DeformRS: Certifying Input Deformations with Randomized Smoothing

Motasem Alfarra [*1], **Adel Bibi** [*2], **Naeemullah Khan** [2], **Philip H.S. Torr** [2], **Bernard Ghanem** [1]

[1] King Abdullah University of Science and Technology (KAUST), [2] University of Oxford
motasem.alfarra@kaust.edu.sa, adel.bibi@eng.ox.ac.uk

Abstract

Deep neural networks are vulnerable to input deformations in the form of vector fields of pixel displacements and to other parameterized geometric deformations *e.g.* translations, rotations, etc. Current input deformation certification methods either (**i**) do not scale to deep networks on large input datasets, or (**ii**) can only certify a specific class of deformations, *e.g.* only rotations. We reformulate certification in randomized smoothing setting for both general vector field and parameterized deformations and propose DEFORMRS-VF and DEFORMRS-PAR, respectively. Our new formulation scales to large networks on large input datasets. For instance, DEFORMRS-PAR certifies rich deformations, covering translations, rotations, scaling, affine deformations, and other visually aligned deformations such as ones parameterized by Discrete-Cosine-Transform basis. Extensive experiments on MNIST, CIFAR10, and ImageNet show competitive performance of DEFORMRS-PAR achieving a certified accuracy of 39% against perturbed rotations in the set $[-10^\circ, 10^\circ]$ on ImageNet.

Introduction

[1]Deep Neural Networks (DNNs) are susceptible to small additive input perturbations, *i.e.* a DNN that correctly classifies x can be fooled into misclassifying $(x + \delta)$, even when δ is so small that x and $(x + \delta)$ are imperceptibly different (Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015a). Even worse, DNNs were shown to be vulnerable to input deformations (Alaifari, Alberti, and Gauksson 2019) such as input rotations and scaling, where such deformations, unlike additive perturbations, can exist due to a slight change in the physical world. This raises a critical concern especially since DNNs are now deployed in safety critical applications, *e.g.* self-driving cars. To address the nuisance of sensitivity to input deformations, one would ideally seek to train DNNs that are *certifiably* free from such adversaries. While there has been impressive progress towards this goal, *i.e.* certifying input deformations, prior art suffers from the limitation of only being able to certify an individual set of deformations, *e.g.* only rotations or only translations etc., or a small composition set of them (Singh et al. 2019; Balunovic et al. 2019; Mohapatra et al. 2020). Only recently has a certification approach been developed for the richer class of smooth vector fields (general displacement of pixels) (Ruoss et al. 2021). However, all previous approaches require solving a mixed-integer or linear program, thus limiting their applicability to small DNNs on small datasets. On the contrary, the only certification methods that scale to larger networks on large datasets (*e.g.* ImageNet) are based on randomized smoothing (Cohen, Rosenfeld, and Kolter 2019). However, such approaches (Fischer, Baader, and Vechev 2020a; Li et al. 2020), similar to many others, are limited to individual deformations, *e.g.* only translations, or to deformations that ought to be *resolvable* limiting the class of certifiable deformations.

In this paper, we revisit the problem of certifying the parameterization of a general class of input deformations through randomized smoothing. Our approach, dubbed DEFORMRS, is general, and it allows for the certification of vector field *and* parameterized deformations. For the class of parameterized deformations, DEFORMRS certifies general affine deformations that cover translation, rotations, scaling, sheering, etc., and any composition of them. Moreover, we show that if the parameterized deformation is represented by the low frequency components of the Discrete Cosine Transform (DCT), DEFORMRS allows for the certification of a set of visually aligned and plausible deformations. Figure 1 presents several examples of the class of deformations DEFORMRS certifies at scale. Our contributions can be summarized as follows. **(i) DEFORMRS-VF.** We extend the formulation of randomized smoothing from pixel intensities to vector field deformations and derive a certification radius R for the deformation vector field. That is to say, DEFORMRS-VF resists all deformations having a vector field with a norm that is smaller than R. **(ii) DEFORMRS-PAR.** We specialize our analysis for parametrizable deformations and propose DEFORMRS-PAR, which grants certification to popular deformations, *e.g.* translation, rotation, scaling, and any composition subset of them, in addition to the general affine class of deformations. We also specialize DEFORMRS-PAR for the set of deformations parameterized by the low-frequency components of DCT, thus certifying a richer class of visually aligned deformations that were not explored in earlier works. **(iii)** We demonstrate the effectiveness of our proposed approach by conducting extensive experiments on MNIST (LeCun 1998), CIFAR10 (Krizhevsky 2012), and ImageNet

[1] Official Code: https://github.com/MotasemAlfarra/DeformRS.

[*] Denotes equal contribution.

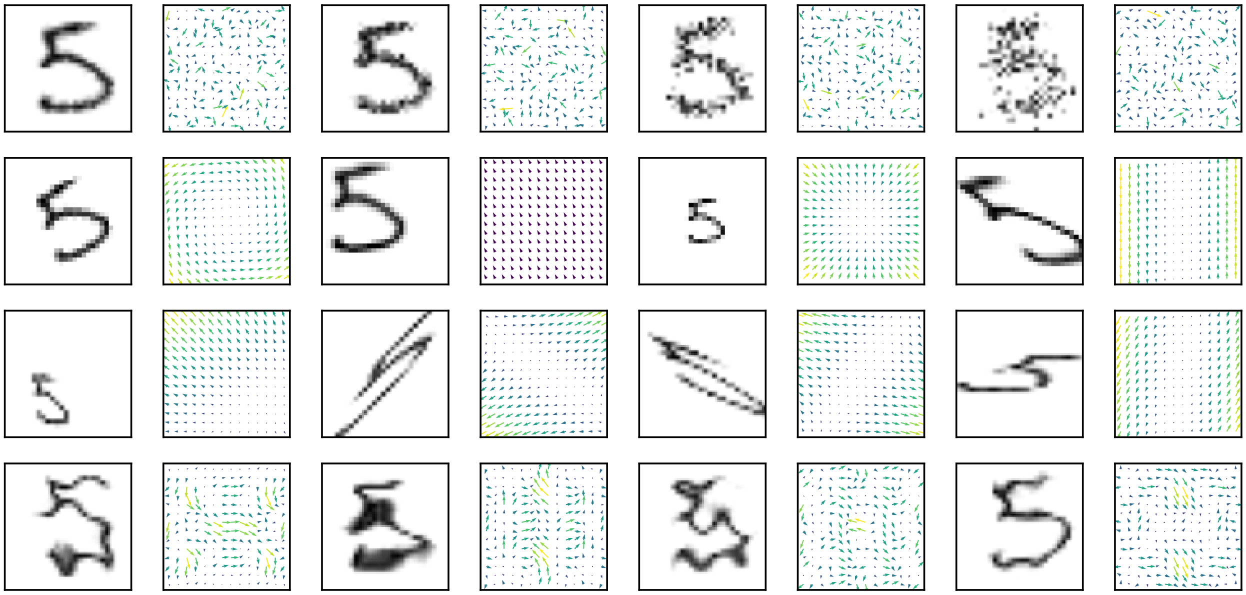

Figure 1: Examples of deformations. We show examples of deformations accompanied with their respective vector fields. First row: Gaussian random deformations. Second row: rotation, translation, scaling, and sheering. Third row: affine deformations. Last row: DCT deformations.

(Russakovsky et al. 2015). DeformRS-VF is capable of providing networks that are certifiably robust against general input deformations. Moreover, DeformRS-Par achieves a certified accuracy of 96.8%, 91.8% and 39% against all rotations in the set $[-30°, 30°]$ for MNIST and $[-10°, 10°]$ on CIFAR10 and ImageNet, respectively. In comparison, a recent work (Mohapatra et al. 2020) achieves a certified accuracy of 21.8% on CIFAR10 under the same rotation perturbation set.

Related Work

Certifying Additive Perturbations. Due to the vulnerability of DNNs to adversarial attacks (Goodfellow, Shlens, and Szegedy 2015b), a stream of work was developed to build models that are certifiable against ℓ_p bounded additive adversaries. This includes methods based on Satisfiability Modulo Theory solvers (Ehlers 2017; Katz et al. 2017; Bunel et al. 2017), interval bound propagation (Gowal et al. 2018), and semi-definite programming (Raghunathan, Steinhardt, and Liang 2018), among many others (Ehlers 2017; Huang et al. 2017). This class of approaches is generally computationally expensive for certifying deeper networks on large dimensional inputs (Tjeng, Xiao, and Tedrake 2019) let alone for using them as part of a training routine (Weng et al. 2018). Recently, randomized smoothing (Lecuyer et al. 2019; Cohen, Rosenfeld, and Kolter 2019) demonstrated to be an effective and scalable approach for probabilistic certification of additive perturbations. Followed by various improvements through incorporating adversarial training (Salman et al. 2019), regularization (Zhai et al. 2020), smoothing distribution optimization (Alfarra et al. 2020; Eiras et al. 2021), randomized smoothing achieved state-of-the-art performance in constructing highly accurate and certifiable networks. Following the favorable properties of randomized smoothing, we leverage it for input deformation certification.

Certifying Image Deformations. In addition to additive input perturbations, DNNs were shown to be susceptible to input deformations. For instance, it was shown that DNNs can be fooled into mispredicting inputs undergoing small imperceptible vector field deformations (pixel displacements) (Alaifari, Alberti, and Gauksson 2019). This was followed by several works that aim to provide empirical evaluation of robustness against such deformations, *e.g.* input translations and rotations, including attacks and defenses (Kanbak, Moosavi-Dezfooli, and Frossard 2017; Wong, Schmidt, and Kolter 2020; Engstrom et al. 2019). Unlike certification of additive input perturbations, certifying input deformations only recently started gaining attention. One of the earliest work performs an abstract interval bound propagation for certification (Singh et al. 2019), which was later followed by a tighter linear program formulation (Balunovic et al. 2019), which certifies geometric transformations such as translation and rotation. Recently, several popular geometric transformations as well as other transformations, such as intensity contrast, were formulated as a piece-wise nonlinear layer (Mohapatra et al. 2020), thus allowing for exact certification based on a tighter formulation of classical ℓ_p certification solvers commonly used for additive perturbations. Moreover, recent work (Ruoss et al. 2021) generated optimal intervals and certify them for general vector fields deformations. However, all previous methods either inherently suffer from scalability limitations, or that they cannot certify a composition of transformations jointly. Alleviating the scalability constraints, randomized smoothing was deployed to certify image transformations that are invariant to interpolation (Levine and Feizi 2019); however, the proposed formulation was restricted to individual transformations like rotation and translation. This was followed by the work of (Li et al. 2020), where networks were verified against individual resolvable transformations by estimating their Lipschitz upper bound. We extend prior art to allow for scalable certification of vector field deformations.

Certifying Deformations with Randomized Smoothing

Background. Randomized smoothing constructs a provably robust classifier $g : \mathbb{R}^n \rightarrow \mathcal{P}(\mathcal{Y})$ from any classifier $f : \mathbb{R}^n \rightarrow \mathcal{P}(\mathcal{Y})$, where $\mathcal{P}(\mathcal{Y})$ is a probability simplex over the set of labels $\mathcal{Y}$. For some distribution $\mathcal{D}$, g is defined as:

$$g(x) = \mathbb{E}_{\epsilon\sim\mathcal{D}}\left[f(x+\epsilon)\right].$$

Suppose that g assigns the class c_A for an input x, we define:

$$p_A = g^{c_A}(x,p) \quad \text{and} \quad p_B = \max_{i\neq c_A} g^i(x,p),$$

where $g^i(x)$ is the i^{th} element of $g(x)$. Then, for Gaussian smoothing, *i.e.* $\mathcal{D} = \mathcal{N}(0,\sigma^2 I)$, g outputs a fixed prediction, *i.e.* $g(x) = g(x+\delta)$, for any perturbation δ satisfying $\|\delta\|_2 \leq \frac{\sigma}{2}(\Phi^{-1}(p_A) - \Phi^{-1}(p_B))$ (Zhai et al. 2020). Here, Φ^{-1} is the inverse CDF of the standard Gaussian. Moreover, for uniform smoothing, *i.e.* $\mathcal{D} = \mathcal{U}[-\lambda,\lambda]^n$, then $g(x) = g(x+\delta)$ for any perturbation δ satisfying $\|\delta\|_1 \leq \lambda(p_A - p_B)$ (Yang et al. 2020). While there has been tremendous progress in robustifying networks against ℓ_p additive attacks, there has been far less progress towards robustness against non-additive perturbations (*e.g.* shadowing, input deformations, etc.).

Threat Model. We focus on the rich class of spatial deformations, *i.e.* perturbations to the pixel coordinates, which cover as a special case translation, rotation, scaling, sheering, etc. Given an input x and a function parametrized by κ that transforms x into x', the threat model aims at finding parameters κ that causes f to mispredict x'. Formally, as proposed earlier (Mohapatra et al. 2020), the threat model solves:

$$\min_{\kappa}\left(f^{c_A}(x') - \max_{c} f^{c\neq c_A}(x')\right) < 0, \quad \text{s.t.} \quad d_\kappa(x,x') < \rho, \tag{1}$$

where $d_\kappa(x,x')$ measures the distance in the parameter space κ (*e.g.* rotation angle). In this setup, the threat model can only access the parameters of the transformation function for a given input x. This important formulation is studied earlier in the literature as it reformulates adversarial attacks to simulators and face recognition systems (*e.g.* attacking pose of a face) (Wu et al. 2020; Hamdi, Mueller, and Ghanem 2020). Here, we leverage randomized smoothing to provide simple general scalable certificates against this threat model.

DEFORMRS-VF: Certifying Vector Fields

Deformations. Let the discrete grid $\Omega^{\mathbb{Z}} \subset \mathbb{Z}^2$, where $\mathbb{Z}$ is the set of integers, represent the domain of images $I : \Omega^{\mathbb{Z}} \rightarrow [0,1]^c$, where c is the number of channels in the image. Then, a domain deformation is defined as $T : \Omega^{\mathbb{Z}} \rightarrow \mathbb{R}^2$, such that for a pixel coordinate $p \in \Omega^{\mathbb{Z}}$, we can write $T(p) = p + v(p)$, where $v : \Omega^{\mathbb{Z}} \rightarrow \mathbb{R}^2$ represents the vector field. Since the deformation T maps pixel coordinates to $\mathbb{R}^2$, one needs to define an associated interpolation function for a deformed image $x \in [0,1]^n$, where $n = c \times |\Omega^{\mathbb{Z}}|$, as $I_T : [0,1]^n \times \mathbb{R}^{2|\Omega^{\mathbb{Z}}|} \rightarrow [0,1]^n$. As such, when $T(p) = p \;\; \forall p \in \Omega^{\mathbb{Z}}$, then we have $I_T(x, T(p)) = x$. For ease of notation, we use p to denote the complete set of the discrete grid $\Omega^{\mathbb{Z}}$. First, we extend the definition of smoothed classifiers to domain deformation smoothed classifiers.

Definition 1. *Given a classifier $f : \mathbb{R}^n \rightarrow \mathcal{P}(\mathcal{Y})$ and an interpolation function $I_T : [0,1]^n \times \mathbb{R}^{2|\Omega^{\mathbb{Z}}|} \rightarrow [0,1]^n$, we define a deformation smoothed classifier as:*

$$\hat{g}(x,p) = \mathbb{E}_{\epsilon\sim\mathcal{D}}\left[f\left(I_T(x, p+\epsilon)\right)\right].$$

Note that contrary to g, which smooths the predictions of f under additive pixel perturbations, $\hat{g}$ smooths predictions of f under pixel coordinate deformations. Similar in spirit to earlier results on randomized smooth for additive perturbations (Cohen, Rosenfeld, and Kolter 2019; Zhai et al. 2020), we can show that $\hat{g}$ is certifiable as per the following Theorem. We leave all proofs to the **Appendix**.

Theorem 1. *Suppose that $\hat{g}$ assigns the class c_A for an input x, i.e. $c_A = arg\max_c \hat{g}^c(x,p)$ with:*

$$p_A = \hat{g}^{c_A}(x,p) \quad \textit{and} \quad p_B = \max_{i\neq c_A} \hat{g}^i(x,p)$$

then $arg\max_c \hat{g}^c(x, p+\psi) = c_A$ for vector field perturbations satisfying:

$$\|\psi\|_1 \leq \lambda\left(p_A - p_B\right) \qquad \textit{for } \mathcal{D} = \mathcal{U}[-\lambda,\lambda],$$
$$\|\psi\|_2 \leq \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right) \qquad \textit{for } \mathcal{D} = \mathcal{N}(0,\sigma^2 I), \tag{2}$$

Theorem 1 states that as long as the ℓ_1 and ℓ_2 norms of the deformation characterized by the vector field ψ are sufficiently small, then $\hat{g}$ enjoys a constant prediction. Note that the ℓ_1 and ℓ_2 certificates are agnostic to the structure of the deformation vector field ψ. That is to say, $\hat{g}$ resists all domain deformations, *e.g.* translation, rotation, scaling, etc., as long as (2) is satisfied. This includes patch level deformations, *i.e.* when ψ is an all zero vector field except for a set of indices representing a patch (*e.g.* a rotation of a patch in the image).

DEFORMRS-PAR: Certifying Parametrizable Deformations

Note that the dimensionality of the deformation vector field ψ is twice (two dimensions of the image) the number of pixel coordinates, *i.e.* $2|\Omega^{\mathbb{Z}}|$, where $|\Omega^{\mathbb{Z}}| = 32 \times 32$ in CIFAR10. As such, the set of deformation vector fields ψ of this large dimensionality satisfying the conditions in (2) might be limited to a set of imperceptible deformations, *i.e.* x and $I_T(x, p+\psi)$ are indistinguishable [2]. However, many popular deformations are parameterized by a much smaller set of parameters. In

[2] Certifying imperceptible deformations is important since adversaries can take this form (Alaifari, Alberti, and Gauksson 2019).

general, consider the deformation $T_\phi(p) = p + v_\phi(p)$, where the dimension of ϕ is much lower than $v_\phi(p)$, where v_ϕ is an element wise function. For example, when the vector field v_ϕ characterizes a translation or a rotation, the parameterization ϕ is of dimensions 2 and 1, respectively. In that regard, we show that a close relative to Theorem 1 also holds for perturbations in the parameters characterizing deformations. We first define a parametric deformation smoothed classifier.

Definition 2. *Given a classifier $f : \mathbb{R}^n \to \mathcal{P}(\mathcal{Y})$ and an interpolation function $I_T : [0,1]^n \times \mathbb{R}^{2|\Omega^{\mathbb{Z}}|} \to [0,1]^n$, we define a parametric deformation smoothed classifier as follows:*

$$\tilde{g}_\phi(x,p) = \mathbb{E}_{\epsilon \sim \mathcal{D}} \left[f\left(I_T\left(x, p + v_{\phi+\epsilon}(p)\right)\right)\right].$$

Unlike Definition 1, $\tilde{g}_\phi$ smooths the prediction of f under a specific class of deformations by perturbing the parameterization ϕ. Next, we analyze the robustness of $\tilde{g}_\phi$.

Corollary 1. *Suppose that $\tilde{g}$ assigns the class c_A for an input x, i.e. $c_A = arg \max_c \tilde{g}_\phi(x,p)$ with:*

$$p_A = \tilde{g}_\phi^{c_A}(x,p) \quad \textit{and} \quad p_B = \max_{i \neq c_A} \tilde{g}_\phi^{i}(x,p),$$

then $arg \max_c \tilde{g}_{\phi+\xi}(x,p) = c_A$ for all parametric domain perturbations satisfying:

$$\begin{aligned} \|\xi\|_1 &\leq \lambda \left(p_A - p_B\right) && \textit{for } \mathcal{D} = \mathcal{U}[-\lambda, \lambda], \\ \|\xi\|_2 &\leq \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right) && \textit{for } \mathcal{D} = \mathcal{N}(0, \sigma^2 I). \end{aligned}$$

Corollary 1 specializes the result of Theorem 1 to the family of parametric deformations. It states that as long as the norm of the perturbations to the deformation parameters is sufficiently small, $\tilde{g}_\phi$ enjoys a constant prediction. Next, we show the parametrization of several popular deformations. Let the pixel grid $\Omega^{\mathbb{Z}}$ be the grid of an image of size $N \times M$, where $p_{n,m} = (n,m) \in \Omega^{\mathbb{Z}}$ is a pixel location and $v_\phi(p_{n,m}) = (u_{n,m}, v_{n,m})$ represents the field at $p_{n,m}$.

Translation. Image translation is only parameterized by two parameters $\phi = \{t_v, t_v\}$, namely $v_\phi(p_{n,m}) = (t_u, t_v) \ \forall p \ \forall n,m$ as per Definition 2 and Corollary 1.

Rotation. 2D rotation is only parameterized by the rotation angle $\phi = \{\theta\}$, where $u_{n,m} = n(cos(\theta) - 1) - msin(\theta)$ and $v_{n,m} = nsin(\theta) + m(cos(\theta) - 1)$.

Scaling. Similar to rotation, scaling is parametrized with one parameter; the scaling factor $\phi = \{\alpha\}$, where $u_{n,m} = (\alpha - 1)n$ and $v_{n,m} = (\alpha - 1)m \ \forall n,m$. That is to say, the vector field has the form $v_\alpha(p) = ((\alpha - 1)n, (\alpha - 1)m) \ \forall p$.

Affine. Our formulation for the certification of the parametric family of deformations is general and covers all affine vector fields as special cases. In particular, affine vector fields are parameterized by 6 parameters, namely $\phi = \{a, b, c, d, e, f\}$, where $u_{n,m} = an + bm + e$ and $v_{n,m} = cn + dm + f$. Note that this class naturally covers composite deformations, such as scaling and translation jointly.

Beyond Affine: DCT- Basis. To address deformations beyond affine vector fields, we also consider certifying a class of deformations represented by the Discrete Cosine Transform (DCT) basis. In particular, we consider the low-frequency component truncated DCT of the vector field $u_{n,m}$ and $v_{n,m}$ with a window size of $k \times k$ (as opposed to the complete size of $N \times M$), where the set is characterized by $2k^2$ parameters.

Experiments

We validate the certified performance of DeformRS following Theorem 1 and Corollary 1, respectively. The goal of this section is to show that **(i)** DeformRS-Par improves certified accuracy against individual deformations that are parameterizable, *e.g.* rotation as compared to (Mohapatra et al. 2020) (MOH), in addition to comparisons against several other individual deformations on several datasets. **(ii)** DeformRS-Par can certify the general class of affine deformations allowing for the certification of a composition of deformations, *e.g.* rotation and sheering jointly. **(iii)** DeformRS-Par can certify deformations that are parameterized by truncated DCT coefficients, a more general class of deformations that can represent visually aligned deformations. **(iv)** Following Theorem 1, DeformRS-VF certifies general vector field deformations that are generally imperceptible. Here, we note that while our work directly compares to MOH in terms of setup and formulation (certifying parameter perturbations as per the threat model in Objective 1), we include other geometric certification approaches, *i.e.* (Li et al. 2020) (LI),(Balunovic et al. 2019) (BAL), and (Fischer, Baader, and Vechev 2020b) (FBV), that are not directly comparable to ours due to a different threat model for full completeness.

Setup. We follow standard practices prior art, *e.g.* LI and FBV, and conduct experiments on MNIST (LeCun 1998), CIFAR10 (Krizhevsky 2012), and ImageNet (Russakovsky et al. 2015) datasets. For experiments on MNIST and CIFAR10, we certify ResNet18 (He et al. 2016) trained for 90 epochs with a learning rate of 0.1, momentum of 0.9, weight decay of 10^{-4}, and learning rate decay at epochs 30 and 60 by a factor of 0.1. For ImageNet experiments, we certify a fine-tuned pretrained ResNet50 for 30 epochs using SGD with a learning rate of 10^{-3} that decays at every 10 epochs by a factor of 0.1. All networks are trained with data augmentation sampled from the respective deformations that are being certified, so as to attain a highly accurate base classifier f under such deformations. Following randomized smoothing methods (Salman et al. 2019; Zhai et al. 2020; Alfarra et al. 2020) and using publicly available code (Cohen, Rosenfeld, and Kolter 2019), all our results are certified with 100 Monte Carlo samples for the selection of the top prediction c_A and $100,000$ samples for the estimation of a lower bound to the prediction probability p_A with a failure probability of 0.001. Throughout all experiments, we choose I_T to be a bi-linear interpolation function. Moreover, since image dimensions vary across datasets (square images of sizes 28, 32, 224 for MNIST, CIFAR10 and ImageNet, respectively), we normalize all image dimensions to $[-1,1] \times [-1,1]$. While our certificate has a probabilistic nature, we compare against both mixed integer and linear program based certification methods (BAL, MOH), as well as randomized smoothing

Certification	MNIST			CIFAR-10		
	R(30°)	S(20%)	T$(\lVert\psi\rVert_2 \leq 5)$	R(10°)	S(20%)	T$(\lVert\psi\rVert_2 \leq 5)$
(BAL,MOH)	87.80(BAL)	-	77.00$^{(\mathbf{a})}$(BAL)	87.80 (BAL), 21.80 (MOH)	-	-
(FBV)	72.75$^{(\mathbf{c})}$	-	95.00$^{(\mathbf{d})}$	42.00$^{(\mathbf{b})}$	-	-
(LI)	95.60	96.80	96.80	63.80	58.40	84.80
DeformRS-Par	**96.85**, **96.10**$^{(\mathbf{c})}$	**98.70**	**99.20**	**91.82**	**90.30**	**88.80**

Table 1: Certifying individual deformations on MNIST and CIFAR10. We compare the certified accuracy of DeformRS-Par against (R)otation, (S)caling and (T)ranslation with that of prior art. We define $\lVert\psi\rVert_2 = (t_u^2 + t_v^2)^{1/2}$ for translation. (a) and (d): Certified accuracy at T$(\lVert\psi\rVert_2 \leq 2)$ and T$(\lVert\psi\rVert_2 \leq 2.41)$, respectively; we adopt these settings from BAL and FBV. (b) and (c): Certified accuracy at R(6.79°) and R(38.24°), respectively; we adopt these settings from FBV. Note that DeformRS-Par achieves a higher certified accuracy than (a, b, d) at a higher radius. Best certified accuracies are highlighted in bold.

Certification	ImageNet		
	R(10°)	S (15%)	T$(\lVert\psi\rVert_2 \leq 5)$
(FBV)	17.25$^{(\mathbf{e})}$	-	-
(LI)	33.00	31.00	**63.30**
DeformRS-Par	**39.00**	**42.80**	48.20

Table 2: Certifying individual deformations on ImageNet. We compare the certified accuracy of DeformRS-Par against (R)otation, (S)caling and (T)ranslation with prior art. (e): Certified accuracy at R(1.86°) (adopted from FBV).

based approaches (FBV, LI) for comprehension.

Evaluation metrics. Following prior art (FBV, LI), we use certified accuracy to compare networks. The certified accuracy at a radius R is the percentage of the test set that is both correctly classified and has a certification radius of at least R. Note that R is computed following Corollary 1 for DeformRS-Par and Theorem 1 for DeformRS-VF. We report the Average Certified Radius (ACR) (Zhai et al. 2020).

Compute Power. In all of our training experiments, we used a single NVIDIA 1080-TI for CIFAR10 and MNIST experiments while we used 2 NVIDIA V100 to fine tune ImageNet models. For the certification experiments, we use a single GPU per experiment (NVIDIA 1080-TI for CIFAR10 and MNIST and NVIDIA V100 for ImageNe).

DeformRS-Par - Paramterizable Deformations [3]

Rotation. Rotation deformations are parameterized with a bounded scalar representing the rotation angle $\theta \in [-\pi, \pi]$. Therefore, we use the Uniform smoothing variant of Corollary 1 resulting in a certification of the form $|\theta| \leq \lambda(p_A - p_B)$. We train several networks with $\lambda \in \{\pi/10, 2\pi/10, \dots, \pi\}$, where each trained network is certified with the corresponding λ used in training. We compare the rotation certified accuracy of DeformRS-Par against that of prior work (BAL, FBV, LI, and MOH) on MNIST and CIFAR10 in Table 1 and on ImageNet in Table 2. Following the common practice in randomized smoothing literature (Salman et al. 2019; Zhai et al. 2020), Tables 1 and 2 report the best certified accuracies for DeformRS-Par cross-validated over λ.

In particular, and as shown in Table 1, DeformRS-Par outperforms its best competitor by 1.25% and 4% on MNIST and CIFAR-10 at rotation radii of 30° (*i.e.* R(30°)) and 10° (*i.e.* R(10°)), respectively. Interestingly, on CIFAR10, the certified accuracy of DeformRS-Par at radius 10° is even better than the accuracy of FBV reported at the smaller angle radius of 6.79°. The improvement is consistent on ImageNet, where DeformRS-Par outperforms the randomized smoothing based approach of LI by 6%, as reported in Table 2. Further, we report an improvement of 70% on the certified accuracy on CIFAR10 at radius 10° against MOH that shares the same threat model to our formulation. We believe that DeformRS-Par outperforms mixed-integer and linear program rotation certification methods due to their high computational cost that results in prohibitive explicit training for improved certification (BAL, MOH). We plot in the first column of Figure 2 the certified accuracy of DeformRS-Par over a subset λ used for training and certification. We leave the rest of the ablations of λ to the **Appendix**. We observe that the certified accuracies of DeformRS-Par at the radii reported in the previous tables are indeed insensitive to the choice of λ. Moreover, we note that DeformRS-Par attains a certified accuracy of at least 80% on both MNIST and CIFAR10 at a radius of 100°. In addition, when $\lambda = 90^\circ$ on MNIST, DeformRS-Par attains an ACR of 85°, *i.e.* the average certified rotation is 85°.

Scaling. Scaling deformations are parameterized by $\alpha \geq 0$. Note that a scaling α can either be a zoom-out $(\alpha > 1)$ or a zoom-in $(0 < \alpha < 1)$. For ease, we consider the bounded scaling factor $\alpha - 1$ instead such that $|\alpha - 1| < 0.7$. Thus, an appropriate smoothing distribution in Corollary 1 is uniform with $\lambda \in \{0.1, 0.2, \dots, 0.7\}$ granting a certificate of the form $|\alpha - 1| \leq \lambda(p_A - p_B)$. We report the certified accuracy at the scale factor of 20% (*i.e.* $0.8 \leq \alpha \leq 1.20$) in Table 1 for MNIST and CIFAR10, and at a scale factor of 15% (*i.e.* $0.85 \leq \alpha \leq 1.15$) for ImageNet in Table 2. The best certified accuracy cross validated over λ for DeformRS-Par outperforms its best competitor (LI) by 1.9% on MNIST, 31.9% on CIFAR10, and 11.8% on ImageNet. Moreover, we plot the certified accuracy in the second column of Figure 2 showing the insensitivity of DeformRS-Par to λ. Moreover, DeformRS-Par enjoys a certified accuracy of at least

[3] Certifying deformations lack standard benchmarks and evaluation protocols. This is why there are several superscripts in Tables 1 and 2 as methods report certified accuracies at different radii.

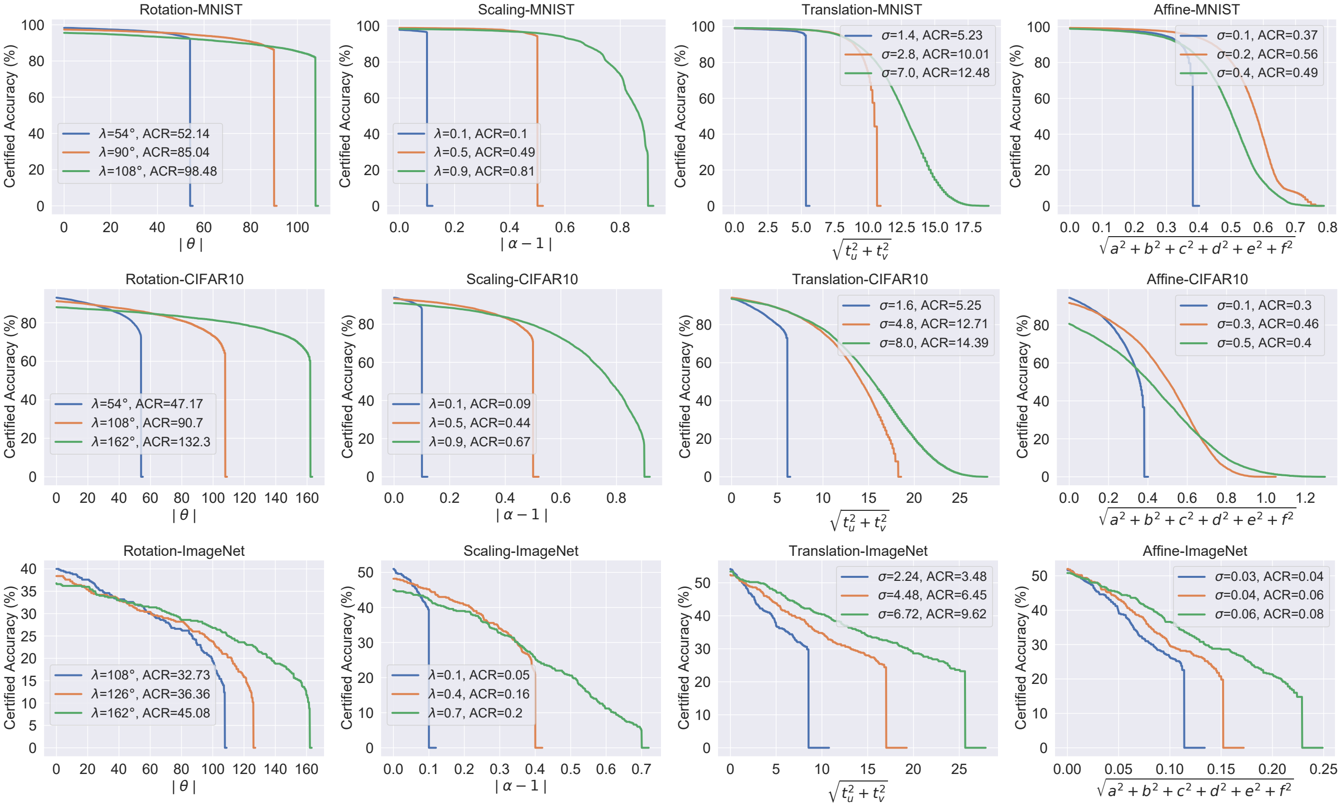

Figure 2: Certified performance of DEFORMRS-PAR. We show the effect of varying the smoothing parameters (λ, σ) on the certified accuracy of DEFORMRS-PAR against rotation, scaling, translation, and affine deformations.

(90%, 80%, 40%) at the larger scaling factors of (0.5, 0.4, 0.2) on MNIST, CIFAR10 and ImageNet, respectively.

Translation. Translation deformations are parameterized by two parameters (t_u, t_v) that can generally be of any value. Thus, we employ two dimensional Gaussian smoothing as per Corollary 1, where $\sigma \in \{0.1, 0.2, \ldots, 0.5\}$ for MNIST and CIFAR10, and $\sigma \in \{0.02, 0.03, \ldots, 0.06\}$ for ImageNet. In this case, the granted certificate is of the form $(t_u^2 + t_v^2)^{1/2} \leq {}^{\sigma}/_{2}(\Phi^{-1}(p_A) - \Phi^{-1}(p_B))$. We compare against BAL, MOH, and LI and report the certified accuracy at a certification radius of at most 5 pixels. [4] As observed from Table 1, DEFORMRS-PAR outperforms its best competitor by 2% and 4% on MNIST and CIFAR10, respectively. However, we observe that DEFORMRS-PAR underperforms on ImageNet attaining 48.2% certified accuracy compared to 63.3% by LI as reported in Table 2. We believe that DEFORMRS-PAR performs worse on ImageNet due to the suboptimal training of the base classifier f on ImageNet. This is evident in the third column of Figure 2, which plots the certified accuracy over a range of different radii for several smoothing σ. Note that the certified accuracy of DEFORMRS-PAR is $\sim 52\%$ at radius 0 over all σ. That is to say, the *accuracy* of DEFORMRS-PAR is already worse than the *certified accuracy* at radius 5 reported by LI. However, the certified accuracy of DEFORMRS-PAR for MNIST and CIFAR10 at radii of 7 and 8 pixels are at least 90% and 80% on MNIST and CIFAR10, respectively.

DEFORMRS-PAR against Affine Deformations

Attaining high certified accuracy for individual deformations, as discussed earlier, requires the training of networks for these particular deformations. Thus, we train a *single* DEFORMRS-PAR network against affine deformations, where we certify it against several specializations of the affine certificate. Recall that the affine deformation is parameterized by 6 parameters. Since generally, there are no restrictions on the values of the affine parameters, we use Gaussian smoothing in Corollary 1 to sample $\{a, b, c, d, e, f\}$ with $\sigma \in \{0.1, 0.2, \ldots, 0.5\}$ for MNIST and CIFAR10 and $\sigma \in \{0.02, 0.03, \ldots, 0.06\}$ on ImageNet. The certificate is thus in the form $\sqrt{a^2 + b^2 + c^2 + d^2 + e^2 + f^2} \leq {}^{\sigma}/_{2}(\Phi^{-1}(p_A) - \Phi^{-1}(p_B))$. The last column of Figure 2 summarizes the certified accuracy of DEFORMRS-PAR on all three datasets. Note, the certified accuracy of DEFORMRS-PAR at affine radius of 0.3 on MNIST is 90%. This is equivalent, under specialization to a translation (*i.e.* $a = b = c = d = 0$), to a certified accuracy of 90% for all translations of radius $0.15 \times 28 = 4.2$ pixels (after unnormalization).

Composition of deformations. We specialize the certifi-

[4] Since the image dimensions in our setting are normalized to $[-1, 1]$, we unnormalize the radius results to pixels in the original image for comparison and ease of interpretation in Figure 2.

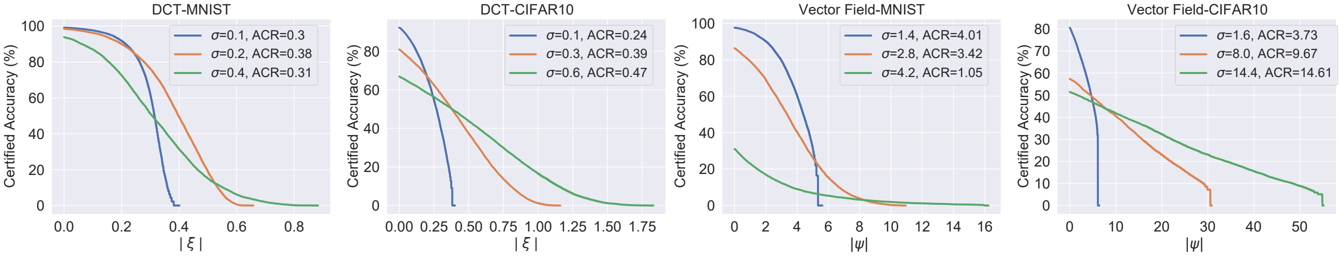

Figure 3: Performance of DeformRS-VF and DeformRS-PAR. We plot the certified accuracy curves of DeformRS-PAR against truncated DCT deformations (left) and DeformRS-VF against general vector field deformations (right).

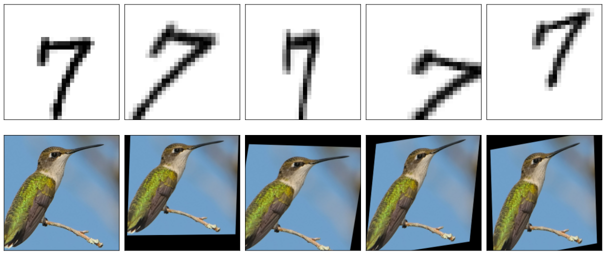

Figure 4: Examples of certified affine deformations. We sample affine parameters satisfying the certification inequality.

cate to a composition of several deformations and compare against the only work certifying deformation compositions (BAL). Following BAL, we consider the composition of shearing with a factor of s followed by rotation with angle θ. The vector field is given as follows:

$$\begin{pmatrix} u_{n,m} \\ v_{n,m} \end{pmatrix} = \underbrace{\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}}_{\text{rotation}} \underbrace{\begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}}_{\text{shear}} \begin{pmatrix} n \\ m \end{pmatrix} - \begin{pmatrix} n \\ m \end{pmatrix}.$$

Note that this composition can be formulated as an affine deformation with $a = \cos(\theta) - 1$, $b = s\cos(\theta) - \sin(\theta)$, $c = \sin(\theta)$, $d = s\sin(\theta) + \cos(\theta) - 1$, and $e = f = 0$. Therefore, Corollary 1 grants the following certificate $\sqrt{s^2 - 2s\sin(\theta) - 4cos(\theta) + 4} \leq \sigma/2(\Phi^{-1}(p_A) - \Phi^{-1}(p_B))$. We compare against BAL, which achieves a certified accuracy of 54.2% on CIFAR10 under the setting $|\theta| \leq 2^\circ$ and $0 \leq s \leq 2\%$. To compute the certified accuracy of DeformRS-Par, note that the left hand side achieves its maximum of 0.0651 at $\theta^* = -2^\circ$ and $s^* = 0.02$; thus, the certified accuracy of DeformRS-Par is the percentage of the test set classified correctly with a radius of at least 0.0651. DeformRS-Par achieves a certified accuracy of 91.28% on CIAFR10 and 43.6% on ImageNet as per the last column in Figure 2, thus outperforming BAL by 37%. Note that, our affine certification allows for the seamless certification of all considered deformations in the literature. This surpasses any need to specialize a certificate for every deformation family of an affine nature. In fact, with a single network trained with DeformRS-Par against affine deformations, we achieve non-trivial certified accuracies against several specialized deformations. Moreover, we consider certifying the same DeformRS-Par network under the composition of a rotation of angle θ, followed by scaling by a factor α, and a translation of parameters (t_u, t_v). The vector field is given as follows:

$$\begin{pmatrix} u_{n,m} \\ v_{n,m} \end{pmatrix} = \underbrace{\begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}}_{\text{scaling}} \underbrace{\begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}}_{\text{rotation}} \begin{pmatrix} n \\ m \end{pmatrix} + \underbrace{\begin{pmatrix} t_u \\ t_v \end{pmatrix}}_{\text{translation}} - \begin{pmatrix} n \\ m \end{pmatrix}.$$

Under such a setting, we have $a = \alpha\cos(\theta) - 1$, $b = -\alpha\sin(\theta)$, $c = \alpha\sin(\theta)$, $d = \alpha\cos(\theta) - 1$, $e = t_u$, and $f = t_v$. Therefore, this grants the following certificate $\sqrt{2 + 2\alpha^2 - 4\alpha\cos(\theta) + t_u^2 + t_v^2} \leq \sigma/2(\Phi^{-1}(p_A) - \Phi^{-1}(p_B))$. We consider certifying DeformRS-Par under the composite deformation of $|\theta| \leq 10^\circ$, $0.8 \leq \alpha \leq 1.2$, and $t_u^2 + t_v^2 \leq 0.1$, where 0.1 corresponds to a radius of 4 and 5 pixels of translation for images in MNIST and CIFAR10, respectively. To that end, we observe that the left hand side of the certificate attains a maximum of 0.503, at which DeformRS-Par enjoys a certified accuracy of 79.78% on MNIST and 50.41% on CIFAR10 as per the last column in Figure 2. To the best of our knowledge, this work is the first to consider such a composite deformation. In Figure 4, we sample several certifiable affine deformations that satisfy the certificate inequality and apply them to MNIST and ImageNet images. We can observe the richness of the certifiable affine maps in both datasets.

DeformRS-Par - Truncated DCT Deformations

We go beyond affine deformations in this section to cover parameterized truncated DCT deformations; particularly, the class of vector field deformations generated by taking the inverse DCT transform of a truncated window of size $k \times k \times 2$. We observe that this class of deformations can generate visually aligned deformations, which are generally not affine, as shown in Figure 1. Since the $k \times k \times 2$ DCT coefficients can take any values, we use Gaussian smoothing with $\sigma \in \{0.1, 0.2, \ldots, 0.5\}$ as per Corollary 1. This grants a certificate of the form $\|\xi\|_2 \leq \sigma/2(\Phi^{-1}(p_A) - \Phi^{-1}(p_B))$, where ξ is the perturbation in the DCT coefficients. For simplicity, we set $k = 2$ for all the experiments in this section. As

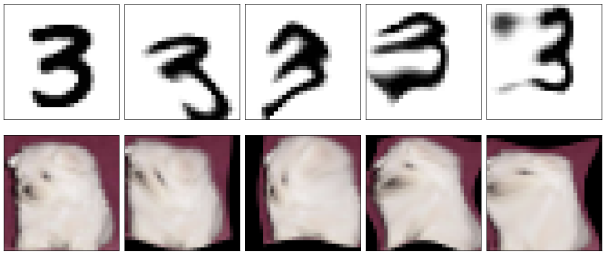

Figure 5: Examples of certified truncated DCT. We sample DCT coefficients satisfying the certification inequality.

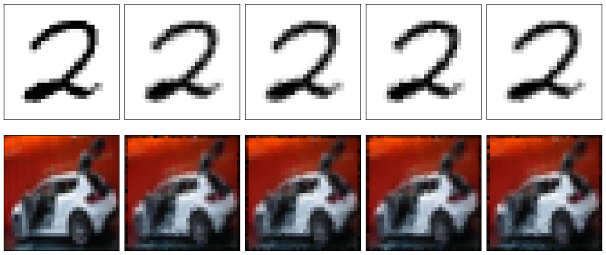

Figure 6: Examples of certified vector field deformations. We sample vector fields satisfying the certification inequality of DeformRS-VF.

per Figure 3, DEFORMRS-PAR certifies perturbations in the DCT coefficients with a certified accuracy of 90% and 80% at radius 0.2 on MNIST and CIFAR10, respectively. Unlike individual deformations or their compositions, it is generally difficult to interpret the certified class of DCT deformations; we instead visualise samples from the certified region of DCT coefficients in Figure 5. We observe interesting certified deformations that are visually aligned resembling different hand written digits in MNIST or ripples in CIFAR10.

DEFORMRS-VF - Vector Field Deformations

We leverage Theorem 1 to certify against a general vector field deformation ψ. Note that such vector fields are in general of size $N \times M \times 2$ and can take any values. Thus, Gaussian smoothing is an appropriate choice, where we set $\sigma \in \{0.1, 0.2, \dots, 0.5\}$. We plot in the second row of Figure 3 the certified accuracy of DEFORMRS-VF for an unnormalized vector field. We observe that DEFORMRS-VF achieves a certified accuracy of 90% and 60% at a radius of 2 pixels on MNIST and CIFAR10, respectively. Note that while vector field deformations can specialize to all previously considered deformations as special cases (*e.g.* rotations), they suffer from the curse of dimensionality (ψ is of size $2NM$) granting certification to only imperceptible deformations (as shown in Figure 6). For instance, consider the vector field generated from a parameterized translation such that $(t_u^2 + t_v^2)^{1/2} \leq 2$. The corresponding vector field will have an energy of at most $\sqrt{2MN}$. That is to say, to certify vector field deformations representing translations of 2 pixels in ℓ_2, the certification radius of the vector field should be at least $\sqrt{2MN}$, which is significantly larger than the radius 2 with the earlier reported accuracy. This is a classical trade-off between the generality of the deformation family and the imperceptibility of the certifiable deformation. We leave the rest of the experiments for the **Appendix**.

Conclusion

We propose DEFORMRS-VF DEFORMRS-PAR to certifying networks against general vector fields and any parameterizable set of deformations, respectively. The parameterizable set of deformations that are certifiable is rich covering individual deformations, *e.g.* rotations, and several compositions, affine deformations and the deformations characterized by truncated DCT coefficients. Both DEFORMRS-VF DEFORMRS-PAR rely on the basic idea that smoothing, through randomized smoothing, the parameterizable space results in a function that is Lipschitz and thus certifiably robust. We conduct several experiments comparing against prior art under against a specific class of deformations. Moreover, we certify against several other composite deformations, i.e. shear and rotations, that follow directly from parameterizable affine defomration certification of DEFORMRS-PAR.

Acknowledgments

This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. OSR-CRG2019-4033, the UKRI grant: Turing AI Fellowship EP/W002981/1 and EPSRC/MURI grant: EP/N019474/1. We would also like to thank the Royal Academy of Engineering.

References

Alaifari, R.; Alberti, G. S.; and Gauksson, T. 2019. ADef: an iterative algorithm to construct adversarial deformations. *International Conference on Learning Representations (ICLR)*.

Alfarra, M.; Bibi, A.; Torr, P. H. S.; and Ghanem, B. 2020. Data Dependent Randomized Smoothing. arXiv:2012.04351.

Balunovic, M.; Baader, M.; Singh, G.; Gehr, T.; and Vechev, M. 2019. Certifying Geometric Robustness of Neural Networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Bunel, R.; Turkaslan, I.; Torr, P. H.; Kohli, P.; and Kumar, M. P. 2017. A unified view of piecewise linear neural network verification. *arXiv preprint arXiv:1711.00455*.

Cohen, J. M.; Rosenfeld, E.; and Kolter, J. Z. 2019. Certified adversarial robustness via randomized smoothing. *International Conference on Machine Learning (ICML)*.

Ehlers, R. 2017. Formal verification of piece-wise linear feed-forward neural networks. In *International Symposium on Automated Technology for Verification and Analysis*.

Eiras, F.; Alfarra, M.; Kumar, M. P.; Torr, P. H. S.; Dokania, P. K.; Ghanem, B.; and Bibi, A. 2021. ANCER: Anisotropic Certification via Sample-wise Volume Maximization. *CoRR*, abs/2107.04570.

Engstrom, L.; Tran, B.; Tsipras, D.; Schmidt, L.; and Madry, A. 2019. Exploring the landscape of spatial robustness. In

International Conference on Machine Learning, 1802–1811. PMLR.

Fischer, M.; Baader, M.; and Vechev, M. 2020a. Certified Defense to Image Transformations via Randomized Smoothing. *Advances in Neural Information Processing Systems (NeurIPS)*.

Fischer, M.; Baader, M.; and Vechev, M. 2020b. Statistical Verification of General Perturbations by Gaussian Smoothing. *OpenReview*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015a. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015b. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572.

Gowal, S.; Dvijotham, K.; Stanforth, R.; Bunel, R.; Qin, C.; Uesato, J.; Arandjelovic, R.; Mann, T. A.; and Kohli, P. 2018. On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models. *CoRR*, abs/1810.12715.

Hamdi, A.; Mueller, M.; and Ghanem, B. 2020. SADA: Semantic Adversarial Diagnostic Attacks for Autonomous Applications. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07): 10901–10908.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Huang, X.; Kwiatkowska, M.; Wang, S.; and Wu, M. 2017. Safety verification of deep neural networks. In *International Conference on Computer Aided Verification*.

Kanbak, C.; Moosavi-Dezfooli, S.-M.; and Frossard, P. 2017. Geometric robustness of deep networks: analysis and improvement. arXiv:1711.09115.

Katz, G.; Barrett, C.; Dill, D. L.; Julian, K.; and Kochenderfer, M. J. 2017. Reluplex: An efficient SMT solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, 97–117. Springer.

Krizhevsky, A. 2012. Learning Multiple Layers of Features from Tiny Images. *University of Toronto*.

LeCun, Y. 1998. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Lecuyer, M.; Atlidakis, V.; Geambasu, R.; Hsu, D.; and Jana, S. 2019. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy (SP)*. IEEE.

Levine, A.; and Feizi, S. 2019. Wasserstein Smoothing: Certified Robustness against Wasserstein Adversarial Attacks. arXiv:1910.10783.

Li, L.; Weber, M.; Xu, X.; Rimanic, L.; Xie, T.; Zhang, C.; and Li, B. 2020. Provable Robust Learning Based on Transformation-Specific Smoothing. arXiv:2002.12398.

Mohapatra, J.; Weng, T.-W.; Chen, P.-Y.; Liu, S.; and Daniel, L. 2020. 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 (CVPR)*.

Raghunathan, A.; Steinhardt, J.; and Liang, P. S. 2018. Semidefinite relaxations for certifying robustness to adversarial examples. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Ruoss, A.; Baader, M.; Balunović, M.; and Vechev, M. 2021. Efficient Certification of Spatial Robustness. arXiv:2009.09318.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*.

Salman, H.; Li, J.; Razenshteyn, I.; Zhang, P.; Zhang, H.; Bubeck, S.; and Yang, G. 2019. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Singh, G.; Gehr, T.; Püschel, M.; and Vechev, M. 2019. An abstract domain for certifying neural networks. *Proceedings of the ACM on Programming Languages*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. arXiv:1312.6199.

Tjeng, V.; Xiao, K.; and Tedrake, R. 2019. Evaluating robustness of neural networks with mixed integer programming. *International Conference on Learning Representations (ICLR)*.

Weng, T.-W.; Zhang, H.; Chen, H.; Song, Z.; Hsieh, C.-J.; Boning, D.; Dhillon, I. S.; and Daniel, L. 2018. Towards fast computation of certified robustness for relu networks. *International Conference on Machine Learning (ICML)*.

Wong, E.; Schmidt, F. R.; and Kolter, J. Z. 2020. Wasserstein Adversarial Examples via Projected Sinkhorn Iterations. arXiv:1902.07906.

Wu, T.; Ning, X.; Li, W.; Huang, R.; Yang, H.; and Wang, Y. 2020. Physical Adversarial Attack on Vehicle Detector in the Carla Simulator. arXiv:2007.16118.

Yang, G.; Duan, T.; Hu, J. E.; Salman, H.; Razenshteyn, I.; and Li, J. 2020. Randomized Smoothing of All Shapes and Sizes. arXiv:2002.08118.

Zhai, R.; Dan, C.; He, D.; Zhang, H.; Gong, B.; Ravikumar, P.; Hsieh, C.-J.; and Wang, L. 2020. Macer: Attack-free and scalable robust training via maximizing certified radius. *International Conference on Learning Representations (ICLR)*.