ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240665

General Lipschitz: Certified Robustness Against Resolvable Semantic Transformations via Transformation-Dependent Randomized Smoothing

Dmitrii Korzh [a,b,*], **Mikhail Pautov** [b, a,c], **Olga Tsymboi** [d, e] **and Ivan Oseledets** [b, a]

[a]Skolkovo Institute of Science and Technology, Moscow, Russia
[b]Artificial Intelligence Research Institute, Moscow, Russia
[c] ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia
[d]Moscow Institute of Physics and Technology, Moscow, Russia
[e]Sber AI Lab, Moscow, Russia

Abstract. Randomized smoothing is the state-of-the-art approach to constructing image classifiers that are provably robust against additive adversarial perturbations of bounded magnitude. However, it is more complicated to compute reasonable certificates against semantic transformations (e.g., image blurring, translation, gamma correction) and their compositions. In this work, we propose General Lipschitz (GL), a new flexible framework to certify neural networks against resolvable semantic transformations. Within the framework, we analyze transformation-dependent Lipschitz-continuity of smoothed classifiers w.r.t. transformation parameters and derive corresponding robustness certificates. To assess the effectiveness of the proposed approach, we evaluate it on different image classification datasets against several state-of-the-art certification methods.

1 Introduction

Deep neural networks show remarkable performance in a variety of computer vision tasks. However, they are drastically vulnerable to specific input perturbations (called adversarial attacks) that might be imperceptible to the human eye as it was initially shown in [36, 4]. Namely, suppose that deep neural network $f : \mathbb{R}^n \to [0,1]^C$ maps input images x to class probabilities. Then, given the classification rule $\hat{f}(x) = \arg\max_{i \in Y} f_i(x)$, where $Y = \{1, 2, \ldots, C\}$, it is possible to craft an adversarial perturbation δ of small magnitude such that x and $x + \delta$ are assigned by $\hat{f}$ to different classes. For some applications, such as self-driving cars [39] and identification systems [21, 32], this vulnerability to the small change in the input data is a serious concern.

Recently, many approaches to create adversarial perturbations were proposed, as well as defense techniques to counteract these approaches, causing an attack-defense arms race [1, 43]. This race, however, barely affected neural network applications where the provably correct behavior of models is required. As a result, more research was conducted in the field of certified robustness, where the goal is to provide provable guarantees on the models' behavior under different input transformations. Randomized smoothing [23, 8] is among the most effective and popular approaches used to build provably robust models. Initially developed as a certification tool against norm-bounded additive perturbations, it was later extended to the cases of semantic perturbations [26, 14], such as brightness shift and translations. In the case of additive perturbations, this approach is about replacing the base classifier f with a smoothed one in the following form:

$$g(x) = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0,\sigma^2)} f(x + \varepsilon). \quad (1)$$

As it is shown in prior works [23, 8], if the base classifier predicts well under Gaussian noise applied to input x, then the smoothed one is guaranteed not to change the predicted class label in some vicinity of x. Lately, it was shown [26, 14] that if the input is subjected to semantic transformation with the parameters sampled from a particular distribution, then it is possible to derive the robustness guarantees for a smoothed model against corresponding semantic transformation.

In this work, we focus on resolvable [26] semantic transformations and their compositions and develop a new framework for certified robustness of classifiers under these perturbations. In a nutshell, we approach smoothing from a different angle and show that a smoothed classifier is Lipschitz continuous with respect to parameters of compositions of resolvable transformations.

Our method is derived with no assumptions on semantic transformation and smoothing distribution (except for the resolvability and smoothness, respectively). It provides a constructive numerical procedure for building a certification against a particular transformation. The proposed approach scales to large datasets at the cost of inference of the smoothed model and can be applied for certification against compositions of resolvable semantic transformations.

Our contributions are summarized as follows:

- We propose a universal certification approach against compositions of resolvable transformations based on randomized smoothing. Our method can be applied for certification against any composition of resolvable semantic transformations, in contrast to the previous studies.
- We propose a numerical procedure to verify the smoothed model's robustness with little to no computation overhead.
- We evaluate our method on different datasets and show that it yields state-of-the-art robustness guarantees in the majority of considered experimental setups.

* Corresponding Author. Email: korzh@airi.net.

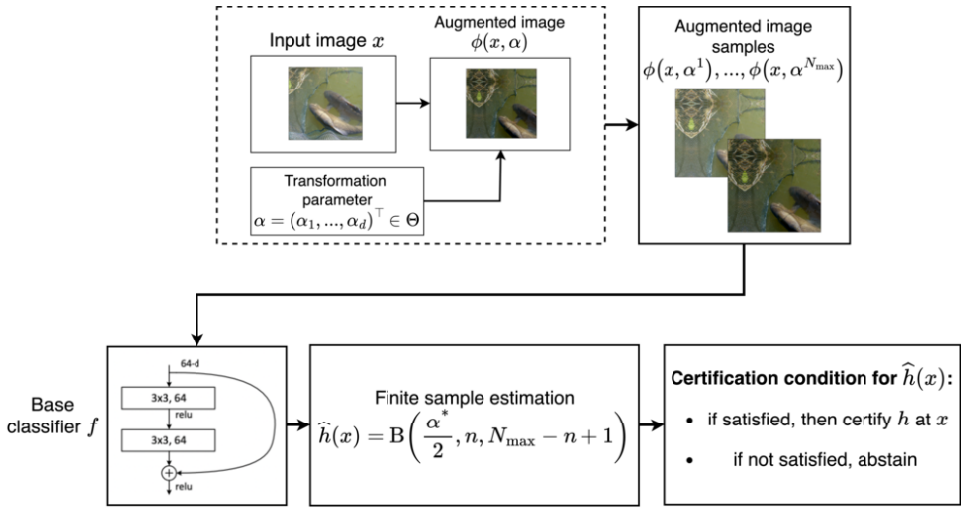


Figure 1: Schematic illustration of the certification procedure: given an input image x of class c and parametric transformation ϕ, we sample $N_{\max}$ augmented images $\{\phi(x,\alpha^1),\dots,\phi(x,\alpha^{N_{\max}})\}$ and compute a lower bound on $h_c(x)$ using Clopper-Pearson test in form $\hat{h}_c(x)=B(\alpha^*/2,n,N_{\max}-n+1)$, where B is Beta distribution and n is the number of augmented images $\phi(x,\alpha^j)$ for which the value of the base classifier $f_c(\phi(x,\alpha^j))>\frac{1}{2}$. Then, the certification condition from Theorem (1) is checked for the value $\hat{h}_c(x)$.

2 Preliminaries

In this work, we focus on the task of image classification. Given $X\subset\mathbb{R}^n$ as the set of input objects and $\mathcal{Y}=\{1,2,\dots C\}$ as the set of classes, the goal is to construct a mapping $\hat{f}:X\to\mathcal{Y}$ assigning a label to each input object. Following the convenient notation, this mapping may be represented as

$$\hat{f}(x)=\underset{i\in\mathcal{Y}}{\operatorname{argmax}}\, f_i(x), \tag{2}$$

where $f:X\to[0,1]^C$ is a classification model and $f_i(x)$ refers to $i-$ th predicted component.

Suppose a parametric mapping $\phi:X\times\Theta\to X$ corresponds to a semantic perturbation of the input of the classification model, where Θ is the space of parameters of the perturbation. The goal of this paper is to construct the framework to certify that a classifier is robust at $x\in X$ to the transformation $\phi(x,\cdot)$ for some set of parameters $\mathcal{B}(\beta_0)$, where $\phi(x,\beta_0)=x$ for all $x\in X$.

A transform $\phi:X\times\Theta\to X$ is called *resolvable* [26] if for any parameter $\alpha\in\Theta$ there exists a continuously differentiable function $\gamma:\Theta\times\Theta\to\Theta$ such that for all $x\in X$ and all $\beta\in\Theta$

$$\phi(\phi(x,\alpha),\beta)=\phi(x,\gamma(\alpha,\beta)). \tag{3}$$

In this work, we analyze the Lipschitz properties of randomized smoothing to certify classification models against compositions of resolvable transformations.

3 Proposed method

This section is devoted to the proposed certification approach and its theoretical analysis.

3.1 Randomized smoothing for semantic transformations

For the given base model $f:X\subset\mathbb{R}^n\to[0,1]^C$, input image $x\in X$, resolvable transformation $\phi:X\times\Theta\to X$ with the resolving function γ as defined in 3, we construct the smoothed classifier $h(x)$ in the form of expectation over perturbation density $\rho(y|x)$ conditioned on the observed sample x:

$$\begin{aligned}h(x)&=\int_\Theta f(\phi(x,\alpha))\rho(\phi(x,\alpha)|x)d\alpha\\&=\int_{\mathbb{R}^n}f(y)\rho(y|x)dy.\end{aligned} \tag{4}$$

The goal of this paper is to present a procedure that guarantees a smoothed model to be robust to semantic perturbations, that is

$$\underset{i\in\{1,2,\dots C\}}{\operatorname{argmax}}\; h_i(x)=\underset{i\in\{1,2,\dots C\}}{\operatorname{argmax}}\; h_i(\phi(x,\beta)), \tag{5}$$

for all $\beta\in\mathcal{B}(\beta_0)$, where $\phi(x,\beta_0)=x$. One way to achieve robustness to parametric perturbation is to bound the Lipschitz constant of the classifier from Eq. (4) with respect to the transformation parameters. For this purpose, the density $\rho(y|x)$ has to be continuously differentiable with respect to perturbation parameters; otherwise, this problem becomes ill-posed. To overcome this issue, we introduce additional Gaussian smoothing.

Assuming that the perturbed sample has the form $\hat{x}=\phi(x,\beta)$, we redefine an auxiliary variable $y=\phi(\hat{x},\alpha)+\varepsilon$, where $\varepsilon\sim\mathcal{N}(0,\sigma^2I_n)$ and the overall conditional probability density $\rho(y|\hat{x})$ in the form:

$$\rho(y|\hat{x})=\frac{\int_\Theta\exp\left\{-\frac{\|y-\phi(\hat{x},\alpha)\|_2^2}{2\sigma^2}\right\}\tau(\alpha)d\alpha}{(2\pi\sigma^2)^{\frac{n}{2}}}, \tag{6}$$

where $\tau(\alpha)$ is the smoothing distribution of the transformation. Following the literature, [14, 26], we sample $\alpha\sim\mathcal{N}(0,\sigma_\alpha^2I_d)$, and then map it to the desired smoothing distribution (see Section 4 for details). Here, $d=\dim(\Theta)$ is the number of transformation parameters.

3.2 Robustness guarantee

In this section, we discuss the main theoretical result. Prior works mainly concentrate on estimating global Lipschitz constants, which may lead to loose guarantees. Instead, we provide certification conditions based on local properties due to using perturbation-dependent smoothing.

Let $x \in X$ be the input object of class c and assume that the smoothed classifier h defined in Eq. (4) correctly classifies x with significant confidence, i.e., $h_c(x) > \frac{1}{2}$. Then, the following result holds.

Theorem 1. *Certification condition.*
Let $\beta(t) : [0,1] \to \Theta$ be a smooth curve such that $\beta(0) = \beta_0$ and $\beta(1) = \beta$. Then there exist mappings $\xi : [0,1] \to \mathbb{R}$ and $\hat{g}(\beta) : \Theta \to \mathbb{R}$ such that if $\hat{g}(\beta) < -\xi(1 - h_c(x)) + \xi(1/2)$, then h is robust at x for all $\beta \in \beta(t)$, where $t \in [0,1]$.

Proof. (Sketch)
Let x be a fixed input of class c. Reassign $h(\beta) = h_c(\beta)$ to manipulate only with $c-$th component of the smoothed classifier that confidently and correctly classifies the ground truth class, namely, let $h_c(\beta) > \frac{1}{2}$.

To construct the certification criteria, we observe that a directional derivative of $h(\beta)$ with respect to β is bounded by the product of two functions, namely $p : [0,1] \to \mathbb{R}$ and $g : \Theta \to \mathbb{R}_{\geq 0}$ such that

$$\langle \nabla_\beta \, h(\beta), u \rangle = \int_{\mathbb{R}^n} f(y) \langle \nabla_\beta \rho(y|\hat{x}), u \rangle dy \leq p(h(\beta)) g(\beta) \quad (7)$$

for all $u : \|u\|_2 = 1$. Note that such $p(\cdot)$ and $g(\cdot)$ exist since $h(\cdot)$ is assumed to be smooth (e.g., $p \equiv 1, g(\beta) \equiv \sup_u \sup_\beta \langle \nabla_\beta h(\beta), u \rangle$).

$$\langle \nabla_\beta h(\beta), u \rangle = \int_{\mathbb{R}^n} f(y) \eta(y, \hat{x}) \rho(y|\hat{x}) dy, \quad (8)$$

where $\eta(y, \hat{x}) = \langle \nabla_\beta \log \rho(y|\hat{x}), u \rangle$ and u is fixed. To estimate the $\tilde{g}(h, \beta) \leq p(h) g(\beta)$ supremum in Eq. (8) given β, we solve an optimization problem with a constraint on current fixed value of h, that change limits of integration. Then, integrating this inequality along a smooth curve $\beta(t) : \beta(0) = \beta_0, \beta(1) = \beta$, we get

$$\int_{\beta(t)} \langle \nabla_\beta h(\beta), u \rangle dt \leq \int_{\beta(t)} p(h) g(\beta) dt. \quad (9)$$

Introducing an auxiliary function

$$\xi(h) = \int \frac{1}{p(h)} dh \quad (10)$$

we get

$$\begin{aligned} \xi(h(\beta)) - \xi(h(\beta_0)) &= \int_{\beta(t)} \langle \nabla_\beta \xi(h(\beta)), u \rangle dt \\ &\leq \int_{\beta(t)} g(\beta) dt = \hat{g}(\beta). \end{aligned} \quad (11)$$

Note that ξ is a monotonically increasing function w.r.t. $h(\beta)$ according to the definition (i.e. $\rho(h) \geq 0$); and $\hat{g}(\beta)$ is non-decreasing along $\beta(t)$ since $g(\beta)$ is non-negative. Assuming that there exists $\beta \in \beta(t)$ such that

$$h_{\neq c}(\beta) > \frac{1}{2}, \quad (12)$$

where $h_{\neq c}(\beta)$ corresponds to the probability of assigning a sample *not* to class c. Finally, using the monotonous property of $\xi(h)$ yields a contradiction, proving the result.

□

Remark 2. The full proof of the theorem is moved to the Appendix [22] so as not to distract the reader. The assumption on $h_c(x) > \frac{1}{2}$ is given to interpret the multiclass classification problem as binary classification (as the one-vs-all setting). The procedure of computing the functions ξ and $\hat{g}$ is described in the numerical evaluation section 3.3. Intuitively, these functions reflect the Lipschitz-continuity of the smoothed classifier w.r.t. transformation parameters.

The theorem states that the smoothed classifier is robust at the point x to transformation ϕ for all parameter values $\beta \in \beta(t)$, if the certification condition is verified for a single parameter value. Note that the certified set of parameters is not necessarily in the ball vicinity of the initial parameter value. This is *the first approach for non-ball-vicinity certification* to our knowledge.

3.3 Numerical evaluation

The Theorem (1) anticipates a numerical procedure to compute certification functions ξ, $\hat{g}$. In this section, we describe this procedure in detail.

Here and below, we assume that the input sample x is fixed and treat smoothed model h as the function of the perturbation parameter β, namely $h(\phi(x, \beta)) \equiv h(\hat{x}) \equiv h(x, \beta) \equiv h(\beta)$ for simplicity. Within our framework, functions ξ, $\hat{g}$ are derived as the ones bounding the smoothed classifier's directional derivative with respect to the perturbation parameter:

$$\langle \nabla_\beta h(\beta), \beta \rangle \leq \tilde{g}(h(\beta), \beta) \leq p(h) g(\beta), \quad (13)$$

where $\tilde{g}(h(\beta), \beta)$ is an upper bound on the directional derivative. This function is also bounded by the product of a function of h and the function of β. If the functions $p(h)$ and $g(\beta)$ are known, the mappings from Theorem 1 have the following form:

$$\xi(h) = \int \frac{1}{p(h)} dh, \quad \hat{g}(\beta) = \int_0^1 g(\beta(t)) dt. \quad (14)$$

It is worth mentioning that the function $\xi(h)$ can be derived analytically, for example, for additive transformations with $\tau(\alpha) \sim \mathcal{N}(0, \kappa^2)$, $d = 1$:

$$\begin{aligned} \log \rho(y \mid \hat{x}) = &- \frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\kappa^2 + \sigma^2) - \\ &\frac{(y - x - \beta)^2}{2(\sigma^2 + \kappa^2)} \end{aligned} \quad (15)$$

$$\eta(y, \hat{x}) = \frac{\partial}{\partial \beta} \log \rho(y \mid \hat{x}) = \frac{y - x - \beta}{\sigma^2 + \kappa^2}, \quad (16)$$

$$\tilde{g}(h, \beta) = \frac{1}{\sqrt{\sigma^2 + \kappa^2}\sqrt{2\pi}} e^{(\mathrm{erf}^{-1}(1-2h))^2} \quad (17)$$

$$\xi(h) = \sqrt{\sigma^2 + \kappa^2} \Phi^{-1}(h), \quad (18)$$

where Φ^{-1} is a standard Gaussian distribution's inverse cumulative density function. This result coincides with the one from [8, 26] if $\sigma = 0$.

3.3.1 Bounding the directional derivative

A bound for the directional derivative in the form from Eq. (13) is used to compute functions $\xi, \hat{g}$ from Theorem (1). However, in the case of a complicated form of conditional density from Eq. (6), it may be unfeasible to construct an exact bound. Instead, we propose to use a numerical procedure to bound directional derivatives of the smoothed model. The gradient of the smoothed classifier with respect to the parameters of transformation has the following form:

$$\begin{aligned} \nabla_\beta h &= \int f(y) \nabla_\beta \rho(y|\hat{x}) dy \\ &= \int f(y) \eta(y, \hat{x}) \rho(y|\hat{x}) dy, \end{aligned} \quad (19)$$

where $\eta(y,\hat{x}) = \nabla_\beta \log \rho(y|\hat{x})$. Given fixed β, the problem of bounding the directional derivative $\langle \nabla_\beta h(\beta), \beta\rangle$ is equivalent to the search of the worst base classifier, i.e., the one with the largest bound. The search for the worst classifier q^* may be formulated as the optimization problem:

$$\begin{aligned} q^* = &\operatorname*{argmax}_{q\in\mathcal{Q}} \int q(y)\eta(y,\hat{x})\rho(y|\hat{x})dy, \\ &\text{s.t.} \quad h(\hat{x}) = \int q(y)\rho(y|\hat{x})dy, \end{aligned} \tag{20}$$

where $\mathcal{Q} = \{q|q : X \to [0,1]\}$ is the set of all binary classifiers. Under the specific choice of resolvable transform ϕ and perturbation distribution $\tau(\alpha)$, the problem from Eq. (20) admits the analytical solution.In general, if the evaluation of $\rho(y|\hat{x})$ and $\eta(y,\hat{x})$ are available, the solution of the problem from Eq. (20) could be obtained numerically.

Namely, suppose that $M \in \mathbb{N}$ is the number of independent and identically distributed variables $(q_1,\eta_1),\dots,(q_M,\eta_M) \sim \rho(y|\hat{x}) \times \eta(y,\hat{x})$. Then, the functional and constraint (respectively) from Eq. (20) can be approximated in the following form:

$$\begin{aligned} &\int q(y)\eta(y,\hat{x})\rho(y|\hat{x})dy \approx \frac{1}{M}\sum_{k=1}^{M} q_k\eta_k, \\ &\int q(y)\rho(y|\hat{x})dy \approx \frac{1}{M}\sum_{k=1}^{M} q_k. \end{aligned} \tag{21}$$

An approximate solution to Eq. (20) is then obtained by sorting $\eta_{i_1} \geq \dots \geq \eta_{i_M}$ and assigning

$$\begin{cases} q_{i_1} = q_{i_2} = \dots = q_{i_k} = 1, \\ q_{i_{k+1}} = q_{i_{k+2}} = \dots = q_M = 0. \end{cases} \tag{22}$$

In the Eq. (22), threshold index k is chosen such that

$$\begin{cases} |h(\hat{x}) - S_k| \leq |h(\hat{x}) - S_{k-1}|, \\ |h(\hat{x}) - S_k| \leq |h(\hat{x}) - S_{k+1}|, \end{cases} \tag{23}$$

where $S_k = \frac{1}{M}\sum_{j=1}^{k} q_{i_j}$.

For sufficiently large M, this scheme yields a tight approximation of q^* from Eq. (20), and, hence, for the bound $w(h,\beta)$ from Eq. (13).

3.3.2 Density Estimation

Since the exact evaluation of the density from Eq. (6) is challenging, we emulate sampling from the conditional density $\rho(y|\hat{x})$ by estimating the gradient of the log-density $\eta(y,\hat{x})$ from Eq. (19).

Namely, we use the first-order approximation for the resolvable transform:

$$\begin{aligned} \phi(\hat{x},\alpha) = &\phi(\hat{x},\alpha_0) + J(\alpha_0)(\alpha-\alpha_0) + \\ &+ \mathcal{O}(\|\alpha-\alpha_0\|^2), \end{aligned} \tag{24}$$

where $J(\alpha) = \frac{\partial\phi}{\partial\alpha}$ to establish Laplace's posterior log-density estimation:

$$\begin{aligned} \log\rho(y|\hat{x}) \approx &\log C - \frac{\|\mu\|^2}{2\sigma^2} + \langle M\alpha_0,\alpha_0\rangle - \\ &- \frac{1}{2}\log\det M, \end{aligned} \tag{25}$$

where $M = J^T J + \sigma^2 I$ and $\mu = y - \phi(\hat{x},\alpha) + J\alpha$ and C is a constant. Finally, an approximation for the initial point α_0 from the Eq. (25) is given via one iteration of the Gauss-Newton method [5]:

$$\alpha_0 \approx \left(J^\top J + \sigma^2 I\right)^{-1}(y - \phi(\hat{x},\alpha) + J\alpha). \tag{26}$$

While the above derivation admits an arbitrary parametric transform ϕ, for the resolvable one, there exists a closed-form limit when $\sigma \to 0$. The last is summarized in the Lemma 1, allowing us to compute log-density either analytically or through automatic differentiation tools.

Lemma 1. *Let $\gamma(\alpha,\beta)$ be the resolving function: $\phi(\phi(x,\beta),\alpha) = \phi(x,\gamma(\alpha,\beta))$. Then, the formula for the logarithm of the conditional density from Eq. (6) has the limit when $\sigma \to 0$ in the form*

$$\log\rho(y|\hat{x}) = -\frac{1}{2}\log\det J^\top J + \log\tau(\alpha), \quad J = \frac{\partial\phi}{\partial\alpha}. \tag{27}$$

If only the log-density $\log\rho(y|\hat{x})$ is known, the expression for $\eta(y,\hat{x}) = \nabla_\beta\log\rho(y|\hat{x})$ is given by the following lemma:

Lemma 2 (Gradient of log-density for resolvable transformations)**.** *Suppose that the log-density $\log\rho(y|\hat{x}) = z(\alpha,\beta) = z(\alpha(\beta),\beta)$ is known. Then*

$$\eta(y|\hat{x}) = \nabla_\beta z = \frac{\partial z}{\partial\beta} - \frac{\partial z}{\partial\alpha}\left(\frac{\partial\gamma}{\partial\alpha}\right)^\dagger\frac{\partial\gamma}{\partial\beta},$$

where γ is a resolving function of the transform: $\phi(\phi(x,\beta),\alpha) = \phi(x,\gamma(\alpha,\beta))$.

Remark 3. Proofs of the lemmas are moved to the Appendix [22] so as not to distract the reader.

The overall procedure is presented in Algorithms 1, 2.

Algorithm 1 Numerical Estimation of ξ and $\hat{g}$ for the Resolvable Transform ϕ

Require: ϕ – resolvable input transformation,
N_s – number of samples for bound estimation,
γ – resolving function of ϕ,
β_0 – identity parameters of ϕ,
α – smoothing parameter,
B – parametric grid of d_b points to estimate bounds on,
d – number of parameters of the transform,
x – a random input point.
Ensure: $\xi(h), \hat{g}(\beta)$ – functions from Theorem 1.
$\{p_i\}_{i=0}^{N_s}, \{g_j\}_{j=0}^{d_b} \leftarrow$
$\leftarrow$ ComputeNormedBounds$(\gamma,\beta_0,B,N_s,\phi,d,x)$
$\{\xi_i\}_{i=0}^{N_s} \leftarrow \frac{1}{N_s}$CumulativeSummation$\left(p_i^{-1}\right)$
$\xi(h) \leftarrow$ Interpolate(linspace$(0,1,N_s),\xi_i)$
$z \leftarrow$ Interpolate $\left(B,\{g_j\}_{j=0}^{d_b}\right)$
for $\beta_j \in B$ **do**
  $\hat{g}_j = \int_0^1 z((1-t)\beta_0 + t\beta_j)dt$
end for
$\hat{g}(\beta) =$ Interpolate$(B,\{\hat{g}_j\})$
return $\xi(h), \hat{g}(\beta)$

4 Experiments

We conducted experiments with different ResNet architectures models on ImageneNet, CIFAR-10, and CIFAR-100 datasets[1]. The models

[1] Our code is publicly available on github.com/dkorzh10/general_lipschitz.

Table 1: Quantitative results on ImageNet dataset. We report smoothing distributions and certified robust accuracy for our approach and competitors' methods. The best results are highlighted in **bold**, underlined denotes equivalent performance. Symbol $''-''$ in the table corresponds to the transformation in which a method does not certify the model against the given distribution parameters. We evaluate the CRA in the fixed parameter range $R_l \leq \beta \leq R_r$ for each transformation type. In the parameter column, $c,\ b,\ \gamma,\ (T_x, T_y),\ r_b$ represent contrast, brightness, gamma-correction, translations, and Gaussian blur attacks' parameters, respectively. CRA-TSS, CRA-MP, and CRA-GS correspond to the certified accuracy of the methods from [26], [30], and [14] respectively. The architecture of the base model is Resnet-50.

Transform	β	R_l	R_r	Distribution	CRA (ours)	CRA-TSS	CRA-MP	CRA-GS
Brightness	b	-0.4	0.4	$\mathcal{N}(0, 0.3)$	**0.69**	0.68	–	0.67
Contrast	c	0.6	1.4	$\mathrm{LogNorm}(0, 0.3)$	0.68	–	0.68	0.67
Blur	r_b	1	4	$\mathrm{Exp}(0.3)$	0.59	0.59	–	0.0
Translation	T_x, T_y	-56	56	$\mathcal{N}(0, 50)$	**0.49**	0.28	–	0.45
Gamma	γ	1.0	2.0	$\mathrm{Rayleigh}(0.1)$	**0.66**	–	0.54	–
Gamma	γ	0.5	1.0	$\mathrm{Rayleigh}(0.1)$	**0.66**	–	0.61	–
Contrast Brightness	c b	0.6 -0.4	1.4 0.4	LogNorm(0, 0.6) $\mathcal{N}(0, 0.6)$	0.62	0.59	–	0.62
Gamma Contrast	γ c	0.8 0.6	1.4 2.0	$\mathrm{Rayleigh}(0, 0.1)$ $\mathrm{LogNorm}(0, 0.1)$	0.62	–	–	–
Brightness Translation	b T_x, T_y	-0.2 -56	0.2 56	$\mathcal{N}(0, 0.4)$ $\mathcal{N}(0, 30)$	**0.46**	0.02	–	–
Contrast Translation	c T_x, T_y	0.8 -25	1.2 25	$\mathrm{LogNorm}(0, 0.4)$ $\mathcal{N}(0, 30)$	0.09	–	–	–
Contrast Brightness Translation	c b T_x, T_y	0.8 -0.2 -15	1.2 0.2 15	$\mathrm{LogNorm}(0, 0.4)$ $\mathcal{N}(0, 04)$ $\mathcal{N}(0, 15)$	0.06	–	–	–
Translation Blur Brightness Contrast	T_x, T_y r_b b c	-3 1 -0.1 0.95	3 3 0.1 1.05	$\mathcal{N}(0, 10)$ $\mathrm{Rayleigh}(1)$ $\mathcal{N}(0, 0.3)$ $\mathrm{LogNorm}(0, 0.3)$	0.20	–	–	–

were modified with an additional normalization layer as in [8, 26]. For a fixed type of semantic transformation ϕ, we train base classifier f with corresponding augmentation with the parameters sampled from the distribution mentioned in Table 1 to make the base classifier f more empirically robust to this type of transformation. Depending on the transformation type, fine-tuning the pre-trained ImageNet models with augmentations takes from 3 to 18 hours on a 1 Nvidia V-100 16GB GPU. The combination of the cross-entropy and consistency losses from [15] is chosen as an optimization objective, and the fine-tuning is conducted for 2 epochs using SGD (with the learning rate of 10^{-3} and momentum of 0.95).

To evaluate our approach, we compute *certified robust accuracy* (CRA) of the smoothed classifier. Certified robust accuracy is a fraction of correctly predicted images x_i from the test set on which the certification condition is met.

CRA is evaluated on 500 images sampled randomly from the test dataset. To estimate the prediction of the smoothed classifier h, we compute the lower bound of the Clopper-Pearson confidence interval [7] over the sample size $N_{\max} = 1000$ and confidence level $\alpha^* = 10^{-3}$ for each initial image x. To estimate mappings $\xi, \hat{g}$, we sample parameters α from the Gaussian distribution and map them to the desired distribution (see Table 1) using the numerical scheme. In our experiments, we sample parameters of additive transformations from Normal distribution, multiplicative transformations – from Log-Normal and Rayleigh distributions, and exponential transformations – from Rayleigh distribution. In our approach, the certification procedure is sample-agnostic: it has to be done only once for a pair "base network – input transform". Then, the certification in a new sample is done at the cost of one forward pass of the smoothed network. To estimate the computational complexity of the proposed method, we report the time required for the certification in Table 2. Visualization of the results of the certification is presented in Figure 2.

Table 2: Computation time in seconds for the certification, ImageNet dataset. We use $N_{max} = 1000$ samples for smoothing. We report construction (constr) and certification (cert) time for our approach. We measured the average time required to certify 500 images for our method, TSS [26] and MP [30].

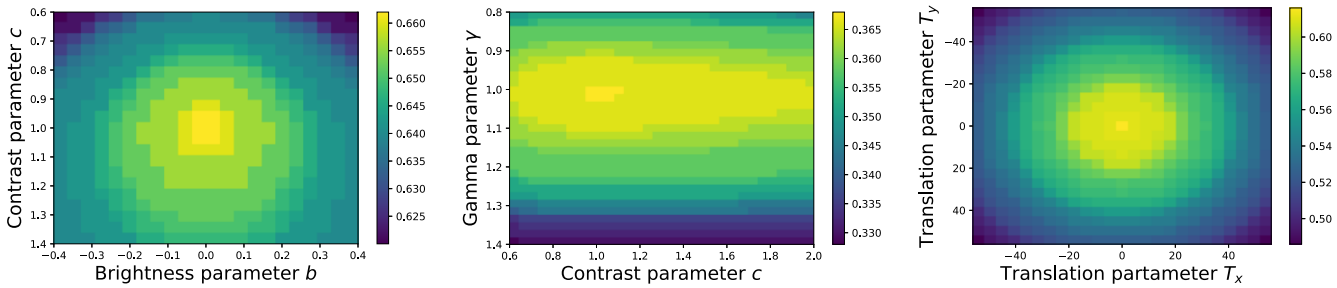
Transform	Ours constr	Ours cert	TSS	MP
Contrast Brightness	170	1350	1675	–
Brightness Translation	200	2905	1500	–
Translation	33.2	1452	1505	–
Gamma	3.4	1450	–	1470

We evaluated our approach against [26, 30, 14] and present the results in Tables 1 and 3. Our method achieves state-of-the-art robustness certificates for the majority of transformations, such as Gamma-Contrast and Contrast-Translation.

It is worth mentioning that for some compositions of transformations (namely, for ones involving both contrast adjustment in a wide range and image translations), the resulting classifier is "over smoothed" – the estimation of probability $h_c(\hat{x})$ of ground truth class is often less than 0.5. Hence, the necessary condition for our method's certification ($h_c(\hat{x}) > 0.5$) is often not satisfied, leading to underestimated CRA.

5 Limitations

This section is devoted to discussing the limitations of the proposed approach.



(a) Contrast and Brightness (b) Gamma and Contrast (c) Translation

Figure 2: Visualization of certified robust accuracy for the subset of parameter space for different transformations, ImageNet dataset. By design of our approach, if the classifier is certified at the input point x for the parameter value β, it is certified for all parameters $\beta^* \in [\beta_0, \beta]$. The values of CRA are presented in the corresponding color bars. Remark: the certified robust accuracy against the given transform in Table 1 is the infimum of CRAs on the corresponding plot.

Algorithm 2 Compute Normed Bounds

Require: ϕ – resolvable input transformation,
N_s – number of samples for bound estimation,
γ – resolving function of ϕ,
β_0 – identity parameters of ϕ,
α – smoothing parameter,
B – parametric grid of d_b points to estimate bounds on,
d – number of parameters of the transform,
x – a random input point.
Ensure: $\{p_i\}_{i=0}^{N_s}, \{g_j\}_{j=0}^{d_b}$ – point-wise estimation of the gradient bound from Eq. (13).

for $\beta_j \in B$ **do**
  $c_j = \mathcal{N}(0, I_d)$
  $z \leftarrow \textsc{LogRho}(x, \gamma, \beta_j, c_j, \phi, N_s)$ {Equation (25), Lemma 1}
  $\eta \leftarrow \textsc{GradLogRho}(x, \gamma, \beta_j, \alpha, \phi, N_s, z)$ {Equation (25), Lemma 2}
  $t \leftarrow (\beta_j - \beta_0)/\|\beta_j - \beta_0\|$
  $\eta \leftarrow \eta t$
  $\eta \leftarrow \eta - \textsc{Mean}(\eta)$
  $\eta \leftarrow \textsc{Sort}(\eta, \text{reverse})$
  $\text{bound}_{j,:} \leftarrow \textsc{CumulativeSummation}(\eta)/N_s$
  $g_j \leftarrow \max_i(\text{bound}_{j,i})\|\beta_0 - \beta_j\|$
  $\text{bound}_{j,:} \leftarrow \text{bound}_{j,:}/\max_i(\text{bound}_{j,i})$
end for
$p_i = \max_j \text{bound}_{ji}$
return $\{p_i\}_{i=0}^{N_s}, \{g_j\}_{j=0}^{d_b}$

5.1 Non-resolvable transformations

The major limitation of the proposed approach is that it is suitable to certify models only against resolvable perturbations. In the case of non-resolvable transformation, the conditional density from Eq. (6) may not be a continuously differentiable function with respect to the transformation parameter in limit $\sigma \to 0$.

5.2 Probabilistic certification

Recall that our approach is based on the randomized smoothing technique; hence, the certified model can not be evaluated exactly. In our experimental setting, for the sample x of class c, the true value of the smoothed classifier $h_c(x)$ is estimated as the lower bound of the Clopper-Pearson confidence interval [7] over $N_{\max}$ samples for some confidence level α^*. Namely, $\hat{h}(x) = B(\alpha^*/2, n, N_{\max} - n + 1)$,

Table 3: Certified robust accuracy (CRA) for some attacks on CIFAR-10 and CIFAR-100 datasets. The best results are highlighted in **bold**, underlined denotes equivalent performance. For the contrast transform, our method vs. MP has 86.2 vs. 86.2 and 45.6 vs. **46.0** for CIFAR-10 and CIFAR-100, respectively. The architecture of the base model is Resnet-110.

Transform	CIFAR-10			CIFAR-100		
	Ours	TSS	GS	Ours	TSS	GS
Brightness	**86.8**	86.6	85.6	**45.6**	43.8	43.2
Contrast	**86.2**	–	85.6	**45.6**	–	43.2
Blur	74.2	**75.4**	0.0	39.8	**41.8**	0.0
CB	85.5	83.4	85.5	**41.6**	38.0	41.4

where B is Beta distribution, $N_{\max}$ is the sample size and n is the number of perturbations for which $f(\phi(x, \alpha^j)) > \frac{1}{2}$. Thus, our approach produces certificates with probability $p \geq 1 - \alpha^*$, where α^* is the upper bound on the probability to return an overestimated lower bound for the value $h(x)$. For comparison, in our settings, we choose $\alpha^* = 10^{-3}$ and $N_{\max} = 1000$.

5.3 Error Analysis

While certain transformations admit analytical solutions, numerical schemes inherently carry errors. However, attaining the desired precision is feasible by augmenting the sample size and refining the approximation with additional points 3, 4. Considering these figures, it is visible that the impact of varying N_s on the behavior of ξ is relatively inconsequential for sufficiently large values. Similar observations can be made for $\hat{g}$. Assuming a fixed brightness parameter of β_0 and varying the contrast parameter, Figure 4 demonstrates that

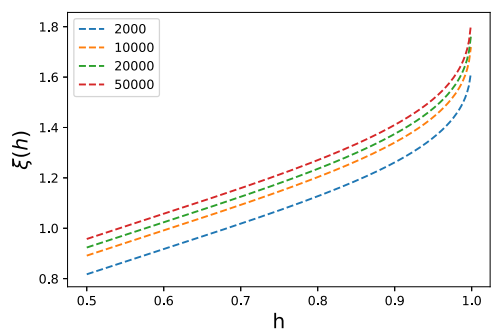


Figure 3: $\xi(h)$ v.s. N_s for the Contrast-Brightness transform

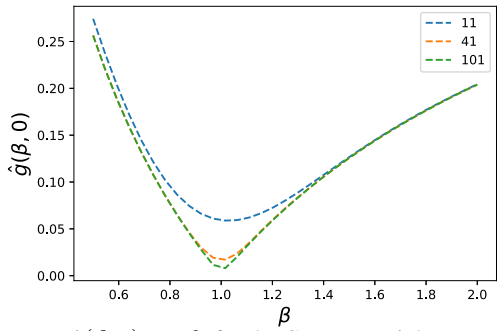


Figure 4: $\hat{g}(\beta, 0)$ v.s. d_b for the Contrast-Brightness transform

while differences exist, they remain insignificant. Analogous results may be anticipated by varying the brightness parameter instead of the contrast.

It is worth mentioning that the estimation of the error of empirical CDF and inverse empirical CDF (22), (22), might be done, for example, by applying Dvoretzky–Kiefer–Wolfowitz inequality.

The error estimation of numerical methods, particularly the more detailed estimation of the error of empirical cumulative distribution function and its inverse derived from numerical methods, is moved to the Appendix [22].

6 Related Work

6.1 *Adversarial attacks and empirical defenses*

The vulnerability of deep learning models to specifically crafted imperceptible additive transformations was discovered in [36, 4]. In [10], the fast gradient sign method was proposed, and the empirical effectiveness of adversarial training against such transformations was discussed. Lately, several ways to exploit this vulnerability in the white-box [29] and black-box [13] settings were proposed. Many applications of adversarial attacks were lately studied, especially in the real-world setting: face recognition [21, 32], detection [20], segmentation [19], and other supervised settings were shown to be affected by this vulnerability. At the same time, empirical defenses [27, 16, 9] against such attacks were proposed. They are mainly narrowed to a specific attack setting or crafting the worst-case examples and including ones in the training set. However, such defenses did not provide enough guarantees to unseen attacks, and consequently, new adaptive attacks were created to overcome previously suggested defenses, causing an arms race.

6.2 *Approaches for provable certification*

Recent research in the field of certified robustness incorporates plenty of verification protocols [25]. New approaches are often introduced during the competition on the verification of neural networks [6]. There are two major approaches to certify classifiers against additive transformations: deterministic [12, 40] and probabilistic [8, 23]. The deterministic approach guarantees that the model is robust at some point x if there is no point from the vicinity of x such that it is classified differently from x.

In contrast, probabilistic approaches are mainly based on randomized smoothing and utilize global [11, 38] or local [42] Lipschitz properties of the smoothed classifier. However, since a smoothed model can not be exactly evaluated, all the robustness guarantees hold with some probability depending on the finite sample estimation of the smoothed model [8]. Randomized smoothing is also applied in different domains, for example, as a defense against text adversarial attacks [45], and automatic speech recognition defense [31]. Solver-based deterministic approaches verify the model's robustness entirely but are not scalable due to computation complexity and usually are restricted to simple architectures [18]. In contrast, linear relaxation approaches do not provide the tightest possible robustness certificates but are model-agnostic and applicable to large datasets [42]. The other deterministic methods are usually based on particular properties of neural networks, such as Lipschitz continuity [24, 38] or curvature of the decision boundary [35]. On the other hand, probabilistic approaches usually utilize an assumption about the smoothness of the model and provide presented state-of-the-art certification results against additive transformations [8, 23].

Semantic transformations are another important class of input perturbations, which fool deep learning models easily [17]. Certified robustness under this threat model is still an open issue [26, 2]. Recently, a few approaches to tackle semantic transformations were proposed that are based on enumeration [34], interval bound propagation [3, 28], and randomized smoothing [26, 14]. It is known that different relaxation approaches provide worse results than the ones based on smoothing [44]. On the other hand, when deterministic guarantees are infeasible, probabilistic approaches [33, 41] to estimate the probability of the model failing when an attack is parameterized provide some insights about the model's robustness. It is worth mentioning that empirical robustness [37] may be improved by incorporating adversarial training, which might be time-consuming on large-scale datasets [37].

A promising way to tackle the certified robustness against semantic perturbations is based on transformation-specific randomized smoothing [26, 14]. The idea of transformation-specific smoothing is to consider the Lipschitz continuity of a smoothed model with respect to the transformation parameters. According to [26], this type of smoothing may be applied to two categories of transformations: *resolvable* and *differentially resolvable*, where the last implies taking into account interpolation errors. However, previous attempts to apply transformation-specific smoothing to certify classifiers against semantic transformations might be infeasible for more complicated transformations [26] or require a surrogate network to represent the transformation and not scalable to large datasets [14]. As a separate application of transformation-dependent smoothing, [30] specifically studied Gamma correction and Contrast change as multiplicative transformations. This work provides asymmetrical guarantees and estimates the certification quality, considering realistic image compression into 8-bit RGB.

7 Conclusion and future work

In this paper, we propose General Lipschitz, a novel framework to certify neural networks against resolvable transformations and their compositions. Based on transformation-dependent randomized smoothing, our approach yields robustness certificates for complex parameterized subsets of parameter space. One of the advantages of the framework is the numerical procedure that produces certificates for a parameter subset by verifying certification condition in a single point of parameter space. Our experimental study shows that the proposed method achieves certified robust accuracy comparable to the state-of-the-art techniques and outperforms them in some experimental settings. Our approach allows us to certify models against resolvable transformations only, so one possible direction for future work is to extend it to the case of differentially resolvable transformations. Another direction is to apply the approach to object detection or segmentation tasks.

Acknowledgements

This work was partially supported by a grant for AI research centres, provided by the Analytical Center in accordance with the subsidy agreement 000000D730321P5Q0002 and the agreement with the ISP RAS dated November 2, 2021 No. 70-2021-00142.

References

[1] N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

[2] M. Alfarra, A. Bibi, N. Khan, P. H. Torr, and B. Ghanem. Deformrs: Certifying input deformations with randomized smoothing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6001–6009, 2022.

[3] M. Balunovic, M. Baader, G. Singh, T. Gehr, and M. Vechev. Certifying geometric robustness of neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[5] Å. Björck. *Numerical methods for least squares problems*. SIAM, 1996.

[6] C. Brix, S. Bak, C. Liu, and T. T. Johnson. The fourth international verification of neural networks competition (vnn-comp 2023): Summary and results. *arXiv preprint arXiv:2312.16760*, 2023.

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

[8] J. Cohen, E. Rosenfeld, and Z. Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

[9] Z. Ge, H. Hu, and T. Zhao. Towards trustworthy nlp: An adversarial robustness enhancement based on perplexity difference. In *ECAI 2023*, pages 803–810. IOS Press, 2023.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[11] H. Gouk, E. Frank, B. Pfahringer, and M. J. Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.

[12] S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, R. Arandjelovic, T. Mann, and P. Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

[13] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pages 2484–2493. PMLR, 2019.

[14] Z. Hao, C. Ying, Y. Dong, H. Su, J. Song, and J. Zhu. Gsmooth: Certified robustness against semantic transformations via generalized randomized smoothing. In *International Conference on Machine Learning*, pages 8465–8483. PMLR, 2022.

[15] J. Jeong and J. Shin. Consistency regularization for certified robustness of smoothed classifiers. *Advances in Neural Information Processing Systems*, 33:10558–10570, 2020.

[16] X. Jia, Y. Zhang, X. Wei, B. Wu, K. Ma, J. Wang, and X. Cao. Prior-guided adversarial initialization for fast adversarial training. In *European Conference on Computer Vision*, pages 567–584. Springer, 2022.

[17] A. Joshi, A. Mukherjee, S. Sarkar, and C. Hegde. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4773–4783, 2019.

[18] G. Katz, D. A. Huang, D. Ibeling, K. Julian, C. Lazarus, R. Lim, P. Shah, S. Thakoor, H. Wu, A. Zeljić, et al. The marabou framework for verification and analysis of deep neural networks. In *Computer Aided Verification: 31st International Conference, CAV 2019*, pages 443–452. Springer, 2019.

[19] S. Kaviani, K. J. Han, and I. Sohn. Adversarial attacks and defenses on ai in medical imaging informatics: A survey. *Expert Systems with Applications*, page 116815, 2022.

[20] E. Kaziakhmedov, K. Kireev, G. Melnikov, M. Pautov, and A. Petiushko. Real-world attack on mtcnn face detection system. In *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, pages 0422–0427. IEEE, 2019.

[21] S. Komkov and A. Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 819–826. IEEE, 2021.

[22] D. Korzh, M. Pautov, O. Tsymboi, and I. Oseledets. General lipschitz: Certified robustness against resolvable semantic transformations via transformation-dependent randomized smoothing. *arXiv preprint arXiv:2309.16710*, 2023.

[23] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy*, pages 656–672. IEEE, 2019.

[24] A. J. Levine and S. Feizi. Improved, deterministic smoothing for l_1 certified robustness. In *International Conference on Machine Learning*, pages 6254–6264. PMLR, 2021.

[25] L. Li, X. Qi, T. Xie, and B. Li. Sok: Certified robustness for deep neural networks. *arXiv preprint arXiv:2009.04131*, 2020.

[26] L. Li, M. Weber, X. Xu, L. Rimanic, B. Kailkhura, T. Xie, C. Zhang, and B. Li. Tss: Transformation-specific smoothing for robustness certification. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 535–557, 2021.

[27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[28] J. Mohapatra, T.-W. Weng, P.-Y. Chen, S. Liu, and L. Daniel. Towards verifying robustness of neural networks against a family of semantic perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–252, 2020.

[29] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[30] N. Muravev and A. Petiushko. Certified robustness via randomized smoothing over multiplicative parameters of input transformations. *arXiv preprint arXiv:2106.14432*, 2021.

[31] R. Olivier and B. Raj. Sequential randomized smoothing for adversarially robust speech recognition. *arXiv preprint arXiv:2112.03000*, 2021.

[32] M. Pautov, G. Melnikov, E. Kaziakhmedov, K. Kireev, and A. Petiushko. On adversarial patches: real-world attack on arcface-100 face recognition system. In *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, pages 0391–0396. IEEE, 2019.

[33] M. Pautov, N. Tursynbek, M. Munkhoeva, N. Muravev, A. Petiushko, and I. Oseledets. Cc-cert: A probabilistic approach to certify general robustness of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7975–7983, 2022.

[34] K. Pei, Y. Cao, J. Yang, and S. Jana. Towards practical verification of machine learning: The case of computer vision systems. *arXiv preprint arXiv:1712.01785*, 2017.

[35] S. Singla and S. Feizi. Second-order provable defenses against adversarial attacks. In *International conference on machine learning*, pages 8981–8991. PMLR, 2020.

[36] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[37] Y.-Y. Tsai, L. Hsiung, P.-Y. Chen, and T.-Y. Ho. Towards compositional adversarial robustness: Generalizing adversarial training to composite semantic perturbations. *arXiv preprint arXiv:2202.04235*, 2022.

[38] Y. Tsuzuku, I. Sato, and M. Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in neural information processing systems*, 31, 2018.

[39] J. Tu, H. Li, X. Yan, M. Ren, Y. Chen, M. Liang, E. Bitar, E. Yumer, and R. Urtasun. Exploring adversarial robustness of multi-sensor perception systems in self driving. *arXiv preprint arXiv:2101.06784*, 2021.

[40] C. Wei and J. Z. Kolter. Certified robustness for deep equilibrium models via interval bound propagation. In *International Conference on Learning Representations*, 2022.

[41] M. Wicker, L. Laurenti, A. Patane, and M. Kwiatkowska. Probabilistic safety for bayesian neural networks. In *Conference on uncertainty in artificial intelligence*, pages 1198–1207. PMLR, 2020.

[42] E. Wong and Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018.

[43] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17:151–178, 2020.

[44] B. Zhang, D. Jiang, D. He, and L. Wang. Rethinking lipschitz neural networks and certified robustness: A boolean function perspective. *Advances in Neural Information Processing Systems*, 35:19398–19413, 2022.

[45] X. Zhang, H. Hong, Y. Hong, P. Huang, B. Wang, Z. Ba, and K. Ren. Text-crs: A generalized certified robustness framework against textual adversarial attacks. *arXiv preprint arXiv:2307.16630*, 2023.