CROCO: CERTIFIED ROBUSTNESS TO OPTICAL CORRUPTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Optical systems degrade in the field through heat, mechanical decentering, and focus actuation, and recent work shows that the resulting aberrations can change the predictions of state-of-the-art image classification models. Empirical defenses reduce the risk but cannot provide guarantees. We present CROCO, the first certified robustness guarantee against physically parametrized lens aberrations. We describe each aberration by its Zernike coefficients, add Gaussian noise to these coefficients, and apply randomized smoothing. This yields an ℓ_2 certificate in coefficient space that makes no assumption about how successive aberrations combine, so the modeled combination may be arbitrarily nonlinear. The certified radius converts into a guaranteed RMS wavefront tolerance and a range of lens sharpness (MTF), which are the quantities used to specify real lenses. The same method certifies classification, semantic segmentation, and the stability of explanations. On Imagenette, ImageNet100, and ImageNet1k, certification costs only a few points of accuracy and is nearly independent of the smoothing level. Certifying each Zernike mode separately shows that models are least robust to rotationally symmetric aberrations.

1 INTRODUCTION

Vision models deployed in safety critical systems, such as the perception stack of an autonomous vehicle (Janai et al., 2021), see the world through optical systems that include a physical lens. The behavior of optical systems changes over time due to aspects such as thermal expansion, mechanical decentering, and imperfect autofocus. As a result, the network receives a blurred input before its first layer runs. Recent benchmarks parametrize these aberrations and report substantial accuracy drops (Bäuerle et al., 2026; Müller et al., 2023). Adverse Lens Corruptions (ALC) (Bäuerle et al., 2026) demonstrate that adversarially optimized lens parameters can effectively change the prediction of an otherwise accurate classifier.

Empirical defenses reduce this risk. Adversarial training (Madry et al., 2019) lowers the attack success rate, and adding an initial information bottleneck to a network can improve its stability under common corruptions and optical aberrations (Bäuerle et al., 2024). However, neither approach can rule out whether a differently parametrized aberration will succeed, and many initially successful, empirical defenses have later been broken by stronger attacks (Athalye et al., 2018a; Croce & Hein, 2020). Certified robustness provides robustness guarantees (Li et al., 2023a), and recent work has begun to extend it toward optical effects. Li et al. (2021) certify isotropic Gaussian blur, which has a single scalar parameter, and Hao et al. (2022) certify defocus blur using a learned surrogate. Both target a scalar or learned notion of blur, which leaves open the anisotropic, physically parametrized aberrations that model real optical degradation, as well as the link between the certified quantity and an optical tolerance an optical engineer could interpret.

The alternative, certifying an ℓ_2 ball in pixel space, does not solve the problem either. Lens aberrations move the image along a low dimensional set with one dimension per optical mode, while an ℓ_2 ball requires robustness in every pixel direction at once. Containing even an aberration so mild that the lens remains near perfect already moves the image by an ℓ_2 distance of about 14.8 (relative to the sharp image), i.e. roughly an order of magnitude beyond what randomized smoothing typically certifies (Appendix I). At the same time, such a ball also admits many perturbations that cannot be produced by a physical lens.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To represent the optical system efficiently, we certify directly in the standard basis for wave aberrations over a circular pupil, in which wave aberrations are parameterized over Zernike modes (Mahajan, 1994). We call the space of their coefficients *coefficient space*. We treat the optical renderer, i.e. the function that translates the coefficients into a blurred version of the image, and the base classifier as a single function, smooth it with Gaussian noise on the coefficients, and certify the resulting prediction inside an ℓ_2 ball in coefficient space (Cohen et al., 2019; Alfarra et al., 2021). Transformation-based certificates typically require that two successive transformations combine into a single one, and physically correct lens rendering does not have this property. Our certificate requires no such compositional assumption and permits the optical renderer to be arbitrarily nonlinear in the aberration parameters.

At the same time, the certified radius has a physical meaning. We convert it into an RMS wavefront tolerance in wavelengths, placing the guarantee on the scale used to characterize real lenses, and relate it to both the conventional Maréchal threshold (Born & Wolf, 1999; Wyant & Creath, 1992) and the modulation transfer function (MTF), a standard measure of lens sharpness.

Our main contributions are the following:

- We introduce **CROCO** (Certified Robustness to Optical COrruptions), to our knowledge, the first certified robustness guarantees against optical aberrations stated directly in the aberration parameter space. The same construction certifies classification, semantic segmentation, and the stability of GradCAM explanations.
- We convert the certified radius into an RMS wavefront tolerance, verified numerically against the optical renderer, and we show why an image side compositional certificate fails.
- Further, we decompose the certificate per Zernike mode and find that, across architectures, models are least robust against rotationally symmetric aberrations. Such findings cannot be produced by pixel space certification.
- We validate the certificate empirically with an expectation over transformation attack in coefficient space and we translate the radius into a modulation transfer function range that can be compared with a lens specification sheet.

Figure 1 illustrates the pipeline. The aberration coefficients $\boldsymbol{w}$ determine the blur the classifier sees. Smoothing over $\boldsymbol{w}$ certifies a ball in coefficient space within which no modeled lens state can change the smoothed prediction.

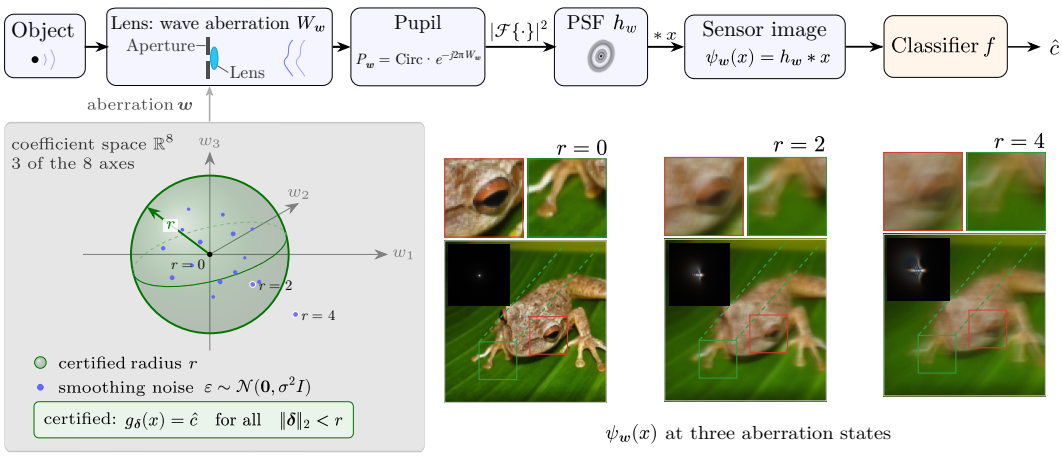


Figure 1: Overview of CROCO: certified robustness against lens aberrations. The aberration coefficients $\boldsymbol{w}$ parameterize the wave aberration $W_{\boldsymbol{w}}$, which determines the phase of the pupil function $P_{\boldsymbol{w}}$, from which the point spread function $h_{\boldsymbol{w}}$ is computed via the squared modulus of the Fourier transform. It defines the characteristic blur in the resulting sensor image $\psi_{\boldsymbol{w}}(x)$. Smoothing with Gaussian noise in coefficient space certifies the prediction inside a ball of radius r, so any modeled aberration whose coefficient vector lies within this ball provably cannot change the prediction. *Right:* The image at three coefficient norms, with the point spread function inset.

2 Related Work

Model robustness has been addressed from various perspectives. In the following, we first give a brief introduction to adversarial and physically realizable attacks and to the benchmarks that measure optical degradation non adversarially, then summarize the certification literature, both the pixel space case and the certificates that target structured semantic and parametric transformations.

Adversarial attacks and optical degradation Adversarial examples were introduced as small worst-case perturbations that flip a prediction (Szegedy et al., 2014; Goodfellow et al., 2015) and are now evaluated with standardized first-order and parameter-free attacks (Madry et al., 2019; Carlini & Wagner, 2017; Athalye et al., 2018a; Croce & Hein, 2020). Spatial and geometric attacks take the first step away from an ℓ_p threat model and toward a parametrized one, flipping predictions by deforming the image (Xiao et al., 2018; Engstrom et al., 2019). Physically realizable attacks go further and act through the imaging process itself, through printed patches and textures (Sharif et al., 2016; Brown et al., 2018; Eykholt et al., 2018; Athalye et al., 2018b), projected or emitted light (Gnanasambandam et al., 2021; Lovisotto et al., 2021; Duan et al., 2021), artifacts on the lens (Li et al., 2019b), the rolling shutter of the sensor (Sayles et al., 2021), the image signal processor (Phan et al., 2021), and physically plausible scene changes (Schmalfuss et al., 2023). Adverse Lens Corruptions moves the attack into the optics, optimizing the parameters of a lens aberration model into a worst-case optical state that changes the prediction (Bäuerle et al., 2026). The perturbation is constrained by a physical process instead of a distance in pixel space, so its natural budget is a physical parameter.

Common corruption benchmarks measure corruptions at fixed severity levels (Hendrycks & Dietterich, 2019; Kar et al., 2022; Hendrycks et al., 2021), and a frequency analysis of them shows that low-pass corruptions such as blur behave differently from high-frequency noise (Yin et al., 2020; Rusak et al., 2020). OpticsBench and LensCorruptions evaluate classifiers under parametrized lens aberrations (Müller et al., 2023; 2025), but at a few discrete severity levels, whereas a certificate is a worst-case statement over a continuous set of optical states. A separate line treats the lens as a differentiable component and optimizes it jointly with the network (Sitzmann et al., 2018; Chang & Wetzstein, 2019; Tseng et al., 2021; Côté et al., 2023) to improve accuracy.

Certified robustness Certification divides into deterministic and probabilistic families (Li et al., 2023a). Deterministic verifiers bound the network output over an input region through complete search, abstract interpretation, convex relaxation, and bound propagation (Katz et al., 2017; Gehr et al., 2018; Singh et al., 2019; Wong & Kolter, 2018; Gowal et al., 2019; Zhang et al., 2018; Xu et al., 2020; Wang et al., 2021). They are sound, but they scale poorly to large models and to the optical transformation spaces we consider. Randomized smoothing instead certifies probabilistically at large scale, from differential privacy and additive noise arguments (Lecuyer et al., 2019; Li et al., 2019a) to the tight ℓ_2 certificate of Cohen et al. (2019) and Lee et al. (2020). Smoothing also extends to dense outputs. SegCertify certifies semantic segmentation by treating every pixel as a separate hypothesis under a Holm correction (Fischer et al., 2022; Holm, 1979), AdaptiveCertify lowers the abstention rate by falling back to coarser labels in a hierarchy (Anani et al., 2024), and median smoothing certifies continuous outputs such as object detection (Chiang et al., 2022). Since attribution maps are fragile under small perturbations and fail basic sanity checks (Ghorbani et al., 2018; Adebayo et al., 2020), certified interpretation (Levine et al., 2019) and certified explanations (Anani et al., 2025) bound which parts of a map remain stable, the latter by recasting the problem as segmentation certification and thereby covering any black-box attribution method. We instantiate both the segmentation and the explanation line in coefficient space, the latter on Grad-CAM attributions (Selvaraju et al., 2019). CROCO is probabilistic, but unlike all of the above it smooths in a low-dimensional physical parameter space and not in pixel space, so the total noise stays small instead of growing with the pixel count (Cohen et al., 2019).

Certification of semantic and parametric transformations A line of work certifies structured transformations. Transformation smoothing certifies rotations and translations, introducing resolvability as the condition a transformation must satisfy to be smoothed (Fischer et al., 2020). Transformation-specific smoothing (TSS) relaxes that condition and certifies isotropic Gaussian blur (Li et al., 2021). That blur certificate is the closest relative of our work, yet it is single parameter and carries no optical interpretation (Appendix I compares the two directly). DeformRS certifies parametric deformations by smoothing in their parameter space (Alfarra et al., 2021) and has been

extended to point clouds (Pérez S. et al., 2022). Generalized smoothing extends to non-resolvable transformations, including defocus, through a learned surrogate, at the price of an approximation error instead of an exact parametric certificate (Hao et al., 2022). Related certificates cover semantic perturbations through interval analysis, transformation-dependent smoothing, and invariances (Mohapatra et al., 2020b; Korzh et al., 2024; Schuchardt & Günnemann, 2023). The deterministic family has been extended to geometric and vector-field deformations, but at a network scale well below the one we certify here (Balunović et al., 2019; Ruoss et al., 2021). Most relevant to our physical framing, camera motion smoothing certifies perception models directly in a parameter space (Hu et al., 2022). Existing blur certificates are thus either exact but isotropic or approximate, whereas ours covers anisotropic aberrations, is exact in the transformation, and has a radius that translates into lens quality metrics such as RMS wavefront error and MTF.

3 METHOD: CROCO

A perfect lens focuses the spherical wave emitted by a scene point onto a single image point. Imperfections of a real lens deform this wave by the wave aberration $W_{\boldsymbol{w}}$, and the finite aperture spreads the image point out (Figure 1). Our optical renderer $\psi_{\boldsymbol{w}}$ simulates the resulting imaging process for given aberration coefficients $\boldsymbol{w}$ and a sharp image x. Propagating the deformed wave from the lens to an intensity sensor gives the point spread function (PSF) (Goodman, 2017; Born & Wolf, 1999),

$$h_{\boldsymbol{w}} \;=\; \big|\mathcal{F}\{P_{\boldsymbol{w}}\}\big|^2, \qquad P_{\boldsymbol{w}} \;=\; \mathrm{Circ}\cdot\exp\!\big(-\,j\,2\pi\,W_{\boldsymbol{w}}\big), \tag{1}$$

where $P_{\boldsymbol{w}}$ is the pupil function, Circ the circular aperture and $2\pi W_{\boldsymbol{w}}$ the phase in radians. To parameterize the wave aberration, we expand $W_{\boldsymbol{w}}$ in the Zernike basis, the standard description of the wave aberration over a circular pupil (Noll, 1976; Mahajan, 1994; Wyant & Creath, 1992), using the primary Fringe modes 4 to 11, where Table 1 shows the corresponding wave aberrations. The rendered kernel is normalized in ℓ_1 for energy conservation, and under incoherent illumination for an isoplanar region the observed image is

$$\psi_{\boldsymbol{w}}(x) \;:=\; h_{\boldsymbol{w}} * x, \tag{2}$$

for a clean image $x \in [0,1]^{3\times H\times W}$, i.e. the optical renderer translates the Zernike coefficients into $h_{\boldsymbol{w}}$ and applies it to the image by convolution.

A base classifier f maps images to a label set Y, and the adversary is a lens, described by a coefficient vector $\boldsymbol{w} \in \mathbb{R}^d$, that renders the observed image through Equation (2). A threat of magnitude r is any lens whose aberration lies in the spanned subspace and whose coefficient vector has ℓ_2 norm at most r, which is the threat model of the ALC attack (Bäuerle et al., 2026). The threat is purely optical and contains no additive component such as sensor noise. The certificate holds for a fixed sharp image x and quantifies over the lens. Its reference state $\boldsymbol{w} = \mathbf{0}$ is an aberration free lens, which still blurs x slightly because light diffracts at the finite aperture, and its point spread function $h_{\mathbf{0}}$ is the Airy pattern, the sharpest image any lens of this aperture can form. The guarantee is therefore a statement about the lens through which x is imaged, not about additional blur applied to an image a camera has already captured. The smoothing noise is a wavefront error added in the same pupil as the aberration (Appendix A), so the guarantee applies to a smoothed imaging pipeline in which the pupil noise is realized before capture, or to an offline audit of a lens design against a fixed classifier.

aberration	Fringe		(n, m)	symmetry
defocus	4		$(2,\ 0)$	rotational
astigmatism $0/90$	5		$(2,\ 2)$	anisotropic
astigmatism ± 45	6		$(2,-2)$	anisotropic
coma x	7		$(3,\ 1)$	anisotropic
coma y	8		$(3,-1)$	anisotropic
spherical	9		$(4,\ 0)$	rotational
trefoil x	10		$(3,\ 3)$	anisotropic
trefoil y	11		$(3,-3)$	anisotropic

Table 1: The eight primary Zernike modes that span our threat model, in Fringe ordering, each shown as its wave aberration over the pupil. Here n is the radial order and m the azimuthal frequency of the Zernike polynomial Z_n^m.

In the achromatic parametrization, which is our main setting, one set of eight modes is shared across the color channels, giving $d = 8$. In the chromatic parametrization each channel carries its own eight modes, giving $d = 24$, which models the fact that a real lens aberrates each wavelength differently. Everything below is stated for a general dimension d, and the certified radius is dimension free, a

prediction we test directly in Appendix C. The only distinction is the physical unit of the radius, which we make precise in Appendix F.

Parameter space smoothing and the certificate Randomized smoothing replaces the prediction of a base classifier with the class that is most likely under noise applied to its input, and certifies that prediction inside an ℓ_2 ball whose radius grows with the margin between the top class probability and the runner up (Cohen et al., 2019). The radius is tight for the information the bound uses (Cohen et al., 2019). The class probabilities are not available in closed form and are estimated by Monte Carlo, with a Clopper and Pearson lower confidence bound and an abstention when that bound does not exceed one half (Clopper & Pearson, 1934). We apply this machinery not in pixel space but in the coefficient space of Equation (2).

For a fixed image x, noise level σ, and shifted aberration $\boldsymbol{\delta} \in \mathbb{R}^d$, define the smoothed classifier in coefficient space by

$$g_{\boldsymbol{\delta}}(x) \;:=\; \operatorname*{arg\,max}_{c \in Y} \; \mathbb{P}_{\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d)} \big[\, f(\psi_{\boldsymbol{\delta}+\boldsymbol{\varepsilon}}(x)) = c \,\big]. \tag{3}$$

Certificate 1 (In coefficient space)**.** *Fix x. Suppose there exist a class $c_A \in Y$ and probabilities $\underline{p_A}, \overline{p_B} \in [0,1]$ with*

$$\mathbb{P}_{\boldsymbol{\varepsilon}}\big(f(\psi_{\boldsymbol{\varepsilon}}(x)) = c_A\big) \;\geq\; \underline{p_A} \;\geq\; \overline{p_B} \;\geq\; \max_{c \neq c_A} \mathbb{P}_{\boldsymbol{\varepsilon}}\big(f(\psi_{\boldsymbol{\varepsilon}}(x)) = c\big). \tag{4}$$

Then $g_{\boldsymbol{\delta}}(x) = c_A$ for every $\boldsymbol{\delta} \in \mathbb{R}^d$ with

$$\|\boldsymbol{\delta}\|_2 \;<\; r \;:=\; \frac{\sigma}{2}\Big(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})\Big). \tag{5}$$

In the binary form with $\overline{p_B} = 1 - \underline{p_A}$ this reduces to $r = \sigma\, \Phi^{-1}(\underline{p_A})$.

Folding the renderer into the base classifier turns the proof into a one line reduction to the classical randomized smoothing bound, as described in Appendix E. The proof uses no property of the optical operator $\psi_\cdot$. It holds without additivity, invertibility, or resolvability, so the entire renderer, including the Fourier transform, the squared magnitude, the binning, and the convolution, may be arbitrarily nonlinear in $\boldsymbol{w}$. The same reduction underlies the parametric certificate of DeformRS (Alfarra et al., 2021). We apply it to a physical parameter space.

Dense prediction and explanations The construction is not tied to a single label, and randomized smoothing has been carried to dense outputs before (Fischer et al., 2022; Chiang et al., 2022). For semantic segmentation we certify every pixel of the output map, where a Holm correction (Holm, 1979) makes the guarantee hold jointly over all pixels of the image, as in the *flat* scheme of SegCertify (Fischer et al., 2022). We also run the *hierarchical* scheme of AdaptiveCertify (Anani et al., 2024), where a pixel that fails at the finest label may still be certified at a coarser one instead of abstaining. The certified quantity is again an RMS wavefront error, now attached to each pixel. This construction also certifies the stability of black box attributions such as GradCAM (Selvaraju et al., 2019). Both extensions inherit Certificate 1 unchanged, since neither touches the optical operator.

Certification and training Algorithm 1 states the certification procedure, which follows the certify routine of Cohen et al. (2019) adapted to kernel rendering. Each sampled aberration requires one point spread function rendering and one convolution.

The base classifier is trained with noise augmentation in coefficient space. Each image is rendered through $\psi_{\boldsymbol{\varepsilon}}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$, using the same sampling distribution that is used at certification time, so that training and certification are matched. We smooth with a Gaussian because the RMS wavefront error of Equation (9) is a weighted ℓ_2 norm of the coefficients, so an ℓ_2 certificate is the one that converts into a tolerance holding in every direction. Uniform smoothing would give an ℓ_1 certificate instead (Alfarra et al., 2021), but an ℓ_1 ball of radius r shrinks by a factor of 2.8 already at

Algorithm 1 CERTIFY in coefficient space

Require: base classifier f, image x, noise σ, sample sizes n_0, n, confidence α

counts$_0 \leftarrow$ class counts of $f(\psi_{\boldsymbol{\varepsilon}}(x))$ over n_0 samples $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$
$\hat{c}_A \leftarrow$ top class of counts$_0$
$n_A \leftarrow$ number of $f(\psi_{\boldsymbol{\varepsilon}}(x)) = \hat{c}_A$ over n samples
$\underline{p_A} \leftarrow$ LOWERCONFBOUND$(n_A, n, 1-\alpha)$
if $\underline{p_A} > \frac{1}{2}$ **then**
    return $\big(\hat{c}_A,\ \sigma\, \Phi^{-1}(\underline{p_A})\big)$
else
    return ABSTAIN
end if

$d = 8$. The dimension stays small, with $d = 8$ in the achromatic case and $d = 24$ in the chromatic case, so the total noise norm $\sigma\sqrt{d}$ stays small and large radii remain attainable at moderate σ, unlike in pixel space where d is the full image resolution (Cohen et al., 2019).

4 EXPERIMENTS

We first report certified accuracy and radius across models and datasets and show that smoothing over aberrations costs little accuracy (Section 4.1). We then certify each Zernike mode separately to see which aberrations models tolerate least (Section 4.2). Section 4.3 tests the certificate against an attack adapted to the defense and against the ALC attack (Bäuerle et al., 2026), and Section 4.4 extends the certification to segmentation and to explanations.

We evaluate on Imagenette (Howard, 2019), ImageNet100 (Tian et al., 2020), and ImageNet1k (Russakovsky et al., 2015), certifying five architectures from four families: ResNet18 and ResNet50 (He et al., 2015), EfficientNet B0 (Tan & Le, 2020), ConvNeXt tiny (Liu et al., 2022b), and SwinV2 tiny (Liu et al., 2022a). All models are trained from scratch with matched coefficient space noise augmentation over $\sigma \in \{0.5, 1.0, \ldots, 5.0\}$. Full training configurations are described in Appendix G. Certification follows Algorithm 1 with $n_0 = 200$, $n = 10^4$, and $\alpha = 10^{-3}$. Certified accuracy at radius zero, written acc@0, is the fraction of evaluated images that are correctly classified and not abstained on. The average certified radius, written ACR, is the mean of the certified radius over the same set of evaluated images, with an abstained or misclassified image contributing a radius of zero, which is the convention of Zhai et al. (2022) and Jeong & Shin (2021). All radii and σ values in this section are in the coefficient units of Equation (7). Equation (10) converts them to a guaranteed RMS wave aberration tolerance, so a certified radius of 1.0 guarantees every lens within 0.125 waves. Sampling the sphere $\|\boldsymbol{w}\|_2 = r$ and rendering turns the same radius into a band of lens sharpness, which is the most direct physical reading the certificate admits: a model certified to $r = 1.0$ covers every lens whose MTF50 falls anywhere in $[0.069, 0.201]$ cycles per pixel, against 0.287 for the diffraction limited system, that is down to 24 percent of nominal sharpness. Appendix F gives the band over the full radius range.

4.1 CERTIFIED ACCURACY AND THE SAMPLING CEILING

Table 2 reports acc@0 and the average certified radius at $\sigma = 1.0$, together with the ACR at two further noise levels. Appendix H plots both metrics against σ for every model and dataset.

			$\sigma = 1.0$						ACR	
Dataset	Model	clean acc. (nominal)	acc@0	abstain	ACR	$r \geq 1$	$r \geq 2$	$r \geq 3$	$\sigma = 2$	$\sigma = 4$
Imagenette	ResNet18	89.4	**86.9**	0.4	**2.55**	**83.7**	**80.0**	**67.5**	**4.71**	**9.35**
	ResNet50	82.3	74.8	1.5	2.08	69.6	63.7	52.5	3.96	8.23
	EfficientNet B0	82.9	79.5	2.8	2.08	73.1	62.5	43.6	3.91	7.70
	SwinV2 tiny	85.0	82.0	2.2	2.24	76.4	68.7	53.4	4.06	7.77
	ConvNeXt tiny	84.4	81.7	1.4	2.38	77.6	72.6	66.1	4.59	9.07
ImageNet100	ResNet18	81.4	**76.6**	3.0	**2.07**	**71.2**	**63.9**	**47.7**	**3.71**	**7.08**
	ResNet50	81.8	74.8	4.6	2.00	68.6	60.9	46.6	3.66	6.59
	SwinV2 tiny	78.1	74.4	4.5	2.00	68.8	60.9	45.1	3.65	7.01
	ConvNeXt tiny	74.6	69.1	4.0	1.90	64.7	57.6	45.7	3.54	6.73
ImageNet1k	ResNet18	66.6	61.2	9.0	1.33	50.5	38.4	18.7	2.39	4.20
	ResNet50	71.5	64.4	9.2	1.41	53.4	41.3	21.1	2.52	4.03
	SwinV2 tiny	79.7	**75.9**	4.2	**2.03**	**70.1**	**61.8**	**45.9**	**3.72**	**7.08**

Table 2: Certified robustness at $\sigma = 1.0$, with the average certified radius at two further noise levels. $r \geq k$ is the fraction of images that are both correctly classified and certified to radius at least k. "Clean acc. (nominal)" is the accuracy of a nominally trained model, without coefficient space augmentation, on clean images. For reference, $r = 1.0$ already guarantees ≥ 0.125 waves RMS in every direction and bounds lens sharpness to MTF50 $\in [0.069, 0.201]$ cycles per pixel.

Certified accuracy changes little with the noise level, which makes the certificate practical. On Imagenette, ResNet18 loses 4.8 points of certified accuracy, from 88.0 to 83.2, and ConvNeXt tiny 6.8 points, while σ grows tenfold from 0.5 to 5.0. On ImageNet1k, SwinV2 tiny goes from 76.8 to 71.0 percent as σ grows from 0.5 to 4.0. Pixel space Gaussian smoothing on ImageNet costs tens of accuracy points already at $\sigma = 1.0$ (Cohen et al., 2019). The aberration is a low pass filter, not additive noise, so it removes high spatial frequencies but preserves the global shape and color statistics on which these classifiers rely most, keeping the augmented training distribution close to the clean one (Yin et al., 2020).

The certified radius grows almost linearly with σ. This is a property of the certification bound and not of the models. With $n = 10^4$ samples, no image can be certified beyond $3.199\,\sigma$ (Equation (8)), and most certified images come close to this limit (Table 7). Certifying the same images with up to a hundred times more samples raises the radius but leaves certified accuracy unchanged (Appendix E). A model trained without coefficient-space noise loses up to 54 points of certified accuracy over the same range of σ (Appendix J).

4.2 Which aberration hurts most

Using the per mode certification of Appendix B, we certify each of the eight Zernike modes in isolation, with the other seven held at zero. This isolates which specific optical defect a network tolerates least, a question that is inaccessible in pixel space. Figure 2 reports the average certified radius per mode, normalized within each configuration to its strongest mode, Table 10 gives the same numbers. In all eight configurations, spanning five architectures and four noise levels, the two weakest modes are the same, defocus followed by spherical aberration. Averaged over all configurations the $m = 0$ modes reach a normalized ACR of 0.935 while the anisotropic $m \neq 0$ modes reach 0.994, a gap of 6.0 percent.

One hypothesis is that a rotationally symmetric aberration attenuates all spatial frequency orientations at once, whereas astigmatism, coma, and trefoil blur along particular axes and leave the orthogonal orientations comparatively sharp. Thus a model can still recover the object from those orientations. This agrees with the adversarial behavior reported for ALC, where defocus dominates the optimized kernel and removing it recovers a large part of the lost accuracy (Bäuerle et al., 2026).

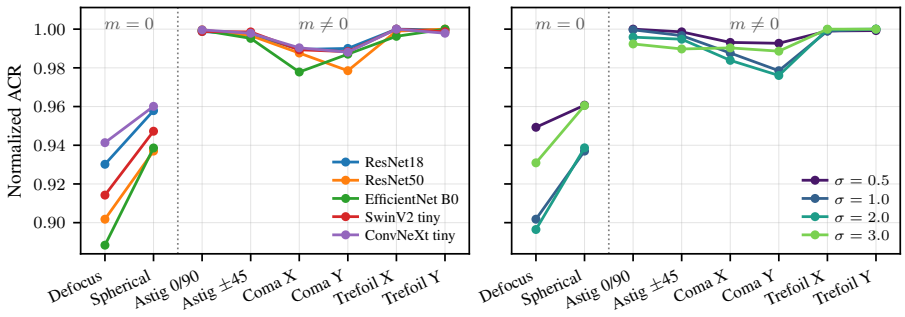


Figure 2: Average certified radius per individual Zernike mode on Imagenette, normalized per configuration to the strongest mode, across architectures at $\sigma = 1.0$ (left) and across noise level for ResNet50 (right). Modes are grouped by azimuthal order, the two rotationally symmetric ones left of the divider. Table 10 gives the underlying values.

Extending the coefficient space to $d = 24$ by giving each color channel its own eight mode wave aberration, the chromatic variant of Section 3, leaves the certified radius nearly unchanged, as the dimension free certificate of Certificate 1 predicts. Appendix C reports the full comparison.

4.3 The certificate under attack

A certificate is only as trustworthy as its implementation, so we test robustness with an attack adapted to the defense (Athalye et al., 2018a; Carlini & Wagner, 2017). We run expectation over transformation PGD (Madry et al., 2019; Athalye et al., 2018b) directly in coefficient space, searching for a wave aberration that flips the smoothed classifier. Because the certificate is granted per

image, the attack budget is too. Each image i is attacked at $\varepsilon = \kappa\, r_i$, where r_i is that image's own certified radius, so any flip at $\kappa \leq 1$ would be a soundness violation. Table 3 reports ResNet50 on Imagenette, and Appendix K gives the protocol and the spread behind the median radii. *Soundness*: the smoothed classifier never flips inside the certified radius, zero flips across 600 attacks spanning three noise levels and two budgets. *Non vacuity*: the undefended base classifier flips inside that same region, on up to 29 images (Figure 9), so the certified region is not one in which nothing happens. *Tightness*: at $\sigma \leq 1.0$ the smoothed model begins to fail immediately outside the certified region. The bound is therefore close to the truth at small σ and markedly conservative at large σ.

Because ALC (Bäuerle et al., 2026) was optimized against an undefended classifier with no knowledge of this construction, we compare its required attack magnitude directly to the certified radius. We take the per image ALC kernels at threshold $\tau = 4$, which are chromatic with $d = 24$, optimized for ResNet50 over all 50,000 ImageNet1k validation images, and place their coefficient norms on the same axis as our certified radii. The per image comparison in Appendix L finds that the certified radius covers 69 to 84 percent of the images the attack actually succeeds on for $\sigma \geq 2$. We additionally evaluate the smoothed classifier directly at the coefficient vector ALC found for each image. Inside the certified radius this never flips, exactly as Certificate 1 guarantees. Outside it, where the proof makes no claim, the flip rate is only 4.7 percent at $\sigma = 1.0$ and zero from $\sigma = 3.0$ on. Most of the coverage gap is therefore due to the conservatism of the bound (see also Appendix L).

		inside		outside		
σ (med. r_i)	clf.	$0.5r_i$	$1.0r_i$	$1.5r_i$	$2.0r_i$	$4.0r_i$
0.5 (1.60)	base	1	3	37	56	79
	ours	**0**	**0**	22	44	77
1.0 (3.20)	base	11	16	50	65	73
	ours	**0**	**0**	12	33	56
2.0 (6.40)	base	12	29	44	47	47
	ours	**0**	**0**	**0**	7	24

Table 3: Flip rate in percent under expectation over transformation PGD, ResNet50 on Imagenette, 100 images per cell. Budget is per image, $\varepsilon = \kappa\, r_i$, so the row label is the *median* r_i, not a shared budget.

4.4 Certifying dense prediction

We train DeepLabV3 (Chen et al., 2017) and FCN (Long et al., 2015) heads on a ResNet50 backbone on VOC2012 (Everingham et al., 2010) at four noise levels and certify with both the flat and the hierarchical scheme of Section 3. Table 4 reports the outcome, where the hierarchical scheme dominates the flat one on certified accuracy in all eight configurations, by 0.40 to 1.65 points, and the margin grows with the noise level.

			flat (SegCertify)			hierarchical (AdaptiveCertify)		
Model	σ	r	cert. ratio	cert. acc.	mIoU	cert. ratio	cert. acc.	CIG
DeepLabV3	0.5	0.337	91.97	88.67	0.672	92.43	**89.07**	0.888
	1.0	0.675	86.14	83.46	0.592	87.18	**84.46**	0.837
	1.5	1.012	84.58	81.58	0.499	85.62	**82.58**	0.819
	2.0	1.349	80.77	78.65	0.490	82.01	**79.84**	0.790
FCN	0.5	0.337	90.75	86.43	0.543	91.55	**87.22**	0.866
	1.0	0.675	86.06	82.73	0.454	87.02	**83.68**	0.830
	1.5	1.012	83.09	80.37	0.424	84.27	**81.54**	0.807
	2.0	1.349	81.09	78.18	0.371	82.81	**79.83**	0.786

Table 4: Certified semantic segmentation on VOC2012 with a ResNet50 backbone ($n_0 = 10$, $n = 100$). Certified ratio and certified accuracy, both in percent, are the fraction of pixels certified and the fraction both certified and correct. CIG is the certified information gain of Anani et al. (2024), measuring how much of the fine grained label survives a fallback to a coarser one.

The accuracy conditional on certification, that is the fraction of certified pixels that are correct, is essentially flat at 95.2 to 97.4 percent across every model, scheme, and noise level. Almost the entire accuracy drop as σ grows comes from pixels abstaining, and not from the certificate becoming less reliable. The mIoU, however, degrades much faster than pixel accuracy. VOC2012 pixels are dominated by background and a few large objects, so pixel accuracy is more stable than mIoU. We

therefore lead with mIoU and with the certified information gain, which measures how much of the fine grained label survives a fallback and stays between 0.786 and 0.888 across all cells.

The dense regime above is two orders of magnitude cheaper than the classification one, which caps the segmentation radii well below the classification ones and is visible in Table 4. Because a pixel is certified by a threshold test rather than a confidence bound, its certified radius is fixed regardless of n. Recertifying DeepLabV3 at a fivefold budget (Appendix M) lowers the abstain rate and raises certified accuracy and mIoU substantially, at a small cost in accuracy conditional on certification, certifying more of the same pixels rather than extending their radius.

Finally we certify which pixels of a GradCAM attribution map remain in the top k percent under every aberration inside the radius, for ResNet50 on Imagenette. Table 5 shows that stability decreases monotonically with σ at every k, and, at every σ, decreases further as k shrinks. At the mildest noise level tested, 26.4 percent of the pixels in the broad top 30 percent mask certify as stable, against only 4.0 percent of the pixels in the narrow top 5 percent mask. Explanations under optical degradation are thus only partially stable.

σ	r	top 30%	top 20%	top 10%	top 5%
0.5	0.337	26.41	17.61	8.39	3.99
1.0	0.675	25.22	16.32	7.65	3.54
2.0	1.349	23.03	14.49	6.42	2.49

Table 5: Fraction in percent of the top k percent GradCAM mask that is certifiably stable under all aberrations inside the certified radius, ResNet50 on Imagenette.

5 Discussion

Our experiments show that a physically grounded threat model exposes structure pixel space certification cannot see, namely which optical modes a network tolerates least, how the guarantee holds against an attack built for the undefended classifier, and how it extends to segmentation and explanations. This guarantee has two limits, the tightness of the sampling bound and how far the certified subspace covers real degradations.

As Table 7 shows, most certified images saturate the bound of Equation (8), so every radius we report is a lower bound on what the same model would certify at a larger n, and differences between models are compressed. Table 8 measures this. Raising n by two orders of magnitude moves the ceiling from $3.199\,\sigma$ to $4.347\,\sigma$ and the ACR by 26 to 30 percent, at a hundredfold cost and with certified accuracy unchanged. In physical terms this shifts the guaranteed tolerance at $\sigma = 1.0$ from 0.400 to 0.543 waves RMS (Equation (10)). The higher order bounds of Mohapatra et al. (2020a) and the second smoothing distribution of Li et al. (2023b) extract a larger radius from the same samples by using more than a single class probability, and both apply here unchanged because the proof uses no property of the optical operator. The attack of Section 4.3 makes the gap measurable.

The certificate holds in the spanned subspace of eight Zernike modes, and in the chromatic setting of eight modes per channel. This subspace covers the primary aberrations of lens design but not the higher order residuals manufacturing can introduce (Bäuerle et al., 2026), so the certificate covers the modeled aberrations rather than every possible lens defect. It excludes sensor noise and point spread functions that vary across the field of view, since we render one kernel for the whole frame and assume the image is a small crop of a larger one. It is also a statement about the lens state relative to the image rather than a runtime guarantee for an already degraded camera image against additional digital blur. The radius also converts into the MTF, a standard measure of optical image quality, so the guarantee can be checked against lens specifications under matched conditions without running the network.

6 Conclusion

We introduced CROCO, certified robustness against adverse lens corruptions, smoothing a classifier with Gaussian noise in coefficient space to obtain a valid ℓ_2 certificate that needs no resolvability assumption. The radius converts in closed form into a guaranteed RMS wavefront tolerance, readable through the Maréchal criterion and the modulation transfer function. A certified radius of 1.0 already covers lenses whose sharpness drops to 24 percent of that of an aberration free lens. The certificate also decomposes per Zernike mode, which pixel space certification cannot do.

ETHICS STATEMENT

This work involves no user studies or other research with human participants and uses only public datasets. The attack in Section 4.3 only checks that the introduced certificate holds and combines already published techniques. The certificate covers only aberrations of the optical model within the certified radius, not other degradations such as sensor noise or motion blur, and thus should not be read as a safety guarantee for a deployed camera.

REPRODUCIBILITY STATEMENT

The proof of Certificate 1 is in Appendix E, and the optical model it is applied to is fully specified in Section 3, with the normalization of the Zernike coefficients in Appendix D. Certification follows Algorithm 1, and we report the sample sizes and confidence level we used in Section 4. Appendix G lists the training configuration for every model and dataset, as well as the settings for segmentation. The baselines, the attacks on the certificate, and the soundness check are described in Appendices I, K and L. All datasets are publicly available.

AI USE STATEMENT

We used a large language model to draft and rephrase text that we then edited and to correct grammar and spelling. We did not use generative AI tools for the research itself, including the theory, the experiments, and the analysis. We have reviewed all AI-assisted text and take full responsibility for the content of this paper.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps, 2020. URL https://arxiv.org/abs/1810.03292.

Motasem Alfarra, Adel Bibi, Naeemullah Khan, Philip H. S. Torr, and Bernard Ghanem. DeformRS: Certifying input deformations with randomized smoothing. *CoRR*, abs/2107.00996, 2021. URL https://arxiv.org/abs/2107.00996.

Alaa Anani, Tobias Lorenz, Bernt Schiele, and Mario Fritz. Adaptive hierarchical certification for segmentation using randomized smoothing, 2024. URL https://arxiv.org/abs/2402.08400.

Alaa Anani, Tobias Lorenz, Mario Fritz, and Bernt Schiele. Pixel-level certified explanations via randomized smoothing, 2025. URL https://arxiv.org/abs/2506.15499.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, 2018a. URL https://arxiv.org/abs/1802.00420.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples, 2018b. URL https://arxiv.org/abs/1707.07397.

Mislav Balunović, Maximilian Baader, Gagandeep Singh, Timon Gehr, and Martin Vechev. Certifying geometric robustness of neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f7fa6aca028e7ff4ef62d75ed025fe76-Abstract.html.

Kai Bäuerle, Patrick Müller, Syed Muhammad Kazim, Ivo Ihrke, and Margret Keuper. Learning the essential in less than 2k additional weights - a simple approach to improve image classification stability under corruptions. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=i2SuGWtIIm.

Kai Bäuerle, Patrick Müller, Ivo Ihrke, and Margret Keuper. Parameterized adverse lens corruptions to probe model robustness to optical tolerances. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=a93BmQRNxC.

Max Born and Emil Wolf. *Principles of Optics*. Cambridge University Press, 7 edition, 1999.

Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch, 2018. URL https://arxiv.org/abs/1712.09665.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017. URL https://arxiv.org/abs/1608.04644.

Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection, 2019. URL https://arxiv.org/abs/1904.08601.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. URL https://arxiv.org/abs/1706.05587.

Ping-yeh Chiang, Michael J. Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and Tom Goldstein. Detection as regression: Certified object detection by median smoothing, 2022. URL https://arxiv.org/abs/2007.03730.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2331986.

Jeremy Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. *CoRR*, abs/1902.02918, 2019. URL http://arxiv.org/abs/1902.02918.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, 2020. URL https://arxiv.org/abs/2003.01690.

Geoffroi Côté, Fahim Mannan, Simon Thibault, Jean-François Lalonde, and Felix Heide. The differentiable lens: Compound lens search over glass surfaces and materials for object detection, 2023. URL https://arxiv.org/abs/2212.04441.

Ranjie Duan, Xiaofeng Mao, A. K. Qin, Yun Yang, Yuefeng Chen, Shaokai Ye, and Yuan He. Adversarial laser beam: Effective physical-world attack to dnns in a blink, 2021. URL https://arxiv.org/abs/2103.06504.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness, 2019. URL https://arxiv.org/abs/1712.02779.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2010. doi: 10.1007/s11263-009-0275-4.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1625–1634, 2018. doi: 10.1109/CVPR.2018.00175.

Marc Fischer, Maximilian Baader, and Martin Vechev. Certified defense to image transformations via randomized smoothing. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8404–8417. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/5fb37d5bbdbbae16dea2f3104d7f9439-Paper.pdf.

Marc Fischer, Maximilian Baader, and Martin Vechev. Scalable certified segmentation via randomized smoothing, 2022. URL https://arxiv.org/abs/2107.00228.

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2018. doi: 10.1109/SP.2018.00058.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile, 2018. URL https://arxiv.org/abs/1710.10547.

Abhiram Gnanasambandam, Alex M. Sherman, and Stanley H. Chan. Optical adversarial attack, 2021. URL https://arxiv.org/abs/2108.06247.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

Joseph W. Goodman. *Introduction to Fourier Optics*. W. H. Freeman and Company, New York, NY, 4th edition, 2017. ISBN 978-1-319-11916-4.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models, 2019. URL https://arxiv.org/abs/1810.12715.

Zhongkai Hao, Chengyang Ying, Yinpeng Dong, Hang Su, Jun Zhu, and Jian Song. Gsmooth: Certified robustness against semantic transformations via generalized randomized smoothing, 2022. URL https://arxiv.org/abs/2206.04310.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. URL https://arxiv.org/abs/1903.12261.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization, 2021. URL https://arxiv.org/abs/2006.16241.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4615733.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet. https://github.com/fastai/imagenette, 2019.

Hanjiang Hu, Zuxin Liu, Linyi Li, Jiacheng Zhu, and Ding Zhao. Robustness certification of visual perception models via camera motion smoothing, 2022. URL https://arxiv.org/abs/2210.04625.

Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state of the art, 2021. URL https://arxiv.org/abs/1704.05519.

Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers, 2021. URL https://arxiv.org/abs/2006.04062.

Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation, 2022. URL https://arxiv.org/abs/2203.01441.

Guy Katz, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks, 2017. URL https://arxiv.org/abs/1702.01135.

Dmitrii Korzh, Mikhail Pautov, Olga Tsymboi, and Ivan Oseledets. General lipschitz: Certified robustness against resolvable semantic transformations via transformation-dependent randomized smoothing, 2024. URL https://arxiv.org/abs/2309.16710.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy, 2019. URL https://arxiv.org/abs/1802.03471.

Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi S. Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers, 2020. URL https://arxiv.org/abs/1906.04948.

Alexander Levine, Sahil Singla, and Soheil Feizi. Certifiably robust interpretation in deep learning, 2019. URL https://arxiv.org/abs/1905.12105.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise, 2019a. URL https://arxiv.org/abs/1809.03113.

Juncheng Li, Frank R. Schmidt, and J. Zico Kolter. Adversarial camera stickers: A physical camera-based attack on deep learning systems, 2019b. URL https://arxiv.org/abs/1904.00759.

Linyi Li, Maurice Weber, Xiaojun Xu, Luka Rimanic, Bhavya Kailkhura, Tao Xie, Ce Zhang, and Bo Li. Tss: Transformation-specific smoothing for robustness certification, 2021. URL https://arxiv.org/abs/2002.12398.

Linyi Li, Tao Xie, and Bo Li. Sok: Certified robustness for deep neural networks, 2023a. URL https://arxiv.org/abs/2009.04131.

Linyi Li, Jiawei Zhang, Tao Xie, and Bo Li. Double sampling randomized smoothing, 2023b. URL https://arxiv.org/abs/2206.07912.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution, 2022a. URL https://arxiv.org/abs/2111.09883.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022b. URL https://arxiv.org/abs/2201.03545.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2015. URL https://arxiv.org/abs/1411.4038.

Giulio Lovisotto, Henry Turner, Ivo Sluganovic, Martin Strohmeier, and Ivan Martinovic. Slap: Improving physical adversarial examples with short-lived adversarial perturbations, 2021. URL https://arxiv.org/abs/2007.04137.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. URL https://arxiv.org/abs/1706.06083.

Virendra N. Mahajan. Zernike circle polynomials and optical aberrations of systems with circular pupils. *Appl. Opt.*, 33(34):8121–8124, Dec 1994. doi: 10.1364/AO.33.008121. URL https://opg.optica.org/ao/abstract.cfm?URI=ao-33-34-8121.

Jeet Mohapatra, Ching-Yun Ko, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Higher-order certification for randomized smoothing, 2020a. URL https://arxiv.org/abs/2010.06651.

Jeet Mohapatra, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Towards verifying robustness of neural networks against semantic perturbations, 2020b. URL https://arxiv.org/abs/1912.09533.

Patrick Müller, Alexander Braun, and Margret Keuper. Classification robustness to common optical aberrations, 2023. URL https://arxiv.org/abs/2308.15499.

Patrick Müller, Alexander Braun, and Margret Keuper. Examining the impact of optical aberrations to image classification and object detection models, 2025. URL https://arxiv.org/abs/2504.18510.

Robert J. Noll. Zernike polynomials and atmospheric turbulence∗. *J. Opt. Soc. Am.*, 66(3):207–211, Mar 1976. doi: 10.1364/JOSA.66.000207. URL https://opg.optica.org/abstract.cfm?URI=josa-66-3-207.

Gabriel Pérez S., Juan C. Pérez, Motasem Alfarra, Silvio Giancola, and Bernard Ghanem. 3DeformRS: Certifying spatial deformations on point clouds, 2022. URL https://arxiv.org/abs/2204.05687.

Buu Phan, Fahim Mannan, and Felix Heide. Adversarial imaging pipelines, 2021. URL https://arxiv.org/abs/2102.03728.

Anian Ruoss, Maximilian Baader, Mislav Balunović, and Martin Vechev. Efficient certification of spatial robustness, 2021. URL https://arxiv.org/abs/2009.09318.

Evgenia Rusak, Lukas Schott, Roland S. Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions, 2020. URL https://arxiv.org/abs/2001.06057.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. URL https://arxiv.org/abs/1409.0575.

Athena Sayles, Ashish Hooda, Mohit Gupta, Rahul Chatterjee, and Earlence Fernandes. Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect, 2021. URL https://arxiv.org/abs/2011.13375.

Jenny Schmalfuss, Lukas Mehl, and Andrés Bruhn. Distracting downpour: Adversarial weather attacks for motion estimation, 2023. URL https://arxiv.org/abs/2305.06716.

Jan Schuchardt and Stephan Günnemann. Invariance-aware randomized smoothing certificates, 2023. URL https://arxiv.org/abs/2211.14207.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pp. 1528–1540, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978392. URL https://doi.org/10.1145/2976749.2978392.

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3(POPL), January 2019. doi: 10.1145/3290354. URL https://doi.org/10.1145/3290354.

Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Trans. Graph.*, 37(4), July 2018. ISSN 0730-0301. doi: 10.1145/3197517.3201333. URL https://doi.org/10.1145/3197517.3201333.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. URL https://arxiv.org/abs/1312.6199.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. URL https://arxiv.org/abs/1905.11946.

Larry N. Thibos, Raymond A. Applegate, James T. Schwiegerling, and Robert Webb. Standards for reporting the optical aberrations of eyes. *Journal of Refractive Surgery*, 18(5):S652–S660, 2002.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2020. URL https://arxiv.org/abs/1906.05849.

Ethan Tseng, Ali Mosleh, Fahim Mannan, Karl St-Arnaud, Avinash Sharma, Yifan Peng, Alexander Braun, Derek Nowrouzezahrai, Jean-François Lalonde, and Felix Heide. Differentiable compound optics and processing pipeline optimization for end-to-end camera design. *ACM Trans. Graph.*, 40(2), June 2021. ISSN 0730-0301. doi: 10.1145/3446791. URL https://doi.org/10.1145/3446791.

Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. Beta-crown: Efficient bound propagation with per-neuron split constraints for complete and incomplete neural network robustness verification, 2021. URL https://arxiv.org/abs/2103.06624.

Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope, 2018. URL https://arxiv.org/abs/1711.00851.

James Wyant and Katherine Creath. Basic wavefront aberration theory for optical metrology. *Appl Optics Optical Eng*, 11, 01 1992.

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples, 2018. URL https://arxiv.org/abs/1801.02612.

Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond, 2020. URL https://arxiv.org/abs/2002.12920.

Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision, 2020. URL https://arxiv.org/abs/1906.08988.

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, 2022. URL https://arxiv.org/abs/2001.02378.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions, 2018. URL https://arxiv.org/abs/1811.00866.

This supplementary material provides detailed information on implementation and dataset details, and additional experimental results. It is structured as follows:

- Why an image space certificate cannot substitute for the coefficient space one in Appendix A
- The per mode certification protocol behind Section 4.2 in Appendix B
- Certifying chromatic aberration in Appendix C
- Coefficient units and the wavefront scale in Appendix D
- The proof of Certificate 1 and what the sampling budget can and cannot express in Appendix E
- The physical, lens sharpness interpretation of the certified radius, with the derivation behind Equation (9), in Appendix F
- Implementation and training details in Appendix G
- Full results across all noise levels, models, and datasets in Appendix H
- Baselines against pixel space and single parameter blur certificates in Appendix I
- Training and threat model controls in Appendix J
- The soundness experiment protocol behind Section 4.3 in Appendix K
- ALC attack magnitude versus the certified radius in Appendix L
- Certified semantic segmentation in Appendix M
- Certified explanations in Appendix N

A WHY AN IMAGE SIDE CERTIFICATE FAILS

An image space certificate would need to compose two renderings, $\psi_{\boldsymbol{\delta}}(\psi_{\boldsymbol{\varepsilon}}(x))$, and treat it as equivalent to rendering the combined state directly, $\psi_{\boldsymbol{\delta}+\boldsymbol{\varepsilon}}(x)$. This equivalence fails, as in general $\psi_{\boldsymbol{\delta}}(\psi_{\boldsymbol{\varepsilon}}(x)) \neq \psi_{\boldsymbol{\delta}+\boldsymbol{\varepsilon}}(x)$. The family $\{\psi_{\boldsymbol{w}}\}$ is not additively closed in $\boldsymbol{w}$, and hence not resolvable in the sense of transformation smoothing (Fischer et al., 2020; Li et al., 2021), because it does not satisfy $\psi_\lambda \circ \psi_\gamma = \psi_{\lambda+\gamma}$. If the renderer convolved with the amplitude point spread function $a_{\boldsymbol{\delta}}$ rather than with its intensity, additivity would hold: By the convolution theorem, convolving $a_{\boldsymbol{\delta}}$ and $a_{\boldsymbol{\varepsilon}}$ corresponds to multiplying their pupil functions, and the phases add exactly, $P_{\boldsymbol{\delta}} \cdot P_{\boldsymbol{\varepsilon}} = \mathrm{Circ} \cdot e^{-j\,2\pi(W_{\boldsymbol{\delta}}+W_{\boldsymbol{\varepsilon}})} \cdot \mathrm{Circ} = P_{\boldsymbol{\delta}+\boldsymbol{\varepsilon}}$, which is precisely the pupil function of the combined state. Thus, amplitude based imaging would therefore be resolvable. Real cameras are illuminated incoherently and record the squared magnitude of the amplitude point spread function rather than the amplitude itself, $h_{\boldsymbol{\delta}} = |a_{\boldsymbol{\delta}}|^2$. Squaring does not commute with convolution, $|a * b|^2 \neq |a|^2 * |b|^2$ in general, so the exact additivity of the amplitude does not survive to the intensity that is actually rendered. A concrete example is a pair of defocus kernels, for which $h_{\boldsymbol{\delta}} * h_{\boldsymbol{\varepsilon}}$ is differently shaped than $h_{\boldsymbol{\delta}+\boldsymbol{\varepsilon}}$, as Figure 3 shows.

None of this touches the certificate of Certificate 1 itself, since its proof needs no resolvability. What it does fix is the semantics of the smoothing noise. The certificate is therefore a tolerance and audit statement about the total lens state relative to the sharp image rather than a runtime guarantee for an already degraded camera image.

B PER MODE CERTIFICATION

Because the coordinates of $\boldsymbol{w}$ carry physical names, the certificate can be restricted to a subspace. Let $S \subseteq \{1, \ldots, d\}$ be a set of modes and let Π_S be the orthogonal projection onto the corresponding coordinates. Setting the smoothing distribution at certification time to $\mathcal{N}(\mathbf{0}, \sigma^2 \Pi_S)$ makes Certificate 1 apply unchanged within the subspace, certifying

$$g_{\boldsymbol{\delta}}(x) = c_A \quad \text{for all } \boldsymbol{\delta} \in \mathrm{range}(\Pi_S) \text{ with } \|\boldsymbol{\delta}\|_2 < r_S. \tag{6}$$

Setting $|S| = 1$ certifies one Zernike mode in isolation and answers the question of which optical defect a given network tolerates least. We certify all eight modes individually using the models

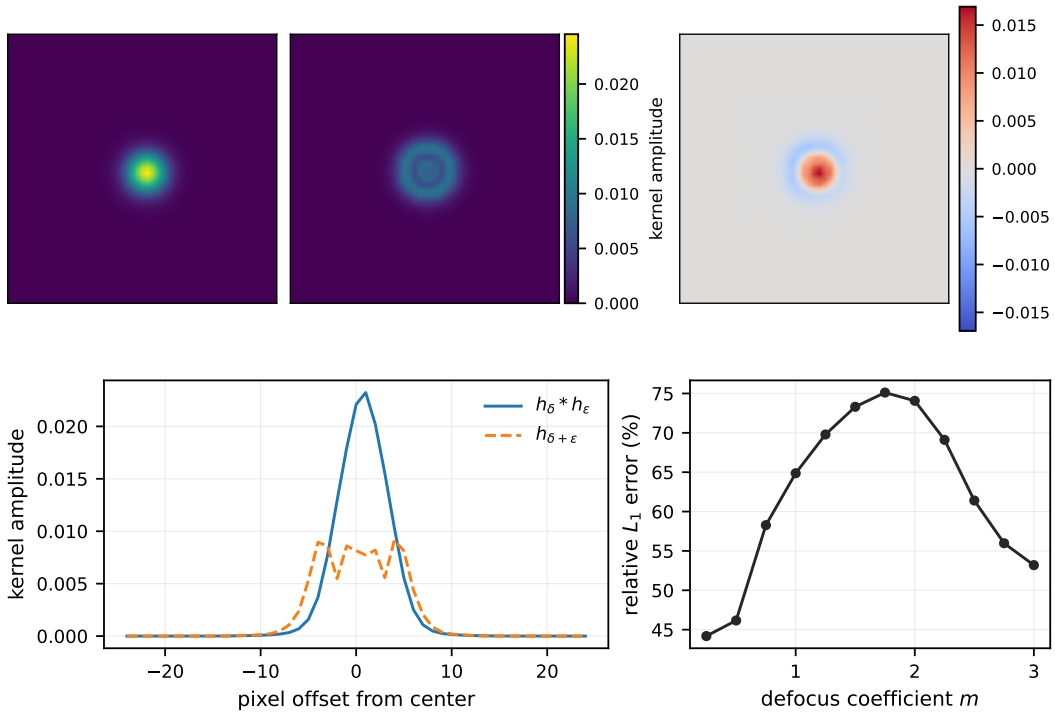


Figure 3: Illustration of the non-resolvability, two pure defocus states $\boldsymbol{\delta} = \boldsymbol{\varepsilon} = 1.0$. *Top*, left to right: the composed kernel $h_{\boldsymbol{\delta}} * h_{\boldsymbol{\varepsilon}}$, the directly rendered $h_{\boldsymbol{\delta}+\boldsymbol{\varepsilon}}$, and their difference. *Bottom*: a central row cross section showing the composed kernel is broader and differently shaped, and the relative L_1 error of $h_m * h_m$ against h_{2m} swept over the defocus magnitude m.

trained on the full eight mode distribution, so the decomposition costs certification time only. This is not possible in pixel space certification, where the coordinates of the perturbation carry no physical meaning.

Orientation pairs certify to the same radius Six of the eight modes form three orientation pairs related by a rotation: astigmatism, coma, and trefoil each contribute two modes that differ only in orientation. Because natural image statistics are close to isotropic, each pair should certify to the same radius. They do, to three or four decimals: 0.9980 against 0.9961 for astigmatism, 0.9878 against 0.9862 for coma, and 0.9993 against 0.9994 for trefoil.

Each pair certifies two directions that should be equivalent, so the disagreement within a pair is a direct estimate of the noise floor of the whole measurement. Over the 24 pair comparisons in Table 10 that disagreement is at most 0.009 in normalized ACR and 0.002 on average, while the gap between the rotationally symmetric and the anisotropic modes is 0.060. The finding is therefore six times the worst case noise floor and twenty five times the typical one, and its direction is identical in all eight configurations, spanning five architectures and four noise levels.

Defocus alone bottlenecks the joint certificate The per mode results certify each mode in isolation, but a per-model radius says nothing about the joint certificate over the full eight-dimensional ball. We test this on the three architectures at $\sigma = 1.0$ on Imagenette, in three configurations. The first drops defocus and spherical aberration and leaves six active modes. The second keeps only those two and leaves two active modes. The third drops each of the eight modes one at a time and leaves seven active modes. Figure 4 reports the resulting change in ACR against the full eight mode ball. Dropping both rotationally symmetric modes raises the ACR by 4.3 to 6.7 percent across the three architectures. Dropping defocus alone already accounts for most of that, 2.5 to 3.6 percent, while dropping spherical aberration alone moves the ACR by at most 0.3 percent, and removing any of the six anisotropic modes individually changes it by at most 0.6 percent in either direction.

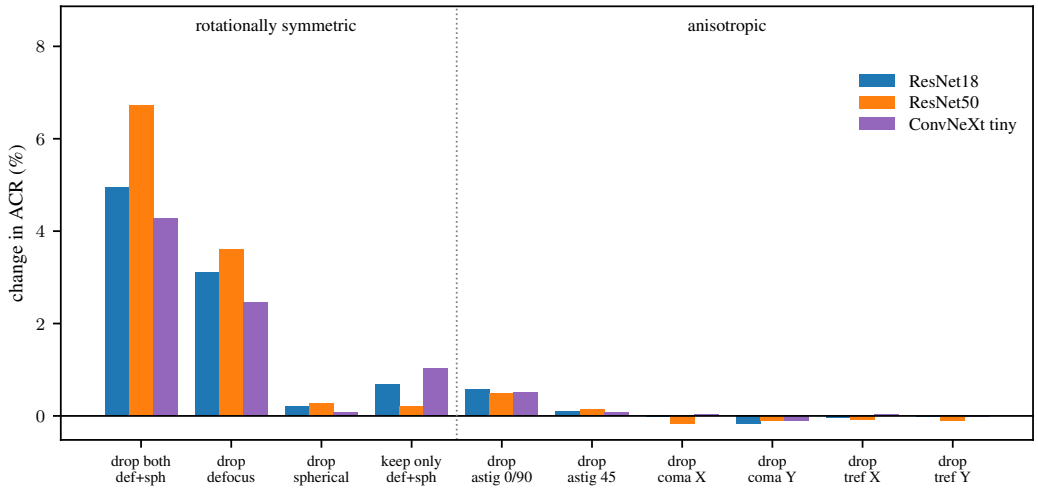


Figure 4: Change in average certified radius when defocus and/or spherical aberration are removed from, or restricted to, the eight dimensional smoothing distribution, and when each anisotropic mode is removed on its own, relative to the full eight mode ball, $\sigma = 1.0$.

C CHANNEL-DEPENDENT WAVE ABERRATION

A real lens can exhibit wavelength-dependent aberrations. The chromatic variant of Section 3 models this by drawing an independent eight mode wavefront per color channel, which triples the coefficient space to $d = 24$. This is a simplified three-channel approximation and does not model spectral integration, lateral chromatic displacement, or wavelength-dependent magnification. Since the Gaussian certificate of Certificate 1 is dimension free, the certified radius should not degrade. Table 6 shows the difference of dimensionality. The practical consequence is that the method extends to richer optical threat models without a robustness penalty, which matters because chromatic aberration is one of the defects real lenses exhibit most reliably.

		ACR		acc@0	
Model	data	$d=8$	$d=24$	$d=8$	$d=24$
ResNet18	Imagenette	2.551	**2.589**	86.9	87.8
ResNet50	Imagenette	2.078	**2.131**	74.8	77.1
EfficientNet B0	Imagenette	2.077	**2.167**	79.5	79.6
SwinV2 tiny	Imagenette	**2.237**	2.236	82.0	81.0
ConvNeXt tiny	Imagenette	2.376	**2.389**	81.7	81.5
ResNet50	ImageNet100	2.001	**2.063**	74.8	75.8
SwinV2 tiny	ImageNet100	**1.996**	1.938	74.4	73.0
ConvNeXt tiny	ImageNet100	**1.899**	1.800	69.1	67.0

Table 6: Chromatic against achromatic certification at matched $\sigma = 1.0$. Tripling the dimension of the perturbation space leaves the certified radius essentially unchanged, as the dimension free bound predicts.

D COEFFICIENT UNITS AND THE WAVEFRONT SCALE

The renderer expands the wave aberration in the unnormalized Fringe polynomials Z_i^{F}, whose mean square over the unit pupil is $1/N_i$, and applies the factor $1/\sqrt{N_i}$ to the coefficient,

$$W_{\boldsymbol{w}} = \sum_i \frac{w_i}{\sqrt{N_i}} Z_i^{\mathrm{F}}, \quad \text{waves}, \tag{7}$$

with Noll weights $N = (3, 6, 6, 8, 8, 5, 8, 8)$ for modes 4 to 11 (Noll, 1976). The same physical aberration takes different numerical values under different conventions, so optical metrology requires the normalization to be named explicitly (Thibos et al., 2002; Mahajan, 1994), which is why we state both units throughout. All radii in this paper are reported in these coefficient units, that is, as the Euclidean norm $\|\boldsymbol{w}\|_2$ of the coefficient vector. This norm is not the root mean square wavefront error, and Equation (9) converts between the two. The wavelength of 532 nm defines the pixel scale of the rendered point spread function.

E THE SAMPLING CEILING

Certificate 1 follows from a direct reduction to the randomized smoothing bound, once the optical renderer is folded into the base classifier. Fix x and define the map $F : \mathbb{R}^d \to Y$ by $F(\boldsymbol{w}) := f(\psi_{\boldsymbol{w}}(x))$. Since f and the rendering operator $\psi_\cdot(x)$ are measurable, F is a measurable classifier on $\mathbb{R}^d$. The smoothed classifiers of Equation (3) are exactly the Gaussian smoothing of F with covariance $\sigma^2 I_d$, evaluated at the base point $\mathbf{0}$ and at the shifted base point $\boldsymbol{\delta}$. Theorem 1 of Cohen et al. (2019) applies to F directly, with no change to its Neyman and Pearson argument, and yields the stated ℓ_2 radius together with its tightness. This step needs nothing about $\psi_\cdot$ beyond measurability, so the renderer may be nonlinear in $\boldsymbol{w}$.

The ceiling below follows from this bound. It follows from one structural property of the Monte Carlo estimate. With n estimation samples and confidence α, the largest attainable lower bound is the Clopper and Pearson bound for a perfect sample, $\underline{p_A}^{\max} = \alpha^{1/n}$, so the certified radius is bounded by

$$r \;\le\; \sigma \cdot \Phi^{-1}\Big(\alpha^{1/n}\Big), \tag{8}$$

independently of how robust the model is. At our setting $n = 10^4$ and $\alpha = 10^{-3}$ this gives $\underline{p_A}^{\max} = 0.999309$ and $r \le 3.199\,\sigma$. At $\sigma = 1.0$ this ceiling is 0.400 waves RMS guaranteed in every direction (Equation (10)). On Imagenette, Table 7 shows the certified radius divided by σ staying close to 3.199 throughout, with ResNet18 attaining 73 to 83 percent of that maximum, so most certified images saturate the Monte Carlo bound rather than the network's robustness. Figure 7 confirms this with the certified accuracy curves for ResNet50 on Imagenette, which flatten out to $r \approx 3$ before falling off a cliff at exactly $3.199\,\sigma$ at every noise level.

Model	0.5	1.0	2.0	3.0	4.0	5.0
ResNet18	2.667	2.551	2.356	2.351	2.338	2.340
ResNet50	2.363	2.078	1.981	1.975	2.056	1.959
EfficientNet B0	2.254	2.077	1.957	1.853	1.925	1.781
SwinV2 tiny	2.403	2.237	2.030	1.945	1.943	1.896
ConvNeXt	2.508	2.376	2.296	2.260	2.268	2.227
ceiling			3.199			

Table 7: Average certified radius divided by σ on Imagenette, for σ from 0.5 to 5.0.

We test Equation (8) by recertifying the same Imagenette images at $n = 10^4$, 10^5, and 10^6, for ResNet18 and ResNet50 at $\sigma \in \{1.0, 2.0\}$. This raises the ceiling from $3.199\,\sigma$ through $3.811\,\sigma$ to $4.347\,\sigma$, and Table 8 shows the ACR rising by 26.1 to 29.6 percent at a hundredfold budget, while acc@0 stays identical at every budget in all four configurations. The extra samples buy radius, not accuracy, exactly as a sampling limited certificate predicts and an accuracy limited one would not.

F INTERPRETATION OF THE RADIUS

Let $\boldsymbol{N} = \mathrm{diag}(N_1, \ldots, N_d)$ hold the Noll weights of the coefficient normalization of Appendix D. Each unit of w_i contributes $1/N_i$ waves, one factor $1/\sqrt{N_i}$ from the normalization of Equation (7) and one from the root mean square of Z_i^{F} over the unit pupil. The Fringe modes are orthogonal, so these contributions add in square,

$$\sigma_W(\boldsymbol{w}) \;=\; \big\|\boldsymbol{N}^{-1}\boldsymbol{w}\big\|_2 \;=\; \sqrt{\textstyle\sum_i (w_i/N_i)^2} \quad \text{waves.} \tag{9}$$

			ACR at $n =$				on ceiling (%)	
Model	σ	acc@0	10^4	10^5	10^6	$\Delta\%$	10^4	10^6
ResNet18	1.0	89.9	2.63	3.05	3.38	+28.5	71.8	53.5
ResNet18	2.0	84.8	4.91	5.66	6.24	+27.1	61.2	53.7
ResNet50	1.0	78.5	2.22	2.59	2.88	+29.6	67.7	53.2
ResNet50	2.0	79.7	4.39	5.03	5.53	+26.1	61.9	52.4

Table 8: Raising the sampling budget on a fixed set of Imagenette images.

This gives the wavefront scale of the certified radius. That ellipsoid $\{\boldsymbol{w} : \|\boldsymbol{w} - \boldsymbol{w}_0\|_2 \le R\}$ maps under $\boldsymbol{N}^{-1}$ to a set with semi axes R/N_i, which contains the ball of radius $R/\max_i N_i$. Thus, a certificate of radius R in coefficient units around $\boldsymbol{w}_0$ covers every aberration whose root mean square wavefront deviation from $\boldsymbol{w}_0$ is at most $R/\max_i N_i = R/8$ waves.

$$\sigma_W^{\mathrm{cert}} \;=\; \frac{R}{\max_i N_i} \;=\; \frac{R}{8} \quad \text{waves}, \tag{10}$$

and covers no aberration whose deviation exceeds $R/\min_i N_i = R/3$ waves.

We verified Equation (9) numerically against the renderer. Setting $\|\boldsymbol{w}\|_2 = 1$ along each single mode and integrating the rendered wavefront over a fine pupil grid reproduces $1/N_i$ to four digits, and over 200 random unit vectors the measured RMS falls in $[0.125, 0.296]$, which is exactly the interval $[1/8, 1/3]$ predicted by the two extreme Noll weights. Because the certificate is a ball in coefficient space, it is an ellipsoid in wavefront space. The guarantee therefore extends furthest along defocus, where $N = 3$, and least far along coma and trefoil, where $N = 8$. We report $R/8$ throughout, since that is the radius that holds in every direction.

The smoothed classifier is guaranteed to predict c_A for every lens whose aberration lies in the spanned subspace and whose RMS wavefront error is at most $R/8$ waves. As a reference anchor, the Maréchal criterion associates an RMS wavefront error of approximately $\lambda/14 \approx 0.071$ waves with the conventional diffraction-limited regime (Born & Wolf, 1999; Wyant & Creath, 1992). Our smallest certified configuration, reaches an average radius of $R = 0.79$ coefficient units, that is 0.1 waves guaranteed in every direction and up to 0.26 waves along defocus, so even there the certified region contains the entire diffraction limited regime.

Figure 5 makes this concrete by showing what the images inside a certified ball look like as the radius grows, together with the point spread function at each radius. It also shows why the guarantee differs by direction. The two rows are pure defocus and a mixed excitation of all eight modes, the two extremes between which the guarantee $R/8$ is the conservative end.

The Maréchal criterion places the radius on an absolute scale but says nothing about what the images inside the certified ball look like. To close that gap we translate the radius into the modulation transfer function, the metric lens manufacturers actually quote. The MTF is the normalized magnitude of the Fourier transform of the point spread function (Goodman, 2017), radially averaged,

$$\mathrm{MTF}_{\boldsymbol{w}}(\rho) \;=\; \frac{1}{|\Omega_\rho|} \sum_{(f_u, f_v) \in \Omega_\rho} \frac{\big|\mathcal{F}\{h_{\boldsymbol{w}}\}(f_u, f_v)\big|}{\big|\mathcal{F}\{h_{\boldsymbol{w}}\}(0,0)\big|}, \tag{11}$$

and MTF50 is the frequency at which it falls to one half. Since the certificate covers the whole sphere $\|\boldsymbol{w}\|_2 = r$, we report the interval spanned by the best and the worst wavefront of that magnitude instead of a single MTF value,

$$\Big[\min_{\|\boldsymbol{w}\|_2 = r} \mathrm{MTF50}(\boldsymbol{w}),\; \max_{\|\boldsymbol{w}\|_2 = r} \mathrm{MTF50}(\boldsymbol{w})\Big], \tag{12}$$

which we estimate by sampling 10^4 random directions per radius. Figure 6 shows such a resulting band. Because it is computed by sampling the sphere and rendering, it is independent of how the coefficient axis is labelled. At $r = 0$ the system is diffraction limited with $\mathrm{MTF50} = 0.287$ cycles per pixel. For a model certified to $r = 1.0$ coefficient units, that is to 0.125 waves RMS in every direction by Equation (10), the covered wavefronts on that sphere span MTF50 values from 0.201 down to 0.069 cycles per pixel, so the certificate tolerates lenses as blurry as 24 percent of diffraction limited sharpness. At $r = 2.0$, or 0.25 waves, the worst case falls to 0.015 cycles per pixel, about 5 percent of nominal, which describes a severely defective lens.



Figure 5: Each column is a certified radius r, labeled with the guaranteed tolerance $r/8$ and the point where it leaves the Maréchal limit. Insets show the point spread function, with the red and green frames enlarging two details. *Top*: pure defocus, at $\mathrm{RMS} = r/3$, the worst case inside the ball. *Bottom*: a mixed excitation of all eight modes, at $\mathrm{RMS} \approx 0.152\, r$.

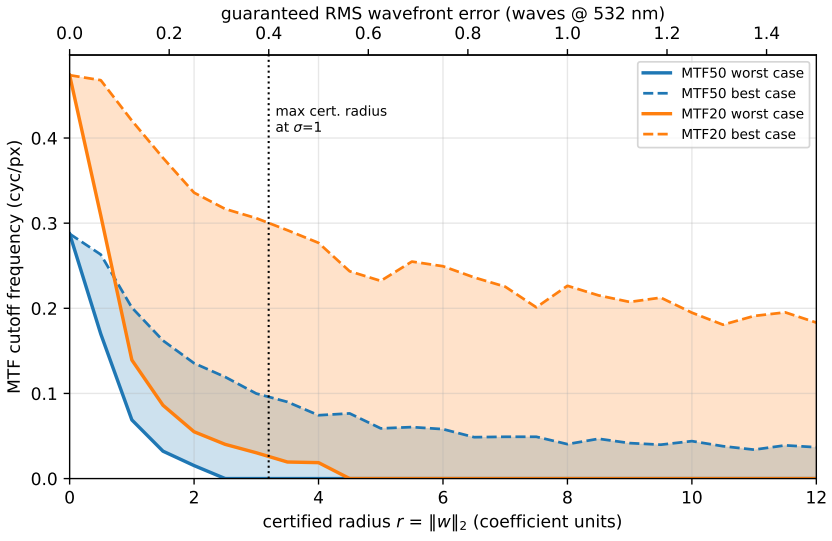


Figure 6: The band of lens sharpness covered by a certified radius, spanning the best and worst wavefront of each coefficient norm (10^4 random directions per radius). The top axis restates the bottom axis in guaranteed waves. The dotted line marks the largest radius expressible at $\sigma = 1$ under our sampling budget.

G Implementation details

Table 9 lists the training configuration. All models are trained from scratch for 90 epochs at seed 0 with the standard recipe of the respective architecture and matched coefficient space noise augmentation. The segmentation heads are trained for 50 epochs on VOC2012 with a ResNet50 backbone.

Classification uses $n_0 = 200$, $n = 10^4$, $\alpha = 10^{-3}$, and every fifth test image. Segmentation uses $n_0 = 10$, $n = 100$, $\tau = 0.75$, and every twentieth image. Certificates are per image and independent.

Model	optimizer	lr	batch	datasets	σ grid
ResNet18	SGD	0.1	128	Imagenette, ImageNet100, ImageNet1k	0.5 . . . 5.0
ResNet50	SGD	0.1	128	Imagenette, ImageNet100, ImageNet1k	0.5 . . . 5.0
EfficientNet B0	RMSprop	0.016	128	Imagenette	0.5 . . . 5.0
SwinV2 tiny	AdamW	5×10^{-4}	64	Imagenette, ImageNet100, ImageNet1k	0.5 . . . 5.0
ConvNeXt tiny	AdamW	5×10^{-4}	128	Imagenette, ImageNet100	0.5 . . . 5.0

Table 9: Training configuration. All models are trained from scratch for 90 epochs with matched coefficient space noise augmentation, using the standard recipe of each architecture rather than a per cell tuned one.

		$m = 0$		$m \neq 0$					
Model	σ	Defoc.	Spher.	Ast. 0/90	Ast. $\pm$45	Coma X	Coma Y	Tref. X	Tref. Y
ResNet18	1.0	0.930	0.958	0.999	0.997	0.990	0.990	1.000	1.000
ResNet50	0.5	0.949	0.961	1.000	0.999	0.993	0.993	0.999	0.999
ResNet50	1.0	0.902	0.937	1.000	0.997	0.988	0.979	0.999	1.000
ResNet50	2.0	0.896	0.939	0.996	0.995	0.984	0.976	1.000	1.000
ResNet50	3.0	0.931	0.961	0.992	0.990	0.990	0.989	1.000	1.000
EfficientNet B0	1.0	0.888	0.939	0.999	0.995	0.978	0.987	0.996	1.000
ConvNeXt tiny	1.0	0.941	0.960	1.000	0.998	0.990	0.988	1.000	0.998
SwinV2 tiny	1.0	0.914	0.947	0.999	0.999	0.989	0.988	1.000	0.999
mean		**0.919**	**0.950**	0.998	0.996	0.988	0.986	0.999	0.999

Table 10: Per mode average certified radius on Imagenette, all eight modes separately, normalized per row to the strongest mode.

H ALL NOISE LEVELS AND MODELS

Table 11 and Table 12 give the complete certification grid of 105 experiment runs. Figure 7 shows the certified accuracy curves, holding ResNet50 on Imagenette fixed and sweeping all ten noise levels of Table 11. Every curve falls to zero at exactly $3.199\,\sigma$, and it does so while still certifying between 43 and 60 percent of the images, so what ends each curve is the Monte Carlo budget of Equation (8) and not the model running out of accuracy.

The main grid trains one checkpoint per noise level. We test whether that is necessary by certifying every ResNet50 checkpoint trained on Imagenette against every noise level in $\{0.5, 1.0, 2.0, 3.0, 5.0\}$, not only the one it was trained for. Figure 8 shows the resulting 5×5 matrix, and the two directions are asymmetric.

I BASELINES

A coefficient norm of 0.05, at most 0.017 waves RMS by Equation (9) and well inside the Maréchal limit, already moves the image by a mean ℓ_2 distance of 14.8. Doubling it to 0.1 moves it only to 14.9, so the displacement comes almost entirely from the diffraction limited point spread function rather than from the aberration. The displacement rises appreciably only for large aberrations, to 16.1 at a coefficient norm of 0.5, 19.5 at 1.0, and 26.9 at 2.0. Cohen et al. certify 49, 37, 19, and 12 percent of ImageNet at radii 0.5, 1.0, 2.0, and 3.0 (Cohen et al., 2019), and that ceiling reflects the noise norm $\sigma\sqrt{d}$ growing with the pixel count d rather than a weaker base classifier. Even the mildest optical aberration we tested therefore moves the image roughly ten times further in pixel space than any published pixel space certificate can cover.

Among certified blur, TSS (Li et al., 2021) comes nearest to our setting with its isotropic Gaussian blur. We reproduce it as a smoothing distribution over a single blur parameter and certify ResNet18

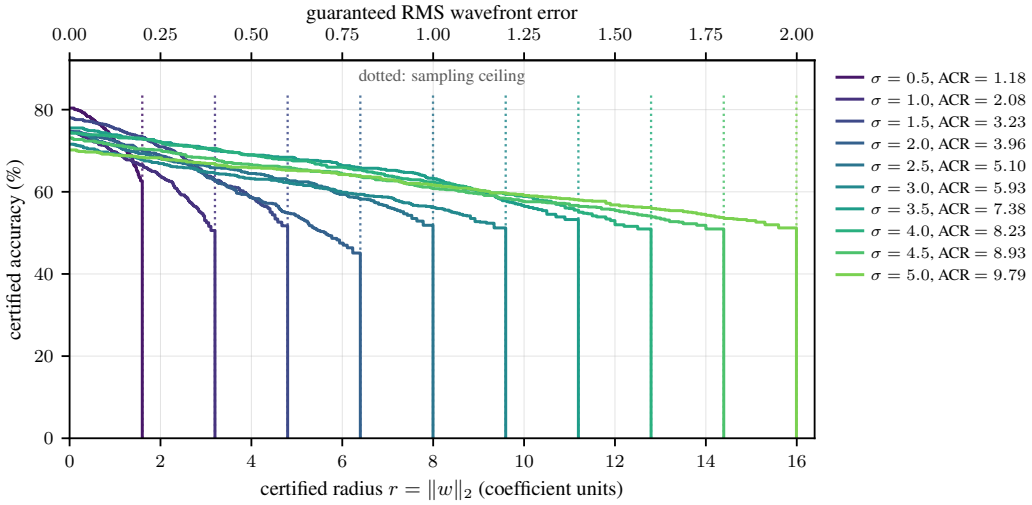


Figure 7: Certified accuracy against certified radius for ResNet50 on Imagenette at all ten noise levels of Table 11. Dotted verticals mark the sampling ceiling $3.199\,\sigma$ of Equation (8), one per noise level.

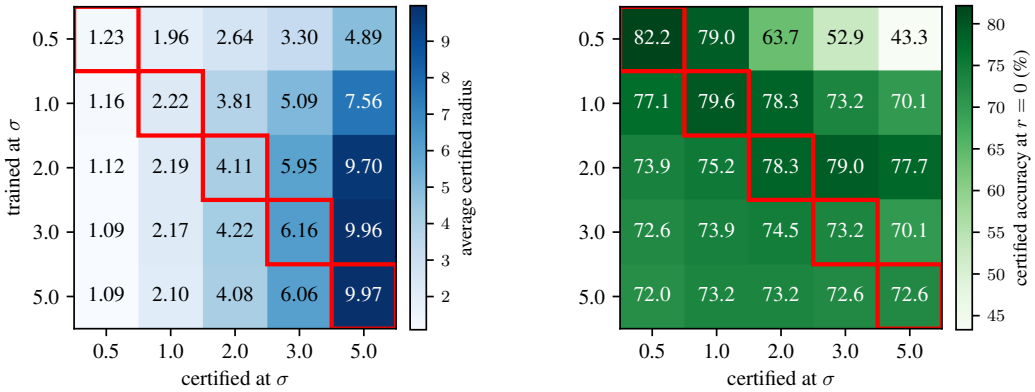


Figure 8: Certifying a ResNet50 / Imagenette checkpoint trained at one noise level (rows) against every noise level in the grid (columns), not only the one it was trained for, on every 25th validation image. Red boxes mark the sigma matched diagonal. It uses the checkpoints of the main grid but only a fifth of the images.

on Imagenette at three noise levels, with everything else held fixed. Table 13 shows that blur certifies slightly higher on both metrics, by 0.04 to 0.44 ACR. A one dimensional isotropic threat model is strictly easier than an eight dimensional anisotropic one, as expected. Our certificate covers a strictly richer perturbation set at a low cost in radius.

J Training and threat model controls

As a control for Section 4.1, Table 14 certifies a nominally trained model over the same noise levels, identical to the main grid except that it never saw coefficient space noise. Certified accuracy collapses by 8.4 to 54.0 points for ResNet18 and 10.1 to 44.5 for ResNet50, and the ratio of nominal to augmented ACR falls from 0.69 to 0.26 and to about 0.33 for ResNet50. Both effects grow with σ, which is common for a model being asked to tolerate noise it was never trained on.

Model		$\sigma = 0.5$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Imagenette											
ResNet18	acc@0	88.0	86.9	86.6	83.8	84.6	84.1	83.3	82.8	81.5	83.2
	ACR	1.334	2.551	3.663	4.711	5.882	7.052	8.152	9.352	10.443	11.702
	abstain	0.8	0.4	0.9	1.4	0.9	1.1	2.5	1.5	1.8	1.7
ResNet50	acc@0	80.4	74.8	78.1	74.6	74.4	71.6	75.5	74.4	73.1	70.2
	ACR	1.181	2.078	3.229	3.962	5.103	5.926	7.383	8.226	8.928	9.793
	abstain	0.6	1.5	1.9	2.2	1.0	2.0	2.5	2.7	2.3	1.9
EfficientNet B0	acc@0	80.8	79.5	76.4	79.2	76.3	74.4	73.5	74.8	76.7	74.1
	ACR	1.127	2.077	2.789	3.914	4.679	5.560	6.305	7.701	8.326	8.904
	abstain	0.9	2.8	3.6	3.7	4.6	4.7	3.7	3.4	4.2	5.4
SwinV2 tiny	acc@0	83.2	82.0	80.8	78.6	78.2	77.7	75.2	76.9	75.3	75.3
	ACR	1.202	2.237	3.116	4.060	4.958	5.836	6.667	7.771	8.387	9.482
	abstain	0.9	2.2	1.9	2.4	2.9	3.9	4.7	4.2	4.3	3.6
ConvNeXt tiny	acc@0	83.9	81.7	79.7	78.9	77.1	77.3	77.5	77.3	76.4	77.1
	ACR	1.254	2.376	3.454	4.593	5.679	6.781	7.881	9.072	10.078	11.136
	abstain	0.6	1.4	0.5	1.0	0.8	1.8	0.9	1.3	1.7	1.4
ImageNet100											
ResNet18	acc@0	79.7	76.6	74.3	72.1	70.8	68.8	69.4	68.5	67.6	68.2
	ACR	1.148	2.070	2.870	3.714	4.520	5.144	6.082	7.084	7.818	8.576
	abstain	1.4	3.0	4.6	5.1	5.7	8.2	7.0	8.7	8.5	8.7
ResNet50	acc@0	79.4	74.8	73.8	71.3	68.9	70.5	66.4	65.7	65.8	66.6
	ACR	1.139	2.001	2.846	3.659	4.373	5.375	5.911	6.586	7.308	8.546
	abstain	1.3	4.6	4.1	6.1	5.7	7.8	7.8	7.8	8.1	7.3
SwinV2 tiny	acc@0	75.8	74.4	72.4	71.8	70.0	71.3	69.5	69.6	70.3	68.5
	ACR	1.089	1.996	2.848	3.655	4.541	5.379	6.151	7.005	8.000	8.784
	abstain	1.8	4.5	5.5	6.6	6.6	7.1	8.0	8.3	7.3	6.2
ConvNeXt tiny	acc@0	70.7	69.1	66.9	67.1	65.5	64.9	64.4	63.2	64.5	61.9
	ACR	1.026	1.899	2.695	3.540	4.378	5.123	5.882	6.729	7.573	8.020
	abstain	1.6	4.0	5.8	5.9	5.3	7.4	7.6	7.4	6.7	7.1

Table 11: Complete certification grid on Imagenette and ImageNet100. acc@0 and the abstain rate are in percent.

Model		0.5	1.0	2.0	3.0	4.0
ResNet18	acc@0	62.8	61.2	56.4	52.3	51.0
	ACR	0.800	1.325	2.387	3.236	4.205
	abst.	4.1	9.0	13.9	18.0	18.0
ResNet50	acc@0	64.5	64.4	58.4	52.1	47.8
	ACR	0.788	1.415	2.522	3.287	4.026
	abst.	4.8	9.2	12.7	16.9	17.6
SwinV2 tiny	acc@0	**76.8**	**75.9**	**73.6**	**72.0**	**71.0**
	ACR	**1.099**	**2.030**	**3.718**	**5.400**	**7.081**
	abst.	1.8	4.2	6.8	8.1	9.2

Table 12: Complete certification grid on ImageNet1k for σ from 0.5 to 4.0.

Table 15 retrains ResNet18 with the consistency objective (Jeong & Shin, 2021), everything else held fixed. It certifies 2.6 to 5.3 percent further at a cost of at most 1.7 points of accuracy, and abstentions fall at every noise level, to zero at $\sigma = 1.0$.

	Gaussian blur, $d = 1$		Zernike, $d = 8$	
σ	acc@0	ACR	acc@0	ACR
0.5	89.3	1.374	88.0	1.334
1.0	88.2	2.673	86.9	2.551
2.0	86.6	5.154	83.8	4.711

Table 13: Single parameter Gaussian blur smoothing against the eight mode Zernike threat model, ResNet18 on Imagenette.

Model		0.5	1.0	2.0	3.0	4.0	5.0
ResNet18	acc@0, (ours)	88.0	86.9	83.8	84.1	82.8	83.2
	acc@0, nominal	79.6	60.1	41.7	35.0	30.7	29.2
	ACR ratio	0.689	0.457	0.333	0.285	0.268	0.260
ResNet50	acc@0, (ours)	80.4	74.8	74.6	71.6	74.4	70.2
	acc@0, nominal	70.3	51.2	38.1	32.9	29.9	27.5
	ACR ratio	0.688	0.536	0.403	0.357	0.326	0.336

Table 14: Nominal training, no coefficient space augmentation, against the main grid on Imagenette, certified over the same noise levels. Columns are σ.

K SOUNDNESS EXPERIMENT

The attack of Section 4.3 is expectation over transformation PGD (Madry et al., 2019; Athalye et al., 2018b) in the eight dimensional coefficient space. Because the certificate is per image, so is the budget, and image i is attacked at $\varepsilon = \kappa r_i$ for five multiples κ, with r_i its certified radius. A cell of Table 3 therefore aggregates 100 attacks at 100 different budgets, and the radius in the row label is the median r_i, not a shared budget. Table 16 gives the distribution of r_i over the 100 attacked images. The majority of images sit exactly on the sampling ceiling of Equation (8), and the share that does so falls from 75 to 53 percent as σ grows, which is the per image counterpart of the aggregate saturation in Table 7. At $\sigma \in \{0.5, 1.0\}$ the smoothed model begins to fail immediately outside the certified region, on 22 and 12 percent of images at $\kappa = 1.5$, so the bound is close to the truth there. At $\sigma = 2.0$ it is not. The smoothed model still flips on zero images at $\kappa = 1.5$ and on only 7 at $\kappa = 2.0$, while the base classifier is already at 44 and 47. The models are therefore robust well beyond the radius we can prove. This follows from the sampling ceiling, where a larger σ multiplies the same capped $\Phi^{-1}(\underline{p_A})$ and leaves proportionally more true robustness unproven.

Non vacuity The base rows of Table 3 count images on which the undefended classifier flips inside the certified radius, up to 29 of 100 at $\sigma = 2.0$. Figure 9 shows two such images per noise level. In each, an aberration inside the certified ball changes the prediction of the base classifier, while the smoothed classifier keeps the correct label.

L ALC ATTACK MAGNITUDE VERSUS THE CERTIFIED RADIUS

This section compares the ALC attack magnitude to the certified radius image by image, the full data behind Section 4.3.

The certified radius r_i used throughout this section is achromatic, over $\mathbb{R}^8$, while the ALC kernel w_i is chromatic, over $\mathbb{R}^{24}$ (Section 3), so comparing them requires a reduction from 24 to 8 dimensions, and the two tables below use different ones. Table 17 uses the mean of the three per channel coefficient norms, $\frac{1}{3}\sum_c \|w_i^{(c)}\|_2$. Table 18 instead evaluates the smoothed classifier at the per channel mean kernel $\bar{w}_i := \frac{1}{3}\sum_c w_i^{(c)}$, applied achromatically, and reports $\|\bar{w}_i\|_2 \le r_i$ as inside.

Attacks that fail against the undefended classifier sit at a median coefficient norm of 1.24 over the full 24 dimensional kernel, attacks that succeed at 3.61, at a 60.0 percent success rate. Flipping an undefended prediction therefore costs a magnitude of the same order as the radii we certify, which is what makes the comparison meaningful. Table 17 reports the per image count at $\sigma = 2.0$ (ACR

	noise augmentation		consistency		
σ	acc@0	ACR	acc@0	ACR	ΔACR%
0.5	88.0	1.334	88.8	1.376	+3.2
1.0	86.9	2.551	85.2	2.617	+2.6
2.0	83.8	4.711	82.3	4.961	+5.3

Table 15: Consistency training against plain noise augmentation, ResNet18 on Imagenette, everything else held fixed.

σ	median r_i	mean r_i	range	on ceiling
0.5	1.5993	1.508	0.320 to 1.599	75 / 100
1.0	3.1986	2.779	0.191 to 3.199	66 / 100
2.0	6.3972	5.200	0.210 to 6.397	53 / 100

Table 16: Distribution of the per image certified radius over the 100 images of the soundness experiment, ResNet50 on Imagenette.

$= 2.52$). The certified radius exceeds that image's own kernel norm on 84.9 percent of certified images, and on 69.4 percent of images where the attack actually succeeded against the undefended model. At $\sigma = 4.0$ this rises to 83.9 percent, at the cost of 1060 fewer certified images. Coverage is thus substantial and tunable, but not complete. Some of the successful ALC kernels still exceed the radius we can certify for their image.

			covers its own attack (%)	
σ	ACR	cert. img.	all cert.	attack succeeded
0.5	0.79	6350	47.0	2.2
1.0	1.41	6240	73.1	45.8
2.0	2.52	5840	84.9	69.4
3.0	3.29	5212	89.0	77.8
4.0	4.03	4780	91.8	83.9

Table 17: Per image coverage of the ALC attack kernel by the certified radius, ResNet50 on ImageNet1k. "Cert. img." are the images carrying a nonzero certificate.

Comparing norms only checks whether the certified radius exceeds the attack's magnitude. It does not test what actually happens to kernels that fall outside the certified ball. We close this gap by evaluating, for every certified image, $g_{\boldsymbol{\delta}_i}(x_i)$, the smoothed classifier's actual decision, at the aberration ALC optimized for that image, rather than only comparing norms. Table 18 reports the result, split by whether the kernel falls inside or outside the certified radius and by whether it fooled the undefended classifier.

Inside the certified radius the flip rate is zero at every noise level and for both outcome classes, exactly what Certificate 1 guarantees and a direct sanity check on the whole pipeline. Outside the certified radius, where the proof makes no claim, the flip rate is far below the 100 percent one might otherwise assume. Among kernels that fooled the undefended classifier it falls from 22.7 percent at $\sigma = 0.5$ to 4.7 at $\sigma = 1.0$, 1.1 at $\sigma = 2.0$, and 0 at $\sigma \geq 3.0$. Kernels that already failed against the undefended classifier flip the smoothed one even less often, at most 8.0 percent, again reaching zero by $\sigma = 2.0$. The coverage gap above is therefore mostly proof conservatism rather than real vulnerability. At $\sigma \geq 3.0$, every stored ALC kernel we tested leaves the smoothed classifier's prediction unchanged, whether inside or outside the certified radius.

M SEGMENTATION

Section 4.4 certifies segmentation with the hierarchical fallback of AdaptiveCertify (Anani et al., 2024) alongside the flat scheme of SegCertify (Fischer et al., 2022). The per level breakdown shows that the label hierarchy does real work but remains a second order correction. At $\sigma = 1.0$,

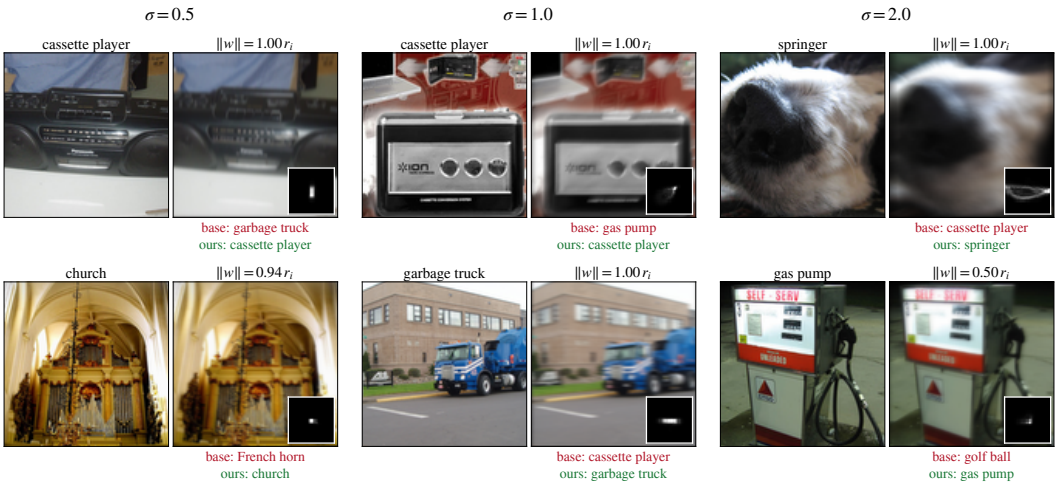


Figure 9: Aberrations inside the certified radius that flip the undefended base classifier, ResNet50 on Imagenette. Each pair shows the clean image with its ground truth label and the aberrated image with its coefficient norm relative to the per image certified radius r_i and its point spread function inset. Below each pair, the prediction of the base classifier (red) and of the smoothed classifier (green) on the aberrated image.

	in certificate		out, attack succeeded		out, attack failed	
σ	n	flip %	n	flip %	n	flip %
0.5	4340	0.00	1598	22.72	412	8.01
1.0	5258	0.00	795	4.65	187	2.67
2.0	5285	0.00	438	1.14	117	0.00
3.0	4848	0.00	262	0.00	102	0.00
4.0	4514	0.00	173	0.00	93	0.00

Table 18: Flip rate of the smoothed classifier g evaluated directly at the optimized ALC attack kernel itself, ResNet50 on ImageNet1k. "In certificate" is $\|w_i\|_2 \leq r_i$, where the certificate guarantees zero flips. "Out" is $\|w_i\|_2 > r_i$, split by whether the kernel flipped the undefended classifier.

DeepLabV3 splits its certified pixels $98.8/0.8/0.4$ percent across the finest, mid, and coarsest label. At $\sigma = 2.0$ this shifts only to $98.5/1.1/0.5$. FCN shows the same pattern with slightly more fallback at $\sigma = 2.0$, at $98.0/1.1/0.9$. The large majority of certified pixels therefore never need to fall back at all, even at the largest noise level tested, yet the fallback is enough to make the hierarchical scheme win on every row of Table 4.

The dense regime of Section 4.4 is two orders of magnitude cheaper than the classification one, $n = 100$ against $n = 10^4$, which caps the segmentation radii well below the classification ones. We test the direction of travel by recertifying DeepLabV3 at $\sigma = 1.0$ with n raised from 100 to 500. A pixel here is certified by a threshold test against τ rather than by an estimated confidence bound, so the radius of a certified pixel is fixed at $\sigma\Phi^{-1}(\tau)$ regardless of n: raising n only changes which pixels pass the test, not how far the ones that do reach. The abstain rate falls from 13.9 to 8.3 percent for the flat scheme and from 12.8 to 7.2 percent for the hierarchical one, certified accuracy rises by 4.3 points in both, and mIoU rises 4.6 points while the certified information gain rises 4.3 points. The one cost is a small drop in accuracy conditional on certification, 1.2 to 1.3 points, since the pixels a larger n newly certifies are the marginal ones the smaller budget correctly declined to certify.

N CERTIFIED EXPLANATIONS

For explanations we certify which pixels remain in the top k percent of a GradCAM attribution map (Selvaraju et al., 2019) under every aberration inside the radius, extending the top k certificate of Levine et al. (2019), which bounds how many pixels remain, to which pixels remain. The reduction that makes it work, treating membership in the top k percent mask as a per pixel hypothesis and reusing the segmentation machinery of Section 3, is the one that Anani et al. (2025) make explicit for attributions in pixel space, and we apply it with the aberration coefficients as the perturbation set instead. Figure 10 shows the qualitative counterpart of Table 5. The certified regions concentrate on the same object that the uncertified saliency map highlights, but shrink to a small high confidence core, which is what the numbers in Table 5 look like on real images.

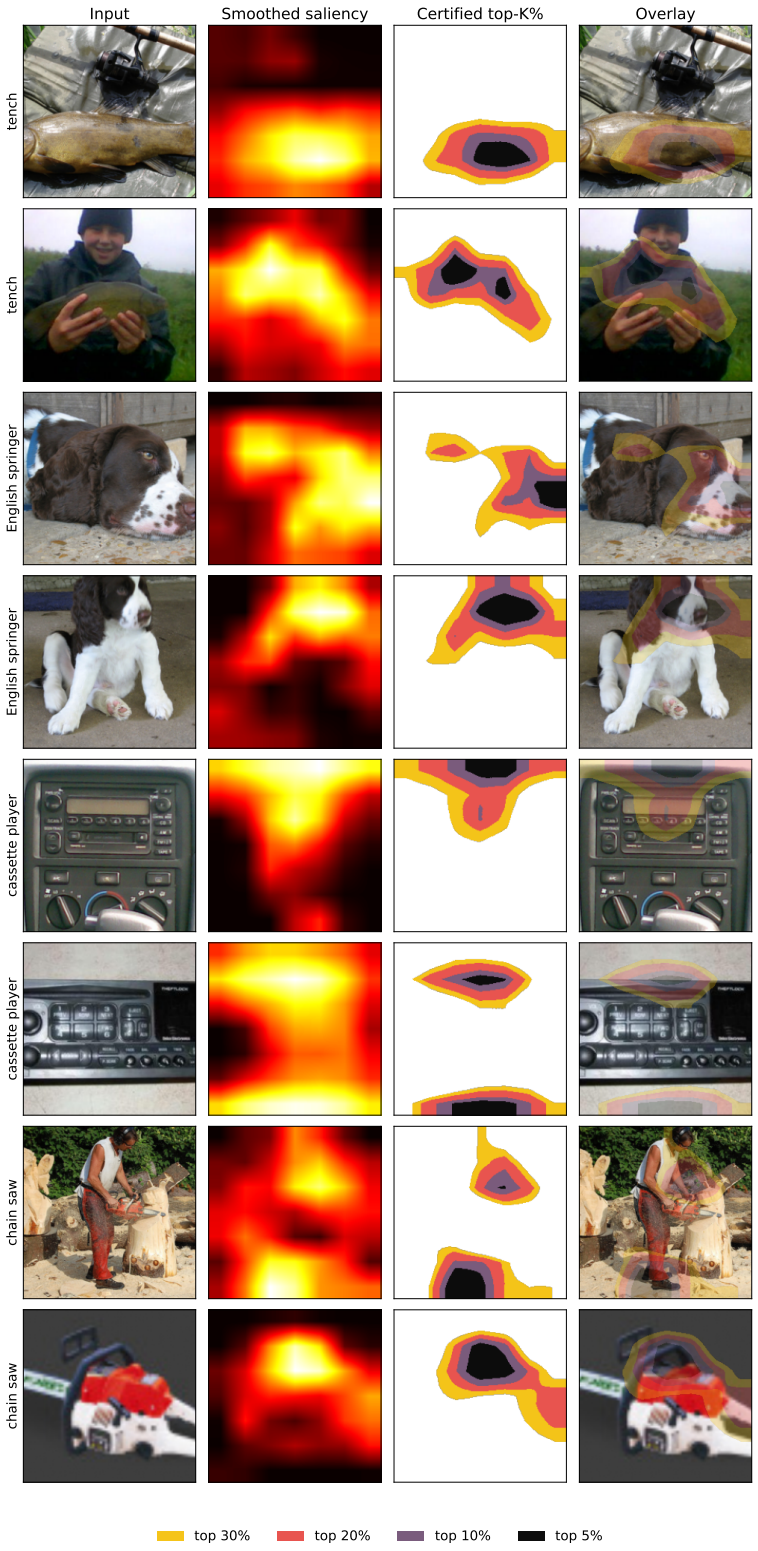


Figure 10: Certified explanations for ResNet50 on Imagenette at $\sigma = 1.0$, $\tau = 0.75$ and a certified radius $r = 0.67$. Columns: input, smoothed saliency, certified top k percent regions for $k \in \{30, 20, 10, 5\}$ nested, and overlay.