Fast and Flexible Robustness Certificates for Semantic Segmentation

Thomas Massena[1,3], Corentin Friedrich[2], Franck Mamalet[2], and Mathieu Serrurier[1]

[1] Institut de Recherche en Informatique de Toulouse, France
[2] IRT Saint Exupéry, Toulouse, France
[3] SNCF DTIPG, Saint-Denis, France

Abstract. Deep neural networks remain vulnerable to imperceptible adversarial perturbations, yet efficient robustness certification for semantic segmentation is largely unexplored. We propose the first real-time compatible certifiably robust semantic segmentation framework based on networks with built-in Lipschitz constraints. Our method achieves competitive pixel accuracy on Cityscapes while providing deterministic robustness guarantees, without the computational overhead of randomized smoothing or formal verification. We introduce a generalized certification framework for segmentation that computes worst-case performance under ℓ_2 attacks of radius ϵ across diverse evaluation metrics. Crucially, our approach is approximately $600\times$ faster than randomized smoothing at inference on an NVIDIA A100 GPU, delivering comparable certificates. We validate our worst-case bounds against state-of-the-art adversarial attacks, demonstrating that Lipschitz-constrained architectures offer a practical and scalable path toward trustworthy dense prediction in safety-critical applications.

Keywords: Certified Robustness · Semantic Segmentation · Lipschitz Networks

1 Introduction

Deep Neural Networks (DNNs) have transformed the landscape of machine learning, driving breakthroughs across perception, reasoning, and decision-making tasks. Yet, their remarkable expressivity comes at the cost of a vulnerability to carefully crafted adversarial perturbations [8]. These weaknesses raise critical concerns when deploying Machine Learning (ML) systems in safety-critical contexts such as autonomous driving or healthcare, where guarantees on the reliability of decisions are crucial [44]. While obtaining robustness certificates for tasks other than classification has been attempted in the context of formal verification methods [10] and randomized smoothing [7, 12], Lipschitz-by-design methods, introduced in Tsuzuku *et al.* [39], were never adapted to provide downstream certificates for tasks more complex than classification or regression. We denote these Lipschitz-by-design networks as LipNets or L-LipNets when the Lipschitz constant is constrained to L. Our contributions are as follows:

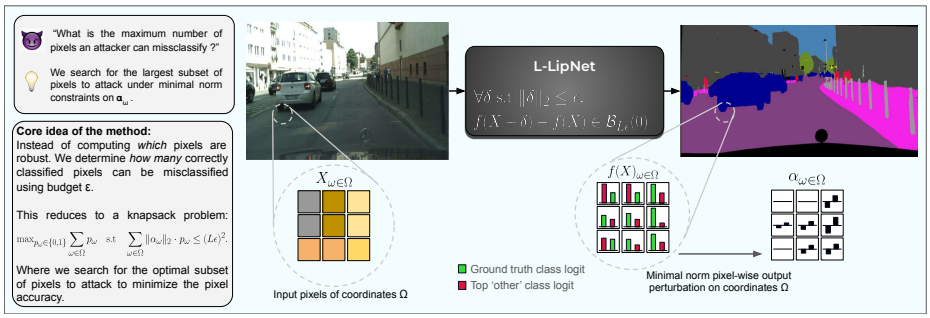

Fig. 1: An overview of our method. We leverage Lipschitz constrained networks and reduce the pixel accuracy degradation objective to a simple knapsack formulation.

- First, we introduce a general class of robustness certificates for segmentation tasks, tailored to certify the robustness of a variety of performance measures to arbitrary adversarial degradations.
- Then, using this framework, we provide low overhead certificate computation methods for the Pixel Accuracy, FNR or Class IoU performance measures under attack for Lipschitz-by-design networks.
- Also, we develop the first efficiently trainable Lipschitz neural network models for semantic segmentation tasks, using a DeepLabV3-like architecture, and scale them to handle challenging tasks like the Cityscapes dataset.
- Leveraging our framework and models, we compute robustness certificates for several safety-critical scenarios. Our networks are suitable for real-time inference and provide the first certifiably robust method that allows the computation of robustness certificates in under 0.1 seconds on 1024 by 1024 images from the Cityscapes dataset.

Additionally, we provide the code for this project on GitHub It allows the training and certification of LipNets in ℓ_2 norm on various semantic segmentation tasks.

2 Background on Adversarial Robustness

First, let us introduce some key concepts and their formulations in the context of classification tasks. Then, we will present relevant certifiable robustness works in the context of semantic segmentation.

2.1 Adversarial Robustness in the Classification Setting

Seminal adversarial deep learning works originally focused on designing adversarial attacks to fool classification models [8]. To this day, a sizeable proportion

[4] All results remain valid for the ℓ_p-norm, for any finite $p \in \mathbb{N}^*$, provided the LipNet is defined w.r.t. the same norm.

Symbol	Description
$\mathcal{K}$	Label (class) set
Ω, $\lvert\Omega\rvert = H \times W$	Pixel coordinate set and image size
$\mathcal{X} := [0,1]^{\lvert\Omega\rvert}$	Input space (images)
$\mathcal{Y} := \mathbb{R}^{\lvert\mathcal{K}\rvert \times \lvert\Omega\rvert}$	Output space (logit maps)
$f : \mathcal{X} \to \mathcal{Y}$	Semantic segmentation model
$f^k(X)_\omega$	Logit of class k at pixel ω
$\hat{Y}_\omega = \arg\max_k f^k(X)_\omega$	Predicted class at pixel ω
$\mathcal{B}_\epsilon(x)$	ℓ_2-ball of radius ϵ around x[4]
h, h_ϵ	Performance metric and its ϵ-worst-case value
κ	Degradation objective (0/1 criterion)

Table 1: Summary of notation.

of the adversarial deep learning literature still focuses on designing maximally efficient attacks on classification tasks. Let us define the key concepts surrounding certifiable adversarial robustness in the context of classification tasks.

Definition 1 (Robustness radius - classification). *For any classifier, $f : \mathcal{X} \to \mathbb{R}^{\lvert\mathcal{K}\rvert}$ and any test point (X, Y), the robustness radius of the model's prediction is defined as:*

$$R(X,Y) = \inf \left\{ \epsilon \in \mathbb{R}^+ \,\middle|\, \exists \delta \in \mathcal{B}_\epsilon(0),\ \operatorname{argmax}_{k \in \mathcal{K}} f^k(X+\delta) \neq Y \right\}. \tag{1}$$

This definition is widely adopted in the context of certifiably robust classification and allows one to characterize how "invariant" a neural network's prediction is to local perturbations. In practice, there are multiple ways of obtaining a lower bound on the value of the robustness radius of Definition 1. We can distinguish three competing approaches to lower-bound the robustness radius of a DNN's prediction locally. Namely, formal-verification methods, randomized smoothing variants, and LipNets stand out as the main approaches.

- **Formal verification methods:** allow for the computation of worst-case logit variations under ϵ-bounded ℓ_p-norm attacks at inference by using formal-verification solvers [16, 38].
- **Randomized smoothing methods:** which define a smoothed classifier g from the base model f, Monte-Carlo estimates then allows probabilistic robustness certificates when using well-chosen sampling distributions [9, 21, 42].
- **Lipschitz neural networks (LipNets):** which leverage a Lipschitz-by-construction weight parametrization methods to ensure bounded worst-case logit variations, incurring no inference-time overhead.

Importantly, all of these methods allow the user to find a conservative lower bound $\underline{R}(X,Y) \leq R(X,Y)$ such that no adversarial attack inferior to ϵ in ℓ_p

norm could misclassify the originally correctly classified sample X. These robustness lower bounds are usually coined as robustness *certificates*.

2.2 Lipschitz Certificates for Classification

A function $f : \mathbb{R}^m \to \mathbb{R}^n$ is said to be L-Lipschitz in ℓ_2 norm if it verifies:

$$\forall (X_1, X_2) \in \mathbb{R}^m, \|f(X_1) - f(X_2)\|_2 \leq L \cdot \|X_1 - X_2\|_2 \tag{2}$$

with $L > 0$. This property effectively bounds how "stable" a predictor's decision remains under attack. Moreover, L-LipNets retain universal approximation for classification (see Béthune *et al.* [4]): For example, a ReLU network is a universal approximator with arbitrarily large Lipschitz constant $\mathcal{L}$; rescaling its outputs by $L/\mathcal{L}$ preserves the argmax, yielding an L-Lipschitz universal classifier. Note that this argument remains valid for a semantic segmentation task.

Classification robustness certificate: For any L-Lipschitz predictive model in ℓ_2 norm, $f : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{K}|}$, we have the following lower bound $\underline{R}(X, Y)$ on $R(X, Y)$:

$$R(X, Y) \geq \underline{R}(X, Y) = \mathbb{1}_{\hat{Y}=Y} \cdot \frac{\mathcal{M}_X(f)}{\sqrt{2}L}. \tag{3}$$

with $\mathcal{M}_X(f) = f^{\text{top1}}(X) - f^{\text{top2}}(X)$, as done in [22]. Here, the $\mathbb{1}_{\hat{Y}=Y}$ term ensures that the robustness radius of originally misclassified samples is zero. LipNets allow for easily computable robustness radius lower bound computations using the global Lipschitz constant of a specially designed neural network.

2.3 Related Works on Robust Segmentation

While adversarial robustness has been extensively investigated for image classification, the robustness of semantic segmentation models has received comparatively less attention. Early work by Arnab *et al.* [2] provided one of the first systematic analyses of adversarial examples on modern semantic segmentation architectures, showing that dense prediction exhibits vulnerability patterns that differ from standard classification. Building on segmentation-specific threat models, Rony *et al.* [32] introduced a proximal-splitting-based adversarial attack that directly optimizes dense pixel-wise perturbations and yields substantially stronger white-box attacks than those adapted from classification, providing a stronger benchmark for assessing segmentation robustness. These works highlight that semantic segmentation robustness remains relatively underexplored compared to the classification case and motivate further study in safety-critical domains such as automated driving perception, for which broader surveys of certified robustness have recently been proposed [44]. In the context of certifiably robust Semantic Segmentation the overall approaches to certification remain the same as for classification. However, the high dimensionality of the neural segmentation network's outputs complicate the analysis. We can identify two main existing approaches to robust semantic segmentation tasks.

Formal verification methods While some formal-verification methods allow for the verification of a segmentation network's predictions on toy datasets [38], formal certifiable methods lack scalability or end up providing vacuous bounds which are not applicable to over-parametrized networks that segment high resolution images. Indeed, these methods usually have time complexities that grow quadratically with the number of neurons and layers of the model [16]. Therefore, even the most prestigious neural network verification competitions only include verification tasks for simple semantic segmentation networks with fewer than a million parameters [27]. Currently, specific approaches to robust semantic segmentation using formal methods include patch-based attack defenses [25] and approaches that can segment MNIST-like images with pixel-wise certificates [29].

Randomized Smoothing In the context of semantic segmentation, Fischer *et al.* [12] first applied randomized smoothing to certify subsets of pixel-wise predictions up to a robustness radius R with probability $1 - \alpha$. The authors find that applying randomized smoothing naively to the $H \times W$ classifications provided by the segmentation network requires too many samples to certify the validity of all pixel classifications with desired error level. To bypass this they approach certification as a multiple-testing problem where the SegCertify method controls the family-wise error rate at level α using a step-down procedure to ensure the validity of the segmentation of a whole image up to a user-chosen error probability α, giving the network the possibility of abstaining if a pixel prediction is unsure or not robust. This method necessitates around $n_{\text{MC}} \approx 10^4$ Monte Carlo (MC) estimates and statistical correction on $H \times W$ pixels, making the method unusable for real-time systems.

More recently, the LocalizedLP method was introduced in Schuchardt *et al.* [36]. This method improves on the SegCertify method by leveraging the local dependencies of semantic segmentation networks via grid-based isotropic Gaussian smoothing. Unfortunately, this improvement comes at a price in terms of sample complexity as LocalizedLP uses around 15× more MC estimations than SegCertify, as described in Section 7.1 of [36]. For reference, the implementation of [36] uses a DeepLabV3 architecture similar to ours on the Cityscapes dataset, and reports 1204 seconds of runtime per image by using 153600 Monte Carlo iterations. Therefore, in this paper, we will not compare the efficiency of our method to the LocalizedLP method given the significant inference efficiency gap. Similarly, Laousy *et al.* [19,20] introduce a method that relies on denoising diffusion models, which further increases runtime by around 43% w.r.t. SegCertify.

3 Unifying Semantic Segmentation Robustness Metrics

The notion of adversarial robustness in semantic segmentation is not clearly defined, and relies on sparse metrics introduced throughout the literature. Adversarial attacks usually aim to increase a proxy loss function with a fixed perturbation budget ϵ in order to decrease the pixel accuracy and the mean Intersection

over Union (mIoU). In contrast, the attack introduced in Rony *et al.* [32] seeks the minimal norm attack to successfully perturb at least $\gamma\%$ of pixels. Regarding defenses, some works focus on certifying pixel-wise classification [12,38], whereas others ensure collective robustness certificates for the classification of pixel subgroups [36]. In this section, we present a general framework for the robustness of performance measures especially for semantic segmentation.

Two paradigms of robustness: In a safety-critical setting, the robustness of a given sample (X, Y) can be addressed from two different perspectives. We express these perspectives as the following questions.

Q1 (Worst Case Performance) - *Given an adversarial budget ϵ, what is the worst achievable performance under attack?*
Q2 (Generalized Robustness Radius) - *Given a target performance degradation, what is the minimal adversarial budget required to reach it?*

Most attacks and defenses try to answer **Q1** by degrading the usual segmentation metrics (i.e. pixel accuracy or mIoU) with a fixed budget. Answering **Q2** is less common. For example, Rony *et al.* [32] consider the degradation of the pixel accuracy metric with the objective of less than $\gamma\%$.

Definition 2 (Worst-case performance). *For any predictive model $f : \mathcal{X} \to \mathcal{Y}$, and performance metric $h : \mathcal{Y} \times \mathcal{K}^{|\Omega|} \to \mathbb{R}$, we define the ϵ worst-case performance measured on a data point (X, Y) as:*

$$h_\epsilon(X, Y) = \min_{\delta \in \mathcal{B}_\epsilon(0)} h\left(f(X + \delta), Y\right). \tag{4}$$

In our setting, we assume that h is positively correlated with system performance, i.e "higher is better".

For example we can set h as the pixel accuracy measure, then, h_ϵ is the minimum reachable pixel accuracy on an image and mask pair (X, Y) under fixed adversarial budget ϵ, also called Robust Pixel Accuracy (RPA) (see Sec. 4.1).

Definition 3 (Generalized robustness radius). *For any predictive model $f : \mathcal{X} \to \mathcal{Y}$, performance metric $h : \mathcal{Y} \times \mathcal{K}^{|\Omega|} \to \mathbb{R}$, and degradation objective $\kappa : \mathbb{R} \to \{0, 1\}$, we define the generalized robustness radius on a data point (X, Y) as:*

$$R_\kappa(X, Y) = \inf\left\{\epsilon \in \mathbb{R}^+ \,\middle|\, \exists \delta \in \mathcal{B}_\epsilon(0), \kappa\left[h(f(X + \delta), Y)\right] = 1\right\}. \tag{5}$$

The degradation objective κ on performance metric h is either unsatisfied (0=failure) or satisfied (1=success).

This definition allows us to define a robustness radius for any desired adversarial degradation objective κ on performance measure h in a variety of sce-

narios[5]. For example, with h being the pixel accuracy, and $\kappa(z) = \mathbb{1}_{z \leq \gamma\%}$ the degradation objective, the radius R_κ is the minimum budget to reach a pixel accuracy below $\gamma\%$ under attack. Computing h_ϵ and R_κ exactly is infeasible in practice. In certified robustness, we answer **Q1** and **Q2** by computing lower bounds (also called certificates) on h_ϵ and R_κ. Secs. 4.1 and 4.2 will illustrate how to compute those certificates for several metrics such as pixel accuracy, FNR, stability and IoU.

4 Fast Estimation of Robustness Certificates

In the following Section, we show how to use Lipschitz pixel-wise certificates to provide lower-bounds for questions **Q1** and **Q2** efficiently.

Computing certificates In order to compute the worst-case performance variations of an L-LipNet under ϵ-bounded attacks, we leverage:

$$h_\epsilon(X, Y) := \min_{\delta \in \mathcal{B}_\epsilon(0)} h(f(X + \delta), Y) \geq \min_{\alpha \in \mathcal{B}_{L\epsilon}(0)} h(f(X) + \alpha, Y). \tag{6}$$

This conversion from input space perturbations in $\mathcal{B}_\epsilon(X)$ to output perturbations in $\mathcal{B}_{L\epsilon}(f(X))$ comes directly from the L-Lipschitz property (Eq. (2)) of the neural network *w.r.t the output vector in* $\mathcal{Y}$ of cardinality $|\mathcal{K}|.|\Omega|$ (for the sake of completeness a proof is given in Appendix A) Importantly, the Lipschitz bound used here aims to determine the minimum possible performance measure under attack, *which differs from pixel-wise prediction robustness measures.*

4.1 Application to the Pixel Accuracy Metric

For a subset of input pixels $S \subseteq \Omega$, we define the pixel accuracy on S by $h : \mathcal{Y} \times \mathcal{K}^{|S|} \to \mathbb{R}$ as:

$$h(f(X), Y) = \frac{1}{|S|} \sum_{\omega \in S} \mathbb{1}_{\hat{Y}_\omega = Y_\omega} \tag{7}$$

with $\hat{Y}_\omega$ the decision at the pixel ω . Note that $S = \Omega$ is the classical pixel accuracy measure on the full image. We define the Robust Pixel Accuracy (RPA) $h_\epsilon(X, Y) = \min_{\delta \in \mathcal{B}_\epsilon(0)} h(f(X + \delta), Y)$ as the exact worst-case pixel accuracy under attack. In practice, following Eq. (6), we consider the Certifiably Robust Pixel Accuracy $\text{CRPA}_\epsilon(X)$ which is a lower bound to the RPA.

Deriving an answer to Q1. Here, we aim to find the worst-case pixel accuracy under attack. This worst-case degradation is reached when a maximum number

[5] Note that κ could also be defined as $\kappa : \mathbb{R} \times \mathbb{R} \to \{0, 1\}$ with $\kappa[h(f(\tilde{X}), Y), h(f(X), Y)]$ to ensure robustness to degradations that are relative to the performance of the network on a clean point (e.g the accuracy must be at least half of the original accuracy).

of pixels is misclassified under budget ϵ. First, based on Eq. (3), we derive the pixel-wise certificate for each pixel $\omega \in S$:

$$\underline{R}^{\omega}(X, Y) = \mathbb{1}_{\hat{Y}_\omega = Y_\omega} \cdot \frac{\mathcal{M}^{\omega}_{X}(f)}{\sqrt{2}L}. \tag{8}$$

Intuitively, the attacker must prioritize the less-robust pixels. Leveraging the L-Lipschitz condition w.r.t the output vector (Eq. (6)), we propose to reformulate the $\mathrm{CRPA}_\epsilon(X)$ computation as a Knapsack Problem [34], which aims to compute the maximum number of pixels that could be misclassified under attack budget ϵ. For each pixel $\omega \in S$, we define a profit variable $p_\omega \in \{0, 1\}$ that equals 1 if the pixel is selected and a corresponding weight $c^{\mathrm{PA}}_\omega = (L\underline{R}^{\omega}(X, Y))^2$ for its misclassification. To ensure the total weight do not exceed capacity $\Lambda = (L\epsilon)^2$, the Knapsack formulation is then:

$$N_{\mathrm{KP}}(X, \Lambda, S, c^{\mathrm{PA}}_\omega) = \max \sum_{\omega \in S} p_\omega \quad \text{s.t.} \quad \sum_{\omega \in S} c^{\mathrm{PA}}_\omega p_\omega \leq \Lambda \tag{9}$$

Here, the linear constraint corresponds to the condition $\alpha \in \mathcal{B}_{L\epsilon}(0)$ in Eq. (6), i.e. $\|\alpha\|^2 = \sum_{\omega \in \Omega} c^{\mathrm{PA}}_\omega p_\omega \leq \Lambda = (L\epsilon)^2$. We can thus derive the following lower bound for the certification of **Q1**:

$$\mathrm{CRPA}_\epsilon(X, Y) = 1 - \frac{N_{\mathrm{KP}}(X, \Lambda, S, c^{\mathrm{PA}}_\omega)}{|S|}. \tag{10}$$

The unidimensional Knapsack Problem described in Eq. (9) can be solved optimally in $\mathcal{O}(|S| \cdot \log|S|)$ time by sorting pixels by ascending contribution c^{PA}_ω and greedily adding them until the capacity is exhausted [34]. This corresponds to the selection of less robust pixels first (already misclassified pixel having $c^{\mathrm{PA}}_\omega = 0$). By denoting $\pi_X : [1, |S|] \to S$ as a map that sorts pixel coordinates of X in ascending order of robustness according to the value of c^{PA}_ω on the (X,Y) pair, the solution of the Knapsack problem in Eq. (9), is given by:

$$N_{\mathrm{KP}}(X, \Lambda, S, c^{\mathrm{PA}}_\omega) = \quad \sup \left\{ n \in [1, |S|] \;\middle|\; \sum_{k=1}^{n} c^{\mathrm{PA}}_{\pi_X(k)} \leq \Lambda \right\}. \tag{11}$$

This estimation process can be efficiently parallelized and run on GPU architectures which results in negligible overhead for the certification process compared to the forward pass of a DNN.

Deriving an answer to Q2. We now seek to answer question Q2 and find the generalized robustness radius of the pixel accuracy metric. First, as defined in our framework and using the previously defined pixel accuracy measure h, we can translate the question into a degradation criterion:

$$\kappa[h(f(X + \delta), Y)] = \mathbb{1}_{h(f(X+\delta),Y) \leq \gamma}. \tag{12}$$

This corresponds to assessing the minimum norm perturbation such that at least $\lceil \gamma.|S| \rceil$ pixels are misclassified. We denote $n_\gamma := \lceil \gamma.|S| \rceil$. Once more, this

can be formulated as a Knapsack Problem[6], with the input perturbation weight $c_\omega^{\mathrm{PA}_2} = \underline{R}^\omega(X,Y)^2$:

$$R_\kappa(X,Y)^2 \geq \min \sum_{\omega \in S} c_\omega^{\mathrm{PA}_2} p_\omega \quad \text{s.t.} \quad \sum_{\omega \in S} p_\omega \geq n_\gamma, \tag{13}$$

Similary as **Q1**, the solution of Eq. (13) can be solved using the pixels sorted by ascending order of robustness:

$$\underline{R}_\kappa(X,Y,S,\underline{R}^\omega,n_\gamma)^2 = \sum_{k=1}^{n_\gamma} \underline{R}^{\pi_X(k)}(X,Y)^2. \tag{14}$$

$\underline{R}_\kappa$ defines the Generalized Robustness lower-bound radius for the γ pixel accuracy over a subset $S \subseteq \Omega$ of the pixels of image X. Note that if the clean accuracy is inferior to the threshold γ, we have $\underline{R}_\kappa = 0$, since $c_\omega^{\mathrm{PA}_2} = 0$ for all misclassified pixels.

4.2 Application to FNR and Stability Metrics

In this section, we consider the certification of other prediction-based performance robustness measures.

FNR: To compute the robustness certificates for the FNR measure on a binary semantic segmentation task, we consider the subset $S_1 = \{\omega \in \Omega, Y_\omega = 1\}$, and the performance measure

$$h_{\mathrm{FNR}} = \frac{1}{|S_1|} \sum_{\omega \in S_1} \mathbb{1}_{\hat{Y}_\omega = Y_\omega} = 1 - \mathrm{FNR}(\hat{Y}, Y). \tag{15}$$

We can derive an answer to **Q1** for the worst-case FNR performance by using the same weight c_ω^{PA} and capacity Λ as for CRPA (Sec. 4.1), we can compute the maximum number of misclassified pixels in S_1 under an ϵ budget using Eq. (11):

$$\begin{gathered} N_{\mathrm{FNR}}(X,\epsilon) = N_{\mathrm{KP}}(X,\Lambda,S_1,c_\omega^{\mathrm{PA}}) \\ \text{therefore, } \mathrm{FNR}(\hat{Y},Y) \leq N_{\mathrm{FNR}}(X,\epsilon)/|S_1| \end{gathered} \tag{16}$$

Considering the question **Q2**, regarding the generalized robustness radius, we consider the degradation criterion $\kappa_{\mathrm{FNR}}[h(f(X+\delta),Y)] = \mathbb{1}_{h(f(X+\delta),Y) \leq 1-\gamma}$. Similarly, we can derive a threshold index $n_\gamma = \gamma.|S_1|$ over the less robust pixels in S_1. We can use Eq. (14), and obtain:

$$R_\kappa(X,Y) \geq \underline{R}_\kappa(X,Y,S_1,\underline{R}^\omega,n_\gamma) \tag{17}$$

which is also an easily computable lower bound for the R_κ answering **Q2** with a degradation criterion κ that aims to drive the FNR above a certain threshold γ.
Stability: Importantly, both PA and FNR require the ground-truth Y values. At inference, we propose the stability measure h_{stab} on subsets of pixels $S \subseteq \Omega$

[6] the reformulation as a maximization problem is classical by setting $q_\omega = 1 - p_\omega$.

to ensure their robustness independently from the ground truth. With $\hat{Y}^*$ as the clean decision (with no perturbation), the stability metric is defined as

$$h_{\text{stab}}(f(X), \hat{Y}^*) = \frac{1}{|S|} \sum_{\omega \in S} \mathbb{1}_{\hat{Y}_\omega = \hat{Y}^*_\omega}. \tag{18}$$

The only difference with the pixel accuracy in Eq. (7) is that the reference is the clean decision $\hat{Y}^*$ instead of the ground truth Y. Similarly to Eq. (8), we denote the stability certificate for any particular pixel of coordinate $\omega \in S$ (the factor $\mathbb{1}_{\hat{Y}_\omega = \hat{Y}^*_\omega}$ is omitted since it is ensured by $top_1 = \hat{Y}^*$).

$$\underline{R}^{\omega}_{\text{stab}}(X, Y) := \frac{\mathcal{M}^{\omega}_{X}(f)}{\sqrt{2}L}. \tag{19}$$

We can now provide answers to **Q1** and **Q2** regarding the stability performance measure. Given the stability weight $c^{\text{stab}}_{\omega} = (L\underline{R}^{\omega}_{stab})^2$ and the capacity $\Lambda = (L\epsilon)^2$, the Certified Robustness Stability is defined as $\text{CRS}_\epsilon(X) = 1 - N_{stab}(X, \epsilon)/|S|$ where:

$$N_{\text{stab}}(X, \epsilon) = N_{\text{KP}}(X, \Lambda, S, c^{\text{stab}}_{\omega}). \tag{20}$$

Here, N_{stab} represents an upper bound of the number of output pixels in S that can change their decision under the budget ϵ. We can also address **Q2**, which concerns the generalized stability radius for a given threshold γ. Setting $n_\gamma = \lceil \gamma.|S| \rceil$, we obtain the following lower bound:

$$R_\kappa(X, Y) \geq \underline{R}_\kappa(X, Y^*, S, \underline{R}^{\omega}_{stab}, n_\gamma). \tag{21}$$

Finally, we also provide robustness certificates for more complex performance measures such as the class IoU. The detailed algorithm is provided in Appendix B.

5 Experimental Validation

In this section, we present a threefold analysis evaluating the quality and efficiency of our robustness certificates. First, we compare our LipNet against optimally tuned SegCertify baselines on the Cityscapes dataset and attain competitive CRPA values with $\sim 600\times$ less compute time. Second, to isolate the certificate computation method, we apply SegCertify directly to our LipNet architecture. This model-agnostic ablation determines the n_{MC} sample budget required for probabilistic smoothing to match our single-pass deterministic certificates. Finally, we quantify the tightness of our deterministic bounds via empirical adversarial attacks on the Oxford-IIIT Pet dataset, contrasting this inherent over-approximation against the clean accuracy drop incurred by tuning σ for SegCertify, where clean accuracy must be sacrificed to certify larger radii. We conclude by demonstrating the practical flexibility of Lipschitz-based certification in a safety-critical scenario. We do not compare to formal-verification methods as they cannot scale to multi-million parameter models with such high dimensional outputs.

ϵ	Method	CRPA	Time	Nb. inferences
0.1	Lipschitz bound (ours)	81.80%	≈ 0.1 s	1
0.1	SegCertify ($\sigma = 0.3$)	$53.48 \pm 0.59\%$	59.8 s $\times 594$	60
0.1	SegCertify ($\sigma = 0.2$)	$83.13 \pm 0.33\%$	62.1 s $\times 624$	80
0.17	Lipschitz bound (ours)	77.34%	≈ 0.1 s	1
0.17	SegCertify ($\sigma = 0.4$)	$38.91 \pm 0.53\%$	60.3 s $\times 594$	60
0.17	SegCertify ($\sigma = 0.2$)	$84.84 \pm 0.73\%$	63.3 s $\times 683$	120

Table 2: CRPA values on Cityscapes [11] with 1024×1024 images. We choose $\alpha = 0.001$ as the failure probability of SegCertify and tune $\sigma \in \{0.15, 0.2, 0.25, 0.3, 0.4, 0.5\}$ for each run. Smoothing methods are run on 100 images as done in [12]. We report the mean and standard deviation of results across 5 runs. We also report the mean per-sample runtime for batched evaluation. LipNets are run on the whole test set.

5.1 Designing LipNets for Semantic Segmentation

The exact computation of a network's Lipschitz constant (see Eq. 2) is known to be NP-hard [40]. As a practical alternative, Anil *et al.* [1] introduced architectures in which every linear layer is constrained such that $\|\nabla_x f(x)\|_2 = 1$ almost everywhere. As a result, the multiplicative bound $L_f = \prod_{i=1}^{L} L_{f_i}$ of layer-wise Lipschitz constants L_{f_i}, becomes a meaningful estimation on the network's global Lipschitz constant. Most of these constraints are ensured by differentiable re-parametrizations on the neural networks weights, which can be performed efficiently, resulting in networks with limited training overhead w.r.t their standard unconstrained counterparts (see Appendix F). Moreover, constraining the Lipschitz constant of a neural network exhibits several other benefits, such as explicit control of the network's robustness [4], or even ensuring stable training [3, 28] for state-of-the-art deep learning tasks. We use DeepLabV3 [6] as our baseline architecture for semantic segmentation, since, to our knowledge, it achieves state-of-the-art performance among convolutional neural network (CNN)–based methods. Attention-based architectures were excluded since standard attention layers are not Lipschitz-continuous, and only recent works have started to address this limitation [18]. To design a LipNet version of DeepLabV3, we leverage the orthogonium and deel-lip libraries introduced in [5] and [37] respectively. More details on this network architecture are provided in Appendix D.

Remark. On Cityscapes, Lipschitz constraints reduce clean pixel accuracy from 94.41% to 92.07%. However, training from scratch with Gaussian noise for randomized smoothing can be more costly; as [12] report a 6 percentage point drop to 87% accuracy for an HRNetV2 model.

5.2 Comparing LipNet certificates to SegCertify

As discussed in Sec. 2.3, only a few methods address robustness in semantic segmentation under the ℓ_2 norm. State-of-the-art methods in this domain are based

on randomized smoothing (SegCertify [12]). We use the CRPA metric to compare these approaches with the proposed method: For LipNets, we use Eq. (10) and Eq. (11) to compute the $\mathrm{CRPA}_{\mathrm{Lip}}$ value. Since SegCertify introduces an additional *abstention* class for non-robust pixels, the $\mathrm{CRPA}_{\mathrm{RS}}$ corresponds to the classical accuracy (because abstention is not a valid ground-truth label Y_ω, abstained pixels are treated as non-robust): $\mathrm{CRPA}_{\mathrm{RS}} := \frac{1}{|\Omega|} \times \sum_{\omega \in \Omega} \mathbb{1}_{\hat{Y}_\omega = Y_\omega}$.

Comparing CRPA in Best Configurations on Cityscapes (Tab. 2) Using the common comparison metric we defined above, we compare a LipNet DeepLabV3 architecture to SegCertify applied to a standard DeepLabV3 model on the Cityscapes dataset [11]. Following the method of Fischer *et al.* [12], we train our unconstrained model with a noise level $\sigma = 0.2$ that is similar to the value we will choose for the smoothing process according to Salman *et al.* [35]. The CRPA certificates given by both Lipschitz constraints and the SegCertify method are presented in Table 2. We also report the time each evaluation took divided by the number of evaluated samples. The runtime does not scale linearly in MC iterations since the computation time of SegCertify depends on both the MC estimation runtime (which can be partially batched, but not fully due to memory limits) and the p-value computation times which are dependent on the image size. More experimental details are given in Appendix E. We observe that LipNets networks allow for rapid inference and certification with performance that remains competitive with that of randomized smoothing methods. Crucially, the performance obtained by using a LipNet is only matched by randomized smoothing methods when they require approximately 600 times more time per inference.

Comparing CRPAs on the same LipNet (Fig. 2, left) Independently, we compare the tightness of worst-case estimations from LipNets and SegCertify on the same LipNet model. This comparison will allow us to compare the tightness of approximation of both methods in a model-agnostic way. Given that increasing the number of MC samples improves the performance of randomized smoothing methods, we will determine the minimum number of MC samples that is necessary for SegCertify to match LipNet certificates under error probability $\alpha = 0.01$. Our results are reported on the left part of Fig. 2. On the Oxford-IIIT Pet dataset with 128×128 images, our results show that randomized smoothing methods need around 10^3 MC samples to become beneficial over using a Lipschitz certificate. This corresponds to an ~ 2000 times faster runtime per image at equal performance level.

Empirical Tightness and the Deterministic Gap (Fig. 2, middle) Having established the superior efficiency of LipNets, we evaluate the absolute tightness of our certificates. Because our models are deterministic, we can directly benchmark them against state-of-the-art white-box segmentation attacks (ALMA [31], ASMA [41], and PDPGD [26] via the `adversarial-library` [33]). To our knowl-

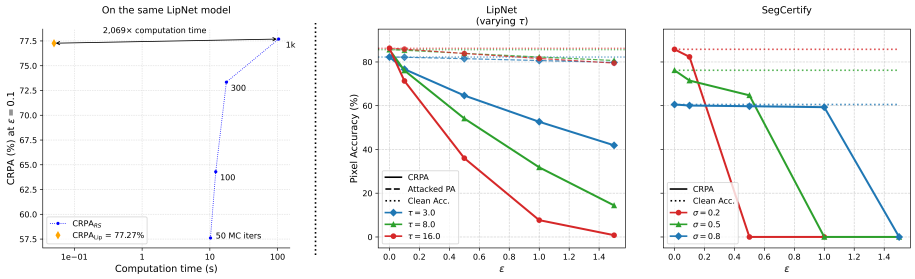

Fig. 2: (left) We evaluate LipNet certificates and SegCertify *on the same LipNet*, for $\epsilon = 0.1,\ \alpha = 0.01$, we tune σ carefully for each n_{MC}. (middle) We plot the pixel accuracy of a different LipNets under SOTA adversarial attacks along with the CRPA values for different ϵ. (right) The clean pixel accuracy (majority vote) and CRPA of SegCertify, using different σ values, with $n_{\text{MC}} = 250$ on a network trained with $\sigma_{\text{train}} = 0.2$, $\alpha = 0.001$.

edge, our method is the only certifiable segmentation framework to explicitly report certificate looseness against empirical attacks.

As shown in Fig. 2 (middle), the gap between the empirical attack performance and our certified lower bound widens as ϵ increases. This is a fundamental consequence of deterministic over-approximation (Gowal *et al.* [13]) and stems directly from the relaxation in Eq. (6). Specifically, our certificate is optimal if their exists an input perturbation $\delta \in \mathcal{B}_\epsilon(X)$ that can realize the theoretical worst-case output perturbation $\alpha^* \in \mathcal{B}_{L\epsilon}(f(X))$. In practice, this can be related to local surjectivity of the neural network outputs, which is structurally unlikely due to the high correlation of logit values of adjacent pixels in dense prediction architectures (as discussed in Section 3.2, Eq. (1) of Schuchardt *et al.* [36]). Crucially, while this assumption creates a visible deterministic gap at larger radii, these bounds remain the only viable solution for fast inference. As demonstrated previously, attempting to recover tighter bounds using SegCertify requires a 2000× computational penalty. To mitigate this phenomenon, we find that controlling the temperature parameter of the training process allows for explicit control of the empirical gap between attacks and certificates, at the slight expense of clean pixel accuracy.

Comparing to SegCertify Across ϵ Values (Fig. 2, right) Randomized Smoothing methods provide statistical guarantees of the robustness of the smoothed model defined according to smoothing parameters σ and n_{MC}. Therefore, these statistical guarantees are valid *only* on the smoothed classifier and not on the base model f. In Fig. 2 (right), we show that the performance of SegCertify heavily depends on σ across ϵ values, exposing a strict trilemma: practitioners can retain at most two of *efficiency*, *clean accuracy*, and *certified robustness*. Certifying larger radii with high accuracy demands massive n_{MC} budgets (sacrificing efficiency); keeping inference fast and accurate forces σ low,

collapsing certificates to zero (sacrificing robustness); and certifying cheaply at large ϵ requires high σ, degrading clean performance (sacrificing accuracy). Our LipNets sidestep this entirely: a single forward pass yields certificates valid across all ϵ values while retaining better clean accuracy, with explicit control of the accuracy–robustness operating point (see Fig. 2 and Appendix C).

5.3 A Safety Critical Use-Case

Here, we demonstrate how LipNets allow for fast and practical worst-case bounds in safety-critical scenarios. We perform a binary segmentation task to detect polyps from endoscopically captured images on the Kvasir-SEG dataset introduced in Jha *et al.* [15]. This is a safety-critical scenario as misdetections might lead to harmful consequences on the patient. Furthermore, the deployment of such a detection model must satisfy some real-time constraints given the time-sensitive nature of the assisted surgeon's task.

Q1		**Q2**	
ϵ	Maximum FNR value	γ	Required ϵ budget
0.1	0.612	0.7	0.423
0.2	0.768	0.85	0.563
0.3	0.871	0.95	0.675

Table 3: Lower bounds for **Q1** and **Q2** certification on the Kvasir-SEG dataset.

Here, we want to certify the worst-case FNR of our segmentation network under adversarial noise bounded by ϵ. In the following, we use h and κ defined as in Sec. 4.2. We expose our robustness certificates in the setting of both **Q1** (worst-case performance) and **Q2** (generalized robustness radius) measures are given in Table 3. Importantly, the forward pass of our LipNet only requires ≈ 0.05 s, and most importantly, our robustness bounds are purposeful and non-vacuous which ensures compatibility with real-time, safety critical applications.

6 Perspectives and Conclusion

In this paper, we propose LipNets for semantic segmentation tasks, which enables real-time compatible certifiably robust segmentation for the first time. Also, we show how to train and use Lipschitz neural networks to obtain general robustness certificates efficiently in a variety of different scenarios.

Limitations While LipNets enable efficient certification of robustness guarantees for semantic segmentation, they struggle to match the performance of smoothed state-of-the-art networks that leverage a high number of MC iterations or diffusion models [19]. Considering the performance-to-inference-time Pareto frontier,

LipNets occupy one extreme while randomized smoothing methods occupy the other. Also, our framework is currently limited to ℓ_2 norm certification.

Future Research Directions Future work could explore hybrid approaches that leverage both model architecture and input dependencies to tighten the bound of Equation 6 through feasibility constraints, while preserving real-time compatibility. Additionally, incorporating input distribution assumptions could yield faster and more meaningful certificates [14, 45].

Acknowledgements

The authors would like to thank Thibaut Boissin for his careful reviews of the paper during the writing process. Similarly, the authors would like to thank Laurent Gardès and Tom Rousseau for their continued support of the project and advisory role in the development of the method.

Our work has benefited from the AI Cluster ANITI and the research program DEEL[7]. ANITI is funded by the France 2030 program under the Grant agreement n°ANR-23-IACL-0002. DEEL is an integrative program of the AI Cluster ANITI, designed and operated jointly with IRT Saint Exupéry, with the financial support from its industrial and academic partners and the France 2030 program under the Grant agreement n°ANR-10-AIRT-01. This project was provided with computing and storage resources by GENCI at IDRIS thanks to the grant 2025-AD011016850 on the supercomputer Jean Zay's A100 and H100 partition

References

1. Anil, C., Lucas, J., Grosse, R.: Sorting out Lipschitz function approximation. In: International conference on machine learning. pp. 291–301. PMLR (2019)
2. Arnab, A., Miksik, O., Torr, P.H.: On the robustness of semantic segmentation models to adversarial attacks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
3. Bansal, N., Chen, X., Wang, Z.: Can we gain more from orthogonality regularizations in training deep networks? In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
4. Béthune, L., Boissin, T., Serrurier, M., Mamalet, F., Friedrich, C., Gonzalez Sanz, A.: Pay attention to your loss: understanding misconceptions about Lipschitz neural networks. Advances in Neural Information Processing Systems **35**, 20077–20091 (2022)
5. Boissin, T., Mamalet, F., Fel, T., Picard, A.M., Massena, T., Serrurier, M.: An adaptive orthogonal convolution scheme for efficient and flexible cnn architectures. In: Forty-second International Conference on Machine Learning (2025)

[7] https://www.deel.ai

6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
7. Chiang, P.y., Curry, M., Abdelkader, A., Kumar, A., Dickerson, J., Goldstein, T.: Detection as regression: Certified object detection with median smoothing. Advances in Neural Information Processing Systems **33**, 1275–1286 (2020)
8. Christian, S., Wojciech, Z., Joan, B., Dumitru, E., Ian, G., Rob, F., et al.: Intriguing properties of neural networks. In: ICLR (2014)
9. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: International Conference on Machine Learning. pp. 1310–1320. PMLR (2019)
10. Cohen, N., Ducoffe, M., Boumazouza, R., Gabreau, C., Pagetti, C., Pucel, X., Galametz, A.: VerifIoU – robustness of object detection to perturbations. In: 2025 AIAA DATC/IEEE 44th Digital Avionics Systems Conference (DASC) (2025)
11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
12. Fischer, M., Baader, M., Vechev, M.: Scalable certified segmentation via randomized smoothing. In: International Conference on Machine Learning. pp. 3340–3351. PMLR (2021)
13. Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., Kohli, P.: On the effectiveness of interval bound propagation for training verifiably robust models. arXiv preprint arXiv:1810.12715 (2018)
14. Hashemi, N., Sasaki, S., Oguz, I., Ma, M., Johnson, T.T.: Scaling data-driven probabilistic robustness analysis for semantic segmentation neural networks. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
15. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26. pp. 451–462 (2020)
16. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient smt solver for verifying deep neural networks. In: International conference on computer aided verification. pp. 97–117. Springer (2017)
17. Kellerer, H., Pferschy, U., Pisinger, D.: Knapsack problems. In: Knapsack problems, pp. 235–283. Springer (2004)
18. Kim, H., Papamakarios, G., Mnih, A.: The Lipschitz constant of self-attention. In: International Conference on Machine Learning. pp. 5562–5571. PMLR (2021)
19. Laousy, O., Araujo, A., Chassagnon, G., Paragios, N., Revel, M.P., Vakalopoulou, M.: Certification of deep learning models for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 611–621. Springer (2023)
20. Laousy, O., Araujo, A., Chassagnon, G., Revel, M.P., Garg, S., Khorrami, F., Vakalopoulou, M.: Towards better certified segmentation via diffusion models. In: The 39th Conference on Uncertainty in Artificial Intelligence (2023)
21. Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., Jana, S.: Certified robustness to adversarial examples with differential privacy. In: 2019 IEEE symposium on security and privacy (SP). pp. 656–672. IEEE (2019)

22. Li, Q., Haque, S., Anil, C., Lucas, J., Grosse, R.B., Jacobsen, J.H.: Preventing gradient attenuation in Lipschitz constrained convolutional networks. In: Advances in Neural Information Processing Systems (2019)
23. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
25. Luo, Y., Ma, J., Han, S., Xie, L.: Benchmarks: Semantic segmentation neural network verification and objection detection neural network verification in perceptions tasks of autonomous driving. In: Bridging the Gap Between AI and Reality: First International Conference, AISoLA 2023, Crete, Greece, October 23–28, 2023, Proceedings. p. 279–290. Springer-Verlag, Berlin, Heidelberg (2023). https://doi.org/10.1007/978-3-031-46002-9_16
26. Matyasko, A., Chau, L.P.: Pdpgd: Primal-dual proximal gradient descent adversarial attack. arXiv preprint arXiv:2106.01538 (2021)
27. Müller, M.N., Brix, C., Bak, S., Liu, C., Johnson, T.T.: The third international verification of neural networks competition (vnn-comp 2022): Summary and results. arXiv preprint arXiv:2212.10376 (2022)
28. Newhouse, L., Hess, R.P., Cesista, F., Zahorodnii, A., Bernstein, J., Isola, P.: Training transformers with enforced Lipschitz constants. arXiv preprint arXiv:2507.13338 (2025)
29. Pal, N., Lee, S., Johnson, T.T.: Benchmark: Formal verification of semantic segmentation neural networks. In: Bridging the Gap Between AI and Reality: First International Conference, AISoLA 2023, Crete, Greece, October 23–28, 2023, Proceedings. p. 311–330. Springer-Verlag, Berlin, Heidelberg (2023)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
31. Rony, J., Granger, E., Pedersoli, M., Ben Ayed, I.: Augmented lagrangian adversarial attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7738–7747 (2021)
32. Rony, J., Pesquet, J.C., Ben Ayed, I.: Proximal splitting adversarial attack for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20524–20533 (2023)
33. Rony, J., Ben Ayed, I.: Adversarial Library (2026). https://doi.org/10.5281/zenodo.5815063, https://github.com/jeromerony/adversarial-library
34. Salkin, H.M., De Kluyver, C.A.: The knapsack problem: a survey. Naval Research Logistics Quarterly **22**(1), 127–144 (1975)
35. Salman, H., Li, J., Razenshteyn, I., Zhang, P., Zhang, H., Bubeck, S., Yang, G.: Provably robust deep learning via adversarially trained smoothed classifiers. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
36. Schuchardt, J., Wollschläger, T., Bojchevski, A., Günnemann, S.: Localized randomized smoothing for collective robustness certification. In: The Eleventh International Conference on Learning Representations (2022)
37. Serrurier, M., Mamalet, F., González-Sanz, A., Boissin, T., Loubes, J.M., Del Barrio, E.: Achieving robustness in classification using optimal transport with hinge regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 505–514 (2021)

38. Tran, H.D., Pal, N., Musau, P., Lopez, D.M., Hamilton, N., Yang, X., Bak, S., Johnson, T.T.: Robustness verification of semantic segmentation neural networks using relaxed reachability. In: International conference on computer aided verification. pp. 263–286. Springer (2021)
39. Tsuzuku, Y., Sato, I., Sugiyama, M.: Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. Advances in neural information processing systems **31** (2018)
40. Virmaux, A., Scaman, K.: Lipschitz regularity of deep neural networks: analysis and efficient estimation. Advances in Neural Information Processing Systems **31** (2018)
41. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 1369–1378 (2017)
42. Yang, G., Duan, T., Hu, J.E., Salman, H., Razenshteyn, I., Li, J.: Randomized smoothing of all shapes and sizes. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 10693–10705. PMLR (13–18 Jul 2020)
43. Yang, G., Simon, J.B., Bernstein, J.: A spectral condition for feature learning. arXiv preprint arXiv:2310.17813 (2023)
44. Yin, H., Zhao, Z., Yan, J., Watzenig, D.: Certified robustness in automated driving perception: A review. Automotive Innovation **8**, 817–837 (2025). https://doi.org/10.1007/s42154-024-00347-3
45. Zargarbashi, S.H., Akhondzadeh, M.S., Bojchevski, A.: One sample is enough to make conformal prediction robust. arXiv preprint arXiv:2506.16553 (2025)