# CONTRS: GENERALIZING RANDOMIZED SMOOTHING TO CONTINUOUS OUTPUTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Randomized smoothing certifies adversarial robustness by aggregating a base model's predictions under injected noise. In classification, its classical formulation compares class probabilities under Gaussian perturbations to certify exact label invariance. For continuous outputs, however, model predictions can remain close without being identical, making a quantitative bound on prediction change more appropriate. Existing extensions typically rely on task-specific constructions and do not directly map a prescribed output tolerance to a certified input radius across different output types. To this end, we introduce **ContRS**, a new unified framework that defines the smoothed prediction as the center of an output neighborhood with maximum probability mass and recovers classical randomized smoothing as a special case. At the population level, we show that a probability margin over competing neighborhoods yields a certified $\ell_2$ radius within which the smoothed prediction changes by at most the prescribed tolerance. For metric and symmetric relaxed-triangle discrepancies, we derive finite-sample certificates using a center selected independently of the certification samples. For general discrepancies, we provide empirical robustness estimates based on sampled competitors. Across diverse continuous-output tasks, ContRS yields larger certified input radii through output-region selection at matched output-region scales and tighter output bounds at matched input budgets, with lower or comparable smoothing error compared with representative baselines.

## 1 INTRODUCTION

Adversarial vulnerability remains a central concern in modern machine learning: even small and carefully crafted input perturbations can induce substantial changes in the outputs of deep learning models Madry et al. (2017); Silva & Najafirad (2020); Kireev et al. (2022). Such vulnerabilities arise across tasks with diverse output structures Calivá et al. (2020) and are especially concerning in high-stakes applications, including face recognition Yang et al. (2021), autonomous driving Croce et al. (2024), and medical imaging MR et al. (2025). Although empirical defenses can improve robustness against specific attacks, they provide no formal guarantee and may fail under stronger or adaptive attacks Bai et al. (2021). **Certified Robustness** instead seeks to prove that a prescribed robustness property holds for every input perturbation within a specified perturbation set, independently of any particular attack algorithm. In classification, this property is typically label invariance, and the resulting certificate specifies an input region within which the predicted label is guaranteed to remain unchanged. Among existing certification techniques, **Randomized Smoothing** Lecuyer et al. (2018); Li et al. (2019); Cohen et al. (2019) is particularly attractive because it constructs a certifiably robust predictor from a black-box base model, requires no modification to the base architecture, and scales to high-dimensional inputs. These properties have made it one of the most widely used frameworks for certified robustness.

Classification remains the predominant application of randomized smoothing because discrete labels admit a natural robustness criterion: invariance of the predicted label over a bounded set of input perturbations. However, *this exact-invariance criterion does not extend naturally to other output types*, as exact output equality is generally unrealistic or overly restrictive for continuous predictions. This mismatch motivates extensions of randomized smoothing beyond classification, but existing approaches remain limited in three respects.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- **Task Dependence.** Median smoothing Chiang et al. (2020), SegCertify Fischer et al. (2021), and randomized $\alpha$-smoothing Miri Rekavandi et al. (2024) respectively adapt randomized smoothing to object detection, segmentation, and regression tasks. However, their reliance on *task-specific* representations or decomposition strategies limits each method to the output type for which it is designed, preventing straightforward generalization to other output spaces.
- **Direction Reversal.** Center smoothing Kumar & Goldstein (2021) defines the smoothed prediction as a center of the output distribution induced by randomized inputs. For a prescribed input perturbation budget, it certifies an upper bound on the resulting output deviation. This input-to-output formulation *reverses* the conventional certification direction of randomized smoothing. As discussed by Shao et al. (2023), adapting this formulation may require repeated evaluations to identify an output range that satisfies application requirements, and the use of the triangle inequality can also lead to conservative output bounds.
- **Indirect Guarantee.** Existing methods formulate robustness in terms of output properties, including coordinate-wise bounds Chiang et al. (2020), component-wise label invariance Fischer et al. (2021), probabilistic coordinate tolerance Miri Rekavandi et al. (2024), and membership in a predefined output region Shao et al. (2023). These methods therefore provide only *indirect* guarantees on prediction change, rather than explicit bounds on the deviation between the original and perturbed smoothed predictions.

Overall, existing methods do not establish an effective mapping from a prescribed scale of output variation to the certified input radius. While classification focuses on label invariance, continuous outputs raise a quantitative question: *given an acceptable output deviation, how large an input perturbation can still be certified as safe?* Addressing this critical question requires translating a bound on output variation into a probabilistic condition estimable from random samples. In classification, randomized smoothing uses the top two class probabilities to certify label invariance. For continuous outputs, however, Gaussian perturbations of the same input typically yield distinct predictions even when they differ only slightly, making the probability of any exact output uninformative about prediction stability. Moreover, certification must account for all outputs whose deviation from the original smoothed prediction exceeds the bound, and this potentially infinite set cannot be exhaustively enumerated.

To this end, we propose **ContRS**, a new unified framework that extends randomized smoothing from discrete labels to continuous outputs. Given a prescribed output tolerance $r$, ContRS first defines the smoothed prediction as the center of a radius-$r/2$ region that maximizes the probability mass of model outputs under Gaussian input perturbations. It then compares this mass with the largest mass captured by any radius-$r/2$ region centered more than $r$ away from the selected prediction. We derive rigorous finite-sample certificates for metrics and symmetric discrepancies satisfying a relaxed triangle inequality. For general measurable discrepancies lacking a suitable geometric bound, we provide empirical robustness estimates based on sampled competitors. In general, our main contributions are as follows. (1) **Generalizable Framework**. ContRS provides a unified certification framework for continuous outputs. This task-agnostic formulation directly maps a prescribed bound on prediction change to a certified input radius and recovers classical randomized smoothing for discrete labels equipped with the discrete distance. (2) **Theoretical Guarantee**. We establish a new robustness guarantee: if the selected radius-$r/2$ output region has greater probability mass than every competing radius-$r/2$ region centered more than $r$ away, then this probability advantage yields a certified $\ell_2$ input radius within which the smoothed prediction changes by at most $r$. (3) **Empirical Evaluation.** ContRS is evaluated on continuous-output tasks using metric and non-metric discrepancies. Compared with representative baselines, ContRS's selected output balls yield larger certified input radii at matched output-region scales, while its output bounds are tighter at matched input budgets, with lower or comparable smoothing error. We also jointly evaluate clean-task performance and certified output stability, and report average output deviations under adversarial attacks.

## 2 Related Work

**Randomized Smoothing for Classification.** Classification remains the predominant application of randomized smoothing, which has become a standard framework for certifying adversarial robustness Cohen et al. (2019). Early studies establish probabilistic robustness guarantees Lecuyer

| Method | Output Setting / Discrepancy | Certified Property | Certification Direction |
|---|---|---|---|
| median smoothing | Coordinate-valued outputs | $L_i(x,R) \le g_i(x+\delta) \le U_i(x,R)$ | $R \longrightarrow \{[L_i,U_i]\}_i$ |
| SegCertify | Segmentation outputs | $g_j(x+\delta) = g_j(x), \quad j \in \widehat{\mathcal{I}}_x$ | $\tau \longrightarrow (R, \widehat{\mathcal{I}}_x)$ |
| randomized $\alpha$-smoothing | Vector-valued outputs | $\Pr(\lvert g_{\alpha,i}(x+\delta) - y_i\rvert \le r_i) \ge P,\ \forall i$ | $(\{r_i\}_i, P) \longrightarrow R$ |
| center smoothing | Metric, pseudometric, or symmetric relaxed-triangle $d$ | $d(g(x+\delta), g(x)) \le \rho$ | $R \longrightarrow \rho$ |
| safe region | Semimetric outputs | $g_{\mathcal{S}}(x+\delta) = 1$ | $\mathcal{S} \longrightarrow R$ |
| **ContRS (ours)** | General outputs; nonnegative measurable $d$ | $d(g(x+\delta), g(x)) \le r$ | $r \longrightarrow R$ |

Table 1: Comparison of randomized-smoothing methods beyond classification Chiang et al. (2020); Fischer et al. (2021); Miri Rekavandi et al. (2024); Kumar & Goldstein (2021); Shao et al. (2023). Throughout, $\|\delta\|_2 < R$, where $R$ is the certified input radius and $g$ is the corresponding smoothed predictor. For median smoothing, $[L_i, U_i]$ bounds output coordinate $i$. For SegCertify, $\tau$ fixes $R = \sigma\Phi^{-1}(\tau)$, and $\widehat{\mathcal{I}}_x$ is the certified non-abstained set. For randomized $\alpha$-smoothing, $y_i$ and $r_i$ are the target and tolerance for coordinate $i$, and $P$ is the prescribed success probability. For center smoothing, $\rho$ is the certified output bound for $R$. For safe region, $g_{\mathcal{S}}$ is the binary smoothed classifier for region $\mathcal{S}$. For ContRS, $r$ is the prescribed output tolerance, and its guarantee for general discrepancies is stated with exact neighborhood probabilities.

et al. (2018); Li et al. (2019), while Cohen et al. (2019) derive a tight $\ell_2$ certificate for Gaussian smoothing based on the probability gap between the most likely and runner-up classes. Subsequent work explores alternative perturbation models and noise distributions Yang et al. (2020); Vorácek & Hein (2023), and develops input-specific or discrete-structure certificates Wang et al. (2021); Chen et al. (2022). Despite these advances, classification-oriented methods share a common certification objective: guaranteeing invariance of the predicted class over a certified perturbation set. But this label-invariance formulation does not extend to continuous outputs.

**Randomized Smoothing Beyond Classification.** Table 1 compares existing extensions of randomized smoothing beyond classification in terms of output settings, certified properties, and certification directions. Median smoothing Chiang et al. (2020) provides coordinate-wise percentile bounds for outputs and applies them to object detection. SegCertify Fischer et al. (2021) certifies the labels of individual pixels or points while abstaining on unstable predictions. Randomized $\alpha$-smoothing Miri Rekavandi et al. (2024) guarantees, with a prescribed probability, that each coordinate of a randomized regression prediction lies within a specified tolerance interval. For more general output types, center smoothing Kumar & Goldstein (2021) bounds the output deviation for a given input perturbation budget, whereas safe region Shao et al. (2023) derives an input radius that preserves membership in a chosen output region. Different from above methods, ContRS starts from a prescribed output tolerance and derives an input radius that directly bounds the change between smoothed predictions before and after perturbation. It applies to diverse output types within a unified framework and shows favorable results in the evaluated comparisons with strong baselines.

## 3 PRELIMINARY

In this section, we briefly review the standard randomized smoothing Lecuyer et al. (2018); Cohen et al. (2019). The central idea is to construct a smoothed classifier from a given base classifier such that the resulting classifier admits a certified robustness guarantee against adversarial perturbations. Consider a base classifier $f : \mathbb{R}^n \to \mathcal{Y}$, where $\mathcal{Y}$ is a discrete label space. Directly certifying the robustness of $f$ is generally challenging because modern deep neural networks are typically highly nonlinear and complex Wong & Kolter (2018); Gowal et al. (2018); Zhang et al. (2018a). Randomized smoothing addresses this challenge by injecting Gaussian noise into the input. Specifically, for an input $x \in \mathbb{R}^n$, let $X \sim \mathcal{N}(x, \sigma^2 I)$. The smoothed classifier is defined as

$$g(x) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}\big(f(X) = c\big). \tag{1}$$

Rather than relying on the prediction of the base classifier at a single input point, the smoothed classifier aggregates its predictions under Gaussian perturbations around $x$. This aggregation enables the formal robustness certification based on class probabilities under Gaussian noise.

Let $c_A$ and $c_B$ denote the most probable and second-most probable classes under the noisy distribution, with probabilities $p_A = \mathbb{P}(f(X) = c_A)$ and $p_B = \mathbb{P}(f(X) = c_B)$, respectively. Randomized

smoothing gives the input certified radius:

$$R = \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right), \tag{2}$$

where $\Phi^{-1}$ denotes the inverse cumulative distribution function of the Gaussian distribution. For every adversarial perturbation $\delta$ satisfying $\|\delta\|_2 < R$, the prediction of the smoothed classifier remains unchanged: $g(x+\delta) = c_A$. This guarantee establishes an exact *label invariance* within a certified $\ell_2$ input ball. Moreover, randomized smoothing is model-agnostic because it can be applied to any base classifier without modifying its model architecture. Despite its theoretical elegance and flexibility, randomized smoothing primarily targets classification with discrete labels. In many applications, however, predictions lie in continuous or more general output spaces, where exact output invariance is often unrealistic or overly restrictive. This limitation motivates us to generalize randomized smoothing from discrete labels to continuous outputs, deriving a certified input radius that guarantees a prescribed bound on prediction change across different output types. The notations used in this paper are summarized in Appendix A.

# 4 METHODOLOGY

In this paper, we propose **ContRS**, which generalizes the probability-based certification mechanism of randomized smoothing from discrete labels to continuous outputs. In classification, each class label defines an event, and certification relies on comparing class probabilities. For continuous outputs, Gaussian perturbations of the same input typically produce distinct predictions, so any exact output may have zero or negligible probability mass and provide little information about prediction stability. ContRS instead aggregates probability mass within *output neighborhoods* defined by a prescribed tolerance and compares the resulting neighborhood probabilities for certification.

Let $f : \mathbb{R}^n \to \mathcal{M}$ be a base predictor, and let $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_{\geq 0}$ be a measurable function that quantifies the discrepancy between two outputs.[1] For a prescribed output tolerance $r > 0$, let $\mathcal{B}(z, r/2) = \{y \in \mathcal{M} : d(z, y) \leq r/2\}$ denote the output neighborhood of radius $r/2$ centered at $z$. Under Gaussian input noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, the smoothed predictor is defined as

$$g(x) \in \arg\max_{z \in \mathcal{M}} \mathbb{P}\left(f(x+\epsilon) \in \mathcal{B}(z, r/2)\right). \tag{3}$$

We assume that the maximum is attained and use a fixed tie-breaking rule when necessary. Each candidate output is thus evaluated by the probability mass of its local neighborhood rather than the probability of that exact output. Here, $r$ specifies the maximum output change to be certified, while $r/2$ determines the *neighborhood radius* used to aggregate nearby predictions.

This formulation recovers standard Gaussian randomized smoothing as a special case. If $\mathcal{M}$ is a finite label set equipped with the discrete distance $d(c, c') = \mathbf{1}\{c \neq c'\}$, then $\mathcal{B}(c, r/2) = \{c\}$ for any $0 < r < 1$. Consequently, Eq.(3) reduces to $g(x) \in \arg\max_{c \in \mathcal{M}} \mathbb{P}(f(x+\epsilon) = c)$, which is the standard Gaussian-smoothed classifier. ContRS thus generalizes classical randomized smoothing by replacing singleton label events with neighborhood events in continuous output spaces.

## 4.1 ROBUSTNESS GUARANTEE

Having defined the smoothed predictor, we now derive a **Certified Input Radius** within which its output changes by at most the prescribed tolerance $r$. Specifically, we compare the probability mass of the selected neighborhood with those of competing neighborhoods whose centers deviate from the original prediction by more than $r$. Let $z^* = g(x)$, and denote the probability mass of the selected neighborhood by $p = \mathbb{P}(f(x+\epsilon) \in \mathcal{B}(z^*, r/2))$. Define the competing mass as

$$q = \sup_{z:\, d(z, z^*) > r} \mathbb{P}\left(f(x+\epsilon) \in \mathcal{B}(z, r/2)\right). \tag{4}$$

Thus, $q$ is the largest mass of any neighborhood centered more than $r$ from $z^*$. It generalizes the runner-up class probability to a potentially infinite set of competing outputs.

[1] Here, we do not require $d$ to be a metric.

**Theorem 4.1** (Certified Input Radius)**.** *Suppose that $p > q$. Then, for any adversarial perturbation $\delta \in \mathbb{R}^n$ satisfying $\|\delta\|_2 < R$, where*

$$R = \frac{\sigma}{2}\left(\Phi^{-1}(p) - \Phi^{-1}(q)\right), \tag{5}$$

*the smoothed predictor satisfies $d(g(x+\delta), g(x)) \leq r$. Here, $\Phi$ denotes the standard Gaussian cumulative distribution function.*

The proof applies the probability bounds derived from the Lipschitz property of Gaussian smoothing (See Appendix B for details.) to the selected neighborhood and every competing neighborhood. Whenever $\|\delta\|_2 < R$, the lower bound for the selected neighborhood strictly exceeds the uniform upper bound for neighborhoods centered beyond the prescribed tolerance. Consequently, no center satisfying $d(z, z^*) > r$ can maximize the neighborhood probability at $x + \delta$. The complete proof is provided in Appendix C. For finite labels equipped with the discrete distance and $0 < r < 1$, $q$ reduces to the runner-up class probability, and the output guarantee becomes label invariance. Theorem 4.1 thus recovers the classical randomized smoothing certificate.

**Remark 4.1** (No Metric Assumption)**.** *The measurable function $d$ need not be a metric; neither symmetry nor the triangle inequality is required. Competing neighborhoods may overlap, since the above proof bounds their probability masses uniformly through the supremum $q$.*

When $d$ is a metric, the certificate admits a simpler form. The triangle inequality implies

$$d(z, z^*) > r \quad \Longrightarrow \quad \mathcal{B}(z, r/2) \cap \mathcal{B}(z^*, r/2) = \varnothing. \tag{6}$$

Every competing neighborhood therefore has probability mass at most $1 - p$, giving $q \leq 1 - p$. Whenever $p > 1/2$, this yields a conservative lower bound on $R$:

$$\underline{R} = \sigma\Phi^{-1}(p) \leq R. \tag{7}$$

This bound requires only $p$, enabling certification without evaluating $q$.

### 4.2 Finite-Sample Certification

The preceding result is stated in terms of population probabilities. In a metric output space, the competing mass satisfies $q \leq 1 - p$, so certification only requires a lower bound on the probability mass of the selected neighborhood. In practice, however, both the center and its neighborhood mass must be estimated from finitely many predictions under Gaussian input noise. We now develop a finite-sample certification and the corresponding Monte Carlo procedure.

Let $\hat{z}$ be an estimated center selected independently of the certification samples, and define its population neighborhood mass as $p_{\hat{z}} = \mathbb{P}(f(x+\epsilon) \in \mathcal{B}(\hat{z}, r/2))$. To estimate $p_{\hat{z}}$, we draw independent Gaussian noises $\epsilon_1, \ldots, \epsilon_{N_2}$ and count

$$k_{\text{cert}} = \sum_{j=1}^{N_2} \mathbf{1}\left\{f(x+\epsilon_j) \in \mathcal{B}(\hat{z}, r/2)\right\}. \tag{8}$$

Since $\hat{z}$ is selected independently of the certification samples, the certification count $k_{\text{cert}}$, conditional on the selected center, follows a binomial distribution with parameters $N_2$ and $p_{\hat{z}}$. We compute $\hat{p}_L$ from $(k_{\text{cert}}, N_2)$ using the one-sided Clopper–Pearson method Clopper & Pearson (1934). This guarantees that $p_{\hat{z}} \geq \hat{p}_L$ with probability at least $1 - \alpha$ over the certification samples.

**Corollary 4.2** (Finite-Sample Radius)**.** *Assume that $(\mathcal{M}, d)$ is a metric space. Let $\hat{z}$ be selected independently of the $N_2$ certification samples, and let $\hat{p}_L$ be the one-sided Clopper–Pearson lower bound computed from Eq. (8) with failure probability $\alpha$. With probability at least $1 - \alpha$, whenever $\hat{p}_L > 1/2$, the radius*

$$\hat{R} = \sigma\Phi^{-1}(\hat{p}_L) \tag{9}$$

*satisfies, for every $\|\delta\|_2 < \hat{R}$, $d\big(g(x+\delta), g(x)\big) \leq r$ and $d\big(g(x+\delta), \hat{z}\big) \leq r$.*

On the confidence event $p_{\hat{z}} \geq \hat{p}_L$, the maximizing center $g(x)$ has neighborhood mass at least $\hat{p}_L$ as well. Thus, the disjointness relation in Eq.(6) and the Gaussian probability-shift bounds apply with

**Algorithm 1** Monte Carlo Prediction and Certification for ContRS

**Input:** data sample $x$; base predictor $f$; a measurable symmetric discrepancy $d$ satisfying the $\kappa$-relaxed triangle inequality in Eq.(10); output tolerance $r$; Gaussian noise scale $\sigma$; known constant $\kappa$; candidate sample size $N_1$; selection sample size $N_{\mathrm{sel}}$; certification sample size $N_2$; failure probability $\alpha \in (0, 1)$.
**Output:** estimated center $\hat{z}$; certified radius $\hat{R}$; lower confidence bound $\hat{p}_L$.
$s \leftarrow r/(2\kappa)$
All noise samples below are mutually independent and follow $\mathcal{N}(0, \sigma^2 I_n)$.
Draw $y_i = f(x + \epsilon_i)$, $i = 1, \ldots, N_1$ *Stage 1: Candidate generation*
Draw $y'_j = f(x + \epsilon'_j)$, $j = 1, \ldots, N_{\mathrm{sel}}$ *Stage 2: Center selection*
**for** $i = 1, \ldots, N_1$ **do**
    $k_i \leftarrow \sum_{j=1}^{N_{\mathrm{sel}}} \mathbf{1}\{d(y_i, y'_j) \le s\}$
**end for**
$\hat{z} \leftarrow y_{i^*}$, where $i^* = \min\{i : k_i = \max_{1 \le \ell \le N_1} k_\ell\}$
Draw $y''_j = f(x + \epsilon''_j)$, $j = 1, \ldots, N_2$ *Stage 3: Probability certification*
$k_{\mathrm{cert}} \leftarrow \sum_{j=1}^{N_2} \mathbf{1}\{d(\hat{z}, y''_j) \le s\}$
$\hat{p}_L \leftarrow \mathrm{ClopperPearsonLower}(k_{\mathrm{cert}}, N_2, \alpha)$
$\hat{R} \leftarrow \sigma\Phi^{-1}(\hat{p}_L)$ **if** $\hat{p}_L > 1/2$; **otherwise**, $\hat{R} \leftarrow 0$
**Return:** $(\hat{z}, \hat{R}, \hat{p}_L)$

either $g(x)$ or $\hat{z}$ as the reference center. The same argument extends to a symmetric discrepancy satisfying

$$d(u, v) \le \kappa\big(d(u, w) + d(w, v)\big), \qquad u, v, w \in \mathcal{M}, \tag{10}$$

for a known $\kappa \ge 1$. Set $s = r/(2\kappa)$ and let $g_s$ be defined as in Eq.(3), with $r/2$ replaced by $s$. Radius-$s$ neighborhoods centered more than $r$ apart are disjoint. Using these neighborhoods for both center selection and certification, the radius in Eq.(9) remains valid, with $\hat{p}_L$ computed for $\mathcal{B}(\hat{z}, s)$. With the same confidence, it guarantees $d\big(g_s(x + \delta), g_s(x)\big) \le r$ and $d\big(g_s(x + \delta), \hat{z}\big) \le r$ for every $\|\delta\|_2 < \hat{R}$. The complete proof is provided in Appendix D.

Algorithm 1 provides a Monte Carlo procedure for ContRS through three stages with independent sample sets: *candidate generation* produces $N_1$ candidate outputs; *center selection* uses $N_{\mathrm{sel}}$ predictions to select the center $\hat{z}$ with the largest empirical neighborhood mass; and *probability certification* uses $N_2$ predictions to compute the lower confidence bound $\hat{p}_L$ and the certified radius $\hat{R}$. Center selection is restricted to the sampled outputs $\{y_i\}_{i=1}^{N_1}$, whereas $g_s(x)$ maximizes population neighborhood mass over all $z \in \mathcal{M}$. The sampled outputs need not contain a full-space maximizer even as $N_1$ increases, which may make $\hat{R}$ conservative. When $\hat{p}_L > 1/2$, the certificate remains valid on the $1 - \alpha$ confidence event, because the neighborhood mass at $g_s(x)$ is at least that at $\hat{z}$ and hence at least $\hat{p}_L$. If $\hat{p}_L \le 1/2$, it returns $\hat{R} = 0$, indicating that no positive radius is certified. For general discrepancies *without* a suitable geometric bound on the competing mass, Appendix E develops an estimate based on sampled competing centers. Its finite-sample guarantees cover the sampled candidates, while the resulting radius is treated as an empirical robustness estimate for the full output space. In addition, Appendix F reports the empirical runtime comparison.

## 5 Experiments

### 5.1 Experimental Setup

**Task Choices.** For continuous outputs, the appropriate discrepancy measure depends on the output type and the changes relevant to the task. We therefore evaluate ContRS on representative prediction tasks using commonly used discrepancy measures. Specifically, we use *Jaccard distance*, *perceptual distance*, and *cosine distance* for face detection Vakhshiteh et al. (2021), image reconstruction Kos et al. (2018), and image-text retrieval Radford et al. (2021), respectively. Jaccard distance is a metric, while the latter two are non-metric discrepancies satisfying a relaxed triangle inequality. We further evaluate image reconstruction using non-metric discrepancies derived from *peak signal-to-noise ratio* (PSNR) Wang & Bovik (2009) and *structural similarity index* (SSIM) Wang et al. (2004).Detailed task choices and discrepancy definitions are provided in Appendix G.

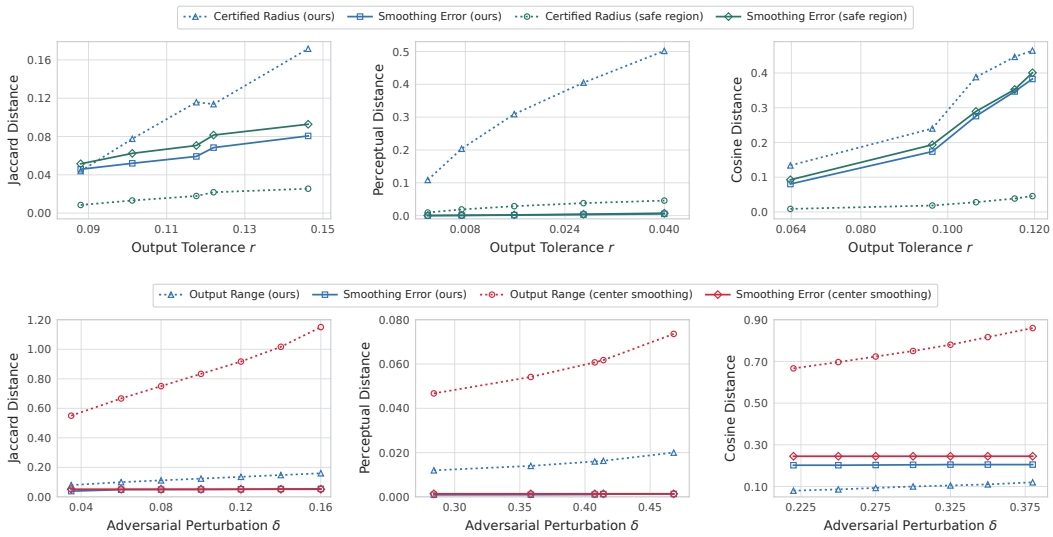



**Figure 1.** Comparison with safe region (top) and center smoothing (bottom) across Jaccard, perceptual, and cosine distances. Dotted curves show certification quantities, and solid curves show smoothing errors.

**Compared Baselines.** Most extensions of randomized smoothing beyond classification rely on task-specific output representations or decomposition strategies Chiang et al. (2020); Fischer et al. (2021); Miri Rekavandi et al. (2024). Their certificates are thus tied to particular output structures and are not applicable to our evaluation, which requires bounds on prediction change across different discrepancy measures and downstream tasks. We therefore focus on *center smoothing* Kumar & Goldstein (2021) and *safe region* Shao et al. (2023), which formulate robustness directly in the output space and allow comparisons using the same task-specific discrepancies as our model. The detailed formulations of these two baselines are provided in Appendix H and Appendix I.

**Implementation Details.** For above tasks, we use standard models and datasets. For face detection, we use a pretrained multi-task cascaded convolutional networks (MTCNN) Zhang et al. (2016) on CelebA Liu et al. (2015), with the predicted bounding box as the output. For image reconstruction, we use a pretrained variational autoencoder (VAE) Kingma & Welling (2013) on CIFAR-10 Krizhevsky et al. (2009), with the reconstructed image as the output. For image-text retrieval, we use the image encoder of contrastive language-image pre-training (CLIP) Radford et al. (2021) on MS-COCO Lin et al. (2014) and normalize its output embeddings before computing cosine dissimilarity. For non-metric experiments, we evaluate VAE reconstruction with PSNR on CIFAR-10 and SSIM on CelebA, respectively. In all settings, ContRS applies Gaussian smoothing to the corresponding base predictor and selects the output center according to the task-specific discrepancy described above. For consistent comparisons, all methods use the same base predictor, Gaussian smoothing noise level, and output discrepancy. All task settings are provided in Appendix J.

### 5.2 Comparison of Certification Performance

We compare ContRS with the two baselines in terms of certification performance and smoothing error. Under the corresponding matched settings, *larger* certified input radii and *smaller* output bounds indicate stronger guarantees, while *lower* smoothing error indicates less deviation from the base predictions Cohen et al. (2019); Salman et al. (2019). For comparison with safe region, we set ContRS's neighborhood radius $r/(2\kappa)$ equal to the radius of safe region's output region and compare the resulting certified input radii. As shown in Fig. 1 (top), ContRS achieves larger radii across all three tasks, with the gap generally widening as the output-region scale increases. Its smoothing errors are lower for Jaccard and cosine distances and comparable for perceptual distance. Since an identical output ball and probability lower bound yield the same Gaussian radius, this difference reflects output-ball selection. For comparison with center smoothing, we set its input perturbation budget to ContRS's certified radius $R$ and compare the resulting certified output bound with ContRS's prescribed output tolerance $r$. Fig. 1 (bottom) shows consistently tighter bounds across all

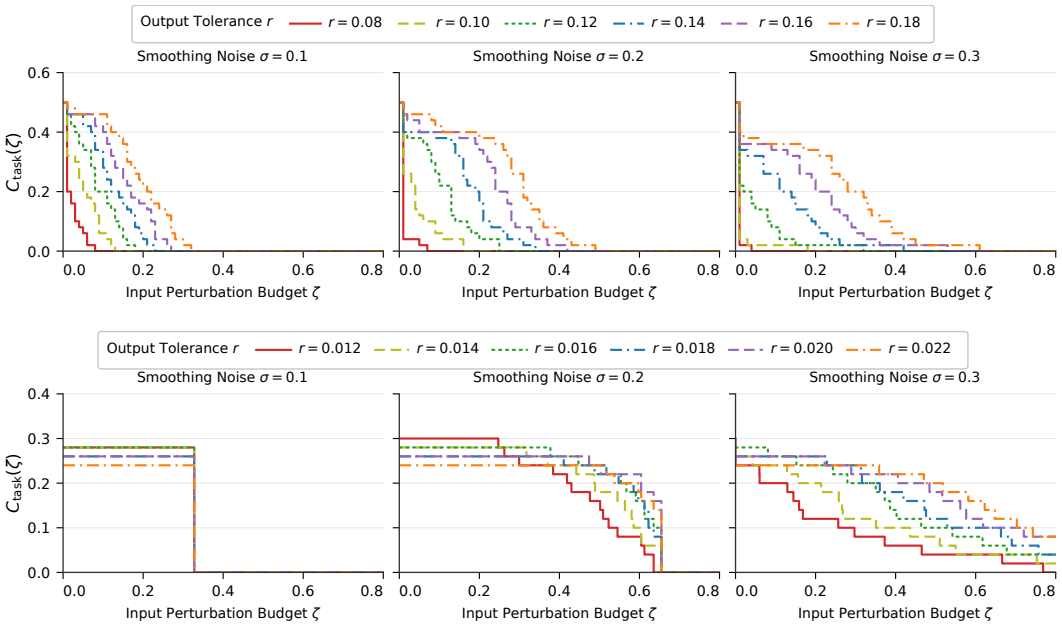



**Figure 2.** Task performance and certified stability of ContRS for face detection (top) and image reconstruction (bottom). Each column fixes the smoothing noise level $\sigma$, while different curves represent output tolerances $r$.

three tasks, with more obvious improvements for Jaccard and cosine distances. Smoothing errors remain comparable under Jaccard and perceptual distances and are lower under cosine distance. These results demonstrate improved certification across the evaluated tasks while maintaining lower or comparable prediction distortion.

### 5.3 Certified Stability Analysis

Beyond certificate tightness, we jointly evaluate downstream task performance and certified output stability. Following certified-accuracy evaluation Cohen et al. (2019) and its adaptation Shao et al. (2023), we report $C_{\text{task}}(\zeta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(R_i \geq \zeta)\mathbb{I}(\mathcal{A}_{\text{task},i})$, where $N$ is the test-set size, $R_i$ is the certified input radius at a prescribed output tolerance, and $\zeta$ is the required input perturbation budget[2]. $\mathbb{I}(\cdot)$ is the indicator function, and $\mathcal{A}_{\text{task},i}$ denotes that sample $i$'s clean prediction meets the task-specific evaluation criterion. Thus, at a fixed $\zeta$, a higher $C_{\text{task}}(\zeta)$ indicates that a larger fraction of samples meet the task criterion on clean inputs and have certified input radii of at least $\zeta$.

We describe the task-specific evaluation criterion for each downstream task. For face detection, we require $\text{IoU}_{i,\text{clean}} > 0.5$ following the standard localization criterion Everingham et al. (2010). Here, $\text{IoU}_{i,\text{clean}}$ measures the overlap between the predicted and ground-truth bounding boxes for clean input $i$, defined as the ratio of their intersection area to their union area. For image reconstruction, we require the clean reconstruction error not to exceed a fixed threshold. Following the error threshold calibration in prior work Meng & Chen (2017), we set this threshold to the empirical 90th percentile of clean reconstruction errors and keep it unchanged across certification settings. For image-text retrieval, let $\mathcal{G}_i$ denote the ground-truth caption set for image $i$ and $\pi_i$ the retrieval ranking induced by its clean smoothed CLIP representation. Following the standard Recall@$K$ evaluation Karpathy & Fei-Fei (2015), retrieval is considered successful if at least one ground-truth caption appears among the top-$K$ results, i.e., $\{\pi_i(1), \ldots, \pi_i(K)\} \cap \mathcal{G}_i \neq \varnothing$. Combining this criterion with the certification requirement $R_i \geq \zeta$ yields $C_{\text{task}}(\zeta)$ in image-text retrieval.

Fig. 2 shows the changes of $C_{\text{task}}(\zeta)$ for face detection and image reconstruction. Across all settings, $C_{\text{task}}(\zeta)$ decreases as the required input perturbation budget increases, since fewer samples satisfy the stronger certification requirement. Furthermore, larger $\sigma$ generally extends the certified input range. This is because ContRS computes the certified radius as $\hat{R} = \sigma\Phi^{-1}(\hat{p}_L)$ whenever $\hat{p}_L > 1/2$,

[2]At $\zeta = 0$, we count only samples with $R_i > 0$.

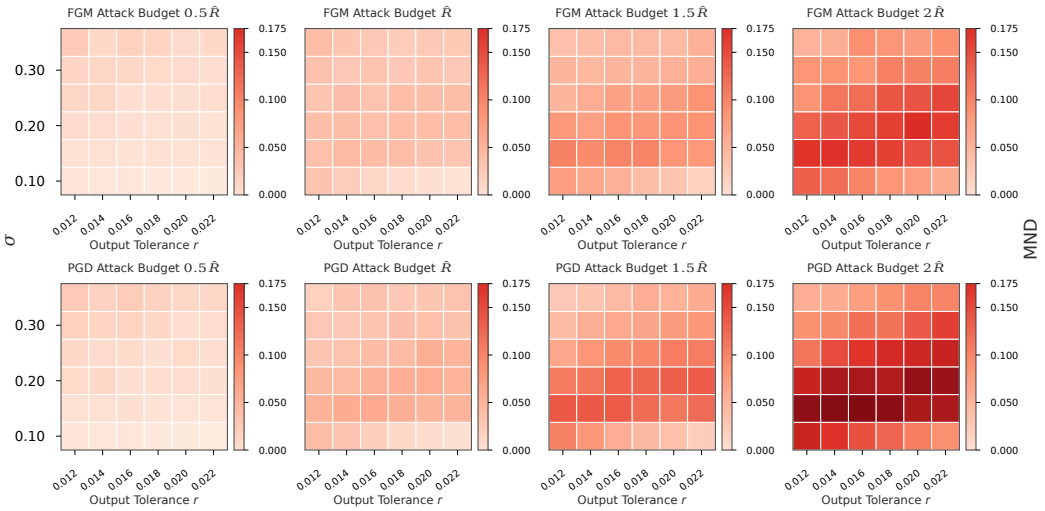



**Figure 3.** MND under FGM (top) and PGD (bottom) attacks. Each column represents a specific perturbation budget for samples with $\hat{R} > 0$, while the heatmap axes vary the smoothing noise $\sigma$ and the output tolerance $r$.

so a larger $\sigma$ can yield a larger input certified radius when $\hat{p}_L$ remains comparable. Additional experimental results are reported in Appendix K.

### 5.4 Empirical Robustness Under Adversarial Attacks

Previous studies Shao et al. (2023) use adversarial attacks as empirical checks of their certified output ranges. Following this approach, we examine whether the output variation of ContRS remains within the prescribed output tolerance under FGM Kireev et al. (2022) and PGD Madry et al. (2017) attacks. Notice that this evaluation is separate from the formal finite-sample certificate and is not intended to empirically verify Corollary 4.2. We evaluate VAE on image reconstruction, using perceptual distance to measure output variation. For each *clean input*, we first obtain the estimated smoothed prediction $\hat{z}$ with its corresponding certified input radius $R$. We then generate *adversarial inputs* using perturbation budgets of $0.5R$, $R$, $1.5R$, and $2R$, covering perturbation scales both within and beyond the certified input radius.The attack settings are provided in Appendix M. For each *adversarial input*, we rerun ContRS to obtain the estimated smoothed prediction $\hat{z}_{\text{adv}}$. We finally report the **Mean Normalized Deviation (MND)**, defined as $\text{MND} = \frac{1}{N_+}\sum_{i=1}^{N_+}\frac{d(\hat{z}_{\text{adv},i},\hat{z}_i)}{r}$, where $N_+$ denotes the number of test inputs with $\hat{R}_i > 0$ under the corresponding $(\sigma, r)$ setting, $d$ denotes the perceptual distance, and $r$ is the prescribed output tolerance. Thus, MND values below and above one indicate that the average prediction discrepancy is below and above $r$, respectively. As shown in Fig. 3, MND remains *below one* under both attacks across all evaluated perturbation budgets. Even at $1.5R$ and $2R$, the average prediction deviation remains below the prescribed output tolerance. Attack-parameter sensitivity and the additional PSNR- and SSIM-based evaluations are reported in Appendices N.

## 6 Conclusion

ContRS generalizes randomized smoothing to continuous outputs while recovering classical Gaussian randomized smoothing as a special case. It compares neighborhood probability masses to map a prescribed output tolerance to a certified input radius. In this paper, we establish guarantees for general measurable discrepancies, with rigorous finite-sample certificates under metric or symmetric relaxed-triangle conditions and empirical estimates when suitable geometric bounds are unavailable. Experiments use Jaccard, perceptual, and cosine distances for face detection, image reconstruction, and image-text retrieval, respectively. ContRS's output-ball selection yields larger certified input radii than safe region at matched region scales and tighter output bounds than center smoothing at matched input budgets, with comparable or lower smoothing error.

## AI USE STATEMENT

In preparing this manuscript, we use large language models (LLMs) to assist with proofreading, grammar correction, and language refinement to improve its clarity, readability, and terminological consistency. We carefully review and edit all AI-assisted content and take full responsibility for the final manuscript.

## REPRODUCIBILITY STATEMENT

To support reproducibility, we specify model assumptions, concrete algorithm, and independent Monte Carlo sampling stages of ContRS, with full derivations and proofs provided in Appendix. We describe used datasets, base models, discrepancy definitions, sample sizes and confidence levels, baseline matching protocols, and adversarial attack settings. We also distinguish finite-sample certificates from the empirical robustness estimates reported for general discrepancies.

## REFERENCES

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021.

Francesco Calivá, Kaiyang Cheng, Rutwik Shah, and Valentina Pedoia. Adversarial robust training of deep learning mri reconstruction models. *arXiv preprint arXiv:2011.00070*, 2020.

Ruoxin Chen, Jie Li, Junchi Yan, Ping Li, and Bin Sheng. Input-specific robustness certification for randomized smoothing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6295–6303, 2022.

Ping-yeh Chiang, Michael Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and Tom Goldstein. Detection as regression: Certified object detection with median smoothing. *Advances in Neural Information Processing Systems*, 33:1275–1286, 2020.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Francesco Croce, Naman D Singh, and Matthias Hein. Towards reliable evaluation and fast training of robust semantic segmentation models. In *European Conference on Computer Vision*, pp. 180–197. Springer, 2024.

Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, June 2010. ISSN 0920-5691. doi: 10.1007/s11263-009-0275-4. URL https://doi.org/10.1007/s11263-009-0275-4.

Marc Fischer, Maximilian Baader, and Martin Vechev. Scalable certified segmentation via randomized smoothing. In *International Conference on Machine Learning*, pp. 3340–3351. PMLR, 2021.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Klim Kireev, Maksym Andriushchenko, and Nicolas Flammarion. On the effectiveness of adversarial training against common corruptions. In *Uncertainty in Artificial Intelligence*, pp. 1012–1021. PMLR, 2022.

Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *2018 ieee security and privacy workshops (spw)*, pp. 36–42. IEEE, 2018.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical report, University of Toronto*, 2009.

Aounon Kumar and Tom Goldstein. Center smoothing: Certified robustness for networks with structured outputs. *Advances in Neural Information Processing Systems*, 34:5560–5575, 2021.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. *arXiv preprint arXiv:1802.03471*, 2018.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. *Advances in neural information processing systems*, 32, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

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.

Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pp. 135–147, 2017.

Aref Miri Rekavandi, Farhad Farokhi, Olga Ohrimenko, and Benjamin I. P. Rubinstein. Certified adversarial robustness via randomized $\alpha$-smoothing for regression models. *Advances in Neural Information Processing Systems*, 37:134127–134150, 2024. doi: 10.52202/079017-4263.

Neethunath MR et al. Adversarial robustness of deep learning in medical imaging: A comprehensive survey and benchmark of state-of-the-art architectures. *International Journal of Advanced Computer Science & Applications*, 16(12):831, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in neural information processing systems*, 32, 2019.

Huaqing Shao, Lanjun Wang, and Junchi Yan. Robustness certification for structured prediction with general inputs via safe region modeling in the semimetric output space. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2010–2022, 2023.

Samuel Henrique Silva and Peyman Najafirad. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*, 2020.

Fatemeh Vakhshiteh, Ahmad Nickabadi, and Raghavendra Ramachandra. Adversarial attacks against face recognition: A comprehensive study. *IEEE Access*, 9:92735–92756, 2021.

Václav Vorácek and Matthias Hein. Improving l1-certified robustness via randomized smoothing by leveraging box constraints. In *International Conference on Machine Learning*, pp. 35198–35222. PMLR, 2023.

Binghui Wang, Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of graph neural networks against adversarial structural perturbation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1645–1653, 2021.

Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, pp. 5286–5295. PMLR, 2018.

Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International conference on machine learning*, pp. 10693–10705. PMLR, 2020.

Lu Yang, Qing Song, and Yingqi Wu. Attacks on state-of-the-art face recognition using attentional adversarial attack generative network. *Multimedia tools and applications*, 80(1):855–875, 2021.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. *Advances in neural information processing systems*, 31, 2018a.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018b.

# A Appendix Overview

The appendices present theoretical derivations, additional experiments, and implementation details that support the main paper. Appendix B describes the Lipschitz property of Gaussian smoothing used in the robustness guarantee, and Appendix C provides the complete proof of Theorem 4.1. Appendix D provides the proof of the finite-sample certificate in Corollary 4.2, including the extension to symmetric discrepancies satisfying a relaxed triangle inequality. For general discrepancies without a suitable geometric bound on the competing mass, Appendix E describes the empirical robustness estimate based on sampled competing centers. Appendix F reports the computational cost. Appendices G–I provide the discrepancy definitions, compared baseline formulations, and baseline comparison settings. Appendices J–K provide the task settings and additional results of $C_{\mathrm{task}}(\zeta)$. Appendix L reports the Monte Carlo sample-size sensitivity analysis. Finally, Appendices M–N provide the adversarial attack settings and additional experiments. Table 2 summarizes the notation used in our paper.

Table 2: Main notations used in this paper.

| Symbol | Description |
|---|---|
| $x, \delta$ | Input and adversarial perturbation. |
| $f, g$ | Base predictor and smoothed predictor. |
| $g_s$ | Population smoothed predictor defined using radius-$s$ output neighborhoods. |
| $\mathcal{M}, d$ | Output space and output discrepancy function. |
| $\epsilon, \sigma$ | Gaussian noise and its standard deviation. |
| $r$ | Prescribed output tolerance. |
| $\zeta$ | Required input perturbation budget. |
| $s, \kappa$ | Neighborhood radius and relaxed triangle inequality constant. |
| $\mathcal{B}(z, s)$ | Output neighborhood centered at $z$ with radius $s$. |
| $z^*, \hat{z}$ | Theoretical smoothed center and its sample-based estimate. |
| $p, q$ | Probability mass of the selected neighborhood and supremum of neighborhood masses over centers beyond the output tolerance. |
| $R, \hat{R}$ | Theoretical and finite-sample certified input radii. |
| $\hat{p}_L$ | Lower confidence bound on the probability mass of $\mathcal{B}(\hat{z}, s)$. |
| $N_1, N_{\text{sel}}, N_2$ | Sample sizes for candidate generation, center selection, and certification. |
| $\alpha$ | Certification failure probability. |

## B Lipschitz Property of Gaussian Smoothing

The proof of Theorem 4.1 requires bounds on how the probability of a fixed output-neighborhood event changes when the input is shifted from $x$ to $x + \delta$. We treat neighborhood membership as a binary event and apply the following standard Gaussian smoothing result.

**Lemma B.1** (Inverse-normal Lipschitz property)**.** *Let $h : \mathbb{R}^d \to [0, 1]$ be measurable and let $\bar{h}(u) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2 I)}[h(u + \epsilon)]$. Then $u \mapsto \Phi^{-1}(\bar{h}(u))$ is $1/\sigma$-Lipschitz; that is, for all $u, v \in \mathbb{R}^d$, we have*

$$\left|\Phi^{-1}(\bar{h}(u)) - \Phi^{-1}(\bar{h}(v))\right| \le \frac{\|u - v\|_2}{\sigma}. \tag{11}$$

*Proof.* Write $\epsilon = \sigma Z$, where $Z \sim \mathcal{N}(0, I_d)$, and set $s(u) = \bar{h}(u) = \mathbb{E}[h(u+\sigma Z)]$. Gaussian convolution makes $s$ smooth even when $h$ is only measurable. For any unit vector $a \in \mathbb{R}^d$, differentiation under the integral gives

$$\partial_a s(u) = \frac{1}{\sigma}\mathbb{E}\left[(a^\top Z)h(u + \sigma Z)\right]. \tag{12}$$

Let $T = a^\top Z \sim \mathcal{N}(0, 1)$ and define $\psi(t) = \mathbb{E}[h(u + \sigma Z) \mid T = t]$. Then $0 \le \psi \le 1$ and $\mathbb{E}[\psi(T)] = s(u)$. Among all functions with these two properties, the quantity $\mathbb{E}[T\psi(T)]$ is maximized by placing all available mass on the upper Gaussian tail and minimized by placing it on the lower tail. This one-dimensional rearrangement argument therefore yields

$$|\mathbb{E}[T\psi(T)]| \le \phi\big(\Phi^{-1}(s(u))\big), \tag{13}$$

where $\phi$ is the standard normal density. For example, the upper extremizer is $\psi(t) = \mathbf{1}\{t \ge \Phi^{-1}(1 - s(u))\}$, whose first moment is $\phi(\Phi^{-1}(s(u)))$.

Combining equation 12 with equation 13 gives

$$|\partial_a s(u)| \le \frac{1}{\sigma}\phi\big(\Phi^{-1}(s(u))\big). \tag{14}$$

Whenever $0 < s(u) < 1$, the chain rule and $(\Phi^{-1})'(p) = 1/\phi(\Phi^{-1}(p))$ imply

$$\left|\partial_a \Phi^{-1}(s(u))\right| \le \frac{1}{\sigma}. \tag{15}$$

Because equation 15 holds for every unit direction $a$, the gradient of $\Phi^{-1} \circ s$ has Euclidean norm at most $1/\sigma$. Integrating this gradient along the line segment from $u$ to $v$ proves equation 11.

If $h$ is not almost everywhere constant, Gaussian convolution gives $0 < s(u) < 1$ for every $u$. The cases in which $h = 0$ or $h = 1$ almost everywhere are trivial and may equivalently be obtained by a limiting argument. This completes the proof. □

This argument is the usual inverse-normal formulation of Gaussian isoperimetry used in randomized smoothing; see, for example, Salman et al. (2019).

To connect the lemma to ContRS, fix a candidate center $z \in \mathcal{M}$ and let $H_z(u)$ denote the probability that a noisy prediction lies within distance $r/2$ of $z$:

$$H_z(u) := \mathbb{P}(f(u+\epsilon) \in \mathcal{B}(z, r/2)), \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 I). \tag{16}$$

This is the smoothed version of the indicator $h_z(v) = \mathbf{1}\{f(v) \in \mathcal{B}(z, r/2)\}$. Applying Lemma B.1 to $h_z$ and comparing $u$ with $u + \delta$ yields

$$\Phi\left(\Phi^{-1}(H_z(u)) - \frac{\|\delta\|_2}{\sigma}\right) \le H_z(u+\delta) \le \Phi\left(\Phi^{-1}(H_z(u)) + \frac{\|\delta\|_2}{\sigma}\right). \tag{17}$$

The lower bound limits how quickly the selected neighborhood can lose probability mass; the upper bound limits how quickly a competing neighborhood can gain mass. Because each neighborhood is treated only as a measurable event, equation 17 does not require symmetry or a triangle inequality for $d$.

## C Proof of Theorem 4.1

We use the notation $H_z$ and the Gaussian probability-shift bounds in Eq.(17) throughout this section.

*Proof.* Fix $x$ and write $z^* = g(x)$. Let

$$\mathcal{C}_x = \{z \in \mathcal{M} : d(z, z^*) > r\}. \tag{18}$$

By definition, $p = H_{z^*}(x)$ and $q = \sup_{z \in \mathcal{C}_x} H_z(x)$, so $H_z(x) \le q$ for every $z \in \mathcal{C}_x$.

For a perturbation $\delta$, let $t = \|\delta\|_2/\sigma$. Applying Eq.(17) to the selected center and to any $z \in \mathcal{C}_x$ gives

$$H_{z^*}(x+\delta) \ge \Phi\big(\Phi^{-1}(p) - t\big), \qquad H_z(x+\delta) \le \Phi\big(\Phi^{-1}(q) + t\big). \tag{19}$$

If $\|\delta\|_2 < R$, where $R = \frac{\sigma}{2}\big(\Phi^{-1}(p) - \Phi^{-1}(q)\big)$, then

$$\Phi^{-1}(p) - t > \Phi^{-1}(q) + t. \tag{20}$$

Since $\Phi$ is strictly increasing, $H_{z^*}(x+\delta) > H_z(x+\delta)$ for every $z \in \mathcal{C}_x$.

Now let $z_\delta = g(x+\delta)$. Since $z_\delta$ maximizes $H_z(x+\delta)$ and $z^*$ is an admissible center, $H_{z_\delta}(x+\delta) \ge H_{z^*}(x+\delta)$. Hence $z_\delta \notin \mathcal{C}_x$, and therefore $d(g(x+\delta), g(x)) \le r$.

The argument above is written for $0 < q < p < 1$. The endpoint cases follow by taking limits. □

## D Proof of Corollary 4.2

*Proof.* We first consider the metric case. Let $\hat{z}$ be selected from the candidate and selection samples, independently of the certification samples, and let

$$p_{\hat{z}} = \mathbb{P}\big(f(x+\epsilon) \in B(\hat{z}, r/2)\big). \tag{21}$$

Conditional on $\hat{z}$, the certification count satisfies $k_{\text{cert}} \sim \text{Binomial}(N_2, p_{\hat{z}})$. Thus, the one-sided Clopper–Pearson lower bound $\hat{p}_L$ satisfies

$$\mathbb{P}(p_{\hat{z}} \ge \hat{p}_L \mid \hat{z}) \ge 1 - \alpha. \tag{22}$$

Averaging over the randomness used to select $\hat{z}$ gives $\mathbb{P}(p_{\hat{z}} \ge \hat{p}_L) \ge 1 - \alpha$.

Work on this event and suppose $\hat{p}_L > 1/2$. For any $z$ satisfying $d(z, \hat{z}) > r$, the metric triangle inequality implies $B(z, r/2) \cap B(\hat{z}, r/2) = \varnothing$. Hence,

$$H_z(x) \le 1 - H_{\hat{z}}(x) = 1 - p_{\hat{z}} \le 1 - \hat{p}_L. \tag{23}$$

Let $t = \|\delta\|_2/\sigma$. By Eq. (17), we have

$$H_{\hat{z}}(x+\delta) \ge \Phi\big(\Phi^{-1}(\hat{p}_L) - t\big), \tag{24}$$

whereas every $z$ with $d(z, \hat{z}) > r$ satisfies

$$H_z(x+\delta) \le \Phi\big(\Phi^{-1}(1-\hat{p}_L) + t\big) = \Phi\big(-\Phi^{-1}(\hat{p}_L) + t\big). \tag{25}$$

If $\|\delta\|_2 < \hat{R} = \sigma\Phi^{-1}(\hat{p}_L)$, the lower bound for $H_{\hat{z}}(x+\delta)$ strictly exceeds the upper bound for every such competing center. Since $g(x+\delta)$ maximizes the neighborhood mass, it follows that

$$d\big(g(x+\delta), \hat{z}\big) \le r. \tag{26}$$

This proves the metric case. Now suppose that $d$ is symmetric and satisfies the $\kappa$-relaxed triangle inequality. Set $s = r/(2\kappa)$ and define

$$H_{z,s}(u) = \mathbb{P}\big(f(u+\epsilon) \in B(z,s)\big), \qquad g_s(u) \in \arg\max_{z\in\mathcal{M}} H_{z,s}(u).$$

Both center selection and probability certification use radius-$s$ neighborhoods. If $B(z,s) \cap B(\hat{z},s) \neq \varnothing$, then for a point $y$ in the intersection, symmetry and the relaxed triangle inequality give

$$d(z,\hat{z}) \le \kappa\big(d(z,y) + d(y,\hat{z})\big) \le 2\kappa s = r. \tag{27}$$

Consequently, whenever $d(z,\hat{z}) > r$, the two neighborhoods are disjoint. On the Clopper–Pearson confidence event $H_{\hat{z},s}(x) \ge \hat{p}_L$, every such competing center has $H_{z,s}(x) \le 1 - \hat{p}_L$. Lemma B.1 applies to $H_{z,s}$, since a radius-$s$ neighborhood is also a measurable event. Repeating the probability comparison above therefore yields

$$d\big(g_s(x+\delta), \hat{z}\big) \le r \quad \text{for all } \|\delta\|_2 < \hat{R} = \sigma\Phi^{-1}(\hat{p}_L),$$

with probability at least $1-\alpha$. □

It remains to establish the bound relative to the unperturbed smoothed prediction. The following argument applies to both the metric case, where $s = r/2$, and the symmetric $\kappa$-relaxed-triangle case, where $s = r/(2\kappa)$. Let $c = g_s(x)$. On the same Clopper–Pearson confidence event used above,

$$H_{\hat{z},s}(x) \ge \hat{p}_L. \tag{28}$$

Because $g_s(x)$ maximizes the neighborhood probability, $H_{c,s}(x) \ge H_{\hat{z},s}(x) \ge \hat{p}_L$. For any center $z$ with $d(z,c) > r$, the geometric condition implies $B(z,s) \cap B(c,s) = \varnothing$. Therefore,

$$H_{z,s}(x) \le 1 - H_{c,s}(x) \le 1 - \hat{p}_L. \tag{29}$$

Putting $t = \|\delta\|_2/\sigma$ and applying Lemma B.1 gives

$$H_{c,s}(x+\delta) \ge \Phi\big(\Phi^{-1}(\hat{p}_L) - t\big), \qquad H_{z,s}(x+\delta) \le \Phi\big(-\Phi^{-1}(\hat{p}_L) + t\big). \tag{30}$$

If $\|\delta\|_2 < \hat{R} = \sigma\Phi^{-1}(\hat{p}_L)$, then $t < \Phi^{-1}(\hat{p}_L)$, so the first bound is strictly larger than the second. No center at distance greater than $r$ from $c$ can maximize the neighborhood probability at $x+\delta$. Consequently,

$$d\big(g_s(x+\delta), g_s(x)\big) \le r. \tag{31}$$

This conclusion and the bound relative to $\hat{z}$ hold simultaneously on the same confidence event, and no additional failure probability or reduction of the certified radius is needed.

## E Empirical Robustness Estimation

This section considers general measurable discrepancies for which no suitable geometric bound on the competing neighborhood mass is available, beyond the metric and relaxed-triangle settings considered in the main text. Without such a bound, the disjointness argument underlying Corollary 4.2 is unavailable. In particular, even if $d(z,z^*) > r$, the neighborhoods $\mathcal{B}(z, r/2)$ and $\mathcal{B}(z^*, r/2)$ may overlap, so the competing probability mass $q = \sup_{d(z,z^*)>r} \mathbb{P}(f(x+\epsilon) \in \mathcal{B}(z, r/2))$ cannot be bounded by $1-p$.

We therefore estimate the competing mass over a finite set of sampled competing centers. Let $\hat{z}$ be the estimated center and let $y''_1, \ldots, y''_{N_2}$ be independent certification outputs. The estimated center and competing-center set are constructed independently of the certification samples. For a set $\mathcal{Z}_{\text{cand}}$

of size at most $M$, define $\hat{q} = \max_{z \in \mathcal{Z}_{\text{cand}}} N_2^{-1} \sum_{j=1}^{N_2} \mathbf{1}\{d(z, y_j'') \leq r/2\}$, where only centers satisfying $d(z, \hat{z}) > r$ are retained, and $\hat{q} = 0$ if the competing-center set is empty.

For the fixed sampled competing-center set, a one-sided Hoeffding bound and a union bound give the Monte Carlo slack $\xi_N = \sqrt{\ln(M/\eta)/(2N_2)}$. Define $q_U = \min\{1, \hat{q} + \xi_N\}$. When $q_U < \hat{p}_L$, we define

$$\hat{R}_{\text{gen}} = \frac{\sigma}{2} \left[ \Phi^{-1}(\hat{p}_L) - \Phi^{-1}(q_U) \right]. \tag{32}$$

If $q_U \geq \hat{p}_L$, we instead return radius zero. The following proposition states exactly what these quantities certify.

**Proposition E.1** (Guarantee over sampled competing centers)**.** *Assume that $\hat{z}$ and $\mathcal{Z}_{\text{cand}}$ are constructed independently of the certification outputs and that $|\mathcal{Z}_{\text{cand}}| \leq M$. With probability at least $1 - \alpha - \eta$, simultaneously*

$$p_{\hat{z}} \geq \hat{p}_L \quad \textit{and} \quad \max_{z \in \mathcal{Z}_{\text{cand}}} \mathbb{P}(f(x + \epsilon) \in \mathcal{B}(z, r/2)) \leq q_U. \tag{33}$$

*On this event, if $q_U < \hat{p}_L$, then every perturbation satisfying $\|\delta\|_2 < \hat{R}_{\text{gen}}$ obeys*

$$H_{\hat{z}}(x + \delta) > H_z(x + \delta) \quad \textit{for every } z \in \mathcal{Z}_{\text{cand}} \textit{ such that } d(z, \hat{z}) > r. \tag{34}$$

*Consequently, if the maximization defining the smoothed prediction is restricted to $\{\hat{z}\} \cup \mathcal{Z}_{\text{cand}}$, it cannot select a sampled competing center whose discrepancy from $\hat{z}$ exceeds $r$.*

*Proof.* Condition on the realized center $\hat{z}$ and competing-center set $\mathcal{Z}_{\text{cand}}$. They are fixed under this conditioning because they were constructed independently of the certification outputs. For each $z \in \mathcal{Z}_{\text{cand}}$, define its population and empirical neighborhood masses by

$$Q_z := \mathbb{P}(f(x + \epsilon) \in \mathcal{B}(z, r/2)), \qquad \widehat{Q}_z := \frac{1}{N_2} \sum_{j=1}^{N_2} \mathbf{1}\{y_j'' \in \mathcal{B}(z, r/2)\}. \tag{35}$$

For a fixed $z$, the variables in the empirical average are independent Bernoulli random variables with mean $Q_z$. The one-sided Hoeffding inequality therefore gives

$$\mathbb{P}\Big( Q_z > \widehat{Q}_z + \xi_N \,\Big|\, \hat{z}, \mathcal{Z}_{\text{cand}} \Big) \leq \exp(-2N_2 \xi_N^2) = \frac{\eta}{M}. \tag{36}$$

A union bound over at most $M$ candidates yields, with conditional probability at least $1 - \eta$,

$$Q_z \leq \widehat{Q}_z + \xi_N \quad \text{for every } z \in \mathcal{Z}_{\text{cand}}. \tag{37}$$

Taking the maximum over $z$ and clipping at one gives $\max_z Q_z \leq q_U$. Because the conditional failure probability is at most $\eta$ for every realized center and competing-center set, the same bound holds unconditionally. If the competing-center set is empty, the statement holds trivially under the convention that its maximum mass is zero.

Independently, the one-sided Clopper–Pearson construction gives $p_{\hat{z}} \geq \hat{p}_L$ with probability at least $1 - \alpha$. The two confidence events need not be independent: another union bound shows that they hold simultaneously with probability at least $1 - \alpha - \eta$, which proves the joint confidence statement in equation 33.

It remains to transfer these bounds from $x$ to $x + \delta$. Put $t = \|\delta\|_2 / \sigma$. On the joint confidence event, the Gaussian shift bounds in equation 17 give

$$H_{\hat{z}}(x + \delta) \geq \Phi\big(\Phi^{-1}(\hat{p}_L) - t\big), \tag{38}$$

$$H_z(x + \delta) \leq \Phi\big(\Phi^{-1}(q_U) + t\big), \qquad z \in \mathcal{Z}_{\text{cand}}. \tag{39}$$

If $\|\delta\|_2 < \hat{R}_{\text{gen}}$, the definition in equation 32 implies

$$\Phi^{-1}(\hat{p}_L) - t > \Phi^{-1}(q_U) + t. \tag{40}$$

By the strict monotonicity of $\Phi$, the bounds in equation 38 and equation 39, together with equation 40, establish equation 34. This proves the stated conclusion for the candidate-restricted maximization rule. □

**Algorithm 2** Prediction and Empirical Robustness Estimation

**Input:** data sample $x$; base predictor $f$; measurable discrepancy $d$; prescribed output tolerance $r$; Gaussian noise scale $\sigma$; sample sizes $N_1, N_{\mathrm{sel}}, N_2$; competing-center sample size $M$; and failure probabilities $\alpha, \eta \in (0,1)$ with $\alpha + \eta < 1$.
**Output:** estimated center $\hat{z}$ and empirical robustness estimate $\hat{R}_{\mathrm{gen}}$.
All noise samples below are mutually independent and follow $\mathcal{N}(0, \sigma^2 I_n)$.
Draw candidate outputs $y_i = f(x + \epsilon_i)$, $i = 1, \ldots, N_1$.
Draw independent selection outputs $y'_j = f(x + \epsilon'_j)$, $j = 1, \ldots, N_{\mathrm{sel}}$.
**for** $i = 1, \ldots, N_1$ **do**
    $k_i \leftarrow \sum_{j=1}^{N_{\mathrm{sel}}} \mathbf{1}\{d(y_i, y'_j) \le r/2\}$.
**end for**
$i^* \leftarrow \min\{i : k_i = \max_{1 \le \ell \le N_1} k_\ell\}$ and $\hat{z} \leftarrow y_{i^*}$.
Draw independent certification outputs $y''_j = f(x + \epsilon''_j)$, $j = 1, \ldots, N_2$.
$k_{\mathrm{cert}} \leftarrow \sum_{j=1}^{N_2} \mathbf{1}\{d(\hat{z}, y''_j) \le r/2\}$.
$\hat{p}_L \leftarrow \mathrm{ClopperPearsonLower}(k_{\mathrm{cert}}, N_2, \alpha)$.
Draw additional candidate outputs $\tilde{y}_t = f(x + \tilde{\epsilon}_t)$, $t = 1, \ldots, M$.
$\mathcal{Z}_{\mathrm{cand}} \leftarrow \{\tilde{y}_t : d(\tilde{y}_t, \hat{z}) > r\}$.
**if** $\mathcal{Z}_{\mathrm{cand}} = \emptyset$ **then**
    $\hat{q} \leftarrow 0$.
**else**
    $\hat{q} \leftarrow \max_{z \in \mathcal{Z}_{\mathrm{cand}}} \frac{1}{N_2} \sum_{j=1}^{N_2} \mathbf{1}\{d(z, y''_j) \le r/2\}$.
**end if**
$\xi_N \leftarrow \sqrt{\ln(M/\eta)/(2N_2)}$.
$q_U \leftarrow \min\{1, \hat{q} + \xi_N\}$.
**if** $q_U \ge \hat{p}_L$ **then**
    $\hat{R}_{\mathrm{gen}} \leftarrow 0$.
**else**
    $\hat{R}_{\mathrm{gen}} \leftarrow \frac{\sigma}{2}\left[\Phi^{-1}(\hat{p}_L) - \Phi^{-1}(q_U)\right]$.
**end if**
**Return:** estimated center $\hat{z}$ and empirical robustness estimate $\hat{R}_{\mathrm{gen}}$.

**Remark E.1** (Scope of the empirical robustness estimate)**.** *The quantity in Eq.(32) is not a finite-sample certificate over the full output space. The slack $\xi_N$ controls estimation error only for the sampled competing centers and does not account for a high-mass competitor that is never sampled. Hence $q_U$ may underestimate the true supremum over $\mathcal{M}$. We therefore report $\hat{R}_{\mathrm{gen}}$ as an empirical robustness estimate with respect to the full output space and evaluate it under the adversarial attacks described in the main text. A rigorous uniform bound over the full output space would require additional structure, such as a suitable covering-number construction.*

The procedure retains the three independent Monte Carlo stages of Algorithm 1 and adds an independent pool of $M$ outputs for searching over sampled competing centers. Algorithm 2 summarizes the resulting estimation procedure.

## F Computational Cost

Table 3 reports the execution time per test input for ContRS and the two baselines across three downstream tasks. All measurements were obtained on a server equipped with NVIDIA H200 NVL GPUs, running NVIDIA driver 570.195.03 and CUDA 12.8. We can observe that ContRS has the lowest measured runtime on all three tasks, with its largest reduction relative to safe region observed on image reconstruction.

| Task | center smoothing | safe region | ContRS |
|---|---|---|---|
| Face Detection | 115.387 | 90.888 | 89.877 |
| Image Reconstruction | 76.427 | 73.702 | 64.176 |
| Image-text Retrieval | 128.673 | 85.894 | 83.567 |

Table 3: Per-input empirical running time (s) for center smoothing, safe region, and ContRS on three downstream tasks.

# G Output Discrepancies

We choose each discrepancy to reflect the continuous output changes relevant to the corresponding task. Jaccard distance measures spatial overlap between bounding boxes, perceptual distance measures feature-level changes between reconstructed images, and cosine distance measures directional changes in normalized retrieval embeddings. We additionally evaluate image reconstruction with PSNR- and SSIM-based discrepancies to study the general setting in which no metric or relaxed-triangle structure is assumed. Together, the five discrepancies cover the three geometric cases considered in this paper: Jaccard distance is a metric; perceptual and cosine distances satisfy a symmetric relaxed triangle inequality with $\kappa = 2$; and the PSNR- and SSIM-based discrepancies are treated as general discrepancies.

Accordingly, ContRS uses neighborhood radius $r/2$ for Jaccard distance, $r/4$ for perceptual and cosine distances, and radius-$r/2$ neighborhoods with sampled competing centers for PSNR and SSIM, yielding $\hat{R}_{\mathrm{gen}}$.

## G.1 Face Detection with Jaccard Distance

The output of face detection is a predicted bounding box. Because meaningful variation in this output is determined by spatial overlap, we use the Jaccard distance. For two bounding boxes $A$ and $B$, it is defined as

$$d_{\mathrm{Jac}}(A, B) = 1 - \mathrm{IoU}(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}. \tag{41}$$

Jaccard distance is a metric, so the metric certification procedure uses output neighborhoods of radius $r/2$.

## G.2 Image Reconstruction with Perceptual Distance

The output of image reconstruction is a reconstructed image. Pixelwise differences need not reflect perceptually meaningful changes, so we measure output variation using perceptual distance Zhang et al. (2018b):

$$d_{\mathrm{perc}}(A, B) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left( \hat{\phi}_l(A)_{hw} - \hat{\phi}_l(B)_{hw} \right) \right\|_2^2. \tag{42}$$

Here, $\hat{\phi}_l$ is the normalized feature representation at layer $l$, $w_l$ contains the corresponding channel weights, and $H_l \times W_l$ is the spatial resolution. This squared weighted feature-space discrepancy satisfies the relaxed triangle inequality with $\kappa = 2$, so ContRS uses output neighborhoods of radius $r/4$.

## G.3 Image-Text Retrieval with Cosine Distance

The output used for image-text retrieval is a normalized image embedding. Because semantic alignment is represented by embedding direction, we use cosine dissimilarity. For normalized embeddings $u$ and $v$, we define

$$d_{\cos}(u, v) = 1 - \frac{u^\top v}{\|u\|_2 \|v\|_2} = \frac{1}{2}\|u - v\|_2^2. \tag{43}$$

Cosine dissimilarity satisfies the relaxed triangle inequality with $\kappa = 2$, so ContRS again uses output neighborhoods of radius $r/4$.

## G.4 Image Reconstruction with a PSNR-Based Discrepancy

For reconstructed images $u, v \in [0, 1]^d$, we define the PSNR-based discrepancy as

$$d_{\mathrm{PSNR}}(u, v) = \frac{1}{1 + \mathrm{PSNR}(u, v)}, \qquad \mathrm{PSNR}(u, v) = 10 \log_{10}\left( \frac{1}{\mathrm{MSE}(u, v)} \right). \tag{44}$$

Smaller values indicate greater similarity between the reconstructed images. We do not assume that $d_{\mathrm{PSNR}}$ satisfies the metric or relaxed-triangle conditions required by the finite-sample certificate. Consequently, we use radius-$r/2$ neighborhoods and the sampled competing centers in Appendix E.

### G.5 Image Reconstruction with an SSIM-Based Discrepancy

For reconstructed images $u$ and $v$, we define

$$d_{\mathrm{SSIM}}(u, v) = 1 - \mathrm{clip}(\mathrm{SSIM}(u, v), 0, 1). \tag{45}$$

Smaller values indicate greater structural similarity. As with the PSNR-based discrepancy, we do not assume that $d_{\mathrm{SSIM}}$ satisfies the metric or relaxed-triangle conditions. We therefore use radius-$r/2$ neighborhoods and report empirical robustness estimates obtained from the sampled competing centers.

## H Compared Baselines

The comparison in Section 5.2 involves methods with different certification targets. Having fixed the output discrepancies above, we next introduce the certification objective of each baseline.

### H.1 Center Smoothing

Center smoothing extends randomized smoothing to metric outputs and also applies to pseudometrics and symmetric discrepancies satisfying a $\kappa$-relaxed triangle inequality. Given a base predictor $f : \mathbb{R}^d \to \mathcal{M}$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, its population smoothed prediction $g_{\mathrm{CS}}(x)$ is a center of a smallest output ball containing at least half of the noisy-output probability mass:

$$\min_{z \in \mathcal{M},\, s \geq 0} s \quad \text{subject to} \quad \Pr(f(x+\epsilon) \in \mathcal{B}(z, s)) \geq \frac{1}{2}. \tag{46}$$

For a prescribed input radius $R$, consider a ball centered at $g_{\mathrm{CS}}(x)$ with radius $t_R$ such that

$$\Pr(f(x+\epsilon) \in \mathcal{B}(g_{\mathrm{CS}}(x), t_R)) \geq \Phi(R/\sigma).$$

The Gaussian probability-shift bound ensures that this ball contains more than half of the noisy-output mass at every $x + \delta$ with $\|\delta\|_2 < R$. The triangle or relaxed-triangle inequality then gives

$$\|\delta\|_2 < R \quad \Longrightarrow \quad d(g_{\mathrm{CS}}(x+\delta), g_{\mathrm{CS}}(x)) \leq \rho, \qquad \rho = 2\kappa t_R, \tag{47}$$

where $\kappa = 1$ for metrics and pseudometrics. Thus, center smoothing takes a prescribed input radius and derives a corresponding output bound. In its finite-sample implementation, the center and output-ball radius are approximated from noisy predictions, and the reported bound accounts for sampling confidence and center approximation.

### H.2 SAFE REGION

Safe region certifies whether noisy predictions remain inside a specified region $S \subseteq \mathcal{M}$. Define

$$p = \mathbb{P}\big(f(x+\epsilon) \in S\big). \tag{48}$$

If $p > 1/2$, Gaussian randomized smoothing gives the certified input radius

$$R = \sigma \Phi^{-1}(p). \tag{49}$$

For every $\delta$ satisfying $\|\delta\|_2 < R$,

$$\mathbb{P}\big(f(x+\delta+\epsilon) \in S\big) > \frac{1}{2}. \tag{50}$$

Thus, safe region certifies majority membership in an output region rather than directly bounding the change between smoothed predictions.

ContRS can also be expressed in this safe-region form. In the metric or relaxed-triangle setting, set

$$s = \frac{r}{2\kappa}, \tag{51}$$

where $\kappa = 1$ for metric discrepancies. The corresponding population smoothed predictor $g_s$ is obtained by replacing $r/2$ with $s$ in Eq. (3):

$$g_s(x) \in \arg\max_{z \in \mathcal{M}} \mathbb{P}\big(f(x+\epsilon) \in B(z, s)\big). \tag{52}$$

Once $g_s(x)$ is selected, it induces the output region

$$S = B(g_s(x), s). \tag{53}$$

**Corollary H.1** (ContRS as a Safe-Region Certificate)**.** *Fix an input $x$ and let $S = B(g_s(x), s)$, where $s = r/(2\kappa)$. If*

$$p = \mathbb{P}\big(f(x+\epsilon) \in S\big) > \frac{1}{2}, \tag{54}$$

*then every perturbation satisfying $\|\delta\|_2 < \sigma\Phi^{-1}(p)$ obeys*

$$\mathbb{P}\big(f(x+\delta+\epsilon) \in S\big) > \frac{1}{2}. \tag{55}$$

*Proof.* After $g_s(x)$ has been selected, the region $S = B(g_s(x), s)$ is fixed. Define $H(u) = \mathbb{P}(f(u+\epsilon) \in S)$. By Lemma B.1, $\Phi^{-1}(H(u))$ is $1/\sigma$-Lipschitz. Therefore,

$$\Phi^{-1}(H(x+\delta)) \geq \Phi^{-1}(H(x)) - \frac{\|\delta\|_2}{\sigma} = \Phi^{-1}(p) - \frac{\|\delta\|_2}{\sigma}. \tag{56}$$

If $\|\delta\|_2 < \sigma\Phi^{-1}(p)$, the right-hand side is positive. Since $\Phi(0) = 1/2$, this implies $H(x+\delta) > 1/2$. □

For finite-sample certification, let $\hat{z}_x$ be selected independently of the certification samples and define the sample-selected region $\hat{S} = B(\hat{z}_x, s)$. A one-sided Clopper–Pearson lower bound $\hat{p}_L$ estimates $\mathbb{P}(f(x+\epsilon) \in \hat{S})$. Whenever $\hat{p}_L > 1/2$, the certified input radius is

$$\hat{R} = \sigma\Phi^{-1}(\hat{p}_L). \tag{57}$$

With confidence at least $1-\alpha$, every $\|\delta\|_2 < \hat{R}$ then satisfies $\mathbb{P}(f(x+\delta+\epsilon) \in \hat{S}) > 1/2$. If $\hat{p}_L \leq 1/2$, no positive radius is certified.

# I Baseline Comparison Settings

The baseline formulations above do not determine a fair comparison, because center smoothing fixes an input radius whereas safe region fixes an output region. To support the comparison reported in Section 5.2, we therefore state the matching rule for each baseline before listing the corresponding parameter settings.

For center smoothing, each method is evaluated at the same input perturbation radius. For every ContRS configuration, we set its prescribed input radius to the finite-sample certified radius returned by ContRS, $R_{\mathrm{CS}} = \hat{R}$, and compare its certified output bound $\rho$ with the prescribed ContRS tolerance $r$. **A smaller output bound at the same input radius represents a tighter certificate.** For safe region, the output set is instantiated as a ball $\mathcal{B}(c, s)$ and the output-region scale is matched. The baseline first returns a radius $r_{\mathrm{SR}}$ at target mass $p = 0.55$. This setting $p = 0.55$ follows the experimental setting used by safe region in its comparison with center smoothing Shao et al. (2023). We then set the ContRS tolerance to $r = 2\kappa r_{\mathrm{SR}}$, where $\kappa = 1$ for Jaccard distance and $\kappa = 2$ for perceptual distance and cosine distance. Consequently, **the ContRS neighborhood radius $r/(2\kappa)$ is exactly $r_{\mathrm{SR}}$, giving a direct comparison of the two safe-region membership certificates.**

Table 4 summarizes the matched parameters for all three task–discrepancy pairs.

| Baseline | Task / Discrepancy | Evaluation parameters |
|---|---|---|
| center smoothing | Face detection / Jaccard distance | $\sigma = 0.1$, $r \in \{0.08, 0.10, 0.12, 0.14, 0.16\}$, input radius $R_{\mathrm{CS}} = \hat{R}$ |
| center smoothing | Image reconstruction / perceptual distance | $\sigma = 0.25$, $r \in \{0.012, 0.014, 0.016, 0.018, 0.020\}$, input radius $R_{\mathrm{CS}} = \hat{R}$ |
| center smoothing | Image-text retrieval / cosine distance | $\sigma = 0.25$, $r \in \{0.080, 0.088, 0.100, 0.108, 0.120\}$, input radius $R_{\mathrm{CS}} = \hat{R}$ |
| safe region | All three tasks | $p = 0.55$, $\sigma \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, and $r = 2\kappa r_{\mathrm{SR}}$ |

Table 4: Baseline comparison settings. The input radius of center smoothing is matched to the finite-sample ContRS radius $\hat{R}$, whereas safe region is matched at the same output-region radius.

# J TASK SETTINGS

This appendix reports the implementation parameters omitted from the main text. For each task, we specify the evaluation-set size, independent Monte Carlo sample sets, confidence level, smoothing levels, output-tolerance grid, and neighborhood radius.

## J.1 FACE DETECTION WITH JACCARD DISTANCE

We evaluate 50 test inputs. For each input, candidate generation, center selection, and probability certification use mutually independent sample sets of $N_1 = 5{,}000$, $N_{\mathrm{sel}} = 5{,}000$, and $N_2 = 10{,}000$ noisy predictions, respectively. The one-sided Clopper–Pearson lower bound uses $\alpha = 0.005$, corresponding to a per-input confidence level of $99.5\%$.

We evaluate smoothing levels $\sigma \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}$ and output tolerances $r \in \{0.08, 0.10, 0.12, 0.14, 0.16, 0.18\}$. Because Jaccard distance is a metric, the neighborhood radius is $r/2$. At $\zeta = 0$, only samples with $\hat{R}_i > 0$ are counted. If the face detector returns multiple boxes, we retain the first; a missing or invalid box is assigned zero IoU.

## J.2 IMAGE RECONSTRUCTION WITH PERCEPTUAL DISTANCE

We evaluate 50 test inputs using independent sample sets with $N_1 = N_{\mathrm{sel}} = 5{,}000$ and $N_2 = 10{,}000$. The one-sided Clopper–Pearson lower bound uses $\alpha = 0.005$. We evaluate $\sigma \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}$ and $r \in \{0.012, 0.014, 0.016, 0.018, 0.020, 0.022\}$. Since perceptual distance satisfies the relaxed triangle inequality with $\kappa = 2$, the neighborhood radius is $r/4$.

For the reconstruction-quality criterion, let $\hat{y}_i$ denote the clean reconstruction of input $x_i$. Sample $i$ satisfies the criterion when

$$d_{\mathrm{perc}}(\hat{y}_i, x_i) \le \tau_{\mathrm{perc}}. \tag{58}$$

We set $\tau_{\mathrm{perc}} = 0.30$, a single rounded global threshold based on the empirical 90th percentile of clean reconstruction errors on a held-out calibration set. With this threshold, approximately $92\%$ of the clean reconstructions satisfy the quality criterion. The threshold is selected once and kept fixed across all smoothing levels, output tolerances, and certification settings. At $\zeta = 0$, only samples with $\hat{R}_i > 0$ are counted.

## J.3 IMAGE-TEXT RETRIEVAL WITH COSINE DISTANCE

We sample 50 images from the official MS COCO 2014 validation set and construct a fixed retrieval bank by including three corresponding human-annotated ground-truth captions for each image. The resulting bank contains 150 captions, with exactly three correct captions associated with each query image. Images and captions are represented by 768-dimensional $\ell_2$-normalized CLIP embeddings. Candidate generation and center selection each use 5,000 noisy predictions, while probability certification uses 10,000 independent predictions. The one-sided Clopper–Pearson lower confidence bound is computed with $\alpha = 0.005$.

We evaluate $\sigma \in \{0.10, 0.15, 0.20, 0.25, 0.30\}$, $r \in \{0.006, 0.008, 0.010, 0.012, 0.014\}$, and retrieval depths $K \in \{1, 5, 10\}$. With $\kappa = 2$, the neighborhood radius is $r/4$, yielding $\{0.0015, 0.0020, 0.0025, 0.0030, 0.0035\}$. At $\zeta = 0$, only samples with $\hat{R}_i > 0$ are counted.

## J.4 IMAGE RECONSTRUCTION WITH PSNR-BASED DISCREPANCY

We evaluate 50 test inputs using four mutually independent sample sets for candidate generation, center selection, probability estimation, and competing-center search. Their sizes are $N_1 = N_{\mathrm{sel}} = N_2 = M = 1{,}000$, where $M$ denotes the competing-center sample size. We set $\alpha = \eta = 0.005$ for the selected mass and sampled-competitor bounds. The evaluated parameter grid is

$$\sigma \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}, \qquad r \in \{0.021, 0.023, 0.025, 0.027, 0.029, 0.031\}. \tag{59}$$

In this experiment, we do not invoke a relaxed-triangle bound for $d_{\mathrm{PSNR}}$. Instead, we use radius-$r/2$ neighborhoods and sampled competing centers, and treat $\hat{R}_{\mathrm{gen}}$ as an empirical robustness estimate with respect to the full output space.

### J.5 Image Reconstruction with SSIM-Based Discrepancy

We again evaluate 50 test inputs with four mutually independent sample sets of size $N_1 = N_{\text{sel}} = N_2 = M = 1{,}000$ and set $\alpha = \eta = 0.005$. The evaluated parameter grid is

$$\sigma \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}, \qquad r \in \{0.006, 0.010, 0.014, 0.018, 0.024, 0.032\}. \quad (60)$$

As in the PSNR-based experiment, we use radius-$r/2$ neighborhoods and report $\hat{R}_{\text{gen}}$ as an empirical robustness estimate over the full output space.

## K Additional Results of Certified Stability Analysis

Section 5.3 presents $C_{\text{task}}(\zeta)$ by fixing the smoothing noise level $\sigma$ and varying the output tolerance $r$. We first complete the face-detection results with the additional noise levels $\sigma \in \{0.15, 0.25, 0.35\}$. We then provide a complementary view that fixes $r$ and compares the evaluated noise levels, thereby isolating the effect of smoothing under the same output requirement. The face-detection and image-reconstruction results are reported together below because both use $C_{\text{task}}(\zeta)$ defined in Section 5.3. We then report $C_{\text{task}}(\zeta)$ for image-text retrieval in the same appendix. In this setting, $C_{\text{task}}(\zeta)$ can be regarded as the clean Recall@K with certified embedding stability. At zero perturbation budget, only samples with a strictly positive certified radius are counted.

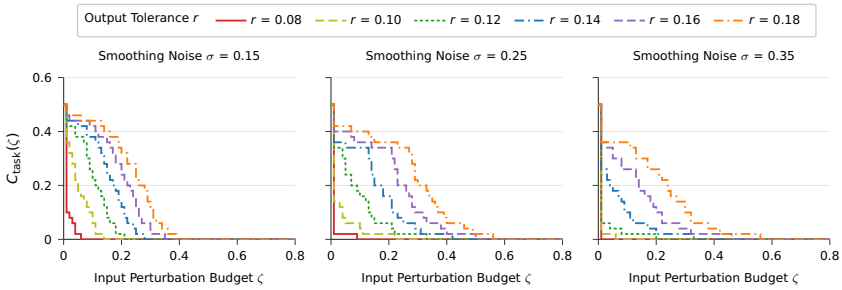



**Figure 4.** Additional results of $C_{\text{task}}(\zeta)$ at $\sigma \in \{0.15, 0.25, 0.35\}$ for face detection. Each panel fixes $\sigma$, and the curves represent output tolerances $r$.

The main-text figure reports $\sigma \in \{0.1, 0.2, 0.3\}$, while Fig. 4 supplies the three intermediate and higher noise levels. Together, these figures cover $\sigma \in \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35\}$ under the fixed-$\sigma$ organization.

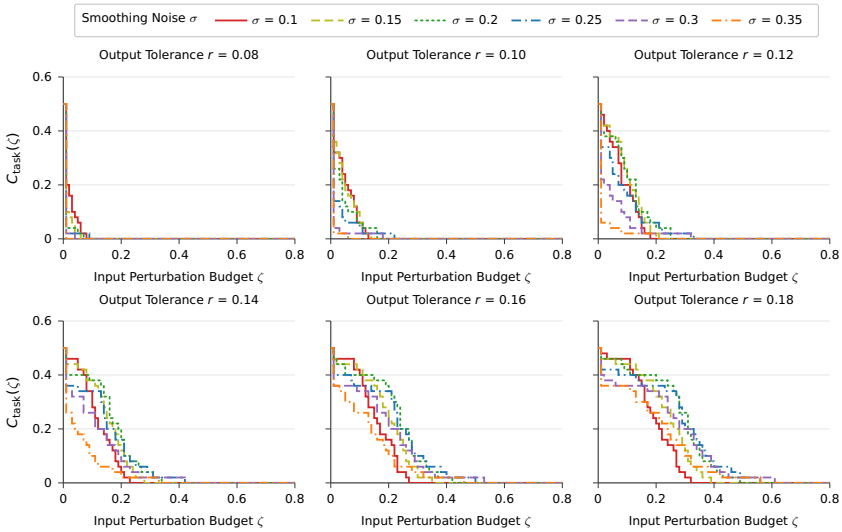



**Figure 5.** Complementary fixed-tolerance view of $C_{\text{task}}(\zeta)$ for face detection. Each panel fixes $r$ and compares all six evaluated smoothing noise levels.

Fig. 5 fixes the output tolerance $r$ in each panel and compares different smoothing noise levels. It shows that $C_{\text{task}}(\zeta)$ is nonincreasing in $\zeta$ by construction. A larger $\sigma$ can extend the range of nonzero certification but can also change the concentration of noisy bounding boxes and the fraction of clean predictions satisfying the localization criterion. The fixed-$r$ view therefore isolates the smoothing trade-off under the same output requirement.

We also examine image reconstruction. We first supplement the fixed-$\sigma$ results and then present the complementary fixed-tolerance view for image reconstruction.

The image-reconstruction panels in Fig. 2 report $\sigma \in \{0.10, 0.20, 0.30\}$, while Fig. **??** supplies $\sigma \in \{0.15, 0.25, 0.35\}$. Together, the two figures cover all six evaluated smoothing noise levels. For every fixed $(\sigma, r)$ configuration, $C_{\text{task}}(\zeta)$ decreases as $\zeta$ increases, because a larger required input perturbation budget imposes a stricter certification requirement and therefore leaves fewer samples satisfying it.

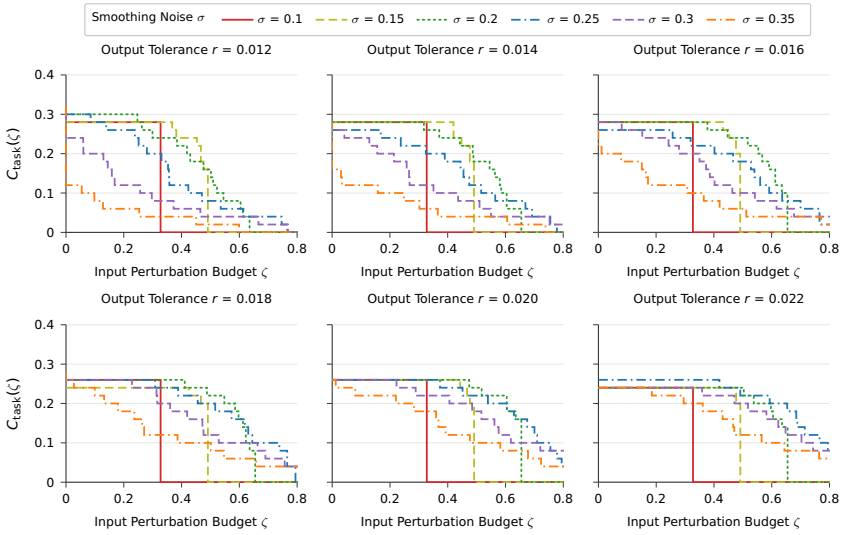



**Figure 6.** Complementary fixed-tolerance view of $C_{\text{task}}(\zeta)$ for image reconstruction. Each panel fixes one of the six evaluated output tolerances $r$ and compares all six smoothing noise levels $\sigma \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}$.

Fig. 6 presents the complete fixed-tolerance view. Increasing $r$ generally makes it easier to obtain a large certified input radius, while the effect of $\sigma$ is not necessarily monotone because smoothing affects both the certified radius and the concentration of reconstructed outputs. The reconstruction-quality criterion in Eq.(58) uses the same fixed threshold $\tau_{\text{perc}} = 0.30$ for every configuration. The curves therefore compare certified coverage under a common clean reconstruction-quality requirement.

We finally report the complete results of $C_{\text{task}}(\zeta)$ for image-text retrieval across the evaluated smoothing noise levels, output tolerances, and retrieval depths $K \in \{1, 5, 10\}$. The corresponding experimental configuration is provided in Appendix J. Fig. 7 summarizes the results across all evaluated smoothing noise levels, output tolerances, and retrieval depths. It shows that $C_{\text{task}}(\zeta)$ decreases as the required input perturbation budget $\zeta$ increases, since fewer examples remain certified. Increasing $K$ relaxes the retrieval-success criterion and therefore increases the fraction of samples counted as successful. Increasing the output tolerance enlarges the neighborhood used by ContRS, which can increase the estimated neighborhood mass and lead to a larger certified radius. The effect of the smoothing noise is less direct: a larger $\sigma$ increases the scale of the certified radius, but may also disperse the noisy embedding distribution and reduce the neighborhood mass.

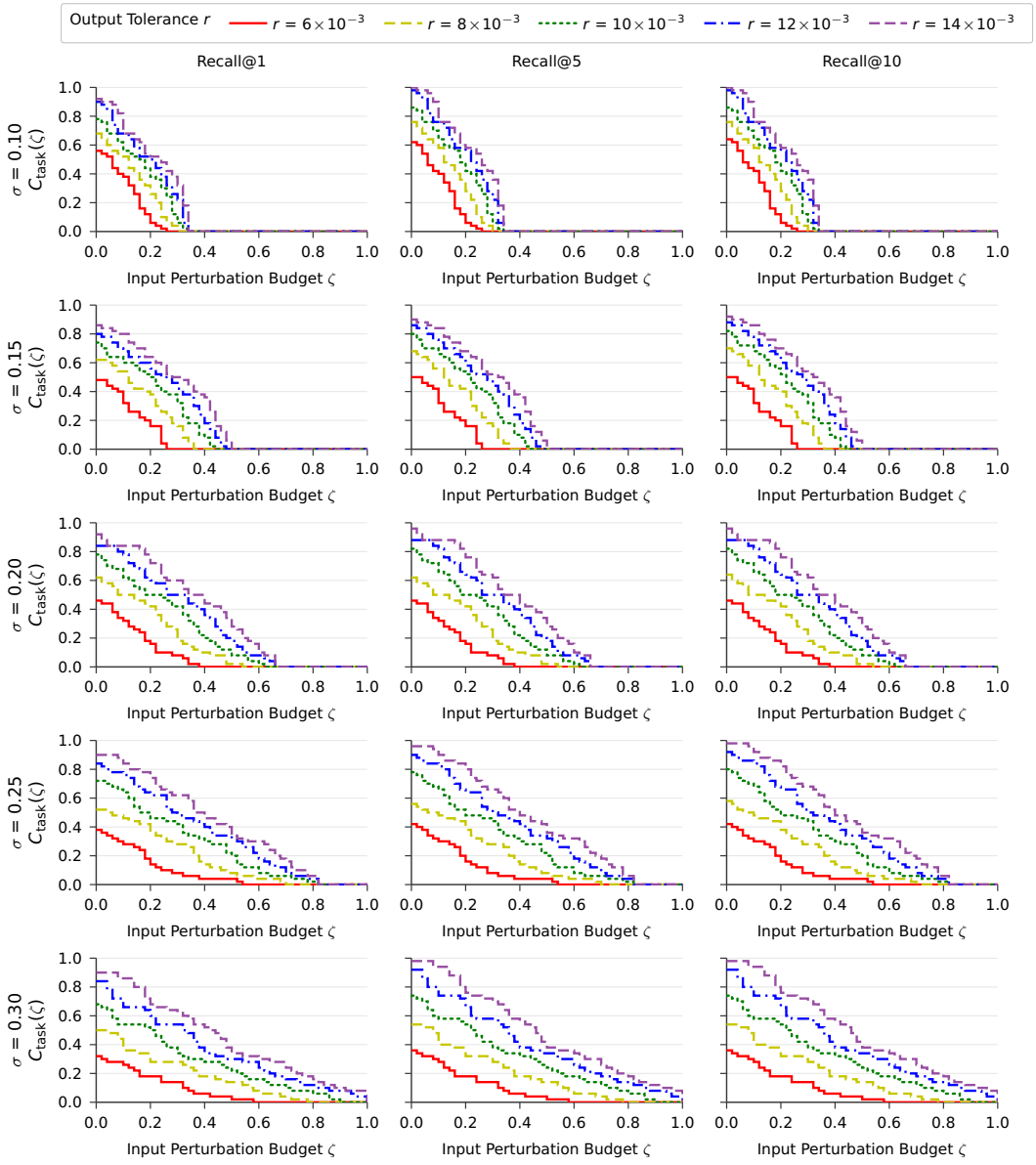



**Figure 7.** $C_{\text{task}}(\zeta)$ for the smoothing levels $\sigma \in \{0.10, 0.15, 0.20, 0.25, 0.30\}$. Rows fix $\sigma$, columns fix the retrieval depth $K$, and curves represent output tolerances $r$. Here, Recall@$K$ denotes the fraction of queries with a correct caption in the clean top-$K$ results and a positive certified embedding-stability radius of at least $\zeta$.

## L Monte Carlo Sample-Size Sensitivity

The preceding experiments use fixed sample sizes $N_1$, $N_{\text{sel}}$, and $N_2$ for candidate generation, center selection, and probability certification, respectively, in Algorithm 1. To assess sensitivity, we vary one sample size at a time while holding all other settings fixed. For each task, we intentionally choose a slightly larger output tolerance outside the main experimental grid: $(\sigma, r) = (0.1, 0.2)$ for face detection, $(0.25, 0.028)$ for image reconstruction, and $(0.25, 0.12)$ for image-text retrieval. These settings are chosen because the mean certified input radius remains strictly positive across all evaluated sample sizes, allowing the effect of changing the Monte Carlo sample sizes on the certified radius and smoothing error to be compared more clearly.

| Task | Candidate Generation | | | Center Selection | | | Probability Certification | | |
|---|---|---|---|---|---|---|---|---|---|
| | $N_1$ | mean $\hat{R}$ | mean smoothing error | $N_{\text{sel}}$ | mean $\hat{R}$ | mean smoothing error | $N_2$ | mean $\hat{R}$ | mean smoothing error |
| Face Detection | 6,000 | 0.2016 | 0.0614 | 6,000 | 0.1974 | 0.0613 | 12,000 | 0.2116 | 0.0591 |
| | 7,000 | 0.2014 | 0.0628 | 7,000 | 0.197038 | 0.0618 | 14,000 | 0.2131 | 0.0585 |
| | 8,000 | 0.2010 | 0.0632 | 8,000 | 0.1978 | 0.0636 | 16,000 | 0.2135 | 0.0584 |
| | 9,000 | 0.2015 | 0.0622 | 9,000 | 0.1988 | 0.0624 | 18,000 | 0.2144 | 0.0581 |
| Image Reconstruction | 6,000 | 0.6053 | 0.0015 | 6,000 | 0.5889 | 0.0016 | 12,000 | 0.6082 | 0.0015 |
| | 7,000 | 0.6050 | 0.0015 | 7,000 | 0.5887 | 0.0016 | 14,000 | 0.6126 | 0.0015 |
| | 8,000 | 0.6062 | 0.0015 | 8,000 | 0.5893 | 0.0016 | 16,000 | 0.6153 | 0.0015 |
| | 9,000 | 0.6058 | 0.0015 | 9,000 | 0.5901 | 0.0016 | 18,000 | 0.6191 | 0.0015 |
| Image-text Retrieval | 6,000 | 0.3819 | 0.2085 | 6,000 | 0.4019 | 0.1980 | 12,000 | 0.3837 | 0.2077 |
| | 7,000 | 0.3851 | 0.2080 | 7,000 | 0.4023 | 0.1981 | 14,000 | 0.3866 | 0.2072 |
| | 8,000 | 0.3862 | 0.2064 | 8,000 | 0.4015 | 0.1982 | 16,000 | 0.3880 | 0.2073 |
| | 9,000 | 0.3908 | 0.2066 | 9,000 | 0.4012 | 0.1985 | 18,000 | 0.3889 | 0.2072 |

Table 5: Sensitivity to Monte Carlo sample sizes in Algorithm 1. We vary the sample sizes for candidate generation ($N_1$), center selection ($N_{\text{sel}}$), and probability certification ($N_2$) one at a time while holding all other settings fixed. We report the mean certified input radius and mean smoothing error for each task.

As shown in Table 5, across all three tasks, increasing the candidate-generation or center-selection sample size leads to only minor changes in the mean certified input radius and mean smoothing error. Increasing the probability-certification sample size modestly improves the mean certified radius, particularly for image reconstruction, while leaving smoothing error nearly unchanged. These results suggest that, under the selected nonzero-radius configurations, moderate changes in the Monte Carlo sample sizes produce only limited variation in the certified radius and smoothing error.

# M Adversarial Attack Settings

Section 5.4 examines whether the output variation of ContRS remains within the prescribed output tolerance under FGM and PGD attacks. In Algorithm 1, center selection uses $N_{\text{sel}}$ predictions and this indicator $\mathbf{1}\{d(y_i, y'_j) \le s\}$ to select the center $\hat{z}$ with the largest empirical neighborhood mass. Because the indicator function is non-differentiable, we use a surrogate objective to generate adversarial inputs. This appendix describes the surrogate objective and the numerical settings used for FGM and PGD in the perceptual-distance evaluation.

## M.1 Surrogate Objective

For each clean input $x$, we first obtain the estimated smoothed prediction $\hat{z}$. We then generate an adversarial input $x_{\text{adv}} = x + \delta$, where the adversarial perturbation $\delta$ satisfies

$$\|\delta\|_2 \le \zeta_{\text{atk}}, \tag{61}$$

where $\zeta_{\text{atk}}$ denotes the perturbation budget. In Algorithm 1, center selection uses $N_{\text{sel}}$ predictions to select the center $\hat{z}$ with the largest empirical neighborhood mass. For each candidate output $y_i$, the neighborhood mass is estimated using

$$k_i = \sum_{j=1}^{N_{\text{sel}}} \mathbf{1}\{d(y_i, y'_j) \le s\}. \tag{62}$$

Because this center-selection step involves the indicator function, it is non-differentiable. We therefore use a surrogate objective based on predictions under Gaussian input noise. Specifically, given $m$ Gaussian noise samples $\epsilon_j \sim \mathcal{N}(0, \sigma^2 I)$, we define

$$\mathcal{L}_{\text{perc}}(x, \delta) = d_{\text{perc}}\left(\frac{1}{m}\sum_{j=1}^{m} f(x + \delta + \epsilon_j), \hat{z}\right). \tag{63}$$

FGM and PGD maximize this objective to generate the adversarial input. For each adversarial input, we rerun ContRS to obtain the estimated smoothed prediction $\hat{z}_{\text{adv}}$ and measure the output variation between $\hat{z}_{\text{adv}}$ and $\hat{z}$.

### M.2 Gradient-Based Attack Methods

We evaluate the two gradient-based attacks used in the main paper: FGM and PGD.

**Fast Gradient Method (FGM).** Following Kireev et al. (2022), we use the $\ell_2$ variant of FGM and perform a single normalized gradient-ascent step on the surrogate objective. Starting from $\delta = 0$, we compute

$$g = \nabla_\delta \mathcal{L}_{\mathrm{perc}}(x, \delta) \tag{64}$$

and set

$$\delta = \zeta_{\mathrm{atk}} \frac{g}{\|g\|_2}. \tag{65}$$

**Projected Gradient Descent (PGD).** Following Madry et al. (2017), PGD starts from a random perturbation in the $\ell_2$ ball and repeatedly performs normalized gradient ascent followed by projection:

$$\delta \leftarrow \Pi_{\|\delta\|_2 \le \zeta_{\mathrm{atk}}} \left( \delta + \gamma \frac{\nabla_\delta \mathcal{L}_{\mathrm{perc}}(x, \delta)}{\|\nabla_\delta \mathcal{L}_{\mathrm{perc}}(x, \delta)\|_2} \right), \tag{66}$$

where $\Pi$ denotes projection onto the $\ell_2$ ball and $\gamma$ is the step size. We use multiple random restarts and retain the adversarial input with the largest value of the surrogate objective. The numerical attack parameters used for the perceptual-distance evaluation are specified below.

### M.3 Perceptual-Distance Attack Settings

The perceptual-distance attack results are reported in Section 5.4. This subsection provides the corresponding numerical settings.

We use the same pretrained VAE, CIFAR-10 test inputs, and perceptual distance as in the image-reconstruction experiments. The same 50 test inputs are used across all attack settings. Candidate generation, center selection, and probability certification use $N_1 = 5{,}000$, $N_{\mathrm{sel}} = 5{,}000$, and $N_2 = 10{,}000$ predictions, respectively.

For perceptual distance, we use $\kappa = 2$, so the neighborhood radius is $r/(2\kappa) = r/4$. We evaluate

$$r \in \{0.012, 0.014, 0.016, 0.018, 0.020, 0.022\} \tag{67}$$

and

$$\sigma \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}. \tag{68}$$

For each input, the perturbation budget is defined relative to its certified input radius $\hat{R}$:

$$\zeta_{\mathrm{atk}} \in \{0.5\hat{R},\ \hat{R},\ 1.5\hat{R},\ 2\hat{R}\}. \tag{69}$$

For every $(\sigma, r)$ setting, we evaluate FGM and PGD at all four perturbation budgets. For samples with $\hat{R} > 0$, the $0.5\hat{R}$ budget keeps every feasible perturbation strictly inside the certified input ball. The $\hat{R}$ budget may reach its boundary, whereas the $1.5\hat{R}$ and $2\hat{R}$ budgets may extend beyond it. The certificate applies only to realized perturbations satisfying $\|\delta\|_2 < \hat{R}$.

For FGM, we use eight Gaussian noise samples to compute the gradient, and the resulting adversarial input is clipped to $[-1, 1]$. For PGD, we use 40 optimization steps, five random restarts, and four Gaussian noise samples per gradient computation. The step size is $\gamma = 2.5\zeta_{\mathrm{atk}}/40$. Each restart is initialized on the $\ell_2$ sphere of radius $\zeta_{\mathrm{atk}}$, and the adversarial input with the largest value of the surrogate objective is retained.

## N Additional Adversarial Experiments

This appendix complements the main perceptual-distance attack evaluation with an attack-parameter sensitivity study and evaluations using the PSNR- and SSIM-based discrepancies. Since $\hat{R}_{\mathrm{gen}}$ is an empirical robustness estimate over the full output space and does not provide a formal finite-sample certificate, we therefore use adversarial attacks as an indirect empirical evaluation, following prior studies Shao et al. (2023). We examine whether the deviations between the outputs of clean and adversarial samples remain within the prescribed tolerance $r$ at perturbation budgets scaled by $\hat{R}_{\mathrm{gen}}$.

### N.1 Attack-Parameter Sensitivity

To assess sensitivity to the attack parameters, we vary the FGM and PGD parameters one at a time while holding the remaining attack settings fixed. We fix the prescribed output tolerance at $r = 0.018$ and evaluate $\sigma \in \{0.15, 0.25, 0.35\}$. For FGM, we vary the number of Gaussian noise samples. For PGD, we vary the number of optimization steps, random restarts, and Gaussian noise samples one at a time. The default settings use eight Gaussian noise samples for FGM and 40 steps, five random restarts, and four Gaussian noise samples for PGD. Only inputs with positive finite-sample certified radii are attacked, giving 50, 41, and 23 samples for the three smoothing noise levels, respectively.

| **Attack Parameter** | **Smoothing Noise** | **MND at the $2\hat{R}$ Attack Budget** | | | |
|---|---|---|---|---|---|
| FGM noise samples $m \in \{1, 4, 8, 16\}$ | $\sigma = 0.15$ | 0.1200 | 0.1280 | 0.1320 | 0.1315 |
| | $\sigma = 0.25$ | 0.1075 | 0.1070 | 0.1100 | 0.1075 |
| | $\sigma = 0.35$ | 0.0735 | 0.0745 | 0.0840 | 0.0830 |
| PGD steps $T \in \{10, 20, 40, 80\}$ | $\sigma = 0.15$ | 0.1650 | 0.1580 | 0.1630 | 0.1570 |
| | $\sigma = 0.25$ | 0.1320 | 0.1295 | 0.1325 | 0.1295 |
| | $\sigma = 0.35$ | 0.0865 | 0.0900 | 0.0830 | 0.0865 |
| PGD restarts $J \in \{1, 3, 5, 10\}$ | $\sigma = 0.15$ | 0.1615 | 0.1610 | 0.1660 | 0.1615 |
| | $\sigma = 0.25$ | 0.1290 | 0.1355 | 0.1400 | 0.1365 |
| | $\sigma = 0.35$ | 0.0885 | 0.0880 | 0.0875 | 0.0840 |
| PGD noise samples $m \in \{1, 4, 8, 16\}$ | $\sigma = 0.15$ | 0.1625 | 0.1575 | 0.1575 | 0.1575 |
| | $\sigma = 0.25$ | 0.1275 | 0.1380 | 0.1415 | 0.1375 |
| | $\sigma = 0.35$ | 0.0820 | 0.0815 | 0.0835 | 0.0810 |

Table 6: Attack-parameter sensitivity for image reconstruction with perceptual distance at prescribed output tolerance $r = 0.018$. Each entry reports MND under the $2\hat{R}$ attack budget. Within each block, the four columns follow the parameter values listed in the first column.

As shown in Table 6, MND remains below one across all evaluated attack configurations. Increasing the number of steps, restarts, or Gaussian noise samples does not produce a systematic increase in the observed deviation. These results indicate limited sensitivity to the evaluated attack parameters. Because $2\hat{R}$ lies outside the certified input region, this analysis is an empirical stress test and does not extend the formal guarantee.

### N.2 PSNR-Based Discrepancy

The PSNR-based reconstruction experiment uses the general-discrepancy setting in Appendix G. Since we do not assume that $d_{\mathrm{PSNR}}$ satisfies the metric or relaxed-triangle conditions required by the finite-sample certificate, we use the sampled competing centers in Appendix E. The resulting $\hat{R}_{\mathrm{gen}}$ is therefore treated as an empirical robustness estimate with respect to the full output space.

We use VAE reconstruction on 50 CIFAR-10 test inputs. Candidate generation, center selection, probability certification, and competing-center search use sample sizes

$$N_1 = N_{\mathrm{sel}} = N_2 = M = 1{,}000, \tag{70}$$

respectively, where $M$ is the competing-center sample size. We set $\alpha = \eta = 0.005$, where $\alpha$ and $\eta$ are the failure probabilities used for the selected neighborhood mass and sampled competing centers, respectively. We evaluate

$$\sigma \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}, \qquad r \in \{0.021, 0.023, 0.025, 0.027, 0.029, 0.031\}, \tag{71}$$

where $\sigma$ is the Gaussian noise scale and $r$ is the prescribed output tolerance. Only inputs with $\hat{R}_{\mathrm{gen}} > 0$ are attacked.

For the PSNR-based discrepancy, we use the same surrogate construction as in Appendix M. Given $m$ Gaussian noise samples $\epsilon_j \sim \mathcal{N}(0, \sigma^2 I)$, we define

$$\mathcal{L}_{\mathrm{PSNR}}(x, \delta) = d_{\mathrm{PSNR}} \left( \frac{1}{m} \sum_{j=1}^{m} f(x + \delta + \epsilon_j), f(x) \right). \tag{72}$$

FGM and PGD maximize this objective to generate the adversarial input. FGM uses $m = 8$ Gaussian noise samples and clips the adversarial input to $[-1, 1]$. PGD uses 40 optimization steps, five random restarts, and $m = 4$ Gaussian noise samples per gradient computation. The step size is $\gamma = 2.5\zeta_{\mathrm{atk}}/40$.

For each input, the perturbation budget is defined relative to its empirical robustness estimate:

$$\zeta_{\mathrm{atk}} \in \{\hat{R}_{\mathrm{gen}},\ 1.25\hat{R}_{\mathrm{gen}},\ 1.5\hat{R}_{\mathrm{gen}},\ 2\hat{R}_{\mathrm{gen}}\}, \tag{73}$$

where $\zeta_{\mathrm{atk}}$ denotes the perturbation budget.

| PSNR-based Discrepancy | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Attack strategy** | | | | | | | |
| $\sigma$ | **Attack** | 0.021 | 0.023 | 0.025 | 0.027 | 0.029 | 0.031 |
| 0.10 | FGM | 0.485 | 0.485 | 0.481 | 0.453 | 0.421 | 0.393 |
| | PGD | 0.485 | 0.487 | 0.484 | 0.453 | 0.422 | 0.395 |
| 0.15 | FGM | – | 0.485 | 0.484 | 0.480 | 0.459 | 0.429 |
| | PGD | – | 0.485 | 0.486 | 0.483 | 0.460 | 0.430 |
| 0.20 | FGM | – | 0.481 | 0.483 | 0.482 | 0.478 | 0.458 |
| | PGD | – | **0.487** | 0.484 | 0.482 | 0.482 | 0.460 |
| 0.25 | FGM | – | – | 0.472 | 0.483 | 0.479 | 0.476 |
| | PGD | – | – | 0.474 | 0.484 | 0.481 | 0.478 |
| 0.30 | FGM | – | – | 0.474 | 0.479 | 0.482 | 0.475 |
| | PGD | – | – | 0.476 | 0.481 | 0.484 | 0.479 |
| 0.35 | FGM | – | – | **0.489** | 0.473 | 0.478 | 0.478 |
| | PGD | – | – | 0.479 | 0.474 | 0.481 | 0.481 |

Table 7: Mean Base-Output Normalized Deviation (MBND) under FGM and PGD for samples with $\hat{R}_{\mathrm{gen}} > 0$ at perturbation budgets $\{\hat{R}_{\mathrm{gen}}, 1.25\hat{R}_{\mathrm{gen}}, 1.5\hat{R}_{\mathrm{gen}}, 2\hat{R}_{\mathrm{gen}}\}$. MBND compares the adversarial and clean base-predictor outputs relative to the prescribed output tolerance. Dashes indicate that no sample has $\hat{R}_{\mathrm{gen}} > 0$; bold cells mark per-attack maxima.

For input $x_i$, let $\delta_{i,\lambda}$ denote the adversarial perturbation generated with perturbation budget $\lambda\hat{R}_{\mathrm{gen},i}$, where $\lambda \in \{1, 1.25, 1.5, 2\}$. For each $(\sigma, r)$ setting and attack method, we define

$$\mathrm{MBND}_P = \frac{1}{4N_+} \sum_{i=1}^{N_+} \sum_{\lambda \in \{1, 1.25, 1.5, 2\}} \frac{d_{\mathrm{PSNR}}(f(x_i + \delta_{i,\lambda}), f(x_i))}{r}, \tag{74}$$

where $N_+$ denotes the number of test inputs with $\hat{R}_{\mathrm{gen},i} > 0$ under the corresponding $(\sigma, r)$ setting. Thus, each reported MBND value averages the normalized deviations over all retained test inputs and all four perturbation budgets.

We report the mean of $\mathrm{BND}_{P,i}$ over the samples with $\hat{R}_{\mathrm{gen}} > 0$ as the Mean Base-Output Normalized Deviation (MBND). Unlike MND in Section 5.4, which compares estimated smoothed predictions, MBND compares base-predictor outputs.

Table 7 reports the PSNR-based attack results. MBND remains below one across all evaluated settings, and FGM and PGD give similar values. Thus, the average deviation between the adversarial and clean base-predictor outputs remains below the prescribed output tolerance under the evaluated attacks for samples with $\hat{R}_{\mathrm{gen}} > 0$. Since a positive $\hat{R}_{\mathrm{gen}}$ requires $q_U < \hat{p}_L$, these samples satisfy the probability-mass separation between the selected neighborhood and the sampled competing centers used by ContRS. The results show stable output behavior under perturbation budgets scaled by the resulting $\hat{R}_{\mathrm{gen}}$.

### N.3 SSIM-BASED DISCREPANCY

The SSIM-based reconstruction experiment uses the general-discrepancy setting in Appendix G. Since we do not assume that $d_{\mathrm{SSIM}}$ satisfies the metric or relaxed-triangle conditions, we use the sampled competing centers in Appendix E. The resulting $\hat{R}_{\mathrm{gen}}$ is treated as an empirical robustness estimate with respect to the full output space.

We use VAE reconstruction on 50 CelebA test inputs. Candidate generation, center selection, probability certification, and competing-center search use sample sizes

$$N_1 = N_{\mathrm{sel}} = N_2 = M = 1{,}000, \tag{75}$$

respectively, where $M$ is the competing-center sample size. We set $\alpha = \eta = 0.005$, where $\alpha$ and $\eta$ are the failure probabilities used for the selected neighborhood mass and sampled competing centers, respectively. We evaluate

$$\sigma \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}, \qquad r \in \{0.006, 0.010, 0.014, 0.018, 0.024, 0.032\}, \tag{76}$$

where $\sigma$ is the Gaussian noise scale and $r$ is the prescribed output tolerance. Only inputs with $\hat{R}_{\mathrm{gen}} > 0$ are attacked.

For the SSIM-based discrepancy, we use the same surrogate construction as in Appendix M. Given $m$ Gaussian noise samples $\epsilon_j \sim \mathcal{N}(0, \sigma^2 I)$, we define

$$\mathcal{L}_{\mathrm{SSIM}}(x, \delta) = d_{\mathrm{SSIM}}\left(\frac{1}{m}\sum_{j=1}^{m} f(x + \delta + \epsilon_j), f(x)\right). \tag{77}$$

FGM and PGD maximize this objective to generate the adversarial input. FGM uses $m = 8$ Gaussian noise samples and clips the adversarial input to $[-1, 1]$. PGD uses 40 optimization steps, five random restarts, and $m = 4$ Gaussian noise samples per gradient computation. The step size is $\gamma = 2.5\zeta_{\mathrm{atk}}/40$.

| SSIM-based Discrepancy | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Attack strategy** | | | | | | | |
| $\sigma$ | **Attack** | 0.006 | 0.010 | 0.014 | 0.018 | 0.024 | 0.032 |
| 0.10 | FGM | 0.336 | 0.310 | 0.278 | 0.228 | 0.175 | 0.134 |
| | PGD | 0.364 | 0.339 | 0.303 | 0.249 | 0.188 | 0.144 |
| 0.15 | FGM | 0.365 | 0.335 | 0.314 | 0.315 | 0.301 | 0.250 |
| | PGD | 0.386 | 0.360 | 0.344 | 0.343 | 0.332 | 0.264 |
| 0.20 | FGM | 0.356 | 0.359 | 0.333 | 0.321 | 0.311 | 0.295 |
| | PGD | 0.372 | 0.376 | 0.369 | 0.349 | 0.344 | 0.328 |
| 0.25 | FGM | 0.382 | 0.359 | 0.356 | 0.348 | 0.326 | 0.309 |
| | PGD | **0.394** | 0.362 | 0.379 | 0.370 | 0.352 | 0.346 |
| 0.30 | FGM | – | **0.383** | 0.358 | 0.351 | 0.335 | 0.312 |
| | PGD | – | 0.376 | 0.377 | 0.377 | 0.368 | 0.342 |
| 0.35 | FGM | – | 0.382 | 0.361 | 0.368 | 0.349 | 0.318 |
| | PGD | – | 0.337 | 0.383 | 0.381 | 0.371 | 0.345 |

Table 8: Mean base-output normalized deviation (MBND) under FGM and PGD, averaged over available attack records with $\hat{R}_{\mathrm{gen}} > 0$ and budgets $\{\hat{R}_{\mathrm{gen}}, 1.25\hat{R}_{\mathrm{gen}}, 1.5\hat{R}_{\mathrm{gen}}, 2\hat{R}_{\mathrm{gen}}\}$. The statistic compares attacked and clean base-predictor outputs relative to the prescribed output tolerance. Dashes indicate no eligible record; bold cells mark per-attack maxima.

For each input, the perturbation budget is defined relative to its empirical robustness estimate:

$$\zeta_{\mathrm{atk}} \in \{\hat{R}_{\mathrm{gen}},\ 1.25\hat{R}_{\mathrm{gen}},\ 1.5\hat{R}_{\mathrm{gen}},\ 2\hat{R}_{\mathrm{gen}}\}, \tag{78}$$

where $\zeta_{\mathrm{atk}}$ denotes the perturbation budget.

For input $x_i$, let $\delta_{i,\lambda}$ denote the adversarial perturbation generated with perturbation budget $\lambda\hat{R}_{\mathrm{gen},i}$, where $\lambda \in \{1, 1.25, 1.5, 2\}$. For each $(\sigma, r)$ setting and attack method, we define

$$\mathrm{MBND}_S = \frac{1}{4N_+}\sum_{i=1}^{N_+}\sum_{\lambda\in\{1,1.25,1.5,2\}} \frac{d_{\mathrm{SSIM}}(f(x_i + \delta_{i,\lambda}), f(x_i))}{r}, \tag{79}$$

where $N_+$ denotes the number of test inputs with $\hat{R}_{\mathrm{gen},i} > 0$ under the corresponding $(\sigma, r)$ setting. Thus, each reported MBND value averages the normalized deviations over all retained test inputs and all four perturbation budgets.

Table 8 reports the SSIM-based attack results. MBND remains below one across all evaluated settings. The values vary across the attack settings, with PGD producing several of the largest values. Nevertheless, the average deviation between the adversarial and clean base-predictor outputs remains below the prescribed output tolerance for samples with $\hat{R}_{\mathrm{gen}} > 0$. Since a positive $\hat{R}_{\mathrm{gen}}$ requires $q_U < \hat{p}_L$, these samples satisfy the probability-mass separation between the selected neighborhood and the sampled competing centers used by ContRS. The results show stable output behavior under the SSIM-based discrepancy and perturbation budgets scaled by the resulting $\hat{R}_{\mathrm{gen}}$.