ICLR 2025

Adversarially Robust Out-of-Distribution Detection Using Lyapunov-Stabilized Embeddings

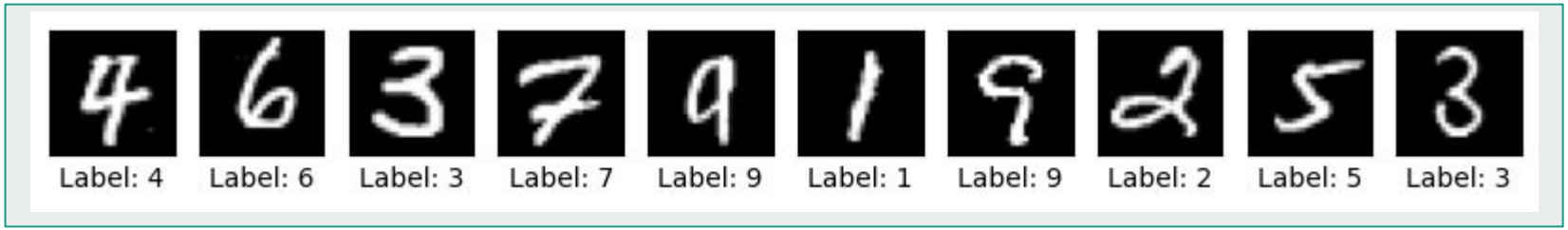
Hossein Mirzaei, Mackenzie W. Mathis

EPFL

OOD Detection

For the OOD detection setup, two different datasets are considered: the ID set, which is available during training, and the OOD set, which needs to be detected during testing.

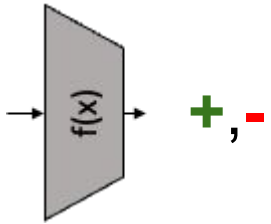
In-distribution (ID) Set: MNIST Dataset



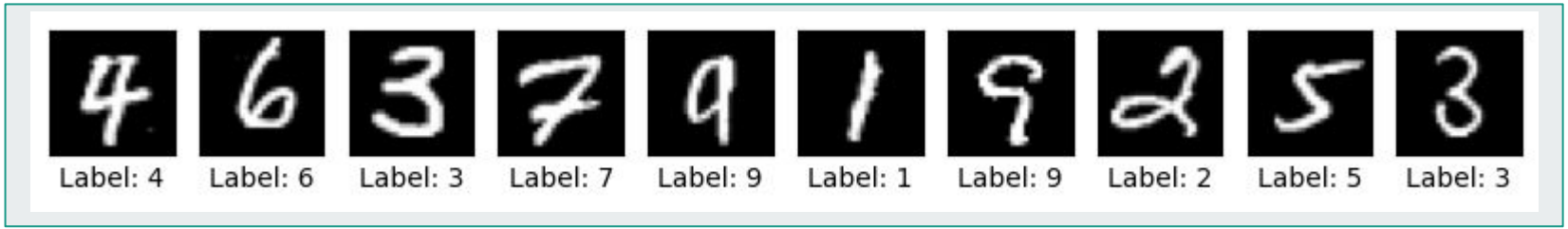
Out-of-Distribution (OOD) Set: FashionMNIST Dataset

Robust OOD Detection

f(x) is an OOD detector, trained on the ID set.



In-distribution (ID) Set: MNIST Dataset



Out-of-Distribution (OOD) Set: FMNIST Dataset

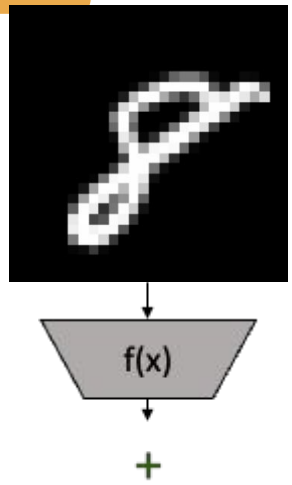
Mathis Laboratory of adaptive intelligence

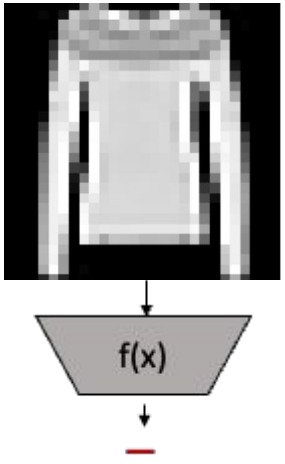
EPFL

Robust OOD Detection

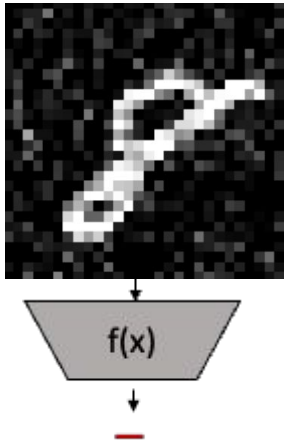
MNIST Vs. FMNIST

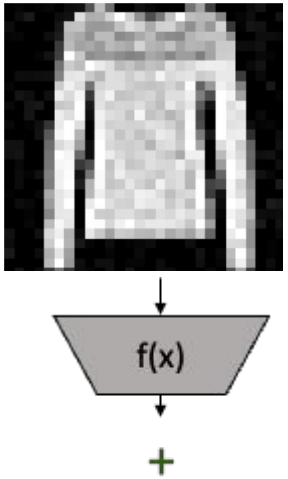
Clen Testing (99.0% AUROC):





Adversarial Testing (0.0% AUROC):





Mathis Laboratory of adaptive intelligence

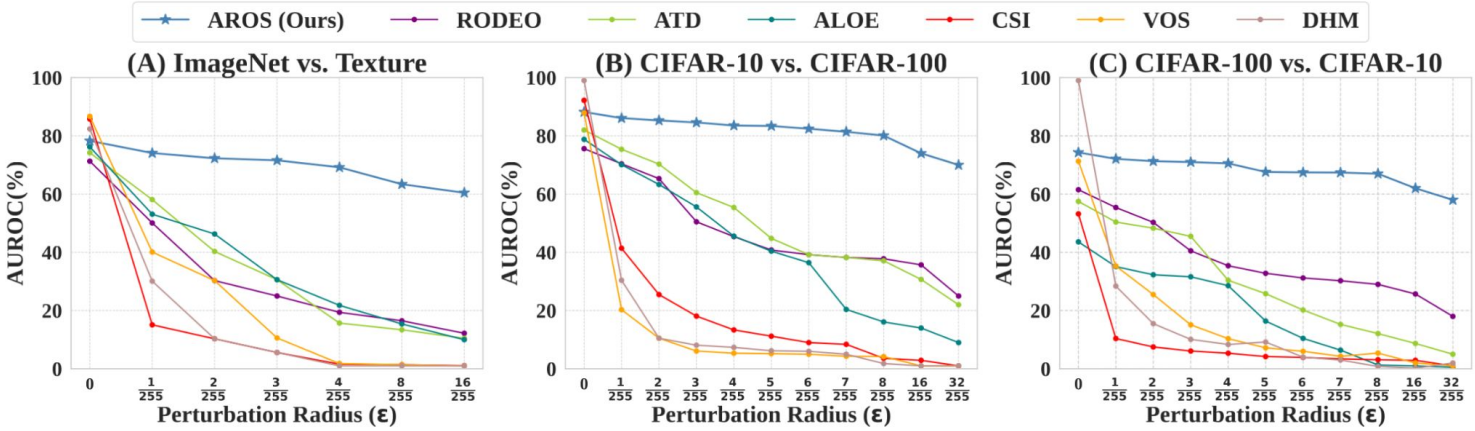
EPFL

Robust OOD Detection

Motivation:

Motivated by the drop in performance of OOD detection methods under adversarial attacks, we aimed to propose an adversarially robust OOD detection method.

First, we propose utilizing stability theorems to ensure a bounded response of the detector. Then, we aim to maximize the distance between OOD and ID samples.

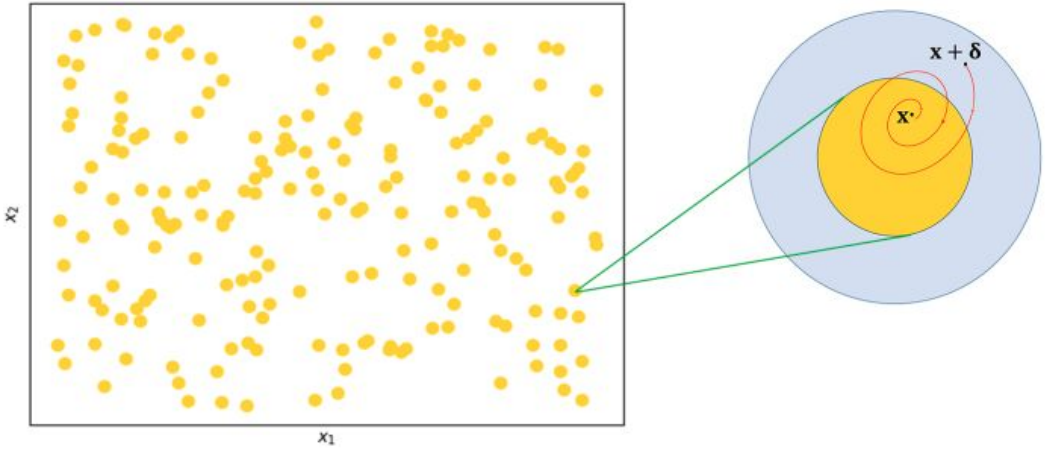


Mathis Laboratory of adaptive intelligence

EPFL

Stability Theorem

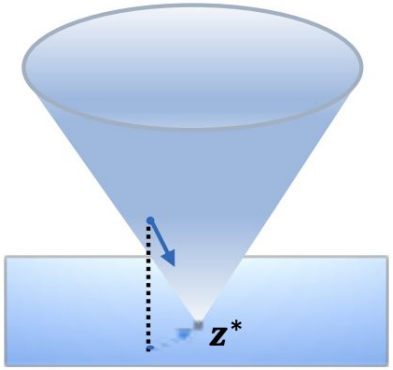
Robustness Through Stable Equilibrium Points

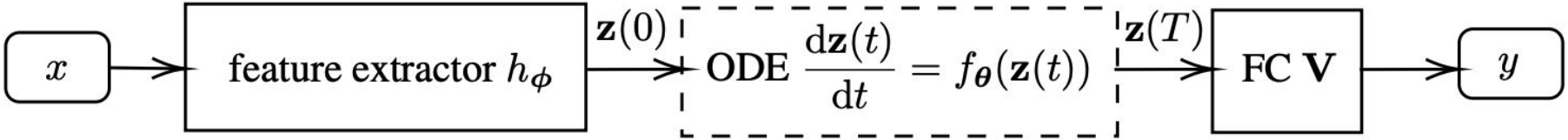


Recently, some methods have leveraged stable neural ordinary differential equations as a defense mechanism. The intuition is to add an ordinary differential equation (ODE) layer that operates over time and to find conditions under which the model remains stable. Here, stability refers to the model's response to perturbed inputs being bounded, which helps prevent successful attacks.

Stability Theorem

Robustness Through Stable Equilibrium Points





Stability Theorem

For a given dynamic system $\frac{dz(t)}{dt} = h_\phi(z(t))$, a state $z^\star$ is an equilibrium point of system if $z^\star$ satisfies $h(z^\star) = 0$. An equilibrium point is stable if the trajectories starting near $z^\star$ remain around it all the time. More formally:

Definition 1: (Lyapunov stability (81)). An equilibrium $z^\star$ is said to be stable in the sense of Lyapunov if, for every $\varepsilon > 0$, there exists $\delta > 0$ such that, if $\|z(0) - z^\star\| < \delta$, then $\|z(t) - z^\star\| < \varepsilon$ for all $t \geq 0$. If $z^\star$ is stable, and $\lim_{t\to\infty} \|z(t) - z^\star\| = 0$, $z^\star$ is said to be asymptotically stable.

Theorem 1: (Hartman–Grobman Theorem (82)). *Consider a time-invariant system with continuous first derivatives, represented by* $\frac{d\mathbf{z}(t)}{dt} = h(\mathbf{z}(t))$. *For a fixed point* $\mathbf{z}^*$, *if the Jacobian matrix* ∇h *evaluated at* $\mathbf{z}^*$ *has no eigenvalues with a real part equal to zero, the behavior of the original nonlinear dynamical system can be analyzed by studying the linearization of the system around this fixed point. The linearized system is given by* $\frac{d\mathbf{z}'(t)}{dt} = \mathbf{A}\mathbf{z}'(t)$, *where* $\mathbf{A}$ *is the Jacobian matrix evaluated at* $\mathbf{z}^*$. *This allows for a simplified analysis of the local dynamics in the vicinity of* $\mathbf{z}^*$.

Theorem 2: (Lyapunov Stability Theorem (81)) *The equation* $\frac{d\mathbf{z}'(t)}{dt} = \mathbf{A}\mathbf{z}'(t)$, *is asymptotically stable if and only if all eigenvalues of* $\mathbf{A}$ *have negative real parts.*

Theorem 3: (Levy–Desplanques Theorem (83)) *Let* $A = [a_{ij}]$ *be an* n*-dimensional square matrix and suppose it is strictly diagonally dominant, i.e.,* $|a_{ii}| \geq \sum_{i\neq j} |a_{ij}|$ *and* $a_{ii} \leq 0$ *for all* i. *Then every eigenvalue of* A *has a negative real part.*

Robust OOD Detection

Difference from the robust classification task: During testing, the model encounters both perturbed OOD samples and ID samples and must effectively distinguish between them. In contrast, in the classification task, OOD data is absent during testing.

AROS (*Proposed Method*)

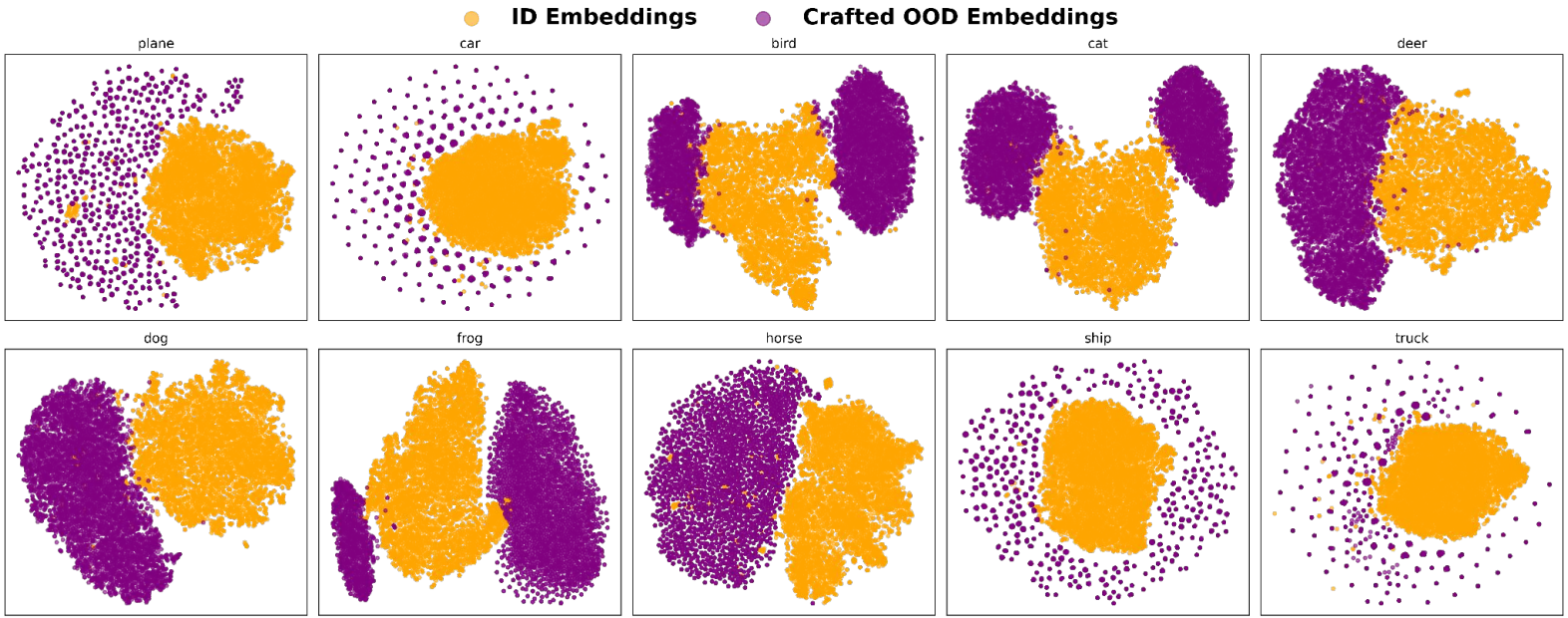
- In this study, we propose AROS: **A**dversarially **R**obust **O**OD detection through **S**tability.

- As mentioned earlier, our detector should be resistant to both perturbed OOD and ID samples. However, OOD samples are absent during training. To address this, we first propose sampling synthetic data from the embedding space to serve as a proxy for real OOD samples.

- To achieve this, we first train a classifier on the ID classes, then extract the representations of the ID samples. Next, we fit a Gaussian distribution to the ID embeddings. We then sample from the low-likelihood regions of the ID embedding space to generate synthetic OOD data—data that has a very low likelihood of belonging to the ID class.

AROS (*Proposed Method*)

t-SNE visualization of CIFAR-10 embeddings and the corresponding crafted OOD embeddings for each class



AROS (*Proposed Method*)

Then, by ensuring that the detector's equilibrium points are Lyapunov-stable, we ensure that perturbed input samples remain within the same stable neighborhood as their unperturbed counterparts.

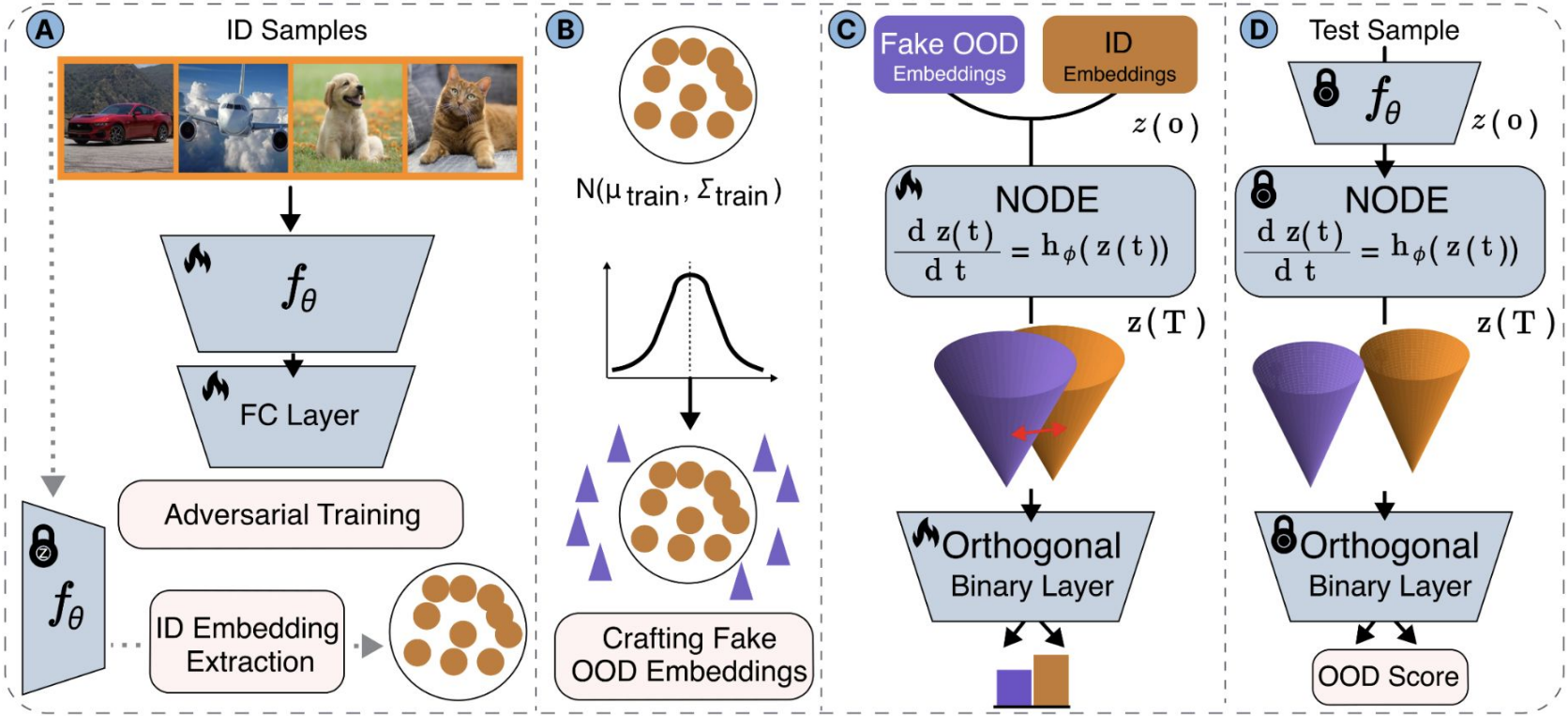
As a result, even when adversarial perturbations are introduced, the system dynamics guide the state back to the stable equilibrium, effectively neutralizing the impact of the perturbation.

$$\mathcal{L}_{\mathrm{SL}} = \min_{\phi,\eta} \frac{1}{|X_{\mathrm{train}}|} \Bigg(\ell_{\mathrm{CE}}(B_\eta(h_\phi(X_{\mathrm{train}})), y) + \gamma_1 \|h_\phi(X_{\mathrm{train}})\|_2 + \gamma_2 \exp\left(-\sum_{i=1}^{n} [\nabla h_\phi(X_{\mathrm{train}})]_{ii} \right) + \gamma_3 \exp\left(\sum_{i=1}^{n} \left(-|[\nabla h_\phi(X_{\mathrm{train}})]_{ii}| + \sum_{j \neq i} |[\nabla h_\phi(X_{\mathrm{train}})]_{ij}| \right) \right) \Bigg)$$

Mathis Laboratory of adaptive intelligence | EPFL

AROS (*Proposed Method*)

An Overview of Method



Comparison of Methods

Dataset		Method								
$\mathcal{D}_{in}$	$\mathcal{D}_{out}$	VOS	DHM	CATEX	CSI	ATOM	ALOE	ATD	RODEO	AROS (Ours)
CIFAR-10	CIFAR-100	87.9/4.2	**100.0**/1.8	88.3/0.8	92.2/3.6	94.2/1.6	78.8/16.1	82.0/37.1	75.6/37.8	88.2/**80.1**
	SVHN	93.3/2.8	**100.0**/ 4.5	91.6/2.3	97.4/1.7	89.2/4.7	83.5/26.6	87.9/39.0	83.0/38.2	93.0/**86.4**
	Places	89.7/5.2	**99.6**/0.0	90.4/ 4.7	93.6/0.1	98.7/5.6	85.1/21.9	92.5/59.8	96.2/70.2	90.8/**83.5**
	LSUN	98.0/7.3	**100.0**/2.6	95.1/0.8	97.7/0.0	99.1/1.0	98.7/50.7	96.0/68.1	99.0/ **85.1**	90.6/ 82.4
	iSUN	94.6/0.5	99.1/2.8	93.2/ 4.4	95.4/3.6	**99.5**/2.5	98.3/49.5	94.8/65.9	97.7/78.7	88.9/**81.2**
CIFAR-100	CIFAR-10	71.3/5.4	**100.0**/2.6	85.1/ 4.0	53.2/0.7	87.5/2.0	43.6/1.3	57.5/12.1	61.5/29.0	74.3/**67.0**
	SVHN	92.6/3.2	**100.0**/0.8	94.6/5.7	90.5/4.2	92.8/5.3	74.0/18.1	72.5/27.6	76.9/31.4	81.5/**70.6**
	Places	75.5/0.0	**100.0**/3.9	87.3/1.4	73.6/0.0	94.8/3.0	75.0/12.4	83.3/40.0	93.0/66.6	77.0/**69.2**
	LSUN	92.9/5.7	**100.0**/1.6	94.0/8.9	63.4/1.8	96.6/1.5	98.7/50.7	96.0/68.1	98.1/63.1	74.3/**68.1**
	iSUN	70.2/4.5	99.6/3.6	81.2/0.0	81.4/3.0	96.4/1.4	**98.3**/49.5	94.8/65.9	95.1/65.6	72.8/**67.9**
ImageNet-1k	Texture	86.7/0.8	82.4/0.0	92.7/0.0	85.8/0.6	**88.9**/7.3	76.2/21.8	74.2/15.7	71.3/19.4	78.3/**69.2**
	iNaturalist	94.5/0.0	80.7/0.0	97.9/2.0	85.2/1.7	83.6/10.5	78.9/19.4	72.5/12.6	72.7/15.0	84.6/**75.3**
	Places	90.2/0.0	76.2/0.4	**90.5**/0.0	83.9/0.2	84.5/12.8	78.6/15.3	75.4/17.5	69.2/18.5	76.2/**68.1**
	LSUN	91.9/0.0	82.5/0.0	**92.9**/0.4	78.4/1.9	85.3/11.2	77.4/ 16.9	68.3/15.1	70.4/16.2	79.4/**69.0**
	iSUN	92.8/2.7	81.6/0.0	**93.7**/0.0	77.5/0.0	80.3/14.1	75.3/11.8	76.6/15.8	72.8/17.3	80.3/**71.6**
	Mean	88.1/2.8	**93.4**/1.6	91.2/2.3	83.3/1.5	91.4/5.6	81.4/25.5	81.6/37.4	82.1/44.4	82.0/**74.0**

Conclusion & Future Research

- We introduced a novel adversarially robust OOD detection approach using Lyapunov-stabilized embeddings, leveraging dynamical system stability to enhance robustness.
- Our method significantly improves the model's resilience against adversarial perturbations, ensuring stable convergence of both ID and OOD samples to equilibrium points.
- Extensive experimental results demonstrate that our approach outperforms current OOD detection methods across various challenging scenarios, achieving superior clean and robust performance.
- These promising results highlight the broader applicability of Lyapunov stability principles, suggesting future extensions to other adversarially-sensitive machine learning applications.