REJECT CLASSES AND ANCHORED ENERGY: BYPASSING THE CLEAN-ROBUST TRADE-OFF IN ADVERSARIAL OOD DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks are often overconfident on out-of-distribution (OOD) inputs, a vulnerability exacerbated by adversarial perturbations designed to evade detection. We reveal that *In-Distribution (ID) adversarial training* is indispensable for genuine OOD robustness—without it, even the strongest OOD-centric defenses collapse under principled attacks. However, it induces a severe clean-robust trade-off: ID adversarial training suppresses the maximum softmax probability (MSP) of clean samples, crippling the separation between ID and OOD data. We establish two structural routes to mitigate this trade-off: (1) replacing MSP with the LogSumExp (LSE) energy score and supplying its missing anchor via energy-margin regularization (RCE+EMR); and (2) GRoOD, which unlocks an auxiliary reject class whose probability p_{K+1} stays insulated from ID confidence degradation. We then uncover a fundamental bottleneck when attempting to transfer EMR across the two routes: the **variance-free vulnerability**—EMR fails on GRoOD because its OOD objective supervises only the reject class, leaving the internal logit variance unconstrained and exposing the LSE operator's extreme-value sensitivity to single-dimension adversarial breakthroughs. Introducing *OOD label smoothing* to architecturally lock this missing variance unleashes EMR's full potential on GRoOD, yielding **GRoOD+OOD-LS+EMR** that attains the best robustness in reject-dominated regimes while providing both detection signals in a single model. Together, the two structural repairs cover both signal regimes, with RCE+EMR leading where LSE detection dominates.

1 INTRODUCTION

Neural networks are prone to overconfidence on Out-of-Distribution (OOD) samples (Hendrycks & Gimpel, 2016), posing risks in safety-critical applications. Semi-supervised methods leveraging auxiliary OOD data (Lee et al., 2018a; Hendrycks et al., 2018; Mohseni et al., 2020) mitigate this overconfidence, yet remain vulnerable to adversarial OOD attacks—perturbations crafted to evade detection (Sehwag et al., 2019; Hein et al., 2019). To enhance robustness, prevailing strategies apply adversarial training on auxiliary OOD data (Megyeri et al., 2021; Azizmalayeri et al., 2022) (e.g., ACET (Hein et al., 2019), ATOM (Chen et al., 2021), RCE (Augustin et al., 2020), ALOE (Chen et al., 2022), RODEO (Mirzaei et al., 2025); see Sec 2). Despite this progress, a key question remains: *is adversarial training on in-distribution (ID) data necessary for genuine OOD robustness?*

Through systematic evaluation with principled attacks (Auto-PGD, APGD (Croce & Hein, 2020)), we identify two complementary insights. First, *the importance of ID adversarial training for OOD robustness has been largely overlooked.* Defenses that omit ID adversarial training (ACET, ATOM) leave the original K ID classes unprotected, creating a structural vulnerability where attackers can easily induce overconfident OOD predictions by exploiting the unprotected ID logits. However, our second insight reveals a severe *clean-robust trade-off*: while ID adversarial training improves adversarial robustness, it systematically degrades the Maximum Softmax Probability (MSP) on clean ID samples. This structural suppression sharply narrows the margin of the very metric by which ID and OOD are typically separated, weakening clean OOD detection performance (Fig 1).

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.

These insights motivate a departure from the prevailing OOD-centric paradigm: rather than solely intensifying OOD adversarial training, *directly bypassing the degraded MSP* is a more effective and principled route to robustness. We pursue two structural routes.

Our first route explores the LogSumExp (LSE) energy score (Liu et al., 2020), which aggregates all K logits and is less sensitive to single-class MSP degradation. Switching to LSE, however, does not recover enforced separability on uniform-target defenses like RCE: RCE's uniform target locks the K logits' relative variance but leaves their absolute magnitude unanchored—a *variance lock* without an *energy anchor*—so the ID–OOD LSE separation is not enforced by training (Sec 5.2). To supply the missing anchor, we propose **RCE+EMR**, which introduces Energy-Margin Regularization (EMR), yielding substantial LSE-based robustness gains (e.g., +6.44% robust AUC on CIFAR-100 (Krizhevsky et al., 2009)).

Our second route, GRoOD (**G**eneralizable **Ro**bust **O**OD **D**etection), modifies the architecture by assigning OOD samples to an auxiliary $(K+1)$-th reject class Geifman & El-Yaniv (2019) during joint adversarial training. The reject class is unlocked by a weak structural mechanism we call the *conservation of probability mass*: although ID adversarial training suppresses the target-class MSP, the detection score $p_{K+1} = 1 - \sum_{k=1}^{K} p_k$ depends only on the *total* ID probability mass, not on its internal distribution. Since the total ID mass remains stable under the trade-off (see Sec 6.2), p_{K+1} stays largely unaffected by MSP degradation. Furthermore, z_{K+1} serves as a natural dynamic energy anchor, inherently bypassing the missing-anchor problem that plagues RCE. The efficacy is substantial: GRoOD achieves a +16.73 robust AUC improvement over RCE on CIFAR-100, scaling with the severity of the trade-off.

The Variance-Free Vulnerability. Given the success of EMR in anchoring RCE, a natural step is to apply it to GRoOD to further strengthen robustness across both detection metrics. Counterintuitively, EMR fails on GRoOD. We trace this to a deeper optimization asymmetry: the **variance-free vulnerability** (Sec 7). We uncover a striking symmetry between the two architectures. RCE acts as an implicit *variance lock* (due to its uniform target) but lacks an *energy anchor* (hence it needs EMR). Conversely, GRoOD possesses a natural *energy anchor* (z_{K+1}) but lacks a *variance lock*. Because GRoOD's OOD objective supervises only the reject class, the first K logits receive no gradient signal and their internal variance is left unconstrained; the resulting extreme-value sensitivity of the LSE operator ($\mathrm{LSE}(\mathbf{z}) \approx \max_j z_j$) lets a *single* fragile dimension dominate the LSE score, diluting EMR's suppressive power under attack (Sec 7).

This diagnosis motivates our final structural repair: **OOD label smoothing** (OOD-LS). By replacing the one-hot OOD target with a soft distribution, we explicitly enforce the missing variance constraint on GRoOD's K logits. Once this architectural "variance lock" is engaged, the single-dimension breakthrough is sealed. Consequently, *the exact same EMR that previously failed* is unleashed with full force on GRoOD (**GRoOD + OOD-LS + EMR**).

Crucially, this variance-locked architecture complements the RCE+EMR repair (Sec 8). Neither route dominates across all datasets—GRoOD's reject-class metric is preferable in spacious feature spaces (CIFAR-10, Tiny-ImageNet-200), while RCE+EMR's anchored LSE metric excels in congested ones (CIFAR-100) (Tab 2)—so our two structural repairs cover both regimes: GRoOD+OOD-LS+EMR provides both detection signals in a single model, retaining GRoOD's reject-class strength where GRoOD leads. Our contributions are:

- **Reframing the Problem.** We reframe adversarial OOD detection through a *pervasive clean-robust trade-off*: ID adversarial training—indispensable for OOD robustness—systematically degrades the MSP on clean ID samples, undermining the signal used for ID–OOD separation. We further show that this trade-off admits *structural mitigation*.
- **Two Structural Routes.** We identify two structural routes around the degraded MSP: an LSE energy score whose uniform-target variant leaves the energy scale unanchored, and a reject-class probability whose insulation from MSP degradation is characterized by a weak probability-mass conservation assumption.
- **The Variance-Free Vulnerability.** We identify a fundamental bottleneck in adversarial optimization: energy-margin regularization (EMR) dramatically lifts RCE but fails on GRoOD. We trace this asymmetry to a training-side root cause: the reject-class objective supervises *only* the reject logit, leaving the first K logits without gradient signal and hence with unconstrained internal variance, which the LogSumExp operator's extreme-value

sensitivity then exposes—formalizing a previously uncharacterized architectural duality (RCE locks variance but floats energy; GRoOD anchors energy but leaves variance free).

- **Two Complementary Structural Repairs.** We supply the LSE route's missing anchor via Energy-Margin Regularization (**RCE+EMR**) and patch the reject-class route's variance-free vulnerability via **OOD label smoothing** (OOD-LS), yielding **GRoOD + OOD-LS + EMR**. The two methods are complementary: RCE+EMR excels where LSE detection dominates, while GRoOD + OOD-LS + EMR attains the best robustness in reject-dominated regimes while providing both detection signals in a single model; the only dataset-specific choice is the reject weight α (selected by a signal-balance rule, Sec J.2).

2 Related Work

We review adversarial robustness and robust OOD detection; a fuller discussion of related paradigms is deferred to Appendix A.

2.1 Adversarial Robustness

Adversarial training (AT) (Madry et al., 2017) is the leading defense against adversarial attacks (Szegedy et al., 2013; Moosavi-Dezfooli et al., 2016; Kurakin et al., 2016; Papernot et al., 2017), solving a min-max problem with PGD as the inner maximizer. While AT improves robustness, it often degrades standard accuracy (Tsipras et al., 2019); refinements explore alternative losses, margin regularization, and structural logit constraints to improve this trade-off. Evaluating robustness requires strong attacks: APGD (Croce & Hein, 2020) adds automatic step-size adjustment, restarts, and momentum to PGD, and we employ it with its multi-targeted variant to evaluate OOD robustness under non-uniform confidence distributions.

2.2 Robust (End-to-End) Out-of-Distribution Detection

Semi-supervised methods with auxiliary OOD data (Hendrycks et al., 2018; Mohseni et al., 2020) reduce OOD overconfidence but remain highly vulnerable to adversarial attacks (Sehwag et al., 2019; Bitterwolf et al., 2020). Subsequent works adversarially train on auxiliary OOD data, varying the OOD target and augmentation (Hein et al., 2019; Chen et al., 2021; Augustin et al., 2020; Chen et al., 2022; Mirzaei et al., 2025), but attach it *ad hoc*, without examining how it *systematically reshapes the detection scores*. Consequently, the *clean-robust trade-off*—the suppression of MSP on clean ID samples and the resulting loss of ID–OOD separability—has not been characterized. We show it is structural and severity-proportional, and admits construction-level bypasses: an anchored energy score and a reject class.

Relation to recent advances. Our end-to-end setting is distinct from three orthogonal lines—preprocessing defenses, vision-language zero-shot detectors, and separate auxiliary detectors—detailed in Appendix A.2. The closest work is RCE (Augustin et al., 2020) (uniform-target joint adversarial training) and RODEO (Mirzaei et al., 2025) (synthetic OOD augmentation), the family our analysis targets.

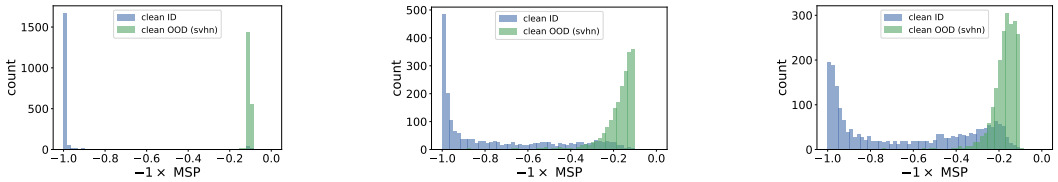
3 Preliminaries

Notation and Detection Scores. Let f_θ output logits $\mathbf{z} \in \mathbb{R}^K$, with $p_i = e^{z_i} / \sum_j e^{z_j}$. OOD detection scores include: the **MSP** $\max_i p_i$; the **LSE energy score** $\log \sum_i e^{z_i}$ (Liu et al., 2020); and, for architectures with an auxiliary $(K+1)$-th reject class (Sec 6), the **reject-class probability** p_{K+1}.

Training Objectives. Our baseline **RCE** (Augustin et al., 2020) adversarially trains both terms with the uniform target $U = [1/K, \ldots, 1/K]$:

$$\min_\theta \frac{1}{N} \sum_{i=1}^{N} \max_{\|\delta_i\|_\infty \le \epsilon} \ell(f_\theta(x_i^{\text{in}} + \delta_i^{\text{in}}), y_i^{\text{in}}) + \frac{\beta}{M} \sum_{j=1}^{M} \max_{\|\delta_j\|_\infty \le \epsilon} \ell(f_\theta(x_j^{\text{o}} + \delta_j^{\text{o}}), U), \tag{1}$$

Setting the perturbation radius to zero (i.e., clean training on both training terms) recovers the standard **OE** (Hendrycks et al., 2018) baseline. The OOD term in Eq (1) is analyzed in Sec 5.2.



(a) OE (b) ID adv. + OOD clean (c) Full adv. (RCE)

Figure 1: MSP distributions on CIFAR-10 (OOD: SVHN) under standard training, ID adversarial training with clean OOD, and full adversarial training (RCE).

Experimental Setup. We train WRN-40-4 (Zagoruyko & Komodakis, 2016) on CIFAR-10/100 and Tiny-ImageNet-200 with per-dataset auxiliary OOD data and 10-step PGD adversarial training (L_∞, $\epsilon = 8/255$), and additionally verify on the ViT architecture DeiT-S (Appendix Sec C). Each ID dataset is evaluated against a mixed OOD test set; full lists, schedules, and metrics are in Appendix B. We report AUC (↑) and FPR-95 (↓).

Threat Model and Attacks. Under the standard L_∞ threat model ($\epsilon = 8/255$), the attacker maximizes the detection score $S(x)$ on OOD samples. We use a multi-targeted APGD (TAPGD^{lgt}) that maximizes each ID-class logit while minimizing the reject-class logit when present, plus the weaker cross-entropy PGD (PGD^{ce}); formal definitions are in Appendix Sec B.5.

Table 1: Detection performance of existing detectors (AUC ↑). PGD^{ce} in gray; TAPGD^{lgt} in bold. Parentheses: improvement over the respective baseline (OE for SLOD; ACET for ATOM; RCE for ALOE/RODEO^U).

D^{in}	Method	Acc	Clean^o	PGD^{ce}	TAPGD^{lgt}
CIFAR10	AT	85.48	83.97	62.06	61.55
	OE	95.16	98.70	8.75	0.14
	SLOD	95.19	98.94 (+0.24)	0.65	0.05
	ACET	95.61	98.81	97.64	30.87
	ATOM	94.56	99.04 (+0.23)	98.56	6.60 (-25.13)
	RCE	85.33	94.50	83.16	**81.56**
	ALOE	93.53	92.27 (-2.23)	92.27	56.68 (-24.88)
	RODEO^U	85.90	94.43 (-0.07)	81.72	79.61 (-1.95)
CIFAR100	AT	60.05	67.65	37.36	35.69
	OE	77.65	90.20	40.87	0.02
	SLOD	76.56	91.27 (+1.07)	0.20	0.02
	ACET	78.62	90.79	90.67	4.84
	ATOM	78.36	93.25 (+2.46)	92.41	0.71 (-4.01)
	RCE	60.20	79.36	55.06	**52.71**
	ALOE	78.29	79.03 (-0.33)	79.03	13.33 (-39.38)
	RODEO^U	60.23	79.66 (+0.30)	52.76	50.57 (-2.14)

4 The Clean-Robust Trade-off in Robust OOD Detection

4.1 ID Adversarial Training Is Indispensable for OOD Robustness

Tab 1 reveals a fundamental tension. ID adversarial training is indispensable: only methods that include it resist strong APGD attacks—RCE (real auxiliary OOD data) and RODEO^U (uniform target on synthetic OOD)[1]. OOD-trained methods without it fail outright: ACET and ATOM, which adversarially train only their OOD term and leave ID classes cleanly trained, collapse to 4.84% and 0.71% under TAPGD (below plain AT's 35.69%); ALOE, which uses perturbed ID samples as surrogate OOD, similarly fails (13.33%). RCE attains the highest TAPGD AUC (52.71%) yet its clean OOD detection (79.36%) trails clean-trained OE (90.20%) and SLOD (91.27%).

The root cause is that ID adversarial training forces the model to distribute probability mass more conservatively, systematically lowering the maximum softmax probability (MSP) on clean ID samples. This mirrors the well-known accuracy-robustness trade-off on ID data (Tsipras et al., 2019) and the reduced confidence of adversarially trained models (Stutz et al., 2020), but with a fatal consequence for OOD detection: as we show next, the suppressed MSP directly undermines ID–OOD separability.

4.2 The Trade-off in OOD Detection: Compromised ID–OOD Separability

The consequence for detection is stark. To isolate the contribution of each training component, Fig 1 compares three models that differ by a single addition: standard training (OE), ID adversarial training with clean OOD, and full adversarial training (RCE). The MSP distributions are evaluated on clean ID and clean OOD (SVHN), without any test-time attack: adding ID adversarial training narrows the ID–OOD margin from 0.81 to 0.55, and joint OOD adversarial training reduces it

[1]RODEO's synthetic OOD augmentation does not surpass RCE's real auxiliary OOD data (CIFAR-10: 79.61% vs. 81.56%; CIFAR-100: 50.57% vs. 52.71%); we adopt real OOD data.

further to 0.46—roughly half the original margin. The same analysis on the Textures domain is in Appendix Sec L.

To bypass the degraded MSP, we pursue two structural routes: the first (Section 5) replaces MSP with a LogSumExp (LSE) energy score explicitly anchored via Energy-Margin Regularization (EMR); the second (Section 6) assigns OOD samples to an auxiliary reject class whose probability is structurally insulated from the MSP degradation.

5 LSE-Based Detection: Supplying the Missing Energy Anchor

Following the first route, we replace the degraded MSP with the LSE energy score, introduced in Sec 3. This section first shows that a direct switch to LSE yields substantial gains yet lacks an enforced margin, and then supplies the missing anchor via Energy-Margin Regularization (EMR).

5.1 Why a Direct Switch to LSE Is Insufficient

The LSE energy score aggregates all K logits via $\text{LSE}(\mathbf{z}) = \log \sum_{k=1}^{K} e^{z_k}$, avoiding the bottleneck of a single maximum and making it inherently less sensitive to internal probability mass redistribution. Empirically, switching RCE's inference metric from MSP to LSE (RCE+LSE-Det) yields substantial gains (Tab 2): robust AUC under TAPGD rises from 81.56% to 84.89% on CIFAR-10, from 52.71% to 67.13% on CIFAR-100, and from 80.24% to 93.27% on Tiny-ImageNet-200. These gains, however, rest on an *unanchored* score: as Sec 5.2 reveals, uniform-target training leaves the LSE separation without any enforced margin.

5.2 Uniform Targets Lock Variance but Leave Energy Unanchored

RCE's cross-entropy loss for the uniform target U (Section 3) decomposes as:

$$\mathcal{L}_{\text{RCE}} = -\frac{1}{K}\sum_{k=1}^{K} \log p_k = \log \sum_{k=1}^{K} e^{z_k} - \frac{1}{K}\sum_{k=1}^{K} z_k = \text{LSE}(z_{1:K}) - \bar{z}, \tag{2}$$

where $\bar{z}$ is the mean of the K logits. By Jensen's inequality, $\text{LSE}(z_{1:K}) - \bar{z} \geq \log K$, with equality iff $z_1 = \cdots = z_K$. Minimizing $\mathcal{L}_{\text{RCE}}$ thus enforces a *relative* constraint (logits toward equality) but is *translation-invariant*—adding a constant to all z_k leaves the softmax probabilities (and hence the loss) unchanged (Appendix Sec E)—so the training signal carries no information about the absolute energy scale.

Consequence: LSE separation is not enforced by training. Translation invariance means the objective imposes *no constraint* on the absolute energy scale of OOD logits, so the ID–OOD LSE separation is not a training objective. Empirically this appears as a small, unstable gap under RCE—why RCE+LSE-Det, despite improving over MSP, still lacks an enforced margin—which we repair in Sec 5.3.

5.3 RCE+EMR: Supplying the Missing Energy Anchor

Since uniform-target architectures lack a natural energy anchor—the uniform target pins the variance but not the absolute scale (Case 1 of Proposition 1)—we supply one explicitly via **RCE with Energy-Margin Regularization (RCE+EMR)**:

$$\mathcal{L}_{\text{EMR}} = \max\left(0, m - (\text{LSE}(z^{\text{ID}}_{1:K}) - \text{LSE}(z^{\text{OOD}}_{1:K}))\right), \tag{3}$$

where $m > 0$ is the desired margin. This establishes a complementary dynamic: **RCE** enforces $z_1 \approx \cdots \approx z_K$, flattening the logit landscape (variance lock); **EMR** supplies the missing energy anchor by encouraging $\text{LSE}_{\text{ID}} - \text{LSE}_{\text{OOD}} \geq m$, turning the unconstrained LSE separation into an explicitly encouraged margin. In practice we apply Eq (3) at the batch level: the hinge is evaluated on the difference between the mean LSE over ID samples and the mean LSE over OOD samples in each training batch, so no explicit ID–OOD pairing is required. The total objective is $\mathcal{L} = \mathcal{L}_{\text{RCE}} + \lambda_{\text{EMR}}\mathcal{L}_{\text{EMR}}$. By default, we set $m = 1$ and $\lambda_{\text{EMR}} = 0.1$ for RCE+EMR on all datasets (GRoOD+OOD-LS+EMR of Sec 8 uses $m = 5$). Hyperparameter sensitivity analysis is provided in Appendix J.

Tab 2 reports the outcome: under TAPGD, robust AUC rises from 84.89% to 86.26% on CIFAR-10, 67.13% to 73.57% on CIFAR-100, and 93.27% to 95.03% on Tiny-ImageNet-200, while clean OOD detection is largely preserved (CIFAR-100: 85.54%, -0.58; Tiny-ImageNet-200: 97.14%, $+2.70$). Virtual-BG (Megyeri et al., 2021), which optimizes an LSE-based loss without structural constraints, attains only 77.07%/56.24%—confirming that explicit anchoring, not merely an LSE objective, drives robustness. Appendix Fig 3 corroborates this, and Appendix Sec B.6 further verifies these conclusions under a direct LSE-maximizing attack, ruling out attack-objective misalignment. RCE+EMR is thus a strong, explicitly anchored LSE-based repair.

Table 2: The two structural routes vs. RCE. Parentheses: improvement over RCE+LSE-Det (RCE+EMR) and over RCE (GRoOD); **bold**: best under TAPGD; AA: AutoAttack.

D^{in}	Method	Acc	AA	Clean^o		TAPGD^{lgt}	
				AUC	FPR-95	AUC	FPR-95
CIFAR-10	RCE	85.33	51.55	94.50	22.93	81.56	46.40
	RCE+LSE-Det	85.33	51.55	95.55	18.77	84.89	44.76
	Virtual-BG (LSE)	86.05	49.95	90.72	38.37	77.07	58.81
	RCE+EMR	85.44	51.64	95.09 (-0.46)	20.65	86.26 (+1.37)	43.85
	GRoOD	85.93	51.44	95.67 (+1.17)	18.38	**88.15** (+6.59)	**39.25**
CIFAR-100	RCE	60.20	28.15	79.36	56.14	52.71	75.84
	RCE+LSE-Det	60.20	28.15	86.12	48.29	67.13	72.04
	Virtual-BG (LSE)	60.73	26.94	80.43	56.76	56.24	77.62
	RCE+EMR	60.12	28.11	85.54 (-0.58)	43.48	**73.57** (+6.44)	**68.39**
	GRoOD	60.57	27.73	85.39 (+6.03)	46.55	69.44 (+16.73)	69.55
Tiny-200	RCE	47.19	28.09	92.12	41.08	80.24	65.07
	RCE+LSE-Det	47.19	28.09	94.44	22.16	93.27	36.85
	RCE+EMR	46.33	28.09	97.14 (+2.70)	9.31	95.03 (+1.76)	12.24
	GRoOD	48.39	28.13	97.72 (+5.60)	8.69	**95.56** (+15.32)	**12.59**

6 GROOD: BYPASSING THE DEGRADED MSP VIA THE REJECT CLASS

Following the second route, we bypass the degraded MSP altogether via a reject class. This section formalizes the reject-class objective (**GRoOD**), analyzes why its probability is structurally insulated from the MSP degradation, and validates it empirically.

6.1 FORMALIZING THE REJECT-CLASS BYPASS

ID adversarial training suppresses the MSP of clean ID samples, degrading the very signal for ID–OOD separation (Sec 4.2). **GRoOD** (Generalizable Robust OOD Detection) follows the objective in Eq (1) unchanged except for the OOD label: instead of the uniform target U, OOD samples are assigned to an auxiliary $(K+1)$-th reject class—the OOD term becomes $\ell\big(f_\theta(x_j^o + \delta_j^o), K+1\big)$. At inference, GRoOD uses the reject-class probability p_{K+1} (Section 3): a sample is flagged OOD if $p_{K+1} > \tau$, and ID classification follows $\arg\max_{k\in\{1,\dots,K\}} p_k(x)$. We study the effect of using multiple reject classes in Appendix Sec J.5.

6.2 MECHANISTIC ANALYSIS: PROBABILITY-MASS CONSERVATION

Intuitively, the reject class should escape the very degradation that cripples MSP: MSP depends on the single maximal class—precisely the quantity ID adversarial training suppresses—whereas the reject-class score is governed by the total ID energy relative to z_{K+1}, and is thus far less affected by the internal probability-mass redistribution behind the trade-off.

Let γ denote the MSP drop that ID adversarial training induces on a clean ID sample (RCE's detection margin is then linearly compressed by γ, degrading proportionally).

GRoOD's Structural Insensitivity. GRoOD's OOD term reduces to $\text{Softplus}\big(\text{LSE}(z_{1:K}) - z_{K+1}\big)$ (Appendix Sec F), so z_{K+1} acts as a **dynamic anchor**—an attacker cannot inflate the ID energy $\text{LSE}(z_{1:K})$ without driving up the penalty through z_{K+1}. Equivalently, the reject-class score is a monotone function of the energy gap:

$$p_{K+1} \;=\; \frac{e^{z_{K+1}}}{e^{z_{K+1}} + \sum_{k=1}^{K} e^{z_k}} \;=\; \sigma\big(z_{K+1} - \text{LSE}(z_{1:K})\big), \tag{4}$$

where σ is the logistic function. GRoOD's score is thus governed entirely by a *single scalar*—the energy gap $\text{LSE}(z_{1:K}) - z_{K+1}$—and is invariant to how that energy is distributed among the K ID

classes: as long as the ID energy stays stable relative to z_{K+1}, p_{K+1} remains low regardless of how γ redistributes confidence among the ID classes.

Two observations support this stability: (i) **by construction**, GRoOD's ID cross-entropy term penalizes reject-class probability on ID samples (its gradient w.r.t. z_{K+1} is $p_{K+1} > 0$), pushing z_{K+1} down there; (ii) **empirically**, p_{K+1} degrades far less than MSP—the MSP margin shrinks by roughly half under RCE, and 25% of ID $-1\times$ MSP scores land in $[-0.4, -0.2]$, whereas under GRoOD only 17% of ID p_{K+1} exceeds 0.5 (Appendix Fig 5).

Thus, while RCE's MSP margin shrinks with γ, GRoOD's score is only mildly affected, so its gain over RCE grows with trade-off severity ($\Delta_{\text{GRoOD}} \propto \gamma$). The same holds for RCE+EMR (CIFAR-100: $+6.44$ vs. CIFAR-10: $+1.37$). Controlled ϵ-sweeps (Sec J.4) confirm this: as ϵ grows, GRoOD's advantage grows systematically.

6.3 Empirical Validation

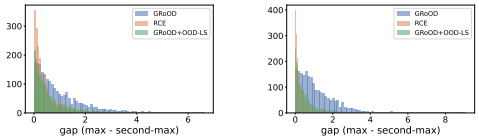
GRoOD achieves substantial gains across all datasets (Tab 2): **+16.73%** robust AUC over RCE on CIFAR-100 (69.44% vs. 52.71%), **+15.32%** on Tiny-ImageNet-200 (95.56% vs. 80.24%), and $+6.59\%$ on CIFAR-10 (88.15% vs. 81.56%). Reject-class methods on synthetic OOD data (RODEO (Mirzaei et al., 2025), Appendix Sec J.1) do not surpass GRoOD on real auxiliary OOD data, which we therefore use throughout. These gains, with improved clean OOD detection, establish GRoOD as a structurally superior alternative to MSP-based detection under ID adversarial training.

Fig 5 verifies this at the score level: under RCE the $-1\times$ MSP distributions of clean ID and attacked OOD overlap in $[-0.4, -0.2]$, whereas GRoOD's reject-class distributions remain far more separable—keeping ID–OOD separable precisely where MSP-based detection degrades. Notably, GRoOD's value lies not in driving p_{K+1} to zero on ID samples but in preserving the ID–OOD separation gap: as long as the gap holds, detection is reliable regardless of how the suppressed mass is redistributed within the ID classes. We contrast GRoOD with prior reject-class methods in Appendix Sec A.3.

7 The Variance-Free Vulnerability: When Energy Regularization Fails to Transfer

Sections 6 and 5 established two structural routes around the degraded MSP: GRoOD's reject-class probability and RCE+EMR's anchored LSE detection. Transferring EMR to GRoOD yields a stark asymmetry—on CIFAR-10, RCE+EMR gains $+1.37$ robust LSE AUC (84.89% $\rightarrow$ 86.26%) versus GRoOD+EMR's $+0.08$; on CIFAR-100, $+6.44$ (67.13% $\rightarrow$ 73.57%) versus $+0.13$.

Why does EMR rescue uniform-target architectures yet fail on the reject-class architecture? We trace this to a training-side bottleneck in adversarial optimization—the reject-class objective leaves the first K logits without supervisory signal—which we term the **variance-free vulnerability**.



(a) SVHN (b) Textures

Figure 2: Single-dimension breakthrough after TAPGD^{lgt} (SVHN, Textures): distributions of the max$-$second-max ID-logit gap. GRoOD's unconstrained variance admits isolated logit spikes (large gap) that dominate the LSE score; RCE's variance lock keeps logits flat; GRoOD+OOD-LS (Section 8.1) recovers most of the lock.

7.1 The Training-Side Root: Zero Supervisory Signal on the First K Logits

As established in Sec 6.2, GRoOD's OOD objective depends on the ID logits only through the aggregate energy LSE$(z_{1:K})$, so its gradient with respect to the first K logits vanishes once the model rejects OOD samples (Case 2 of Proposition 1): those logits receive *zero supervisory signal* at convergence, and their internal variance is unconstrained—an artifact of initialization and optimization path. RCE's uniform target, by contrast, supplies a per-class signal driving the K logits toward mutual equality (a *variance lock*, Case 1).

This variance-free state becomes exploitable through the extreme-value sensitivity of the LSE operator: for any logit vector,

$$\max_j z_j \leq \text{LSE}(\mathbf{z}) \leq \max_j z_j + \log K, \tag{5}$$

so $\text{LSE}(\mathbf{z})$ is dominated by its maximum element, and an attacker can inflate the LSE score by exploiting a *single* fragile dimension rather than all K. The landscape is severely asymmetric: EMR's aggregate push must hold down the full logit vector, while the attacker only re-spikes one already-prominent dimension—so EMR's gradient is perpetually diluted by single-dimension breakthroughs (Fig 2). Notably, this vulnerability concerns only the LSE score (and hence EMR); GRoOD's reject-class detection remains anchored under the same attack (Appendix Fig 5).

Proposition 1 (Fixed-point characterization of the OOD targets)**.** *Let $\mathbf{z} \in \mathbb{R}^K$ be the ID-class logits of an OOD sample, and let z_{K+1} be the reject logit. Consider the set of loss-minimizing configurations of the OOD term.*

1. ***Uniform target (RCE).*** *Minimizing $\mathcal{L}_{RCE} = LSE(z_{1:K}) - \bar{z}$ yields $LSE(z_{1:K}) - \bar{z} = \log K$ (Jensen's inequality, as in Sec 5.2), with equality* iff *$z_1 = \cdots = z_K$. Hence every loss-minimizing configuration has* zero variance *among the K ID logits, and the loss is invariant to a global translation $z_{1:K} \to z_{1:K} + c\mathbf{1}$. The minimizer set is therefore the one-dimensional manifold $\{c\mathbf{1} : c \in \mathbb{R}\}$—**variance locked, scale unanchored**.*
2. ***Reject target (GRoOD).*** *Minimizing $\mathcal{L}_{GRoOD} = Softplus(LSE(z_{1:K}) - z_{K+1})$ drives $LSE(z_{1:K}) - z_{K+1} \to -\infty$; since the loss is strictly decreasing in this gap, its infimum 0 is approached (not attained) as the gap grows. Because the loss depends on $\mathbf{z}_{1:K}$ only through the* aggregate *LSE, for every $\delta > 0$ the δ-sublevel set of the loss contains configurations with* arbitrary *internal structure among the first K logits, as long as their LSE matches the target gap. The loss-minimizing* family *is therefore the $(K-1)$-dimensional equal-LSE surface $\{\mathbf{z}_{1:K} : LSE(\mathbf{z}_{1:K}) = const\}$ with the gap to z_{K+1} driven to $+\infty$—**anchored in the energy gap, variance free**.*

The proof is provided in Appendix G.

Corollary 1 (Why EMR fails on GRoOD but succeeds on RCE)**.** *The energy-margin loss $\mathcal{L}_{EMR}$ (Eq (3)) pushes down the* aggregate *LSE of OOD samples. Under the uniform target (Case 1), the trained OOD logits lie on $\{c\mathbf{1}\}$, so the aggregate push translates the configuration uniformly downward—EMR is fully effective, supplying the anchor without internal distortion. Under the reject target (Case 2), the ID logits may carry arbitrary internal variance, so EMR's push leaves the* relative *configuration untouched: an adversary can spike a single already-prominent ID logit to recover the LSE score, and the near-zero gradient $p_{ID}p_j^{ID}$ offers no training-time resistance. This explains the asymmetry above.*

This asymmetry directly motivates the architectural variance lock introduced in Sec 8.1.

Proposition 1 also reveals a formal *architecture–objective duality*: each architecture solves one optimization flaw while exposing itself to the other, explaining why naive integration fails and motivating GRoOD+OOD-LS+EMR of Sec 8; see Appendix Sec H for the full discussion.

8 Variance Locking Unleashes EMR on GRoOD

The variance-free vulnerability diagnosed in Sec 7 admits an elegant remedy: if unconstrained logit variance is the bottleneck, architecturally patching this gap should unleash EMR on GRoOD.

8.1 OOD Label Smoothing: Architecting the Missing Variance Lock

GRoOD's extreme-value sensitivity stems from its strict one-hot OOD target $[0, \ldots, 0, 1]$, which instructs the optimizer solely to maximize the reject class, leaving the internal distribution of the K ID logits unconstrained. We address this with **OOD Label Smoothing (OOD-LS)**, which applies the classical label-smoothing idea (Szegedy et al., 2016) to the OOD target, replacing it with a soft distribution that enforces relative equality among the K ID logits:

$$\text{OOD-LS target}: \quad \mathbf{y}_{\text{soft}}^{\text{OOD}} = \left[\frac{1-\alpha}{K}, \frac{1-\alpha}{K}, \ldots, \frac{1-\alpha}{K}, \alpha\right], \tag{6}$$

Table 3: Unleashing EMR on GRoOD via OOD label smoothing.

D^{in}	Method	ID Acc	AA	Clean^o AUC		TAPGD^{lgt} AUC	
				LSE	Reject	LSE	Reject
CIFAR-10	GRoOD	85.93	51.44	–	95.67	–	88.15
	RCE+EMR	85.44	51.64	95.09	–	86.26	–
	GRoOD + OOD-LS + EMR	86.29	52.01	95.19	**96.01**	87.94	**88.52**
CIFAR-100	GRoOD	60.57	27.73	–	85.39	–	69.44
	RCE+EMR	60.12	28.11	85.54	–	**73.57**	–
	GRoOD + OOD-LS + EMR	60.03	28.16	81.60	**84.90**	**72.66**	69.49
Tiny-200	GRoOD	48.39	28.13	–	97.72	–	95.56
	RCE+EMR	46.33	28.09	97.14	–	95.03	–
	GRoOD + OOD-LS + EMR	47.80	28.10	96.90	**98.04**	95.04	**95.76**

where $\alpha \in (0, 1)$ is the weight of the reject class in the OOD target (the *reject weight* in our implementation). Setting $\alpha = 1$ recovers the vulnerable vanilla GRoOD (the one-hot reject target), while a smaller α allocates more supervision to the K ID classes and enforces stricter variance control. Because a larger α emphasizes the reject-class signal and a smaller one the LSE signal, we set the default per dataset based on which signal is stronger (Appendix Sec J.2); the exact values are summarized in Appendix Sec B.1.

This soft target supplies a strict restoring force on the ID logits—whenever an APGD adversary spikes a single logit, the resulting gradient penalizes the deviation and pulls it back toward the mean (derivation in Appendix Sec I)—re-establishing the missing **variance lock** within the reject-class architecture.

8.2 The Completed Model: GRoOD+OOD-LS+EMR

With the variance lock engaged, the previously futile EMR is fully unleashed: applying the EMR loss (Eq (3), $m = 5$) to the variance-locked GRoOD architecture lets OOD-LS flatten the logit landscape while EMR pushes the OOD energy manifold downward (Fig 2). Table 3 reports robust detection under TAPGD^{lgt} for **GRoOD + OOD-LS + EMR** alongside its two route-level counterparts, GRoOD (reject-class) and RCE+EMR (anchored LSE); LSE robustness under APGD^{LSE} is in Appendix Sec B.6.

8.3 Complementarity of the Two Structural Repairs

The two structural repairs are complementary. In spacious feature spaces (CIFAR-10, Tiny-ImageNet-200), where the reject-class signal is stronger, GRoOD+OOD-LS+EMR attains the best robust detection under TAPGD^{lgt} (88.52% and 95.76%) while its LSE metric stays competitive with the RCE+EMR specialist (87.94% vs. 86.26%), and it is likewise best at larger scale (ImageNet-200, Appendix Sec D). The sole exception is congested CIFAR-100, where the RCE+EMR specialist wins on its home LSE metric (73.57% vs. 72.66%), though GRoOD+OOD-LS+EMR still improves the reject-class metric over plain GRoOD (69.49% vs. 69.44%). The default per-dataset α (Appendix Sec J.2) lets each model emphasize its stronger signal; in practice, RCE+EMR suits LSE-dominated CIFAR-100 and GRoOD+OOD-LS+EMR elsewhere.

9 Conclusion

This work re-examines adversarial OOD detection through a pervasive clean-robust trade-off: ID adversarial training is indispensable for robustness, yet it systematically degrades the MSP of clean samples and undermines ID–OOD separation. We established two structural routes around the degraded MSP—**GRoOD** and **RCE+EMR**—and uncovered the **variance-free vulnerability** that blocks energy-margin regularization from transferring to GRoOD, patching it with **OOD label smoothing** to yield **GRoOD+OOD-LS+EMR**. Together, the two complementary repairs cover both signal regimes: RCE+EMR excels where LSE detection dominates, and GRoOD+OOD-LS+EMR where reject-class detection dominates.

AI USE STATEMENT

In this work, we used generative AI tools for editing the manuscript to improve readability, including grammar, spelling, and word choice (a task with recommended disclosure). We have not used generative AI tools for tasks with required disclosure—namely, designing the proposed methods, formulating the theoretical analysis, implementing the methods, writing the experimental code, running the experiments, or analyzing the experimental results; these tasks were performed by the authors. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Training settings, hyperparameters (architecture choices, learning rate schedule, batch sizes, and attack parameters), and evaluation protocols are fully specified in Sec 3 and Appendix B.1. Attack formulations are provided in Appendix B.5. All datasets used are publicly available. An anonymized code release is provided in the supplementary material.

REFERENCES

Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in-and out-distribution improves explainability. In *European Conference on Computer Vision (ECCV)*, pp. 228–245. Springer, 2020.

Mohammad Azizmalayeri, Arshia Soltani Moakhar, Arman Zarei, Reihaneh Zohrabi, Mohammad Manzuri, and Mohammad Hossein Rohban. Your out-of-distribution detection method is not robust! *Advances in Neural Information Processing Systems*, 35:4887–4901, 2022.

Julian Bitterwolf, Alexander Meinke, and Matthias Hein. Provable worst case guarantees for the detection of out-of-distribution data, 2020. arXiv preprint arXiv:2007.08473.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. *In Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2021.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Robust out-of-distribution detection for neural networks. In *The AAAI-22 Workshop on Adversarial Machine Learning and Beyond*, 2022.

Wenxi Chen, Raymond A Yeh, Shaoshuai Mou, and Yan Gu. Leveraging perturbation robustness to enhance out-of-distribution detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4724–4733, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 2206–2216, 2020.

Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

Nicola Franco, Daniel Korth, Jeanette Miriam Lorenz, Karsten Roscher, and Stephan Guennemann. Diffusion denoised smoothing for certified and adversarial robust out-of-distribution detection. *arXiv preprint arXiv:2303.14961*, 2023.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, pp. 2151–2159. PMLR, 2019.

Sven Gowal, Jonathan Uesato, Chongli Qin, Po-Sen Huang, Timothy Mann, and Pushmeet Kohli. An alternative surrogate loss for pgd-based adversarial testing. *arXiv preprint arXiv:1910.09338*, 2019.

Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41–50, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Wang, Dongdong Li, Dawn Song, Jacob Steinhardt, and Justin Gilmer. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, pp. 8759–8773, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009. Technical report, University of Toronto.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018a.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018b.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

István Megyeri, István Hegedüs, and Márk Jelasity. Robust classification combined with robust out-of-distribution detection: An empirical analysis. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

Hossein Mirzaei, Mohammad Jafari, Hamid Reza Dehbashi, Ali Ansari, Sepehr Ghobadi, Masoud Hadi, Arshia Soltani Moakhar, Mohammad Azizmalayeri, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. Rodeo: Robust outlier detection via exposing adaptive out-of-distribution samples. *arXiv preprint arXiv:2501.16971*, 2025.

Yichuan Mo, Dongxian Wu, Yifei Wang, Yiwen Guo, and Yisen Wang. When adversarial training meets vision transformers: Recipes from training to architecture. *Advances in Neural Information Processing Systems*, 35:18599–18611, 2022.

Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5216–5223, 2020.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Vikash Sehwag, Arjun Nitin Bhagoji, Liwei Song, Chawin Sitawarin, Daniel Cullina, Mung Chiang, and Prateek Mittal. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pp. 105–116, 2019.

Naman Deep Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. *Advances in Neural Information Processing Systems*, 36:13931–13955, 2023.

David Stutz, Matthias Hein, and Bernt Schiele. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In *International Conference on Machine Learning*, pp. 9155–9166, 2020.

Yiyou Sun, Chuan Guo, and Yinar Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 144–157, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Lin Tao, Kaiwen Yang, Yige Feng, Yixuan Wang, and Xing Xie. Benchmarking robustness of out-of-distribution detection. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Sachin Vernekar, Ashish Gaurav, Taylor Denouden, Buu Phan, Vahdat Abdelzad, Rick Salay, and Krzysztof Czarnecki. Analysis of confident-classifiers for out-of-distribution detection. *arXiv preprint arXiv:1904.12220*, 2019.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.