# One Model, Many Behaviors: Training-Induced Effects on Out-of-Distribution Detection

## Supplementary Material

This supplementary document expands on the main manuscript. It provides full experimental details (Appendix A), comprehensive results that support and extend our analyses (Appendix B), and additional experiments with Vision Transformers (ViT) (Appendix C).

## A. Experimental Details

### A.1. Implementation Details

**Software stack.** Our experimental framework is built upon the OpenOOD [56, 60, 61] framework. Specifically, we utilize the public fork from Humblot-Renaux [24], as it provides a GRAM [44] implementation that follows the official implementation details, which also leverages information from intermediate layers. We extend the model zoo by integrating 56 ImageNet checkpoints, adapted from the collection provided by [12]. The full list of all models used in this study, along with their training categories and performance metrics, is detailed in Tab. 2. All evaluations in this study are executed within this unified framework.

To broaden the evaluation scope, we also enrich the data layer with two additional OOD categories: (i) *extreme-OOD* including MNIST [28] and Fashion-MNIST [54], and (ii) *synthetic-OOD* including the unit-test data provided by NINCO [1]. This setup ensures reproducibility and fair comparison with a broad set of diverse training strategies, OOD test sets, and existing state-of-the-art OOD detection methods.

**Hardware and system configuration.** All experiments were executed on a workstation equipped with an Intel Core i9-9900X (10 cores, 3.5 GHz) and two NVIDIA GPUs (RTX 2080Ti + RTX 3090). The software environment consisted of Ubuntu 22.04, Python 3.10, PyTorch 2.0.1, and CUDA 11.8.

**OOD detection methods.** Many OOD detection methods require a configuration phase prior to evaluation, for which we strictly follow the OpenOOD benchmark protocols to ensure comparability. This process includes two types of setup: some methods are calibrated on the ID training set to compute statistics or other parameters, while others have crucial hyperparameters that are tuned on a held-out validation set containing both ID and OOD samples. Although detectors ship with default hyperparameters, these defaults are typically tuned to a vanilla training recipe, which can risk biasing the comparison. Re-optimizing all parameters, therefore, provides a fair test across the diverse training strategies evaluated here. The exact settings and hyperparameter search spaces adopted for each method are detailed in Tab. 1.

| **Method** | **Hyperparameter Search Space** |
|---|---|
| MSP [18] | |
| MLS [21] | |
| EBO [34] | |
| ODIN [31] | temperature $T \in \{1, 10, 100, 1000\}$<br>perturbation mag. $\sigma \in \{0.0014, 0.0028\}$ |
| TempScale [15] | |
| KLM [21] | |
| GEN [35] | gamma $\in \{0.01, 0.1, 0.5, 1, 2, 5, 10\}$<br>top-M classes $\in \{10, 50, 100, 200, 500, 1000\}$ |
| KNN [48] | $K \in \{50, 100, 200, 500, 1000\}$ |
| MDS [29] | |
| RMDS [42] | |
| SHE [59] | |
| ViM [51] | |
| ASH [9] | percentile $\in \{65, 70, 75, 80, 85, 90, 95\}$ |
| ReAct [47] | percentile $\in \{70, 80, 85, 90, 95, 99\}$ |
| DICE [46] | percentile $\in \{10, 30, 50, 70, 90\}$ |
| SCALE [55] | percentile $\in \{65, 70, 75, 80, 85, 90, 95\}$ |
| NNGuide [39] | |
| fDBD [33] | normalization $\in \{\text{true}, \text{false}\}$ |
| GRAM [44] | |
| ATS [27] | |
| GradNorm [23] | |

Table 1. Overview of hyperparameter search space for all considered OOD detection methods.

## B. Detailed Results

**Does higher ID accuracy imply better OOD detection?** To validate that our findings are not an artifact of the AUROC metric, we perform an equivalent analysis using the False Positive Rate at 95% True Positive Rate (FPR95). As shown in Fig. 1, this analysis plots ID classification accuracy against FPR95, where lower values signify better OOD detection performance.

This analysis quantitatively confirms the visually observed mirrored fall-then-rise pattern. Consistent with the AUROC results, the overall relationship between accuracy and FPR95 yields a weak global correlation (Spearman's $\rho = -0.04, p \ll 0.001$). Similar to the AUROC analysis, in the low-to-baseline accuracy regime, performance is primarily driven by adversarially trained models, which exhibit a strong negative correlation between ID classification accuracy and FPR95 (OOD performance improves). Con-

|  | Model | Category | In-Distribution |  | Out-of-Distribution |  |
|---|---|---|---|---|---|---|
|  |  |  | Accuracy ↑ | ECE ↓ | AUROC ↑ | FPR95 ↓ |
| • | Original Baseline [16] | Baseline | 76.19 | 3.62 | 89.07$^{\pm 9.95}$ | 42.05$^{\pm 27.62}$ |
| • | PGD-AT ($l_2, \epsilon = 0$) [36, 43] | Adversarial Training | 75.90 | 3.50 | 88.90$^{\pm 10.22}$ | 42.27$^{\pm 28.36}$ |
| • | PGD-AT ($l_2, \epsilon = 0.01$) [36, 43] | Adversarial Training | 75.69 | 3.02 | 88.50$^{\pm 9.97}$ | 44.42$^{\pm 27.39}$ |
| • | PGD-AT ($l_2, \epsilon = 0.03$) [36, 43] | Adversarial Training | 75.88 | 2.80 | 88.37$^{\pm 9.76}$ | 44.96$^{\pm 26.67}$ |
| • | PGD-AT ($l_2, \epsilon = 0.05$) [36, 43] | Adversarial Training | 75.58 | 2.68 | 88.86$^{\pm 10.05}$ | 42.97$^{\pm 28.03}$ |
| • | PGD-AT ($l_2, \epsilon = 0.1$) [36, 43] | Adversarial Training | 74.86 | 2.21 | 88.53$^{\pm 10.54}$ | 44.48$^{\pm 28.62}$ |
| • | PGD-AT ($l_2, \epsilon = 0.25$) [36, 43] | Adversarial Training | 74.15 | 1.79 | 87.90$^{\pm 10.14}$ | 46.97$^{\pm 26.49}$ |
| • | PGD-AT ($l_2, \epsilon = 0.5$) [36, 43] | Adversarial Training | 73.22 | 1.58 | 87.50$^{\pm 10.55}$ | 48.06$^{\pm 26.35}$ |
| • | PGD-AT ($l_2, \epsilon = 1$) [36, 43] | Adversarial Training | 70.50 | 3.35 | 85.99$^{\pm 10.97}$ | 51.61$^{\pm 26.49}$ |
| ● | PGD-AT ($l_2, \epsilon = 3$) [36, 43] | Adversarial Training | 62.86 | 9.06 | 79.85$^{\pm 10.42}$ | 65.71$^{\pm 21.19}$ |
| ● | PGD-AT ($l_2, \epsilon = 5$) [36, 43] | Adversarial Training | 56.15 | 12.65 | 74.45$^{\pm 9.75}$ | 71.33$^{\pm 18.70}$ |
| ▪ | PGD-AT ($l_\infty, \epsilon = 0.5$) [36, 43] | Adversarial Training | 73.75 | 1.23 | 87.25$^{\pm 9.86}$ | 48.39$^{\pm 23.86}$ |
| ▪ | PGD-AT ($l_\infty, \epsilon = 1.0$) [36, 43] | Adversarial Training | 72.13 | 2.78 | 85.86$^{\pm 9.93}$ | 53.04$^{\pm 25.30}$ |
| ▪ | PGD-AT ($l_\infty, \epsilon = 2.0$) [36, 43] | Adversarial Training | 69.13 | 4.80 | 83.42$^{\pm 10.24}$ | 58.99$^{\pm 25.04}$ |
| ■ | PGD-AT ($l_\infty, \epsilon = 4.0$) [36, 43] | Adversarial Training | 63.94 | 8.92 | 80.26$^{\pm 10.39}$ | 65.47$^{\pm 20.56}$ |
| ■ | PGD-AT ($l_\infty, \epsilon = 8.0$) [36, 43] | Adversarial Training | 54.55 | 13.28 | 70.62$^{\pm 11.43}$ | 73.24$^{\pm 16.29}$ |
| • | AutoAugment (270Ep) [7] | Augmentations | 77.52 | 2.74 | 89.54$^{\pm 10.06}$ | 40.22$^{\pm 29.92}$ |
| ● | FastAutoAugment (270Ep) [32] | Augmentations | 77.69 | 3.58 | 88.77$^{\pm 9.94}$ | 42.82$^{\pm 29.14}$ |
| Y | StyleAugment [14, 61] | Augmentations | 74.68 | 1.91 | 88.35$^{\pm 10.17}$ | 43.92$^{\pm 28.09}$ |
| ♦ | RandAugment (270Ep) [8] | Augmentations | 77.65 | 3.26 | 88.78$^{\pm 9.57}$ | 43.16$^{\pm 28.80}$ |
| ⅄ | AugMix (180Ep) [19] | Augmentations | 77.63 | 1.88 | 89.72$^{\pm 9.51}$ | 40.78$^{\pm 27.29}$ |
| × | DeepAugment [20] | Augmentations | 76.76 | 2.37 | 88.03$^{\pm 9.49}$ | 45.72$^{\pm 25.13}$ |
| ◆ | DeepAugment + AugMix [20] | Augmentations | 75.89 | 2.82 | 88.56$^{\pm 11.00}$ | 41.10$^{\pm 30.16}$ |
| ⬟ | RegMixup [41] | Augmentations | 76.69 | 2.94 | 88.14$^{\pm 9.30}$ | 45.87$^{\pm 26.13}$ |
| ▲ | Diffusion-like Noise [25] | Augmentations | 67.26 | 1.79 | 84.24$^{\pm 11.48}$ | 53.83$^{\pm 25.72}$ |
| ◀ | NoisyMix [11] | Augmentations | 77.14 | 12.92 | 86.41$^{\pm 10.25}$ | 50.67$^{\pm 27.04}$ |
| ▶ | OpticsAugment [38] | Augmentations | 74.25 | 3.02 | 88.88$^{\pm 10.67}$ | 41.10$^{\pm 28.71}$ |
| ▼ | PRIME [37] | Augmentations | 76.99 | 2.79 | 88.63$^{\pm 9.43}$ | 44.04$^{\pm 26.30}$ |
| ■ | PixMix (90Ep) [22] | Augmentations | 77.43 | 1.49 | 88.07$^{\pm 8.94}$ | 44.60$^{\pm 26.33}$ |
| + | PixMix (180Ep) [22] | Augmentations | 78.18 | 2.19 | 87.29$^{\pm 8.62}$ | 46.45$^{\pm 25.65}$ |
| ★ | MixUp [58] | Augmentations | 77.55 | 20.40 | 84.13$^{\pm 11.08}$ | 52.89$^{\pm 28.08}$ |
| ⬢ | CutMix [57] | Augmentations | 78.62 | 18.79 | 79.65$^{\pm 11.57}$ | 58.89$^{\pm 24.11}$ |
| ✚ | ShapeNet (SIN) [14] | Augmentations | 60.22 | 6.80 | 84.65$^{\pm 12.28}$ | 47.16$^{\pm 34.05}$ |
| \| | ShapeNet (SIN+IN) [14] | Augmentations | 76.74 | 4.82 | 88.62$^{\pm 9.23}$ | 44.79$^{\pm 26.48}$ |
| _ | ShapeNet (SIN+IN → IN) [14] | Augmentations | 74.68 | 1.91 | 88.34$^{\pm 10.18}$ | 43.96$^{\pm 28.08}$ |
| ≺ | Texture/Shape debiased [30] | Augmentations | 76.92 | 3.21 | 87.72$^{\pm 10.02}$ | 46.02$^{\pm 27.19}$ |
| ≻ | Texture/Shape-Shape biased [30] | Augmentations | 76.31 | 2.38 | 88.26$^{\pm 9.84}$ | 44.72$^{\pm 27.46}$ |
| ✖ | Texture/Shape-Texture biased [30] | Augmentations | 75.31 | 3.22 | 89.03$^{\pm 10.05}$ | 41.51$^{\pm 28.97}$ |
| • | Dinov1 [3] | SSL | 75.32 | 2.04 | 87.46$^{\pm 11.91}$ | 40.19$^{\pm 32.33}$ |
| ● | MoCo v3 (100Ep) [5] | SSL | 68.99 | 3.79 | 84.11$^{\pm 12.19}$ | 50.33$^{\pm 27.68}$ |
| Y | MoCo v3 (300Ep) [5] | SSL | 72.84 | 3.44 | 85.34$^{\pm 11.21}$ | 50.55$^{\pm 26.64}$ |
| ♦ | MoCo v3 (1000Ep) [5] | SSL | 74.62 | 2.34 | 85.58$^{\pm 10.73}$ | 49.76$^{\pm 26.13}$ |
| ⅄ | SimCLRv2 [4] | SSL | 74.96 | 3.55 | 87.99$^{\pm 11.96}$ | 41.76$^{\pm 30.60}$ |
| × | SwAV [2] | SSL | 75.33 | 2.49 | 84.07$^{\pm 11.18}$ | 53.38$^{\pm 27.46}$ |
| ◆ | SupCon [26] | SSL | 77.37 | 6.53 | 79.02$^{\pm 6.27}$ | 63.43$^{\pm 13.17}$ |
| • | timm A1 [52, 53] | Improved Training | 80.14 | 8.71 | 84.65$^{\pm 7.91}$ | 57.85$^{\pm 19.18}$ |
| ● | timm A1h [52, 53] | Improved Training | 80.15 | 43.78 | 75.69$^{\pm 6.49}$ | 69.05$^{\pm 13.40}$ |
| Y | timm A2 [52, 53] | Improved Training | 79.86 | 8.77 | 86.74$^{\pm 9.40}$ | 55.11$^{\pm 23.14}$ |
| ♦ | timm A3 [52, 53] | Improved Training | 77.45 | 6.60 | 79.07$^{\pm 6.64}$ | 72.55$^{\pm 11.76}$ |
| ⅄ | timm B1k [52, 53] | Improved Training | 79.25 | 14.44 | 82.72$^{\pm 11.73}$ | 51.60$^{\pm 29.58}$ |
| × | timm B2k [52, 53] | Improved Training | 79.30 | 14.86 | 83.49$^{\pm 12.16}$ | 49.40$^{\pm 30.26}$ |
| ◆ | timm C1 [52, 53] | Improved Training | 79.78 | 22.05 | 83.05$^{\pm 11.71}$ | 47.89$^{\pm 29.53}$ |
| ⬟ | timm C2 [52, 53] | Improved Training | 79.97 | 15.91 | 83.42$^{\pm 12.01}$ | 47.52$^{\pm 29.54}$ |
| ▲ | timm D [52, 53] | Improved Training | 79.95 | 2.97 | 83.47$^{\pm 7.04}$ | 57.26$^{\pm 19.27}$ |
| ◀ | TorchVision 2 [40, 50] | Improved Training | 80.92 | 41.27 | 74.33$^{\pm 7.45}$ | 62.90$^{\pm 18.86}$ |
| ■ | Frozen Random Filters [13] | Freezing | 74.87 | 2.91 | 79.83$^{\pm 7.48}$ | 66.16$^{\pm 17.12}$ |

Table 2. Performance summary for the 56 ResNet-50 models evaluated in our study. For each model, the table lists its unique visual identifier used consistently throughout all figures: color denotes the training Category (*e.g.*, Augmentations), while marker shape identifies the specific model. We report ID metrics (Accuracy, ECE) and OOD metrics (AUROC, FPR95). OOD performance is shown as mean ± standard deviation across all 21 OOD detection methods and eight OOD datasets. Arrows (↑/↓) indicate whether higher or lower values are better, and all values are reported as percentages.

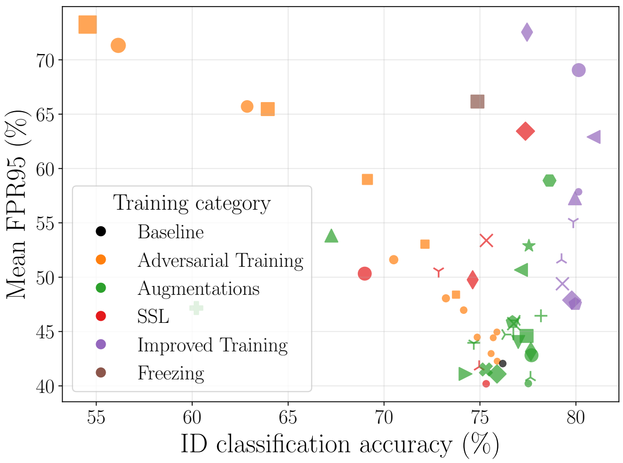


Figure 1. Relationship between ID classification accuracy and OOD detection performance, measured by the mean False Positive Rate at 95% True Positive Rate (FPR95). Each point represents one of 56 ResNet-50 models trained with a diverse strategy. The reported FPR95 for each model is the average across all 21 OOD detection methods and four OOD categories. Color indicates the model's training category, while the marker shape uniquely identifies each model within that category.

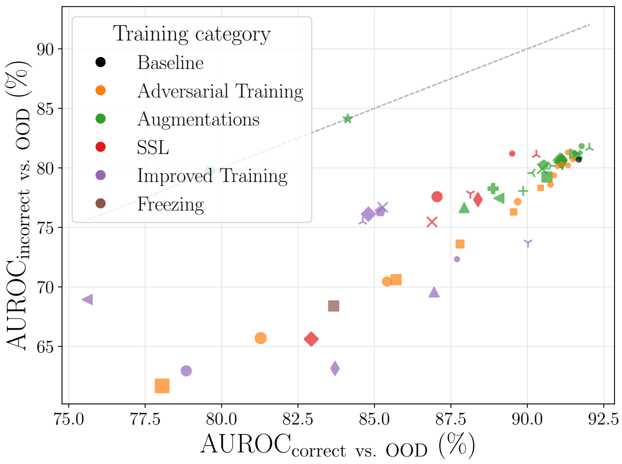


Figure 2. Relationship between AUROC$_{\text{correct vs. OOD}}$ and AUROC$_{\text{incorrect vs. OOD}}$. Each point represents one of 56 models, with performance averaged across all 21 OOD detection methods and four OOD categories. Color indicates the model's training category, while the marker shape uniquely identifies each model within that category.

versely, for high-performing models, advanced augmentations and regularization techniques reverse this relationship, leading to a degradation in OOD performance (an increase in FPR95).

This result provides strong evidence that the complex, non-monotonic relationship between ID accuracy and OOD performance is a general phenomenon, independent of the evaluation metric.

**Are OOD detectors merely identifying misclassified samples?** We revisit the claim that post-hoc detectors succeed largely because they separate correctly classified ID samples from OOD inputs. Fig. 2 confirms the strong positive correlation between OOD performance on correctly versus incorrectly classified ID data (Spearman's $\rho = 0.88$, $p \ll 0.001$). It also makes the consistent performance gap visually apparent, as nearly all points lie below the $x = y$ identity line, showing that performance is systematically higher on correctly classified samples. A notable exception are models trained with MixUp or CutMix, where points for all detectors lie on (or very close to) the identity line, indicating similar OOD performance when conditioning on correct vs. incorrect ID predictions. However, the magnitude of this performance gap is highly method-dependent, as detailed in Fig. 3 and Fig. 4.

Classification-based methods like MSP, which are highly sensitive to classification correctness (*i.e.*, high AUROC$_{\text{correct vs. incorrect}}$), exhibit a large performance drop when evaluating on misclassified samples, since their scores are tightly coupled to prediction confidence, these methods risk confusing hard ID examples with true OOD data. In contrast, methods that leverage richer feature-space representations, like NNGuide and GRAM, show almost no performance gap. Their near-chance failure detection performance (AUROC$_{\text{correct vs. incorrect}} \approx 50\%$) implies their OOD scoring is largely decoupled from the correctness of the ID classification.

In Figs. 5 and 6, we correlate the OOD detection performance (AUROC) with the ID classification accuracy. This analysis is performed for all ID samples, and we further dissect the behavior by also considering the subsets of correctly and incorrectly classified samples separately (Fig. 5). The Spearman correlation coefficients (Fig. 6) reveal a consistently weak or statistically non-significant relationship across all three groups (*i.e.*, all, correct, and incorrect), echoing the main manuscript finding. This result diverges from prior work [24]; while they performed a similar analysis, they observed a strong overall correlation that was almost entirely driven by the performance on correctly classified ID samples (AUROC$_{\text{correct vs. OOD}}$)

While AUROC$_{\text{incorrect vs. OOD}}$ can approach random chance for some model–method pairs, this is not the case for well-matched configurations (Fig. 5). For the baseline model—representing the default benchmark setting where the model-method fit is strong—every single detector performs significantly better than random guess-

| **Effect** | **F-value** | **p-value** | **Variance Share (%)** |
|---|---|---|---|
| Method | 156.69 | $\ll 0.001$ | 7.08 |
| Model | 77.90 | $\ll 0.001$ | 9.69 |
| OOD Category | 5045.22 | $\ll 0.001$ | 34.22 |
| Model × Method | 8.47 | $\ll 0.001$ | 21.05 |
| Model × OOD Category | 9.10 | $\ll 0.001$ | 3.39 |
| Method × OOD Category | 39.05 | $\ll 0.001$ | 5.30 |
| Model × Method × OOD Group | 1.16 | $\ll 0.001$ | 8.66 |
| Residual | — | — | 10.62 |

Table 3. Three-way ANOVA decomposition of AUROC variance across models, OOD detection methods, and OOD dataset categories. The table reports the F-statistic, significance level (p-value), and proportion of explained variance for each main effect and interaction.

ing. This demonstrates a genuine ability to distinguish true OOD samples from a model's own most challenging ID examples, proving that—while misclassifications impair performance—these methods are fundamentally more than mere failure detectors.

**Where does the AUROC variance come from?** Tab. 3 lists the complete F-values, p-values, and variance shares of the three-way ANOVA; all main effects and interactions are significant ($p \ll 0.001$). To rule out a single OOD category artifact, we reran the ANOVA four times, each time omitting one OOD category. Tab. 4 shows the variance shares. Leaving out the hardest split (near-OOD) drops the OOD category main effect to $6.59\%$, but the *model* × *method* interaction increases to $33.36\%$, revealing model-detector coupling that had been masked by uniformly low AUROC on the toughest OOD category. When far- or extreme-OOD is omitted, the OOD-category term remains dominant ($\approx 40\%$) while the interaction never falls below $16\%$. The residual variance is stable across all runs. Thus, no single dataset dictates the conclusions; indeed, model–OOD method compatibility becomes more salient once the most challenging category is removed, underscoring the need for a diverse OOD benchmark.

**How robust are detection methods across training variants?** The robustness of OOD detection methods also depends on the nature of the distributional shift. Fig. 7 shows the OOD detection performance for each method across the four OOD categories, revealing several key insights.

First, as expected, performance is generally lowest for the most challenging near-OOD datasets, where the semantic similarity with the ID data is highest. Most methods struggle to achieve high AUROC scores in this setting, confirming the difficulty of this benchmark.

Second, and more surprisingly, the variance in performance across our 56 models is often highest for the supposedly easier extreme- and synthetic-OOD categories. This suggests that the choice of training strategy can have a more pronounced and unpredictable impact on a method's effectiveness when the domain shift is large but structurally simple (*e.g.*, ImageNet vs. MNIST).

This highlights a critical aspect of robustness: a method that appears stable and effective on near-OOD data may become unreliable on other types of shifts, and vice versa. For example, the high variance of some model enhancement methods on extreme- and synthetic-OOD data may not just stem from a sensitivity to low-level statistics, but also from operating on final-layer features where discriminative information for structurally simple OOD data might be diminished. This hypothesis is supported by prior work [27], which showed that simpler OOD tasks are often more easily solved in a model's earlier layers. The notable robustness of GRAM, which leverages intermediate features, on these same categories lends further support to this idea, suggesting that access to earlier representations is key for handling such shifts. This underscores the necessity of benchmarking on a wide range of OOD test sets to gain a complete picture of a method's generalization capabilities.

**Relationship with Model Calibration** To investigate if other ID metrics are better predictors of OOD performance than accuracy, we analyzed the relationship between Expected Calibration Error (ECE) and the OOD detection performance (Fig. 8). Globally, we observe a weak negative correlation (Spearman's $\rho = -0.17, p \ll 0.001$), which, while more consistent than the correlation with ID classification accuracy ($\rho = 0.04$), remains a poor proxy for OOD detection performance.

A breakdown by training category (Fig. 9) reveals that this global correlation is a misleading artifact. The trend is driven almost entirely by the adversarial training regime ($\rho = -0.33$). At the same time, models trained with augmentations, SSL, or improved recipes show little to no correlation between their calibration and OOD detection performance.

This finding underscores that OOD detection performance is too complex to be reliably predicted by a single ID metric, such as accuracy or calibration. While correlations may appear within specific subgroups (*e.g.*, training strategies or OOD detection methods), such as adversarially trained models, they do not imply causality and fail to generalize across the diverse landscape of training strategies, making them unreliable as universal proxies.

**Feature-Space Analysis and Robustness of OOD Detection Methods.** To better understand why advanced training recipes degrade OOD detection, we analyze feature-space statistics for four ResNet-50 models: the baseline [16], MixUp [58], CutMix [57], and the TorchVision 2 recipe [40, 50] (which includes MixUp, CutMix together with additional regularization such as label smooth-

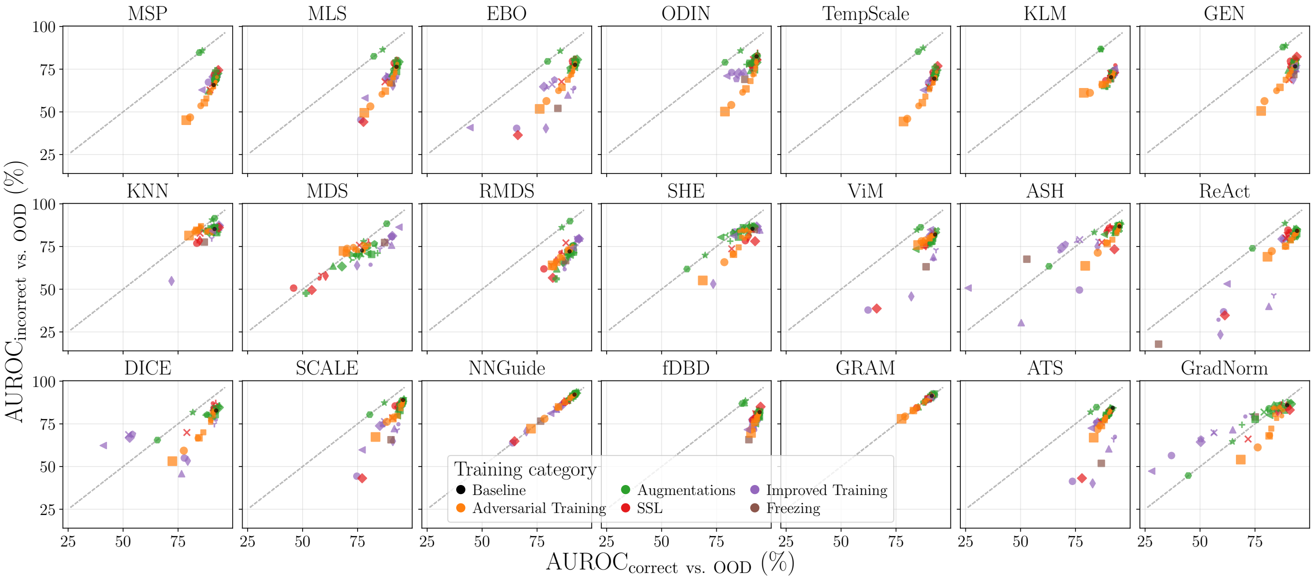


Figure 3. Relation between OOD Detection Performance on correct versus incorrect ID samples for each OOD detection method. Each point represents one of the 56 ResNet-50 models, averaged over eight OOD datasets. Color indicates the model's training category, while the marker shape uniquely identifies each model within that category.

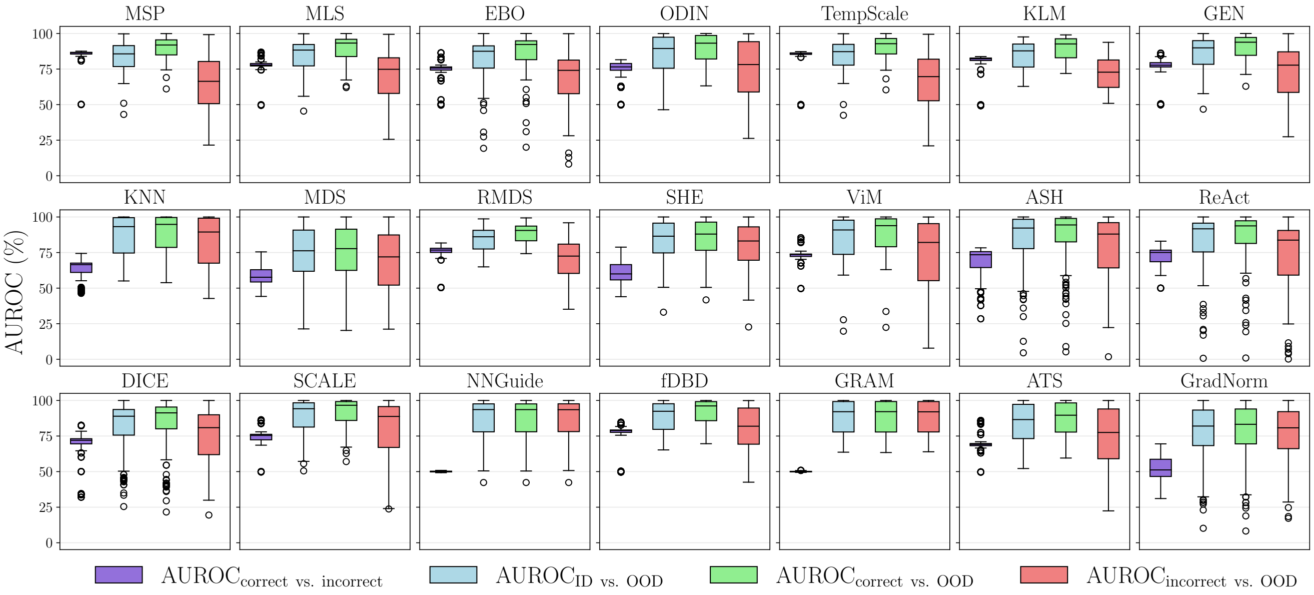


Figure 4. Performance comparison of all 21 OOD detection methods across multiple AUROC-based evaluation metrics. AUROC$_{\text{correct vs. incorrect}}$ evaluates failure prediction on ID data only, distinguishing between correctly and incorrectly classified samples. The remaining metrics assess OOD detection, either across all ID samples, only correctly classified ones, or only misclassified ones. Each boxplot shows the distribution over 56 models and four OOD categories.

ing, stronger augmentation, and EMA; see Fig. 10). We report five complementary metrics: total variance (spread of embeddings), participation ratio (effective dimensionality), sparsity (fraction of near-zero activations), and the mean and standard deviation of feature norms.

While MixUp, CutMix, and TorchVision 2 achieve higher ID accuracy than the baseline, their internal representations become progressively more compressed. From the baseline through MixUp and CutMix to TorchVision 2, we observe a clear progression. MixUp reduces variance

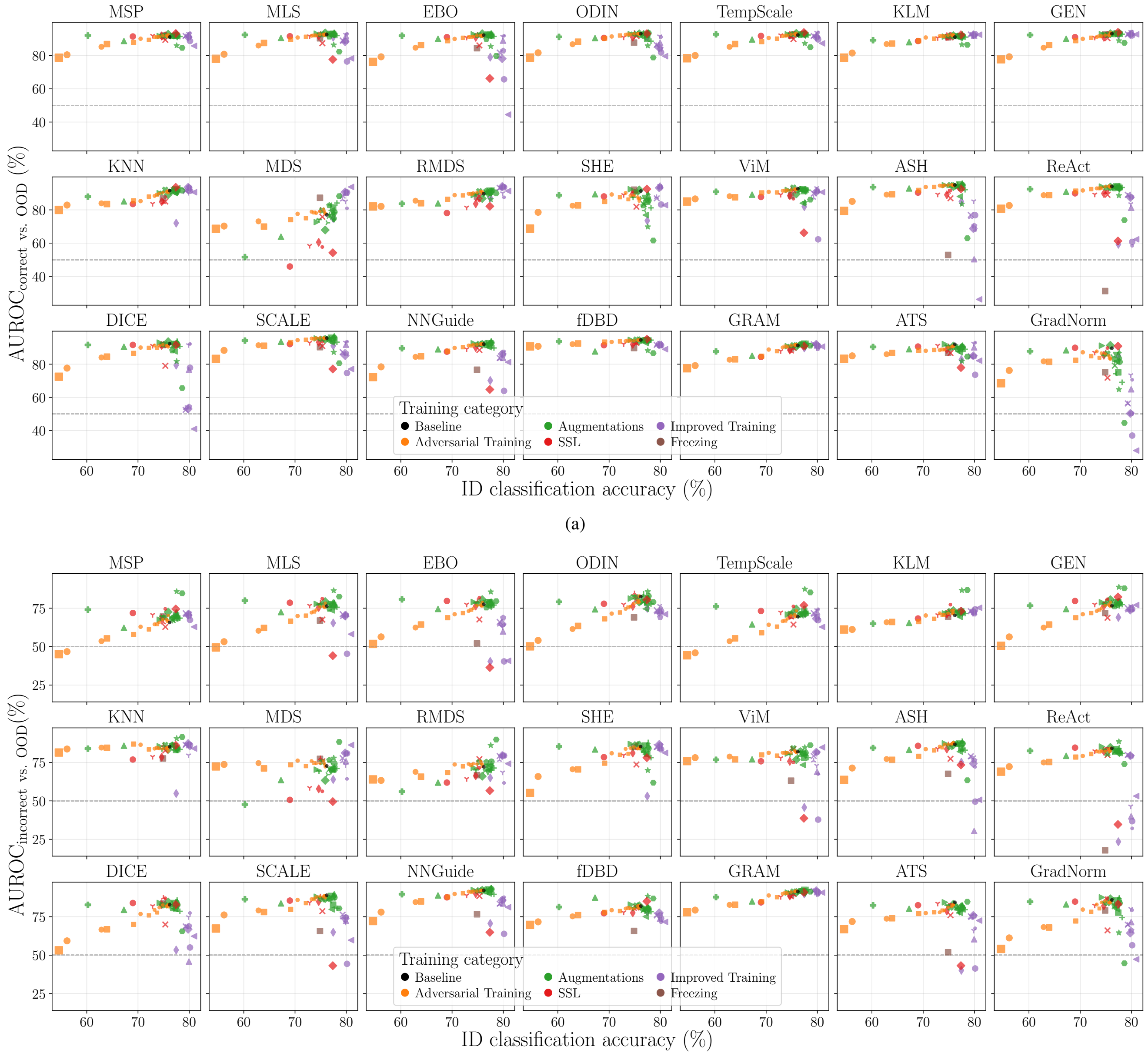
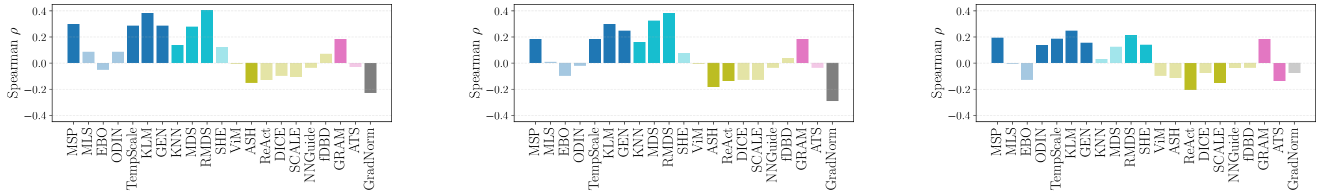


Figure 5. Relation between ID classification accuracy and OOD detection performance. Subfigure (a) shows the AUROC for distinguishing correctly classified ID samples from OOD samples, while (b) focuses on incorrectly classified ID samples. Each point represents one of the 56 ResNet-50 models, averaged over eight OOD datasets. Color indicates the model's training category, while the marker shape uniquely identifies each model within that category.

| Left-Out | Model | Method | OOD Category | Model×Method | Model×OOD Category | Method×OOD Category | 3-Way Interaction | Residual |
|---|---|---|---|---|---|---|---|---|
| Near | 16.66 | 10.02 | 6.78 | 33.87 | 4.33 | 6.41 | 11.43 | 10.55 |
| Far | 9.59 | 5.16 | 42.15 | 16.51 | 3.44 | 6.25 | 8.45 | 8.44 |
| Extreme | 8.41 | 8.62 | 40.23 | 19.79 | 2.17 | 2.60 | 6.46 | 11.67 |
| Synthetic | 11.49 | 8.04 | 31.49 | 22.49 | 2.76 | 4.63 | 6.39 | 12.77 |

Table 4. Explained variance from leave-one-out 3-way ANOVA (factors: model, method, OOD category). Each row excludes one OOD group and recomputes variance proportions. All reported values are in percentage and statistically significant ($p \ll 0.001$).



(a) Correlation between ID classification accuracy and AUROC.

(b) Correlation between ID classification accuracy and $\text{AUROC}_{\text{correct vs. OOD}}$.

(c) Correlation between ID classification accuracy and $\text{AUROC}_{\text{incorrect vs. OOD}}$.

Figure 6. Relationship between ID classification accuracy and OOD detection performance. Spearman rank correlation ($\rho$) between ID classification accuracy and OOD-detection AUROC for each detector: (a) all ID samples, (b) only correctly classified ID samples, and (c) only misclassified ID samples. Bars are sorted and color-coded according to the method's OOD detection category (● classification-based, ● feature-based, ● hybrid, ● intermediate-feature, ● gradients). Non-significant correlations ($p \geq 0.05$) are shown with reduced opacity. Statistics are computed over 56 models and four OOD categories.

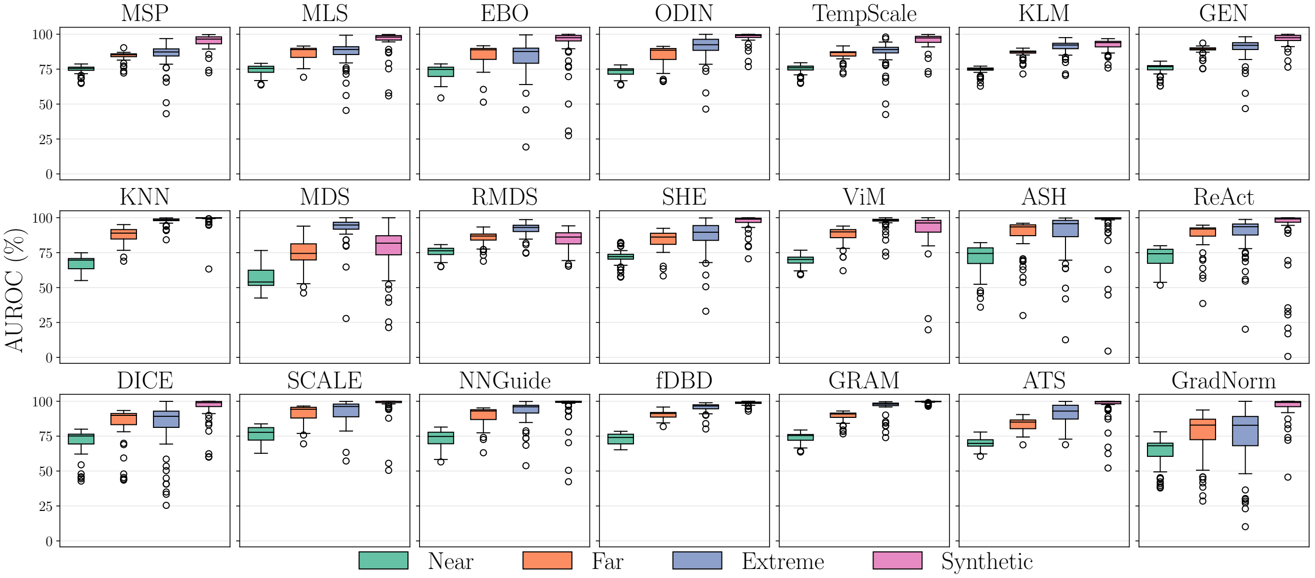


Figure 7. OOD detection performance across different OOD categories (near, far, extreme, and synthetic). Each boxplot shows the distribution over 56 models.

and feature norms while lowering the participation ratio, suggesting a lower-rank embedding. CutMix shows similar but slightly stronger effects, with variance/norm reduced further and sparsity moderately increased. TorchVision 2 amplifies these trends: variance and norms collapse, sparsity increases by more than two orders of magnitude, and the representation is flattened. Thus, while all three advanced recipes achieve higher ID accuracy than the baseline, they also progressively compress and sparsify the embedding space.

These shifts are also mirrored in the logit and embedding space (see Figs. 11 and 12). The max-logit distributions become narrower and show increasing ID-OOD overlap: baseline leaves a clear margin (FPR95 = 30.62%), MixUp reduces separation (57.70%), CutMix worsens it further (70.93%), and TorchVision 2 nearly eliminates it (77.52%). Likewise, the penultimate-layer activation distributions show that the characteristic pattern described by Sun *et al.* [47]—a near-constant mean activation for ID samples and lower but more variable activations for OOD samples, which ReAct exploits via activation clipping—progressively changes under MixUp, CutMix, and TorchVision 2. As a result, activation-shaping detectors such as ReAct—whose efficacy depends on clipping high activations—lose discriminative power, reflected in a significant performance drop: FPR95 increases from 16.75% (baseline) to 40.46% (MixUp), 58.51% (CutMix), and 88.55% (TorchVision 2).

In contrast, feature-based methods (*e.g.*, KNN, GRAM, RMD) that leverage distances or higher-order statistics rather than specific activation characteristics, and therefore remain comparatively robust under increasing regulariza-

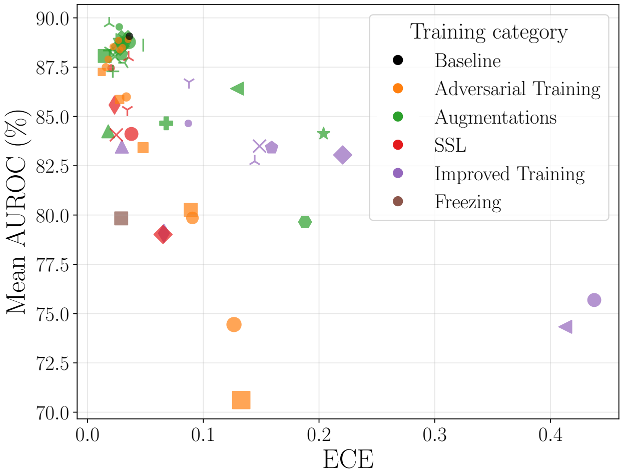


Figure 8. Relationship between the expected calibration error (ECE) and the OOD detection performance. Each point represents one of 56 models, with performance averaged across all 21 OOD detection methods and eight OOD datasets.

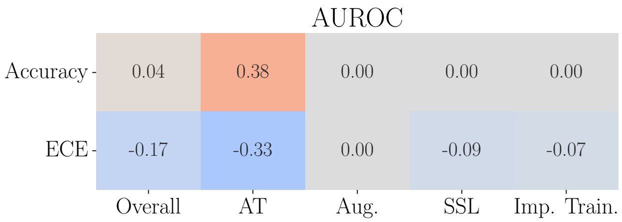


Figure 9. Spearman correlation coefficients ($\rho$) between OOD detection performance (AUROC) and in-distribution (ID) performance metrics (accuracy and ECE), computed across all models, OOD methods, and OOD categories (*Overall*), and separately for adversarial training (*AT*), data augmentations (*Aug.*), self-supervised learning (*SSL*), and improved training recipes (*Imp. Train.*). Non-significant correlations ($p \geq 0.05$) are set to 0. Note that Spearman $r$ reflects monotonic relationships and may not capture non-monotonic trends.

tion. Altogether, these results show that although MixUp, CutMix, and TorchVision 2 improve ID accuracy, they also systematically reshape the feature space in ways that disadvantage activation-based detectors while leaving geometry-based or magnitude-agnostic approaches more stable. This provides further evidence for our central finding that improvements in ID accuracy do not necessarily yield better OOD detection, underscoring the strong dependency between the underlying model and the effectiveness of a given OOD detection method.

## C. Results on ViT

To test whether our findings extend beyond ResNets, we also evaluate Vision Transformer (ViT-B/16) mod-

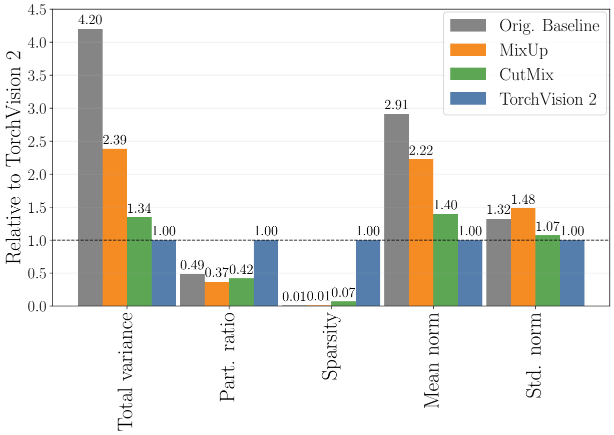


Figure 10. Feature-space metrics for ResNet-50 baseline, MixUp, CutMix, and TorchVision 2 on the ImageNet test set, computed from penultimate-layer embeddings and shown relative to TorchVision 2 (set to 1.0). We report five complementary statistics: i) total variance, the trace of the covariance matrix measuring overall spread of embeddings; ii) participation ratio, the effective dimensionality of the feature space; iii) sparsity, the fraction of activations below $10^{-3}$; iv) mean feature norm, the average $l_2$-norm of embedding vectors; and v) standard deviation of feature norms, capturing variability in embedding magnitudes.

els [10], that originate from AugReg [45], Masked Autoencoders (MAE) [17], Data-Efficient Image Transformers (DeiT) [49], and Sharpness-Aware Minimization (SAM) [6]. As with ResNet, all models are trained exclusively on the ILSVRC2012 subset of ImageNet to prevent OOD contamination.

Consistent with our ResNet results, ViTs achieve higher ID accuracy but do not exhibit improved OOD detection performance (see Fig. 13). At the OOD detection method level (Fig. 14), we again observe a clear dichotomy: feature-based methods that rely on distances or higher-order statistics (*e.g.*, KNN, RMDS, GRAM) remain comparatively robust, while model-enhancement methods that depend on shaping specific activation patterns degrade substantially.

These findings reinforce our central claim that better ID accuracy does not guarantee better OOD detection, even for more modern, higher-capacity architectures. They also support recent evidence [61] that many OOD detection methods have been implicitly tuned to CNN-style representations, and may overfit to the activation characteristics of ResNets rather than transfer robustly to other architectures.

## References

[1] Julian Bitterwolf, Alexander Meinke, and Matthias Hein. Certifiably Adversarially Robust Detection of Out-of-Distribution Data. In *NeurIPS*, 2020. 1

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Pi-

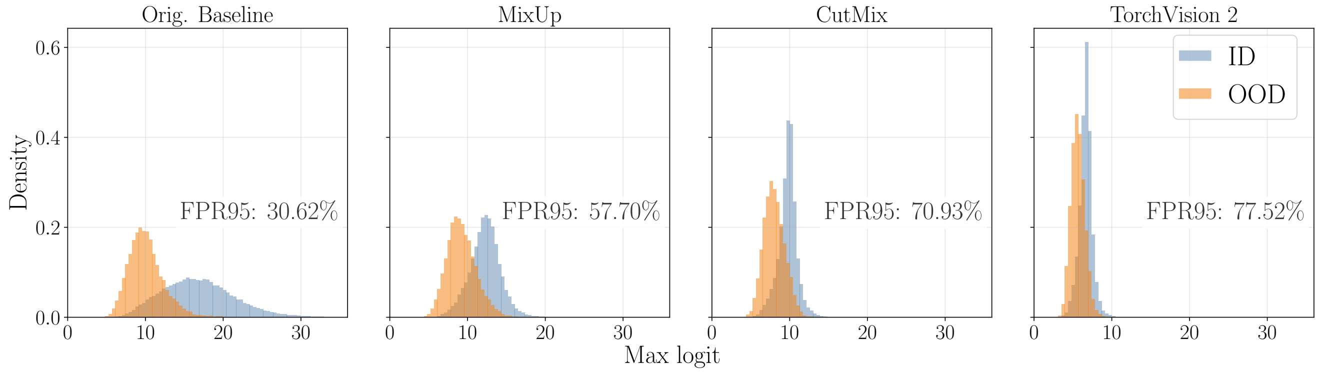


Figure 11. Max-logit distributions for ResNet-50 baseline, MixUp, CutMix, and TorchVision 2 with ImageNet (ID) and iNaturalist (OOD). The plots show the distribution of the maximum predicted logit for ID and OOD samples, together with the corresponding FPR95 values.

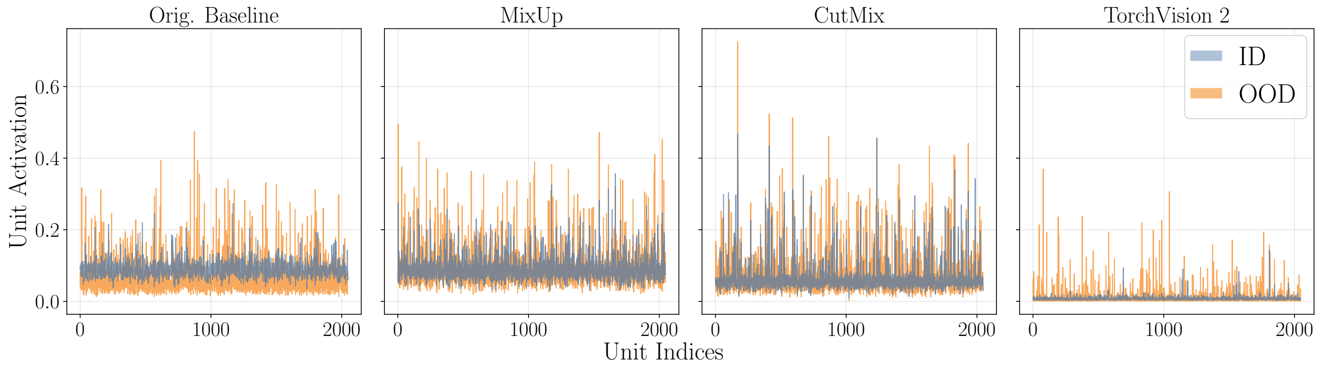


Figure 12. Distribution of per-unit activations in the penultimate layer for ImageNet (ID) and iNaturalist (OOD) across ResNet-50 baseline, MixUp, CutMix, and TorchVision 2.

otr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. ICCV*, 2021. 2

[4] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020. 2

[5] Xinlei Chen, Saining Xie, and Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers . In *Proc. ICCV*, 2021. 2

[6] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations. In *Proc. ICLR*, 2022. 8

[7] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *Proc. CVPR*, 2019. 2

[8] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In *NeurIPS*, 2020. 2

[9] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely Simple Activation Shaping for Out-of-Distribution Detection. In *Proc. ICLR*, 2023. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proc. ICLR*, 2021. 8

[11] Benjamin Erichson, Soon Hoe Lim, Winnie Xu, Francisco Utrera, Ziang Cao, and Michael Mahoney. NoisyMix: Boosting model robustness to common corruptions. In *Proc. AISTATS*, 2024. 2

[12] Paul Gavrikov and Janis Keuper. Can biases in imagenet models explain generalization? In *Proc. CVPR*, 2024. 1

[13] Paul Gavrikov and Janis Keuper. The Power of Linear Combinations: Learning with Random Convolutions, 2024. 2

[14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proc. ICLR*, 2019. 2

[15] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger.

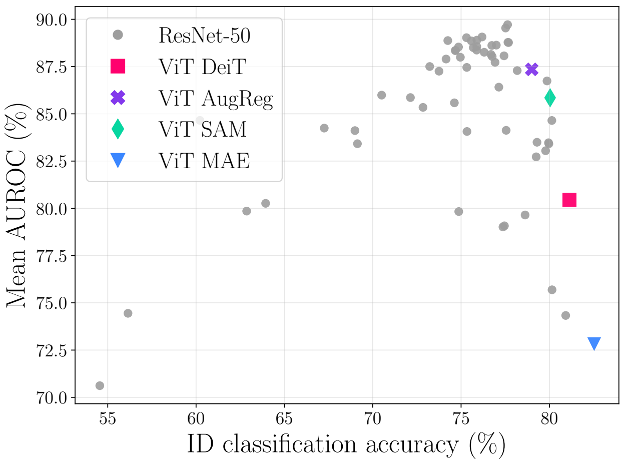


Figure 13. Relationship between ID accuracy and OOD detection performance (AUROC) for 56 ResNet-50 and four ViT-B/16 models. Each point corresponds to a specific training strategy on ImageNet (ID), with OOD performance averaged over 21 detection methods and eight OOD datasets.

On Calibration of Modern Neural Networks. In *Proc. ICML*, 2017. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proc. CVPR*, 2016. 2, 4

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proc. CVPR*, 2022. 8

[18] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *Proc. ICLR*, 2017. 1

[19] Dan Hendrycks*, Norman Mu*, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple method to improve robustness and uncertainty under data shift. In *Proc. ICLR*, 2020. 2

[20] 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. *Proc. ICCV*, 2021. 2

[21] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling Out-of-Distribution Detection for Real-World Settings. In *Proc. ICML*, 2022. 1

[22] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. *Proc. CVPR*, 2022. 2

[23] Rui Huang, Andrew Geng, and Yixuan Li. On the Importance of Gradients for Detecting Distributional Shifts in the Wild. In *NeurIPS*, 2021. 1

[24] Galadrielle Humblot-Renaux, Sergio Escalera, and Thomas B. Moeslund. A noisy elephant in the room: Is your out-of-distribution detector robust to label noise? In *Proc. CVPR*, 2024. 1, 3

[25] Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *Proc. ICLR*, 2024. 2

[26] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 2

[27] Gerhard Krumpl, Henning Avenhaus, Horst Possegger, and Horst Bischof. ATS: Adaptive Temperature Scaling for Enhancing Out-of-Distribution Detection Methods. In *Proc. WACV*, 2024. 1, 4

[28] Yann Lecun, Lé'on Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based Learning Applied to Document Recognition. *IEEE*, 86(11):2278–2324, 1998. 1

[29] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *NeurIPS*, 2018. 1

[30] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and cihang xie. Shape-texture debiased neural network training. In *Proc. ICLR*, 2021. 2

[31] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *Proc. ICLR*, 2018. 1

[32] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. In *NeurIPS*, 2019. 2

[33] Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. *ICML*, 2024. 1

[34] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based Out-of-distribution Detection. In *NeurIPS*, 2020. 1

[35] Xixi Liu, Yaroslava Lochman, and Zach Christopher. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proc. CVPR*, 2023. 1

[36] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proc. ICLR*, 2018. 2

[37] Apostolos Modas, Rahul Rade, Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Prime: A few primitives can boost robustness to common corruptions. In *Proc. ECCV*, 2022. 2

[38] Patrick Müller, Alexander Braun, and Margret Keuper. Classification robustness to common optical aberrations. In *Proc. ICCV Workshops*, 2023. 2

[39] Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *Proc. ICCV*, 2023. 1

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, 2019. 2, 4

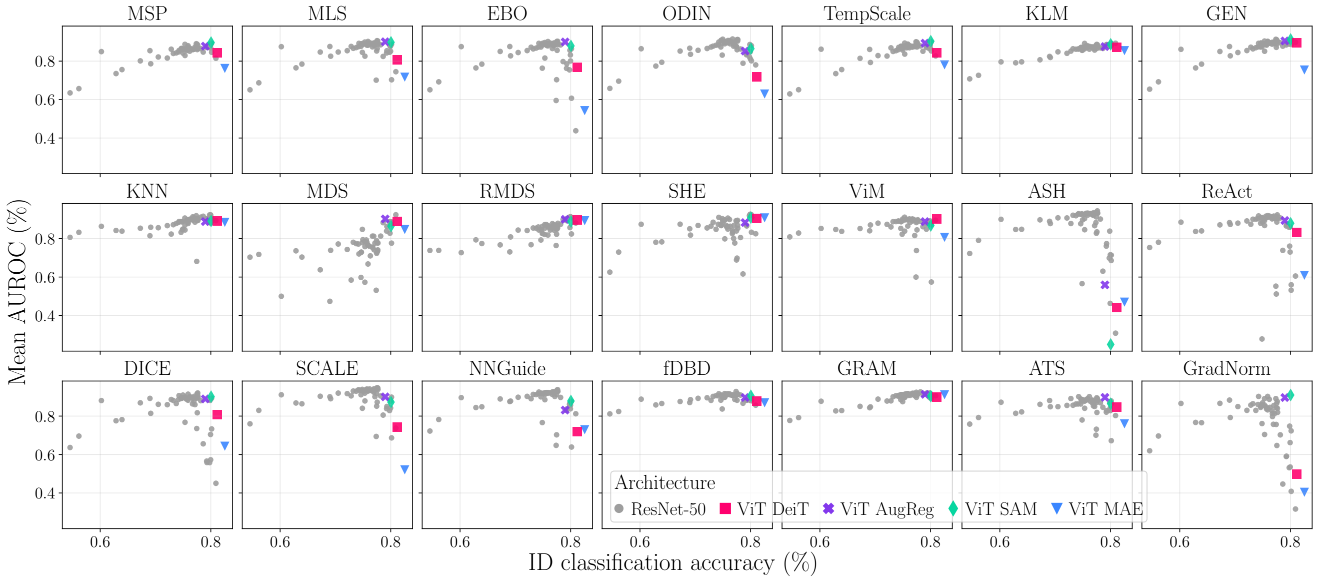


Figure 14. Relationship between ID accuracy and OOD detection performance (AUROC) for ResNet-50 and ViT-B/16 across individual OOD detection methods. Each panel shows one OOD detection method, with points corresponding to different training strategies (21 ResNet-50 and four ViT-B/16 models trained on ImageNet). OOD performance is averaged across eight OOD datasets.

[41] Francesco Pinto, Harry Yang, Ser-Nam Lim, Philip Torr, and Puneet K. Dokania. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. In *NeurIPS*, 2022. 2

[42] Jie Jessie Ren, Stanislav Fort, Jeremiah Zhe Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection. *arXiv preprint arXiv:2106.09022*, 2021. 1

[43] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do Adversarially Robust ImageNet Models Transfer Better? In *NeurIPS*, 2020. 2

[44] Chandramouli Shama Sastry and Sageev Oore. Detecting Out-of-Distribution Examples with Gram Matrices. In *Proc. ICML*, 2020. 1

[45] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. In *TMLR*, 2022. 8

[46] Yiyou Sun and Yixuan Li. DICE: Leveraging Sparsification for Out-of-Distribution Detection. In *Proc. ECCV*, 2022. 1

[47] Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution Detection With Rectified Activations. In *NeurIPS*, 2021. 1, 7

[48] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution Detection with Deep Nearest Neighbors. In *Proc. ICML*, 2022. 1

[49] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers amp; distillation through attention. In *Proc. ICML*, 2021. 8

[50] Vasilis Vryniotis. How to Train State-Of-The-Art Models Using TorchVision's Latest Primitives, 2023. https://pytorch.org/blog/how-to-train-state-of-the-art-models-using-torchvision-latest-primitives/ [Accessed: 2025-07-01]. 2, 4

[51] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-Of-Distribution with Virtual-logit Matching. In *Proc. CVPR*, 2022. 1

[52] Ross Wightman. Pytorch image models. https://github.com/huggingface/pytorch-image-models, 2019. 2

[53] Ross Wightman, Hugo Touvron, and Herve Jegou. ResNet strikes back: An improved training procedure in timm. In *NeurIPS Workshops*, 2021. 2

[54] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 1

[55] Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for Training Time and Post-hoc Out-of-distribution Detection Enhancement. In *Proc. ICLR*, 2024. 1

[56] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WENXUAN PENG, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking Generalized Out-of-Distribution Detection. In *NeurIPS Datasets and Benchmarks Track*, 2022. 1

[57] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features . In *Proc. ICCV*, 2019. 2, 4

[58] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and

David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. ICLR*, 2018. 2, 4

[59] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Xiaoguang Liu, Shi Han, and Dongmei Zhang. Out-of-Distribution Detection based on In-Distribution Data Patterns Memorization with Modern Hopfield Energy. In *Proc. ICLR*, 2023. 1

[60] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023. 1

[61] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. *J. of DMLR*, 2024. 1, 2, 8