TSS: Transformation-Specific Smoothing for Robustness Certification

Linyi Li*
University of Illinois
Urbana, Illinois, USA
linyi2@illinois.edu

Maurice Weber*
ETH Zürich
Zürich, Switzerland
maurice.weber@inf.ethz.ch

Xiaojun Xu
University of Illinois
Urbana, Illinois, USA
xiaojun3@illinois.edu

Luka Rimanic
ETH Zürich
Zürich, Switzerland
luka.rimanic@inf.ethz.ch

Bhavya Kailkhura
Lawrence Livermore National Laboratory
Livermore, California, USA
kailkhura1@llnl.gov

Tao Xie
Key Laboratory of High Confidence Software Technologies, MoE (Peking University)
Beijing, China
taoxie@pku.edu.cn

Ce Zhang
ETH Zürich
Zürich, Switzerland
ce.zhang@inf.ethz.ch

Bo Li
University of Illinois
Urbana, Illinois, USA
lbo@illinois.edu

ABSTRACT

As machine learning (ML) systems become pervasive, safeguarding their security is critical. However, recently it has been demonstrated that motivated adversaries are able to mislead ML systems by perturbing test data using semantic transformations. While there exists a rich body of research providing provable robustness guarantees for ML models against ℓ_p norm bounded adversarial perturbations, guarantees against semantic perturbations remain largely underexplored. In this paper, we provide **TSS**—*a unified framework for certifying ML robustness against general adversarial semantic transformations*. First, depending on the properties of each transformation, we divide common transformations into two categories, namely *resolvable* (e.g., Gaussian blur) and *differentially resolvable* (e.g., rotation) transformations. For the former, we propose transformation-specific randomized smoothing strategies and obtain strong robustness certification. The latter category covers transformations that involve interpolation errors, and we propose a novel approach based on stratified sampling to certify the robustness. Our framework **TSS** leverages these certification strategies and combines with consistency-enhanced training to provide rigorous certification of robustness. We conduct extensive experiments on over ten types of challenging semantic transformations and show that **TSS** significantly outperforms the state of the art. Moreover, to the best of our knowledge, **TSS** is the first approach that achieves nontrivial certified robustness on the large-scale ImageNet dataset. For instance, our framework achieves 30.4% certified robust accuracy against rotation attack (within ±30°) on ImageNet. Moreover, to consider a broader range of transformations, we show **TSS** is also robust against adaptive attacks and unforeseen image corruptions such as CIFAR-10-C and ImageNet-C.

CCS CONCEPTS

• **General and reference** → **Verification**; • **Security and privacy** → **Logic and verification**; *Software security engineering*; • **Computing methodologies** → *Neural networks.*

KEYWORDS

Certified Robustness, Semantic Transformation Attacks

ACM Reference Format:
Linyi Li*, Maurice Weber*, Xiaojun Xu, Luka Rimanic, Bhavya Kailkhura, Tao Xie, Ce Zhang, and Bo Li. 2021. TSS: Transformation-Specific Smoothing for Robustness Certification . In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21), November 15–19, 2021, Virtual Event, Republic of Korea.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3460120.3485258

1 INTRODUCTION

Recent advances in machine learning (ML) have enabled a plethora of applications in tasks such as image recognition [19] and game playing [36, 46]. Despite all of these advances, ML systems are also found exceedingly vulnerable to adversarial attacks: image recognition systems can be adversarially misled [16, 48, 59], and malware detection systems can be evaded easily [52, 62].

The existing practice of security in ML has fallen into the cycle where new empirical defense techniques are proposed [31, 54], followed by new adaptive attacks breaking these defenses [1, 12, 16, 60]. In response, recent research has attempted to provide *provable robustness guarantees* for an ML model. Such certification usually follows the form that the ML model is provably robust against

* Equal contribution.

CCS '21, November 15–19, 2021, Virtual Event, Republic of Korea

ACM ISBN 978-1-4503-8454-4/21/11…$15.00
https://doi.org/10.1145/3460120.3485258

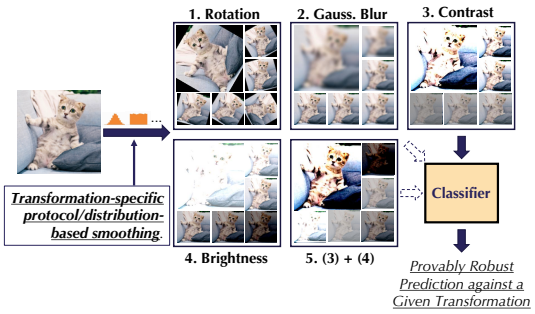


Figure 1: ***T**ransformation-**S**pecific **S**moothing-based robustness certification*, a general robustness certification framework against various semantic transformations. We develop a range of different transformation-specific smoothing protocols and various techniques to provide substantially better certified robustness bounds than state-of-the-art approaches on large-scale datasets.

arbitrary adversarial attacks, as long as the perturbation magnitude is below a certain threshold. Different certifiable defenses and robustness verification approaches have provided nontrivial robustness guarantees against ℓ_p perturbations where the perturbation is bounded by small ℓ_p norm [7, 30, 51, 57, 61].

However, certifying robustness only against ℓ_p perturbations is not sufficient for attacks based on semantic transformation. For instance, image rotation, scaling, and other semantic transformations are able to mislead ML models effectively [11, 14, 15, 60]. These transformations are common and practical [4, 20, 38]. For example, it has been shown [22] that brightness/contrast attacks can achieve 91.6% attack success on CIFAR-10, and 71%-100% attack success rate on ImageNet [20]. In practice, brightness- and contrast-based attacks have been demonstrated to be successful in autonomous driving [38, 50]. These attacks incur large ℓ_p-norm differences and are thus beyond the reach of existing certifiable defenses [3, 18, 25, 43]. *Can we provide provable robustness guarantees against these semantic transformations?*

In this paper, we propose theoretical and empirical analyses to certify the ML robustness against a wide range of semantic transformations beyond ℓ_p bounded perturbations. The theoretical analysis is nontrivial given different properties of the transformations, and our empirical results set the new state-of-the-art robustness certification for a range of semantic transformations, exceeding existing work by a large margin. In particular, we propose **T**ransformation-**S**pecific **S**moothing-based robustness certification — a *general framework* based on function smoothing providing certified robustness for ML models against a range of adversarial transformations (Figure 1). To this end, we first categorize semantic transformations as either *resolvable* or *differentially resolvable*. We then provide certified robustness against *resolvable* transformations, which include brightness, contrast, translation, Gaussian blur, and their composition. Second, we develop novel certification techniques for *differentially resolvable* transformations (e.g., rotation and scaling), based on the *building block* that we have developed for resolvable transformations.

For resolvable transformations, we leverage the framework to *jointly* reason about (1) function smoothing under different smoothing distributions and (2) the properties inherent to each specific transformation. To our best knowledge, this is the first time that the interplay between smoothing distribution and semantic transformation has been analyzed as existing work [7, 27, 63] that studies different smoothing distributions considers only ℓ_p perturbations. Based on this analysis, we find that against certain transformations such as Gaussian blur, exponential distribution is better than Gaussian smoothing, which is commonly used in the ℓ_p-case.

For differentially resolvable transformations, such as rotation, scaling, and their composition with other transformations, the common **challenge** is that they naturally induce *interpolation error*. Existing work [2, 13] can provide robustness guarantees but it cannot rigorously certify robustness for ImageNet-scale data. We develop a collection of novel techniques, including stratified sampling and Lipschitz bound computation to provide a tighter and sound upper bound for the interpolation error. We integrate these novel techniques into our **TSS** framework and further propose a progressive-sampling-based strategy to accelerate the robustness certification. We show that these techniques comprise a scalable and general framework for certifying robustness against differentially resolvable transformations.

We conduct extensive experiments to evaluate the proposed certification framework and show that our framework significantly outperforms the state-of-the-art on different datasets including the large-scale ImageNet against a series of practical semantic transformations. In summary, this paper makes the following contributions:

(*1*) We propose a *general* function smoothing framework, **TSS**, to certify ML robustness against general semantic transformations.
(*2*) We categorize common adversarial semantic transformations in the literature into *resolvable* and *differentially resolvable* transformations and show that our framework is general enough to certify both types of transformations.
(*3*) We theoretically explore different smoothing strategies by sampling from different distributions including non-isotropic Gaussian, uniform, and Laplace distributions. We show that for specific transformations, such as Gaussian blur, smoothing with exponential distribution is better.
(*4*) We propose a pipeline, **TSS-DR**, including a stratified sampling approach, an effective Lipschitz-based bounding technique, and a progressive sampling strategy to provide rigorous, tight, and scalable robustness certification against differentially resolvable transformations such as rotation and scaling.
(*5*) We conduct extensive experiments and show that our framework **TSS** can provide significantly higher certified robustness compared with the state-of-the-art approaches, against a range of semantic transformations and their composition on MNIST, CIFAR-10, and ImageNet.
(*6*) We show that **TSS** also provides much higher empirical robustness against adaptive attacks and unforeseen corruptions such as CIFAR-10-C and ImageNet-C.

The code implementation and all trained models are publicly available at https://github.com/AI-secure/semantic-randomized-smoothing.

2 BACKGROUND

We next provide an overview of different semantic transformations and explain the intuition behind the randomized smoothing [7] that has been proposed to certify the robustness against ℓ_p perturbations.

Semantic Transformation Based Attacks. Beyond adversarial ℓ_p perturbations, a realistic threat model is given by image transformations that preserve the underlying semantics. Examples for these types of transformations include changes to contrast or brightness levels, or rotation of the entire image. These attacks share three common characteristics: (1) The perturbation stemming from a successful semantic attack typically has higher ℓ_p norm compared to ℓ_p-bounded attacks. However, these attacks still preserve the underlying semantics (a car image rotated by 10° still contains a car). (2) These attacks are governed by a low-dimensional parameter space. For example, the rotation attack chooses a single-dimensional rotation angle. (3) Some of such adversarial transformations would lead to high interpolation error (e.g., rotation), which makes it challenging to certify. Nevertheless, these types of attacks can also cause significant damage [20, 22] and pose realistic threats for practical ML applications such as autonomous driving [38]. We remark that our proposed framework can be extended to certify robustness against other attacks sharing these characteristics even beyond the image domain, such as GAN-based attacks against ML based malware detection [23, 55], where a limited dimension of features of the malware can be manipulated in order to preserve the malicious functionalities and such perturbation usually incurs large ℓ_p differences for the generated instances.

Randomized Smoothing. On a high level, randomized smoothing [7, 26, 27] provides a way to certify robustness based on randomly perturbing inputs to an ML model. The intuition behind such randomized classifier is that noise smoothens the decision boundaries and suppresses regions with high curvature. Since adversarial examples aim to exploit precisely these high curvature regions, the vulnerability to this type of attack is reduced. Formally, a base classifier h is smoothed by adding noise ε to a test instance. The prediction of the smoothed classifier is then given by the most likely prediction under this smoothing distribution. Subsequently, a tight robustness guarantee can be obtained, based on the noise variance and the class probabilities of the smoothed classifier. It is guaranteed that, as long as the ℓ_2 norm of the perturbation is bounded by a certain amount, the prediction on an adversarial vs. benign input will stay the same. This technique provides a powerful framework to study the robustness of classification models against adversarial attacks for which the primary figure of merit is a low ℓ_p norm with a simultaneously high success rate of fooling the classifier [10, 63]. However, semantic transformations incur large ℓ_p perturbations, which renders classical randomized smoothing infeasible [3, 18, 25], making it of great importance to generalize randomized smoothing to this kind of threat model.

3 THREAT MODEL & TSS OVERVIEW

In this section, we first introduce the notations used throughout this paper. We then define our **threat model**, the **defense goal** and outline the **challenges** for certifying the robustness against semantic transformations. Finally, we will provide a brief **overview** of our **TSS** certification framework.

We denote the space of inputs as $\mathcal{X} \subseteq \mathbb{R}^d$ and the set of labels as $\mathcal{Y} = \{1, \ldots, C\}$ (where $C \geq 2$ is the number of classes). The set of transformation parameters is given by $\mathcal{Z} \subseteq \mathbb{R}^m$ (e.g., rotation angles). We use the notation $\mathbb{P}_X$ to denote the probability measure induced by the random variable X and write f_X for its probability density function. For a set S, we denote its probability by $\mathbb{P}_X(S)$. A classifier is defined to be a deterministic function h mapping inputs $x \in \mathcal{X}$ to classes $y \in \mathcal{Y}$. Formally, a classifier learns a conditional probability distribution $p(y|x)$ over labels and outputs the class that maximizes p, i.e., $h(x) = \arg\max_{y \in \mathcal{Y}} p(y|x)$.

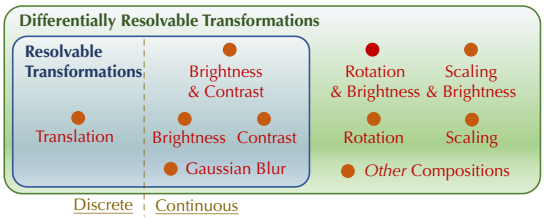


Figure 2: We provide strong robustness certification for both resolvable transformations and differentially revolvable transformations. These two categories cover common adversarial semantic transformations.

3.1 Threat Model and Certification Goal

Semantic Transformations. We model semantic transformations as deterministic functions $\phi : \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$, transforming an image $x \in \mathcal{X}$ with a $\mathcal{Z}$-valued parameter α. For example, we use $\phi_R(x, \alpha)$ to model a rotation of the image x by α degrees counter-clockwise with bilinear interpolation. We further partition semantic transformations into two different categories, namely resolvable and differentially resolvable transformations. We will show that these two categories could cover commonly known semantic attacks. This categorization depends on whether or not it is possible to write the composition of the transformation ϕ with itself as applying the same transformation just once, but with a different parameter, i.e., whether for any $\alpha, \beta \in \mathcal{Z}$ there exists γ such that $\phi(\phi(x, \alpha), \beta) = \phi(x, \gamma)$. Precise definitions are given in Sections 5 and 6. Figure 2 presents an overview of the transformations considered in this work.

Threat model. We consider an adversary that launches a semantic attack, a type of data evasion attack, against a given classification model h by applying a semantic transformation ϕ with parameter α to an input image $x \to \phi(x, \alpha)$. We allow the attacker to choose an *arbitrary* parameter α within a predefined (attack) parameter space $\mathcal{S}$. For instance, a naïve adversary who randomly changes brightness from within ±40% is able to reduce the accuracy of a state-of-the-art ImageNet classifier from 74.4% to 21.8% (Table 2). While this attack is an example random adversarial attack, our threat model also covers other types of semantic attacks and we provide the first taxonomy for semantic attacks (i.e., resolvable and differentially resolvable) in detail in Sections 5 and 6.

Certification Goal. Since the only degree of freedom that a semantic adversary has is the parameter, **our goal** is to characterize a set of parameters for which the model under attack is guaranteed to be robust. Formally, we wish to find a set $\mathcal{S}_{\text{adv}} \subseteq \mathcal{Z}$ such that, for a classifier h and adversarial transformation ϕ, we have

$$h(x) = h(\phi(x, \alpha)) \quad \forall\, \alpha \in \mathcal{S}_{\text{adv}}. \tag{1}$$

Challenges for Certifying Semantic Transformations. Certifying ML robustness against semantic transformations is nontrivial and

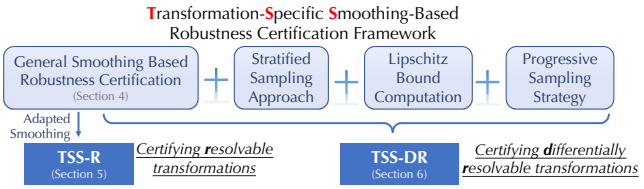


Figure 3: An overview of **TSS**.

requires careful analysis. We identify the following two main challenges that we aim to address in this paper:

- **(C1)** The absolute difference between semantically transformed images in terms of ℓ_p-norms is typically high. This factor causes existing certifiable defenses against ℓ_p bounded perturbations to be inapplicable [3, 18, 25, 43].
- **(C2)** Certain semantic transformations incur additional *interpolation errors*. To derive a robustness certificate, it is required to bound these errors, an endeavour that has been proven to be hard both analytically and computationally. This challenge applies to transformations that involve interpolation, such as rotation and scaling.

We remark that it is in general not feasible to use brute-force approaches such as grid search to enumerate all possible transformation parameters (e.g., rotation angles) since the parameter spaces are typically continuous. Given that different transformations have their own unique properties, it is crucial to provide a *unified* framework that takes into account transformation-specific properties in a general way.

To address these challenges, we generalize randomized smoothing via our proposed *function smoothing framework* to certify arbitrary input transformations via different smoothing distributions, paving the way to robustness certifications that go beyond ℓ_p perturbations. This result addresses challenge **(C1)** in a unified way. Based on this generalization and depending on specific transformation properties, we address challenge **(C2)** and propose a series of smoothing strategies and computing techniques that provide robustness certifications for a diverse range of transformations.

We next introduce our generalized function smoothing framework and show how it can be leveraged to certify semantic transformations. We then categorize transformations as either *resolvable* transformations (Section 5) such as Gaussian blur, or *differentially resolvable* transformations (Section 6) such as rotations.

3.2 Framework Overview

An overview of our proposed framework **TSS** is given in Figure 3. We propose the function smoothing framework, a generalization of randomized smoothing, to provide robustness certifications under general smoothing distributions (Section 4). This generalization enables us to smooth the model on specific transformation dimensions. We then consider two different types of transformation attacks. For *resolvable* transformations, using function smoothing framework, we adapt different smoothing strategies for specific transformations and propose **TSS-R** (Section 5). We show that some smoothing distributions are more suitable for certain transformations. For *differentially resolvable* transformations, to address the interpolation error, we combine function smoothing with the proposed stratified sampling approach and a novel technique for Lipschitz bound computation to compute a rigorous upper bound of the error. We then develop a progressive sampling strategy to accelerate the certification. This pipeline is termed **TSS-DR**, and we provide details and the theoretical groundwork in Section 6.

4 TSS: TRANSFORMATION SPECIFIC SMOOTHING BASED CERTIFICATION

In this section, we extend randomized smoothing and propose a function smoothing framework **TSS** (**T**ransformation-**S**pecific **S**moothing-based robustness certification) for certifying robustness against semantic transformations. This framework constitutes the main building block for **TSS-R** and **TSS-DR** against specific types of adversarial transformations.

Given an arbitrary base classifier h, we construct a smoothed classifier g by randomly transforming inputs with parameters sampled from a smoothing distribution. Specifically, given an input x, the smoothed classifier g predicts the class that h is most likely to return when the input is perturbed by some random transformation. We formalize this intuition in the following definition.

Definition 1 (ε-Smoothed Classifier). *Let $\phi\colon \mathcal{X}\times\mathcal{Z}\to\mathcal{X}$ be a transformation, $\varepsilon\sim\mathbb{P}_\varepsilon$ a random variable taking values in $\mathcal{Z}$ and let $h\colon\mathcal{X}\to\mathcal{Y}$ be a base classifier. We define the ε-smoothed classifier $g\colon\mathcal{X}\to\mathcal{Y}$ as $g(x;\varepsilon)=\arg\max_{y\in\mathcal{Y}} q(y|\,x;\,\varepsilon)$ where q is given by the expectation with respect to the smoothing distribution ε, i.e.,*

$$q(y|\,x;\,\varepsilon) := \mathbb{E}(p(y|\,\phi(x,\,\varepsilon))). \tag{2}$$

A key to certifying robustness against a specific transformation is the choice of transformation ϕ in the definition of the smoothed classifier (2). For example, if the goal is to certify the Gaussian blur transformation, a reasonable choice is to use the same transformation in the smoothed classifier. However, for other types of transformations this choice does not lead to the desired robustness certificate, and a different approach is required. In Sections 5 and 6, we derive approaches to overcome this challenge and certify robustness against a broader family of semantic transformations.

General Robustness Certification. Given an input $x\in\mathcal{X}$ and a random variable ε taking values in $\mathcal{Z}$, suppose that the base classifier h predicts $\phi(x,\varepsilon)$ to be of class y_A with probability at least p_A and the second most likely class with probability at most p_B (i.e., (4)). Our goal is to derive a robustness certificate for the ε-smoothed classifier g, i.e., we aim to find a set of perturbation parameters $\mathcal{S}_{\text{adv}}$ depending on p_A, p_B, and smoothing parameter ε such that, for all possible perturbation $\alpha\in\mathcal{S}_{\text{adv}}$, it is guaranteed that

$$g(\phi(x,\,\alpha);\,\varepsilon) = g(x;\,\varepsilon) \tag{3}$$

In other words, the prediction of the smoothed classifier can never be changed by applying the transformation ϕ with parameters α that are in the robust set $\mathcal{S}_{\text{adv}}$. The following theorem provides a generic robustness condition that we will subsequently leverage to obtain conditions on transformation parameters. In particular, this result addresses the first challenge **(C1)** for certifying semantic transformations since this result allows to certify robustness beyond additive perturbations.

Theorem 1. *Let $\varepsilon_0\sim\mathbb{P}_0$ and $\varepsilon_1\sim\mathbb{P}_1$ be $\mathcal{Z}$-valued random variables with probability density functions f_0 and f_1 with respect to a*

measure μ on $\mathcal{Z}$ and let $\phi\colon \mathcal{X}\times\mathcal{Z} \to \mathcal{X}$ be a semantic transformation. Suppose that $y_A = g(x;\, \varepsilon_0)$ and let $p_A,\, p_B \in [0,\, 1]$ be bounds to the class probabilities, i.e.,

$$q(y_A|\, x,\, \varepsilon_0) \geq p_A > p_B \geq \max_{y \neq y_A} q(y|\, x,\, \varepsilon_0). \tag{4}$$

For $t \geq 0$, let $\underline{S}_t,\, \overline{S}_t \subseteq \mathcal{Z}$ be the sets defined as $\underline{S}_t := \{f_1/f_0 < t\}$ and $\overline{S}_t := \{f_1/f_0 \leq t\}$ and define the function $\xi\colon [0,\, 1] \to [0,\, 1]$ by

$$\begin{aligned} \xi(p) &:= \sup\{\mathbb{P}_1(S)\colon\, \underline{S}_{\tau_p} \subseteq S \subseteq \overline{S}_{\tau_p}\} \\ \text{where} \quad \tau_p &:= \inf\{t \geq 0\colon\, \mathbb{P}_0(\overline{S}_t) \geq p\}. \end{aligned} \tag{5}$$

Then, if the condition

$$\xi(p_A) + \xi(1 - p_B) > 1 \tag{6}$$

is satisfied, then it is guaranteed that $g(x;\, \varepsilon_1) = g(x;\, \varepsilon_0)$.

A detailed proof for this statement is provided in Appendix C. At a high level, the condition (4) defines a family of classifiers based on class probabilities obtained from smoothing the input x with the distribution ε_0. Based on the Neyman Pearson Lemma from statistical hypothesis testing, shifting $\varepsilon_0 \to \varepsilon_1$ results in bounds to the class probabilities associated with smoothing x with ε_1. For class y_A, the lower bound is given by $\xi(p_A)$, while for any other class the upper bound is given by $1 - \xi(1 - p_B)$, leading to the the robustness condition $\xi(p_A) > 1 - \xi(1 - p_B)$. It is a more general version of what is proved by Cohen et al. [7], and its generality allows us to analyze a larger family of threat models. Notice that it is not immediately clear how one can obtain the robustness guarantee (3) and deriving such a guarantee from Theorem 1 is nontrivial. We will therefore explain in detail how this result can be instantiated to certify semantic transformations in Sections 5 and 6.

5 TSS-R: Resolveable Transformations

In this section, we define resolvable transformations and then show how Theorem 1 is used to certify this class of semantic transformations. We then proceed to providing a robustness verification strategy for each specific transformation. In addition, we show how the generality of our framework allows us to reason about the best smoothing strategy for a given transformation, which is beyond the reach of related randomized smoothing based approaches [13, 63].

Intuitively, we call a semantic transformation resolvable if we can separate transformation parameters from inputs with a function that acts on parameters and satisfies certain regularity conditions.

Definition 2 (Resolvable transform). *A transformation $\phi\colon \mathcal{X}\times\mathcal{Z} \to \mathcal{X}$ is called resolvable if for any $\alpha \in \mathcal{Z}$ there exists a resolving function $\gamma_\alpha\colon \mathcal{Z} \to \mathcal{Z}$ that is injective, continuously differentiable, has non-vanishing Jacobian and for which*

$$\phi(\phi(x,\, \alpha),\, \beta) = \phi(x,\, \gamma_\alpha(\beta)) \quad x \in \mathcal{X},\, \beta \in \mathcal{Z}. \tag{7}$$

Furthermore, we say that ϕ is additive, if $\gamma_\alpha(\beta) = \alpha + \beta$.

The following result provides a more intuitive view on Theorem 1, expressing the condition on probability distributions as a condition on the transformation parameters.

Corollary 1. *Suppose that the transformation ϕ in Theorem 1 is resolvable with resolving function γ_α. Let $\alpha \in \mathcal{Z}$ and set $\varepsilon_1 := \gamma_\alpha(\varepsilon_0)$ in the definition of the function ξ. Then, if α satisfies condition (6), it is guaranteed that $g(\phi(x, \alpha);\, \varepsilon_0) = g(x;\, \varepsilon_0)$.*

This corollary implies that for resolvable transformations, after we choose the smoothing distribution for the random variable ε_0, we can infer the distribution of $\varepsilon_1 = \gamma_\alpha(\varepsilon_0)$. Then, by plugging in ε_0 and ε_1 into Theorem 1, we can derive an explicit robustness condition from (6) such that for any α satisfying this condition, we can certify the robustness. In particular, for additive transformations we have $\varepsilon_1 = \gamma_\alpha(\varepsilon_0) = \alpha + \varepsilon_0$. For common smoothing distributions ε_0 along with additive transformation, we derive robustness conditions in Appendix D.

In the next subsection, we focus on specific resolvable transformations. For certain transformations, this result can be applied directly. However, for some transformations, e.g., the composition of brightness and contrast, more careful analysis is required. We remark that this corollary also serves a stepping stone to certifying more complex transformations that are in general not resolvable, such as rotations as we will present in Section 6.

5.1 Certifying Specific Transformations

Here we build on our theoretical results from the previous section and provide approaches to certifying a range of different semantic transformations that are resolvable. We state all results here and provide proofs in appendices.

5.1.1 Gaussian Blur. This transformation is widely used in image processing to reduce noise and image detail. Mathematically, applying Gaussian blur amounts to convolving an image with a Gaussian function

$$G_\alpha(k) = \frac{1}{\sqrt{2\pi\alpha}} \exp\left(-k^2/(2\alpha)\right) \tag{8}$$

where $\alpha > 0$ is the *squared* kernel radius. For $x \in \mathcal{X}$, we define Gaussian blur as the transformation $\phi_B\colon \mathcal{X} \times \mathbb{R}_{\geq 0} \to \mathcal{X}$ where

$$\phi_B(x,\, \alpha) = x * G_\alpha \tag{9}$$

and $*$ denotes the convolution operator. The following lemma shows that Gaussian blur is an *additive transform*. Thus, existing robustness conditions for additive transformations shown in Appendix D are directly applicable.

Lemma 1. *The Gaussian blur transformation is additive, i.e., for any $\alpha,\, \beta \geq 0$, we have $\phi_B(\phi_B(x,\, \alpha),\, \beta) = \phi_B(x,\, \alpha + \beta)$.*

We notice that the Gaussian blur transformation uses only positive parameters. We therefore consider uniform noise on $[0,\, a]$ for $a > 0$, folded Gaussians and exponential distribution for smoothing.

5.1.2 Brightness and Contrast. This transformation first changes the brightness of an image by adding a constant value $b \in \mathbb{R}$ to every pixel, and then alters the contrast by multiplying each pixel with a positive factor e^k, for some $k \in \mathbb{R}$. We define the brightness and contrast transformation $\phi_{BC}\colon \mathcal{X} \times \mathbb{R}^2 \to \mathcal{X}$ as

$$(x,\, \alpha) \mapsto \phi_{BC}(x,\, \alpha) := e^k(x + b), \quad \alpha = (k,\, b)^T \tag{10}$$

where $k,\, b \in \mathbb{R}$ are contrast and brightness parameters, respectively. We remark that ϕ_{BC} is resolvable; however, it is not additive and applying Corollary 1 directly using the resolving function γ_α leads to analytically intractable expressions. On the other hand, if the parameters k and b follow independent Gaussian distributions, we can circumvent this difficulty as follows. Given $\varepsilon_0 \sim \mathcal{N}(0,\, \mathrm{diag}(\sigma^2,\, \tau^2))$,

we compute the bounds p_A and p_B to the class probabilities associated with the classifier $g(x; \varepsilon_0)$, i.e., smoothed with ε_0. In the next step, we identify a distribution ε_1 with the property that we can map any lower bound p of $q(y|\, x; \varepsilon_0)$ to a lower bound on $q(y|\, x; \varepsilon_1)$. Using ε_1 as a bridge, we then derive a robustness condition, which is based on Theorem 1, and obtain the guarantee that $g(\phi_{BC}(x, \alpha); \varepsilon_0) = g(x; \varepsilon_0)$ whenever the transformation parameters satisfy this condition. The next lemma shows that the distribution ε_1 with the desired property (lower bound to the classifier smoothed with ε_1) is given by a Gaussian with transformed covariance matrix.

Lemma 2. *Let $x \in \mathcal{X}$, $k \in \mathbb{R}$, $\varepsilon_0 \sim \mathcal{N}(0, \operatorname{diag}(\sigma^2, \tau^2))$ and $\varepsilon_1 \sim \mathcal{N}(0, \operatorname{diag}(\sigma^2, e^{-2k}\tau^2))$. Suppose that $q(y|\, x; \varepsilon_0) \geq p$ for some $p \in [0, 1]$ and $y \in \mathcal{Y}$. Let Φ be the cumulative density function of the standard Gaussian. Then*

$$q(y|\, x; \varepsilon_1) \geq \begin{cases} 2\Phi\left(e^k \Phi^{-1}\left(\frac{1+p}{2}\right)\right) - 1 & k \leq 0 \\ 2\left(1 - \Phi\left(e^k \Phi^{-1}(1 - \frac{p}{2})\right)\right) & k > 0. \end{cases} \tag{11}$$

Now suppose that $g(\cdot; \varepsilon_0)$ makes the prediction y_A at x with probability at least p_A. Then, the preceding lemma tells us that the prediction confidence of $g(\cdot; \varepsilon_1)$ satisfies the lower bound (11) for the same class. Based on these confidence levels, we instantiate Theorem 1 with the random variables ε_1 and $\alpha + \varepsilon_1$ to get a robustness condition.

Lemma 3. *Let ε_0 and ε_1 be as in Lemma 2 and suppose that*

$$q(y_A|\, x; \varepsilon_1) \geq \tilde{p}_A > \tilde{p}_B \geq \max_{y \neq y_A} q(y|\, x; \varepsilon_1). \tag{12}$$

Then it is guaranteed that $y_A = g(\phi_{BC}(x, \alpha); \varepsilon_0)$ as long as $\alpha = (k, b)^T$ satisfies

$$\sqrt{(k/\sigma)^2 + \left(b/(e^{-k}\tau)\right)^2} < \frac{1}{2}\left(\Phi^{-1}(\tilde{p}_A) - \Phi^{-1}(\tilde{p}_B)\right). \tag{13}$$

In practice, we apply this lemma by replacing $\tilde{p}_A$ and $\tilde{p}_B$ in (13) with the bound computed from (11) based on the class probability bounds p_A and p_B associated with the classifier $g(x; \varepsilon_0)$.

5.1.3 Translation. Let $\bar{\phi}_T : \mathcal{X} \times \mathbb{Z}^2 \to \mathcal{X}$ be the transformation moving an image k_1 pixels to the right and k_2 pixels to the bottom with reflection padding. In order to handle continuous noise distributions, we define the translation transformation $\phi_T : \mathcal{X} \times \mathbb{R}^2 \to \mathcal{X}$ as $\phi_T(x, \alpha) = \bar{\phi}_T(x, [\alpha])$ where $[\cdot]$ denotes rounding to the nearest integer, applied element-wise. We note that ϕ_T is an *additive transform*, allowing us to directly apply Corollary 1 and derive robustness conditions. We note that if we use black padding instead of reflection padding, the transformation is not additive. However, since the number of possible translations is finite, another possibility is to use a simple brute-force approach that can handle black padding, which has already been studied extensively [35, 39].

5.1.4 Composition of Gaussian Blur, Brightness, Contrast, and Translation. Interestingly, the composition of all these four transformations is still resolvable. Thus, we are able to derive the explicit robustness condition for this composition based on Corollary 1, as shown in details in Appendix B. Based on this robustness condition, we compute practically meaningful robustness certificates as we will present in experiments in Section 7.

Table 1: Summary of the Robustness Certification Strategies for Resolvable Transformations.

Transformation	Step 1	Step 2
Gaussian Blur	Compute p_A and p_B with Monte-Carlo Sampling	Check via Corollary 8 (in Appendix D)
Brightness		Check via Corollary 7 (in Appendix D)
Translation		Check via Corollary 7 (in Appendix D)
Brightness and Contrast		Compute p_A via Lemma 2, then check via Lemma 3 (detail in Appendix B.1)
Gaussian Blur, Brightness, Contrast and Translation		Compute p_A via Corollary 3, then check via Lemma B.1 (detail in Appendix B.2)

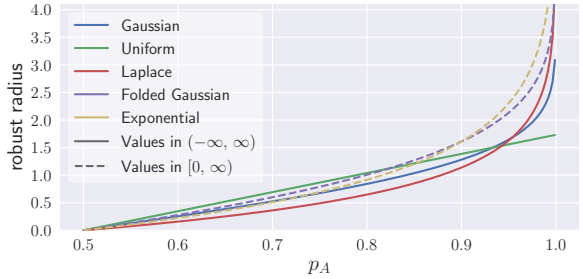


Figure 4: Robust radius comparison for different noise distributions, each with unit variance and dimension.

5.1.5 Robustness Certification Strategies. With these robustness conditions, for a given clean input x, a transformation ϕ, and a set of parameters $\mathcal{S}_{\text{adv}}$, we certify the robustness of the smoothed classifier g with two steps: 1) estimate p_A and p_B (Equation (4)) with Monte-Carlo sampling and high-confidence bound following Cohen et al. [7]; and 2) leverage the robustness conditions to obtain the certificate. A summary for each transformation including the used robustness conditions are shown in Table 1.

5.2 Properties of Smoothing Distributions

The robustness condition in Theorem 1 is generic and leaves a degree of freedom with regards to which smoothing distribution should be used. Previous work mainly provides results for cases in which this distribution is Gaussian [7, 64], while it is nontrivial to extend it to other distributions. Here, we aim to answer this question and provide results for a range of distributions, and discuss their differences. As we will see, *for different scenarios, different distributions behave differently and can certify different radii.* We instantiate Theorem 1 with an arbitrary transformation ϕ and with $\varepsilon_1 := \alpha + \varepsilon_0$ where ε_0 is the smoothing distribution and α is the transformation parameter. The robust radius is then derived by solving condition (6) for α.

Figure 4 illustrates robustness radii associated with different smoothing distributions, each scaled to have unit variance. The bounds are derived in Appendix D and summarized in Table 5. We emphasize that the contribution of this work is not merely these results on different smoothing distributions but, more importantly, the *joint study between different smoothing mechanisms and different semantic transformations.* To compare the different radii for a fixed base classifier, we assume that *the smoothed classifier $g(\cdot; \varepsilon)$ always has the same confidence p_A for noise distributions with equal variance.* Finally, we provide the following conclusions and we will verify them empirically in Section 7.3.1.

(1) *Exponential noise can provide larger robust radius.* We notice that smoothing with exponential noise generally allows for larger adversarial perturbations than other distributions. We also observe that, while all distributions behave similarly for low confidence levels, it is only non-uniform noise distributions that converge toward $+\infty$ when $p_A \to 1$ and exponential noise converges quickest.

(2) *Additional knowledge can lead to larger robust radius.* When we have additional information on the transformation, e.g., all perturbations in Gaussian blur are positive, we can take advantage of this additional information and certify larger radii. For example, under this assumption, we can use folded Gaussian noise for smoothing instead of a standard Gaussian, resulting in a larger radius.

6 TSS-DR: Differentially Resolvable Transformations

As we have seen, our proposed function smoothing framework can directly deal with resolvable transformations. However, due to their use of interpolation, some important transformations do not fall into this category, including rotation, scaling, and their composition with resolvable transformations. In this section, we show that they belong to the more general class termed *differentially resolvable transformations* and to address challenge **(C2)**, we propose a novel pipeline **TSS-DR** to provide rigorous robustness certification using our function smoothing framework as a central building block.

Common semantic transformations such as rotations and scaling do not fall into the category of resolvable transformations due to their use of interpolation. To see this issue, consider for example the rotation transformation denoted by ϕ_R. As shown in Figure 5b, despite very similar, the image rotated by 30° is different from the image rotated separately by 15° and then again by 15°. The reason is the bilinear interpolation occurring during the rotation. Therefore, if the attacker inputs $\phi_R(x, 15)$, the smoothed classifier in Section 5 outputs

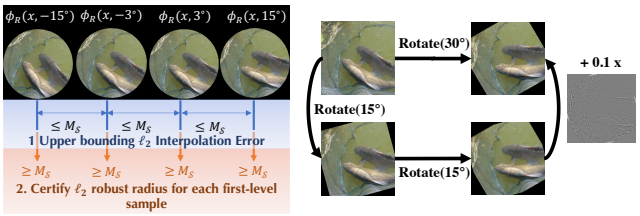
$$g(\phi_R(x,15);\,\varepsilon) = \arg\max_{y\in\mathcal{Y}} \mathbb{E}\left(p(y|\,\phi_R(\phi_R(x,15),\varepsilon))\right), \tag{14}$$

which is a weighted average over the predictions of the base classifier on the randomly perturbed set $\{\phi_R(\phi_R(x,15),\alpha)\colon \alpha \in \mathcal{Z}\}$. However, in order to use Corollary 1 and to reason about whether this prediction agrees with the prediction on the clean input (i.e., the average prediction on $\{\phi_R(x,\alpha)\colon \alpha \in \mathcal{Z}\}$), we need ϕ_R to be resolvable. As it turns out, this is not the case for transformations that involve interpolation such as rotation and scaling.

To address these issues, we define a transformation ϕ to be *differentially resolvable*, if it can be written in terms of a resolvable transformation ψ and a parameter mapping δ.

Definition 3 (Differentially resolvable transform). *Let $\phi\colon \mathcal{X}\times\mathcal{Z}_\phi \to \mathcal{X}$ be a transformation with noise space $\mathcal{Z}_\phi$ and let $\psi\colon \mathcal{X}\times\mathcal{Z}_\psi \to \mathcal{X}$ be a resolvable transformation with noise space $\mathcal{Z}_\psi$. We say that ϕ can be resolved by ψ if for any $x \in \mathcal{X}$ there exists function $\delta_x\colon \mathcal{Z}_\phi \times \mathcal{Z}_\phi \to \mathcal{Z}_\psi$ such that for any $\alpha \in \mathcal{Z}_\phi$ and any $\beta \in \mathcal{Z}_\phi$,*

$$\phi(x,\,\alpha) = \psi(\phi(x,\,\beta),\,\delta_x(\alpha,\,\beta)). \tag{15}$$



(a) (b)

Figure 5: (a) High-level illustration of our robustness certification pipeline **TSS-DR** for differentially resolvable transformations; (b) interpolation error.

This definition leaves open a certain degree of freedom with regard to the choice of resolvable transformation ψ. For example, we can choose the resolvable transformation corresponding to additive noise $(x,\,\delta) \mapsto \psi(x,\,\delta) := x+\delta$, which lets us write any transformation ϕ as $\phi(x,\,\alpha) = \phi(x,\,\beta) + (\phi(x,\,\alpha) - \phi(x,\,\beta)) = \psi(\phi(x,\,\beta),\,\delta)$ with $\delta = (\phi(x,\,\alpha) - \phi(x,\,\beta))$. In other words, $\phi(x,\alpha)$ can be viewed as first being transformed to $\phi(x,\beta)$ and then to $\phi(x,\beta)+\delta$.

6.1 Overview of TSS-DR

Here, we derive a general robustness certification strategy for differentially resolvable transformations. Suppose that our goal is to certify the robustness against a transformation ϕ that can be resolved by ψ and for transformation parameters from the set $\mathcal{S} \subseteq \mathcal{Z}_\phi$. To that end, we first sample a set of parameters $\{\alpha_i\}_{i=1}^N \subseteq \mathcal{S}$, and transform the input (with those sampled parameters) that yields $\{\phi(x,\alpha_i)\}_{i=1}^N$. In the second step, we compute the class probabilities for each transformed input $\phi(x,\,\alpha_i)$ with the classifier smoothed with the resolvable transformation ψ. Finally, the intuition is that, if every $\alpha \in \mathcal{S}$ is close enough to one of the sampled parameters, then the classifier is guaranteed to be robust against parameters from the set $\mathcal{S}$. In the next theorem, we show the existence of such a "proximity set" for general δ_x.

Theorem 2. *Let $\phi\colon \mathcal{X}\times\mathcal{Z}_\phi \to \mathcal{X}$ be a transformation that is resolved by $\psi\colon \mathcal{X}\times\mathcal{Z}_\psi \to \mathcal{X}$. Let $\varepsilon \sim \mathbb{P}_\varepsilon$ be a $\mathcal{Z}_\psi$-valued random variable and suppose that the smoothed classifier $g : \mathcal{X} \to \mathcal{Y}$ given by $q(y|\,x;\varepsilon) = \mathbb{E}(p(y|\,\psi(x,\,\varepsilon)))$ predicts $g(x;\,\varepsilon) = y_A = \arg\max_y q(y|\,x;\varepsilon)$. Let $\mathcal{S} \subseteq \mathcal{Z}_\phi$ and $\{\alpha_i\}_{i=1}^N \subseteq \mathcal{S}$ be a set of transformation parameters such that for any i, the class probabilities satisfy*

$$q(y_A|\,\phi(x,\,\alpha_i);\,\varepsilon) \geq p_A^{(i)} \geq p_B^{(i)} \geq \max_{y\neq y_A} q(y|\,\phi(x,\,\alpha_i);\,\varepsilon). \tag{16}$$

Then there exists a set $\Delta^* \subseteq \mathcal{Z}_\psi$ with the property that, if for any $\alpha \in \mathcal{S}$, $\exists \alpha_i$ with $\delta_x(\alpha,\,\alpha_i) \in \Delta^*$, then it is guaranteed that

$$q(y_A|\,\phi(x,\,\alpha);\varepsilon) > \max_{y\neq y_A} q(y|\,\phi(x,\,\alpha);\varepsilon). \tag{17}$$

In the theorem, the smoothed classifier $g(\cdot;\,\varepsilon)$ is based on the resolvable transformation ψ that serves as a starting point to certify the target transformation ϕ. To certify ϕ over its parameter space $\mathcal{S}$, we input N transformed samples $\phi(x,\alpha_i)$ to the smoothed classifier $g(\cdot;\,\varepsilon)$. Then, we get Δ^*, the certified robust parameter set for resolvable transformation ψ. This Δ^* means that for any $\phi(x,\alpha_i)$, if we apply the transformation ψ with any parameter $\delta \in \Delta^*$, the resulting instance $\psi(\phi(x,\alpha_i),\delta)$ is robust for $g(\cdot;\varepsilon)$. Since ϕ is resolvable by ψ, i.e., for any $\alpha \in \mathcal{S}$, there exists an α_i and $\delta \in \Delta^*$ such

that $\phi(x, \alpha) = \psi(\phi(x, \alpha_i), \delta)$, we can assert that for any $\alpha \in \mathcal{S}$, the output of $g(\cdot; \varepsilon)$ on $\phi(x, \alpha)$ is robust.

The key of using this theorem for a specific transformation is to choose the resolvable transformation ψ that can enable a tight calculation of Δ^* under a specific way of sampling $\{\alpha_i\}_{i=1}^N$. First, we observe that a large family of transformations including rotation and scaling can be resolved by the additive transformation $\psi\colon \mathcal{X} \times \mathcal{X} \to \mathcal{X}$ defined by $(x, \delta) \mapsto x + \delta$. Indeed, any transformation whose pixel value changes are continuous (or with finite discontinuities) with respect to the parameter changes are differentially resolvable—they all can be resolved by the additive transformation. Choosing isotropic Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_d)$ as smoothing noise then leads to the condition that the maximum ℓ_2-interpolation error between the interval $\mathcal{S} = [a, b]$ (which is to be certified) and the sampled parameters α_i must be bounded by a radius R. This result is shown in the next corollary, which is derived from Theorem 2.

Corollary 2. *Let $\psi(x, \delta) = x + \delta$ and let $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_d)$. Furthermore, let ϕ be a transformation with parameters in $\mathcal{Z}_\phi \subseteq \mathbb{R}^m$ and let $\mathcal{S} \subseteq \mathcal{Z}_\phi$ and $\{\alpha_i\}_{i=1}^N \subseteq \mathcal{S}$. Let $y_A \in \mathcal{Y}$ and suppose that for any i, the ε-smoothed classifier defined by $q(y|\, x; \varepsilon) := \mathbb{E}(p(y|\, x + \varepsilon))$ has class probabilities that satisfy*

$$q(y_A|\, \phi(x, \alpha_i);\, \varepsilon) \geq p_A^{(i)} \geq p_B^{(i)} \geq \max_{y \neq y_A} q(y|\, \phi(x, \alpha_i);\, \varepsilon). \quad (18)$$

Then it is guaranteed that $\forall \alpha \in \mathcal{S}\colon\ y_A = \arg\max_y q(y|\, \phi(x, \alpha); \varepsilon)$ if the maximum interpolation error

$$M_{\mathcal{S}} := \max_{\alpha \in \mathcal{S}} \min_{1 \leq i \leq N} \|\phi(x, \alpha) - \phi(x, \alpha_i)\|_2 \quad (19)$$

$$\textit{satisfies} \quad M_{\mathcal{S}} < R := \frac{\sigma}{2} \min_{1 \leq i \leq N} \left(\Phi^{-1}\left(p_A^{(i)}\right) - \Phi^{-1}\left(p_B^{(i)}\right) \right). \quad (20)$$

In a nutshell, this corollary shows that if the smoothed classifier classifies all samples of transformed inputs $\{\phi(x, \alpha_i)\}_{i=1}^N$ consistent with the original input and the smallest gap between confidence levels $p_A^{(i)}$ and $p_B^{(i)}$ is large enough, then it is guaranteed to make the same prediction on transformed inputs $\phi(x, \alpha)$ for *any* $\alpha \in \mathcal{S}$.

The main challenge now lies in computing a tight and scalable upper bound $M \geq M_{\mathcal{S}}$. Given this bound, a set of transformation parameters $\mathcal{S}$ can then be certified by computing R in (20) and checking that $R > M_{\mathcal{S}}$. With this methodology, we address challenge **(C2)** and provide means to certify transformations that incur interpolation errors. Figure 5a illustrates this methodology on a high level for the rotation transformation as an example. In the following, we present the general methodology that provides an upper bound of the interpolation error $M_{\mathcal{S}}$ and provide closed form expressions for rotation and scaling. In Appendix B, we further extend this methodology to certify transformation compositions such as rotation + brightness change + ℓ_2 perturbations.

We remark that dealing with the interpolation error has already been tried before [2, 13]. However, these approaches either leverage explicit linear or interval bound propagation – techniques that are either not scalable or not tight enough. Therefore, on large datasets such as ImageNet, they can provide only limited certification (e.g., against certain random attack instead of any attack).

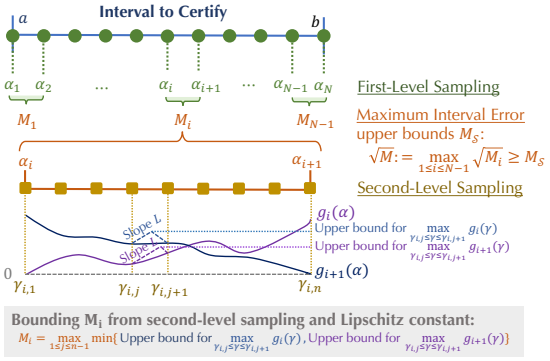


Figure 6: An overview of our interpolation error bounding technique based on stratified sampling and Lipschitz computation.

6.2 Upper Bounding the Interpolation Error

Here, we present the general methodology to compute a rigorous upper bound of the interpolation error introduced in Corollary 2. The methodology presented here is based on stratified sampling and is of a general nature; an explicit computation is shown for the case of rotation and scaling toward the end of this subsection.

Let $\mathcal{S} = [a, b]$ be an interval of transformation parameters that we wish to certify and let $\{\alpha_i\}_{i=1}^N$ be parameters dividing $\mathcal{S}$ uniformly, i.e.,

$$\alpha_i = a + (b - a) \cdot \frac{i - 1}{N - 1}, \quad i = 1, \ldots, N. \quad (21)$$

The set of these parameters corresponds to the first-level samples in stratified sampling. With respect to these first-level samples, we define the functions $g_i\colon [a, b] \to \mathbb{R}_{\geq 0}$ as

$$\alpha \mapsto g_i(\alpha) := \|\phi(x, \alpha) - \phi(x, \alpha_i)\|_2^2 \quad (22)$$

corresponding to the squared ℓ_2 interpolation error between the image x transformed with α and α_i, respectively. For each first-level interval $[\alpha_i, \alpha_{i+1}]$ we look for an upper bound M_i such that

$$M_i \geq \max_{\alpha_i \leq \alpha \leq \alpha_{i+1}} \min\{g_i(\alpha), g_{i+1}(\alpha)\}. \quad (23)$$

It is easy to see that $\max_{1 \leq i \leq N-1} M_i \geq M_{\mathcal{S}}^2$ and hence setting

$$\sqrt{M} := \max_{1 \leq i \leq N-1} \sqrt{M_i} \quad (24)$$

is a valid upper bound to $M_{\mathcal{S}}$. The problem has thus reduced to computing the upper bounds M_i associated with each first-level interval $[\alpha_i, \alpha_{i+1}]$. To that end, we now continue with a second-level sampling within the interval $[\alpha_i, \alpha_{i+1}]$ for each i. Namely, let $\{\gamma_{i,j}\}_{j=1}^n$ be parameters dividing $[\alpha_i, \alpha_{i+1}]$ uniformly, i.e.,

$$\gamma_{i,j} = \alpha_i + (\alpha_{i+1} - \alpha_i) \cdot \frac{j - 1}{n - 1}, \quad j = 1, \ldots, n. \quad (25)$$

Now, suppose that L is a global Lipschitz constant for all functions $\{g_i\}_{i=1}^N$. By definition, for any $1 \leq i \leq N - 1$, L satisfies

$$L \geq \max\left\{ \max_{c,d \in [\alpha_i, \alpha_{i+1}]} \left| \frac{g_i(c) - g_i(d)}{c - d} \right|, \max_{c,d \in [\alpha_i, \alpha_{i+1}]} \left| \frac{g_{i+1}(c) - g_{i+1}(d)}{c - d} \right| \right\}. \quad (26)$$

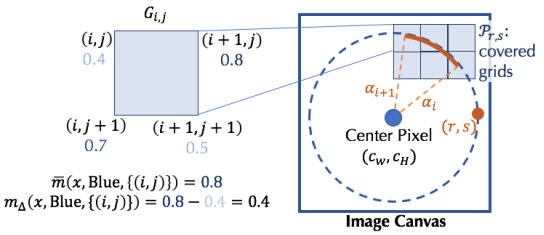


Figure 7: An illustration of Grid Pixel Generator $G_{i,j}$, Color Extractors $\bar{m}$ and m_Δ (take blue channel as example), and the set $\mathcal{P}_{r,s}$.

In the following, we will derive explicit expressions for L for rotation and scaling. Given the Lipschitz constant L, one can show the following closed-form expression for M_i:

$$M_i = \frac{1}{2} \max_{1 \le j \le n-1} \Big(\min \big\{ g_i(\gamma_{i,j}) + g_i(\gamma_{i,j+1}), \\ g_{i+1}(\gamma_{i,j}) + g_{i+1}(\gamma_{i,j+1}) \big\} \Big) + L \cdot \frac{b-a}{(N-1)(n-1)}. \tag{27}$$

An illustration of this bounding technique using stratified sampling is shown in Figure 6. We notice that, as the number N of first-level samples is increased, the interpolation error M_i becomes smaller by shrinking the sampling interval $[\alpha_i, \alpha_{i+1}]$; similarly, increasing the number of second-level samples n makes the upper bound of the interpolation error M_i tighter since the term $L(b-a)/((N-1)(n-1))$ decreases. Furthermore, it is easy to see that as $N \to \infty$ or $n \to \infty$ we have $M \to M_S^2$, i.e., our interpolation error estimation is *asymptotically tight*. Finally, this tendency also highlights an important advantage of our two-level sampling approach: without stratified sampling, it is required to sample $N \times n$ α_i's in order to achieve the same level of approximation accuracy. As a consequence, these $N \times n$ α_i's in turn require to evaluate the smoothed classifier in Corollary 2 $N \times n$ times, compared to just N times in our case.

It thus remains to find a way to efficiently compute the Lipschitz constant L for different transformations. In the following, we derive closed form expressions for rotation and scaling transformations.

6.3 Computing the Lipschitz Constant

Here, we derive a global Lipschitz constant L for the functions $\{g_i\}_{i=1}^N$ defined in (22), for rotation and scaling transformations. In the following, we define K-channel images of width W and height H to be tensors $x \in \mathbb{R}^{K \times W \times H}$ and define the region of valid pixel indices as $\Omega := [0, W-1] \times [0, H-1] \cap \mathbb{N}^2$. Furthermore, for $(r, s) \in \Omega$, we define $d_{r,s}$ to be the ℓ_2-distance to the center of an image, i.e.,

$$d_{r,s} = \sqrt{(r - (W-1)/2)^2 + (s - (H-1)/2)^2}. \tag{28}$$

For ease of notation we make the following definitions that are illustrated in Figure 7.

Definition 4 (Grid Pixel Generator). *For pixels $(i, j) \in \Omega$, we define the grid pixel generator G_{ij} as*

$$G_{ij} := \{(i, j), (i+1, j), (i, j+1), (i+1, j+1)\}. \tag{29}$$

Definition 5 (Max-Color Extractor). *We define the operator that extracts the channel-wise maximum pixel wise on a grid $S \subseteq \Omega$ as the map $\bar{m}\colon \mathbb{R}^{K \times W \times H} \times \{0, \dots, K-1\} \times 2^\Omega \to \mathbb{R}$ with*

$$\bar{m}(x, k, S) := \max_{(i,j) \in S} \big(\max_{(r,s) \in G_{ij}} x_{k,r,s} \big). \tag{30}$$

Definition 6 (Max-Color Difference Extractor). *We define the operator that extracts the channel-wise maximum change in color on a grid $S \subseteq \Omega$ as the map $m_\Delta\colon \mathbb{R}^{K \times W \times H} \times \{0, \dots, K-1\} \times 2^\Omega \to \mathbb{R}$ with*

$$m_\Delta(x, k, S) := \max_{(i,j) \in S} \big(\max_{(r,s) \in G_{ij}} x_{k,r,s} - \min_{(r,s) \in G_{ij}} x_{k,r,s} \big). \tag{31}$$

Rotation. The rotation transformation is defined as rotating an image by an angle α counter-clock wise, followed by bilinear interpolation I. Clearly, when rotating an image, some pixels may be padded that results in a sudden change of pixel colors. To mitigate this issue, we apply black padding to all pixels that are outside the largest centered circle in a given image (see Figure 5a for an illustration). We define the rotation transformation ϕ_R as the (raw) rotation $\tilde{\phi}_R$ followed by interpolation and the aforementioned preprocessing step P so that $\phi_R = P \circ I \circ \tilde{\phi}_R$ and refer the reader to Appendix H for details. We remark that our certification is independent of different rotation padding mechanisms, since these padded pixels are all refilled by black padding during preprocessing. The following lemma provides a closed form expression for L in (27) for rotation. A detailed proof is given in Appendix I.

Lemma 4. *Let $x \in \mathbb{R}^{K \times W \times H}$ be a K-channel image and let $\phi_R = P \circ I \circ \tilde{\phi}_R$ be the rotation transformation. Then, a global Lipschitz constant L for the functions $\{g_i\}_{i=1}^N$ is given by*

$$L_r = \max_{1 \le i \le N-1} \sum_{k=0}^{K-1} \sum_{r,s \in V} 2 d_{r,s} \cdot m_\Delta(x, k, \mathcal{P}_{r,s}^{(i)}) \cdot \bar{m}(x, k, \mathcal{P}_{r,s}^{(i)}) \tag{32}$$

where $V = \big\{(r,s) \in \mathbb{N}^2 \,|\, d_{r,s} < \frac{1}{2}(\min\{W, H\} - 1)\big\}$. The set $\mathcal{P}_{r,s}^{(i)}$ is given by all integer grid pixels that are covered by the trajectory of source pixels of (r, s) when rotating from angle α_i to α_{i+1}.

Scaling. In Appendix A we introduce how to compute the Lipschitz bound for the scaling transformation and provide the certification. The process is similar to that for rotation.

Computational complexity. We provide pseudo-code for computing bound M in Appendix J. The algorithm is composed of two main parts, namely the computation of the Lipschitz constant L, and the computation of the interpolation error bound M based on L. The former is of computational complexity $O(N \cdot KWH)$, and the latter is of $O(NR \cdot KWH)$, for both scaling and rotation. We note that $\mathcal{P}_{r,s}$ contains only a constant number of pixels since each interval $[\alpha_i, \alpha_{i+1}]$ is small. Thus, the bulk of costs come from the transformation operation. We improve the speed by implementing a fast and fully-parallelized `C` kernel for rotation and scaling of images. As a result, on CIFAR-10, the algorithm takes less than 2 s on average with 10 processes for rotation with $N = 556$ and $n = 200$ and the time for scaling is faster. We refer readers to Section 7 for detailed experimental evaluation. Also, we remark that the algorithm is model-independent. Thus, we can precompute M for test set and reuse for any models that need a certification.

6.4 Discussion

Here, we briefly summarize the computation procedure of robustness certification, introduce an acceleration strategy—progressive sampling—and discuss the extensions beyond rotation and scaling.

6.4.1 Computation of Robustness Certification. With the methodology mentioned above, for differentially resolvable transformations such as rotation and scaling, computing robustness certification follows two steps: (1) computing the interpolation error bound M; (2) generate transformed samples $\{\phi(x, \alpha_i)\}_{i=1}^{N}$, compute $p_A^{(i)}$ and $p_B^{(i)}$ for each sample, and check whether $M_{\mathcal{S}} < R$ holds for each sample according to Corollary 2.

6.4.2 Acceleration: Progressive Sampling. In step (2) above, we need to estimate $p_A^{(i)}$ and $p_B^{(i)}$ for each sample $\phi(x, \alpha_i)$ to check whether $M_{\mathcal{S}} < R$. In the brute-force approach, to obtain a high-confidence bound on $p_A^{(i)}$ and $p_B^{(i)}$, we typically sample $n_s = 10,000$ or more [7] then apply the binomial statistical test. In total, we thus need to sample the classifier's prediction $N \times n_s$ times, which is costly.

To accelerate the computation, we design a *progressive sampling strategy* from the following two insights: (1) we only need to check whether $R > M_{\mathcal{S}}$, but are not required to compute R precisely; (2) for any sample $\phi(x, \alpha_i)$ if the check fails, the model is not certifiably robust and there is no need to proceed. Based on (1), for the current $\phi(x, \alpha_i)$, we sample n_s samples in batches and maintain high-confidence lower bound of R based on existing estimation. Once the lower bound exceeds $M_{\mathcal{S}}$ we proceed to the next $\phi(x, \alpha_{i+1})$. Based on (2), we terminate early if the check $R > M_{\mathcal{S}}$ for the current $\phi(x, \alpha_i)$ fails. More details are provided in Appendix J.

6.4.3 Extension to More Transformations. For other transformations that involve interpolation, we can similarly compute the interpolation error bound using intermediate results in our above lemmas. For transformation compositions, we extend our certification pipeline for the composition of (1) rotation/scaling with brightness, and (2) rotation/scaling with brightness and ℓ_p-bounded additive perturbations. These compositions simulate an attacker who does not precisely perform the specified transformation. We present these extensions in Appendix B.3 and Appendix B.4 in detail, and in Appendix B.5 we discuss how to analyze possible new transformations and then extend **TSS** to provide certification.

7 EXPERIMENTS

We validate our framework **TSS** by certifying robustness over semantic transformations experimentally. We compare with state of the art for each transformation, highlight our main results, and present some interesting findings and ablation studies.

7.1 Experimental Setup

7.1.1 Dataset. Our experiments are conducted on three classical image classification datasets: MNIST, CIFAR-10, and ImageNet. For all images, the pixel color is normalized to $[0, 1]$. We follow common practice to resize and center cropping the ImageNet images to 224×224 size [7, 24, 40, 63]. To our best knowledge, we are the *first* to provide rigorous certifiable robustness against semantic transformations on the large-scale *standard* ImageNet dataset.

7.1.2 Model. The undefended model is very vulnerable even under simple random semantic attacks. Therefore, we apply existing data augmentation training [7] combined with consistency regularization [24] to train the base classifiers. We then use the introduced smoothing strategies, to obtain the models for robustness certification. On MNIST and CIFAR-10, the models are trained from scratch while on ImageNet, we either finetune undefended models in `torchvision` library or finetune from state-of-the-art certifiably robust models against ℓ_2 perturbations [42]. Details are available in Appendix K.1. We remark that our framework focuses on robustness certification and did not fully explore the training methods for improving the certified robustness or tune the hyperparameters.

7.1.3 Implementation and Hardware. We implement our framework **TSS** based on `PyTorch`. We improve the running efficiency by tensor parallelism and embedding `C` modules. Details are available in Appendix K.2. All experiments were run on 24-core Intel Xeon Platinum 8259CL CPU and one Tesla T4 GPU with 15 GB RAM.

7.1.4 Evaluation Metric. On each dataset, we uniformly pick 500 samples from the test set and evaluate all results on this *test subset* following Cohen et al [7]. In line with related work [7, 24, 42, 63], we report the **certified robust accuracy** that is defined as the fraction of samples (within the test subset) that are both *certified robust* and *classified correctly*, and set the certification confidence level to $p = 0.1\%$. We use $n_s = 10^5$ samples to obtain a confidence lower bound $\underline{p_A}$ for resolvable transformations, and $n_s = 10^4$ samples to obtain each $\underline{p_A}^{(i)}$ for differentially resolvable transformations. Due to Progressive Sampling (Algorithm 2), the actual samples used for differentially resolvable transformations are usually far fewer than n_s. In addition, we report the **benign accuracy** in Appendix K.5 defined as the fraction of *correctly classified* samples when no attack is present, and the **empirical robust accuracy**, defined as the fraction of samples in the test subset that are classified correctly under either a simple random attack (following [2, 13]) or two adaptive attacks (namely Random+ Attack and PGD Attack). We introduce all these attacks in Appendix K.3 and provide a detailed comparison in Appendix K.7. Note that the empirical robust accuracy under any attacks is lower bounded by the certified accuracy.

7.1.5 Notations for Robust Radii. In the tables, we use these notations: α for squared kernel radius for Gaussian blur; $\sqrt{\Delta x^2 + \Delta y^2}$ for translation distance; b and c for brightness shift and contrast change respectively as in $x \mapsto (1+c)x+b$; r for rotation angle; s for size scaling ratio; and $\|\delta\|_2$ for ℓ_2 norm of additional perturbations.

7.1.6 Vanilla Models and Baselines. We compare with vanilla (undefended) models and baselines from related work. The vanilla models are trained to achieve high accuracy only on clean data. For fairness, on all datasets we use the same model architectures as in our approach. On the test subset, the *benign accuracy* of vanilla models is 98.6%/88.6%/74.4% on MNIST/CIFAR-10/ImageNet. We also report their empirical robust accuracy under attacks in Table 2. Since vanilla models are not smoothed, we cannot have certified robust accuracy for them. In terms of baselines, we consider the approaches that provide certification against semantic transformations: DeepG [2], Interval [47], VeriVis [39], Semanify-NN [35], and DistSPT [13]. In Appendix K.4, we provide more detailed discussion and comparison with these baseline approaches, and list how we run these approaches for fair comparison.

Table 2: Comparison of certified robust accuracy achieved by our framework **TSS** and other known baselines and empirical robust accuracy achieved by **TSS** and vanilla models under random and adaptive attacks. "-" denotes the settings where the baselines cannot support. The parentheses show the weaker baseline settings. For certified robust accuracy, the existing state of the art is **bolded**. For empirical robust accuracy, the higher accuracy under each setting are **bolded**.

Transformation	Type	Dataset	Attack Radius	Certified Robust Accuracy						Empirical Robust Accuracy			
										Random Attack		Adaptive Attacks	
				TSS	DeepG [2]	Interval [47]	VeriVis [39]	Semanify-NN [35]	DistSPT [13]	**TSS**	Vanilla	**TSS**	Vanilla
Gaussian Blur	Resolvable	MNIST	Squared Radius $\alpha \leq 36$	**90.6%**	-	-	-	-	-	**91.4%**	12.2%	**91.2%**	12.2%
		CIFAR-10	Squared Radius $\alpha \leq 16$	**63.6%**	-	-	-	-	-	**65.8%**	3.4%	**65.8%**	3.4%
		ImageNet	Squared Radius $\alpha \leq 36$	**51.6%**	-	-	-	-	-	**52.8%**	8.4%	**52.6%**	8.2%
Translation (Reflection Pad.)	Resolvable, Discrete	MNIST	$\sqrt{\Delta x^2 + \Delta y^2} \leq 8$	**99.6%**	-	-	98.8%	98.8%	-	**99.6%**	0.0%	**99.6%**	0.0%
		CIFAR-10	$\sqrt{\Delta x^2 + \Delta y^2} \leq 20$	**80.8%**	-	-	65.0%	65.0%	-	**86.2%**	4.4%	**86.0%**	4.2%
		ImageNet	$\sqrt{\Delta x^2 + \Delta y^2} \leq 100$	**50.0%**	-	-	43.2%	43.2%	-	**69.2%**	46.6%	**69.2%**	46.2%
Brightness	Resolvable	MNIST	$b \pm 50\%$	**98.2%**	-	-	-	-	-	**98.2%**	96.6%	**98.2%**	96.6%
		CIFAR-10	$b \pm 40\%$	**87.0%**	-	-	-	-	-	**87.2%**	44.4%	**87.4%**	42.6%
		ImageNet	$b \pm 40\%$	**70.0%**	-	-	-	-	-	**70.4%**	19.6%	**70.4%**	18.4%
Contrast and Brightness	Resolvable, Composition	MNIST	$c \pm 50\%, b \pm 50\%$	**97.6%**	$\leq 0.4\%$ $(c, b \pm 30\%)$	0.0% $(c, b \pm 30\%)$	-	$\leq 74\%$ $(c \pm 5\%, b \pm 50\%)$	-	**98.0%**	94.6%	**98.0%**	93.2%
		CIFAR-10	$c \pm 40\%, b \pm 40\%$	**82.4%**	0.0% $(c, b \pm 30\%)$	0.0% $(c, b \pm 30\%)$	-	-	-	**86.0%**	21.0%	**85.8%**	9.6%
		ImageNet	$c \pm 40\%, b \pm 40\%$	**61.4%**	-	-	-	-	-	**68.4%**	1.2%	**68.4%**	0.0%
Gaussian Blur, Translation, Brightness, and Contrast	Resolvable, Composition	MNIST	$\alpha \leq 1, \sqrt{\Delta x^2 + \Delta y^2} \leq 5, c, b \pm 10\%$	**90.2%**	-	-	-	-	-	**97.2%**	0.4%	**97.0%**	0.4%
		CIFAR-10	$\alpha \leq 1, \sqrt{\Delta x^2 + \Delta y^2} \leq 5, c, b \pm 10\%$	**58.2%**	-	-	-	-	-	**67.6%**	9.6%	**67.8%**	5.6%
		ImageNet	$\alpha \leq 10, \sqrt{\Delta x^2 + \Delta y^2} \leq 10, c, b \pm 20\%$	**32.8%**	-	-	-	-	-	**48.8%**	9.4%	**47.4%**	4.0%
Rotation	Differentially Resolvable	MNIST	$r \pm 50°$	**97.4%**	$\leq 85.8\%$ $(r \pm 30°)$	$\leq 6.0\%$ $(r \pm 30°)$	-	$\leq 92.48\%$	82%	**98.4%**	12.2%	**98.2%**	11.0%
		CIFAR-10	$r \pm 10°$	**70.6%**	62.5%	20.2%	-	-	37%	**76.6%**	65.6%	**76.4%**	65.4%
			$r \pm 30°$	**63.6%**	10.6%	0.0%	-	$\leq 49.37\%$	22%	**69.2%**	21.6%	**69.4%**	21.4%
		ImageNet	$r \pm 30°$	**30.4%**	-	-	-	-	16% (rand. attack)	37.8%	**40.0%**	**37.8%**	37.0%
Scaling	Differentially Resolvable	MNIST	$s \pm 30\%$	**97.2%**	85.0%	16.4%	-	-	-	**99.2%**	90.2%	**99.2%**	89.2%
		CIFAR-10	$s \pm 30\%$	**58.8%**	0.0%	0.0%	-	-	-	**67.2%**	51.6%	**67.0%**	51.2%
		ImageNet	$s \pm 30\%$	**26.4%**	-	-	-	-	-	37.4%	**50.0%**	36.4%	**49.8%**
Rotation and Brightness	Differentially Resolvable, Composition	MNIST	$r \pm 50°, b \pm 20\%$	**97.0%**	-	-	-	-	-	**98.2%**	11.0%	**98.0%**	10.4%
		CIFAR-10	$r \pm 10°, b \pm 10\%$	**70.2%**	-	-	-	-	-	**76.6%**	59.4%	**76.0%**	56.8%
			$r \pm 30°, b \pm 20\%$	**61.4%**	-	-	-	-	-	**68.4%**	13.0%	**68.2%**	9.0%
		ImageNet	$r \pm 30°, b \pm 20\%$	**26.8%**	-	-	-	-	-	**37.4%**	22.4%	**36.8%**	21.2%
Scaling and Brightness	Differentially Resolvable, Composition	MNIST	$s \pm 50\%, b \pm 50\%$	**96.6%**	-	-	-	-	-	**97.8%**	24.8%	**97.8%**	15.6%
		CIFAR-10	$s \pm 30\%, b \pm 30\%$	**54.2%**	-	-	-	-	-	**67.2%**	17.4%	**66.8%**	11.6%
		ImageNet	$s \pm 30\%, b \pm 30\%$	**23.4%**	-	-	-	-	-	**36.4%**	16.0%	**36.0%**	8.8%
Rotation, Brightness, and ℓ_2	Differentially Resolvable, Composition	MNIST	$r \pm 50°, b \pm 20\%, \|\delta\|_2 \leq .05$	**96.6%**	-	-	-	-	-	**97.6%**	10.8%	**97.4%**	9.0%
		CIFAR-10	$r \pm 10°, b \pm 10\%, \|\delta\|_2 \leq .05$	**64.2%**	-	-	-	-	-	**71.6%**	31.8%	**71.2%**	29.6%
			$r \pm 30°, b \pm 20\%, \|\delta\|_2 \leq .05$	**55.2%**	-	-	-	-	-	**65.2%**	0.8%	**64.0%**	0.4%
		ImageNet	$r \pm 30°, b \pm 20\%, \|\delta\|_2 \leq .05$	**26.6%**	-	-	-	-	-	**37.0%**	17.6%	**36.4%**	14.0%
Scaling, Brightness, and ℓ_2	Differentially Resolvable, Composition	MNIST	$s \pm 50\%, b \pm 50\%, \|\delta\|_2 \leq .05$	**96.4%**	-	-	-	-	-	**97.6%**	22.2%	**97.6%**	12.2%
		CIFAR-10	$s \pm 30\%, b \pm 30\%, \|\delta\|_2 \leq .05$	**51.2%**	-	-	-	-	-	**65.0%**	4.4%	**61.8%**	2.6%
		ImageNet	$s \pm 30\%, b \pm 30\%, \|\delta\|_2 \leq .05$	**22.6%**	-	-	-	-	-	**36.0%**	7.4%	**35.6%**	4.8%

7.2 Main Results

Here, we present our main results from five aspects: (1) certified robustness compared to baselines; (2) empirical robustness comparison; (3) certification time statistics; (4) empirical robustness under unforeseen physical attacks; (5) certified robustness under attacks exceeding the certified radii.

7.2.1 Certified Robustness Compared to Baselines. Our results are summarized in Table 2. For each transformation, we ensure that our setting is either the same as or strictly stronger than all other baselines. When our setting is strictly stronger, the baseline setting is shown in corresponding parentheses, and our certified robust accuracy implies a higher or equal certified robust accuracy in the corresponding baseline setting. To our best knowledge, we are the first to provide certified robustness for Gaussian blur, brightness, composition of rotation and brightness, etc. Moreover, on the large-scale standard ImageNet dataset, we are the first to provide nontrivial certified robustness against certain semantic attacks. Note that DistSPT [13] is theoretically feasible to provide robustness certification for the ImageNet dataset. However, its certification is not tight enough to handle ImageNet and it provides robustness certification for only a certain random attack instead of arbitrarily worst-case attacks [13, Section 7.4]. We observe that, across transformations, our framework *significantly* outperforms the state of the art, if present, in terms of robust accuracy. For example, on the composition of contrast and brightness, we improve the certified robust accuracy from 74% to 97.6% on MNIST, from 0.0% (failing to certify) to 82.4% on CIFAR-10, and from 0% (absence of baseline) to 61.4% on ImageNet. On the rotation transformation, we improve the certified robust accuracy from 92.48% to 97.4% on MNIST, from 49.37% to 63.6% on CIFAR-10 (rotation angle within 30°), and from 16% against a certain random attack to 30.4% against arbitrary attacks on ImageNet. Some baselines are able to provide certification under other certification goals and the readers can refer to Appendix K.4 for a detailed discussion.

7.2.2 Comparison of Empirical Robust Accuracy. In Table 2, we report the empirical robust accuracy for both (undefended) vanilla models and trained **TSS** models. The empirical robust accuracy is either evaluated under random attack or two adaptive attacks–Random+ and PGD attack. When it is under adaptive attacks, we report the lower accuracy to evaluate against stronger attackers.

(1) For almost all settings, **TSS** models have significantly higher *empirical robust accuracy*, which means that **TSS** models are also practical in terms of defending against existing attacks. The only exception is rotation and scaling on ImageNet. The

reason is that a single rotation/scaling transformation is too weak to attack even an undefended model. At the same time, our robustness certification comes at the cost of benign accuracy, which also affects the empirical robust accuracy. This exception is eliminated when rotation and scaling are composed with other transformations.

(2) Similar observations arise when comparing the *empirical robust accuracy of the vanilla model with the certified robust accuracy of ours*. Hence, even compared to *empirical* metrics, our *certified* robust accuracy is nontrivial and guarantees high accuracy.

(3) Our *certified* robust accuracy is always lower or equal compared to the *empirical* one, verifying the validity of our robustness certification. The gaps range from ∼ 2% on MNIST to ∼ 10% - 20% on ImageNet. Since empirical robust accuracy is an upper bound of the certified accuracy, this implies that our certified bounds are usually tight, particularly on small datasets.

(4) The adaptive attack decreases the empirical accuracy of **TSS** models *slightly*, while it decreases that of vanilla models significantly. Taking contrast and brightness on CIFAR-10 as example, **TSS** accuracy decreases from 86% to 85.8% while the vanilla model accuracy decreases from 21.0% to 9.6%. Thus, **TSS** is still robust against adaptive attacks. Indeed, **TSS** has robustness guarantee against any attack within the certified radius.

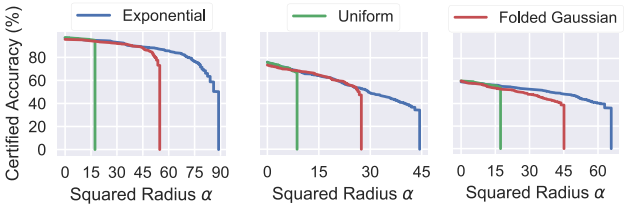
7.2.3 Certification Time Statistics. Our robustness certification time is usually less than 100 s on MNIST and 200 s on CIFAR-10; on ImageNet it is around 200 s - 2000 s. Compared to other baselines, ours is slightly faster and achieves much higher certified robustness. For fairness, we give 1000 s time limit per instance when running baselines on MNIST and CIFAR-10. Note that other baselines cannot scale up to ImageNet. Our approach is scalable due to the black-box nature of smoothing-based certification, the tight interpolation error upper bound, and the efficient progressive sampling strategy. Details on hyperparameters including smoothing variance and average certification time are given in Appendix K.6.

*7.2.4 Generalization to Unforeseen Common Corruptions. Are **TSS** models still more robust when it comes to potential unforeseen physical attacks?* To answer this question, we evaluate the robustness of **TSS** models on the realistic CIFAR-10-C and ImageNet-C datasets [20]. These two datasets are comprised of corrupted images from CIFAR-10 and ImageNet. They apply around 20 types of common corruptions to model *physical attacks*, such as fog, snow, and frost. We evaluate the *empirical robust accuracy* against the highest corruption level (level 5) to model the strongest physical attacker. We apply **TSS** models trained against a transformation composition attack, Gaussian blur + brightness + contrast + translation, to defend against these corruptions. We select two baselines: vanilla models and AugMix [21]. AugMix is the state of the art model on CIFAR-10-C and ImageNet-C [8].

The results are shown in Table 3. The answer is *yes*—**TSS** models are more robust than undefended vanilla models. It even exceeds the state of the art, AugMix, on CIFAR-10-C. On ImageNet-C, **TSS** model's empirical accuracy is between vanilla and AugMix. We emphasize that in contrast to **TSS**, both vanilla and AugMix fail to provide robustness certification. Details on evaluation protocols and additional findings are in Appendix K.8.

Table 3: Comparison of **empirical accuracy** of different models under physical corruptions (CIFAR-10-C and ImageNet-C) and **certified accuracy** against composition of transformations. **TSS** achieves higher or comparable empirical accuracy against unforeseen corruptions and significantly higher certified accuracy (under attack radii in Table 2).

	CIFAR-10			ImageNet		
	Vanilla	AugMix [21]	**TSS**	Vanilla	AugMix [21]	**TSS**
Empirical Accuracy on CIFAR-10-C and ImageNet-C	53.9%	65.6%	**67.4%**	18.3%	**25.7%**	21.9%
Certified Accuracy against Composition of Gaussian Blur, Translation, Brightness, and Contrast	0.0%	0.4%	**58.2%**	0.0%	0.0%	**32.8%**



(a) MNIST **(b) CIFAR-10** **(c) ImageNet**

Figure 8: Certified accuracy for different smoothing distributions for Gaussian blur. On MNIST/CIFAR-10/ImageNet the noise std. is 10/5/10.

Table 4: Study of the impact of different smoothing variance levels on certified robust accuracy and benign accuracy on **ImageNet** for **TSS**. The attack radii are consistent with Table 2. "Dist." refers to both training and smoothing distribution. The variance used in Table 2 is labeled in gray.

Transformation	Attack Radii	Certified Accuracy and Benign Accuracy under Different Variance Levels			
Gaussian Blur	$\alpha \le 36$	Dist. of α	Exp(1/5)	Exp(1/10)	Exp(1/20)
		Cert. Rob. Acc.	0.0%	**51.6%**	48.4%
		Benign Acc.	**63.4%**	59.2%	53.2%
Translation (Reflection Pad.)	$\sqrt{\Delta x^2 + \Delta y^2} \le 100$	Dist. of $(\Delta x, \Delta y)$	$\mathcal{N}(0, 20^2 I)$	$\mathcal{N}(0, 30^2 I)$	$\mathcal{N}(0, 40^2 I)$
		Cert. Rob. Acc.	0.0%	50.0%	**55.4%**
		Benign Acc.	70.0%	**72.6%**	70.0%
Brightness	$b \pm 40\%$	Dist. of (c, b)	$\mathcal{N}(0, 0.3^2 I)$	$\mathcal{N}(0, 0.4^2 I)$	$\mathcal{N}(0, 0.5^2 I)$
		Cert. Rob. Acc.	**70.2%**	70.0%	67.6%
		Benign Acc.	**73.2%**	72.2%	69.4%
Contrast	$c \pm 40\%$	Dist. of (c, b)	$\mathcal{N}(0, 0.3^2 I)$	$\mathcal{N}(0, 0.4^2 I)$	$\mathcal{N}(0, 0.5^2 I)$
		Cert. Rob. Acc.	58.4%	63.6%	**65.0%**
		Benign Acc.	**72.8%**	71.4%	68.6%
Rotation	$r \pm 30°$	Dist. of ε	$\mathcal{N}(0, 0.25^2 I)$	$\mathcal{N}(0, 0.50^2 I)$	$\mathcal{N}(0, 1.00^2 I)$
		Cert. Rob. Acc.	9.8%	**30.4%**	20.0%
		Benign Acc.	**55.6%**	46.2%	32.2%
Scaling	$s \pm 30\%$	Dist. of ε	$\mathcal{N}(0, 0.25^2 I)$	$\mathcal{N}(0, 0.50^2 I)$	$\mathcal{N}(0, 1.00^2 I)$
		Cert. Rob. Acc.	7.2%	**26.4%**	17.4%
		Benign Acc.	**58.8%**	50.8%	33.8%

7.2.5 Evaluation on Attacks Beyond Certified Radii. The semantic attacker in the physical world may not constrain itself to be within the specified attack radii. In Appendix K.9 we present a thorough evaluation of **TSS**'s robustness when the attack radii go beyond the certified ones. We show, for example, for **TSS** model defending against ±40% brightness change on ImageNet, when the radius increases to 50%, the certified accuracy only slightly drops from 70.4% to 70.0%. In a nutshell, there is no significant or immediate degradation on both certified robust accuracy and empirical robust accuracy when the attack radii go beyond the certified ones.

7.3 Ablation Studies

Here, we provide two ablation studies: (1) Comparison of different smoothing distributions; (2) Comparison of different smoothing variances. In Appendix K.11, we present another ablation study on

different numbers of samples for differentially resolvable transformations, which reveals a tightness-efficiency trade-off.

7.3.1 Comparison of Smoothing Distributions. To study the effects of different smoothing distributions, we compare the certified robust accuracy for Gaussian blur when the model is smoothed by different smoothing distributions. We consider three smoothing distributions, namely exponential (blue line), uniform (green line), and folded Gaussian (red line). On each dataset, we adjust the distribution parameters such that each distribution has the same variance. All other hyperparameters are kept the same throughout training and certification. As shown Figure 8, we notice that on all three datasets, the exponential distribution has the highest average certified radius. This observation is in line with our theoretical reasoning in Section 5.2.

7.3.2 Comparison of Different Smoothing Variances. The variance of the smoothing distribution is a hyperparameter that controls the accuracy-robustness trade-off. In Table 4, we evaluate different smoothing variances for several transformations on ImageNet and report both the certified accuracy and benign accuracy. The results on MNIST and CIFAR-10 and more discussions are in Appendix K.10. From these results, we observe that usually, when the smoothing variance increases, the benign accuracy drops and the certified robust accuracy first rises and then drops. This tendency is also observed in classical randomized smoothing [7, 63]. However, the range of acceptable variance is usually wide. Thus, without careful tuning the smoothing variances, we are able to achieve high certified and benign accuracy as reported in Table 2 and Table 6.

8 RELATED WORK

Certified Robustness against ℓ_p perturbations. Since the studies of adversarial vulnerability of neural networks [16, 48], there has emerged a rich body of research on evasion attacks (e.g., [1, 5, 53, 59]) and empirical defenses (e.g., [32, 44, 45]). To provide robustness certification, different robustness training and verification approaches have been proposed. In particular, interval bound propagation [17, 66], linear relaxations [33, 56–58, 61], and semidefinite programming [9, 41] have been applied to certify NN robustness. Recently, robustness certification based on randomized smoothing has shown to be scalable and with tight guarantees [7, 26, 27]. With improvements on optimizing the smoothing distribution [10, 49, 63] and better training mechanisms [6, 24, 42, 64], the verified robustness of randomized smoothing is further improved. A recent survey summarizes certified robustness approaches [28].

Semantic Attacks for Neural Networks. Recent work has shown that semantic transformations are able to mislead ML models [14, 22, 60]. For instance, image rotations and translations can attack ML models with 40% - 99% degradation on MNIST, CIFAR-10, and ImageNet on both vanilla models and models that are robust against ℓ_p-bounded perturbations [11]. Brightness/contrast attacks can achieve 91.6% attack success on CIFAR-10, and 71%-100% attack success rate on ImageNet [20]. Our evaluation on empirical robust accuracy (Table 2) for vanilla models also confirms these observations. Moreover, brightness attacks have been shown to be of practical concern in autonomous driving [38]. Empirical defenses against semantic transformations have been investigated [11, 20].

Certified Robustness against Semantic Transformations. While heuristic defenses against semantic attacks have been proposed, *provable* robustness requires further investigation. Existing certified robustness against transformations is based on heuristic enumeration, interval bound propagation, linear relaxation, or smoothing. Efficient enumeration in VeriVis [39] can handle only discrete transformations. Interval bound propagation has been used to certify common semantic transformations [2, 13, 47]. To tighten the interval bounds, linear relaxations are introduced. DeepG [2] optimizes linear relaxations for given semantic transformations, and Semantify-NN [35] encodes semantic transformations by neural networks and applies linear relaxations for NNs [56, 66]. However, linear relaxations are loose and computationally intensive compared to our **TSS**. Recently, Fischer et al [13] have applied a smoothing scheme to provide provable robustness against transformations but on the large ImageNet dataset, it can provide certification only against random attacks that draw transformation parameters from a pre-determined distribution. More details are available in Appendix K.4.

9 CONCLUSION

In this paper, we have presented a unified framework, **TSS**, for certifying ML robustness against general semantic adversarial transformations. Extensive experiments have shown that **TSS** significantly outperforms the state of the art or, if no previous work exists, set new baselines. In future work, we plan to further improve the efficiency and tightness of our robustness certification and explore more transformation-specific smoothing strategies.

ACKNOWLEDGMENTS

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, Lawrence Livermore National Security, LLC. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes. This work was supported by LLNL Laboratory Directed Research and Development project 20-ER-014 and released with LLNL tracking number LLNL-CONF-822465. This work is supported in part by NSF under grant no. CNS-2046726, CCF-1910100, CCF-1816615, and 2020 Amazon Research Award.

CZ and the DS3Lab gratefully acknowledge the support from the Swiss National Science Foundation (Project Number 200021_184628, and 197485), Innosuisse/SNF BRIDGE Discovery (Project Number 40B2-0_187132), European Union Horizon 2020 Research and Innovation Programme (DAPHNE, 957407), Botnar Research Centre for Child Health, Swiss Data Science Center, Alibaba, Cisco, eBay, Google Focused Research Awards, Kuaishou Inc., Oracle Labs, Zurich Insurance, and the Department of Computer Science at ETH Zurich.

The authors thank the anonymous reviewers for valuable feedback, Adel Bibi (University of Oxford) for pointing out a bug in the initial implementation, and the authors of [13], especially Marc Fischer and Martin Vechev, for insightful discussions and support on experimental evaluation.

REFERENCES

[1] Anish Athalye, Nicholas Carlini, and David Wagner. 2018. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *2018 International Conference on Machine Learning (ICML)*. PMLR, 274–283.

[2] Mislav Balunovic, Maximilian Baader, Gagandeep Singh, Timon Gehr, and Martin Vechev. 2019. Certifying Geometric Robustness of Neural Networks. In *2019 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 15287–15297.

[3] Avrim Blum, Travis Dick, Naren Manoj, and Hongyang Zhang. 2020. Random Smoothing Might be Unable to Certify ℓ_∞ Robustness for High-Dimensional Images. *Journal of Machine Learning Research (JMLR)* 21, 211 (2020), 1–21.

[4] Saikiran Bulusu, Bhavya Kailkhura, Bo Li, Pramod K Varshney, and Dawn Song. 2020. Anomalous Example Detection in Deep Learning: A Survey. *IEEE Access* 8 (2020), 132330–132347.

[5] Nicholas Carlini and David Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 39–57.

[6] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. 2019. Unlabeled Data Improves Adversarial Robustness. In *2019 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 11192–11203.

[7] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified Adversarial Robustness via Randomized Smoothing. In *2019 International Conference on Machine Learning (ICML)*. PMLR, 1310–1320.

[8] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. 2020. RobustBench: A Standardized Adversarial Robustness Benchmark. *arXiv preprint arXiv:2010.09670* 10 (2020).

[9] Sumanth Dathathri, Krishnamurthy Dvijotham, Alexey Kurakin, Aditi Raghunathan, Jonathan Uesato, Rudy Bunel, Shreya Shankar, Jacob Steinhardt, Ian Goodfellow, Percy Liang, et al. 2020. Enabling Certification of Verification-Agnostic Networks via Memory-Efficient Semidefinite Programming. In *2020 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 5318–5331.

[10] Krishnamurthy Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Kai Xiao, Sven Gowal, and Pushmeet Kohli. 2020. A Framework for Robustness Certification of Smoothed Classifiers using f-Divergences. In *2020 International Conference on Learning Representations (ICLR)*. OpenReview.

[11] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. 2019. Exploring the Landscape of Spatial Robustness. In *2019 International Conference on Machine Learning (ICML)*. PMLR, 1802–1811.

[12] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust Physical-World Attacks on Deep Learning Visual Classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1625–1634.

[13] Marc Fischer, Maximilian Baader, and Martin Vechev. 2020. Certified Defense to Image Transformations via Randomized Smoothing. In *2020 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 8404–8417.

[14] Amin Ghiasi, Ali Shafahi, and Tom Goldstein. 2020. Breaking Certified Defenses: Semantic Adversarial Examples with Spoofed Robustness Certificates. In *2020 International Conference on Learning Representations (ICLR)*. OpenReview.

[15] Tejas Gokhale, Rushil Anirudh, Bhavya Kailkhura, Jayaraman J Thiagarajan, Chitta Baral, and Yezhou Yang. 2021. Attribute-Guided Adversarial Training for Robustness to Natural Perturbations. In *2021 AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 35. Association for the Advancement of Artificial Intelligence Press, 7574–7582.

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *2015 International Conference on Learning Representations (ICLR)*. OpenReview.

[17] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. 2019. Scalable Verified Training for Provably Robust Image Classification. In *2019 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 4842–4851.

[18] Jamie Hayes. 2020. Extensions and Limitations of Randomized Smoothing for Robustness Guarantees. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 3413–3421.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 1026–1034.

[20] Dan Hendrycks and Thomas Dietterich. 2018. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *2018 International Conference on Learning Representations (ICLR)*. OpenReview.

[21] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2020. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *2020 International Conference on Learning Representations (ICLR)*. OpenReview.

[22] Hossein Hosseini and Radha Poovendran. 2018. Semantic Adversarial Examples. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 1614–1619.

[23] Weiwei Hu and Ying Tan. 2017. Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN. *arXiv preprint arXiv:1702.05983* 02 (2017).

[24] Jongheon Jeong and Jinwoo Shin. 2020. Consistency Regularization for Certified Robustness of Smoothed Classifiers. In *2020 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 10558–10570.

[25] Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. 2020. Curse of Dimensionality on Randomized Smoothing for Certifiable Robustness. In *2020 International Conference on Machine Learning (ICML)*. PMLR, 5458–5467.

[26] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified Robustness to Adversarial Examples with Differential Privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 656–672.

[27] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. 2019. Certified Adversarial Robustness with Additive Noise. In *2019 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 9459–9469.

[28] Linyi Li, Xiangyu Qi, Tao Xie, and Bo Li. 2020. SoK: Certified Robustness for Deep Neural Networks. *arXiv preprint arXiv:2009.04131* 09 (2020).

[29] Linyi Li, Maurice Weber, Xiaojun Xu, Luka Rimanic, Tao Xie, Ce Zhang, and Bo Li. 2020. TSS: Transformation-Specific Smoothing for Robustness Certification. *arXiv preprint arXiv:2002.12398* 02 (2020). https://arxiv.org/abs/2002.12398

[30] Linyi Li, Zexuan Zhong, Bo Li, and Tao Xie. 2019. Robustra: Training Provable Robust Neural Networks over Reference Adversarial Space. In *2019 International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, 4711–4717.

[31] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. 2018. Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality. In *2018 International Conference on Learning Representations (ICLR)*. OpenReview.

[32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *2018 International Conference on Learning Representations (ICLR)*. OpenReview.

[33] Matthew Mirman, Timon Gehr, and Martin Vechev. 2018. Differentiable Abstract Interpretation for Provably Robust Neural Networks. In *2018 International Conference on Machine Learning (ICML)*. PMLR, 3575–3583.

[34] Jeet Mohapatra, Ching-Yun Ko, Lily Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. 2021. Hidden Cost of Randomized Smoothing. In *2021 International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 4033–4041.

[35] Jeet Mohapatra, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. 2020. Towards Verifying Robustness of Neural Networks Against A Family of Semantic Perturbations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 244–252.

[36] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. 2017. DeepStack: Expert-Level Artificial Intelligence in Heads-Up No-Limit Poker. *Science* 356, 6337 (2017), 508–513.

[37] OpenCV. 2020. OpenCV: Transformations of Images. https://docs.opencv.org/master/dd/d52/tutorial_js_geometric_transformations.html.

[38] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. 2017. DeepXplore: Automated Whitebox Testing of Deep Learning Systems. In *2017 Symposium on Operating Systems Principles (SOSP)*. Association for Computing Machinery, 1–18.

[39] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. 2017. Towards Practical Verification of Machine Learning: The Case of Computer Vision Systems. *arXiv preprint: arXiv:1712.01785* 12 (2017).

[40] PyTorch. 2021. torchvision.models — Torchvison 0.10.0 documentation. https://pytorch.org/vision/stable/models.html.

[41] Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. 2018. Semidefinite Relaxations for Certifying Robustness to Adversarial Examples. In *2018 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 10877–10887.

[42] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. 2019. Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers. In *2019 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 11289–11300.

[43] Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, and Pengchuan Zhang. 2019. A Convex Relaxation Barrier to Tight Robustness Verification of Neural Networks. In *2019 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 9835–9846.

[44] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. 2018. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. In *2018 International Conference on Learning Representations (ICLR)*. OpenReview.

[45] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial Training for Free!. In *2019 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 3358–3369.

[46] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the Game of Go without Human Knowledge. *Nature* 550, 7676 (2017), 354–359.

[47] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. 2019. An Abstract Domain for Certifying Neural Networks. *Proc. ACM Program. Lang.* 3, POPL, Article 41 (Jan. 2019), 30 pages.

[48] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing Properties of Neural Networks. In *2014 International Conference on Learning Representations (ICLR)*. OpenReview.

[49] Jiaye Teng, Guang-He Lee, and Yang Yuan. 2020. ℓ_1 Adversarial Robustness Certificates: A Randomized Smoothing Approach. https://openreview.net/forum?id=H1lQIgrFDS

[50] Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray. 2018. DeepTest: Automated Testing of Deep-Neural-Network-Driven Autonomous Cars. In *2018 International Conference on Software Engineering (ICSE)*. Association for Computing Machinery, 303–314.

[51] Vincent Tjeng, Kai Y Xiao, and Russ Tedrake. 2018. Evaluating Robustness of Neural Networks with Mixed Integer Programming. In *2018 International Conference on Learning Representations (ICLR)*. OpenReview.

[52] Liang Tong, Bo Li, Chen Hajaj, Chaowei Xiao, Ning Zhang, and Yevgeniy Vorobeychik. 2019. Improving Robustness of ML Classifiers against Realizable Evasion Attacks Using Conserved Features. In *2019 USENIX Conference on Security Symposium (USENIX Security)*. USENIX Association, 285–302.

[53] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On Adaptive Attacks to Adversarial Example Defenses. In *2020 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 1633–1645.

[54] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2018. Ensemble Adversarial Training: Attacks and Defenses. In *2018 International Conference on Learning Representations (ICLR)*. OpenReview.

[55] Yaxiao Wang, Yuanzhang Li, Quanxin Zhang, Jingjing Hu, and Xiaohui Kuang. 2019. Evading PDF Malware Classifiers with Generative Adversarial Network. In *2019 International Symposium on Cyberspace Safety and Security (CSS)*. Springer International Publishing, 374–387.

[56] Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. 2018. Towards Fast Computation of Certified Robustness for ReLU Networks. In *2018 International Conference on Machine Learning (ICML)*. PMLR, 5276–5285.

[57] Eric Wong and Zico Kolter. 2018. Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope. In *2018 International Conference on Machine Learning (ICML)*. PMLR, 5286–5295.

[58] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. 2018. Scaling Provable Adversarial Defenses. In *2018 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 8400–8409.

[59] Chaowei Xiao, Bo Li, Jun Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. 2018. Generating Adversarial Examples with Adversarial Networks. In *2018 International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization, 3905–3911.

[60] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. 2018. Spatially Transformed Adversarial Examples. In *2018 International Conference on Learning Representations (ICLR)*. OpenReview.

[61] Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. 2020. Automatic Perturbation Analysis for Scalable Certified Robustness and Beyond. In *2020 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 1129–1141.

[62] Weilin Xu, Yanjun Qi, and David Evans. 2016. Automatically Evading Classifiers: A Case Study on PDF Malware Classifiers. In *2016 Network and Distributed Systems Symposium (NDSS)*, Vol. 10. Internet Society.

[63] Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. 2020. Randomized Smoothing of All Shapes and Sizes. In *2020 International Conference on Machine Learning (ICML)*. PMLR, 10693–10705.

[64] Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. 2020. MACER: Attack-Free and Scalable Robust Training via Maximizing Certified Radius. In *2020 International Conference on Learning Representations (ICLR)*. OpenReview.

[65] Dinghuai Zhang, Mao Ye, Chengyue Gong, Zhanxing Zhu, and Qiang Liu. 2020. Black-Box Certification with Randomized Smoothing: A Functional Optimization Based Framework. In *2020 Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2316–2326.

[66] Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho-Jui Hsieh. 2020. Towards Stable and Efficient Training of Verifiably Robust Neural Networks. In *2020 International Conference on Learning Representations (ICLR)*. OpenReview.

Appendix A introduces our method for Lipschitz bound computation for scaling transformation. **Appendix B** introduces the certification procedure of **TSS** for common transformation compositions and discusses the extension to more compositions. **Appendix C** contains the proofs for **TSS** general framework, which is introduced in Section 4. **Appendix D** contains the theorem statements and proofs for the robustness conditions derived for common smoothing distributions. These statements are instantiations of Theorem 1, and serve for both certifying resolvable transformations and differentially resolvable transformations. **Appendix E** compares the closed-form expressions of the robustness radii derived for different smoothing distributions, which corresponds to Figure 4. **Appendix F** contains the proofs of robustness conditions for various resolvable transformations. **Appendix G** contains the proofs of general theorems for certifying differentially resolvable transformations. **Appendix H** formally defines the rotation and scaling transformations, two typical differentially resolvable transformations. **Appendix I** proves our supporting theorems for computing the interpolation bound, which is used when certifying differentially resolvable transformations. **Appendix J** contains concrete algorithm descriptions for certifying differentially resolvable transformations. Finally, **Appendix K** presents the omitted details in experiments, including experiment settings, detailed discussion of baseline approaches, implementation details, and additional results.

A COMPUTING THE LIPSCHITZ BOUND FOR SCALING TRANSFORMATION

The scaling transformation ϕ_S first stretches height and width of the input image by a factor $\alpha \in \mathbb{R}_+$ where values $\alpha < 1\,(> 1)$ correspond to shrinking (enlarging) an image. Then, bilinear interpolation is applied, followed by black padding to determine pixel values. We refer the reader to Appendix H for a formal definition. Due to black padding, the functions g_i may contain discontinuities. To circumvent this issue, we enumerate all these discontinuities as $\mathcal{D}$. It can be shown that $\mathcal{D}$ contains at most $H + W$ elements. Hence, for large enough N, the interval $[\alpha_i,\, \alpha_{i+1}]$ contains at most one discontinuity. We thus modify the upper bounds M_i in (27) as

$$M_i := \begin{cases} \max\limits_{\alpha_i \le \alpha \le \alpha_{i+1}} \min\{g_i(\alpha),\, g_{i+1}(\alpha)\} & [\alpha_i,\, \alpha_{i+1}] \cap \mathcal{D} = \varnothing \\ \max\left\{ \max\limits_{\alpha_i \le \alpha \le t_i} g_{i+1}(\alpha),\, \max\limits_{t_i \le \alpha \le \alpha_{i+1}} g_i(\alpha) \right\} & [\alpha_i,\, \alpha_{i+1}] \cap \mathcal{D} = \{t_i\} \end{cases} \tag{33}$$

In either case, the quantity M_i can again be bounded by a Lipschitz constant. With this definition, the following lemma provides a closed form expression for the Lipschitz constant L in (27) for scaling. A detailed proof is given in Appendix I.

Lemma 5. *Let $x \in \mathbb{R}^{K \times W \times H}$ be a K-channel image and let ϕ_S be the scaling transformation. Then, a global Lipschitz constant L for the functions $\{g_i\}_{i=1}^N$ is given by*

$$L_s = \max_{1 \le i \le N-1} \sum_{k=0}^{K-1} \sum_{r,s \in \Omega \cap \mathbb{N}^2} \frac{\sqrt{2} d_{r,s}}{a^2} \cdot m_\Delta(x, k, \mathcal{P}_{r,s}^{(i)}) \cdot \overline{m}(x,\, k,\, \mathcal{P}_{r,s}^{(i)}) \tag{34}$$

where $\Omega = [0,\, W-1] \times [0,\, H-1]$ and a is the lower boundary value in $\mathcal{S} = [a,\, b]$. The set $\mathcal{P}_{r,s}^{(i)}$ is given by all integer grid pixels that are covered by the trajectory of source pixels of (r, s) when scaling with factors from α_{i+1} to α_i.

B CERTIFICATION OF TRANSFORMATION COMPOSITIONS

Here we state how **TSS** certifies typical transformation compositions in detail and discuss how **TSS** can be directly extended for providing robustness certificates of other transformations or their compositions.

B.1 Brightness and Contrast

As noted in Section 5.1.2, we certify the composition of brightness and contrast based on Lemma 2 and Lemma 3. To this end, we first obtain p_A, a lower bound of $q(y_A | x,\, \varepsilon_0)$ by Monte-Carlo sampling, where $\varepsilon_0 \sim \mathcal{N}(0,\, \mathrm{diag}(\sigma^2,\, \tau^2))$ is the smoothing distribution. For the given $k, b \in \mathbb{R}$ that we would like to certify $g(\phi_{BC}(x,\, (k,\, b)^T);\, \varepsilon_0) = y_A$, we then trigger Lemma 2 to get $\tilde{p}_A$, a lower bound of $q(y_A | x,\, \varepsilon_1)$, and set $\tilde{p}_B = 1 - \tilde{p}_A$. Finally, we use the explicit condition in Lemma 3 to obtain the certification.

In the actual computation, instead of certifying a single pair $(k,\, b)$, we usually certify the robustness against a set of transformation parameters

$$\mathcal{S}_{\text{adv}} = \{(k,\, b) | \, k \in [-k_0,\, k_0], b \in [-b_0,\, b_0]\}, \tag{35}$$

which stands for any contrast change within e^{k_0} and brightness change within b_0. It is infeasible to check every $(k,\, b) \in \mathcal{S}_{\text{adv}}$. To mitigate this, we relax the robustness condition in Lemma 3 from

$$\sqrt{(k/\sigma)^2 + \left(b/(e^{-k}\tau)\right)^2} < \frac{1}{2}\left(\Phi^{-1}(\tilde{p}_A) - \Phi^{-1}(\tilde{p}_B)\right) \tag{36}$$

to

$$\sqrt{(k/\sigma)^2 + \left(b/(\min\{e^{-k},\, 1\}\tau)\right)^2} < \frac{1}{2}\left(\Phi^{-1}(\tilde{p}_A) - \Phi^{-1}(\tilde{p}_B)\right). \tag{37}$$

Thus, we only need to check the condition (37) for $(k_0,\, b_0)$ and $(-k_0,\, b_0)$ to certify the robustness for any $(k,\, b)$ in (35). This is because the LHS of (37) is monotonically increasing w.r.t. $|k|$ and $|b|$, and the RHS of (37) is equal to $\Phi^{-1}(\tilde{p}_A)$ that is monotonically decreasing w.r.t. $|k|$. Throughout the experiments, we use this strategy for certification of brightness and contrast.

B.2 Gaussian Blur, Brightness, Contrast, and Translation

The certification generally follows the same procedure as in certifying brightness and contrast. In the following, we first provide a formal definition of this transformation composition. Specifically, the transformation ϕ_{BTBC} is defined as:

$$\phi_{BTBC}(x, \alpha) := \phi_B(\phi_T(\phi_{BC}(x, \alpha_k, \alpha_b), \alpha_{Tx}, \alpha_{Ty}), \alpha_B), \tag{38}$$

where ϕ_B, ϕ_T and ϕ_{BC} are Gaussian blur, translation, and brightness and contrast transformations respectively as defined before; $\alpha := (\alpha_k, \alpha_b, \alpha_{Tx}, \alpha_{Ty}, \alpha_B)^T \in \mathbb{R}^4 \times \mathbb{R}_{\ge 0}$ is the transformation parameter.

Our certification relies on the following corollary (extended from Lemma 2) and lemma, which are proved in Appendix F.

Corollary 3. *Let $x \in \mathcal{X}$, $k \in \mathbb{R}$ and let $\varepsilon_0 := (\varepsilon_0^a, \varepsilon_0^b)^T$ be a random variable defined as*

$$\varepsilon_0^a \sim \mathcal{N}(0, \mathrm{diag}(\sigma_k^2,\, \sigma_b^2,\, \sigma_T^2,\, \sigma_T^2)) \text{ and } \varepsilon_0^b \sim \mathrm{Exp}(\lambda_B). \tag{39}$$

Similarly, let $\varepsilon_1 := (\varepsilon_1^a, \varepsilon_1^b)$ *be a random variable with*

$$\varepsilon_1^a \sim \mathcal{N}(0, \mathrm{diag}(\sigma_k^2,\ e^{-2k}\sigma_b^2,\ \sigma_T^2,\ \sigma_T^2)) \text{ and } \varepsilon_1^b \sim \mathrm{Exp}(\lambda_B). \tag{40}$$

For either random variable (denoted as ε*), recall that* $q(y|x;\ \varepsilon) := \mathbb{E}(p(y|\phi_{BTBC}(x,\ \varepsilon)))$*. Suppose that* $q(y|x;\ \varepsilon_0) \geq p$ *for some* $p \in [0,\ 1]$ *and* $y \in \mathcal{Y}$*. Then* $q(y|x;\ \varepsilon_1)$ *satisfies Eq. (11).*

Lemma B.1. *Let* ε_0 *and* ε_1 *be as in Corollary 3 and suppose that*

$$q(y_A|x;\ \varepsilon_1) \geq \tilde{p}_A > \tilde{p}_B \geq \max_{y \neq y_A} q(y|x;\ \varepsilon_1). \tag{41}$$

Then it is guaranteed that $y_A = g(\phi_{BTBC}(x,\ \alpha); \varepsilon_0)$ *as long as* $p_A' > p_B'$*, where*

$$p_A' = \begin{cases} 0, & \text{if } \tilde{p}_A \leq 1 - \exp(-\lambda_B \alpha_B), \\ \Phi\left(\Phi^{-1}\left(1 - (1-\tilde{p}_A)\exp(\lambda_B\alpha_B)\right)\right. & \\ \quad \left. -\sqrt{\alpha_k^2/\sigma_k^2 + \alpha_b^2/(e^{-2\alpha_k}\sigma_b^2) + (\alpha_{Tx}^2+\alpha_{Ty}^2)/\sigma_T^2}\right) & \text{otherwise} \end{cases} \tag{42}$$

and

$$p_B' = \begin{cases} 1, & \text{if } \tilde{p}_B \geq \exp(-\lambda_B \alpha_B), \\ 1 - \Phi\left(\Phi^{-1}\left(1 - \tilde{p}_B\exp(\lambda_B\alpha_B)\right)\right. & \\ \quad \left. -\sqrt{\alpha_k^2/\sigma_k^2 + \alpha_b^2/(e^{-2\alpha_k}\sigma_b^2) + (\alpha_{Tx}^2+\alpha_{Ty}^2)/\sigma_T^2}\right). & \text{otherwise} \end{cases} \tag{43}$$

The ε_0 specified by (39) is the smoothing distribution. Similar as in brightness and contrast certification, we first obtain p_A, a lower bound of $q(y_A|\ x,\ \varepsilon_0)$ by Monte-Carlo sampling. For a given transformation parameter $\alpha := (\alpha_k,\ \alpha_b,\ \alpha_{Tx},\ \alpha_{Ty},\ \alpha_B)^T$, we then trigger Corollary 3 to get $\tilde{p}_A$, a lower bound of $q(y_A|\ x,\ \varepsilon_1)$ and set $\tilde{p}_B = 1 - \tilde{p}_A$. Finally, we use the explicit condition in Lemma B.1 to obtain the certification. Indeed, with $\tilde{p}_B = 1 - \tilde{p}_A$, Lemma B.1 can be simplified to the following corollary.

Corollary 4. *Let* ε_0 *and* ε_1 *be as in Corollary 3 and suppose that*

$$q(y_A|x;\ \varepsilon_1) \geq \tilde{p}_A. \tag{44}$$

Then it is guaranteed that $y_A = g(\phi_{BTBC}(x,\ \alpha); \varepsilon_0)$ *as long as*

$$\tilde{p}_A > 1 - \exp(-\lambda_B\alpha_B)\left(1 - \Phi\left(\sqrt{\frac{\alpha_k^2}{\sigma_k^2} + \frac{\alpha_b^2}{e^{-2\alpha_k}\sigma_b^2} + \frac{\alpha_{Tx}^2+\alpha_{Ty}^2}{\sigma_T^2}}\right)\right). \tag{45}$$

To certify against a set of transformation parameters

$$\begin{aligned} \mathcal{S}_{\text{adv}} = \{&(\alpha_k,\ \alpha_b,\ \alpha_{Tx},\ \alpha_{Ty},\ \alpha_B)^T | \\ &\alpha_k \in [-k_0,\ k_0], \alpha_b \in [-b_0,\ b_0], \\ &\|(\alpha_{Tx}, \alpha_{Ty})\|_2 \leq T, \alpha_B \leq B_0\}, \end{aligned} \tag{46}$$

we relax the robust condition in (45) to

$$\tilde{p}_A > 1 - \exp(-\lambda_B\alpha_B)\left(1 - \Phi\left(\sqrt{\frac{\alpha_k^2}{\sigma_k^2} + \frac{\alpha_b^2}{\min\{e^{-2\alpha_k},\ 1\}\sigma_b^2} + \frac{\alpha_{Tx}^2+\alpha_{Ty}^2}{\sigma_T^2}}\right)\right). \tag{47}$$

The LHS of Equation (47) is monotonically decreasing w.r.t. $|\alpha_k|$ and the RHS is monotonically increasing w.r.t. $|\alpha_k|$, $|\alpha_b|$, $\|(\alpha_{Tx}, \alpha_{Ty}\|_2$, and $|\alpha_B|$, and the RHS is symmetric w.r.t. α_b and $\|(\alpha_{Tx}, \alpha_{Ty})\|_2$. As a result, we only need to check the condition for $(-k_0,\ b_0,\ T,\ 0,\ B_0)$ and $(k_0,\ b_0,\ T,\ 0,\ B_0)$ to certify the entire set $\mathcal{S}_{\text{adv}}$. Throughout the experiments, we use this strategy for certification.

B.3 Scaling/Rotation and Brightness

To certify the composition of scaling and brightness or rotation and brightness, we follow the same methodology as certifying scaling or rotation alone and reuse the computed interpolatation error $M_{\mathcal{S}}$. We only make the following two changes: (1) alter the smoothing distribution from additive Gaussian noise $\psi(x, \delta) = x + \delta$ where $\delta \sim \mathcal{N}(0, \sigma^2\mathbb{1}_d)$ to additive Gaussian noise and Gaussian brightness change $\psi(x, \delta, \delta_b) = x + \delta + b \cdot \mathbb{1}_d$ where $\delta \sim \mathcal{N}(0, \sigma^2\mathbb{1}_d), b \sim \mathcal{N}(0, \sigma_b^2)$; (2) change the robustness condition from $R > M_{\mathcal{S}}$ in Corollary 2 to $R > \sqrt{M_{\mathcal{S}}^2 + (\sigma^2/\sigma_b^2)b_0^2}$. We formalize this robustness condition in the following corollary, and the proof is entailed in Appendix G.

Corollary 5. *Let* $\psi_B(x,\ \delta,\ b) = x + \delta + b \cdot \mathbb{1}_d$ *and let* $\varepsilon \sim \mathcal{N}(0,\ \sigma^2\mathbb{1}_d)$, $\varepsilon_b \sim \mathcal{N}(0,\ \sigma_b^2)$*. Furthermore, let* ϕ *be a transformation with parameters in* $\mathcal{Z}_\phi \subseteq \mathbb{R}^m$ *and let* $\mathcal{S} \subseteq \mathcal{Z}_\phi$ *and* $\{\alpha_i\}_{i=1}^N \subseteq \mathcal{S}$*. Let* $y_A \in \mathcal{Y}$ *and suppose that for any i, the* $(\varepsilon, \varepsilon_b)$*-smoothed classifier* $q(y|\ x;\ \varepsilon, \varepsilon_b) := \mathbb{E}(p(y|\ \psi_B(x,\ \varepsilon,\ \varepsilon_b))$ *satisfies*

$$q(y_A|\ x;\ \varepsilon, \varepsilon_b) \geq p_A^{(i)} > p_B^{(i)} \geq \max_{y \neq y_A} q(y|\ x;\ \varepsilon, \varepsilon_b). \tag{48}$$

for each i. Let

$$R := \frac{\sigma}{2} \min_{1 \leq i \leq N} \left(\Phi^{-1}\left(p_A^{(i)}\right) - \Phi^{-1}\left(p_B^{(i)}\right)\right) \tag{49}$$

Then, $\forall \alpha \in \mathcal{S}$ *and* $\forall b \in [-b_0, b_0]$ *it is guaranteed that* $y_A = \arg\max_y q(y|\ \phi(x, \alpha) + b \cdot \mathbb{1}_d;\ \varepsilon, \varepsilon_b)$ *as long as*

$$R > \sqrt{M_{\mathcal{S}}^2 + \frac{\sigma^2}{\sigma_b^2}b_0^2}, \tag{50}$$

where $M_{\mathcal{S}}$ *is defined as in Corollary 2.*

B.4 Scaling/Rotation, Brightness, and ℓ_2 Perturbations

We use the same smoothing distribution as above, and the following corollary directly allows us to certify the robustness against the composition of scaling/rotation, brightness, and an additional ℓ_2-bounded perturbations—we only need to change the robustness condition from $R > \sqrt{M_{\mathcal{S}}^2 + (\sigma^2/\sigma_b^2)b_0^2}$ to $R > \sqrt{(M_{\mathcal{S}} + r)^2 + (\sigma^2/\sigma_b^2)b_0^2}$. The proof is given in Appendix G.

Corollary 6. *Under the same setting as in Corollary 5, for* $\forall \alpha \in \mathcal{S}$*,* $\forall b \in [-b_0, b_0]$ *and* $\forall \delta \in \mathbb{R}^d$ *such that* $\|\delta\|_2 \leq r$*, it is guaranteed that* $y_A = \arg\max_k q(y|\ \phi(x, \alpha) + b \cdot \mathbb{1}_d + \delta;\ \varepsilon, \varepsilon_d)$ *as long as*

$$R > \sqrt{(M_{\mathcal{S}} + r)^2 + \frac{\sigma^2}{\sigma_b^2}b_0^2}, \tag{51}$$

where $M_{\mathcal{S}}$ *is defined as in Corollary 2.*

B.5 Discussion on More Transformations and Compositions

Our **TSS** is not limited to specific transformations. Here we briefly discuss how to extend **TSS** for any new transformations or new compositions.

For a new transformation, we first identify the parameter space $\mathcal{Z}$, where the identified parameter should completely and deterministically decide the output after transformation for any given input. Then, we use Definition 2 to check whether the transformation is resolvable. If so, we can write down the function γ_α. Next, we choose a smoothing distribution, i.e., the distribution of the random variable ε_0, and identify the distribution of $\varepsilon_1 = \gamma_\alpha(\varepsilon_0)$. Finally, we use Theorem 1 to derive the robustness certificates and follow the two-step template (Section 5.1.5) to compute the robustness certificate.

If the transformation is not resolvable, we identify a dimension in $\mathcal{Z}$ for which the transformation is *resolvable*. For example, the composition of rotation and brightness has a rotation and a brightness axis, where the brightness axis is itself resolvable. As a result, we can write the parameter space as Cartesian product of non-resolvable subspace and resolvable subspace: $\mathcal{Z} := \mathcal{Z}_{\text{no-resolve}} \times \mathcal{Z}_{\text{resolve}}$. We perform smoothing on the resolvable subspace and sample enough points in the non-resolvable subspace. Next, we bound the interpolation error between sampled points and arbitrary points in the non-resolvable subspace, using either ℓ_p difference as we did for rotation and scaling or other regimes. Specifically, our Lemma 8 shown in Appendix I is a useful tool to bound ℓ_p difference caused by interpolation error. Finally, we instantiate Theorem 2 to compute the robustness certificate.

Theoretically, we can certify against the composition of all the discussed transformations: Gaussian blur, brightness, contrast, translation, rotation, and scaling. However, as justified in [21, Figure 3], the composition of more than two transformations leads to unrealistic images that are even hard to distinguish by humans. Moreover, if the composition contains too many transformations, the parameter space would be no longer low dimensional. Therefore, there would be much more axes that are differentially resolvable (instead of resolvable). As a consequence, much more samples are required to obtain a small bound on interpolation error (which is necessary for a nontrivial robustness certification). Therefore, we mainly evaluate on single transformation or composition of two transformations to simulate a practical attack.

C PROOFS FOR THE GENERAL CERTIFICATION FRAMEWORK

Here we provide the proof for Theorem 1. For that purpose, recall the following definition from the main part of this paper:

Definition 1 (restated). *Let $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ be a transformation, $\varepsilon \sim \mathbb{P}_\varepsilon$ a random variable taking values in $\mathcal{Z}$ and let $h\colon \mathcal{X} \to \mathcal{Y}$ be a base classifier. We define the ε-smoothed classifier $g\colon \mathcal{X} \to \mathcal{Y}$ as $g(x;\varepsilon) = \arg\max_{y\in\mathcal{Y}} q(y|\, x;\, \varepsilon)$ where q is given by the expectation with respect to the smoothing distribution ε, i.e.,*

$$q(y|\, x;\, \varepsilon) := \mathbb{E}(p(y|\, \phi(x,\, \varepsilon))). \tag{52}$$

Here, we additionally define the notion of level sets separately. These sets originate from statistical hypothesis testing correspond to rejection regions of likelihood ratio tests.

Definition 7 (Lower level sets). *Let $\varepsilon_0 \sim \mathbb{P}_0$, $\varepsilon_1 \sim \mathbb{P}_1$ be $\mathcal{Z}$-valued random variables with probability density functions f_0 and f_1 with respect to a measure μ. For $t \geq 0$ we define lower and strict lower level sets as*

$$\underline{S_t} := \{z \in \mathcal{Z}\colon\ \Lambda(z) < t\}, \quad \overline{S_t} := \{z \in \mathcal{Z}\colon\ \Lambda(z) \leq t\}, \quad \text{where} \quad \Lambda(z) := \frac{f_1(z)}{f_0(z)}. \tag{53}$$

We also make the following definition in order to reduce clutter and simplify the notation. This definition will be used throughout the proofs presented here.

Definition 8 ((p_A, p_B)-Confident Classifier). *Let $x \in \mathcal{X}$, $y_A \in \mathcal{Y}$ and p_A, $p_B \in [0,\, 1]$ with $p_A > p_B$. We say that the ε-smoothed classifier q is $(p_A,\, p_B)$-confident at x if*

$$q(y_A|\, x;\, \varepsilon) \geq p_A \geq p_B \geq \max_{y \neq y_A} q(y|\, x;\, \varepsilon). \tag{54}$$

C.1 Auxiliary Lemmas

Lemma 6. *Let ε_0 and ε_1 be random variables taking values in $\mathcal{Z}$ and with probability density functions f_0 and f_1 with respect to a measure μ. Denote by Λ the likelihood ratio $\Lambda(z) = f_1(z)/f_0(z)$. For $p \in [0,\, 1]$ let $\tau_p := \inf\{t \geq 0\colon\ \mathbb{P}_0(\overline{S_t}) \geq p\}$. Then, it holds that*

$$\mathbb{P}_0\left(\underline{S}_{\tau_p}\right) \leq p \leq \mathbb{P}_0\left(\overline{S}_{\tau_p}\right). \tag{55}$$

Proof. We first show the RHS of inequality (55). This follows directly from the definition of τ_p if we show that the function $t \mapsto \mathbb{P}_0\left(\overline{S}_t\right)$ is right-continuous. For that purpose, let $t \geq 0$ and let $\{t_n\}_n$ be a sequence in $\mathbb{R}_{\geq 0}$ such that $t_n \downarrow t$. Define the sets $A_n := \{z\colon\ \Lambda(z) \leq t_n\}$ and note that $A_{n+1} \subseteq A_n$. Clearly, if $z \in \overline{S}_t$, then $\forall\, n\colon\ \Lambda(z) \leq t \leq t_n$, thus $z \in \cap_n A_n$ and hence $\overline{S}_t \subseteq \cap_n A_n$. If on the other hand $z \in \cap_n A_n$, then $\forall n\colon \Lambda(z) \leq t_n \to t$ as $n \to \infty$ and thus $z \in \overline{S}_t$, yielding $\overline{S}_t = \cap_n A_n$. Hence for any $t \geq 0$ we have that

$$\lim_{n\to\infty} \mathbb{P}_0\left(A_n\right) = \mathbb{P}_0\left(\bigcap_n A_n\right) = \mathbb{P}_0\left(\underline{S}_t\right). \tag{56}$$

Thus, the function $t \mapsto \mathbb{P}_0\left(\overline{S}_t\right)$ is right continuous and in particular it follows that $\mathbb{P}_0\left(\overline{S}_{\tau_p}\right) \geq p$. We now show the LHS of inequality (55). Consider the sets $B_n := \{z\colon\ \Lambda(z) < \tau_p - 1/n\}$ and note that $B_n \subseteq B_{n+1}$. Clearly, if $z \in \cup_n B_n$, then $\exists\, n$ such that $\Lambda(z) < \tau_p - 1/n < \tau_p$ and thus $z \in \underline{S}_{\tau_p}$. If on the other hand $z \in \underline{S}_{\tau_p}$, then we can choose n large enough such that $\Lambda(z) < \tau_p - 1/n$ and thus $z \in \cup_n B_n$ yielding $\underline{S}_{\tau_p} = \cup_n B_n$. Furthermore, by the definition of τ_p and since for any $n \in \mathbb{N}$ we have that $\mathbb{P}_0(B_n) = \mathbb{P}_0\left(\underline{S}(\tau_p - 1/n)\right) < p$ it follows that

$$\mathbb{P}_0\left(\underline{S}_{\tau_p}\right) = \mathbb{P}_0\left(\bigcup_n B_n\right) = \lim_{n\to\infty} \mathbb{P}_0\left(B_n\right) \leq p \tag{57}$$

concluding the proof. □

Lemma 7. *Let ε_0 and ε_1 be random variables taking values in $\mathcal{Z}$ and with probability density functions f_0 and f_1 with respect to a measure μ. Let $h\colon \mathcal{Z} \to [0,\, 1]$ be a determinstic function. Then, for any $t \geq 0$ the following implications hold:*

(i) For any $S \subseteq \mathcal{Z}$ with $\underline{S_t} \subseteq S \subseteq \overline{S_t}$ the following implication holds:

$$\mathbb{E}[h(\varepsilon_0)] \geq \mathbb{P}_0(S) \Rightarrow \mathbb{E}[h(\varepsilon_1)] \geq \mathbb{P}_1(S). \tag{58}$$

(i) For any $S \subseteq \mathcal{Z}$ with $\overline{S_t}^c \subseteq S \subseteq \underline{S_t}^c$ the following implication holds:

$$\mathbb{E}[h(\varepsilon_0)] \leq \mathbb{P}_0(S) \Rightarrow \mathbb{E}[h(\varepsilon_1)] \leq \mathbb{P}_1(S). \tag{59}$$

Proof. We first prove (i). For that purpose, consider

$$\begin{aligned}
\mathbb{E}[f(\varepsilon_1)] - \mathbb{P}_1(S) &= \int h f_1 \, d\mu - \int_S f_1 \, d\mu && (60)\\
&= \int_{S^c} h f_1 \, d\mu - \left(\int_S (1-h) f_1 \, d\mu \right) && (61)\\
&= \int_{S^c} h \Lambda f_0 \, d\mu - \left(\int_S (1-h) \Lambda f_0 \, d\mu \right) && (62)\\
&\geq t \cdot \int_{S^c} h f_0 \, d\mu - t \cdot \left(\int_S (1-h) f_0 \, d\mu \right) && (63)\\
&= t \cdot \left(\int h f_0 \, d\mu - \int_S f_0 \, d\mu \right) && (64)\\
&= t \cdot (\mathbb{E}[f(\varepsilon_0)] - \mathbb{P}_0(S)) \geq 0. && (65)
\end{aligned}$$

The inequality in (63) follows from the fact that whenever $z \in S^c$, then $f_1(z) \geq t \cdot f_0(z)$ and if $z \in S$, then $f_1(z) \leq t \cdot f_0(z)$ since S is a lower level set. Finally, the inequality in (65) follows from the assumption. The proof of (ii) is analogous and omitted here. □

C.2 Proof of Theorem 1

Theorem 1 (restated). *Let $\varepsilon_0 \sim \mathbb{P}_0$ and $\varepsilon_1 \sim \mathbb{P}_1$ be $\mathcal{Z}$-valued random variables with probability density functions f_0 and f_1 with respect to a measure μ on $\mathcal{Z}$ and let $\phi \colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ be a semantic transformation. Suppose that $y_A = g(x;\, \varepsilon_0)$ and let $p_A,\, p_B \in [0,\, 1]$ be bounds to the class probabilities, i.e.,*

$$q(y_A|\, x,\, \varepsilon_0) \geq p_A > p_B \geq \max_{y \neq y_A} q(y|\, x,\, \varepsilon_0). \tag{66}$$

For $t \geq 0$, let $\underline{S}_t,\, \overline{S}_t \subseteq \mathcal{Z}$ be the sets defined as $\underline{S}_t \coloneqq \{f_1/f_0 < t\}$ and $\overline{S}_t \coloneqq \{f_1/f_0 \leq t\}$ and define the function $\xi \colon [0,\, 1] \to [0,\, 1]$ by

$$\begin{aligned}
\xi(p) &\coloneqq \sup\{\mathbb{P}_1(S) \colon \underline{S}_{\tau_p} \subseteq S \subseteq \overline{S}_{\tau_p}\}\\
\text{where} \quad \tau_p &\coloneqq \inf\{t \geq 0 \colon \mathbb{P}_0(\overline{S}_t) \geq p\}.
\end{aligned} \tag{67}$$

If the condition

$$\xi(p_A) + \xi(1 - p_B) > 1 \tag{68}$$

is satisfied, then it is guaranteed that $g(x;\, \varepsilon_1) = g(x;\, \varepsilon_0)$.

Proof. For ease of notation, let ζ be the function defined by

$$t \mapsto \zeta(t) \coloneqq \mathbb{P}_0(\overline{S}_t) \tag{69}$$

and notice that $\tau_p = \zeta^{-1}(p)$ where ζ^{-1} denotes the generalized inverse of ζ. Furthermore, let $\tau_A \coloneqq \tau_{p_A}$, $\tau_B \coloneqq \tau_{1-p_B}$, $\underline{S}_A \coloneqq \underline{S}(\tau_A)$, $\underline{S}_B \coloneqq \underline{S}(\tau_B)$, $\overline{S}_A \coloneqq \overline{S}(\tau_A)$ and $\overline{S}_B \coloneqq \overline{S}(\tau_B)$. We first show that $q(y_A|\, x,\, \varepsilon_1)$ is lower bounded by $\xi(\tau_A)$. For that purpose, note that by Lemma 6 we have that $\zeta(\tau_A) = \mathbb{P}_0(\overline{S}_A) \geq p_A \geq \mathbb{P}_0(\underline{S}_A)$. Thus, the collection of sets

$$\mathcal{S}_A \coloneqq \{S \subseteq \mathcal{Z} \colon \underline{S}_A \subseteq S \subseteq \overline{S}_A,\, \mathbb{P}_0(S) \leq p_A\} \tag{70}$$

is not empty. Pick some $A \in \mathcal{S}_A$ arbitrary and note that, since by assumption $g(\cdot;\, \varepsilon_0)$ is $(p_A,\, p_B)$-confident at x it holds that

$$\mathbb{E}(p(y_A|\, \phi(x,\, \varepsilon_0))) = q(y_A|\, x;\, \varepsilon_0) \geq p_A \geq \mathbb{P}_0(A). \tag{71}$$

Since $\underline{S}_A \subseteq A \subseteq \overline{S}_A$ we can apply part (i) of Lemma 7 and obtain the lower bound

$$q(y_A|\, x,\, \varepsilon_1) = \mathbb{E}(p(y_A|\, \phi(x,\, \varepsilon_1))) \geq \mathbb{P}_1(A). \tag{72}$$

Since $A \in \mathcal{S}_A$ was arbitrary, we take the sup over all $A \in \mathcal{S}_A$ and obtain

$$q(y_A|\, x;\, \varepsilon_1) \geq \sup_{A \in \mathcal{S}_A} \mathbb{P}_1(A) = \xi(p_A) \tag{73}$$

We now show that for any $y \neq y_A$ the prediction $q(y|\, x;\, \varepsilon_1)$ is upper bounded by $1 - \xi(1 - p_B)$. For that purpose, note that by Lemma 6 we have that $\zeta(\tau_B) = \mathbb{P}_0(\overline{S}_A) \geq 1 - p_B \geq \mathbb{P}_0(\underline{S}_B)$. Thus, the collection of sets

$$\mathcal{S}_B \coloneqq \{S \subseteq \mathcal{Z} \colon \underline{S}_B \subseteq S \subseteq \overline{S}_B,\, \mathbb{P}_0(S) \leq 1 - p_B\} \tag{74}$$

is not empty. Pick some $B \in \mathcal{S}_A$ arbitrary and note that, since by assumption $g(\cdot;\, \varepsilon_0)$ is $(p_A,\, p_B)$-confident at x it holds that

$$\begin{aligned}
\mathbb{E}(p(y|\, \phi(x,\, \varepsilon_0))) &= q(y|\, x;\, \varepsilon_0) \leq p_B\\
&= 1 - (1 - p_B) \leq 1 - \mathbb{P}_0(B).
\end{aligned} \tag{75}$$

Since $\underline{S}_B^c \subseteq B^c \subseteq \overline{S}_B^c$ we can apply part (ii) of Lemma 7 and obtain the upper bound

$$q(y|\, x;\varepsilon_1) = \mathbb{E}(p(y|\, \phi(x,\, \varepsilon_1))) \leq 1 - \mathbb{P}_1(B). \tag{76}$$

Since $B \in \mathcal{S}_B$ was arbitrary, we take the inf over all $B \in \mathcal{S}_B$ and obtain

$$q(y|\, x;\varepsilon_1) \leq \inf_{B \in \mathcal{S}_B} (1 - \mathbb{P}_1(B)) = 1 - \xi(1 - p_B). \tag{77}$$

Combining together (77) and (73), we find that, whenever

$$\xi(p_A) + \xi(1 - p_B) > 1 \tag{78}$$

it is guaranteed that

$$q(y_A|\, x;\, \varepsilon_1) > \max_{y \neq y_A} q(y_|\, x;\, \varepsilon_1) \tag{79}$$

what concludes the proof. □

D PROOFS FOR CERTIFICATION WITH DIFFERENT SMOOTHING DISTRIBUTIONS

Hereby we only list selected main results and leave the proofs to the arXiv version [29] due to space limit.

Corollary 7 (Non-Isotropic Gaussian Smoothing). *Suppose $\mathcal{Z} = \mathbb{R}^m$, $\Sigma \coloneqq \operatorname{diag}(\sigma_1^2,\, \ldots,\, \sigma_m^2)$, $\varepsilon_0 \sim \mathcal{N}(0,\, \Sigma)$ and $\varepsilon_1 \coloneqq \alpha + \varepsilon_0$ for some $\alpha \in \mathbb{R}^m$. Suppose that the ε_0-smoothed classifier g is $(p_A,\, p_B)$-confident at $x \in \mathcal{X}$ for some $y_A \in \mathcal{Y}$. Then, it holds that $q(y_A|\, x;\, \varepsilon_1) > \max_{y \neq y_A} q(y|\, x;\, \varepsilon_1)$ if α satisfies*

$$\sqrt{\sum_{i=1}^m \left(\frac{\alpha_i}{\sigma_i} \right)^2} < \frac{1}{2} \left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B) \right). \tag{80}$$

Corollary 8 (Exponential Smoothing). *Suppose $\mathcal{Z} = \mathbb{R}^m_{\geq 0}$, fix some $\lambda > 0$ and let $\varepsilon_{0,i} \overset{\text{iid}}{\sim} \operatorname{Exp}(1/\lambda)$, $\varepsilon_0 \coloneqq (\varepsilon_{0,1},\, \ldots,\, \varepsilon_{0,m})^T$ and $\varepsilon_1 \coloneqq \alpha + \varepsilon_0$ for some $\alpha \in \mathbb{R}^m_{\geq 0}$. Suppose that the ε_0-smoothed classifier g is $(p_A,\, p_B)$-confident at $x \in \mathcal{X}$ for some $y_A \in \mathcal{Y}$. Then, it holds that $q(y_A|\, x;\, \varepsilon_1) > \max_{y \neq y_A} q(y|\, x;\, \varepsilon_1)$ if α satisfies*

$$\|\alpha\|_1 < -\frac{\log(1 - p_A + p_B)}{\lambda}. \tag{81}$$

Algorithm 1 Interpolation Error M Computation for Rotation Transformation.

Input: clean input image x;
interval of rotation angle to certify $[a,\ b]$;
number of first-level samples N;
number of second-level samples n
Output: rotation angle samples $\{\alpha_i\}_{i=1}^N$;
upper bound M of squared ℓ_2-interpolation error

$$M_{\mathcal{S}}^2 = \arg\max_{\alpha\in[a,b]} \min_{1\le i\le N} \|\tilde{\phi}_R(x,\alpha) - \tilde{\phi}_R(x,\alpha_i)\|_2^2.$$

/ Compute Lipschitz constant L_r (32) */*
$\alpha_1 \leftarrow a$
for $i = 1, \ldots, N-1$ **do**
  $\alpha_{i+1} \leftarrow a + (b-a)\cdot \frac{i}{N-1}$ (21)
  for all $(r,s) \in V$ **do**
    / V and $\mathcal{P}_{r,s}^{(i)}$ are defined in Lemma 4 */*
    Compute trajectory covered grid pixels $\mathcal{P}_{r,s}^{(i)}$
    for $k = 0, \ldots, K-1$ **do**
      Compute $2d_{r,s}\cdot m_\Delta(x,k,\mathcal{P}_{r,s}^{(i)})\cdot \bar{m}(x,k,\mathcal{P}_{r,s}^{(i)})$ (32)
    end for
  end for
  $L_{r,i} \leftarrow \sum_{k=0}^{K-1}\sum_{(r,s)\in V} 2d_{r,s}\cdot m_\Delta(x,k,\mathcal{P}_{r,s}^{(i)})\cdot \bar{m}(x,k,\mathcal{P}_{r,s}^{(i)})$.
end for
$L_r \leftarrow \max_{1\le i\le N-1} L_{r,i}$ (32)
/ Compute interpolation error bound M (24) from stratified sampling */*
for $i = 1, \ldots, N-1$ **do**
  for $j = 1, \ldots, n$ **do**
    / Second-level sampling */*
    $\gamma_{i,j} \leftarrow \alpha_i + (\alpha_{i+1}-\alpha_i)\cdot\frac{j-1}{n-1}$ (25)
  end for
  $M_i \leftarrow 0$
  for $j = 1, \ldots, n-1$ **do**
    Compute $g_i(\gamma_{i,j})$, $g_i(\gamma_{i,j+1})$, $g_{i+1}(\gamma_{i,j})$, and $g_{i+1}(\gamma_{i,j+1})$ (22)
    $M_i \leftarrow \max\{M_i, \min\{g_i(\gamma_{i,j}) + g_i(\gamma_{i,j+1}),$
      $g_{i+1}(\gamma_{i,j}) + g_{i+1}(\gamma_{i,j+1})\}\}$
  end for
  $M_i \leftarrow \frac{1}{2}M_i + L\cdot\frac{b-a}{(N-1)(n-1)}$ (27)
end for
Return: $M \leftarrow \max_{1\le i\le N-1} M_i$ (24)

Table 5: Comparison of certification radii with $p_A + p_B = 1$, the variance and noise dimension are set to 1 for each distribution.

Distribution	Value Space	Robust Radius
Gaussian$(0,\ 1)$	$(-\infty,\ \infty)$	$\Phi^{-1}(p_A)$
Laplace$(0,\ 1/\sqrt{2})$	$(-\infty,\ \infty)$	$-\log(2-2p_A)/\sqrt{2}$
Uniform$[-\sqrt{3},\ -\sqrt{3}]$	$(-\infty,\ \infty)$	$2\sqrt{3}\cdot(p_A - 1/2)$
Exponential(1)	$[0,\ \infty)$	$-\log(2-2p_A)$
FoldedGaussian$(0,\ \sqrt{\frac{\pi}{\pi-2}})$	$[0,\ \infty)$	$\sqrt{\frac{\pi}{\pi-2}}\cdot\left(\Phi^{-1}\left(\frac{1+p_A}{2}\right) - \Phi^{-1}\left(\frac{3}{4}\right)\right)$

E COMPARISON OF DIFFERENT SMOOTHING DISTRIBUTIONS

Here, we provide robustness radii derived from different smoothing distributions in Table 5, each scaled to have unit variance. The figure comparison is in Figure 4 in main text.

F PROOFS FOR CERTIFYING RESOLVABLE TRANSFORMATIONS

We leave the proofs to the arXiv version [29] due to space limit.

G PROOFS FOR DIFFERENTIALLY RESOLVABLE TRANSFORMATIONS

In the corresponding appendix of the arXiv version [29], we provide proofs and technical details for theoretical results about certifying differentially resolvable transformations.

H TRANSFORMATION DETAILS

Due to space limit, we provide detailed definitions of rotation and scaling transformation in the corresponding appendix of the arXiv version [29].

I PROOFS FOR INTERPOLATION BOUND COMPUTATION

We state the proofs for the theoretical results governing our approach to certifying rotations and scaling transformations using randomized smoothing in the corresponding appendix of the arXiv version [29]. As a quick glance, the following auxiliary lemma is used for both rotation and scaling:

Lemma 8. *Let $x \in \mathbb{R}^{K\times W\times H}$, $-\infty < t_1 < t_2 < \infty$ and suppose $\rho\colon [t_1,\ t_2] \to [0,\ W-1]\times[0,\ H-1]$ is a curve of class C^1. Let*

$$\psi_k\colon [t_1,\ t_2] \to \mathbb{R},\ \ \psi_k(t) \coloneqq Q_x(k,\ \rho_1(t),\ \rho_2(t)) \tag{82}$$

where $k \in \Omega_K$ and Q_x denotes bilinear interpolation. Then ψ_k is L_k-Lipschitz continuous with constant

$$L_k = \max_{t\in[t_1,t_2]} \left(\sqrt{2}\,\|\dot{\rho}(t)\|_2 \cdot m_\Delta(x,\ k,\ \lfloor\rho(t)\rfloor)\right) \tag{83}$$

J ALGORITHM DESCRIPTION FOR DIFFERENTIALLY RESOLVABLE TRANSFORMATIONS

Algorithm 1 presents a pseudo-code for interpolation error M computation, taking rotation transformation as the example. It corresponds to the description in Section 6.2. Algorithm 2 presents a pseudo-code for progressive sampling. It corresponds to the description in Section 6.4.2. We remark that in practice, we sample in mini-batches with batch size B. The error tolerance T is set to $M_{\mathcal{S}}$ (20) if certifying rotation or scaling, is set to $\sqrt{M_{\mathcal{S}}^2 + \sigma^2/\sigma_b^2\cdot b_0^2}$ (50) if certifying the composition of rotation or scaling with brightness change within $[-b_0,\ b_0]$; and is set to $\sqrt{(M_{\mathcal{S}}+r)^2 + \sigma^2/\sigma_b^2\cdot b_0^2}$ (51) if certifying the composition of rotation or scaling, brightness change $[-b_0,\ b_0]$, and ℓ_2 bounded perturbations within r. The two algorithms jointly constitute our pipeline **TSS-DR** for certifying against differentially resolvable transformations as shown in Figure 5a.

K OMITTED EXPERIMENT DETAILS

Here we provide all omitted details about experiment setup, implementation, discussion about baselines, evaluation protocols, results, findings, and analyses. Due to the space limit, most of the results are omitted to the corresponding appendix of the arXiv version [29].

Algorithm 2 Progressive Sampling for Certification.

Input: clean input image x with true class k_A; first-level parameter samples $\{\alpha_i\}_{i=1}^{N}$; perturbation random variable ε with variance σ^2; ℓ_2 error tolerance T; batch size B; sampling size limit n_s; confidence level p.
Output: with probability $1-p$, whether $g(\cdot;\varepsilon)$ is certifiably robust at $\phi(x,\alpha)$.

```
for i = 1, ..., N do
  x^(i) ← φ(x, α_i)
  j ← 0
  while j ≤ n_s do
    Sample B instances of φ(x^(i), ε), and use them to update empirical
    mean q̂(y_A | x^(i); ε).
    j ← j + B.
    /* Lower confidence interval bound with these j samples */
    p_A^(i) = LowerConfBound(q̂(y_A | x^(i); ε), j, 1 − p/N).
    if R_i = σΦ^{-1}(p_A^(i)) > T then
      /* Already get the certification that R_i > T, break */
      Break
    end if
  end while
  if R_i = σΦ^{-1}(p_A^(i)) ≤ T then
    /* Cannot ensure that R_i > T. So cannot ensure that R = min R_i > T.
    Early halt */
    Return: false
  end if
end for
Return: true
```

K.1 Model Preparation

Due to the space limit, the full description of the model preparation including all hyperparameters are omitted to the corresponding section of the arXiv version [29].

K.2 Implementation Details

We implement the whole approach along with the training scripts in a tool based on `PyTorch`. For resolvable transformations, we extend the smoothing module from Cohen et al [7] to accommodate various smoothing strategies and smoothing distributions. The predict and certify modules are kept the same. For differentially resolvable transformations, since the stratified sampling requires $N \times n$ times of transformation to compute the interpolation error bound (where N is the number of first-level samples and n the number of second-level samples), we implement a fast `C` module and integrate it to our `Python`-based tool. It empirically achieves $3-5$x speed gain compared with `OpenCV`[37]-based transformation. For Lipschitz upper bound computation, since the loop in `Python` is slow, we reformulate the computation by loop-free tensor computations using `numpy`. It empirically achieves $20-40$x speed gain compared to the plain loop-based implementation. The full code implementation of our **TSS** tool along with all trained models are publicly available at https://github.com/AI-secure/semantic-randomized-smoothing.

K.3 Details on Attacks

We use the following three attacks to evaluate the empirical accuracy of both **TSS** models and vanilla models: Random Attack, Random+ Attack, and PGD Attack. The Random Attack is used in previous work [2, 13] but does not consider the intrinsic characteristics of semantic transformations. Thus, we propose Random+ Attack and PGD Attack as the alternatives since they are adaptively designed for our smoothed **TSS** models and also consider the intrinsic characteristics of these transformations. The detailed description of these three attacks can be found in the arXiv version [29].

K.4 Details on Baseline Approaches

DeepG [2] is based on linear relaxations. The code is open-sourced, and we utilize it to provide a direct comparison. The code provides trained models on MNIST and CIFAR-10, while on ImageNet the method is too slow and memory-consuming to run. On both MNIST and CIFAR-10, we use the provided trained models from the code. In terms of computation time, since our approach uses far less than 1000 s for certification per input on MNIST and CIFAR-10, we tune the hyperparameters to let the code spend roughly 1000 s for certification.

Interval [47] is based on interval bound propagation. We also utilize the open-sourced code to provide a direct comparison. The settings are the same as in DeepG.

VeriVis [39] provides an enumeration-based solution when the number of possible transformation parameters or the number of possible transformed images is finite. In our evaluation, only translation satisfies this property. Therefore, as the baseline, we implement the enumeration-based robustness certification algorithm for our trained robust models.

Semanify-NN [35] proposes to insert a preprocessing layer to reduce the verification against semantic transformations to the verification against classical ℓ_p noises. To our knowledge, the code has not been open-sourced yet. Therefore, we directly compare with the numbers reported in their paper. Since they report the average of certified robust magnitude, we apply Markov's inequality to obtain an upper bound of their certified robust accuracy. For example, they report 46.24 degrees as the average certified robust rotation angles. It means that $\mathbb{P}[r \geq 50^\circ] \leq \mathbb{E}[r]/50 = 92.48\%$, i.e., the certified robust accuracy is no larger than 92.48% when fixing the rotation angle to be 50°.

DistSPT [13] combines randomized smoothing and interval bound propagation to provide certified robustness against semantic transformations. Concretely, the approach leverages interval bound propagation to compute the upper bound of interpolation error and then applies randomized smoothing. On small datasets such as MNIST and CIFAR-10, the approach is able to provide nontrivial robustness certification. Though the certified robust accuracy is inferior than **TSS** as reflected in Table 2. We use their reported numbers in [13, Table 4] for DistSPT^x for comparison, since the certification goal and evaluation protocol are the same as ours. On ImageNet, as described in [13, Section 7.4], the interval bound propagation is computationally expensive and loose. Therefore, they use sampling to estimate the interpolation error, which makes the robustness certification no longer hold against arbitrary attacks but just a certain random attack ("worst-of-10" attack).

IndivSPT [13] provides a different certification goal from the above approaches. At a high level, the approach uses a transformed image as the input where the transformation parameter is within predefined threshold. Then the approach certifies whether the prediction for the transformed image and the prediction for the original image are the same. In contrast, **TSS** and other baseline approaches

take original image as the input and certifies whether there exists no transformed image that can mislead the model. Due to different certification goals, **TSS** is not comparable with IndivSPT.

K.5 Benign Accuracy

Table 6 shows the benign accuracy of our models corresponding to Table 2. For comparison, the vanilla trained models have benign accuracy 98.6% on MNIST, 88.6% on CIFAR-10, and 74.4% on ImageNet.

We observe that though the trade-off between accuracy and (certified) robustness is widely reported both theoretically [34, 63, 65] and empirically (e.g., [7, 24, 64]) in classical ℓ_p threat model, it does not always exist in our semantic defense setting. Specifically, for resolvable transformations, we do not observe an apparent loss of benign accuracy in our certifiably robust models; while for differentially resolvable transformations (those involving scaling and rotation), there is no loss on MNIST, slight losses on CIFAR-10, and apparent losses on ImageNet. When there does exist a trade-off between benign accuracy and certified robust accuracy, we show that smoothing variance levels control it in Appendix K.10.

Table 6: Benign accuracy of our **TSS** models corresponding to those in Table 2. Certified robust accuracy shown as reference.

Transformation	Dataset	Attack Radius	Certified Robust Acc.	Benign Acc.
Gaussian Blur	MNIST	Squared Radius $\alpha \leq 36$	90.6%	96.8%
	CIFAR-10	Squared Radius $\alpha \leq 16$	63.6%	76.2%
	ImageNet	Squared Radius $\alpha \leq 36$	51.6%	59.2%
Translation (Reflection Pad.)	MNIST	$\sqrt{\Delta x^2 + \Delta y^2} \leq 8$	99.6%	99.6%
	CIFAR-10	$\sqrt{\Delta x^2 + \Delta y^2} \leq 20$	80.8%	87.0%
	ImageNet	$\sqrt{\Delta x^2 + \Delta y^2} \leq 100$	50.0%	73.0%
Brightness	MNIST	$b \pm 50\%$	98.2%	98.2%
	CIFAR-10	$b \pm 40\%$	87.0%	87.8%
	ImageNet	$b \pm 40\%$	70.0%	72.2%
Contrast and Brightness	MNIST	$c \pm 50\%, b \pm 50\%$	97.6%	98.0%
	CIFAR-10	$c \pm 40\%, b \pm 40\%$	82.4%	86.8%
	ImageNet	$c \pm 40\%, b \pm 40\%$	61.4%	72.2%
Gaussian Blur, Translation, Brightness, and Contrast	MNIST	$\alpha \leq 1, c, b \pm 10\%, \sqrt{\Delta x^2 + \Delta y^2} \leq 5$	90.2%	98.2%
	CIFAR-10	$\alpha \leq 1, c, b \pm 10\%, \sqrt{\Delta x^2 + \Delta y^2} \leq 5$	58.2%	77.6%
	ImageNet	$\alpha \leq 10, c, b \pm 20\%, \sqrt{\Delta x^2 + \Delta y^2} \leq 10$	32.8%	61.6%
Rotation	MNIST	$r \pm 50^\circ$	97.4%	99.4%
	CIFAR-10	$r \pm 10^\circ$	70.6%	83.2%
		$r \pm 30^\circ$	63.6%	82.6%
	ImageNet	$r \pm 30^\circ$	30.4%	46.2%
Scaling	MNIST	$s \pm 30\%$	99.0%	99.4%
	CIFAR-10	$s \pm 30\%$	58.8%	79.8%
	ImageNet	$s \pm 30\%$	26.4%	50.8%
Rotation and Brightness	MNIST	$r \pm 50^\circ, b \pm 20\%$	97.0%	99.4%
	CIFAR-10	$r \pm 10^\circ, b \pm 10\%$	70.2%	83.0%
		$r \pm 30^\circ, b \pm 20\%$	61.4%	82.6%
	ImageNet	$r \pm 30^\circ, b \pm 20\%$	26.8%	45.8%
Scaling and Brightness	MNIST	$s \pm 50\%, b \pm 50\%$	96.6%	99.4%
	CIFAR-10	$s \pm 30\%, b \pm 30\%$	54.2%	79.6%
	ImageNet	$s \pm 30\%, b \pm 30\%$	23.4%	50.8%
Rotation, Brightness, and ℓ_2	MNIST	$r \pm 50^\circ, b \pm 20\%, \|\delta\|_2 \leq .05$	96.6%	99.4%
	CIFAR-10	$r \pm 10^\circ, b \pm 10\%, \|\delta\|_2 \leq .05$	64.2%	83.0%
		$r \pm 30^\circ, b \pm 20\%, \|\delta\|_2 \leq .05$	55.2%	82.6%
	ImageNet	$r \pm 30^\circ, b \pm 20\%, \|\delta\|_2 \leq .05$	26.6%	45.8%
Scaling, Brightness, and ℓ_2	MNIST	$s \pm 50\%, b \pm 50\%, \|\delta\|_2 \leq .05$	96.4%	99.4%
	CIFAR-10	$s \pm 30\%, b \pm 30\%, \|\delta\|_2 \leq .05$	51.2%	79.6%
	ImageNet	$s \pm 30\%, b \pm 30\%, \|\delta\|_2 \leq .05$	22.6%	50.8%

K.6 Smoothing Distributions and Running Time Statistics

In Table 7, we present the smoothing distributions with concrete parameters and average certification computing time per sample for results in main table (Table 2). In the table, α is for squared kernel radius for Gaussian blur; Δx and Δy are for translation displacement on horizontal and vertical direction; b and c are for brightness shift and contrast change respectively as in $x \mapsto (1 + c)x + b$; r is for rotation angle; s is for size scaling ratio; ε is for additive noise vector; and $\|\delta\|_2$ for ℓ_2 norm of permitted additional perturbations. Specifically, "Training Distribution" stands for the distributions for data augmentation during training the base classifiers; and "Smoothing Distribution" stands for the distributions for constructing the smoothed classifiers for certification.

We select these distributions according to the principles in Appendix K.1.

K.7 Comparison of Random Attack and Adaptive Attacks: Detail

Detailed results are omitted to the corresponding section of the arXiv version [29].

K.8 Empirical Robustness against Unforeseen Attacks: Evaluation Protocol and Result Breakdown

Detailed results are omitted to the corresponding section of the arXiv version [29].

K.9 Certified Accuracy Beyond Certified Radii

Detailed results are omitted to the corresponding section of the arXiv version [29].

K.10 Different Smoothing Variance Levels: More Results

Detailed results are omitted to the corresponding section of the arXiv version [29].

K.11 Tightness-Efficiency Trade-Off

Detailed results are omitted to the corresponding section of the arXiv version [29].

Table 7: Detailed smoothing distributions and running time statistics for our **TSS**. $\mathcal{N}(\mu, \Sigma)$ is the normal distribution, $\exp(\lambda)$ is the exponential distribution, $\mathcal{U}([a, b])$ is the uniform distribution. Random variable ε is the elementwise noise as in Corollary 2. "Cert." means certification.

Transformation	Dataset	Attack Radius	Training Distribution	Smoothing Distribution	Avg. Cert. Time per Sample
Gaussian Blur	MNIST	Squared Radius $\alpha \leq 36$	$\alpha \sim \text{Exp}(1/10)$		7.9 s
	CIFAR-10	Squared Radius $\alpha \leq 16$	$\alpha \sim \text{Exp}(1/5)$		30.9 s
	ImageNet	Squared Radius $\alpha \leq 36$	$\alpha \sim \text{Exp}(1/10)$		45.7 s
Translation (Reflection Pad.)	MNIST	$\sqrt{\Delta x^2 + \Delta y^2} \leq 8$	$(\Delta x, \Delta y) \sim \mathcal{N}(0, 10^2 I)$		10.2 s
	CIFAR-10	$\sqrt{\Delta x^2 + \Delta y^2} \leq 20$	$(\Delta x, \Delta y) \sim \mathcal{N}(0, 15^2 I)$		39.4 s
	ImageNet	$\sqrt{\Delta x^2 + \Delta y^2} \leq 100$	$(\Delta x, \Delta y) \sim \mathcal{N}(0, 30^2 I)$		161.9 s
Brightness	MNIST	$b \pm 50\%$	$b \sim \mathcal{N}(0, 0.6^2)$		2.1 s
	CIFAR-10	$b \pm 40\%$	$b \sim \mathcal{N}(0, 0.3^2)$		4.4 s
	ImageNet	$b \pm 40\%$	$b \sim \mathcal{N}(0, 0.4^2)$		45.1 s
Contrast and Brightness	MNIST	$c \pm 50\%, b \pm 50\%$	$(c, b) \sim \mathcal{N}(0, 0.6^2 I)$	$(c, b) \sim \mathcal{N}(0, 1.0^2 I)$	9.8 s
	CIFAR-10	$c \pm 40\%, b \pm 40\%$	$(c, b) \sim \mathcal{N}(0, 0.4^2 I)$	$(c, b) \sim \mathcal{N}(0, 0.6^2 I)$	45.0 s
	ImageNet	$c \pm 40\%, b \pm 40\%$	$(c, b) \sim \mathcal{N}(0, 0.4^2 I)$		325.6 s
Gaussian Blur, Translation, Brightness, and Contrast	MNIST	$\alpha \leq 5, c, b \pm 10\%$, $\sqrt{\Delta x^2 + \Delta y^2} \leq 5$	$\alpha \sim \text{Exp}(1/10)$, $(\Delta x, \Delta y) \sim \mathcal{N}(0, 10^2 I)$, $(c, b) \sim \mathcal{N}(0, 0.3^2 I)$, $\varepsilon \sim \mathcal{N}(0, 0.05^2 I)$	$\alpha \sim \text{Exp}(1/10)$, $(\Delta x, \Delta y) \sim \mathcal{N}(0, 10^2 I)$, $(c, b) \sim \mathcal{N}(0, 0.3^2 I)$	12.9 s
	CIFAR-10	$\alpha \leq 1, c, b \pm 10\%$, $\sqrt{\Delta x^2 + \Delta y^2} \leq 5$	$\alpha \sim \text{Exp}(1)$, $(\Delta x, \Delta y) \sim \mathcal{N}(0, 10^2 I)$, $(c, b) \sim \mathcal{N}(0, 0.3^2 I)$, $\varepsilon \sim \mathcal{N}(0, 0.01^2 I)$	$\alpha \sim \text{Exp}(1)$, $(\Delta x, \Delta y) \sim \mathcal{N}(0, 10^2 I)$, $(c, b) \sim \mathcal{N}(0, 0.3^2 I)$	43.1 s
	ImageNet	$\alpha \leq 10, c, b \pm 20\%$, $\sqrt{\Delta x^2 + \Delta y^2} \leq 10$	$\alpha \sim \text{Exp}(1/5)$, $(\Delta x, \Delta y) \sim \mathcal{N}(0, 20^2 I)$, $(c, b) \sim \mathcal{N}(0, 0.4^2 I)$, $\varepsilon \sim \mathcal{N}(0, 0.01^2 I)$	$\alpha \sim \text{Exp}(1/5)$, $(\Delta x, \Delta y) \sim \mathcal{N}(0, 20^2 I)$, $(c, b) \sim \mathcal{N}(0, 0.4^2 I)$	238.1 s
Rotation	MNIST	$r \pm 50°$	Same as Rotation and Brightness	$\varepsilon \sim \mathcal{N}(0, 0.12^2 I)$	20.1 s
	CIFAR-10	$r \pm 10°$		$\varepsilon \sim \mathcal{N}(0, 0.05^2 I)$	52.8 s
		$r \pm 30°$		$\varepsilon \sim \mathcal{N}(0, 0.05^2 I)$	141.0 s
	ImageNet	$r \pm 30°$		$\varepsilon \sim \mathcal{N}(0, 0.5^2 I)$	2358.1 s
Scaling	MNIST	$s \pm 30\%$	Same as Scaling and Brightness	$\varepsilon \sim \mathcal{N}(0, 0.12^2 I)$	17.7 s
	CIFAR-10	$s \pm 30\%$		$\varepsilon \sim \mathcal{N}(0, 0.12^2 I)$	42.2 s
	ImageNet	$s \pm 30\%$		$\varepsilon \sim \mathcal{N}(0, 0.5^2 I)$	1201.2 s
Rotation and Brightness	MNIST	$r \pm 50°, b \pm 20\%$	$r \sim \mathcal{U}([-55, 55])$, $\varepsilon \sim \mathcal{N}(0, 0.12^2 I)$, $b \sim \mathcal{N}(0, 0.2^2)$	$\varepsilon \sim \mathcal{N}(0, 0.12^2 I)$, $b \sim \mathcal{N}(0, 0.2^2)$	31.4 s
	CIFAR-10	$r \pm 10°, b \pm 10\%$	$r \sim \mathcal{U}([-12.5, 12.5])$, $\varepsilon \sim \mathcal{N}(0, 0.05^2 I)$, $b \sim \mathcal{N}(0, 0.2^2)$	$\varepsilon \sim \mathcal{N}(0, 0.05^2 I)$, $b \sim \mathcal{N}(0, 0.2^2)$	62.3 s
		$r \pm 30°, b \pm 20\%$	$r \sim \mathcal{U}([-35, 35])$, $\varepsilon \sim \mathcal{N}(0, 0.05^2 I)$, $b \sim \mathcal{N}(0, 0.2^2)$	$\varepsilon \sim \mathcal{N}(0, 0.05^2 I)$, $b \sim \mathcal{N}(0, 0.2^2)$	157.0 s
	ImageNet	$r \pm 30°, b \pm 20\%$	$r \sim \mathcal{U}([-35, 35])$, $\varepsilon \sim \mathcal{N}(0, 0.5^2 I)$, $b \sim \mathcal{N}(0, 0.2^2)$	$\varepsilon \sim \mathcal{N}(0, 0.5^2 I)$, $b \sim \mathcal{N}(0, 0.2^2)$	2475.6 s
Scaling and Brightness	MNIST	$s \pm 50\%, b \pm 50\%$	$s \sim \mathcal{U}([0.45, 1.55])$, $\varepsilon \sim \mathcal{N}(0, 0.12^2 I)$, $b \sim \mathcal{N}(0, 0.5^2)$	$\varepsilon \sim \mathcal{N}(0, 0.12^2 I)$, $b \sim \mathcal{N}(0, 0.5^2)$	74.9 s
	CIFAR-10	$s \pm 30\%, b \pm 30\%$	$s \sim \mathcal{U}([0.65, 1.35])$, $\varepsilon \sim \mathcal{N}(0, 0.12^2 I)$, $b \sim \mathcal{N}(0, 0.3^2)$	$\varepsilon \sim \mathcal{N}(0, 0.12^2 I)$, $b \sim \mathcal{N}(0, 0.3^2)$	44.5 s
	ImageNet	$s \pm 30\%, b \pm 30\%$	$s \sim \mathcal{U}([0.65, 1.35])$, $\varepsilon \sim \mathcal{N}(0, 0.5^2 I)$, $b \sim \mathcal{N}(0, 0.3^2)$	$\varepsilon \sim \mathcal{N}(0, 0.5^2 I)$, $b \sim \mathcal{N}(0, 0.3^2)$	1401.6 s
Rotation, Brightness, and ℓ_2	MNIST	$r \pm 50°, b \pm 20\%, \|\delta\|_2 \leq .05$	Same as Rotation and Brightness	Same as Rotation and Brightness	35.1 s
	CIFAR-10	$r \pm 10°, b \pm 10\%, \|\delta\|_2 \leq .05$			132.5 s
		$r \pm 30°, b \pm 20\%, \|\delta\|_2 \leq .05$			520.2 s
	ImageNet	$r \pm 30°, b \pm 20\%, \|\delta\|_2 \leq .05$			3463.8 s
Scaling, Brightness, and ℓ_2	MNIST	$s \pm 50\%, b \pm 50\%, \|\delta\|_2 \leq .05$	Same as Scaling and Brightness	Same as Scaling and Brightness	75.1 s
	CIFAR-10	$s \pm 30\%, b \pm 30\%, \|\delta\|_2 \leq .05$			50.0 s
	ImageNet	$s \pm 30\%, b \pm 30\%, \|\delta\|_2 \leq .05$			1657.7 s