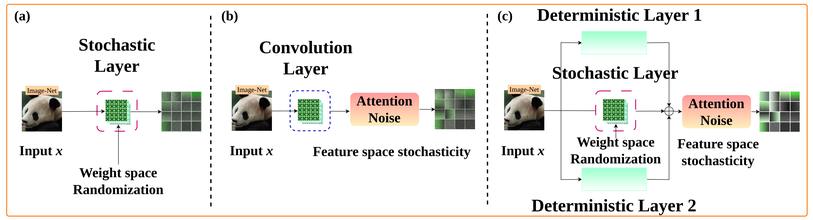
Certified vs. Empirical Adversarial Robustness via Hybrid Convolutions with Attention Stochasticity

Joy Dhar, IIT Ropar, **Song Xia*,** NTU Singapore, **Manish Kumar Pandey*,** RoentGen Health, **Maryam Haghighat,** QUT Australia, **Azadeh Alavi,** RMIT Australia, **Ferdous Sohel,** Murdoch University, **Wenyu Zhang,** Indepedent Researcher, **Nayyar Zaidi** Deakin University

ICLR International Conference On Learning Representations

Concept



High-level comparison of adversarial defense architectures. Prior work focuses on either (a) weight-space stochasticity in convolutional filters (e.g., RPF [1], CTRW [2]) or (b) feature-space stochasticity via post-convolution attention noise (e.g., GATN [3]), whereas (c) our hybrid defense combines both stochastic approaches with deterministic principles to achieve leading certified and empirical robustness.

	Empirical Defenses (AT, PNI)	Certified Defenses (RS, LOT)	HyCAS (The Synthesis)
Worst-case Guarantees	✗ No	✓ Yes, rigorous math	✓ Yes, ≤ 2-Lipschitz bounds
Resistance to ℓ_∞ Attacks	✓ High	✗ Low/Brittle	✓ High, dual stochasticity
Clean Accuracy Preservation	— Moderate	✗ Sacrificed for large radii	✓ State-of-the-art preservation
High-Res Medical Efficacy	— Varies	✗ Rarely benchmarked	✓ Dominant, generalized

Table. 1: Research gap analysis. Scope of representative certified, empirical, and hybrid defences

Contributions

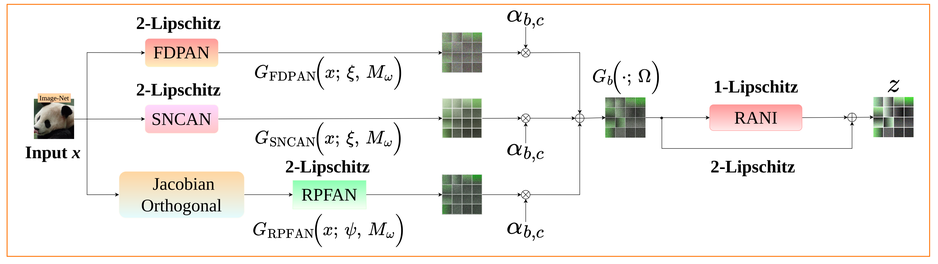
- **Hybrid defense:** HyCAS is a randomized Lipschitz-constrained defense that jointly delivers certified $\ell 2$ robustness and strong empirical $\ell\infty$ robustness.
- **Theoretical guarantees.** HyCAS offers a tight $\ell 2$ certificate guarantees while ensuring empirical robustness against $\ell\infty$ attacks.
- **Plug-and-play design.** HyCAS jointly exploits a 1-Lipschitz aware deterministic spectrally normalized core with two stochastic modules: random-projection filters and randomized attention noise, thereby incorporating refined stochasticity into the network, which yielding a certifiable ≤2-Lipschitz network.
- **Comprehensive evaluation.** Across diverse natural and medical imaging benchmarks, HyCAS outperforms prior defenses and offers a tunable trade-off between certification strength and empirical robustness.

Implementation

Our code and paper are publicly available !

Code **Paper**

Method

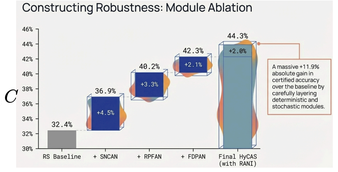


Overview of HyCAS mechanism. It consists of three parallel streams—FDPAN, SNCAN, and RPFAN—each built from 1-Lipschitz cores with RANI module. Per-channel convex gating fuses the streams, where each stream,and their pipekinefusion is 2-Lipschitz; .

Hybrid Convolutions with Attention Stochasticity

We define the smoothed classifier induced by HyCAS with having deterministic and stochastic parameters

$$\Omega = (\xi, \psi, M_\omega) \rightarrow \quad \text{s.t.} \quad g_\theta(x) = \arg\max_{c\in\mathcal{Y}} P_{\varepsilon,\Omega}[f_\theta(x+\varepsilon;\Omega)=c]$$



Then we formulate the HyCAS block output is:

$$z(x)_{:,:,c} = \sum_{b\in\mathcal{B}} \alpha_{b,c}\,[G_b(x;\Omega)]_{:,:,c} + R\Big(\sum_{b\in\mathcal{B}} \alpha_{b,c}\,[G_b(x;\Omega)]_{:,:,c} M_\omega\Big) \quad c = 1,\cdots,C$$

By Theorem 1, we formulate

$$G_b(x;\xi,\psi,\omega) = \mathcal{D}_\xi(\mathcal{T}_\psi(x)) + M_\omega(\mathcal{D}_\xi(\mathcal{T}_\psi(x))),$$ which is provably 2-Lipschitz on each forward pass.

The HyCAS--integrated network is optimised as:

$$\mathcal{L}_{HyCAS} = \zeta \odot \mathcal{L}_{FDPAN} + \varphi \odot \mathcal{L}_{SNCAN} + \nu \odot \mathcal{L}_{RPFAN} + \kappa \odot \mathcal{L}_{RANI}$$

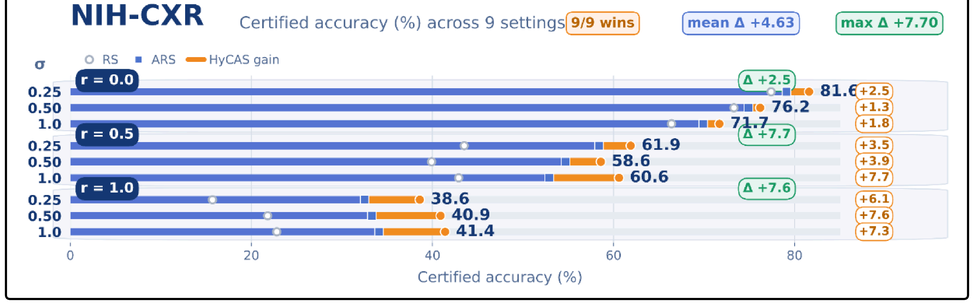
where:

FDPAN, **SNCAN**, **RPFAN**, and **RANI** minimises the objective of **HyCAS**, thereby incorporating stochasticity into the network:

$$\mathcal{L}_{\text{FDPAN}}(\theta) = \min_\theta \mathbb{E}_{(x,y)}\, \mathbb{E}_{\substack{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)\\ \xi,\, M_{\omega_i}}} \ell\Big(f_\theta(x+\varepsilon;\,\xi,\,M_{\omega_i}),\,y\Big); \mathcal{L}_{\text{SNCAN}}(\theta) = \min_\theta \mathbb{E}_{(x,y)}\, \mathbb{E}_{\substack{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)\\ \xi,\, M_{\omega}}} \ell\Big(f_\theta(x+\varepsilon;\,\xi,\,M_{\omega}),\,y\Big);$$

$$\mathcal{L}_{\text{RPFAN}}(\theta) = \min_\theta \mathbb{E}_{(x,y)}\, \mathbb{E}_{\substack{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)\\ \Omega}} \ell\Big(f_\theta(x+\varepsilon;\,\Omega),\,y\Big); \quad \mathcal{L}_{\text{RANI}}(\theta) = \min_\theta \mathbb{E}_{(x,y)}\, \mathbb{E}_{\substack{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)\\ M_\omega}} \ell\Big(f_\theta(x+\varepsilon;\,M_\omega),\,y\Big).$$

Results I

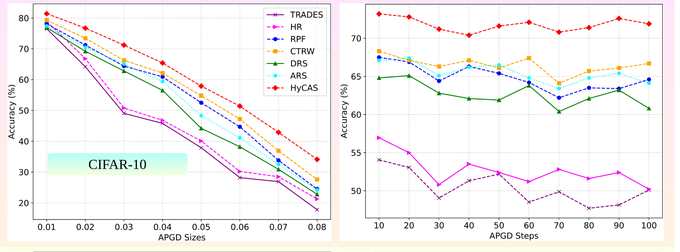


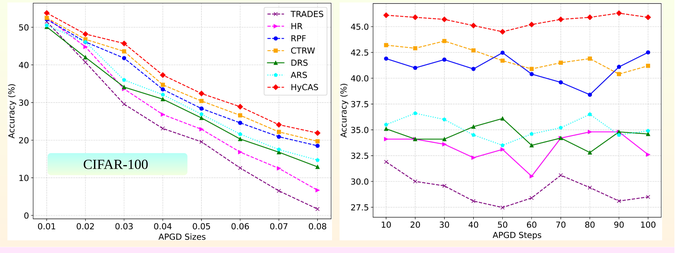
Results II

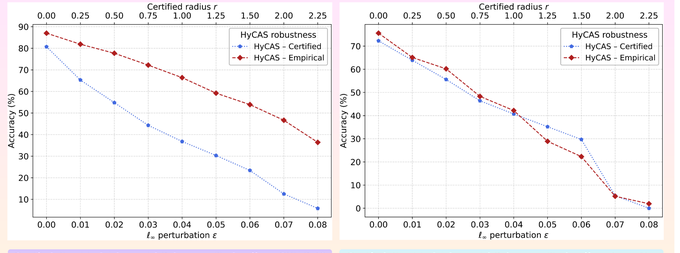
Certified Accuracy (%) of HyCAS and Prior Baselines on CIFAR-10 and ImageNet

Bold denotes the best in each column across all noise–radius pairs. All methods are evaluated at two noise levels.

		CIFAR-10								ImageNet							
Approach	σ	0.00	0.25	0.50	0.75	1.00	1.25	1.50	2.0	0.00	0.25	0.50	0.75	1.00	1.25	1.50	2.0
RS	0.25	75.3	60.2	43.4	26.1	0	0	0	0	67.1	48.7	0	0	0	0	0	0
	0.50	65.2	54.1	41.3	32.4	23.2	14.7	9.34	0	57.3	45.9	36.8	28.7	0	0	0	0
IRS	0.25	78.6	63.2	47.5	30.8	19.6	10.3	5.72	0	68.4	58.5	46.2	38.7	32.1	19.3	10.8	0
	0.50	71.3	58.5	44.1	33.3	24.1	15.7	11.4	2.2	62.4	50.9	41.5	34.7	27.3	20.2	13.8	6.31
DRS	0.25	83.4	65.8	50.2	34.5	24.7	15.8	10.5	0	70.6	61.2	51.8	42.7	38.4	32.6	25.4	0
	0.50	78.1	62.1	48.7	35.8	24.5	17.9	12.9	4.6	67.6	58.2	49.6	42.8	35.6	33.2	29.8	21.3
ARS	0.25	84.1	67.3	51.4	39.1	30.9	21.1	16.2	0	71.1	61.4	52.7	43.1	39.1	33.4	26.7	0
	0.50	78.4	63.7	50.2	38.9	31.8	23.3	19.7	8.47	68.1	58.7	50.3	43.4	39.1	34.5	30.6	22.4
LOT	—	80.5	64.7	48.6	34.3	23.6	15.2	9.14	0	69.7	60.6	50.9	42.2	37.1	30.5	21.8	0
		76.7	60.4	46.3	35.1	24.9	17.3	12.1	6.25	66.1	57.4	48.9	42.8	38.4	32.9	28.3	19.8
SLL	—	81.4	65.3	49.9	33.1	23.6	14.7	9.94	0	70.2	57.7	48.4	41.8	37.6	31.9	24.3	0
		77.9	62.6	48.7	34.5	24.4	16.2	13.7	5.83	67.3	55.5	49.1	42.8	39.1	34.5	26.7	21.3
HyCAS	0.25	85.4	70.1	56.7	44.3	36.5	29.6	22.9	8.52	72.3	63.9	55.6	46.4	40.7	35.2	29.7	5.42
	0.50	80.7	65.3	54.8	44.3	36.8	30.3	23.4	12.5	69.2	60.6	53.9	45.6	41.1	36.3	32.7	24.8







Certified–Empirical Adversarial Robustness Trade-offs on CIFAR-10 | Certified–Empirical Adversarial Robustness Trade-offs on ImageNet