Deep Learning at the Intersection: Certified Robustness as a Tool for 3D Vision

Gabriel Pérez S.[1*], Juan C. Pérez[2*],
Motasem Alfarra[2], Jesús Zarzar[2], Sara Rojas[2], Bernard Ghanem[2], Pablo Arbeláez[1]

[1]Universidad de Los Andes, [2]KAUST

Abstract

This paper presents preliminary work on a novel connection between certified robustness in machine learning and the modeling of 3D objects. We highlight an intriguing link between the Maximal Certified Radius (MCR) of a classifier representing a space's occupancy and the space's Signed Distance Function (SDF). Leveraging this relationship, we propose to use the certification method of randomized smoothing (RS) to compute SDFs. Since RS' high computational cost prevents its practical usage as a way to compute SDFs, we propose an algorithm to efficiently run RS in low-dimensional applications, such as 3D space, by expressing RS' fundamental operations as Gaussian smoothing on pre-computed voxel grids. Our approach offers an innovative and practical tool to compute SDFs, validated through proof-of-concept experiments in novel view synthesis. This paper bridges two previously disparate areas of machine learning, opening new avenues for further exploration and potential cross-domain advancements.

1. Introduction

This paper explores preliminary observations on a connection between certified robustness and 3D object modeling in machine learning. Certified robustness studies guarantees about model prediction stability under input variations [13, 2], while 3D object modeling focuses on computational representations like meshes, voxel grids [16], and signed distance functions (SDFs) [3, 11].

Our work centers on the link we observe between SDFs and a fundamental concept in certified robustness. In particular, we consider the concept of "Maximal Certified Radius" (MCR) [17] for certifying a classifier f at an input x: that is, the radius r of the largest ball, around x, inside of which f's predictions remain constant. Based on that concept, our main theoretical observation is that *computing the MCR for a classifier f that represents the space's occupancy is equivalent to computing the space's SDF*. We visualize this notion in Figure 1 (left), where we consider

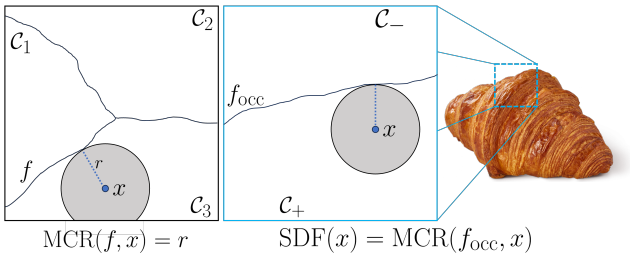


Figure 1: **Connection between certified robustness and 3D modeling via signed distance function (SDF).** We observe an equivalence between the Maximal Certified Radius (MCR) of a space's occupancy function f_{occ} at a point x and the value of $\text{SDF}(x)$, *i.e.* the (signed) distance to the closest surface.

computing the MCR of a three-way classifier f at an input x, resulting in the radius r for which the prediction of f remains unchanged. We note that, if the occupancy in space was modeled via a binary classifier f_{occ}, then the MCR r of f_{occ} at x corresponds precisely to the distance to the closest surface, which is, by definition, the SDF.

Given the connection between SDFs and MCRs, we turn to the literature on certified robustness for certification algorithms, where we find the method of randomized smoothing (RS) [2]. However, we find that RS' high computational cost hinders its usage for computing SDFs in practice. To circumvent this, we further make the observation that numerous applications of SDFs are for the 3D world [5, 1, 7, 11], and thus focus our attention on this (low-dimensional) space. In this setup, we propose an algorithm that can efficiently run RS in three dimensions by expressing RS' fundamental computations in terms of inexpensive Gaussian smoothing on pre-computed voxel grids.

We validate our preliminary observations via proof-of-concept experiments that test the benefits of using certification to compute SDFs in 3D applications. Specifically, we observe that our efficient approach can be integrated into the task of novel view synthesis [9], where we model scene geometry through an occupancy grid [8] whose SDF is easy to compute via our certification algorithm. Our experiments show that our method can be added to novel view synthe-

[*]These authors contributed equally to this work

sis techniques to learn a useful representation of the scene while retaining desirable visual results.

Our contributions can be summarized as follows:

- Identifying a novel link between certified robustness and 3D object modeling, two previously distinct areas in machine learning, providing insights that can benefit both domains.
- Proposing an algorithm to efficiently compute randomized smoothing certificates, *i.e.* compute SDFs, in low-dimensional applications (such as the physical world) by leveraging Gaussian smoothing on voxel grids.
- Validating the practical applicability of our approach via proof-of-concept experiments. In particular, we find that our algorithm can effectively be used to represent geometry in pipelines for novel view synthesis.

We underscore that this manuscript presents work in progress whose details we are still studying.

2. Preliminaries: Certified Robustness

Given the vulnerability of Deep Neural Networks (DNNs) to small imperceptible perturbations known as adversarial attacks [4, 15], several works devised approaches with provable robustness guarantees. That is, given a classifier $f_\theta : \mathcal{X} \to \mathbb{R}^k$ where $\mathcal{X}$ is the input space and k is the number of classes, certified robustness aims at making f_θ output a fixed prediction for an $\|\ell\|_p$ region around a given x. In other words, we say that f_θ is certifiably robust at a given x with radius $r > 0$ if the following statement is achievable:

$$\arg\max_i f_\theta^i(x) = \arg\max_i f_\theta^i(x+\delta) \quad \forall \|\delta\| \leq r,$$

where $f_\theta^i(x)$ is the i^{th} element of the vector $f_\theta(x)$. We note that quantifying the certified radius r is, in general, a very challenging problem for DNNs. However, recently, a tractable approach, known as Randomized Smoothing [2], alleviated this problem via a probabilistic approach.

2.1. Randomized Smoothing

Randomized Smoothing (RS) [2] is a technique that constructs a smoothed classifier g derived from a base classifier f_θ. At its core, g assigns an input x the class most likely predicted by the base classifier f when x is subjected to perturbations with isotropic Gaussian noise. More formally, and for a binary classifier $f_\theta : \mathcal{X} \to \{0,1\}$*, the smoothed classifier is given by:

$$g(x) = \arg\max_{c \in \{0,1\}} \mathbb{P}_{\epsilon \sim \mathcal{N}(0,\sigma^2\mathrm{I})}(f_\theta(x+\epsilon) = c).$$

where σ^2 controls the trade-off between robustness and accuracy for the smooth classifier. Cohen *et al.* [2] showed that g is certifiably robust at least with radius R given by

$$R = \sigma\, \Phi^{-1}(p_A) \tag{1}$$

with Φ^{-1} being the inverse CDF of the standard Gaussian distribution and $p_A = \max_{c \in \{0,1\}} \mathbb{P}_{\epsilon \sim \mathcal{N}(0,\sigma^2\mathrm{I})}(f_\theta(x+\epsilon) = c)$. That is, $g(x) = g(x+\delta)\ \forall \|\delta\|_2 \leq R$. While estimating p_A is generally challenging for complex DNNs, Cohen *et al.* proposed a tractable approach based on Monte-Carlo sampling with confidence bounds [2]. Notably, the certified radius provided by RS is exact in the convex case, and a lower bound otherwise.

3. Efficient Randomized Smoothing in Low-dimensional Spaces

Let $k \in [0,1]^d$ be a grid representing the soft occupancy of an object, where $k^i \in [0,1]$ represents the probability of the i^{th} voxel being occupied with an object. Then, $f_\theta(x)$ may continuously represent the soft occupancy of an object by trilinearly interpolating voxel values of k around x. We observe that the certified radius of each voxel, the distance for the predicted class to flip, represents the Signed Distance Function (SDF). Our main aim is to leverage RS to compute the SDF. However, conducting the Monte-Carlo approach from [2] on each voxel results in intractable computation.

To that end, we leverage the equivalency between subjecting the input of the base classifier to isotropic Gaussian noise, and convolving the output with a Gaussian distribution [14]. That is, one can rewrite the smooth classifier as: $g(x) = \arg\max_c \hat{f}^c(x)$ with

$$\hat{f}(x) = \left(f_\theta * \mathcal{N}(0, \sigma^2\mathrm{I})\right)(x),$$

where $*$ denotes the convolution operator.

While such computation is impractical for classification problems, we unlock its potential in low-dimensional spaces. In particular, we note that $\hat{f}(x)$ can be efficiently approximated via inexpensive Gaussian smoothing on a voxel grid discretizing space, which is tractable for low-dimensional spaces. This formulation allows efficient calculation of certified radii via RS, offering a promising avenue for applying to new domains.

4. SDF as Certified Radius: Applications in 3D Vision

Our key insight is the equivalence between a space's Signed Distance Function (SDF) and the Maximal Certified Radius (MCR) of the space's occupancy function. In particular, we note that computing the SDF value at a point in space is equivalent to computing the MCR of the occupancy function at that same point. Formally, for any given x in a space whose occupancy is described by the occupancy function f_{occ}, *i.e.* a binary classifier, the following holds:

$$\text{SDF}(x) \equiv \text{MCR}(f_{\text{occ}}, x), \tag{2}$$

*Extension to a larger number of classes follows directly (see [2]).

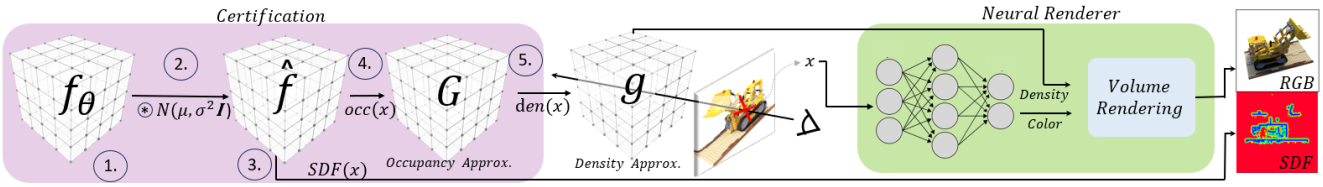


Figure 2: **Rendering pipeline**: A voxel grid f_θ is trained within [0, 1]. Gaussian smoothing via 3D convolution produces $\hat{f}$. Utilizing this, a Weak Signed Distance Function (SDF) incorporates $\hat{f}$, the normal distribution's CDF, and σ. High-eccentricity sigmoid application to $\hat{f}$ generates $G(x)$ for occupancy. Rendering density, obtained from $G(x)$, is calculated using a differentiable density activation function resembling a negative logarithm asymptotically approaching $1 + \epsilon$, yielding $g(x)$. The renderer queries $g(x)$ for density.

where $\text{SDF}(x)$ denotes the signed distance function at point x and $\text{MCR}(f_{\text{occ}}, x)$ denotes the maximal certified radius for the occupancy function at x. Please refer to Figure 1 for a visual guide to our observation. We note that, while Eq. (2) holds true, the radius given by Eq. 1 is, in the general case, a lower bound, and so in practice we estimate a *weak* SDF, *i.e.* a *lower bound* to the SDF.

4.1. Novel View Synthesis

An important application in 3D computer vision which can benefit from the use of SDFs is novel view synthesis. This task consists on predicting novel views of a scene from a set of posed images of the scene. NeRFs [9] successfully tackled this task by learning two fields in 3D: a radiance field $\hat{L}_o(\mathbf{x}, \mathbf{d}; \theta) : \mathbb{R}^3 \times \mathbb{R}^2 \to \mathbb{R}^3$, representing outgoing radiance in each point $\mathbf{x}$ in direction $\mathbf{d}$, and a density field $\sigma(\mathbf{x}; \theta) : \mathbb{R}^3 \to \mathbb{R}$ which captures the scene's geometry. Given these functions, one can leverage the volume rendering integral

$$\hat{C}(\mathbf{r}; \theta) = \int_{t_n}^{t_f} T(t)\, \sigma(\mathbf{r}(t))\, \hat{L}_o(\mathbf{r}(t), \mathbf{d})\, \mathrm{d}t, \qquad (3)$$

to compute pixel colors $C(\mathbf{r})$ along rays $\mathbf{r}(t)$ in space. The parameterized radiance and density functions can be trained by minimizing a reconstruction loss measuring dissimilarity between rendered ray values and original image pixels. Please refer to [9] for further details.

We demonstrate an application of our method by generating a density field from the occupancy we learn, and incorporating this density within a popular framework for novel view synthesis [10]. Our rendering pipeline works as follows:

1. **Learning a Voxel Grid**: We train a voxel grid, representing a field, which we denote as f_θ. By design, the values of f_θ are clamped between 0 and 1.
2. **Gaussian Smoothing**: We perform Gaussian smoothing with a kernel of standard deviation σ to derive $\hat{f}$. Given the constraints on f_θ, $\hat{f}$ invariably retains values within the 0 to 1 range.
3. **Deriving the *weak* SDF**: At any point x of interest, a weak Signed Distance Function (SDF) can be computed through the equation:

$$\text{SDF}(x) = \sigma \times \Phi^{-1} \hat{f}(x))$$

where Φ denotes the CDF of the Gaussian distribution.

4. **Achieving Hard Classification for Occupancy**: Direct extraction of a hard version from $\hat{f}$ would involve the argmax operation. However, to allow for back-propagation, we approximate the argmax operator with a high-eccentricity sigmoid function

$$G(x) = \text{occupancy}(x) = \text{sigmoid}\left(\alpha \cdot \left(\hat{f}(x) - 1/2\right)\right).$$

5. **Mapping to Density**: Rendering requires density, which we model as a (monotonically increasing) transformation of occupancy via a differentiable function h:

$$g(x) = \text{density}(x) = -30 \cdot \ln(1 + \epsilon - G(x))$$

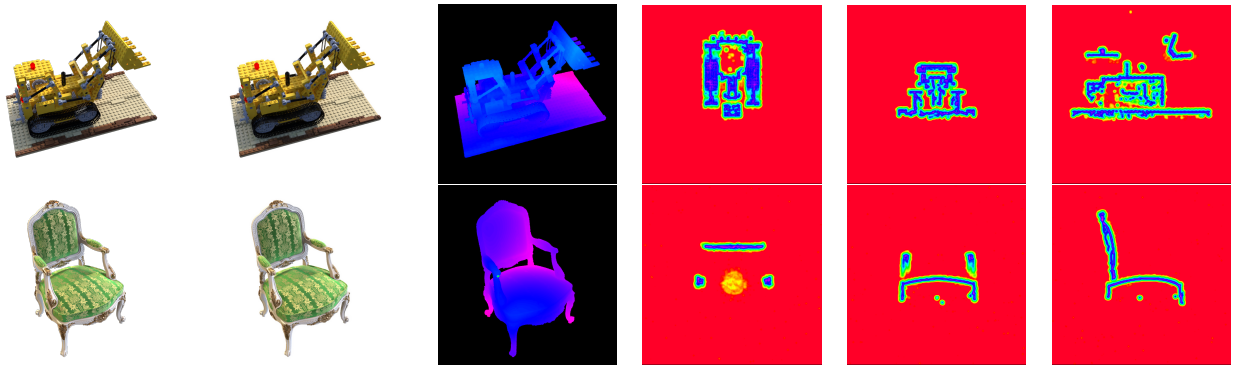
5. Experiments & Results

5.1. Setup and Implementation Details

Dataset We demonstrate our approach using synthetic scenes from NeRF [9]. This dataset consists of 8 synthetic scenes with 100 training images rendered from various poses around the object.

Implementation We build on top of Nerfacc's [6] PyTorch code base [12] implementation of I-NGP [10] with its default parameters. We set the Gaussian smoothing's σ to 1.1, sigmoid's α to 19, use 10k optimization iterations for I-NGP, 5k iterations for fine-tuning the voxel grid, and 500 optimization iterations for the final fine-tuning. We test three resolutions for the discrete learned representation F: 192^3, 256^3 and 384^3.

3D object extraction We extract object meshes by employing marching cubes on top of the weak SDF our method generates. In our experiments, we run marching cubes (MC) searching for a slightly negative isosurface of SDF $= -0.1$, with the purpose of compensating for locally sharp irregularities in points with certified radii of exactly 0. Comparison to I-NGP requires extracting a surface from its density field via MC. To determine a good density threshold for I-NGP, we run MC with threshold values in multiples of 10 from 0 to 100, and report its best-performing numbers.



(a) Ground Truth (b) Rendered Image (c) Depth Map (d) SDF x cut (e) SDF y cut (f) SDF z cut

Figure 3: **Qualitative results** Ground truth test image, rendered image, radiance field's depth map and SDF axis cuts are presented for the Lego and Chair scenes.

Grid res.	chair	drums	ficus	hotdog	lego	mat.	mic	ship	avg.
192	32.65	23.75	28.96	33.83	29.74	26.84	32.50	26.66	29.36
256	32.99	24.20	30.00	34.15	30.33	26.93	32.84	27.07	29.81
384	33.30	24.60	30.89	34.08	31.01	26.60	33.10	27.17	30.09
I-NGP	35.43	25.40	33.37	36.86	35.29	29.20	36.36	29.35	32.65

Table 1: **Test set average PSNR (↑) on the Blender synthetic dataset.**Our method provides inexpensive access to a *guaranteed* weak SDF while simultaneously providing competitive rendering quality.

Grid res.	chair	drums	ficus	hotdog	lego	mat.	mic	ship	avg.
192	661	583	594	614	609	591	678	597	616
256	1216	1070	1084	1104	1123	1125	1262	1091	1134
384	3519	3116	3144	3270	3131	3281	3930	3173	3321
I-NGP	160	148	120	166	153	163	142	156	151

Table 2: **Training time (in seconds).** All experiments were run on a Quadro RTX 8000 GPU.

Grid res.	chair	drums	ficus	hotdog	lego	mat.	mic	ship	avg.
192	51.7	1409.0	590.0	229.9	39.4	82.6	61.4	1639.8	513.0
256	39.7	1202.0	468.7	208.5	36.0	66.4	30.9	1563.6	452.0
384	39.2	754.0	87.1	110.5	33.4	45.2	53.5	1607.1	341.3
I-NGP	260.7	3954.5	1526.4	1284.5	85.1	213.3	5975.0	1422.5	1840.2

Table 3: **Chamfer distance (↓).**

Metrics We measure render quality with Peak Noise to Signal Ratio (PSNR) of the generated images, geometry reconstruction via Chamfer distance (with point clouds of 100k samples), and computational cost via training time.

5.2. Results and Analysis

We report proof-of-concept experiments demonstrating our method's capability to generate guaranteed weak SDFs while being able to generate renders of competitive quality in novel view synthesis.

Qualitative Results We report qualitative results in Figure 3. We show comparisons between GT test set images, the rendered image with our pipeline, and the depth map of the learned density field. These results show that our method is capable of correctly capturing the scene's geometry as well as generating competitive renders. We highlight that our method's most important capacity is providing inexpensive access to a weak SDF of the scene. We visualize this in subfigures (d), (e) and (f) of Figure 3 with color-mapped transversal cuts of the values computed by running RS via our efficient algorithm from Section 3.

SDF Quality We report Chamfer distances in Table 3. With respect to this metric, our method displays clear superiority against I-NGP. That is, even after choosing the best-performing density threshold for I-NGP for each scene, the average Chamfer distance is usually at least three times as large as that of our approach.

Rendering Quality Table 1 reports an average PSNR of 30.09 when using a resolution of 384, which represents a drop of 2.51 points of PSNR compared to I-NGP. This drop in rendering quality is to be contrasted with the fact that our method *guarantees* the output of a weak SDF.

Efficiency The correlation between resolution and both quality and performance becomes critical, as shown Table 2. Space complexity scales cubically, and time complexity seems to follow a similar trend. Our results show that, when considering training time against rendering quality, our method rapidly achieves diminishing returns.

6. Conclusions

In summary, our work demonstrates a novel connection between certified robustness and 3D object modeling, leading to an efficient algorithm for computing weak SDFs. We showcase the utility of this connection by employing it for the task of novel view synthesis where it guarantees learning a weak SDF while maintaining rendering quality. We anticipate that this synergy between robustness and geometry will drive further advancements in machine learning, computer graphics, and related domains, inspiring researchers to explore the untapped potential at this intersection.

References

[1] Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 608–625. Springer, 2020. 1

[2] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019. 1, 2

[3] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 1

[4] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 2

[5] Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Diffusion-sdf: Text-to-shape via voxelized diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12642–12651, June 2023. 1

[6] Ruilong Li, Hang Gao, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: Efficient sampling accelerates nerfs. *arXiv preprint arXiv:2305.04966*, 2023. 3

[7] Daniel Mayost. *Applications of the signed distance function to surface geometry*. University of Toronto (Canada), 2014. 1

[8] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 1

[9] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 3

[10] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 3

[11] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 1

[12] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[13] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018. 1

[14] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[15] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 2

[16] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[17] Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. In *International Conference on Learning Representations*, 2020. 1