Scalable Certified Segmentation via Randomized Smoothing

Marc Fischer, Maximilian Baader, Martin Vechev

We present a new certification method for image and point cloud segmentation based on randomized smoothing [1]. Key to our approach is the ability to **abstain** from classifying single components (e.g. pixel or point cloud points) and reliance on established **multiple-testing correction** mechanisms.

Robust Classifier For a segmentation model $f\colon \mathbb{R}^N \to \mathbb{R}^N$ and threshold $\tau \in [\frac{1}{2}, 1)$ we define

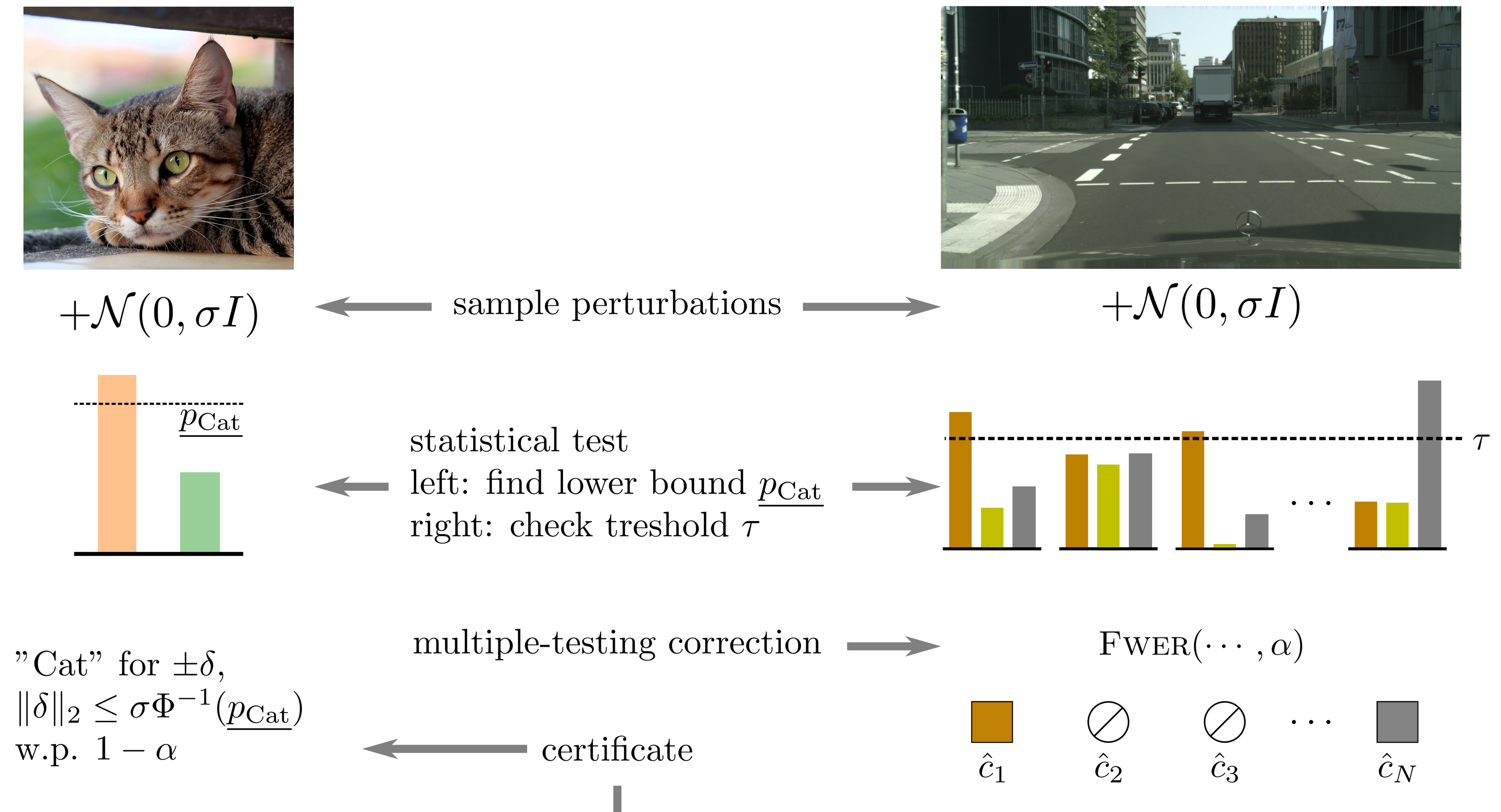
$$\bar{f}_i^\tau(\boldsymbol{x}) = \begin{cases} c_{A,i} & \text{if } \mathbb{P}_{\epsilon\sim\mathcal{N}(0,\sigma)}(f_i(\boldsymbol{x}+\epsilon) = c_{A,i}) > \tau \\ \oslash & \text{else} \end{cases}.$$

Then, letting $\mathcal{I}_{\boldsymbol{x}} = \{i \mid \bar{f}_i^\tau(\boldsymbol{x}) \neq \oslash, i \in 1, \dots, N\}$ denote the set of non-abstain indices for $\bar{f}^\tau(\boldsymbol{x})$,

$$\bar{f}_i^\tau(\boldsymbol{x}+\delta) = \bar{f}_i^\tau(\boldsymbol{x}), \quad \forall i \in \mathcal{I}_{\boldsymbol{x}}$$

for $\delta \in \mathbb{R}^{N\times m}$ with $\|\delta\|_2 \leq R := \sigma\Phi^{-1}(\tau)$.

Classification [1] / Segmentation (this work)



Segmentation Certificate

$$\hat{\mathcal{I}} = \{i \mid \hat{c}_i \neq \oslash\}$$

$$\bar{f}_i^\tau(\boldsymbol{x}) = \bar{f}_i^\tau(\boldsymbol{x}+\delta) = \hat{c}_i \quad \forall i \in \hat{\mathcal{I}}, \forall\delta \text{ with } \|\delta\|_2 \leq \sigma\Phi^{-1}(\tau)$$

w.p. $1-\alpha$

github.com/eth-sri/segmentation-smoothing

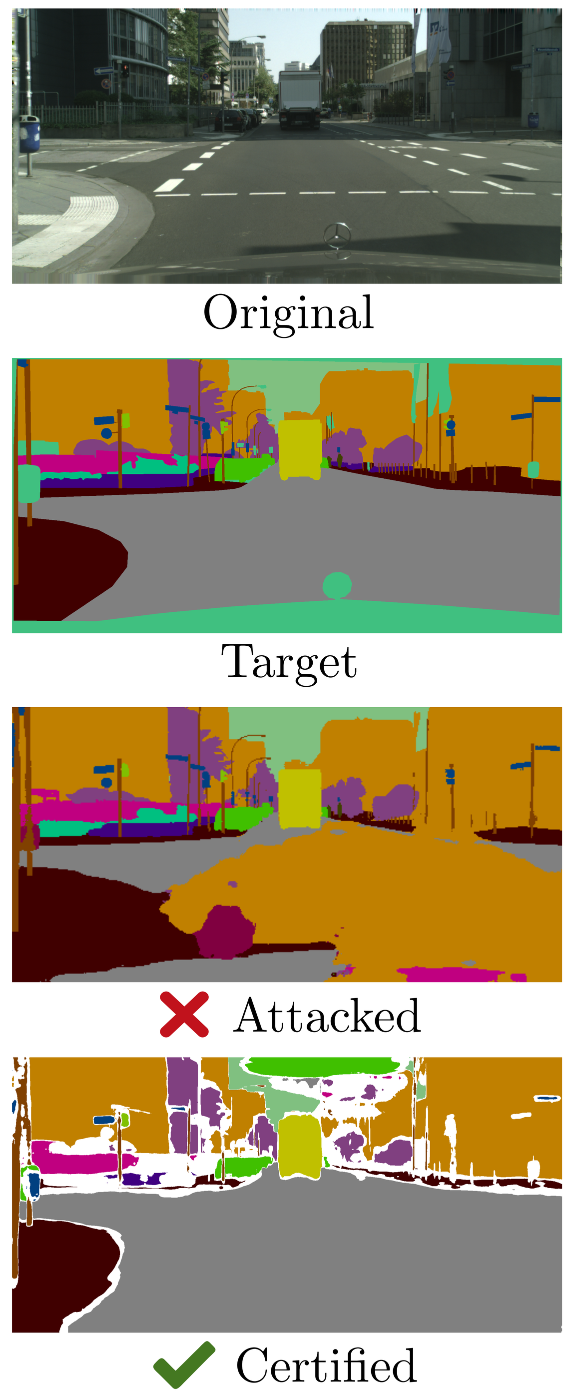


Figure 1: Image from [2]. White pixels show abstentions.

Table 1: Segmentation results for 100 images. acc. shows mean per-pixel accuracy, mIoU the mean intersection over union, %⊘ abstentions and t runtime in seconds. All SegCertify results are certifiably robust at radius R w.h.p. ($n_0 = 10, \alpha = 0.001$)

				Cityscapes			
scale		σ	R	acc.	mIoU	%⊘	t
0.25	non-robust model	-	-	0.93	0.60	0.00	0.38
	base model	-	-	0.87	0.42	0.00	0.37
	SegCertify $n = 100, \tau = 0.75$	0.25	0.17	0.84	0.43	0.07	70.00
		0.33	0.22	0.84	0.44	0.09	70.21
		0.50	0.34	0.82	0.43	0.13	71.45
0.5	non-robust model	-	-	0.96	0.76	0.00	0.39
	base model	-	-	0.89	0.51	0.00	0.39
	SegCertify $n = 100, \tau = 0.75$	0.25	0.17	0.88	0.54	0.06	75.59
		0.33	0.22	0.87	0.54	0.08	75.99
		0.50	0.34	0.86	0.54	0.10	75.72
1.0	non-robust model	-	-	0.97	0.81	0.00	0.52
	base model	-	-	0.91	0.57	0.00	0.52
	SegCertify $n = 100, \tau = 0.75$	0.25	0.17	0.88	0.59	0.11	92.75
		0.33	0.22	0.78	0.43	0.20	92.85
		0.50	0.34	0.34	0.06	0.40	92.48
multi	non-robust model	-	-	0.97	0.82	0.00	8.98
	base model	-	-	0.92	0.60	0.00	9.04
	SegCertify $n = 100, \tau = 0.75$	0.25	0.17	0.88	0.57	0.09	1040.55

Scales to large Semnatic Segmentation problems like Cityscapes [2] or Pascal Context [3].

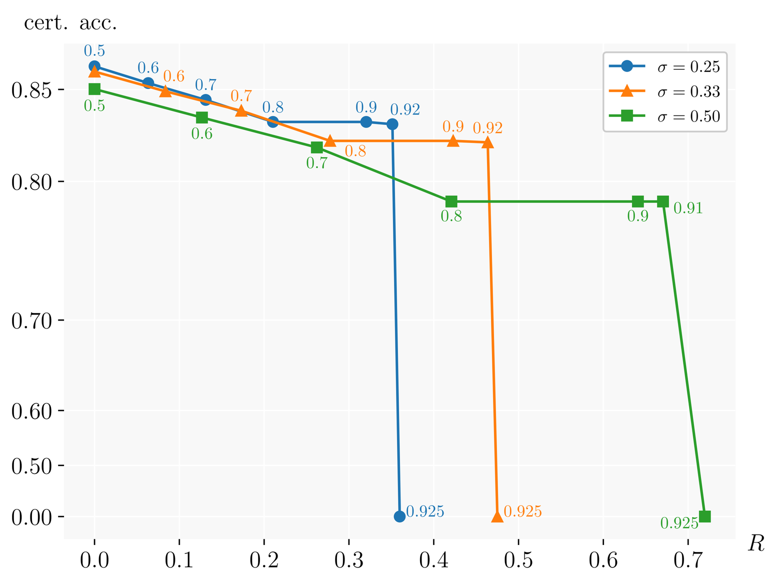


Figure 2: Radius versus certified mean per-pixel accuracy for semantic segmentation on Cityscapes at scale 0.25. Numbers next to dots show τ. The y-axis is scaled to the fourth power for clarity. ($n = 300$)

Table 2: Same setting as Table 1, for the Pascal Context dataset at scale 0.5.

			Pascal Context			
	σ	R	acc.	mIoU	%⊘	t
non-robust model	-	-	0.93	0.60	0.00	0.38
base model	-	-	0.87	0.42	0.00	0.37
SegCertify $n = 100, \tau = 0.75$	0.25	0.17	0.48	0.16	0.09	16.29
	0.33	0.22	0.50	0.17	0.11	16.08
	0.50	0.34	0.36	0.10	0.21	16.14
SegCertify $n = 300, \tau = 0.90$	0.25	0.32	0.46	0.15	0.17	27.17
	0.33	0.42	0.47	0.16	0.21	27.17
	0.50	0.64	0.31	0.10	0.38	27.32
SegCertify $n = 500, \tau = 0.95$	0.25	0.41	0.45	0.19	0.21	38.27
	0.33	0.52	0.46	0.16	0.26	38.43
	0.50	0.82	0.30	0.09	0.44	38.37

Applicable to Point Cloud Part Segmentation, such as ShapeNet [4].

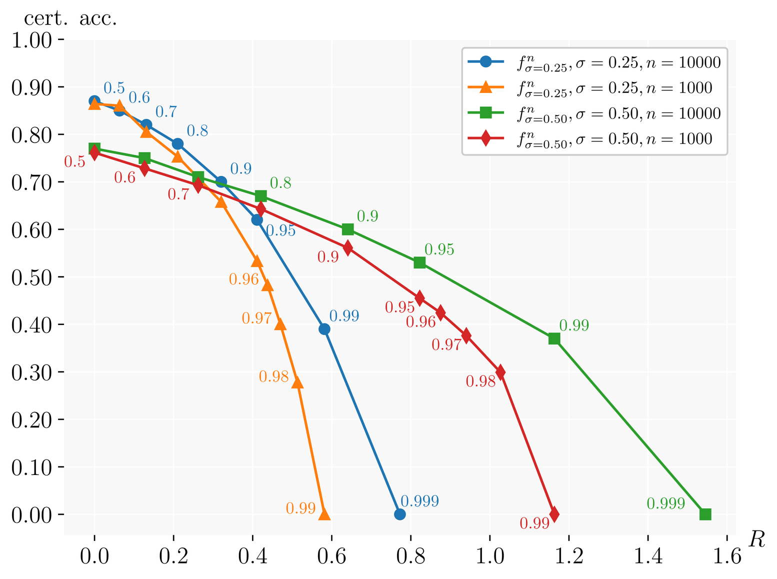


Figure 3: Radius versus certified accuracy at different radii for Point Cloud part segmentation. Numbers next to dots show τ.

Table 3: Results on 100 point clouds part segmentations, showing the mean per-point accuracy (acc), abstentions (%⊘) and t runtime in seconds.

	n	τ	σ	acc	%⊘	t
non-robust model	-	-	-	0.91	0.00	0.57
base model $f^n_{\sigma=0.25}$	-	-	-	0.86	0.00	0.57
SegCertify $f^n_{\sigma=0.25}$	1000	0.75	0.25	0.78	0.16	54.46
	1000	0.85	0.25	0.71	0.25	54.41
	10000	0.95	0.25	0.62	0.35	496.65
	10000	0.99	0.25	0.39	0.60	496.68
base model $f^n_{\sigma=0.50}$	-	-	-	0.86	0.00	0.57
SegCertify $f^n_{\sigma=0.50}$	1000	0.75	0.50	0.67	0.21	54.58
	1000	0.85	0.50	0.61	0.30	54.44
	10000	0.95	0.50	0.53	0.41	497.40
	10000	0.99	0.50	0.37	0.60	497.87

Certification of challenging specifications (such as rotations) beyond ℓ_2.

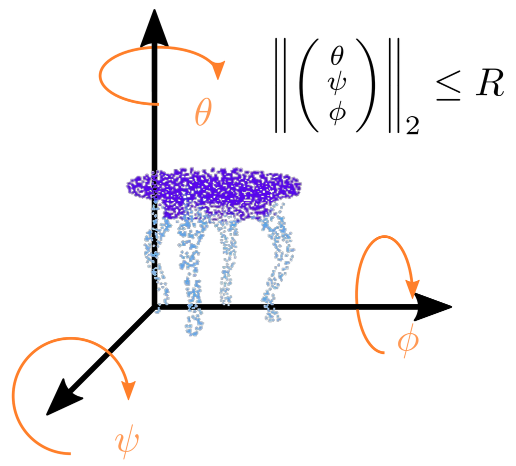


Figure 4: Certification of 3d-rotations via [5].

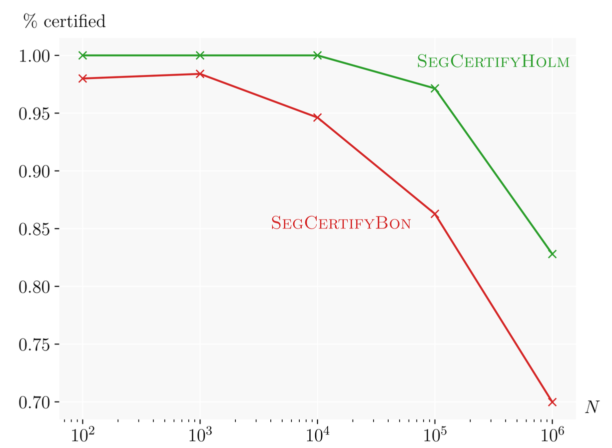


Figure 5: Difference between Bonferroni and Holm FwerControl for SegCertify.

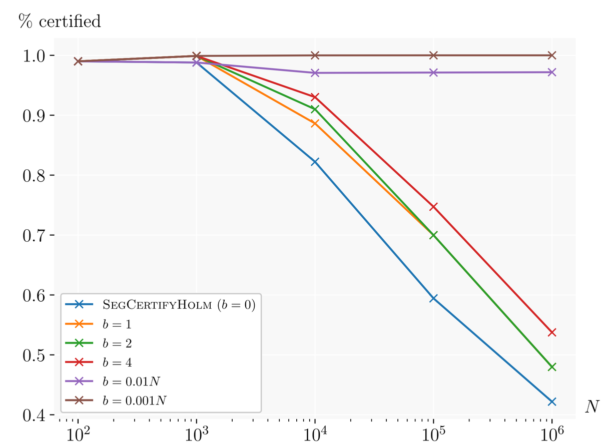


Figure 6: SegCertify with k-FwerControl that allows for up to b errors.

Error Rate Control and Error Budgets Figure 5, shows the impact of different choices of FwerControl on the number of certified components when f is an oracle with error rate 0.05. FwerControl limits the probability of making any error. k-FwerControl, displayed in Figure 6, allows up to b errors.

Advanced FwerControl allows to improve statistical power, e.g. by allowing errors for a small fraction of components.

[1] Cohen et al. Certified adversarial robustness via randomized smoothing ICML'19
[2] Cordts et al. The cityscapes dataset for semantic urban scene under-standing CVPR'16
[3] Mottaghi et al. The role of context for object detection and semantic segmentation in the wild CVPR'14
[4] Chang et al. Shapenet: An information-rich 3d model repository. 2015
[5] Fischer et al. Certified defense to image transformations via randomized smoothing NeurIPS'20

ETH zürich SRILAB