DUAL RANDOMIZED SMOOTHING: BEYOND GLOBAL NOISE VARIANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized Smoothing (RS) is a prominent technique for certifying the robustness of neural networks against adversarial perturbations. With RS, achieving high accuracy at small radii requires a small noise variance, while achieving high accuracy at large radii requires a large noise variance. However, the global noise variance used in the standard RS formulation leads to a fundamental limitation: there exists no global noise variance that simultaneously achieves strong performance at both small and large radii. To break through the global variance limitation, we propose a dual RS framework which enables input-dependent noise variances. To achieve that, we first prove that RS remains valid with input-dependent noise variances, provided the variance is locally constant around each input. Building on this result, we introduce two components which form our dual RS framework: (i) a variance estimator first predicts an optimal noise variance for each input, (ii) this estimated variance is then used by a standard RS classifier. The variance estimator is independently smoothed via RS to ensure local constancy, enabling flexible design. We also introduce training strategies to iteratively optimize the two components involved in the framework. Extensive experiments on the CIFAR-10 dataset demonstrate that our dual RS method provides strong performance for both small and large radii—unattainable with global noise variance—while incurring only a 60% computational overhead at inference. Moreover, it consistently outperforms prior input-dependent noise approaches across most radii, with particularly large gains at radii 0.5, 0.75, and 1.0, achieving relative improvements of 19.2%, 24.2%, and 20.6%, respectively. On IMAGENET, dual RS remains effective across all radii, with roughly 1.5x performance advantages at radii 0.5, 1.0 and 1.5. Additionally, the proposed dual RS framework naturally provides a routing perspective for certified robustness, improving the accuracy-robustness trade-off with off-the-shelf expert RS models.

1 INTRODUCTION

Deep neural networks have achieved remarkable success across diverse tasks but remain highly vulnerable to adversarial attacks—small, carefully crafted perturbations that can lead to incorrect or unexpected predictions. This vulnerability has made adversarial robustness, which ensures consistent model outputs under small perturbations, a critical research focus. As heuristic defenses are often unreliable (Athalye et al., 2018; Croce & Hein, 2020), methods with provable robustness guarantees have become increasingly important.

Randomized Smoothing (RS) is a prominent technique for certifying robustness against ℓ_2-norm adversarial perturbations. It constructs a smoothed classifier by adding Gaussian noise to the input and taking the majority vote of predictions, thereby ensuring consistent outputs within a certified neighborhood. Prior work has primarily focused on two directions: (1) training-based RS, which improves robustness by explicitly training the base classifier on noisy inputs (Cohen et al., 2019; Salman et al., 2019; Jeong & Shin, 2020; Zhai et al., 2020; Jeong et al., 2021; 2023), and (2) denoised smoothing, where noisy inputs are first denoised before classification (Salman et al., 2020; Carlini et al., 2023). Recent advances in deep learning, particularly diffusion models, have significantly enhanced denoised smoothing approaches, enabling state-of-the-art certified accuracy at small perturbation radii (Carlini et al., 2023; Xiao et al., 2023; Zhang et al., 2023).

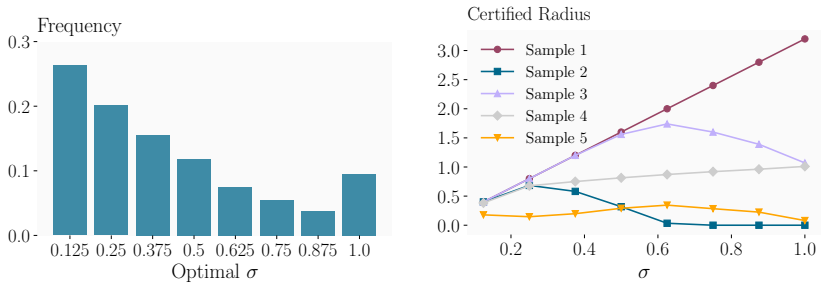


Figure 1: Left: The distribution of the optimal σ on CIFAR-10 test set, where the base model is fixed to the pretrained denoised smoothing model from Carlini et al. (2023). The optimal σ for each input is defined as the σ that maximizes the certified radius under the standard RS certification. Right: The certified radii curve of five independent samples against σ.

Table 1: Comparison of key features of the literature and the proposed Dual RS.

	Literature	Flexible σ	No test-time memorization	Flexible routing
Certified Routing	Mueller et al. (2021)	NA	✓	Restricted
Adaptive RS	Alfarra et al. (2022)	✓	✗	NA
	Wang et al. (2021)	✓	✗	NA
	Súkeník et al. (2022)	Restricted	✓	NA
	Jeong & Shin (2024)	Biased	✓	NA
	This work	✓	✓	✓

Despite recent advances, RS remains limited by a fundamental accuracy-robustness trade-off. Achieving a larger certified radius requires increasing the noise variance, which often reduces certified accuracy at smaller radii. This trade-off arises because prior methods apply a global noise variance shared across all inputs (Cohen et al., 2019). As illustrated in Fig. 1, the noise variance that maximizes the certified radius varies substantially across samples. Recent work has explored input-dependent RS to mitigate this issue, but existing approaches either rely on test-time memorization (Alfarra et al., 2022; Wang et al., 2021), permit only limited adaptivity (Súkeník et al., 2022), or systematically over-estimate the optimal variance (Jeong & Shin, 2024).

Motivated by these limitations, we propose *Dual Randomized Smoothing* (Dual RS), a novel framework that enables RS certification with input-dependent noise variances. Our key insight is that RS certification remains valid, with appropriate confidence adjustments, as long as the noise variance is locally constant within the certified region rather than globally fixed across all inputs.

Main Contributions. Our key contributions are:

- A generalization of RS certification to locally constant noise variances, enabling flexible models to predict an optimal variance for each input. This generalization expands the applicability of RS and supports more favorable accuracy-robustness trade-offs.
- A dual RS framework consisting of a variance estimator and a standard RS classifier. The variance estimator predicts the optimal σ for each input, which is then used by the classifier for RS inference. We develop an iterative training procedure that jointly optimizes both components. An alternative routing perspective is also discussed, where the variance estimator acts as a router that selects an appropriate off-the-shelf expert RS classifier based on the input. Table 1 compares key features of prior works with our proposed method.
- An extensive experimental evaluation of Dual RS, showing that Dual RS achieves strong performance across both small and large radii, outperforming prior input-dependent noise methods at most radii while adding roughly 60% computational overhead at inference, compared to standard RS. Comparing against prior works, relative improvements of 19.2%, 24.2%, and 20.6% are achieved at radii 0.5, 0.75, and 1.0 on CIFAR-10, respectively, and roughly 1.5x performance is delivered on IMAGENET at radii 0.5, 1.0, and 1.5.

2 RELATED WORK

Provable Adversarial Robustness Empirical defenses against adversarial attacks are often unreliable (Athalye et al., 2018; Croce & Hein, 2020), motivating research on *provable adversarial*

robustness. Existing approaches fall into two categories: deterministic and probabilistic. Deterministic methods provide exact guarantees but do not scale well to large models (Gowal et al., 2018; Mirman et al., 2018; Shi et al., 2021; Mueller et al., 2023; De Palma et al., 2024; Mao et al., 2023; 2024; Balauca et al., 2024; Mao et al., 2025). Therefore, *Randomized Smoothing* (RS) (Lécuyer et al., 2019; Cohen et al., 2019) becomes the most widely used probabilistic method due to its scalability. Many works have improved RS by developing better training algorithms (Salman et al., 2019; Jeong & Shin, 2020; Zhai et al., 2020; Jeong et al., 2021; 2023), leveraging pretrained models to construct base classifiers (Salman et al., 2020; Carlini et al., 2023), extending RS to different norms and noise distributions (Yang et al., 2020; Kumar et al., 2020), designing alternative certification procedures (Xia et al., 2024; Cullen et al., 2022; Li et al., 2022), proposing new evaluation metrics (Sun et al., 2025), and exploring ensemble techniques (Horváth et al., 2022; Liu et al., 2021). However, a common limitation of these works is the use of a global noise variance in the smoothing distribution for all inputs, which leads to an inherent accuracy-robustness trade-off.

Input-dependent Randomized Smoothing To mitigate the accuracy-robustness trade-off, recent works have explored adapting the noise variance per input. However, existing methods have notable limitations. Some rely on test-time memorization and are computationally expensive (Wang et al., 2021; Alfarra et al., 2022). Súkeník et al. (2022) provide theoretical guarantees for varying σ with severely limited adaptivity. Jeong & Shin (2024) propose a multi-scale RS framework that cascades models with fixed variances, yet it always selects the largest variance that certifies an input, which often yields suboptimal results (Fig. 1). Finally, Lyu et al. (2024) introduce a two-stage framework for ℓ_∞ norm by splitting a fixed noise budget, but it lacks flexible per-input adaptiveness and fails to generalize to ℓ_2 norms.

3 Background

This section introduces the key concepts of adversarial robustness and randomized smoothing (RS).

Adversarial Robustness. A model f is adversarially robust if it produces consistent outputs under small perturbations. Given input x and label y with $f(x) = y$, f is robust (with regard to ℓ_2 norm) if $f(x') = f(x)$ for all x' in $S(x) = \{x' \mid \|x' - x\|_2 \leq \epsilon\}$, where ϵ defines the perturbation magnitude.

Randomized Smoothing. RS provides certified robustness by constructing a smoothed classifier $g_c(x) = \arg\max_{y \in \mathcal{Y}} \mathbb{P}_{\delta \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})}[f(x + \delta) = y]$, where f is the base classifier. The classifier g_c is certifiably robust within an ℓ_2 ball if the predicted class has probability greater than 0.5 (Cohen et al., 2019). Improving the probability margin enhances the certified radius.

Denoised Smoothing. Denoised smoothing (Salman et al., 2020) applies a denoiser before classification, i.e., $f(x + \delta) = f_{\text{cls}}(\texttt{denoise}(x + \delta))$, where `denoise` removes noise from the perturbed input and f_{cls} performs classification. This approach serves as a powerful paradigm for constructing RS base classifiers. Diffusion models have proven to be highly effective denoisers (Carlini et al., 2023), achieving state-of-the-art performance with off-the-shelf components. Following Carlini et al. (2023); Jeong & Shin (2024), we adopt diffusion-based denoised smoothing to build base classifiers in our framework.

4 Certification with Locally Constant Noise Variance

In this section, we formalize the main theoretical contribution of this work: we prove that RS certification remains valid when the noise variance is input-dependent, as long as it is constant within the certified region. This result provides the theoretical foundation for our dual RS framework.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the input space, $\mathcal{Y}$ the output space, and $f_c : \mathcal{X} \to \mathcal{Y}$ the base classifier. The classifier smoothed with a Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ is defined as $g_c(\boldsymbol{x}, \sigma) := \arg\max_{y \in \mathcal{Y}} \mathbb{P}_{\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})}(f(\boldsymbol{x} + \boldsymbol{\delta}) = y)$. Let p_σ be the probability of the most likely class, i.e., $p_\sigma := \max_{y \in \mathcal{Y}} \mathbb{P}_{\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})}(f(\boldsymbol{x} + \boldsymbol{\delta}) = y)$. Cohen et al. (2019) prove that with a global σ constant in $\mathcal{X}$, the smoothed classifier g_c is certifiably robust within an ℓ_2 ball $\mathbb{B}(\boldsymbol{x}, R(\boldsymbol{x}, \sigma))$ of radius $R(\boldsymbol{x}, \sigma) := \sigma \Phi^{-1}(p_\sigma)$ centered at the input $\boldsymbol{x}$, where Φ is the cumulative distribution function of the standard Gaussian distribution. We replace the global σ with an input-dependent function $\sigma : \mathcal{X} \to \Sigma$, where $\Sigma \subset \mathbb{R}^+$ is the discrete set of allowed values, and denote the smoothed classifier

with input-dependent noise variance as $g_c(\boldsymbol{x}, \sigma(\boldsymbol{x}))$. Building on this setup, we present a certification theorem that refines the result of Cohen et al. (2019) by relaxing the assumption on σ from being globally constant to locally constant.

Theorem 4.1 (Certification with Locally Constant σ)**.** Fix $\boldsymbol{x}_0 \in \mathcal{X}$ and f_c. Assume $\sigma(\boldsymbol{x})$ is constant within the ℓ_2 ball $\mathbb{B}(\boldsymbol{x}_0, R_\sigma)$. Then for all $\boldsymbol{x}$ such that $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \leq \min(R_\sigma, R(\boldsymbol{x}, \sigma(\boldsymbol{x}_0)))$, we have $g_c(\boldsymbol{x}, \sigma(\boldsymbol{x})) = g_c(\boldsymbol{x}_0, \sigma(\boldsymbol{x}_0))$.

The proof follows by carefully adapting the alternative argument of Salman et al. (2019) on the result of Cohen et al. (2019), which leverages Lipschitz continuity, to remove the reliance on the global constancy of σ. The detailed proof is deferred to App. B.1.

Practically, the assumption that σ is constant within a neighborhood of $\boldsymbol{x}_0$ can be satisfied in two ways: (1) by designing $\sigma(\boldsymbol{x})$ to be piecewise constant (Wang et al., 2021; Alfarra et al., 2022), or (2) by certifying that $\sigma(\boldsymbol{x})$ is locally constant using deterministic certification methods (Singh et al., 2019; Wong & Kolter, 2018; Müller et al., 2022; Shi et al., 2024). Approaches in the former category typically rely on test-time memorization, which is undesirable in practice. In contrast, approaches in the latter category, though extensively developed, are usually computationally expensive and less scalable. Therefore, in this work, we seek a certification of $\sigma(\boldsymbol{x})$ that both scales well and eliminates test-time memorization. To this end, we propose to use a separate RS model to learn effective $\sigma(\boldsymbol{x})$ and certify the local constancy. To achieve this, we need to extend Theorem 4.1 to a probabilistic setting, since RS in practice only provides probabilistic guarantees.

Before presenting the theorem, we extend the notion of RS to the practical setting, where p_σ is lower bounded with uncertainty α. Given N trials of the event $I(f(\boldsymbol{x}+\boldsymbol{\delta}) = y)$ and a predefined threshold α, we can derive a lower bound $\hat{p_\sigma}$ such that $\mathbb{P}(p_\sigma \geq \hat{p_\sigma}) \geq 1 - \alpha$ (Cohen et al., 2019). Consequently, the smoothed classifier $g_c(\boldsymbol{x}, \sigma)$ is certifiably robust within the ℓ_2 ball $\mathbb{B}(\boldsymbol{x}, \sigma\Phi^{-1}(\hat{p_\sigma}))$ with probability at least $1 - \alpha$. Now we are ready to present the probabilistic version of Theorem 4.1.

Theorem 4.2 (Probabilistic Guarantee with Confidence Adjustment)**.** Fix $\boldsymbol{x}_0 \in \mathcal{X}$ and f_c. Assume $g_c(\boldsymbol{x}, \sigma(\boldsymbol{x}_0))$ is certifiably robust within $\mathbb{B}(\boldsymbol{x}_0, R_c)$ with probability at least $1 - \alpha$, and $\sigma(\boldsymbol{x})$ is constant within $\mathbb{B}(\boldsymbol{x}_0, R_\sigma)$ with probability at least $1 - \beta$. Then for all $\boldsymbol{x}$ such that $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \leq \min(R_\sigma, R_c)$, we have $g_c(\boldsymbol{x}, \sigma(\boldsymbol{x})) = g_c(\boldsymbol{x}_0, \sigma(\boldsymbol{x}_0))$ with probability at least $1 - \alpha - \beta$.

The proof follows by applying union bound to upper bound the failure probability. The detailed proof is deferred to App. B.2. Note that Theorem 4.2 does not assume independence between the two failure events, and therefore remains valid even when the two failure events are correlated, e.g., correlated noise samples may be used in two certifications.

Comparison with Prior Work. Although not explicitly formalized, the idea of using a locally constant σ has been explored in prior work (Wang et al., 2021; Alfarra et al., 2022). Wang et al. (2021) partition $\mathcal{X}$ into a collection of ℓ_2 balls, referred to as robust regions, and assign a constant σ to each region. These regions are allocated and stored at test time, which prevents parallel inference and leads to dependence on the prior test cases. Similar strategies are adopted by Alfarra et al. (2022). Beyond formalization and rigorous proof, Theorem 4.1 further improves by eliminating the need for test-time memorization and instead ensuring local constancy through certifying $\sigma(\boldsymbol{x})$, which can be any learned model or hand-crafted function.

Separately, Súkeník et al. (2022) also study RS with input-dependent $\sigma(\boldsymbol{x})$ and show that proofs based on Neyman-Pearson lemma cannot allow reasonably flexible $\sigma(\boldsymbol{x})$. We circumvent this limitation by leveraging a proof based on Lipschitz continuity, similar to Salman et al. (2019); Jeong & Shin (2024), which enables much greater flexibility in defining $\sigma(\boldsymbol{x})$. Note that our result does not restrict the behavior of $\sigma(\boldsymbol{x})$ outside the certified region, which can be arbitrarily complex. We provide a comparison table of prior works in Table 1.

Despite these advantanges, Theorem 4.2 introduces a confidence penalty of β to account for the probabilistic guarantee of $\sigma(\boldsymbol{x})$ being locally constant. This cost is inevitable when using any certification method that is not deterministic. However, in practice, we find that this cost is negligible when using RS to certify $\sigma(\boldsymbol{x})$. We list a few numerical examples under different configurations in Table 5 in App. D, confirming that β has minimal impact on the certified radius.

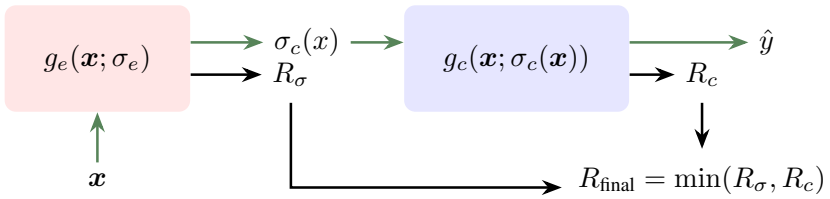


Figure 2: The dual RS framework. First, a RS model g_e smoothed with a global σ_e is deployed to estimate $\sigma_c(\boldsymbol{x})$ and return a certified radius for the estimation, R_σ. Second, another RS model is smoothed with $\sigma_c(\boldsymbol{x})$, and then perform a standard classification and return a certified radius for the classification, R_c. The final prediction is the result of the second stage, with a final certified radius $R_{\text{final}} = \min(R_\sigma, R_c)$. The green arrows indicate activated paths during inference.

5 The Dual Randomized Smoothing Framework

In this section, we present the dual RS framework implementing RS with input-dependent noise variances. We first give an overview of the framework, followed by details of the inference and certification process. Then, we describe the training methods to optimize the performance. Finally, we discuss an alternative routing perspective of the dual RS framework.

5.1 Inference & Certification

Fig. 2 illustrates the dual RS framework. Given an input $\boldsymbol{x}$, a *variance estimator* predicts an appropriate variance, σ_c, followed by a *classifier* smoothed with σ_c to perform the final classification. Intuitively, the variance estimator partitions the input space into disjoint subsets associated with different values of σ_c, and assign the input (ideally also its neighborhood) to the corresponding subset. This formulation exactly matches the definition of robustness, in the task of predicting the optimal σ_c. Therefore, a separate RS model is applied, which uses a pre-defined global noise variance to certify the estimated σ_c. With the estimated σ_c, another base model can be smoothed via RS with σ_c to perform certified classification. The final certified radius is then guaranteed by Theorem 4.2. Note that one needs to ensure $\sigma_e \geq \max_{\boldsymbol{x}} \sigma_c(\boldsymbol{x})$ to not limit the final certified radius inherently.

Unless otherwise noted, we use diffusion denoised smoothing to build both the variance estimator and the RS classifier, for the simplicity and efficiency. Formally, we denote the two RS models as:

$$g_e(\boldsymbol{x}, \sigma_e) := \arg\max_{\sigma_i \in \Sigma} \mathbb{P}_{\boldsymbol{\delta_e} \sim \mathcal{N}(0, \sigma_e^2 I)}(h_e(\texttt{denoise}(\boldsymbol{x} + \boldsymbol{\delta_e})) = \sigma_i),$$
$$g_c(\boldsymbol{x}, \sigma_c) := \arg\max_{\hat{y} \in \mathcal{Y}} \mathbb{P}_{\boldsymbol{\delta_c} \sim \mathcal{N}(0, \sigma_c^2 I)}(h_c(\texttt{denoise}(\boldsymbol{x} + \boldsymbol{\delta_c})) = \hat{y}),$$

where `denoise` represents a single-step denoising using an off-the-shelf diffusion denoiser, and h_e and h_c are base models for variance estimation and classification, respectively.

At inference time, given an input $\boldsymbol{x}$, we sample noise samples $\{\boldsymbol{\delta_e}\}$ from $\mathcal{N}(0, \sigma_e^2 I)$, and use the PREDICT function from Cohen et al. (2019) to predict the noise variance $\sigma_c(\boldsymbol{x})$ with uncertainty $\alpha/2$. Then, we sample noise samples $\{\boldsymbol{\delta_c}\}$ from $\mathcal{N}(0, \sigma_c(\boldsymbol{x})^2 I)$, and use the PREDICT function again to predict the class label $\hat{y}$ with uncertainty $\alpha/2$. The final prediction is $\hat{y}$, with a total uncertainty of α, using the union bound again on the failure events, similar to the proof of Theorem 4.2. To certify the prediction, we use the CERTIFY function from Cohen et al. (2019) to certify the local constancy of $\sigma_c(\boldsymbol{x})$ with uncertainty $\alpha/2$, and certify the classification with uncertainty $\alpha/2$. The final certified radius is $R_{\text{final}} = \min(R_\sigma, R_c)$, where R_σ is the certified radius for the estimation of $\sigma_c(\boldsymbol{x})$, and R_c is the certified radius for the classification. The total uncertainty is α, as guaranteed by Theorem 4.2. We note that for simplicity, we use the same uncertainty level $\alpha/2$ for both certifications, but they can be adjusted flexibly as long as the total uncertainty does not exceed α.

5.2 Training methods

5.2.1 Training the Variance Estimator

Building the Training Dataset. Training h_e requires ground-truth labels for the optimal noise $\sigma_c(\boldsymbol{x})$ of each input. Given a candidate set Σ and a fixed h_c, we evaluate for each input the certified

radius under each $\sigma_i \in \Sigma$. The label for the optimal noise $\sigma_c(\boldsymbol{x})$ is then $\arg\max_i R_c(\boldsymbol{x}, \sigma_i)$. This step is usually the most computationally expensive part of the training, as it requires multiple certifications for each input. However, it only needs to be performed once before training h_e, and can be parallelized across multiple devices. In practice, one may also use a smaller budget N than required during certification to estimate $R_c(\boldsymbol{x}, \sigma_i)$ and train only on a subset of the train data, to reduce the computational cost. In App. E.4, we conduct detailed studies on these two strategies to show that they can significantly reduce the training cost with minimal performance degradation.

Training with Soft Labels. Estimating optimal variance is formulated as a classification task, but it has certain special properties. Even if the estimated σ_c is not optimal, a non-zero certified radius is still likely. For example, assume that given $\Sigma = \{0.25, 0.5, 1.0\}$, the certified radii of x_1 are 0.0, 1.6 and 0.0, respectively, while those of x_2 are 0.3, 0.4 and 0.3, respectively. Choosing the wrong σ for x_1 is more harmful than for x_2 intuitively, as the latter still has a reasonably close certified radius. Motivated by this, we propose to use soft labels introduced below to train the variance estimator. Formally, the soft label for the variance estimation is defined as:

$$y_i = \frac{\exp(R_c(\boldsymbol{x}, \sigma_i))}{\sum_{\sigma_j \in \Sigma} \exp(R_c(\boldsymbol{x}, \sigma_j))}.$$

A standard cross-entropy loss is then applied between the soft labels and the predicted class probabilities to evaluate the estimation performance.

Consistency Regularization. Many strategies have been proposed to increase the certified radius in the standard RS training. We choose one of them, consistency regularization (Jeong & Shin, 2020), to further improve the certified radius of the estimated σ. Formally,

$$\mathcal{L}_{\text{con}}(\boldsymbol{x}) := \lambda \mathbb{E}_\delta \left[\text{KL}(\hat{f}(\boldsymbol{x}) \| f(\boldsymbol{x} + \boldsymbol{\delta})) \right] + \eta \text{H}(\hat{f}(\boldsymbol{x})),$$

where $\hat{f}(x) = \mathbb{E}(f(x + \delta))$, KL is the Kullback-Leibler divergence, H is the entropy, and λ and η are hyperparameters controlling the trade-off between accuracy and robustness. We remark that any other RS training strategies can be alternatively applied; we choose consistency because it is the fastest while being competitive in performance (Jeong et al. (2023), Appendix E).

Overall Objective. The overall loss function to train the variance estimator is a weighted average between the soft-label cross-entropy loss and the consistency loss:

$$\mathcal{L}_\sigma = \mathbb{E}_{\boldsymbol{x}} \left[w_e(\boldsymbol{x}) \left(\mathcal{L}_{\text{softCE}}(\boldsymbol{x}) + w_r(\boldsymbol{x}) \mathcal{L}_{\text{con}}(\boldsymbol{x}) \right) \right],$$

where $w_e(\boldsymbol{x})$ and $w_r(\boldsymbol{x})$ are two weighting functions. We introduce a balancing weight $w_e(\boldsymbol{x})$ because the distribution of optimal σ_c is usually skewed. Formally, assume the fraction of training samples with optimal noise σ_i is q_i, then $w_e(\boldsymbol{x}) = 1/q_i$ if the optimal noise for $\boldsymbol{x}$ is σ_i. $w_r(\boldsymbol{x}) := \max_{\sigma_i} R_c(\boldsymbol{x}, \sigma_i)/C$ puts more consistency regularization for inputs with larger optimal certified radii, rescaled to $[0, 1]$ by a constant C.

5.2.2 Adapting the Classifier to the Variance Estimator

Prior work (Carlini et al., 2023) have shown that finetuning the off-the-shelf classifier with regard to the RS framework can significantly improve the performance. In this section, we follow a similar approach, showing how to adapt the classifier to the dual RS framework.

Given a fixed variance estimator g_e, we finetune the classifier h_c under the estimated noise variances. Formally, given an input $\boldsymbol{x}$, we first query the noise variance $\sigma_c(\boldsymbol{x})$ from g_e. Then, we sample noise $\boldsymbol{\delta_c} \sim \mathcal{N}(0, \sigma_c(\boldsymbol{x})^2 I)$, and apply the denoising step to obtain $\tilde{\boldsymbol{x}} = \texttt{denoise}(\boldsymbol{x} + \boldsymbol{\delta_c})$. Finally, we apply a standard cross-entropy loss between the prediction $h_c(\tilde{\boldsymbol{x}})$ and the ground-truth label y. This procedure follows Carlini et al. (2023) with only one difference: the noise variance is input-dependent, estimated by g_e, instead of being a global constant.

The described training process naturally leads to an alternating training scheme, where we iteratively train the variance estimator and finetune the classifier. In practice, we find that one round of classifier finetuning is usually sufficient to achieve good performance, i.e., training the variance estimator from scratch based on the off-the-shelf classifier, followed by one round of classifier finetuning. More rounds of alternating training may lead to marginal improvements, but at a much higher computational cost (c.f. App. E.2).

5.3 Routing with Expert RS Models

Routing is to select the best model from a pool of expert models for each input. It has been widely studied in the context of mixture-of-experts, especially for large language models (Varangot-Reille et al., 2025). In this section, we present a novel perspective of the proposed dual RS framework as a router among a pool of pretrained expert RS models.

§5.2.1 proposes strategies to train the variance estimator to predict the best σ_c for a fixed base classifier h_c. This naturally requires h_c to perform well under all $\sigma_i \in \Sigma$, each for a subset of inputs. However, as well-known in the RS literature (e.g., Sun et al. (2025)), no single model wins uniformly across all noise levels. Luckily, Theorem 4.2 does not restrict h_c to be the same model under different σ_i. Therefore, we can define $g_c(\boldsymbol{x}, \sigma(\boldsymbol{x}))$ to be the best expert among a pool of models. Formally, let $\mathcal{H} := \{\mathcal{H}_{\sigma_i}\}$ be the pool of the pretrained expert models where $\mathcal{H}_{\sigma_i}$ are expert models performing well under σ_i. Define $\mathcal{X}_{\sigma_i} := \{\boldsymbol{x} \mid g_e(\boldsymbol{x}, \sigma_e) = \sigma_i\}$ to be the subset of inputs assigned to σ_i by the variance estimator. Then we define $g_c(\boldsymbol{x}, \sigma(\boldsymbol{x})) := \mathcal{H}_{\sigma_i}(\boldsymbol{x}, \sigma_i)$ for all $\boldsymbol{x} \in \mathcal{X}_{\sigma_i}$. In other words, the variance estimator g_e serves not only as a predictor for the optimal noise variance, but also as a router to select the best expert RS model for each input. The training process of g_e remains unchanged, except that the certified radius $R_c(\boldsymbol{x}, \sigma_i)$ is now evaluated using the corresponding expert model $\mathcal{H}_{\sigma_i}$. Note that we do not evaluate the performance of the expert models except with the corresponding variance, i.e., $\mathcal{H}_{\sigma_i}$ is not evaluated with σ_j for $j \neq i$.

The proposed routing perspective of dual RS has several advantages. First, it allows leveraging existing expert models without the need for training a new base classifier that performs well under all noise levels. This is particularly useful when the training cost is prohibitively high. Second, it enables the use of specialized models that excel in specific noise regimes, potentially improving overall performance. Third, it provides a flexible framework that can easily incorporate new expert models, with the minimal effort of re-training the variance estimator. This is because certification under dual RS has much smaller overhead given the certified radii of the expert models since the variance estimator is usually lightweight. Fourth, assuming the expert models are trained independently, improving expert models usually leads to a strict improvement in the overall performance, as we will demonstrate in §6. However, due to the routing nature, the performance of dual RS is upper bounded by the performance of the expert model $\mathcal{H}_{\sigma_i}$ within each $\mathcal{X}_{\sigma_i}$.

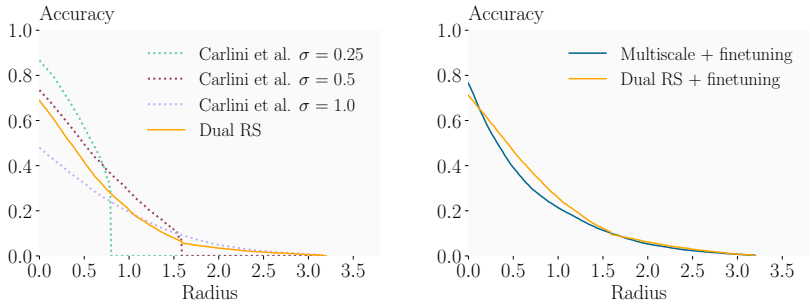
As a final remark, the routing perspective of RS is not limited to the dual RS framework, and can be extended to deterministic certification methods as well. Given a pool of expert models (potentially trained with different algorithms and hyperparameters), offering different trade-offs between accuracy and robustness, one can train a standard RS model to route each input to the best expert model, then certify the routing choice by RS. The final certified radius is the minimum between the certified radius of the routing RS model and that of the selected expert model. This generalization opens up new possibilities for combining the strengths of various certification methods within a unified framework. We leave the exploration of this direction to future work.

Comparison with Prior Work. Mueller et al. (2021) presents a similar idea which routes among a standard network and a robust network using a deterministically certified router. We generalize their idea in the following ways: (i) they only considers routing between two models due to the design of their router, while our framework allows routing among multiple models natively; (ii) they uses a deterministic certification method to certify a heuristically trained router, while our framework uses RS to train and certify the router, which is more scalable and flexible; (iii) they focus on improving the accuracy-robustness trade-off under the given radii, while our objective is to optimize the overall performance across all radii.

6 Experimental Evaluation

In this section, we extensively evaluate the proposed dual RS method on CIFAR-10 and ImageNet. The results demonstrate that dual RS can achieve strong performance across different radii, which is unattainable with a global noise variance. Further, it incurs only a modest computational overhead compared to standard RS. We include all implementation details in App. C and only highlight key experimental settings here.

Baselines. We compare our method against two baselines: (i) diffusion denoised smoothing with global noise variances (Carlini et al., 2023), which we use as the base classifiers, and (ii) the state-of-



(a) Comparison with Standard RS (b) Comparison with Input-dependent RS

Figure 3: Certified accuracy on CIFAR-10 across radii.

Table 2: Certified accuracy on CIFAR-10 across different certification radii. **Bold** entries indicate whenever Dual RS outperforms Multiscale.

Method	σ	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
Carlini et al.	0.25	86.61	73.90	57.02	35.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.5	73.49	62.23	49.46	38.20	28.58	19.54	11.22	0.00	0.00	0.00	0.00
	1.0	47.98	39.85	32.17	25.16	19.34	14.49	10.30	7.32	4.89	3.35	2.15
Multiscale + finetuning	{0.25, 0.5, 1.0}	76.51	54.78	39.15	28.46	21.33	15.95	11.40	7.91	5.31	3.63	2.34
Dual RS	{0.25, 0.5, 1.0}	68.91	**55.45**	**41.62**	**29.48**	20.26	13.29	7.95	4.70	3.45	2.40	1.70
Dual RS + finetuning		71.12	**58.93**	**46.55**	**35.42**	**25.75**	**18.35**	**12.22**	**8.20**	**6.18**	**4.35**	**2.96**

the-art input-dependent RS method (Jeong & Shin, 2024), denoted as *Multiscale*. Unless otherwise stated, all results are reported with $N = 10{,}000$ noise samples for certification with the overall uncertainty level $\alpha = 0.001$.

CIFAR-10 Setup. Unless otherwise stated, Σ is set to $\{0.25, 0.5, 1.0\}$. Following the baselines, we employ a 50M-parameter diffusion model (Nichol & Dhariwal, 2021) as the denoiser, and a 87M-parameter ViT model (Dosovitskiy et al., 2020) as the classifier. A ResNet-110 (He et al., 2016) is used as the base model for the variance estimator, and $N = 10{,}000$ is used to estimate $R_c(\boldsymbol{x}; \sigma_i)$ during training.

IMAGENET Setup. Unless otherwise stated, Σ is set to $\{0.5, 1.0\}$. Following Carlini et al., we utilize a 552M-parameter class-unconditional diffusion model (Dhariwal & Nichol, 2021) as the denoiser and a 305M-parameter BEiT model (Bao et al., 2021) as the classifier. A ResNet-50 is used as the variance estimator, and $N = 100$ is used to estimate $R_c(\boldsymbol{x}; \sigma_i)$ during training. We remark that here the Multiscale results are cited from Jeong & Shin (2024) due to a repetitive exception thrown by their public code, thus are based on a different classifier: a ViT-B/16 pre-trained via CLIP (Radford et al., 2021) and finetuned on ImageNet using FT-CLIP (Dong et al., 2022). Since performance of classifiers on IMAGENET are weaker than on CIFAR-10, we adopt two additional modifications in training the variance estimator on IMAGENET: (1) *Failure case filtering*: samples that cannot be certified by any candidate noise variance are removed from training data, and (2) *Weaker consistency loss*: instead of setting $w_r(\boldsymbol{x}) := \max_{\sigma_i} R_c(\boldsymbol{x}, \sigma_i)/C$, we set $w_r(\boldsymbol{x}) := R_c(\boldsymbol{x}, \hat{\sigma})/C$, where $\hat{\sigma}$ is the minimum variance predicted by the variance estimator among all noisy samples. The former strategy prevents the variance estimator from being trained on extremely hard samples, while the latter avoids over-regularizing the variance estimator towards the global optimal variance, which is less frequently predicted on IMAGENET.

6.1 KEY RESULTS

Dual RS with Single Pretrained Classifier. We first evaluate the performance of dual RS with a pretrained global classifier, as described in §5.2.1. Fig. 3a and Table 2 compare the pretrained diffusion denoised smoothing model with different global noise variances and dual RS using the same classifier on CIFAR-10. While the baseline models with a small global noise variance (e.g., $\sigma = 0.25$ or 0.5) achieve high certified accuracy at small radii, they fail to provide non-trivial guarantees at larger radii. Conversely, the model with a large global noise variance ($\sigma = 1.0$) attains a large certified radius but suffers from low accuracy at small radii. In contrast, dual RS has a strong

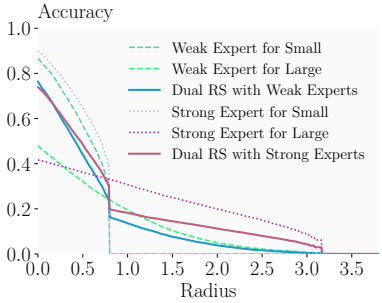


Figure 4: Comparison between dual RS built on weak and strong experts, respectively, along with the experts.

Table 3: Certified accuracy on IMAGENET, structured in the same way as Table 2.

Method	σ	0.0	0.5	1.0	1.5	2.0
Carlini at al.	0.25	80.4	70.6	0.0	0.0	0.0
	0.5	74.4	64.4	52.4	34.8	0.0
	1.0	56.0	47.8	37.4	29.4	24.0
Multiscale†	{0.25, 0.5, 1.0}	69.0	42.4	26.6	19.0	14.6
Multiscale + finetuning†		72.5	44.0	27.5	19.9	14.1
Dual RS	{0.5, 1.0}	67.8	54.6	40.4	28.0	15.6
Dual RS + finetuning		**74.0**	**60.6**	**48.0**	**33.6**	**17.0**

† Results cited from Jeong & Shin (2024), which use a different classifier.

performance across all radii, achieving a superior accuracy-robustness trade-off. This demonstrates that dual RS can effectively leverage the pretrained classifier.

Dual RS with Single Classifier Finetuning. Fig. 3b and Table 2 compare dual RS with Multiscale (Jeong & Shin, 2024), the state-of-the-art input-dependent RS method. For a fair comparison, we finetune the classifier in dual RS as described in §5.2.2, while Multiscale adopts the finetuned diffusion denoiser described in Jeong & Shin (2024). The result shows that dual RS consistently outperforms Multiscale across most radii, with especially strong improvements in the small-radius region. At radii 0.5, 0.75, and 1.0, dual RS improves certified accuracy by 19.2%, 24.2%, and 20.6%, respectively. On a single NVIDIA RTX 4090 GPU with batch size 1000 and $N = 10{,}000$, certifying with dual RS requires 22.58 seconds per input on average, compared to 14.07 seconds for standard RS and 20.21 seconds for Multiscale. Thus, dual RS incurs only a modest computational overhead relative to standard RS, while achieving significant performance gains. We remark that Multiscale requires multiple rounds of certification for some inputs, leading to a higher worst-case certification time (14.07, 28.14, and 42.21 seconds on average for 1, 2, and 3 rounds, respectively).

Dual RS with Multiple Pretrained Experts (Routing). We further evaluate the efficacy of dual RS as a routing mechanism with multiple pretrained expert classifiers, as discussed in §5.3. Specifically, we consider two experts: one specialized for $\sigma = 0.25$ and another specialized for $\sigma = 1.0$. For $\sigma = 0.25$, we define the *Weak Expert for Small* to be an off-the-shelf denoised smoothing model, and the *Strong Expert for Small* to be a finetuned denoised smoothing model by Carlini et al. (2023) on $\sigma = 0.25$. For $\sigma = 1.0$, we define the *Weak Expert for Large* to be the same off-the-shelf model, and the *Strong Expert for Large* to be another off-the-shelf model, pretrained by Sun et al. (2025), which achieves the state-of-the-art performance for large radii. Fig. 4 compares these four experts and dual RS built upon weak and strong experts, respectively. The results show that dual RS effectively leverages the improved performance of the strong experts, achieving a better accuracy-robustness trade-off than that of the weak experts. This demonstrates that dual RS can flexibly incorporate different expert models to further enhance performance.

Dual RS on Large Datasets. We further evaluate dual RS on IMAGENET. As shown in Table 3, Dual RS achieves strong certified accuracy across all radii, consistent to the improvements shown on CIFAR-10. Compared to Multiscale, the state-of-the-art input-dependent RS, dual RS gets roughly 1.5x performance advantage at radii 0.5, 1.0 and 1.5. Overall, these results show that Dual RS scales effectively to large datasets and high-dimensional input spaces.

We further conduct ablation studies on three aspects: (1) different choices of σ candidate sets, (2) strategies for constructing the train set of the variance estimator, and (3) different architectures of the variance estimator, detailed in App. E.3, App. E.4, and App. E.5, respectively. The key

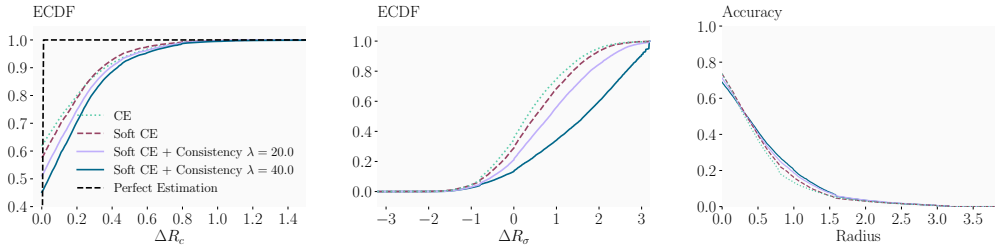


Figure 5: Comparison of dual RS models with different variance estimators.

findings are: (1) the candidate set Σ strongly affects the favored radii, similar to the observation in the global variance methods; (2) the variance estimator can be trained with minimal performance degradation on a much smaller N (up to 99% cost reduction) or a much smaller train set (up to 80% cost reduction); and (3) the performance of dual RS is robust to the architecture of the variance estimator.

6.2 Delving into Dual RS

In this section, we conduct an in-depth study on dual RS on CIFAR-10. We use diffusion denoised smoothing as the classifier with off-the-shelf models and $\Sigma = \{0.25, 0.5, 1.0\}$.

To evaluate the performance of the variance estimation, we define $\Delta R_c := R_c^*(x) - R_c(x)$, where $R_c^*(x)$ is the maximum classification certified radius among all candidate noise variances for input x. This metric reflects how much R_c is reduced due to the suboptimal variance estimation. Fig. 5a shows the empirical cumulative distribution function (ECDF) of this metric for variance estimators trained with different loss functions and hyperparameters on CIFAR-10. The intercept at $\Delta R_c = 0$ indicates the proportion of samples for which the variance estimator predicts the optimal noise variance, and the area between the curve and the the perfect estimation (black dash line) reflects the overall loss in the certified radius due to suboptimal variance estimation. We observe that using soft cross-entropy (CE) loss instead of standard CE loss reduces the variance estimation accuracy, as it encourages the model to predict a suboptimal noise variance that yields a similar certified radius rather than the optimal one. However, fewer inputs are constrained significantly when using soft CE loss, as the curve is closer to the perfect estimation line when ΔR_c is large. Further, adding the consistency loss reduces the variance estimation accuracy in general, since it puts additional regularization on the robustness of the variance estimator.

Since the final certified radius is the minimum between the classification certified radius and the variance certified radius, the alignment between these two radii is of interest as well. We define ΔR_σ as $\Delta R_\sigma = R_\sigma - R_c$. A negative ΔR_σ means that the final radius is constrained by the R_σ, while a positive ΔR_σ means it is constrained by R_c. Ideally, we want ΔR_σ to be positive for as many samples as possible, so that the final certified radius is not constrained by R_σ. Fig. 5b shows the ECDF of ΔR_σ. The intercept at $\Delta R_\sigma = 0$ indicates the proportion of samples constrained by R_σ. We observe that using soft CE loss decreases this ratio, and adding consistency loss further decreases it significantly. This aligns with our intuition in the design.

As a reference, Fig. 5c shows the accuracy-radius curves for these models. Using soft CE loss almost improves over the standard CE loss uniformly, while adding consistency loss slightly degrades the performance at small radii but improves it at large radii. Overall, the model trained with soft CE loss and consistency loss achieves the best accuracy-robustness trade-off.

7 Conclusion

In this work, we address the fundamental trade-off between certified accuracy and certified radius in Randomized Smoothing (RS). We prove that RS remains valid under input-dependent noise variances, provided the variance is locally constant within the certified region. Building on this result, we introduce a dual RS framework, which achieves strong performance across both small and large radii, unattainable with fixed noise variance, while incurring modest computational overhead. Our method consistently outperforms prior input-dependent noise approaches across most radii. Further, the dual RS framework offers a novel routing perspective for certified robustness, enhancing the accuracy-robustness trade-off using off-the-shelf expert RS models.

8 Reproducibility Statement

We have made extensive efforts to ensure the reproducibility of our work. For theoretical results, we provide precise definitions and formal statements in §4, with complete proofs given in App. B. For the proposed framework, detailed descriptions of the inference, certification, and training procedures are presented in §5. Experimental settings, including architectures, hyperparameters, and training details, are reported in §6 and App. C. To further support reproducibility, we include our code in the supplementary material, along with links to download the experiment checkpoints and data used in this paper.

References

Motasem Alfarra, Adel Bibi, Philip HS Torr, and Bernard Ghanem. Data dependent randomized smoothing. In *Uncertainty in Artificial Intelligence*, pp. 64–74. PMLR, 2022.

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

Stefan Balauca, Mark Niklas Müller, Yuhao Mao, Maximilian Baader, Marc Fischer, and Martin T Vechev. Overcoming the paradox of certified training with gaussian smoothing. *CoRR*, 2024.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J Zico Kolter. (certified!!) adversarial robustness for free! In *The Eleventh International Conference on Learning Representations*, 2023.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proc. of ICML*, volume 97, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 2020.

Andrew Cullen, Paul Montague, Shijie Liu, Sarah Erfani, and Benjamin Rubinstein. Double bubble, toil and trouble: enhancing certified robustness through transitivity. *Advances in Neural Information Processing Systems*, 35:19099–19112, 2022.

Alessandro De Palma, Rudy Bunel, Krishnamurthy Dvijotham, M Pawan Kumar, Robert Stanforth, and Alessio Lomuscio. Expressive losses for verified robustness via convex combinations. In *ICLR 2024-International Conference on Learning Representations*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Shuyang Gu, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet. *arXiv preprint arXiv:2212.06138*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

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.

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

Miklós Z Horváth, Mark Niklas Mueller, Marc Fischer, and Martin Vechev. Boosting randomized smoothing with variance reduced classifiers. In *International Conference on Learning Representations*, 2022.

Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers. In *NeurIPS*, 2020.

Jongheon Jeong and Jinwoo Shin. Multi-scale diffusion denoised smoothing. *Advances in Neural Information Processing Systems*, 36, 2024.

Jongheon Jeong, Sejun Park, Minkyu Kim, Heung-Chang Lee, Do-Guk Kim, and Jinwoo Shin. Smoothmix: Training confidence-calibrated smoothed classifiers for certified robustness. In *NeurIPS*, pp. 30153–30168, 2021.

Jongheon Jeong, Seojin Kim, and Jinwoo Shin. Confidence-aware training of smoothed classifiers for certified robustness. In *AAAI*, pp. 8005–8013. AAAI Press, 2023.

Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. In *International Conference on Machine Learning*, pp. 5458–5467. PMLR, 2020.

Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, 2019. doi: 10.1109/SP.2019.00044.

Linyi Li, Jiawei Zhang, Tao Xie, and Bo Li. Double sampling randomized smoothing. In *International Conference on Machine Learning*, pp. 13163–13208. PMLR, 2022.

Chizhou Liu, Yunzhen Feng, Ranran Wang, and Bin Dong. Enhancing certified robustness via smoothed weighted ensembling. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

Saiyue Lyu, Shadab Shaikh, Frederick Shpilevskiy, Evan Shelhamer, and Mathias Lécuyer. Adaptive randomized smoothing: Certifying multi-step defences against adversarial examples. *arXiv e-prints*, pp. arXiv–2406, 2024.

Yuhao Mao, Mark Müller, Marc Fischer, and Martin Vechev. Connecting certified and adversarial training. *Advances in Neural Information Processing Systems*, 36:73422–73440, 2023.

Yuhao Mao, Mark Niklas Mueller, Marc Fischer, and Martin Vechev. Understanding certified training with interval bound propagation. In *The Twelfth International Conference on Learning Representations*, 2024.

Yuhao Mao, Stefan Balauca, and Martin Vechev. Ctbench: A library and benchmark for certified training. In *Forty-second International Conference on Machine Learning*, 2025.

Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, pp. 3578–3586. PMLR, 2018.

Mark Niklas Mueller, Mislav Balunović, and Martin Vechev. Boosting certified robustness of deep networks via a compositional architecture. In *International Conference on Learning Representations*, 2021.

Mark Niklas Mueller, Franziska Eckert, Marc Fischer, and Martin Vechev. Certified training: Small boxes are all you need. In *The Eleventh International Conference on Learning Representations*, 2023.

Mark Niklas Müller, Gleb Makarchuk, Gagandeep Singh, Markus Püschel, and Martin T. Vechev. PRIMA: general and precise neural network certification via scalable convex hull approximations. *Proc. ACM Program. Lang.*, 6(POPL), 2022. doi: 10.1145/3498704.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

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 P. Razenshteyn, Pengchuan Zhang, Huan Zhang, Sébastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Proc. of NeurIPS*, 2019.

Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J. Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. In *NeurIPS*, 2020.

Zhouxing Shi, Yihan Wang, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Fast certified robust training with short warmup. *Advances in Neural Information Processing Systems*, 34:18335–18349, 2021.

Zhouxing Shi, Qirui Jin, Zico Kolter, Suman Jana, Cho-Jui Hsieh, and Huan Zhang. Neural network verification with branch-and-bound for general nonlinearities. *CoRR*, abs/2405.21063, 2024. doi: 10.48550/ARXIV.2405.21063. URL https://doi.org/10.48550/arXiv.2405.21063.

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin T. Vechev. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3(POPL), 2019. doi: 10.1145/3290354.

Peter Súkeník, Aleksei Kuvshinov, and Stephan Günnemann. Intriguing properties of input-dependent randomized smoothing. In *International Conference on Machine Learning*. PMLR, 2022.

Chenhao Sun, Yuhao Mao, Mark Niklas Müller, and Martin T. Vechev. Average certified radius is a poor metric for randomized smoothing. In *The Forty-Second International Conference on Machine Learning*, 2025.

Clovis Varangot-Reille, Christophe Bouvard, Antoine Gourru, Mathieu Ciancone, Marion Schaeffer, and François Jacquenet. Doing more with less: A survey on routing strategies for resource optimisation in large language model-based systems. 2025.

Lei Wang, Runtian Zhai, Di He, Liwei Wang, and Li Jian. Pretrain-to-finetune adversarial training via sample-wise randomized smoothing. 2021.

Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proc. of ICML*, volume 80, 2018.

Song Xia, Yi Yu, Xudong Jiang, and Henghui Ding. Mitigating the curse of dimensionality for certified robustness via dual randomized smoothing. In *ICLR*, 2024.

Chaowei Xiao, Zhongzhu Chen, Kun Jin, Jiongxiao Wang, Weili Nie, Mingyan Liu, Anima Anandkumar, Bo Li, and Dawn Song. Densepure: Understanding diffusion models for adversarial robustness. In *The Eleventh International Conference on Learning Representations*, 2023.

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.

Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. MACER: attack-free and scalable robust training via maximizing certified radius. In *ICLR*. OpenReview.net, 2020.

Jiawei Zhang, Zhongzhu Chen, Huan Zhang, Chaowei Xiao, and Bo Li. {DiffSmooth}: Certifiably robust learning via diffusion models and local smoothing. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 4787–4804, 2023.

Table 4: GPU hours on a single NVIDIA RTX 4090 for the main components of our training pipeline. The costs of building the optimal-variance dataset, training variances estimator and finetuning the classifier are reported as (number of parallel GPUs $\times$ wall-clock time in hours).

Component	CIFAR-10	IMAGENET
Building optimal variances dataset	48×14.6	128×42.2
Training variance estimator	1×2.5	8×63
Generating estimated variances	1.5	9.9
Finetuning the classifier	1×1.0	8×0.7

A USAGE OF LARGE LANGUAGE MODELS

We used a large language model (GPT-5) solely to assist with polishing and grammar correction of the paper. The LLM was not involved in other aspects of this paper.

B DEFERRED PROOFS

B.1 PROOF OF THEOREM 4.1

We first cite the following lemma from Salman et al. (2019), using the formulation as in Lemma D.1 of Jeong & Shin (2024). Note that we adapt the notation to be consistent with our paper. We slightly abuse the notation and let p_σ be the probability of a certain class, i.e., $p_\sigma(\boldsymbol{x}) := \mathbb{P}_{\boldsymbol{\delta}\sim\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I})}(f(\boldsymbol{x}+\boldsymbol{\delta}) = y)$ for some y.

Lemma B.1. $h_y(\boldsymbol{x}) := \sigma\Phi^{-1}(p_\sigma(\boldsymbol{x}))$ is 1-Lipschitz with respect to the ℓ_2 norm.

Lemma B.1 can be extended to locally constant $\sigma(\boldsymbol{x})$ as follows.

Lemma B.2. Let $\mathcal{X}$ be partitioned into non-overlapping subsets $\bigcup_{i\in I}\mathcal{X}_i \subseteq \mathcal{X}$, and assume $\sigma(\boldsymbol{x})$ is constant within each $\mathcal{X}_i$. Let $h_y(\boldsymbol{x}) := \sigma(\boldsymbol{x})\Phi^{-1}(p_{\sigma(\boldsymbol{x})}(\boldsymbol{x}))$. Then for all $i \in I, \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}_i$, we have $|h_y(\boldsymbol{x}_1) - h_y(\boldsymbol{x}_2)| \le \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$.

Proof. For any $i \in I$, let σ_i be the constant value of $\sigma(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathcal{X}_i$. Then for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}_i$, we have

$$\begin{aligned} |h_y(\boldsymbol{x}_1) - h_y(\boldsymbol{x}_2)| &= |\sigma_i\Phi^{-1}(p_{\sigma_i}(\boldsymbol{x}_1)) - \sigma_i\Phi^{-1}(p_{\sigma_i}(\boldsymbol{x}_2))| \\ &\le \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2, \end{aligned}$$

where the last inequality follows from Lemma B.1. □

Now we are ready to prove Theorem 4.1, restated below for convenience.

Theorem 4.1 (Certification with Locally Constant σ)**.** Fix $\boldsymbol{x}_0 \in \mathcal{X}$ and f_c. Assume $\sigma(\boldsymbol{x})$ is constant within the ℓ_2 ball $\mathbb{B}(\boldsymbol{x}_0, R_\sigma)$. Then for all $\boldsymbol{x}$ such that $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \le \min(R_\sigma, R(\boldsymbol{x}, \sigma(\boldsymbol{x}_0)))$, we have $g_c(\boldsymbol{x}, \sigma(\boldsymbol{x})) = g_c(\boldsymbol{x}_0, \sigma(\boldsymbol{x}_0))$.

Proof. Let $\mathcal{X}_i = \{\boldsymbol{x} | \sigma(\boldsymbol{x}) = \sigma_i\}$, where σ_i are distinct values taken by $\sigma(\boldsymbol{x})$. Then $\mathcal{X} = \bigcup_{i\in I}\mathcal{X}_i$ is a partition of $\mathcal{X}$ into non-overlapping subsets. By Lemma B.2, for any $i \in I, \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}_i$, we have $|h_y(\boldsymbol{x}_1) - h_y(\boldsymbol{x}_2)| \le \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$. Further, given $\boldsymbol{x}$, there exists exactly one $j \in I$ such that $\boldsymbol{x} \in \mathcal{X}_j$. This implies $\mathbb{B}(\boldsymbol{x}_0, R_\sigma) \subseteq \mathcal{X}_j$. If there is no adversarial perturbation $\boldsymbol{\delta}$ such that $\|\boldsymbol{\delta}\|_2 \le R_\sigma$ and $g_c(\boldsymbol{x}_0 + \boldsymbol{\delta}) \ne g_c(\boldsymbol{x}_0)$, then the claim holds trivially. In the following, we consider the case where such adversarial perturbation $\boldsymbol{\delta}$ exists.

Given the smoothing distribution $\mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$ where $\sigma = \sigma_j$, let A be the most likely class at $\boldsymbol{x}_0$, and B be any other class. Let $p_A(\boldsymbol{x})$ be the probability of class A at $\boldsymbol{x}$ under the smoothing distribution, and $p_B(\boldsymbol{x})$ be the probability of class B at $\boldsymbol{x}$. Therefore, $\forall \boldsymbol{x}_0 + \boldsymbol{\delta} \in \mathcal{X}_i$, we have $\sigma\Phi^{-1}(p_A(\boldsymbol{x}_0)) - \sigma\Phi^{-1}(p_A(\boldsymbol{x}_0 + \boldsymbol{\delta})) = h_A(\boldsymbol{x}_0) - h_A(\boldsymbol{x}_0 + \delta) \le \|\delta\|_2$. Let $\boldsymbol{\delta}$ be an adversarial perturbation such that

$\|\boldsymbol{\delta}\|_2 \leq R_\sigma$ and let B be the most likely class at $\boldsymbol{x}_0 + \boldsymbol{\delta}$. Then, since $p_A(\boldsymbol{x}_0 + \boldsymbol{\delta}) \leq p_B(\boldsymbol{x}_0 + \boldsymbol{\delta})$ and $\Phi^{-1}(t)$ is monotonically increasing in t, we have

$$\begin{aligned}\sigma\Phi^{-1}(p_A(\boldsymbol{x}_0)) - \sigma\Phi^{-1}(p_B(\boldsymbol{x}_0 + \boldsymbol{\delta})) &\leq \sigma\Phi^{-1}(p_A(\boldsymbol{x}_0)) - \sigma\Phi^{-1}(p_A(\boldsymbol{x}_0 + \boldsymbol{\delta})) \\ &\leq \|\boldsymbol{\delta}\|_2.\end{aligned}$$

Further, applying Lemma B.2 again gives $\sigma\Phi^{-1}(p_B(\boldsymbol{x}_0 + \boldsymbol{\delta})) - \sigma\Phi^{-1}(p_B(\boldsymbol{x}_0)) = h_B(\boldsymbol{x}_0 + \boldsymbol{\delta}) - h_B(\boldsymbol{x}_0) \leq \|\boldsymbol{\delta}\|_2$. Combining the two inequalities gives

$$\sigma\Phi^{-1}(p_A(\boldsymbol{x}_0)) - \sigma\Phi^{-1}(p_B(\boldsymbol{x}_0)) \leq 2\|\boldsymbol{\delta}\|_2.$$

Thus, we have

$$\begin{aligned}\|\boldsymbol{\delta}\|_2 &\geq \frac{\sigma}{2}\left(\Phi^{-1}(p_A(\boldsymbol{x}_0)) - \Phi^{-1}(p_B(\boldsymbol{x}_0))\right) \\ &\geq \frac{\sigma}{2}\left(\Phi^{-1}(p_A(\boldsymbol{x}_0)) - \Phi^{-1}(1 - p_A(\boldsymbol{x}_0))\right) \\ &= \sigma\Phi^{-1}(p_A(\boldsymbol{x}_0)) \\ &= R(\boldsymbol{x}, \sigma_j) \\ &= R(\boldsymbol{x}, \sigma(\boldsymbol{x}_0)).\end{aligned}$$

This completes the proof. □

B.2 Proof of Theorem 4.2

We restate Theorem 4.2 below for convenience.

Theorem 4.2 (Probabilistic Guarantee with Confidence Adjustment)**.** Fix $\boldsymbol{x}_0 \in \mathcal{X}$ and f_c. Assume $g_c(\boldsymbol{x}, \sigma(\boldsymbol{x}_0))$ is certifiably robust within $\mathbb{B}(\boldsymbol{x}_0, R_c)$ with probability at least $1 - \alpha$, and $\sigma(\boldsymbol{x})$ is constant within $\mathbb{B}(\boldsymbol{x}_0, R_\sigma)$ with probability at least $1 - \beta$. Then for all $\boldsymbol{x}$ such that $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \leq \min(R_\sigma, R_c)$, we have $g_c(\boldsymbol{x}, \sigma(\boldsymbol{x})) = g_c(\boldsymbol{x}_0, \sigma(\boldsymbol{x}_0))$ with probability at least $1 - \alpha - \beta$.

Proof. Let F_1 be the event that $g_c(\boldsymbol{x}_0 + \boldsymbol{\delta}) \neq g_c(\boldsymbol{x}_0)$ for some $\boldsymbol{\delta}$ such that $\|\boldsymbol{\delta}\|_2 \leq R_c$. Let F_2 be the event that $\sigma(\boldsymbol{x}_0 + \boldsymbol{\delta}) \neq \sigma(\boldsymbol{x}_0)$ for some $\boldsymbol{\delta}$ such that $\|\boldsymbol{\delta}\|_2 \leq R_c$. Then we have $\mathbb{P}(F_1) \leq \alpha$ and $\mathbb{P}(F_2) \leq \beta$ by the assumption. Let $F = F_1 \cup F_2$. Then we have $\mathbb{P}(F) \leq \mathbb{P}(F_1) + \mathbb{P}(F_2) \leq \alpha + \beta$, where the first inequality follows from the union bound. Applying Theorem 4.1, the complement of F implies that $g_c(\boldsymbol{x}_0 + \boldsymbol{\delta}) = g_c(\boldsymbol{x}_0)$ for all $\boldsymbol{\delta}$ such that $\|\boldsymbol{\delta}\|_2 \leq \min(R_c, R_\sigma)$. The result follows. □

C Experiment Details

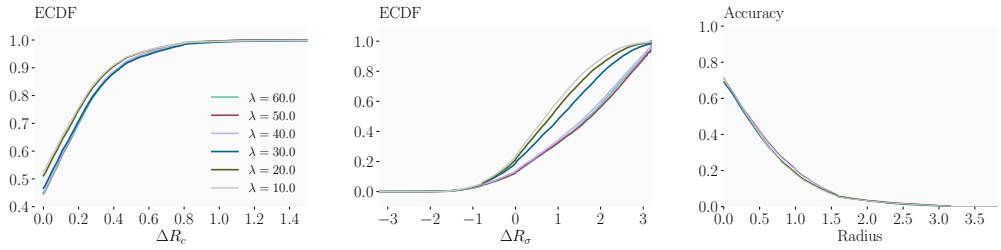
C.1 Experiment Setup

CIFAR-10 In the main experiments, we use $N = 10^4$ to calculate the certified radius for each sample and σ candidate to construct the optimal variances dataset. The variance estimator model is trained from scratch for 90 epochs with a batch size of 256. We use the AdamW optimizer with an initial learning rate of 0.01 and a weight decay of 0.01. The learning rate is decayed by a factor of 0.5 every 30 epochs. Unless otherwise stated, we set $\lambda = 40$ and $\eta = 0.5$ for the consistency loss, and use $\sigma_e = 1.0$ for variance estimation certification. To compute the consistency loss, we always use two noise samples ($m = 2$) following Jeong & Shin (2020). For classifier finetuning, we apply the Cross-Entropy loss on denoised images $\texttt{denoise}(\boldsymbol{x} + \boldsymbol{\delta})$. The classifier is finetuned for 15 epochs with a batch size of 128 using AdamW with a learning rate of 2×10^{-5} and a weight decay of 0.01.

ImageNet We use $N = 100$ to calculate the approximate certified radii to construct the optimal variances dataset. The detailed approach is described in App. E.4. The variance estimator model is trained from scratch for 9 epochs with a batch size of 200. We use the AdamW optimizer with an initial learning rate of 0.005 and a weight decay of 0.01. The learning rate is decayed by a factor of 0.5 every 3 epochs. Unless otherwise stated, we set $\lambda = 10$ and $\eta = 0.5$ for the consistency loss, and use $\sigma_e = 1.0$ for variance estimation certification. For the classifier finetuning, we randomly choose 2% of the training set to finetune the classifier for 1 epoch with a batch size of 32 using AdamW with a learning rate of 2×10^{-5} and a weight decay of 0.01. On ImageNet, after finetuning the classifier, we do not retrian the variance estimator due to the high computational cost.

Table 5: Numerical examples of the confidence penalty β under the given uncertainty level. We assume large enough R_σ, such that the final certified radius equals R_c. When $\alpha : \beta = 1 : 0$, it matches the standard RS setting. The budget is fixed to $N = 10^5$, σ is fixed to 1.0 and $\alpha + \beta$ is fixed to 0.001, following the standard certification setting.

$\alpha : \beta$	$\hat{p_\sigma}$	certified radius
1 : 0	0.99	2.2900
1 : 1	0.99	2.2877
1 : 4	0.99	2.2848
1 : 0	0.8	0.8277
1 : 1	0.8	0.8267
1 : 4	0.8	0.8256
1 : 0	0.6	0.2409
1 : 1	0.6	0.2401
1 : 4	0.6	0.2391



(a) (b) (c)

Figure 6: Abaltion study on λ in the consistency loss. The figures are organized in the same way as Fig. 5.

C.2 Training Cost

We report the computational cost of training our dual RS framework. All experiments are conducted on NVIDIA RTX 4090 GPUs.Table 4 summarizes the GPU hours required for each major component of the training pipeline. The dominant cost arises from constructing the optimal variances dataset, which involves performing classification certification on the full dataset under all three noise variances. For the CIFAR-10 main experiments, two variance estimators are trained and one classifier finetuning is performed, resulting in approximately 1517 GPU hours on a single RTX 4090.

In practice, on CIFAR-10 we parallelized the dataset construction step across 48 GPUs, reducing its wall-clock time to 14.6 hours. Consequently, the overall training pipeline requires approximately 36.7 hours. On ImageNet, we parallelized the dataset construction step across 128 GPUs. The overall training pipeline requires approximately 115.8 hours.

D Numerical Examples for the Confidence Penalty

We list numerical examples of the confidence penalty β under different uncertainty levels in Table 5.

E Additional Studies

In this section, we present additional studies to further investigate different components of our dual RS framework on CIFAR-10.

E.1 Ablation on Consistency Loss Hyperparameter λ

We employ the off-the-shelf diffusion denoiser and classifier in this study. Fig. 6 illustrates the effect of λ in the consistency loss. As λ increases, the accuracy of variance estimation decreases and fewer samples are constrained by R_σ. Beyond $\lambda > 40.0$, the impact of further increases becomes

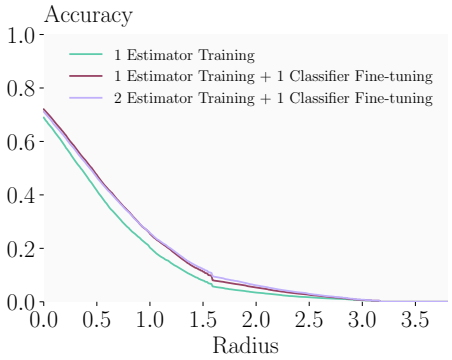


Figure 7: Effect of training rounds for the variance estimator and the classifier. *1 Estimator Training* trains the variance estimator using the off-the-shelf classifier. *1 Estimator Training + 1 Classifier Finetuning* finetunes the classifier using the estimated variances by the trained variance estimator. *2 Estimator Training + 1 Classifier Finetuning* further re-train the variance estimator based on the finetuned classifier.

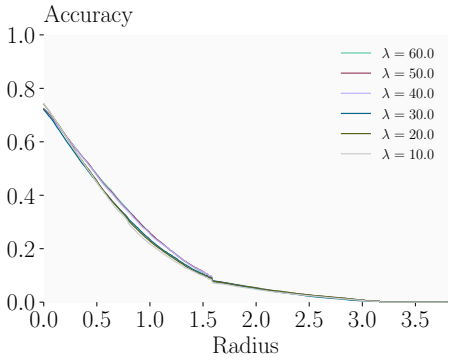


Figure 8: Accuracy–R_{final} curves after *1 Estimator Training + 1 Classifier Finetuning* with different λ in the consistency loss.

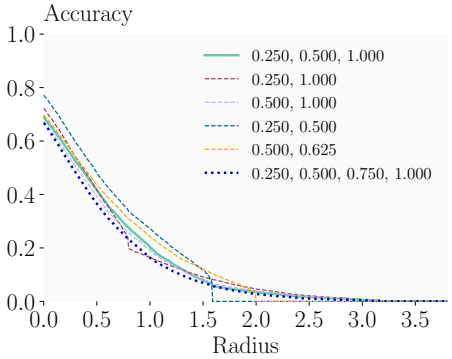
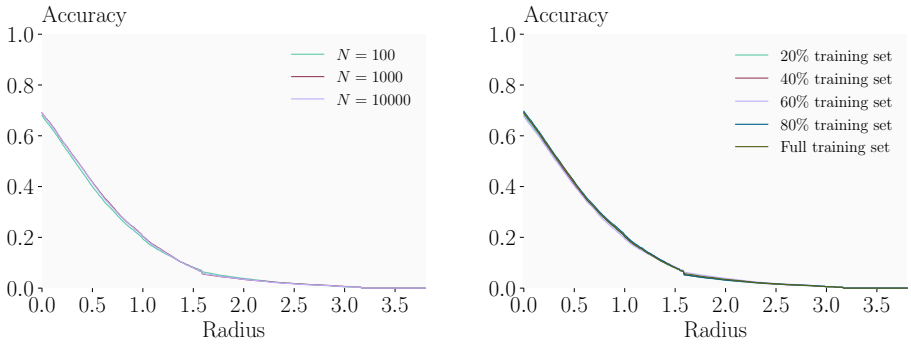


Figure 9: Certified accuracy R_{final} with different σ candidate sets.

Table 6: Certified accuracy (%) at different radii with different σ candidate sets. The denoiser and classifier are fixed (off-the-shelf), and only the variance estimator is trained. The best performance at each radius is highlighted in **bold**, and the worst and second worst are grayed.

σ candidates set	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
{0.25, 0.5, 1.0}	68.91	55.45	41.62	29.48	20.26	13.29	7.95	4.70	3.45	2.40	1.70
{0.25, 1.0}	72.26	57.17	41.74	26.74	16.36	12.62	9.16	6.63	**4.63**	**3.23**	**2.10**
{0.5, 1.0}	67.31	53.37	39.29	27.59	18.99	12.77	8.13	5.40	3.90	2.78	1.90
{0.25, 0.5}	**77.18**	**62.66**	**48.07**	**36.14**	**27.20**	**19.08**	11.92	0.00	0.00	0.00	0.00
{0.5, 0.625}	69.71	56.91	43.64	32.78	24.24	17.11	**11.96**	**7.75**	0.00	0.00	0.00
{0.25, 0.5, 0.75, 1.0}	66.77	51.35	36.62	25.11	16.35	10.72	6.97	4.34	2.74	1.81	1.08



(a) (b)

Figure 10: Study on reducing the training set construction cost. (a) Accuracy R_{final} curves with different number of samples N when calculating the certified radius for each σ candidate. (b) Accuracy R_{final} curves with different portion of training data used to train the variance estimator.

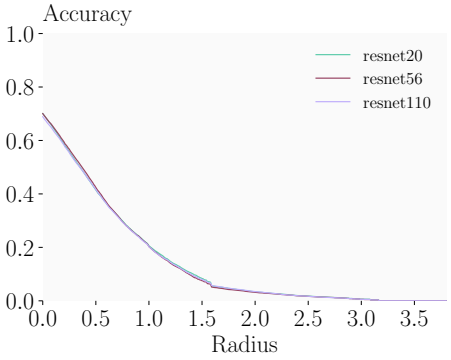


Figure 11: Accuracy R_{final} curves with different σ estimator architectures.

negligible. With a moderate value (e.g., $\lambda = 40.0$), dual RS achieves strong performance in the medium-radius region, while incurring a slight performance drop in the small-radius region.

E.2 Iterative Training

Fig. 7 shows training the variance estimator and finetune the classifier for different number of times. Finetuning the classifier with the estimated variances significantly improves the performance of dual RS across all radii. Moreover, re-training the variance estimator after classifier finetuning yields additional gains, particularly in the medium-radius region. Fig. 8 reports the Accuracy–R_{final} curves after *1 Estimator Training + 1 Classifier Finetuning* with different values of λ in the consistency loss. Compared with Fig. 6c, which uses the off-the-shelf classifier, the influence of λ becomes larger after finetuning. This occurs because larger λ induces larger R_σ, thereby constraining fewer samples after finetuning, which amplifies the performance gap across different λ values.

E.3 Choice of σ Candidates

Dual RS requires selecting a set of candidate noise variances. We conduct an ablation study to study how this choice affects certified accuracy on CIFAR-10. In this ablation study, we use the off-the-shelf denoiser and classifier, and train the variance estimator only. In addition to the original candidate set $\{0.25, 0.5, 1.0\}$, we evaluate five alternative candidate sets: $\{0.25, 0.5\}$, $\{0.5, 0.625\}$, $\{0.25, 1.0\}$, $\{0.5, 1.0\}$, and $\{0.25, 0.5, 0.75, 1.0\}$. For each set, we use the maximum of the candidates as the estimator's global noise level, σ_e. Fig. 9 shows the accuracy - R_{final} curves and Table 6 presents the numerical results.

Compared with $\{0.25, 0.5, 1.0\}$, using only two candidates ($\{0.25, 0.5\}$, $\{0.5, 0.625\}$, $\{0.25, 1.0\}$, or $\{0.5, 1.0\}$) leads to performance degradation at radii unfavored by the candidate set, but improving the performance at radii favored by the candidate set. Specifically, the candidate set $\{0.25, 0.5\}$ and $\{0.5, 0.625\}$ cannot achieve non-trivial accuracy at radii larger than 2.0, but achieve stronger performance at radii smaller than 2.0. The candidate set $\{0.25, 1.0\}$ leads to reduced accuracy in the medium-radii range (radii from 0.75 to 1.25), but improves at both small and large radii (radii smaller than 0.75 or larger than 1.25).

Increasing the number of candidates, e.g., using $\{0.25, 0.5, 0.75, 1.0\}$, does not improve performance over $\{0.25, 0.5, 1.0\}$, potentially due to the increased difficulty of accurately estimating the optimal σ and obtaining a sufficiently large certified radius for the estimated σ.

E.4 Reducing Training Cost

As shown in Table 4, constructing the optimal variance dataset contributes the majority of the training cost. To reduce this overhead, we explore two strategies: (1) using a smaller number of samples N to calculate the certified radius when constructing the training dataset, and (2) using only a subset of the training data to train the variance estimator. All experiments here use off-the-shelf denoiser and classifier, on the CIFAR-10 dataset.

In the main CIFAR-10 experiments, we use $N = 10^4$ samples to estimate the radius used for training. However, an approximated radius might be sufficient for training: a much smaller N can be used to estimate $\hat{p_A}$, then this weaker estimation can be plugged into the radius formula to compute an approximation of the certified radius. To see if reduction on N is possible, we additionally explore $N = 100$ and $N = 1000$ in this paradigm. This reduces the dataset construction cost by 99% and 90%, respectively. As shown in Fig. 10a, reducing N has minimal effect on performance, demonstrating that the train set of the variance estimator can be constructed much more efficiently. For ImageNet, we adopt the same strategy, using $N = 100$ when building the training set, while still achieving strong performance (see §6.1). We note that this cost then matches the cost of inference, which typically also requires hundreds of queries to make a single prediction.

We also study whether the variance estimator can be trained only on a subset of the training data. Specifically, we randomly sample $\{20\%, 40\%, 60\%, 80\%\}$ of the training data and train the variance estimator solely based on the sampled training subset, thereby cutting down the training cost respectively. As shown in Fig. 10b, using a significantly smaller training set for the variance estimator has

minimal effect on the performance. This shows that a relatively small portion of the training data is sufficient to train a high-quality variance estimator.

In summary, both strategies substantially reduce the training cost while maintaining estimator performance. Therefore, in practice, we suggest to start with a small subset of the training data and estimate the radius based on small N, then progressively grow the size and the estimation quality until the performance gain diminishes.

E.5 Architecture of the Variance Estimator

In the main CIFAR-10 experiments, we used a ResNet-110 estimator, which is a standard choice in training-based RS. We additionally evaluate smaller architectures (ResNet-20 and ResNet-56) while keeping the denoiser and classifier fixed. As shown in Fig. 11, using a smaller variance estimator has minimal effect on the accuracy - R_{final} curves. This indicates that even though smaller estimators have lower representational capacity, they remain sufficiently expressive to approximate locally constant variance in practice.