Proceedings of the Third Workshop on New Frontiers in Adversarial Machine Learning, 2024.
12. A. Rekavandi, F. Farokhi, O. Ohrimenko, and B. Rubinstein, *Certified Adversarial Robustness via Randomized α-Smoothing for Regression Models*, Advances in Neural Information Processing Systems **37** (2024), 134127–134150.
13. V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, *KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment*, IEEE Transactions on Image Processing **29** (2020), 4041–4056.
14. Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, *From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality*, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3575–3585.
15. S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, *Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network*, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3667–3676.
16. I. J. Goodfellow, J. Shlens, and C. Szegedy, *Explaining and Harnessing Adversarial Examples*, in: Proceedings of the International Conference on Learning Representations (ICLR), 2015.
17. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, *Towards Deep Learning Models Resistant to Adversarial Attacks*, in: Proceedings of the International Conference on Learning Representations (ICLR), 2018.

Поступило 4 февраля 2026 г.

Trusted AI Research Center, RAS,
109004 Moscow, Russia
E-mail: `ekaterina.shumitskaya@graphics.cs.msu.ru`

MSU Institute for Artificial Intelligence,
119991 Moscow, Russia
E-mail: `dmitriy@graphics.cs.msu.ru`

Lomonosov Moscow State University,
119991 Moscow, Russia
E-mail: `aantsiferova@graphics.cs.msu.ru`