WHICH OOD DETECTOR SHOULD YOU DEPLOY WHEN TEST-SET GAINS DISAGREE?

Anonymous authors
Paper under double-blind review

ABSTRACT

A deployed out-of-distribution (OOD) detector meets unfamiliar inputs whose mix is unknown and changing. Its effectiveness is highly sensitive to this mix. Yet on benchmarks, detectors that gain on some test sets perform poorly on others, sometimes below a vanilla cross-entropy baseline. How should one choose a detector whose performance holds across the mixes it may meet? We find that these losses are reproducible effects of training. Predictions fixed before training hold across seeds, scoring functions, architectures and datasets, with and without auxiliary OOD data. We compare a new detector with every existing choice, including the baseline under each scoring function, and compute in closed form how far the mix can shift before it stops being the top performer. At the published means of four benchmarks, the baseline wins under this comparison for about half of the training gains measured under a shared scoring function. None of those gains holds under every mix. A detector that passed this comparison held an 8.8-point lead in area under the ROC curve on new images and halved false acceptances. Gains that failed it fell short on new runs and images. OOD detectors should therefore be chosen by their stability over plausible mixes.

1 INTRODUCTION

An out-of-distribution (OOD) detector flags inputs outside a classifier's training distribution (Hendrycks & Gimpel, 2017; Liu et al., 2020; Sun et al., 2022; Yang et al., 2024). Once deployed, it meets unfamiliar inputs whose mix is unknown and changes over time. Its effectiveness is highly sensitive to this mix. Benchmarks evaluate detectors on several OOD test sets (Zhang et al., 2024). Yet detectors that gain on some test sets perform poorly on others, sometimes below a vanilla cross-entropy baseline, and rankings vary across datasets, seeds and splits (Tajwar et al., 2021; Szyc et al., 2023).

How should one choose a detector whose performance holds across the mixes it may meet?

Consider Outlier Exposure (OE), which trains a classifier to assign low confidence to auxiliary OOD images (Hendrycks et al., 2019a). With CIFAR-100 as the in-distribution (ID) dataset and ReAct as the activation-clipping scorer, OE improves detection of TinyImageNet by 17.0 points in area under the receiver operating characteristic curve (AUROC). On CIFAR-10, it falls 9.4 points below the same scorer on a model trained with cross-entropy alone, whereas the near-OOD average over the two test sets nearest the ID data favors OE by 3.8 points. The auxiliary pool and the TinyImageNet test images both come from ImageNet, with disjoint classes. Figure 1 shows the same opposing changes under all seven scorers.

We find that these losses are reproducible effects of training. Predictions fixed before training are confirmed on ten new CIFAR-100 ResNet-18 seeds, recur with WideResNet-40-2, and extend to an iNaturalist loss on five ImageNet-200 seeds. LogitNorm (Wei et al., 2022), which uses no auxiliary OOD data, shows the same CIFAR-10 loss on new runs, at a cost of 2.4 points of ID accuracy. OE improves both near sources in three of four CIFAR-10 comparisons, and both sources of our exploratory text experiment.

We propose to compare a new detector with every existing choice, including the baseline under each scorer, and to compute how far the mix of OOD sources can shift before the choice stops

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.

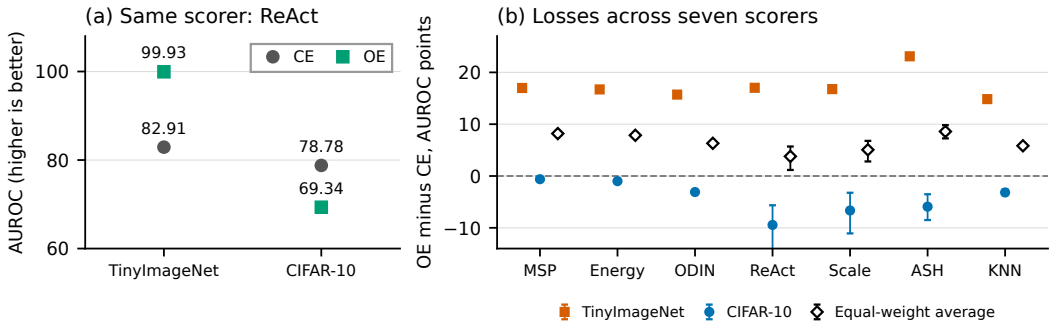


Figure 1: **Outlier Exposure (OE) improves detection on TinyImageNet but falls below cross-entropy on CIFAR-10.** CIFAR-100 is the in-distribution dataset, with ResNet-18 trained for 200 epochs and ten new seeds per arm. (a) Area under the receiver operating characteristic curve (AUROC) for the ReAct scorer after cross-entropy (CE) training or OE with TinyImageNet-597 auxiliaries. (b) OE minus CE for each out-of-distribution (OOD) source and their equal-weight average, across seven scorers, with pointwise 95% paired seed-bootstrap intervals. Scorers include maximum softmax probability (MSP), activation shaping (ASH) and nearest-neighbour scoring (KNN). All seven source losses survive Holm correction under paired and Welch tests (Table 13).

being the top performer. Our central insight is that a test-set gain matters only where it changes the best available detector. Because existing detectors remain available, a new one adds value only at mixes where it beats all of them. For fixed detectors and ID data, AUROC is linear in the source weights, so these mixes and that distance follow exactly from per-test-set results, as in weighted-score optimization (Roijers et al., 2015; Mazurek & Strzałka, 2022). A new detector should replace the current choice when three conditions hold (Definition 1). It must (i) beat every existing choice at a stated reference mixture, (ii) keep its lead under the shifts specified in advance, and (iii) on separate images, meet a requirement on AUROC gain and ID accuracy fixed before that evaluation. Beyond the variation across individual OOD test sets, this measure lets one reason about how stable a detector's advantage is under any mix of them.

Contributions.

1. **We show that test-set conflicts are reproducible effects of training.** Losses predicted before training recur for OE on ten new seeds under seven scorers, with a second architecture and on ImageNet-200, and for LogitNorm, selected from archived runs, without auxiliary OOD data (Section 4.1, Appendix D.6).
2. **We show that a gain matters only where it changes the best available detector, and locate those mixes in closed form.** AUROC is linear in the mix, so per-test-set results determine exactly where a new detector beats every existing choice and how far that lead extends (Definition 1). At the published means of four benchmarks, only 26 of 178 candidates can lead at some mix of sources. Of the 39 training gains, 19 survive comparison with the baseline under its best scorer, and none holds under every mix (Section 4.3).
3. **We test the definition on new data, where it separates gains that hold from gains that fall short.** OE with CIFAR-10 as the ID dataset beats every existing choice on every original test set, and in a read-out of saved outputs it leads by 8.8 points on new images and halves false acceptances. On ImageNet-200, OE's lead of 0.44 points over the best original misses the requirement fixed before five new runs, and a blend of auxiliary pools misses it on new images (Sections 4.1 and 4.4).

2 Comparing Detector Choices Across Source Mixtures

A detector here is a trained classifier together with a scorer and the scorer's settings. When detectors perform better on different sources, their ranking depends on the source mixture. We compare them by varying the source weights while holding the trained models, scorers, settings and ID population

fixed. Let a_{dk} denote detector d's AUROC on source k. Conditioning on the OOD source gives

$$A_d(w) = \sum_k w_k a_{dk}, \qquad w \in \mathcal{S}_K, \tag{1}$$

where $\mathcal{S}_K$ contains the nonnegative weights summing to one (Appendix A). The identity also holds when a_{dk} averages AUROC across training runs. Our estimated boundaries use these mean profiles. They need not equal a single checkpoint's boundary or the mean of per-run boundaries.

Value of an additional choice. Let $\mathcal{B}$ contain the original choices, the detectors existing before an intervention, and $\mathcal{N}$ the new candidates it produces. Define their margin and the gain from expanding the choice set by

$$M(w) = \max_{n \in \mathcal{N}} A_n(w) - \max_{b \in \mathcal{B}} A_b(w), \qquad V(w) = \max\{0, M(w)\}. \tag{2}$$

A positive $M(w)$ means that a new candidate beats every original choice at that mixture. If originals remain available, a negative margin gives zero gain $V(w)$. ID accuracy and application costs remain separate outcomes. Maximizing the margin over allowed weights gives a retrospective estimate of the intervention's best gain in that source family. Linear programs compute the point optimum, and a weight grid supports resampling (Appendix G).

Ranking and relative-deficit boundaries. For a reference leader d_0 at w_0, measure source-weight changes by total variation (TV), $\mathrm{TV}(w, w_0) = \frac{1}{2}\sum_k |w_k - w_{0k}|$. For an allowed set $\mathcal{W} \subseteq \mathcal{S}_K$, the boundary is

$$r_\delta(w_0; d_0) = \inf\left\{ \mathrm{TV}(w, w_0) : w \in \mathcal{W},\ \max_{j \neq d_0} A_j(w) - A_{d_0}(w) \geq \delta \right\}. \tag{3}$$

The comparator set is stated for each analysis. Here r_0, the tie radius, is the distance to a tie or loss, and r_δ for $\delta > 0$ is the distance to a relative deficit. Unattainable boundaries have infimum $+\infty$ and are reported separately. AUROC uses the 0 to 100 scale. One point is a reporting margin whose practical importance depends on the application. A relative deficit does not measure the detector's absolute decline from its reference performance.

For two sources, let Δ_k be a detector's AUROC minus its comparator's. Opposite signs give the crossover

$$\Delta(w) = w\Delta_1 + (1-w)\Delta_2, \qquad w^* = -\Delta_2/(\Delta_1 - \Delta_2). \tag{4}$$

The robustness interval contains weights where $\Delta(w) \geq 0$. A crossover coordinate of 0.716 lies a TV distance of 0.216 from equal weights. Appendix A also defines mixture-admissible candidates, which can lead some mixture, and regret relative to the best candidate.

Definition 1 (When a gain changes detector choice)**.** An intervention changes detector choice at a reference mixture w_0 when three conditions hold. First, a new candidate has a positive margin $M(w_0)$ against every original choice, with each original model compared under every scorer and every recorded selection of scorer settings. Second, the margin survives the source-weight changes a practitioner is prepared to defend: the new leader's tie boundary $r_0(w_0)$ exceeds an allowed shift stated before the boundary is computed. Third, on separate images the gain meets a requirement on AUROC gain and ID-accuracy cost fixed before that evaluation, with the operating point, the false-positive rate at a fixed ID-acceptance threshold, reported beside it.

Where no shift was stated in advance, we report the boundary and leave the second condition open. The first two conditions use the measured sources, and the third tests transfer. Section 4 scores each intervention against them.

Relation to established tools. The optimization is standard: the mixture-admissible set is the convex coverage set of Roijers et al. (2015), and the boundary is a weight-stability radius (Mazurek & Strzałka, 2022). Equation (1) is an identity for AUROC at fixed detectors rather than an assumed linear utility, so the boundaries are properties of the measured table. The radius is measured from a stated reference and separates a tie from a stated deficit. Uncertainty is reported with the leader fixed and reselected, and the three conditions above join AUROC with ID accuracy and report the operating point beside them.

Scope and operating points. These comparisons hold scoring rules and source distributions fixed. The false-positive rate at 95% ID acceptance (FPR@95) is also linear when its threshold is fixed on ID data. OpenOOD's published FPR@95 instead uses source-specific thresholds for 95% OOD detection. Averaging those columns need not give the mixture's rate (Appendix A), and our operating-point claims use a threshold fixed on ID data.

3 Experimental Design

We use matched retraining to test whether source losses recur, then change the auxiliary pool and scorer settings to examine their sensitivity. Published tables extend the comparison to a larger set of detectors. We then test a blend of the two pools on separate images against the original choices.

Networks. A network maps an input to logits $\mathbf{z} \in \mathbb{R}^C$, one for each of C classes.

3.1 Training Objectives

Exposure and controls. Each CIFAR auxiliary training configuration (arm) replaces a fraction $\rho_{\text{aux}} = 0.2$ of ID rows per epoch with auxiliary images and keeps the original epoch length. Matched arms share the data schedule set by each random seed (Appendix C). In a batch of size B, ID labels satisfy $y_i \geq 0$ and the B_{aux} auxiliary rows have $y_j = -1$. Vanilla (VAN) trains only on ID images with cross-entropy (CE). The mixing-only control (MIX) uses auxiliary replacement but applies no loss to those rows. OE encourages equal class probabilities on auxiliary images by adding the uniform-target cross-entropy $-C^{-1} \sum_{c=1}^{C} \log \operatorname{softmax}(\mathbf{z}_j)_c$ on each auxiliary row with weight 0.5. Row losses are summed and divided by B.

3.2 Experimental Protocol

Scorers. Maximum softmax probability (MSP) and Energy use the logits (Hendrycks & Gimpel, 2017; Liu et al., 2020). ODIN rescales logits by a temperature and perturbs the input (Liang et al., 2018). ReAct, activation shaping (ASH) and Scale truncate or rescale activations before the final classifier, then compute an energy score (Sun et al., 2021; Djurisic et al., 2023; Xu et al., 2024). The nearest-neighbour scorer (KNN) uses the distance to the k-th nearest ℓ_2-normalized ID training feature (Sun et al., 2022). Tunable settings are selected per run.

Training and evaluation. The CIFAR experiments train ResNet-18 from scratch for 200 epochs on CIFAR-10 or CIFAR-100 and evaluate on OpenOOD v1.5's near-OOD and far-OOD splits (Krizhevsky, 2009; He et al., 2016; Zhang et al., 2024). We use ten seeds per arm. Near-OOD AUROC averages the near pair, the other CIFAR dataset and TinyImageNet, or SSB-hard (Vaze et al., 2022) and NINCO (Bitterwolf et al., 2023) for ImageNet-200. Auxiliary-objective comparisons match trainer, budget, seed indices, data schedule, pool and exposure. Clusters A, B and C (the third cluster) are computing environments with different machines and software, using the same evaluation sets. Appendix C gives their allocation and the separate CIFAR-100 confirmation seeds. ImageNet-200 uses OpenOOD's 90-epoch recipe at 224×224, five seeds per arm, and ImageNet-800 batches paired with ID batches.

Pool and scorer interventions. The default auxiliary pool is TinyImageNet-597, comprising 597 ImageNet classes disjoint from the test categories (Zhang et al., 2024). ImageNet also supplies the scorer-selection data and half of the near-OOD average. OpenOOD's authors link OE's near-OOD gain to this similarity. On the third cluster we retrain OE with a size-matched subsample of 300K Random Images (Hendrycks et al., 2019a), then repeat it with a disjoint draw. The replacement comes from the 80 Million Tiny Images collection with CIFAR-class images removed (Torralba et al., 2008). That collection was withdrawn for offensive content (Torralba et al., 2020) (see the ethics statement). Both CIFAR datasets originated in that collection (Krizhevsky, 2009). We also train OE on four blends of the two pools, with TinyImageNet-597 shares of 25, 50 and 75 percent and the union of both lists, ten seeds per arm with either CIFAR dataset as ID (Appendix C.1). Default scorer settings maximize validation AUROC between ID and TinyImageNet, over the grids in Appendix D. On saved trained models (checkpoints) we also hold settings constant or re-select

ReAct, ASH, Scale and KNN on Textures using the same grids. The TinyImageNet-selected and Textures-selected settings form the two settings regimes compared below. ODIN retains its original settings because its input perturbations cannot be recomputed from saved features. MSP and Energy have no tunable setting. Textures is excluded from evaluation targets whenever used for selection.

Metrics and uncertainty. Higher AUROC is better. For the saved-feature CIFAR read-out, the false-positive rate at 95% ID acceptance (FPR@95) is the percentage of OOD inputs accepted as ID at the fifth percentile of ID-test confidence. Lower is better. Counts of gains and losses use a two-standard-error threshold. We also control false positives across each family of seven per-set comparisons with Holm correction at 5%, using paired tests on matched seeds and unpaired Welch tests. The seven scorers share checkpoints and test images, so they are not independent replicates. Figure intervals resample paired seed differences and cover each comparison separately. Appendix C distinguishes registered comparisons, whose rules were fixed and committed before their runs, from exploratory read-outs.

Selection exercise. We retrospectively compare selecting by the near-OOD average and by the worst near-OOD set. Candidates combine training objectives and scorers (Appendix D.3). Selection regret is the best candidate's AUROC minus the selected candidate's on the mean over MNIST handwritten digits, Street View House Numbers (SVHN) and Places365, withheld from selection.

Exploratory decision analysis. The expanded-choice and reference-sensitivity analyses use saved evaluations. Original CIFAR-10 choices combine vanilla and the two pure-pool OE configurations with seven scorers and both settings regimes. CIFAR-100 adds the two mixing-only controls. New candidates combine the four blends with seven default scorers. The five-source extension excludes Textures.

We resample complete seed records jointly across sources and candidates, using $5{,}000$ replicates for the expanded-choice analyses and $10{,}000$ for direct pool-change contrasts. Ranges use the 2.5th and 97.5th percentiles and are conditional on the measured evaluation images. Appendix G gives the candidate counts, optimization and noise assumptions.

Prespecified evaluations. The subsequent study freezes the blend choice and one test mixture before evaluating separate images (Section 4.4). Its primary uncertainty resamples images within source and category strata, sharing indices across candidates and checkpoints. A separate sensitivity analysis also resamples training runs. We use $5{,}000$ replicates and an intersection-union rule, under which every condition must hold. Every original must be beaten by more than 0.5 AUROC points, and the ID-accuracy cost must remain below 0.5 percentage points at their one-sided 95% bounds. Thresholds are calibrated on separate ID images and remain fixed for OOD evaluation. No new training or scorer selection uses these data. A second protocol, committed before its runs, applies the same requirement to OE with MSP against vanilla under every scorer on five further ImageNet-200 runs per arm, with Welch bounds (Appendix C.4).

4 Results

4.1 Source Losses Recur Across Retraining and Architectures

After matched retraining on CIFAR-100, OE remains below vanilla cross-entropy on CIFAR-10 across all seven scorers on ten new ResNet-18 seeds (Figure 1). Each scorer gains on TinyImageNet and the near-OOD average. Table 1 summarizes the replications and their ranking boundaries, and Figure 2 sets our matched differences beside the published training-method rows. Both directions survive Holm correction across scorers under paired and Welch tests. With WideResNet-40-2, all seven point estimates retain the pattern. Six source losses survive both corrected tests, with KNN the exception. Predictions fixed before training also hold for LogitNorm (Wei et al., 2022), which uses no auxiliary images. On five new runs per arm under MSP, it raises MNIST AUROC over cross-entropy by 11.1 points and lowers CIFAR-10 by 4.2 without overlap between the arms, at a cost of 2.4 points of ID accuracy. With CIFAR-10 as ID data it improves all six sources at a cost of 0.6 points (Appendix D.6).

Table 1: **Registered average gains coexist with source losses, with different consequences for the ranking.** Outlier Exposure (OE) minus vanilla area under the receiver operating characteristic curve (AUROC), in points, using the maximum softmax probability scorer (MSP) for both. The loss source is CIFAR-10 for CIFAR-100 ID, and iNaturalist for ImageNet-200 ID. The tie weight is on that source, with the remainder on TinyImageNet or, for ImageNet-200, split equally over SSB-hard and NINCO, and Textures and OpenImage-O have zero weight on that path. Brackets give 95% seed-bootstrap ranges for boundaries of mean performance. The source-loss predictions preceded training, and boundary read-outs use the saved results. WRN denotes WideResNet.

ID / network	Seeds	Near gain	Source loss	Tie weight [95%]
CIFAR-100 / ResNet-18	10	$+8.19$	-0.61	0.97 [0.96, 0.98]
CIFAR-100 / WRN-40-2	10	$+9.19$	-2.43	0.90 [0.88, 0.91]
ImageNet-200 / ResNet-18	5	$+2.02$	-3.60	0.36 [0.33, 0.38]

The ImageNet-200 prediction preceded training and was confirmed on five seeds: MSP gains 2.02 points on the near pair and loses 3.60 on iNaturalist. Yet the decision consequences differ. The CIFAR-100 ResNet-18 MSP advantage persists until CIFAR-10 has about 97 percent of the near-pair weight. The ImageNet-200 advantage ends at about 36 percent iNaturalist weight when the remainder is split equally over SSB-hard and NINCO, with Textures and OpenImage-O excluded. Under equal weights over all five sources, OE minus vanilla is -0.231 points, with a seed-bootstrap range of $[-0.486, -0.031]$. Our ImageNet-200 vanilla models reach 84.41 percent ID accuracy against 86.37 for OpenOOD's released checkpoint, and a near-OOD AUROC 1.6 points lower (Appendix C.4). The OE-minus-vanilla contrasts are internal to our runs.

Table 17 gives tie and one-point-deficit boundaries for all matched image comparisons. Against Definition 1, OE with MSP on CIFAR-100 meets the first at the equal near pair against every vanilla scorer. Its lead over the strongest of them, vanilla with Scale, ends at CIFAR-10 weight 0.934, a total-variation shift of 0.43. It therefore meets the second condition for any allowed shift below that, a bound we report because no shift was stated in advance. Per seed the MSP boundary ranges from 0.94 to 0.98, and one seed shows no loss (Table 17). On ImageNet-200, OE's margin at the equal five-source reference is negative, so it fails the first condition there. A replication protocol committed before five further ImageNet-200 runs per arm (Appendix C.4, Table 10) reproduces the iNaturalist loss, -3.31 points with MSP against a threshold of 0.44, and the near-pair gain over vanilla with MSP, 2.06. Against vanilla's best scorer on that pair, Scale, the lead is 0.44 points (one-sided 95% bounds 0.28 to 0.60) at an ID-accuracy cost of 0.26 points (upper bound 0.56). The requirement fixed before those runs (Section 3.2) is therefore not met, and the upper bound does not place the gain below 0.5 either. OE's lead over every original scorer ends once iNaturalist carries 0.061 of the weight.

The source losses depend on the ID dataset and scorer settings. In three of the four CIFAR-10 architecture/pool comparisons, OE improves AUROC over vanilla on both near sources with all seven scorers. We tested one such consistent gain against all three conditions, and it meets them. With ResNet-18 and TinyImageNet-597, OE with MSP beats every vanilla choice on every measured source, and on the separate images of Section 4.4 it leads vanilla's best scorer by 8.81 points (one-sided lower bound 8.52) at an ID-accuracy cost of 0.07 (upper bound 0.23), and it accepts 34.8 percent of OOD images against 70.4 (Appendix H.1, a read-out of saved outputs). On CIFAR-100 ResNet-18, reselection on Textures reduces the largest loss among the four reselected scorers to 1.2 points, and ODIN retains its settings and approximately 3.1-point loss. Appendices D.4 and D.5 report the controls and complete source profiles, and an exploratory text experiment improves both of its sources (Appendix E).

4.2 Repairing a Source Can Make the Ranking Less Stable

In the CIFAR-100 ResNet-18 confirmation, replacing TinyImageNet-597 auxiliaries with the 300K pool improves Scale's OE-minus-vanilla difference on CIFAR-10 by 5.05 points (Table 2). The TinyImageNet gain falls more, so the CIFAR-10 crossover decreases from 0.716 to 0.581. From equal weights on the pair, the TV shift needed to reach the tie falls from 0.216 to 0.081.

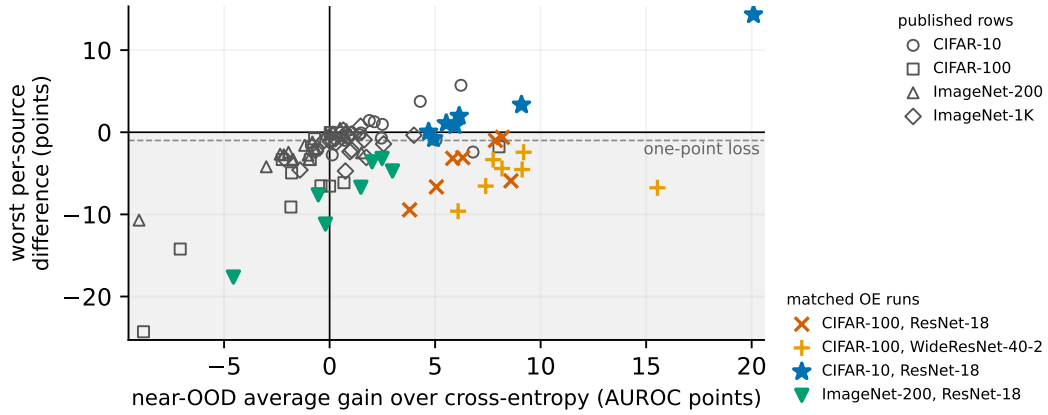


Figure 2: **Average gains sit beside per-source losses in the published tables and in matched retraining.** Each open marker is a training-method row of an OpenOOD v1.5 AUROC sheet with a cross-entropy row under the same scorer (67 rows): the near-OOD average gain over that row against the smallest per-source difference over all measured sources. The other markers are OE minus vanilla per scorer from our matched runs, seed means. Of the 39 published rows that raise the near-OOD average, 29 fall below cross-entropy on some source, and 13 by more than a point at the published means (dashed line; 12 also beyond twice the published standard error, Section 4.3). Against cross-entropy under its best published scorer, 19 of the 39 gains remain, and none leads every such scorer on every source. *Abbreviations:* area under the receiver operating characteristic curve (AUROC); Outlier Exposure (OE); out-of-distribution (OOD).

Table 2: **A source repair can reduce ranking stability.** CIFAR-100 in-distribution, ResNet-18, Scale with original settings. Differences are Outlier Exposure minus vanilla area under the receiver operating characteristic curve (AUROC), in points. Crossover weights are on CIFAR-10 within its pair with TinyImageNet. TV is the total-variation distance between source weights.

Quantity	TinyImageNet pool	300K pool
CIFAR-10 source difference	-6.653	-1.602
TinyImageNet source difference	$+16.771$	$+2.221$
CIFAR-10 weight at tie	0.716	0.581
TV shift from equal pair to tie	0.216	0.081
CIFAR-10 weight at one-point deficit	0.759	0.843

Joint seed resampling gives a source-repair range of $[1.709, 9.553]$ points and a crossover-change range of $[-0.260, -0.014]$. The comparison is a contrast at a fixed scorer. Under the first condition the repaired Scale detector is not the repaired model's best choice, since its MSP is higher on both sources (78.74 and 85.71 against 78.15 and 85.40). The example therefore describes how a per-scorer boundary moves rather than a switching case. ODIN gives a second example of source repair with a lower crossover. Other scorers differ (Appendix Figure 6). The repair shrinks the advantage on the stronger source by more than it improves the weaker one, so the tie moves closer. This pattern is specific to the measured comparisons. The one-point-deficit weight moves from 0.759 to 0.843, with a range $[-0.050, +0.213]$ that includes zero, so tolerance to a stated deficit is a separate objective and here an uncertain one.

4.3 Published Rankings Depend on the Candidate Set and Source Mixture

We apply the analysis to four OpenOOD v1.5 AUROC tables, with 41 to 47 candidates per benchmark (Zhang et al., 2024). The published screen of Figure 2 identifies 39 near-average gains among 67 matched-scorer training-method comparisons. Of these gains, 13 accompany a loss of at least one point on a measured source at the published means. Twelve of them also lose beyond twice the standard error where the table publishes one, including four within the near pair. Seven of the 12 come from ImageNet-1K, whose table omits standard deviations, so their classification uses the

one-point threshold alone. These counts are descriptive (Table 5). Against cross-entropy under its best published scorer, 19 of the 39 gains remain on the near-OOD average, and none leads every cross-entropy scorer on every source.

At the published point estimates, only 5 of 44, 12 of 41, 6 of 47 and 3 of 46 candidates are mixture-admissible, able to lead at some source mixture, on CIFAR-10, CIFAR-100, ImageNet-200 and ImageNet-1K, respectively.

The equal-six-source reference gives RotPred a tie radius of 0.023, whereas OE with MSP leads at the equal near-pair reference with radius 0.361 (Table 3). Changing the reference can change the leader, restricting within-group movement changes the feasible shifts and hence the radius, and a source-specific loss can coexist with a stable average lead.

At the published standard-error scale, the equal-source leader keeps first place in 88 to 89 percent of perturbed tables and its fixed-leader tie radius ranges from 0 to 0.047. The near-pair leader stays first in every replicate even at twice that scale. These are sensitivity calculations under Gaussian noise models (Appendix G). The existing selection exercises show no general withheld-source benefit. Choosing by the worst near source increases far-source regret in one matched environment, and the published-table minimax-regret rule fails its prespecified success criterion (Appendices B and D.3). These outcomes limit the use of measured-source comparisons as selection policies for new sources.

4.4 A Blend Adds a Small Gain That Does Not Justify Switching

We tested an exploratory blend of the two auxiliary pools (75 percent TinyImageNet-597), scored with MSP, against the original CIFAR-10 choices. Those choices are vanilla and OE with either pure pool, under all seven scorers and both settings regimes, and we call this comparison the discovery analysis (Appendix C.2). On the near pair, the blend improves on all originals by at most 0.870 AUROC points (bootstrap range 0.814 to 0.937), and the margin remains positive on every held-out seed at an ID-accuracy cost of roughly 0.29 percentage points. At equal weight over five sources, the near pair with MNIST, SVHN and Places365, its margin is -0.287.

To test the switching requirement, we froze the blend, the 42 original choices with their ten checkpoints and settings, and a CIFAR-100 test weight of 0.65. We then evaluated them on separate images: $9{,}600$ CIFAR-10.2 images as ID, $2{,}400$ separate images for threshold calibration, and $10{,}000$ images from each near-OOD source drawn from separate splits of the measured categories after overlap screening (Lu et al., 2020). On these separate images, the blend retains an AUROC margin of 0.46 points over the best original (paired image-bootstrap range $[0.244, 0.645]$). Its ID-accuracy cost is 0.647 percentage points (range $[0.440, 0.852]$), and it raises OOD false acceptance, the FPR@95 of Section 3.2 with thresholds set on separate ID images, by 3.03 points (range $[0.859, 6.069]$).

The requirement of Section 3.2, fixed before this evaluation and after the discovery result, is not met. The smallest gain lower bound is 0.281 against 0.5, the accuracy-cost upper bound is 0.819 against 0.5, and the requirement still fails when images and training runs are resampled jointly. ID-accuracy ceilings of 0, 0.25 and 0.5 points retain an original, and a one-point ceiling would admit the blend. Against the three conditions, the blend meets the first at the equal near pair on the discovery images, fails the first at the equal five-source reference, and fails the third on separate images (Table 19).

5 Related Work

Evaluating OOD detectors. Shafaei et al. (2019) showed that evaluation on development OOD sets can overstate performance and proposed tests on withheld OOD sources. OpenOOD (Yang et al., 2022; Zhang et al., 2024) separates validation from test OOD categories and publishes per-dataset results, surveys organize the field by the kind of shift (Yang et al., 2024), and near-OOD sources remain challenging (Fort et al., 2021). Rankings of OOD detectors are known to be inconsistent across datasets and sensitive to seeds and splits (Tajwar et al., 2021; Szyc et al., 2023), and OOD test sets themselves have needed repair (Bitterwolf et al., 2023). Benchmark conclusions in other fields change with the choice and aggregation of tasks (Dehghani et al., 2021; Zhang & Hardt, 2024) and with variance across training runs (Bouthillier et al., 2021; Agarwal et al., 2021). Our work differs in two ways. We test with registered retraining whether a training method's loss on one

Table 3: **The reference and allowed shifts change the boundary.** Published CIFAR-100 means. Tie and Deficit are the smallest total-variation shifts of the source weights at which a rival ties the leader or leads it by one point of area under the receiver operating characteristic curve. RotPred is rotation prediction (Hendrycks et al., 2019b). Both 50/50 references weight sources equally within each near/far group. Only the grouped-shift row holds those proportions fixed during reweighting. Outlier Exposure (OE) uses maximum softmax probability (MSP).

Reference and allowed shifts	Leader	Tie	Deficit
Equal six sources, unrestricted	RotPred	0.023	0.056
50% near / 50% far, unrestricted	OE + MSP	0.080	0.112
Same 50/50 reference, group shifts only	OE + MSP	0.129	0.182
Equal near pair, shifts within pair	OE + MSP	0.361	0.413

source recurs, and we compute how far the mix of sources can shift before the preferred detector changes.

Training and scoring for OOD detection. OE trains on real auxiliary OOD images, from which its authors removed test-set overlap (Hendrycks et al., 2019a). Later work samples informative auxiliary images (Ming et al., 2022), transforms or augments the auxiliary set (Wang et al., 2023b;a), or learns from unlabeled wild data (Katz-Samuels et al., 2022; Du et al., 2024), and Bitterwolf et al. (2022) compare auxiliary objectives on shared OOD training data and analyse what each estimates. Energy-bounded training (Liu et al., 2020), MixOE (Zhang et al., 2023; 2018), Pursuing Feature Separation (PFS) (Wu et al., 2025), VOS and NPOS (Du et al., 2022; Tao et al., 2023), and LogitNorm and CIDER (Wei et al., 2022; Ming et al., 2023) change the objective through energy margins, mixing, feature separation, synthetic outliers, or regularized logits and embeddings. These methods propose objectives, and our conditions apply to the gain of any of them, with OE and LogitNorm as registered cases and the published screen covering the others. Post-hoc scorers (Section 3.2) also change a fixed model's per-source profile, as do others in the published tables that use Mahalanobis distances, gradient norms, virtual logits or sparsified weights (Lee et al., 2018; Huang et al., 2021; Wang et al., 2022; Sun & Li, 2022). The first condition therefore compares a new detector with the original model under every scorer. On the ImageNet-200 replication runs, this comparison reduces OE's near-OOD lead from 2.06 points over vanilla with MSP to 0.44 over vanilla with Scale.

Shifts in the mix of sources. A change in how often each OOD source occurs is a prior shift over sources, the setting of label-shift estimation (Saerens et al., 2002; Lipton et al., 2018). A guarantee over a set of such mixtures is the aim of distributionally robust optimization (DRO) (Ben-Tal et al., 2013; Duchi & Namkoong, 2021), which group DRO applies to the worst mixture of training groups (Sagawa et al., 2020). We keep the detectors fixed and ask the converse question, how far the mixture can move before the preferred detector changes. Label-shift estimates from deployment data could supply the reference mixture that Definition 1 takes as given.

6 Conclusion

A benchmark gain for OOD detection describes one mix of unfamiliar inputs, and deployment may not supply that mix. Registered comparisons show that such a gain can come with a reproducible loss below cross-entropy on another test set, for OE across two CIFAR-100 architectures and ImageNet-200 and for LogitNorm without auxiliary OOD data. A gain matters only where it changes the best available detector. Whether a new detector should replace the current one then depends on how far the mix can shift before it stops being superior to every existing choice, including the baseline under each scorer. The answer also depends on whether its lead holds on new data. This comparison changes conclusions. At the published means, about half of the training gains do not survive comparison with the baseline under its best scorer, and none holds under every mix. On ImageNet-200, OE's lead over that scorer is 0.44 points, short of the requirement fixed before five new runs, and a blend of auxiliary pools misses its requirement on new images. With CIFAR-10 as the ID dataset, OE's gain holds on every measured source and keeps an 8.8-point lead on new images. OOD detectors should therefore be chosen by their stability over plausible mixes, and gains reported with the range of mixes over which they hold.

Ethics statement

This work adheres to the ICLR Code of Ethics. The image experiments use publicly available benchmarks (CIFAR-10, CIFAR-10.2, CIFAR-100, TinyImageNet and ImageNet-200, with the OOD test sets of OpenOOD v1.5) and classifiers trained on them. The separate text experiment uses SST-2 (Socher et al., 2013), IMDB (Maas et al., 2011), RTE from GLUE (Wang et al., 2019) and WikiText-2 (Merity et al., 2017). The auxiliary image sets (TinyImageNet-597, a subsample of 300K Random Images, and Places365 and SVHN in archived runs) are public. 300K Random Images is the cleaned subset that the OE authors published of 80 Million Tiny Images, a collection its creators withdrew in 2020 over offensive labels and images. We use the published subset only, as images without labels, and did not access the withdrawn collection itself. The study involves no human subjects and collects no personal data.

Reproducibility statement

The supplemental zip contains the registered configurations, recorded per-seed evaluations and analysis scripts, with an evidence index and a file manifest. Its README identifies path configuration needs and artifacts not included in the archive. The exploratory decision analyses include a frozen input snapshot, saved outputs and a portable claim checker (Appendix G). The prespecified transfer evaluation includes its protocol, sample hashes, replay checks, per-checkpoint metrics and compact bootstrap arrays that reproduce the reported decision summaries (Appendix H). The original checkpoints and per-example score arrays are not included in the zip. It also contains the script that computes robustness intervals from any table of per-set results, with the parsed published tables of Section 4.3. The evaluation protocol is in Section 3.2. Appendix C records the predictions and decision rules fixed before evaluation, their original verdicts, and the distinction between exploratory analyses and subsequent confirmation. Archived tables (Appendix F) survive as aggregates only and support no significance claim. Registered comparisons that the paper does not detail are listed with their verdicts in Appendix C, and their files are in the zip.

LLM use statement

We used a large language model (LLM) to aid or polish writing of selected passages, and for code autocompletion, generation, documentation, debugging, exhaustive testing, and project management. We submitted the complete manuscript to a locally served LLM for adversarial review and feedback. To preserve confidentiality, we did not submit the full paper to any LLM hosted by a third party.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, 2021.

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Julian Bitterwolf, Alexander Meinke, Maximilian Augustin, and Matthias Hein. Breaking down out-of-distribution detection: Many methods based on OOD training data estimate a combination of the same core quantities. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2041–2074, 2022.

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing ImageNet out-of-distribution detection evaluation. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2471–2506, 2023.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, et al. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, 2021.

Matthew Cook, Alina Zare, and Paul Gader. Outlier detection through null space analysis of neural networks. *arXiv preprint arXiv:2007.01263*, 2020. ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ndYXTEL6nR.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: Learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TW7d65uYu5M.

Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does unlabeled data provably help out-of-distribution detection? In *International Conference on Learning Representations*, 2024.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3), 2021.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. doi: 10.1109/CVPR.2016.90.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019a. URL https://openreview.net/forum?id=HyxCxhRcY7.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, volume 32, 2019b.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Julian Katz-Samuels, Julia B. Nakhleh, Robert Nowak, and Yixuan Li. Training OOD detectors in their natural habitats. In *International Conference on Machine Learning*, 2022.

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

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, 2018.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Shangyun Lu, Bradley Nott, Aaron Olson, Alberto Todeschini, Hossein Vahabi, Yair Carmon, and Ludwig Schmidt. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020. URL https://www.gatsby.ucl.ac.uk/~balaji/udl2020/accepted-papers/UDL2020-paper-101.pdf.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, 2011.

Jiří Mazurek and Dominik Strzałka. On the monte carlo weights in multiple criteria decision analysis. *PLOS ONE*, 17(10):e0268950, 2022.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Yifei Ming, Ying Fan, and Yixuan Li. POEM: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, pp. 15650–15665, 2022.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=aEFaE0W5pAd.

Diederik M. Roijers, Shimon Whiteson, and Frans A. Oliehoek. Computing convex coverage sets for faster multi-objective coordination. *Journal of Artificial Intelligence Research*, 52:399–443, 2015.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Alireza Shafaei, Mark Schmidt, and James J. Little. A less biased evaluation of out-of-distribution sample detectors. In *British Machine Vision Conference (BMVC)*, 2019.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.

Yiyou Sun and Yixuan Li. DICE: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, 2022. doi: 10.1007/978-3-031-20053-3_40.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20827–20840, 2022. URL https://proceedings.mlr.press/v162/sun22d.html.

Kamil Szyc, Tomasz Walkowiak, and Henryk Maciejewski. Why out-of-distribution detection experiments are not reliable: Subtle experimental details muddle the OOD detector rankings. In *Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2078–2088, 2023.

Fahim Tajwar, Ananya Kumar, Sang Michael Xie, and Percy Liang. No true state-of-the-art? OOD detection methods are inconsistent across datasets. *arXiv preprint arXiv:2109.05554*, 2021.

Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JHklpEZqduQ.

Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.

Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: Letter of withdrawal. https://groups.csail.mit.edu/vision/TinyImages/, 2020.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=5hLP5JY9S2d.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.00487.

Qizhou Wang, Zhen Fang, Yonggang Zhang, Feng Liu, Yixuan Li, and Bo Han. Learning to augment distributions for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2023a.

Qizhou Wang, Junjie Ye, Feng Liu, Quanyu Dai, Marcus Kalander, Tongliang Liu, Jianye Hao, and Bo Han. Out-of-distribution detection with implicit outlier transformation. In *International Conference on Learning Representations*, 2023b.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23631–23644, 2022. URL https://proceedings.mlr.press/v162/wei22d.html.

Yingwen Wu, Ruiji Yu, Xinwen Cheng, Zhengbao He, and Xiaolin Huang. Pursuing feature separation based on neural collapse for out-of-distribution detection. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sNEMqdReel.

Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nanyAujl6e.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kun Ding, et al. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision*, 131:2607–2622, 2023.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132:5635–5662, 2024.

Guanhua Zhang and Moritz Hardt. Inherent trade-offs between diversity and stability in multi-task benchmarks. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58984–59002, 2024.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. doi: 10.1109/WACV56688.2023.00549.

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. *Journal of Data-centric Machine Learning Research*, 2(3):1–32, 2024. URL https://data.mlr.press/assets/pdf/v02-3.pdf.

A Proof of the Mixture Identity and Its Consequences

A detector is a score function s, larger for inputs judged in-distribution. For an ID distribution P_0 and an OOD distribution Q, with $X \sim P_0$ and $Y \sim Q$ independent,

$$\mathrm{AUROC}(s; P_0, Q) = \Pr[s(X) > s(Y)] + \tfrac{1}{2}\Pr[s(X) = s(Y)].$$

The simplex $\mathcal{S}_K$ is the set of nonnegative source weights that sum to one. For weights w in this simplex write $Q_w = \sum_k w_k P_k$.

Lemma 1. $\mathrm{AUROC}(s; P_0, Q_w) = \sum_k w_k \,\mathrm{AUROC}(s; P_0, P_k)$.

Proof. Draw $Y \sim Q_w$ by first drawing a source index κ with $\Pr[\kappa = k] = w_k$ and then $Y \sim P_\kappa$. Both probabilities in the definition are expectations over κ of the same probabilities conditional on κ, and X is independent of (κ, Y), so each term splits over sources with weights w_k. □

The same identity holds exactly for test sets. The empirical AUROC compares every ID score with every OOD score and gives half credit to ties (the Mann–Whitney statistic). For an ID test set against a pooled OOD test set in which source k contributes n_k images it is equal to $\sum_k (n_k / \sum_j n_j)\,\mathrm{AUROC}_k$, so an unweighted benchmark average is the AUROC against a pool in which every source has the same size.

Proof of the Mixture Identity. By Lemma 1, $\Delta(w) = \sum_k w_k \Delta_k$. A linear function on $\mathcal{S}_K$ attains its minimum at a vertex, so $\Delta(w) \geq 0$ for every w if and only if $\Delta_k \geq 0$ for every k, and $\Delta(w) > 0$ for every w if and only if $\Delta_k > 0$ for every k. For $K = 2$ and weight w on P_1, $\Delta(w) = w\Delta_1 + (1-w)\Delta_2$ is affine in w. When $\Delta_1\Delta_2 < 0$ it vanishes only at $w^* = -\Delta_2/(\Delta_1 - \Delta_2) \in (0,1)$, so $\Delta(w)$ has the sign of $\Delta(\frac{1}{2})$ exactly for the w on the same side of w^* as $\frac{1}{2}$, provided $\Delta(\frac{1}{2}) \neq 0$. The excluded case is $w^* = \frac{1}{2}$, where the benchmark itself is a tie and both signs occur in every neighbourhood of it, and the intervals reported in Section 4.3 are closed, so a weight at which a rival draws level is inside the interval. □

Let $a_d \in \mathbb{R}^K$ collect detector d's per-source AUROC, with $K \geq 1$, and let $\mathcal{D}$ be a finite nonempty set of detectors. A singleton family has zero regret and no rival boundary. The comparisons below assume at least two detectors. The following consequences are used in Section 4.3.

Mixture-admissible detectors. Detector d is best for some mixture if and only if some $w \in \mathcal{S}_K$ satisfies $w \cdot (a_d - a_j) \geq 0$ for every $j \in \mathcal{D}$. This condition is the feasibility problem solved by linear programming in Section 4.3. By the minimax theorem for the bilinear function $w \cdot (a_d - \sum_j q_j a_j)$ on two simplices,

$$\max_{w \in \mathcal{S}_K} \min_{j \neq d} w \cdot (a_d - a_j) = \min_{q} \max_{k} \Big(a_{dk} - \sum_{j \neq d} q_j a_{jk} \Big),$$

where q ranges over probability vectors on the other detectors. Hence d is best under no mixture if and only if some such q has $\sum_j q_j a_{jk} > a_{dk}$ for every source k. For each w, some detector j with

$q_j > 0$ then has $w \cdot a_j > w \cdot a_d$, so the randomization is only a certificate: under every weighting some single published detector beats d, and which one may change with w. Because zero source weights and ties are allowed, a weakly dominated detector can still be admissible on a boundary face where its deficits receive zero weight. The set counts reported below are checked directly.

Worst-case regret. The regret of d at w is $R_d(w) = \max_j w \cdot a_j - w \cdot a_d$, a maximum of linear functions of w and therefore convex. Its maximum over $\mathcal{S}_K$ is attained at a vertex: $\max_w R_d(w) = \max_k \big(\max_j a_{jk} - a_{dk} \big)$.

FPR at 95% true-positive rate. If the threshold τ of a detector is set on ID data alone, its false-positive rate against Q_w is $\Pr_{Y \sim Q_w}[s(Y) \geq \tau] = \sum_k w_k \Pr_{Y \sim P_k}[s(Y) \geq \tau]$, because τ does not depend on w. Every statement of this section therefore holds for FPR@95 under that condition, with lower values better. Our CIFAR read-out satisfies it, since it fixes τ at the fifth percentile of ID confidence and reports the share of OOD inputs above it. OpenOOD's published FPR@95 does not. Its implementation treats OOD as the positive class and reports the share of ID inputs below the threshold at which 95 percent of that source's OOD inputs are detected, so every column carries its own threshold and the threshold moves with the source. A weighted average of those columns is the benchmark's own average of separately calibrated operating points, and it is not the false-positive rate against the mixture. The two differ by more than rounding. With ID confidences uniform on $[0, 1]$ and two sources uniform on $[0, 0.2]$ and $[0.8, 1]$, the per-source rates at 95 percent detection are 0.19 and 0.99, averaging 0.59, while detecting 95 percent of the equal mixture needs a threshold that leaves 0.98. There is no fixed ordering: with the same sources weighted 0.96 and 0.04, the average is 0.222, whereas the mixture's threshold is $0.2(0.95/0.96)$ and its rate is approximately 0.198. The linear mixture identity also need not hold for the area under the precision–recall curve, whose precision term depends on the pooled OOD set.

The robustness radius. For reference weights u and the detector L that is best at u, which Section 2 writes w_0 and d_0, the radius in total-variation distance (TV) is

$$r(u) = \min_{j \neq L} \inf \big\{ \mathrm{TV}(w, u) : \; w \in \mathcal{S}_K, \; w \cdot (a_L - a_j) \leq 0 \big\}, \qquad \mathrm{TV}(w, u) = \tfrac{1}{2} \|w - u\|_1.$$

We set the infimum of an empty feasible set to $+\infty$. Every nonempty feasible set is compact, so its infimum is attained. This radius is a distance between source-weight vectors: $\mathrm{TV}(Q_w, Q_u) \leq \mathrm{TV}(w, u)$, with equality not guaranteed when sources overlap. The radius is the smallest shift that permits a tie, and it does not claim that every shift of that size causes one. Fix j and let $m = a_L - a_j$, choose $k^* \in \arg\min_k m_k$, and set $g = u \cdot m$. If $g \leq 0$ the rival is already at least as good and the radius is 0. If $g > 0$ and $m_{k^*} > 0$ the leader is ahead of j on every source and j contributes no boundary. If $g > 0$ and $m_{k^*} = 0$ the rival draws level only where all the weight sits on the sources it ties on, at distance $1 - \sum_{k : m_k = 0} u_k$. No rival of any leader is of this kind in any sheet we analyse, which we checked. The remaining case is $g > 0$ and $m_{k^*} < 0$. Any $w \in \mathcal{S}_K$ arises from u by removing mass $\epsilon_s \leq u_s$ from some sources and adding the same total t to others, and $\mathrm{TV}(w, u) = t$. The decrease of $w \cdot m$ below $u \cdot m$ is at most $\sum_s \epsilon_s (m_s - m_{k^*})$, with equality when all mass is added at k^*. Reaching a decrease of $u \cdot m$ with the least total t is then a fractional knapsack problem, solved by removing mass from sources in decreasing order of m_s and adding it at k^*. This greedy rule is the closed form used in Section 4.3, and $r(u)$ is its minimum over j. The closed form was checked against a linear program on 200 continuous random instances and on discrete instances with ties and zero reference weights. For a required rival lead of $\delta \geq 0$ AUROC points, replace m by $m + \delta \mathbf{1}$ throughout, since $w \cdot \mathbf{1} = 1$. This gives the tie radius when $\delta = 0$, and the smallest shift to a deficit of at least δ otherwise. A rival that only ties on a boundary has finite tie radius but no strictly better mixture.

B The Published-Table Analysis

The analysis of §4.3 uses the results sheet that OpenOOD v1.5 links from its report, exported on 2026-09-19 (the supplement holds the export, its checksum and the parser). The sheet includes entries added after the report, such as RotPred with PixMix on CIFAR-10. A detector is a row with a published AUROC on every OOD source of its benchmark. The plan, fixed before the quantities below were computed, names what the authors had already seen: the near-OOD pairs, their reversals

and the ten leaders per benchmark. A detector is mixture-admissible when a linear program finds source weights for which no other detector has higher AUROC. Region shares use $200{,}000$ weight vectors drawn uniformly from the simplex.

Seed-noise sensitivity. The seed-noise check draws $1{,}000$ replicate tables with independent normal noise at each cell's standard error. This assumes independent noise, ignoring potential covariance between detectors evaluated on shared test sets. Under this assumption, it yields 5th–95th percentile ranges of 4 to 6, 10 to 14 and 6 to 8 admissible detectors on CIFAR-10, CIFAR-100 and ImageNet-200. It leaves the published near-OOD leader the most frequent winner across uniformly drawn all-source mixtures in every replicate on CIFAR-10, in 15 percent of replicates on CIFAR-100 and in none on ImageNet-200. The ImageNet-1K sheets publish no standard deviations, so their cells take no noise and every replicate of them is the published table, with 3 admissible detectors and its leader always the most frequent winner. That is a property of the missing standard deviations and not a stability result.

Near-OOD intervals and held-out selection. On the near-OOD pair alone two detectors are admissible on three benchmarks and one on ImageNet-1K. The leader's interval is $[0, 0.86]$ in the weight on CIFAR-10 for CIFAR-100, $[0.04, 1]$ in the weight on SSB-hard for ImageNet-200 and $[0.49, 1]$ in the weight on CIFAR-100 for CIFAR-10, where the leader and OE are 0.03 points apart. Near-OOD leaders can remain preferable despite a per-source loss. At the published means OE with MSP leads the near-OOD average by 7.0 points on CIFAR-100, ranks first on TinyImageNet and 19th of 41 on CIFAR-10 as OOD. Within the pair its lead holds until CIFAR-10 carries 0.86 of the weight, and a rival is ahead by one point only after 0.413 of the weight moves from the equal average. For the leave-one-source-out test each rule selects a detector from the published means on all sources but one, and its held-out regret is the best published AUROC on the remaining source minus the selected detector's. Mean held-out regret for selection by the average, by the worst source and by minimax regret is 0.80, 0.80 and 0.91 on CIFAR-10, 13.99, 12.18 and 11.10 on CIFAR-100, 2.70, 4.20 and 1.96 on ImageNet-200 and 0.29 for all three on ImageNet-1K. Minimax regret and the average select different detectors in 7 of 22 cases, with minimax regret lower in 5. The plan's label required a lower mean on three benchmarks, so both comparisons are labelled no clear difference.

Radius conventions. Each radius is computed on the simplex of the leader's own sources, with the leader's margin beside it, and a lead is called a tie when the 5th percentile of its resampled radius is zero. The resampling is repeated with noise shared within a training method, one draw per training method and source, as a second assumption about the dependence that the sheet does not publish. Neither model captures dependence across methods through shared test images, the two give nearly the same intervals, and Table 5 reports the wider. The radius needs no distribution over weights, but like the shares it depends on how a benchmark divides its sources, because the reference weights are uniform over the listed sources.

Dominance. A detector is dominated if another is at least as good on every source and strictly better on at least one. On CIFAR-100, 22 of 41 detectors are not dominated on all sources and 12 are admissible, so ten undominated detectors are best under no mixture at the published means. The distinction is fragile: in an exploratory check, the margin by which other detectors beat such a row under every weighting (Appendix A) stays above half a point in 95 percent of resampled tables for one of the ten on CIFAR-100 and one of seven on ImageNet-200. Every sheet analysed is in Table 5.

Ties and consequential deficits. The robustness radius is computed in closed form by moving weight from the sources where the leader's margin over a rival is largest to the source where it is smallest, and was checked against a linear program. A radius records where a leader stops being uniquely best, and an exploratory computation asks the next question, how far the weights move before a rival leads by a stated margin. On the all-source simplex a one-point deficit needs a shift of 0.056 on CIFAR-100, 0.286 on CIFAR-10, 0.397 on ImageNet-200 and 0.711 on ImageNet-1K, against ties at 0.023, 0.138, 0.097 and 0.442. Within the near-OOD pair the CIFAR-100 leader concedes a point after 0.413 and no rival of the ImageNet-200 leader reaches a point anywhere on that simplex. Where a tie comes early and a point comes soon after, as on CIFAR-100, the ranking and the performance are both fragile. Where a tie comes early and a point never arrives, the leaders are close and the choice between them matters little.

Table 4: **Tie and relative-deficit boundaries with source-table sensitivity.** Area under the receiver operating characteristic curve (AUROC) from published source means. Radii measure total-variation distance between source weights from the named equal-weight reference. The deficit is one AUROC point relative to a rival at the shifted mixture. Ranges span the conditional 5th to 95th percentiles over finite boundaries in the independent and shared noise models, with the original leader fixed and 2,000 perturbations per model. These are sensitivity ranges without selection-adjusted coverage guarantees. ImageNet-1K has no published standard deviations. None denotes an unattainable point-estimate boundary.

Benchmark	Sources averaged	Shift to tie [range]	Shift to deficit [range]
CIFAR-10	All sources	0.138 $[0.119, 0.156]$	0.286 $[0.252, 0.322]$
CIFAR-100	All sources	0.023 $[0.000, 0.047]$	0.056 $[0.027, 0.076]$
ImageNet-200	All sources	0.097 $[0.055, 0.133]$	0.397 $[0.378, 0.418]$
ImageNet-1K	All sources	0.442	0.711
CIFAR-100	Near pair	0.361 $[0.352, 0.369]$	0.413 $[0.403, 0.421]$
ImageNet-200	Near pair	0.461 $[0.212, 0.487]$	None

On the ImageNet-200 near pair, no tie boundary exists in 41.95% of independent-noise replicates and 41.50% of shared-noise replicates. No one-point-deficit boundary exists in 100% and 99.95%, respectively. The sole finite deficit replicate under shared noise has radius 0.482. All other reported noisy rows have finite boundaries in every replicate. Appendix G defines the noise models.

Loss counts and other metrics. The comparison of every published row with its cross-entropy baseline counts a loss or gain of at least one AUROC point that also exceeds twice the standard error $\sqrt{(s^2_{\text{row}} + s^2_{\text{base}})/3}$ of the difference of two three-seed means, where the sheet publishes standard deviations, which the ImageNet-1K sheets do not. On ImageNet-200, OE raises the near-OOD average by 1.51 points and lowers iNaturalist AUROC by 2.50 at the same scorer. The admissible sets, shares and radii are also computed on the two full-spectrum sheets (Yang et al., 2023), whose ID group also includes shifted images of familiar classes, and on the ViT-B-16 and Swin-T transformer sheets (Table 5). These are published three-seed means without per-seed rows, the sources of a benchmark share test protocols, and a held-out benchmark source is not a deployment distribution. Section 4.3 reports descriptive counts, without multiplicity correction or a population-frequency interpretation.

C Predictions, Decision Rules and Analysis History

Table 6 records the principal planned comparisons reported in this paper, their original labels and supporting results. Registration files, verdict files and the scripts that compute the labels from per-seed rows are in the supplement. The main text is organized around the empirical findings, and this record preserves the distinction between discovery, prediction and confirmation.

The original labels use the following vocabulary. Norm-targeted outlier exposure (NT-OE) is the diagnostic penalty with a target feature length conditioned on cosine similarity to class means. The constant-target version is NT-OE_0, and OE+NT adds the conditioned penalty to OE. The band is the largest minus smallest AUROC across scorers, and the floor is the smallest, computed per seed and then averaged. Breadth requires the floor to rise and the band to narrow beyond the decision threshold, with no scorer worse beyond it. Conditioning compares the two targets, and composition compares their combination with OE alone. A verdict is the label returned by the rule fixed before evaluation. Here noise names the two-standard-error threshold used for decisions. The readable component is the part of a feature that changes relative class logits, a centred-head variant of the row-space split of Cook et al. (2020). Its mean norm on OOD inputs divided by its nonzero mean on ID test inputs is the readable ratio.

Design and original decision rules. The registered experiment trains ResNet-18 from scratch for 200 epochs on CIFAR-10 and evaluates with OpenOOD v1.5 (Krizhevsky, 2009; He et al., 2016; Zhang et al., 2024). Near-OOD AUROC averages the CIFAR-100 and TinyImageNet test sets. Each scorer's hyperparameter is selected per run by AUROC on the ID validation split against a TinyImageNet validation split, over the grids of Appendix D. For each seed, the band is the largest

Table 5: **The published-table analysis on every sheet analysed.** OpenOOD v1.5's published means (Zhang et al., 2024). Pareto and admissible: detectors not dominated on all out-of-distribution (OOD) sources, and detectors that are best for some mixture of them. Radius: smallest total-variation shift of uniform weights at which another detector is at least as good, on the simplex of the leader's own sources (the near-OOD pair for the near-OOD leader, all sources for the all-source leader), beside the leader's margin over the runner-up in points. In parentheses, the 5th to 95th percentiles over 1,000 tables resampled within the published seed standard errors, the wider of independent noise and noise shared within a training method (the ImageNet-1K sheets publish no standard deviations). ‡A tie: the 5th percentile is zero. None, and a resampled value of 1: the leader is ahead on every source of the set. That value is a marker for the absence of a boundary and not a distance, since no weighting is further than $(K-1)/K$ from uniform weights on K sources. Share: fraction of uniformly drawn all-source mixtures on which the near-OOD leader is best. Gains: training-method rows that raise the near-OOD average over cross-entropy (CE) under the same scorer, with the number that lose on some source and, after the slash, inside the near-OOD pair. A loss is at least 1 point of area under the receiver operating characteristic curve (AUROC), and beyond twice the standard error where it is published. The ViT-B-16 and Swin-T sheets hold post-hoc scorers only. Descriptive, with no test.

	Near-OOD pair			All sources		
Sheet	Leader	Margin	Radius	Leader	Margin	Radius
AUROC						
CIFAR-10	RotPred + PixMix	0.03	0.01 (0.00 to 0.06)‡	RotPred + PixMix	0.39	0.14 (0.12 to 0.16)
CIFAR-100	OE + MSP	6.98	0.36 (0.35 to 0.37)	RotPred	0.71	0.02 (0.00 to 0.05)‡
ImageNet-200	OE + MSP	0.72	0.46 (0.23 to 1.00)	AugMix + ASH	0.40	0.10 (0.05 to 0.13)
ImageNet-1K, ResNet-50	AugMix + ASH	2.23	none	AugMix + ASH	1.57	0.44
ImageNet-200, full-spectrum	PixMix + RMDS	0.67	0.42 (0.05 to 1.00)	NPOS + KNN	0.69	0.04 (0.02 to 0.05)
ImageNet-1K, ResNet-50, full-spectrum	AugMix + SHE	1.39	0.26	AugMix + SHE	1.44	0.07
ImageNet-1K, ViT-B-16	CE + RMDS	1.05	none	CE + RMDS	0.42	0.21
ImageNet-1K, Swin-T	CE + GEN	0.62	none	CE + RMDS	0.57	0.11

Sheet	Detectors	Pareto	Admissible	Share	Gains (with a loss / inside the pair)
AUROC					
CIFAR-10	44	5	5	0.860	13 (1 / 0)
CIFAR-100	41	22	12	0.310	5 (3 / 1)
ImageNet-200	47	13	6	0.046	5 (1 / 0)
ImageNet-1K, ResNet-50	46	5	3	0.999	16 (7 / 3)
ImageNet-200, full-spectrum	53	19	6	0.076	21 (8 / 1)
ImageNet-1K, ResNet-50, full-spectrum	46	12	5	0.581	17 (11 / 3)
ImageNet-1K, ViT-B-16	16	5	4	0.908	–
ImageNet-1K, Swin-T	17	4	3	0.099	–

minus the smallest of the seven AUROCs and the floor is the smallest, both averaged over seeds. All auxiliary arms share trainer, epochs, seeds, data schedule, auxiliary set and the fraction of ID rows replaced by auxiliary rows, and differ in the loss applied to the auxiliary rows. Cluster A trained vanilla cross-entropy, OE, NT-OE and OE+NT, and later NT-OE$_0$ under its own registration. Cluster B, a different machine and software stack, trained OE, NT-OE, NT-OE$_0$, OE+NT and a mixing-only control (MIX) from the same configurations and seed indices, and a third cluster trained the pool intervention of Section 4.2. The two clusters share seed indices, and same-seed runs on the two clusters differ about as much as runs with different seeds (per-seed near-OOD AUROC correlations between -0.30 and $+0.48$). The replication changes the training environment but reuses the evaluation sets, so it does not test new-source generalization.

In the matched arms the mixture draw at epoch e is seeded by $(\text{run seed} \ll 20) + e$, so arms compared seed by seed see the same data schedule and different seeds draw different subsets.

Each registered comparison was decided by a rule fixed and committed before its arms were evaluated. The rule counts a difference when $|\Delta| \geq 2\sqrt{s_a^2/n_a + s_b^2/n_b}$ with sample standard deviations, which at $n = 10$ per arm corresponds to a two-sided Welch threshold between approximately 6.1 and 7.7% and carries no multiplicity control. Beside each registered label we therefore report a Holm correction at 5% over its family (nine endpoints for a breadth rule, seven scorers for a per-set rule), under paired tests on per-seed differences, since matched seeds share data schedules, and under unpaired Welch tests. Seed uncertainty is conditional on the fixed evaluation sets. A breadth rule asks whether one arm has a higher floor, a narrower band and no scorer lower beyond the threshold, and returns confirmed, refuted or partial. A refuted verdict covers both losses beyond the threshold and a gain that is not detected, and we say which in each case, since the rule tests no equivalence. A per-set rule asks whether at least five of seven scorers move in the predicted direction beyond the threshold and none against it. A replication was registered to succeed if the second cluster returned the same label with every counted scorer on the same side. Table 6 retains the original labels even when the main text describes the constituent effects directly. A failed compound rule is not evidence that all constituent effects are absent. Analyses computed after a verdict are identified in the result descriptions below.

Discovery and confirmation. The CIFAR-100 OE-versus-vanilla pattern was identified after the matched comparison and then specified as a prediction before the confirmation runs on ten new seeds. The CIFAR-10 per-set reversal was first observed on cluster A and predicted before evaluation on cluster B. The constant target's comparison with OE was exploratory on cluster B and subsequently registered for cluster A and the third cluster. The pool replacement and its second draw had plans fixed before their runs. The scorer-selection analysis was exploratory on cluster A and its pattern predicted for cluster B. The regret plan was written before computing regret, but after the authors had seen aggregate far-OOD results. Its targets were withheld from the selection procedures and were known to the authors (Appendix D.3). The statistical families and original verdicts are retained throughout.

Registered comparisons without a dedicated subsection. The remaining rows of Table 6 concern a diagnostic auxiliary objective and probes of where such a loss acts. The objective replaces OE's uniformity term with a penalty on auxiliary feature norms at the same exposure, with a constant target (NT-OE$_0$) or a target conditioned on the cosine to the nearest class mean (NT-OE). Their verdicts are listed with the rows. In summary, the penalty and OE rank oppositely across the two CIFAR-10 near-OOD sets as registered, and the gap shrinks by 85 to 97 percent with the 300K pool. Both targets are below OE with CIFAR-100 as ID on five scorers with wider bands, conditioning on the cosine lowers AUROC against the constant target on five scorers, and on the third cluster every arm with an auxiliary loss costs ID accuracy against OE (93.57 to 94.89 against 95.28). Three outcomes are disclosed here because the rows compress them. Energy-bounded training (Liu et al., 2020) was run from scratch with margins published for fine-tuning. The ID-only norm hinge lowered CIFAR-100 ID accuracy from 77.77 to 75.88 against its control. An archived placement experiment on three size-matched TinyImageNet-597 subsets failed its registered predictions for the logit scorers, and its per-seed rows were not recovered. Registration and verdict files for every row are in the supplement.

Table 6: Registered comparisons and verdicts. Clusters A, B and C denote computing environments (Section 3.2). "Same label" requires the earlier outcome and directions of counted scorers. Breadth compares minima and ranges across scorers. Readable ratio compares the mean norm of the feature component visible to softmax between out-of-distribution (OOD) and in-distribution (ID) inputs. Supp.: registration, verdict and per-seed files in the supplemental archive, which does not include the archived placement comparison. *Abbreviations:* Outlier Exposure (OE); auxiliary-image mixing with no auxiliary loss (MIX); the cosine-conditioned norm penalty (NT-OE); the constant-target norm penalty (NT-OE_0).

Comparison	Rule	Label	Where
NT-OE against OE, CIFAR-10, cluster A	breadth	partial	Supp.
NT-OE against OE, cluster B	same label	replicated	Supp.
NT-OE against OE per set, cluster B	per set, CIFAR-100 higher and TinyImageNet lower	held, held	Supp.
Test-time geometry, cluster B	higher cosine on TinyImageNet, lower norm ratio on CIFAR-100	held	Supp.
Textures-selected settings, cluster B	prediction of the cluster A pattern	held	§4.3
NT-OE against NT-OE_0, cluster B (primary)	conditioning rule	conditioning hurts	Supp.
NT-OE against NT-OE_0, cluster A	same label	conditioning hurts	Supp.
NT-OE and NT-OE_0 against MIX, cluster B	penalty rule	penalty helps, penalty helps	Supp.
NT-OE_0 against OE, cluster A	breadth, per set	confirmed, both held	Supp.
OE+NT against OE, cluster A	breadth	mixed	Supp.
OE+NT against OE, cluster B	breadth, per set	composition helps, TinyImageNet failed	Supp.
OE with 300K against OE with TIN-597, cluster C	source effect per set	source effect, both sets	§4.2
Objective-by-pool interaction, NT-OE_0 and NT-OE, cluster C	follows the pool	follows the pool, both	Supp.
Readable ratio under the pool change, cluster C	source effect, follows the pool	source effect, follows the pool (both)	§4.2
NT-OE_0 against OE with TIN-597, cluster C	breadth, per set	confirmed, both held	Supp.
NT-OE_0 against OE with 300K, cluster C	breadth	refuted	Supp.
Target scale $\alpha \in \{0, 0.1, 0.4\}$ against 0.2, cluster C	scale rule	lower, no difference, lower	Supp.
OE plus NT-OE_0 against OE and against NT-OE_0, cluster C	breadth	confirmed, refuted	Supp.
Readable-component probes, cluster C	per set against NT-OE_0	mixed	Supp.
Second, disjoint draw of the 300K subsample, cluster C	source effect, follows the pool, draw against draw	replicates, replicates, draws differ (constant target, TinyImageNet)	§4.2
CIFAR-100 confirmation on ten new seeds, both pools, with a mixing-only control	OE below vanilla on CIFAR-10 as OOD; mixing and OE-term contrasts; pool rules	confirmed; mixed and lower (no attribution); pool effect on TinyImageNet, none counted on CIFAR-10, OE with 300K below vanilla	§4.1, Tab. 13
Energy-bounded training against OE, both pools, cluster C	reverses by test set, follows the pool, breadth	mixed, mixed, refuted; accuracy cost	Supp.
Matched CIFAR-100, clusters A and B	breadth (two), conditioning, composition	refuted, refuted, inert, hurts	Supp.
ID-only hinge, CIFAR-100, five seeds per arm	replication with a hinge-off control	not replicated; ID-accuracy gate failed	Supp.
Penalty on synthetic rows without auxiliary images, four arms, cluster B	penalty rule against a masking control, pair-generator rule, against real auxiliaries	mixed (pixel), hurts (feature, shuffle), generator hurts, below real auxiliaries (both)	Supp.
Second architecture, WideResNet-40-2, eleven arms	the CIFAR-100 confirmation rule, mixing and OE-term contrasts, four reversal rules, source effect, follows the pool (two)	confirmed; no mixing effect and OE term lowers (attributed to OE's loss term); replicated (four); source effect (two); follows the pool (two)	§4.1, Tab. 14, Supp.
Published OpenOOD tables, leave-one-source-out selection	minimax regret against the average, worst set against the average	no clear difference (two)	§4.3, App. B
Selection regret on held-out sets, three benchmarks	one-sample rule on per-seed regret differences	no difference (B), worse (C), no difference on CIFAR-100 far-OOD, helps on CIFAR-10 as OOD, where the rule selects on the target	§4.3, App. D.3
ImageNet-200 confirmation, OE against vanilla, five seeds per arm	iNaturalist lower beyond noise (primary), near-OOD average higher	confirmed, as predicted	Tab. 8, App. C.4
Blended auxiliary pools, CIFAR-10 and CIFAR-100, four blends	half blend against the chord (original); MNIST lower and far-OOD average higher, quarter blend (amendment, primary); each blend against the hull of the original choices, MSP (amendment)	no clear departure from the chord; confirmed; all four outside the hull on CIFAR-10 and inside it on CIFAR-100	§4.4, App. C.1
Auxiliary images carrying no loss, CIFAR-100, three arms	TinyImageNet higher and MNIST lower on at least five of seven scorers	reversed: both rise	App. C.3
Auxiliary placement, three size-matched TinyImageNet-597 subsets, CIFAR-10, 100 epochs (archived)	registered predictions for the logit scorers	failed; per-seed rows not recovered	none

C.1 Blended Auxiliary Pools

The two pure pools of Section 4.2 are endpoints. Choosing one endpoint detector at random once for a whole evaluation gives, in expectation, the weighted mean of their per-source AUROCs, with weights equal to the selection probabilities. We call this mean the chord. Choosing independently per input introduces cross-detector score comparisons and is a different object. We train OE on blends with TinyImageNet-597 shares 0.25, 0.50 and 0.75, and on the union of both full lists. Each ID dataset has ten seeds per arm. Each configuration changes only the auxiliary list of its same-seed endpoint configuration.

The original chord comparison. The primary rule compared the half blend with the chord at weight 0.5 on each near-OOD source. It counted a source above or below the chord when at least five of seven scorers differed beyond the threshold with none against, and it required an above-chord source with no scorer below on the other. On CIFAR-10 the half blend is above the chord on CIFAR-100 and below it on TinyImageNet, both CIFAR-100 labels are mixed, and neither dataset meets the compound rule, so the original label is no clear departure and no recommendation to blend pools. The full union is labelled higher than the half blend on both CIFAR-10 sources, but under Holm correction only KNN's gain survives, none survives with CIFAR-100 as ID, and the union costs 0.19 ID-accuracy points against the half blend.

Comparison with every original choice. An amendment committed before any blend result was read asks whether a blend can lead some mixture of the two near-OOD sources. For CIFAR-10, the original choices are vanilla and OE with either pure pool. For CIFAR-100, they also include both mixing-only controls. Each blend is tested separately against this fixed set, with MSP carrying the verdict and the other scorers reported separately. The new blends are excluded from one another's registered comparison.

Let b be a blend's two-source AUROC vector and E_j those of the original choices. The optimized margin is

$$g = \max_{w \in \Delta_2} \min_j w \cdot (b - E_j).$$

It is the negative of the registered convex-combination margin, by the minimax identity of Appendix A. A positive g means that the blend beats every original choice at some weighting. A negative g means that at every weighting some original choice beats the blend. The uncertainty calculation resamples the same seed indices for every arm, with $2{,}000$ replicates. Its intervals concern each optimized margin separately. They are neither simultaneous bands over weights nor guarantees after choosing the largest of the four margins.

Table 7: Optimized gain g over the original choices, in area under the receiver operating characteristic curve (AUROC) points, for maximum softmax probability (MSP). Means from ten seeds per arm, with 95% seed-bootstrap intervals. Because g is a plug-in estimator maximizing a minimum over sample means on the simplex, its sample value may be subject to the optimizer's curse. Columns identify the in-distribution (ID) dataset. The other CIFAR dataset and TinyImageNet are the two out-of-distribution (OOD) sources. Each row excludes the other new blends.

TinyImageNet-597 share	CIFAR-10 ID	CIFAR-100 ID
0.25	0.591 $[0.504, 0.699]$	-0.530 $[-0.682, -0.334]$
0.50	0.799 $[0.747, 0.866]$	-0.768 $[-0.957, -0.604]$
0.75	0.946 $[0.894, 1.000]$	-0.723 $[-0.828, -0.624]$
Full union	0.798 $[0.729, 0.874]$	-0.532 $[-0.724, -0.353]$

All four CIFAR-10 blends receive the registered outside-the-hull label, and all four CIFAR-100 blends receive the inside-the-hull label (Table 7). These historical labels concern the sign of the optimized margin. Each CIFAR-10 blend beats every original choice at some weight. For each CIFAR-100 blend, an original choice is better at every weight, although its identity can change. The calculation does not test membership in the candidates' convex hull. This separates the question of adding a winning candidate from the original requirement of improving one source without losing on the other. The largest estimated gain remains below one AUROC point. The rounded upper interval endpoint of 1.000 does not establish a gain of at least one point.

An exploratory calculation from these means shows where each blend leads. The 0.75 blend leads every original CIFAR-10 candidate once CIFAR-100's test weight exceeds 0.381, with its maximum margin at 0.619. Including all new blends changes the comparison. At fixed MSP, the half blend is never best, the 0.75 blend leads for weights 0.381 to 0.772, the union for 0.772 to 0.959, and the 0.25 blend above 0.959. The latter two regions have margins below 0.03 points over their strongest rivals. These point-estimate regions carry no simultaneous uncertainty guarantee.

Dependence on the scorer. The CIFAR-100 exclusion in Table 7 applies to MSP. With another scorer fixed for every candidate the comparison can differ.

MNIST within the far-OOD average. The same amendment predicted lower MNIST AUROC and a higher far-OOD average for the quarter blend with CIFAR-100 ID. With MSP, the differences from vanilla are -11.71 and $+4.65$ points. Both directions hold on six of seven scorers at default settings and survive Holm correction under paired and Welch tests within each seven-scorer source family. KNN shows neither a detected conjunction nor a reversal. With Textures-selected settings, three of seven scorers confirm the conjunction and none reverses it, so that read-out does not meet the across-scorer rule. That read-out also excludes Textures from the far-OOD average because it was used for selection. The other three blends repeat the MSP conjunction. This confirmation concerns a measured-source loss within the average, without establishing a general benefit from blending.

C.2 The Blend's Discovery Analysis

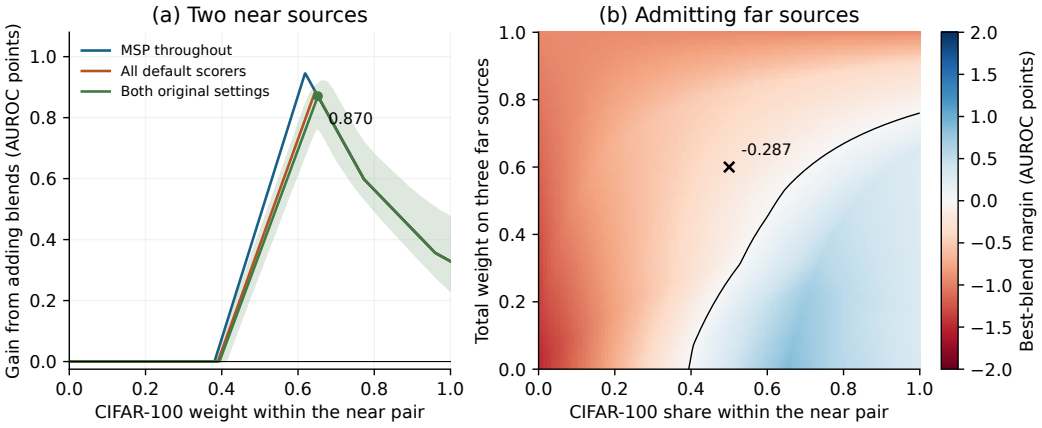


Figure 3: **A blend adds a small gain on the near pair.** CIFAR-10 is the in-distribution dataset. (a) Gain in area under the receiver operating characteristic curve (AUROC) from adding blends as the out-of-distribution (OOD) mixture varies between CIFAR-100 and TinyImageNet. MSP denotes maximum softmax probability. The strongest comparison gives originals seven scorers and both settings regimes. Shading is its pointwise 95% seed-bootstrap range. (b) Signed best-blend margin after admitting MNIST handwritten digits, Street View House Numbers (SVHN) and Places365, which share the far-source mass equally. The cross marks equal weights over all five sources, where the margin is -0.287 points. Panel (a) retains originals and clips negative margins to zero. Panel (b) shows point estimates.

For CIFAR-10 as ID, the original choices are vanilla and OE with either TinyImageNet-597 or 300K auxiliaries, under all seven scorers and both recorded settings regimes. We compare them with four blends of the two pools on the CIFAR-100 and TinyImageNet pair (Figure 3a). The blend with a 75 percent TinyImageNet share, scored with MSP, has a maximum margin of 0.870 points at CIFAR-100 test weight 0.652, with a nominal 95% bootstrap range of 0.814 to 0.937 and optimization repeated in each resample. Choosing the mixture and detectors on nine seeds and evaluating on the tenth gives 0.865 points, positive on every fold. The blend costs about 0.29 percentage points of ID accuracy against the original TinyImageNet-pool model. For CIFAR-100 as ID, where the originals also include both mixing-only controls, no blend improves on the near pair (margin -0.612, range

$[-0.680, -0.464]$). These expanded-choice analyses are exploratory, and Appendix C.1 preserves the earlier registered MSP-only comparison.

Admitting MNIST, SVHN and Places365 at equal weight over the five common sources turns the CIFAR-10 margin to -0.287 points, with range $[-0.439, -0.137]$, and the preferred scorers change with the reference (Figure 3b). On CIFAR-100 the five-source margin is $+0.387$ with range $[-0.628, +1.127]$. Appendix G reports both source families and the held-out-seed checks.

C.3 Auxiliary Images That Carry No Loss

A registered CIFAR-100 experiment trained three arms that differ only in what fills the unlabelled fifth of each batch: TinyImageNet-597 images, CIFAR-100 training images or nothing. It predicted that the TinyImageNet-filled arm would be higher on TinyImageNet and lower on MNIST on at least five of seven scorers. The label is REVERSED: against the CIFAR-100-filled arm, TinyImageNet AUROC rises on six of seven scorers, by 3.9 to 8.0 points, and MNIST also rises on six, by 6.5 to 12.9. Re-estimating each checkpoint's normalization statistics on either arm's mixture leaves the MSP contrasts nearly unchanged ($+3.89$ and $+6.50$ as trained, $+3.92$ to $+4.14$ and $+6.21$ to $+6.58$ recalibrated), so the stored batch statistics do not explain it, and the read-out does not isolate the training-time paths.

C.4 The ImageNet-200 Confirmation

The rows of Table 8 for ImageNet-200 were registered before their arms existed. The prediction, taken from the published sheet, is that OE lowers iNaturalist AUROC against vanilla under MSP and raises the near-OOD average, and the registered rule reads the primary label from the iNaturalist difference against the two-standard-error threshold of five seeds per arm. Training is OpenOOD's own ImageNet-200 recipe at 224×224 for 90 epochs, and evaluation is OpenOOD's evaluator through a script fixed by checksum before the first run. Before any arm was read, that script was run on OpenOOD's released ImageNet-200 vanilla checkpoint, which it must reproduce within one point of the published near-OOD 83.34 and far-OOD 90.13. It returns 83.42 and 90.19, and no source differs from its published mean by more than 0.18 points, which matters because the registered difference is 2.50 points on a single source and an aggregate tolerance would not protect it. The same script on the same checkpoint bytes on a second machine agrees to 0.0003 points on every source. Our five-seed ImageNet-200 vanilla training averages 81.80 near-OOD AUROC, which is 1.6 points below OpenOOD's released checkpoint (83.42) under the same stated configuration, and their ID accuracy is 84.41 against the published 86.37 percent. The cause of this gap remains unresolved, and it recurs on the five further runs below.

Both registered predictions hold. With MSP, iNaturalist AUROC falls from 91.93 to 88.33, a difference of -3.60 against a two-standard-error threshold of 0.32, and the near-OOD average rises from 81.80 to 83.82, $+2.02$ against 0.40. The iNaturalist loss appears on six scorers, with no detected change under KNN, and under the full-spectrum protocol, which includes shifted images of familiar classes in the ID group, the MSP difference is -12.05 against 1.24. Table 9 gives every source and both averages. With MSP, OE raises both near-OOD sources and lowers all three far-OOD sources, so the near-OOD average rises by 2.02 while the far-OOD average falls by 1.73, and the registered loss lies outside the average that rises. Applying the mixture identity to the two, OE stays ahead of vanilla on a mixture of the near-OOD average and iNaturalist while iNaturalist carries less than 0.36 of the weight (seed bootstrap 0.34 to 0.38), which is the crossover this paper's construction computes from per-source results alone. Here the two near-OOD sources retain equal weights within their group, and Textures and OpenImage-O receive zero weight. An exploratory calculation places the one-point OE deficit at iNaturalist weight 0.537 (95% seed-bootstrap interval 0.516 to 0.563), using the same 4,000 resamples as the planned crossover. At weight 0.70, the estimated deficit is 1.92 points. These boundaries use the standard ID distribution, separately from the full-spectrum read-out. Table 8 lists the registered predictions taken from the published sheet. Retraining confirms selected losses and reverses one published comparison, the retrospective CIFAR-10 MNIST loss, whose published difference was near zero against fifteen standard errors elsewhere.

Replication on new training runs. A second protocol, committed before any of its runs existed, trained five further vanilla and five further OE models (seeds 50 to 54) under the same recipe on two

Table 8: **Average gains and per-source losses tested in matched training.** Outlier Exposure (OE) minus cross-entropy (vanilla), maximum softmax probability (MSP) scorer, area under the receiver operating characteristic curve (AUROC) in points. Published: OpenOOD v1.5's three-seed means (Zhang et al., 2024), with the difference in units of its standard error computed from the published standard deviations. Matched: our runs, mean difference and the two-standard-error threshold of the registered rules. ResNet-18: CIFAR-100 uses 200 epochs and seeds 10–19, ImageNet-200 uses the OpenOOD recipe for 90 epochs and seeds 10–14, and CIFAR-10 uses seeds 0–9. The CIFAR-10 row is retrospective, and matched training reverses it. The other predictions were registered before training. *Abbreviations:* out-of-distribution (OOD); in-distribution (ID).

ID data	OOD source	Published (s.e. units)	Matched (threshold)	Verdict
CIFAR-100	CIFAR-10	-1.77 (-15.1)	-0.61 (0.26)	confirmed (seven scorers)
ImageNet-200	iNaturalist	-2.50 (-14.6)	-3.60 (0.32)	confirmed
ImageNet-200	near-OOD average	$+1.51$ ($+17.3$)	$+2.02$ (0.40)	as predicted
CIFAR-10	MNIST	-2.41 (-2.0)	$+1.98$ (0.77)	not reproduced (no registered rule)

Table 9: **The registered ImageNet-200 confirmation, every source.** Outlier Exposure (OE) and vanilla cross-entropy under OpenOOD's ImageNet-200 recipe at 224×224, 5 and 5 seeds, area under the receiver operating characteristic curve (AUROC) in points with the maximum softmax probability (MSP) scorer. Δ is OE minus vanilla and noise is the registered $2\sqrt{s_a^2/n + s_b^2/n}$ threshold. The registered primary label is read from iNaturalist (Appendix C.4), and the other rows are reported without a test. Our five-seed ImageNet-200 vanilla training averages 81.80 near-OOD AUROC, which is 1.6 points below OpenOOD's released checkpoint (83.42) under the identical configuration. *Abbreviations:* out-of-distribution (OOD); in-distribution (ID).

Source	Vanilla	OE	Δ	noise	counts
SSB-hard	$78.89_{\pm0.18}$	$81.49_{\pm0.37}$	$+2.60$	0.37	yes
NINCO	$84.72_{\pm0.45}$	$86.16_{\pm0.29}$	$+1.43$	0.47	yes
Near-OOD average	$81.80_{\pm0.31}$	$83.82_{\pm0.32}$	$+2.02$	0.40	yes
iNaturalist	$91.93_{\pm0.08}$	$88.33_{\pm0.35}$	-3.60	0.32	yes
Textures	$87.47_{\pm0.60}$	$86.45_{\pm0.42}$	-1.02	0.66	yes
OpenImage-O	$88.02_{\pm0.33}$	$87.46_{\pm0.33}$	-0.56	0.42	yes
Far-OOD average	$89.14_{\pm0.30}$	$87.41_{\pm0.26}$	-1.73	0.35	yes
ID accuracy	84.41	84.22	-0.19		

computing platforms. It froze OE with MSP as the new choice against vanilla under all seven scorers and fixed the requirement of Appendix H: on the equal SSB-hard/NINCO pair, every gain's one-sided 95% lower bound above 0.5 AUROC points and the ID-accuracy cost's upper bound below 0.5 percentage points, with Welch bounds over training runs. These are new training runs on the existing evaluation images, so they test the choice across runs and not on separate images. Table 10 gives the seven comparisons and every source. The registered pattern recurs: with MSP, OE lowers iNaturalist AUROC by 3.31 points (two-standard-error threshold 0.44; per run -2.77 to -4.00) and raises the near pair by 2.06 (lower bound 1.82, above the requirement), and it also lowers Textures by 1.42 and OpenImage-O by 0.65. Every OE run is above every vanilla run with MSP on both near-OOD sources and below every vanilla run on the three far-OOD sources. Against vanilla's best near-pair scorer, Scale, the lead is 0.44 points with bounds 0.28 to 0.60, and ID accuracy is 83.94 against 84.20 percent, a cost of 0.26 with upper bound 0.56. The requirement therefore fails on both parts, and the Scale upper bound does not place the gain below 0.5 either, so the test leaves the practical benefit unestablished rather than reversed. On the iNaturalist path against every vanilla scorer, OE's lead ends at iNaturalist weight 0.061 (image-bootstrap 2.5th to 97.5th percentiles 0.016 to 0.102) and a one-point deficit arrives at 0.198 (0.154 to 0.239). At equal weight over the five sources OE trails vanilla with MSP by 0.25 points and vanilla with Scale by 3.56, range $[-3.86, -3.25]$. At

thresholds calibrated on separate ID images to 95 percent acceptance, OE with MSP accepts 66.2 percent of near-OOD images against 72.1 for vanilla with MSP and 62.6 for vanilla with Scale. The 3.6-point excess over Scale has image-bootstrap range $[0.55, 6.54]$, and the range includes zero once training runs are also resampled ($[-0.52, 7.46]$). The five new vanilla runs reach 84.20 percent ID accuracy on both platforms, so the gap to the published 86.37 recurs across runs and platforms and is not seed variation. Per-run values, the bootstrap arrays and the protocol are in the supplement.

Table 10: **Replication of the ImageNet-200 choice on five further training runs per arm.** Seeds 50 to 54, with the protocol committed before training. Outlier Exposure (OE) with maximum softmax probability (MSP) is the frozen new choice, against vanilla cross-entropy under every scorer on the equal SSB-hard/NINCO pair. (a) Near-pair gains in area under the receiver operating characteristic curve (AUROC), means over runs with one-sided 95% Welch bounds. The requirement fixed before the runs asks every lower bound to exceed 0.5 and the in-distribution (ID) accuracy cost's upper bound to stay below 0.5. (b) AUROC per out-of-distribution source for vanilla and OE with MSP, means over runs, with the two-standard-error threshold of the paired difference. New training runs on the existing evaluation images.

(a)

Vanilla scorer	Vanilla near pair	OE (MSP) minus vanilla	Bounds
MSP	81.66	+2.06	[1.82, 2.30]
Energy	80.74	+2.98	[2.78, 3.18]
ODIN	78.61	+5.11	[4.88, 5.33]
ReAct	79.43	+4.29	[3.98, 4.60]
Scale	83.27	+0.44	[0.28, 0.60]
ASH	79.51	+4.21	[4.06, 4.35]
KNN	79.13	+4.59	[4.23, 4.94]
ID accuracy (%), vanilla and OE	84.20	83.94	cost 0.26, upper 0.56

(b)

Source	Vanilla (MSP)	OE (MSP)	Difference (2 s.e.)
SSB-hard	78.75	81.56	+2.82 (0.24)
NINCO	84.56	85.87	+1.30 (0.29)
iNaturalist	91.59	88.28	−3.31 (0.44)
Textures	87.72	86.30	−1.42 (0.48)
OpenImage-O	87.95	87.30	−0.65 (0.29)
Near pair	81.66	83.72	+2.06 (0.26)
Equal five sources	86.11	85.86	−0.25 (0.26)

D Numerical Results and Sensitivity Analyses

D.1 Scorer Selection

The scorer grids are: ReAct searches the $\{85, 90, 95, 99\}$th activation percentiles, ASH and Scale search $\{65, 70, \ldots, 95\}$, KNN searches $k \in \{50, 100, 200, 500, 1000\}$, and ODIN searches temperatures $\{1, 10, 100, 1000\}$ and perturbation magnitudes $\{0.0014, 0.0028\}$.

Per-run selection can amplify apparent differences between detectors. On cluster A, OE's selected Scale percentile ranges from 65 to 95 across seeds, and its Scale AUROC is 92.50 ± 3.17 under per-run selection against 95.07 ± 0.20 at a fixed 65th percentile. Selecting the ReAct, ASH and Scale percentiles and KNN's k on Textures, which is disjoint from the auxiliary images and both test sets, changes results on the same checkpoints. This pattern was first seen on cluster A and then predicted for cluster B, where OE's Scale AUROC is 95.02 ± 0.14 with Textures selection and 92.40 ± 3.95 with TinyImageNet selection (Figure 4). Textures changes both source and difficulty, which prevents an attribution to provenance. At fixed checkpoints, 12 ordered scorer pairs combine a higher near-OOD average with lower AUROC on one constituent source, both beyond the threshold.

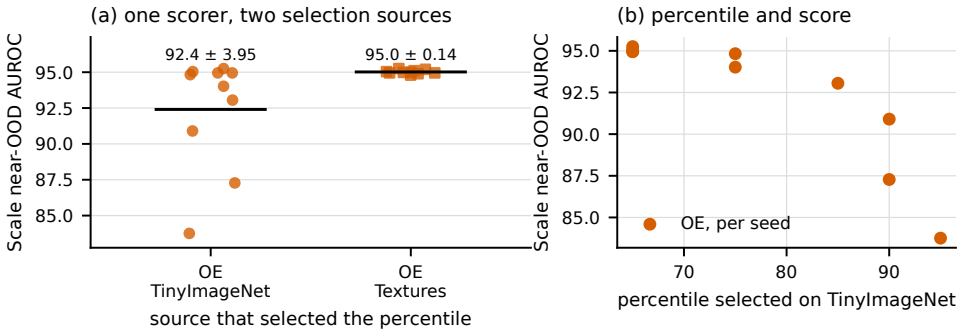


Figure 4: **Scorer selection changes both the mean and the spread.** CIFAR-10, ResNet-18, computing environment B, ten runs per training configuration, fixed checkpoints. (a) Scale's area under the receiver operating characteristic curve (AUROC), averaged over the near out-of-distribution (OOD) sets CIFAR-100 and TinyImageNet, for Outlier Exposure (OE). Percentiles are selected on TinyImageNet validation data or on Textures, which is disjoint from the auxiliary images and both test sets. Annotations give seed means and sample standard deviations, and horizontal lines mark means. (b) Selected percentiles and near-OOD AUROC under TinyImageNet selection. The figure is a post hoc read-out of saved features.

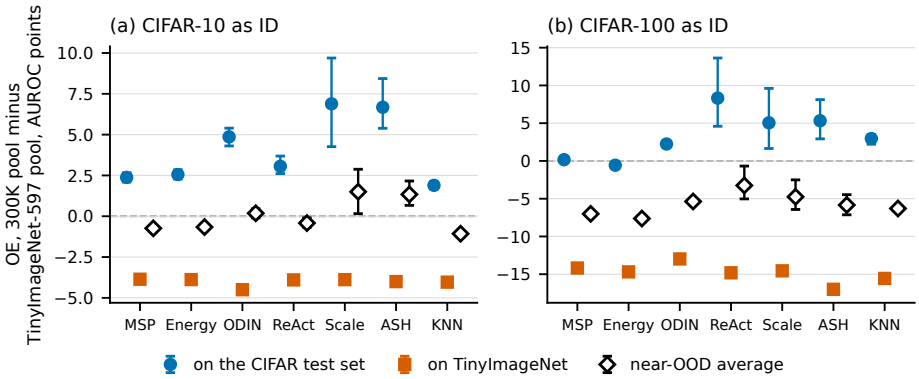


Figure 5: **Replacing the auxiliary pool moves the two per-set results in opposite directions.** ResNet-18, ten runs per training configuration. Points show Outlier Exposure (OE) trained with a size-matched subsample of 300K Random Images minus OE trained with TinyImageNet-597, in points of area under the receiver operating characteristic curve (AUROC), on the CIFAR test set (filled circles, CIFAR-100 in panel a and CIFAR-10 in panel b), on TinyImageNet (filled squares) and on their average (open diamonds). Intervals use 95% paired resampling of runs for each comparison separately. Panels use different vertical ranges. (a) CIFAR-10 as in-distribution (ID) data, seeds 0 to 9. (b) CIFAR-100 as ID, seeds 10 to 19. Seeds, exposure and the selection procedure are shared within a panel (Section 4.2, Appendix D.2). *Abbreviations:* out-of-distribution (OOD).

D.2 The Auxiliary-Pool Intervention

Table 11 gives the registered intervention of §4.2. The first 300K subsample holds 29,850 images, the size of TinyImageNet-597, drawn with a fixed seed. We audited it against the evaluation images at 32×32 pixels. No auxiliary image is an exact duplicate of a CIFAR-100 or CIFAR-10 test or validation image, one is a near-duplicate of a CIFAR-100 test image (per-pixel root mean squared error (RMSE) 0.6 on a 0 to 255 scale) and no other lies within RMSE 10 of any of them. This audit finds little pixel-level overlap, but does not isolate the effect of the near-duplicate or test semantic similarity. TinyImageNet-597 has no image within RMSE 6 of a TinyImageNet test image. The registered labels are a source effect for OE on TinyImageNet and on CIFAR-100 and a source effect in OE's readable ratio (Appendix C), which on TinyImageNet is 0.016 with TinyImageNet-597 and

0.112 with the other pool. The labels for the other registered arms are in Table 6. Every label is unchanged at Textures-selected settings. A second draw of 29,850 images from the 270,150 images of 300K Random Images outside the first, registered before its runs existed on the same cluster, commit and seeds, replicates the source effect on both sets, and OE differs between the two draws on no scorer of either set. On far-OOD sets OE with the 300K subsample is lower than OE with TinyImageNet-597 on six scorers, by 0.9 to 1.5 points.

Table 11: **Registered auxiliary-source intervention**, CIFAR-10, ResNet-18, third computing environment, n=10 per training configuration, area under the receiver operating characteristic curve (AUROC) per near out-of-distribution (OOD) set, means over seeds. TIN-597 is TinyImageNet-597 and 300K a size-matched subsample (29,850 images) of 300K Random Images. Ratio is the readable ratio, defined in the text below, on CIFAR-100 and on TinyImageNet. All eleven arms passed every validation check, and the scoring job reproduced the registered computing environment A rows exactly. *Abbreviations:* in-distribution (ID); Outlier Exposure (OE); vanilla cross-entropy (VAN); maximum softmax probability (MSP); activation shaping (ASH); nearest-neighbour scoring (KNN).

		CIFAR-100							TinyImageNet								Ratio	
Arm	Auxiliary	MSP	Energy	ODIN	ReAct	Scale	ASH	KNN	MSP	Energy	ODIN	ReAct	Scale	ASH	KNN	ID acc	C-100	TIN
VAN	none	88.30	88.62	83.73	87.80	87.13	71.93	89.83	90.11	90.77	85.28	90.19	89.10	73.86	91.67	95.34	0.765	0.737
OE	TIN-597	90.99	90.72	87.49	90.15	86.36	86.26	91.34	99.71	99.73	99.70	99.72	99.72	99.70	99.54	95.28	0.397	0.016
OE	300K	93.37	93.27	92.35	93.22	93.25	92.93	93.23	95.85	95.85	95.20	95.82	95.84	95.70	95.51	94.99	0.184	0.112

The pool that shrinks one loss creates another. With the 300K pool, OE on CIFAR-100 falls below vanilla on MNIST on all seven scorers, by 1.5 to 28.0 points. With TinyImageNet-597 it is above vanilla on four scorers beyond the threshold and below on KNN alone, so the pool switch creates the loss on six scorers and deepens it on one. Its far-OOD average rises on six, by up to 4.5. This loss was found in an exploratory reading and then predicted for the blended pools before any blend was evaluated (Appendix C.1). The intervention establishes sensitivity to the auxiliary pool. It does not separate data origin from class content, diversity or image history, which the two pools change at once (Section 3.2), and archived arms trained with Places365 or SVHN as the pool leave OE below vanilla on CIFAR-10 on every scorer (Appendix F). Pool and scorer choices move the CIFAR-100 crossover coordinates differently across scorers (Figure 6).

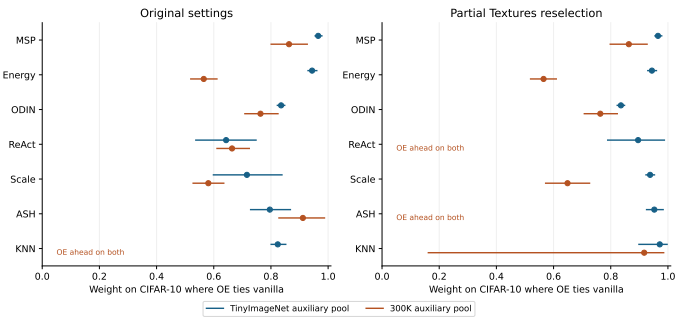


Figure 6: **Pool and scorer choices change crossover coordinates.** CIFAR-100 is the in-distribution dataset, with ResNet-18. Points mark the CIFAR-10 out-of-distribution weight at which Outlier Exposure (OE) ties vanilla on the CIFAR-10/TinyImageNet pair. Bars are conditional 95% paired seed-bootstrap ranges among resamples with a finite crossover. These values are coordinates on the CIFAR-10 weight axis and do not measure distance from equal weights. The right panel reselects ReAct, ASH, Scale and KNN on Textures. MSP and Energy are untuned, and ODIN retains its original setting. *OE ahead on both* denotes no crossover in the point estimates. *Abbreviations:* maximum softmax probability (MSP); activation shaping (ASH); nearest-neighbour scoring (KNN).

D.3 Selection Regret

The analysis plan was registered before the regret was computed. Candidates are the arms of a benchmark crossed with the seven scorers, each at TinyImageNet-selected settings and, on the two

CIFAR-10 benchmarks, at Textures-selected settings. On cluster B the candidates are OE, NT-OE, NT-OE_0 and OE+NT (the mixing-only control has no Textures-selected rows). On the third cluster they are the eleven registered arms of the pool intervention (Appendix D.2) other than the zero target, which a one-point ID-accuracy floor removes, and on CIFAR-100 the five matched arms (vanilla, OE, the two norm targets and OE+NT). A procedure picks, per seed, the candidate that maximizes its criterion, and its regret is the target score of the best candidate over both settings regimes minus that of its pick. The far-OOD target is the mean AUROC over MNIST, SVHN and Places365, which no procedure selects on. The far-OOD means of the registered arms had been reported before this plan was written, so the target was withheld from the procedures and not from the authors. Table 12 gives the mean regret per procedure with the seed standard deviation. The registered contrasts are three. The per-set worst case with Textures-selected settings against the average with TinyImageNet-selected settings shows no difference on cluster B (mean difference $+0.13$ with noise 0.51) and is worse on the third cluster ($+2.20$ with noise 0.70). The worst case at TinyImageNet-selected settings receives the same labels ($+0.13$ and $+1.99$), and the average at Textures-selected settings makes identical picks on both clusters. On CIFAR-100 the worst case against the average is no difference on the far-OOD target ($+1.17$ with noise 1.85) and a gain on CIFAR-10 as OOD (-1.29 with noise 0.32), where the worst case is that set for every auxiliary arm. Per-seed selections and the best candidates on each evaluation target are in the supplement. With the 300K pool, the arms the worst-set rule selects on the third cluster are balanced across the near-OOD sets but lower on far-OOD sets. Scored per seed by the worst of the three far-OOD sources, the criterion a worst-set rule targets, the worst-set rule is again no better than the average rule. The paired difference is -0.08 (two standard errors 1.41) on cluster B, $+4.56$ (2.10) on the third cluster, where it is worse on every seed, and -0.50 (3.86) on CIFAR-100 (a read-out computed after the verdict).

Table 12: Selection regret is the best candidate's performance minus the selected candidate's on the evaluation target. Mean selection regret in area under the receiver operating characteristic curve (AUROC) points, with the sample standard deviation over ten seeds, for four procedures on three benchmarks. The target is the far out-of-distribution (OOD) mean over MNIST, SVHN and Places365 unless stated. Textures-selected settings exist on the CIFAR-10 benchmarks only. The worst-case rule's zero regret with CIFAR-10 as the target is not independent evidence, since that rule then selects on the target.

Procedure	CIFAR-10, cluster B	CIFAR-10, third cluster	CIFAR-100	CIFAR-100, CIFAR-10 as OOD
average, TinyImageNet-selected	$1.27_{\pm.84}$	$1.04_{\pm.30}$	$7.88_{\pm1.56}$	$1.29_{\pm.50}$
worst case, TinyImageNet-selected	$1.41_{\pm.43}$	$3.03_{\pm1.05}$	$9.05_{\pm2.70}$	$0.00_{\pm.00}$
average, Textures-selected	$1.27_{\pm.84}$	$1.04_{\pm.30}$		
worst case, Textures-selected	$1.41_{\pm.43}$	$3.24_{\pm1.01}$		

D.4 CIFAR-100 Confirmation

Table 13 gives the registered confirmation of §4.1, registered after the matched comparison's verdict and before any of its runs existed. Its arms are vanilla, OE and a mixing-only control with TinyImageNet-597, and OE and the control with the 300K subsample, ten new seeds each on the third cluster, from the code commit of the pool intervention. Every validation check passed for all five arms, the matched comparison's vanilla and OE rows reproduced exactly in the same scoring job, and the frozen verdict script was run on the rows that job wrote (run records in the supplement). The prediction, OE below vanilla on CIFAR-10 as OOD, is confirmed on all seven scorers, where OE loses 0.6 (MSP) to 9.4 (ReAct) points on CIFAR-10, and at Textures-selected settings, which shrink OE's loss to 0.6–3.1 on five scorers. The largest remaining loss, ODIN's, is at settings that this read-out cannot re-select from saved features, and the largest among the four re-selected scorers is 1.2 (ASH). The two contrasts registered to decide an attribution are the control against vanilla (three scorers higher, none lower) and OE against the control (all seven lower), and the registration writes an attribution only when one of them is null. With the 300K subsample OE's TinyImageNet gain falls by 13 to 17 points, and its CIFAR-10 AUROC rises on five scorers, by up to 8.3 points (ReAct), and falls on Energy by 0.6, so the response to the pool is not uniform. The rule yields no pool effect for CIFAR-10 as OOD, because Energy moves against the other five. OE is then below vanilla on CIFAR-10 on five scorers under the registered selection and mixed at Textures-selected settings, and it costs ID accuracy (77.20 against 77.63). The weights on CIFAR-10 as OOD beyond

which vanilla leads OE are 0.97 for MSP, 0.94 for Energy, 0.84 for ODIN, 0.64 for ReAct, 0.72 for Scale, 0.80 for ASH and 0.82 for KNN, with 95% seed-bootstrap intervals $[0.95, 0.98]$, $[0.93, 0.96]$, $[0.82, 0.85]$, $[0.54, 0.75]$, $[0.60, 0.84]$, $[0.73, 0.87]$ and $[0.80, 0.85]$.

Table 13: **On ten new seeds, as predicted, Outlier Exposure (OE) with TinyImageNet-597 gains on the near-OOD average and loses on CIFAR-10 as out-of-distribution (OOD) inputs.** CIFAR-100 as in-distribution (ID), ResNet-18, 200 epochs from scratch, third computing environment, means over seeds 10 to 19. All five arms share the trainer, budget and seeds, and the four auxiliary arms replace the same fraction of ID rows per epoch. Area under the receiver operating characteristic curve (AUROC) and ID accuracy are in percent. TIN-597 is TinyImageNet-597 and 300K is the size-matched subsample of 300K Random Images. VAN is vanilla cross-entropy and MIX the mixing-only control. Every OE–VAN difference with TIN-597 in both panels survives Holm correction under paired and Welch tests. Per-seed rows and their standard deviations are in the supplement. Details are in Appendix D.4. *Abbreviations:* maximum softmax probability (MSP); activation shaping (ASH); nearest-neighbour scoring (KNN).

Scorer Auxiliary pool	VAN none	OE TIN-597	MIX TIN-597	OE 300K	MIX 300K
(a) CIFAR-10 as OOD					
MSP	79.18	78.57	79.45	78.74	79.16
Energy	79.53	78.54	79.59	77.98	79.15
ODIN	78.72	75.62	79.06	77.87	78.53
ReAct	78.78	69.34	78.91	77.66	78.38
Scale	79.75	73.09	79.85	78.15	79.51
ASH	75.11	69.19	76.79	74.52	75.70
KNN	77.33	74.16	77.28	77.12	77.24
(b) TinyImageNet as OOD					
MSP	82.90	99.89	86.37	85.71	84.33
Energy	83.22	99.93	87.45	85.24	84.74
ODIN	81.99	97.70	86.37	84.72	83.49
ReAct	82.91	99.93	87.31	85.12	84.46
Scale	83.18	99.95	86.93	85.40	84.65
ASH	76.85	99.95	84.95	82.95	80.84
KNN	83.56	98.40	83.43	82.84	83.54
ID accuracy	77.63	77.73	77.64	77.20	77.61

Our auxiliary exposure differs from OpenOOD's OE trainer, which pairs every ID batch with an auxiliary batch. The loss also appears under OpenOOD's recipe: the published table shows it with MSP, and archived runs with OpenOOD's trainer show it on all seven scorers, by 1.2 to 9.6 points (Appendix F).

The loss survives giving each objective its best scorer. Even when an oracle selects the optimal scorer for each objective by the test average, Energy for OE gains 7.8 points over Scale for vanilla on the near-OOD average and loses 1.2 on CIFAR-10 as OOD, on every confirmation seed. At 95% ID acceptance, these detectors accept 83.40% and 78.91% of CIFAR-10 outliers, respectively. OE's false-positive rate is higher on all ten seeds, by 4.49 points against a threshold of 0.64, or about 45 extra false acceptances per $1{,}000$ outliers from this measured source. With the same scorer for both objectives, OE's excess false-positive rate is 1.0 to 10.6 points across all seven scorers.

D.5 Second Architecture

Table 14 gives the registered replication of §4.1 on WideResNet-40-2 (depth 40, width 2, no dropout, 128-dimensional features, 2.2 million parameters), registered before any of its runs existed. Each config is the registered ResNet-18 config of the same arm and seed with the network line changed and nothing else, asserted by the generator, and the decision rules are the ResNet-18 rules with the arm names substituted. The verdict script reproduced all 18 registered ResNet-18 labels in a self-test before any run was extracted. All eleven arms passed every validation check, including a gate on the exact parameter count in every run log, and the scoring job reproduced the registered ResNet-18

rows exactly. The verdict of record comes from the frozen script run on the complete, checksum-matched run logs, and no label changed from an earlier partial read (run records in the supplement). The registered labels are: the CIFAR-100 loss confirmed, with TinyImageNet and the near-OOD average as predicted; no mixing effect on CIFAR-10 as OOD and OE's term lowering it, for which the registered rule writes an attribution to OE's loss term on this network; a pool effect for OE on CIFAR-100 on both sets, with OE with the 300K subsample against vanilla on CIFAR-10 mixed; and a source effect on both sets on CIFAR-10. The labels for the other arms are in Table 6. ID accuracy is 75.73 for vanilla and 75.50 for OE on CIFAR-100, and 95.03 and 94.83 on CIFAR-10. The constants and the recipe were set on ResNet-18 and were not tuned for this network.

Control comparisons. The rule counts no mixing effect when fewer than two scorers differ beyond the threshold, and mixed otherwise. On ResNet-18 three scorers differ (mixed, no attribution), and on WideResNet-40-2 one (KNN, -0.58), so the rule counts no mixing effect and records an attribution to OE's loss term. The contrast shows an incremental effect of the term beyond mixing on six scorers. On KNN the term raises AUROC by 0.21 while the full configuration stays 0.37 below vanilla. OE is below the control on six scorers, by up to 9.6 points, and 0.21 above it on KNN (Appendix Table 14).

Table 14: **The per-set findings replicate on WideResNet-40-2.** Differences in area under the receiver operating characteristic curve (AUROC), in points, ten runs per training configuration, the registered ResNet-18 recipe, constants and decision rules carried over without tuning (network changed only). OE is Outlier Exposure, VAN vanilla cross-entropy, MIX the mixing-only control and OOD out-of-distribution. *Beyond the two-standard-error threshold. †Beyond the threshold but not surviving Holm correction at 5% under both paired and Welch tests. Registered labels: the CIFAR-100 loss confirmed, no mixing effect on CIFAR-10 with OE's term lowering it, a pool effect on both CIFAR-100 near-OOD sets and a source effect on both sets. These labels are unchanged at Textures-selected settings, and the labels of the other registered arms are in Table 6. *Abbreviations:* in-distribution (ID); maximum softmax probability (MSP); activation shaping (ASH); nearest-neighbour scoring (KNN).

Difference	MSP	Energy	ODIN	ReAct	Scale	ASH	KNN
(a) CIFAR-100 as ID, seeds 10 to 19							
OE − VAN, CIFAR-10 as OOD	-2.43^{*}	-4.41^{*}	-4.53^{*}	-6.55^{*}	-9.61^{*}	-6.76^{*}	$-0.37^{*\dagger}$
OE − VAN, TinyImageNet	$+20.82^{*}$	$+20.75^{*}$	$+22.78^{*}$	$+21.35^{*}$	$+21.79^{*}$	$+37.83^{*}$	$+15.87^{*}$
OE − VAN, near-OOD average	$+9.19^{*}$	$+8.17^{*}$	$+9.12^{*}$	$+7.40^{*}$	$+6.09^{*}$	$+15.54^{*}$	$+7.75^{*}$
MIX − VAN, CIFAR-10 as OOD	-0.05	-0.22	-0.32	-0.77	-0.03	$+0.59$	-0.58^{*}
OE − MIX, CIFAR-10 as OOD	-2.38^{*}	-4.19^{*}	-4.21^{*}	-5.78^{*}	-9.58^{*}	-7.35^{*}	$+0.21$
OE with 300K − OE with TinyImageNet-597, TinyImageNet	-16.64^{*}	-17.55^{*}	-17.01^{*}	-17.62^{*}	-17.44^{*}	-19.78^{*}	-15.60^{*}
OE with 300K − OE with TinyImageNet-597, CIFAR-10 as OOD	$+1.62^{*}$	$+1.69^{*}$	$+4.25^{*}$	$+4.68^{*\dagger}$	$+7.92^{*}$	$+14.65^{*}$	-0.07
(b) CIFAR-10 as ID, OE with the 300K subsample minus OE with TinyImageNet-597							
TinyImageNet	-4.50^{*}	-4.59^{*}	-5.44^{*}	-4.66^{*}	-4.58^{*}	-5.11^{*}	-4.41^{*}
CIFAR-100	$+3.57^{*}$	$+4.14^{*}$	$+9.11^{*}$	$+5.74^{*}$	$+5.01^{*}$	$+13.25^{*}$	$+2.20^{*}$

Per-set FPR@95. Table 15 repeats the CIFAR-100 per-set result at a 95% ID true-positive rate. It is a read-out computed after the registered verdicts from the saved features with the pinned scorers, whose per-set AUROC and selected settings reproduce the registered rows exactly. On CIFAR-100, OE's false-positive rate on CIFAR-10 as OOD exceeds vanilla's on all seven scorers, by 1.0 to 10.6 points (0.5 to 6.3 with Textures-selected settings), and falls by 64 to 76 points on TinyImageNet. Each objective's best scorer by the near-OOD AUROC average, Energy for OE and Scale for vanilla, gives a CIFAR-10 difference of 4.5 points (noise 0.64), higher for OE on every seed.

D.6 A Training Objective Without Auxiliary Data

This replication asks whether the source inconsistency recurs with a training objective that uses no auxiliary images. It compares LogitNorm (Wei et al., 2022), with temperature 0.04, against cross-entropy. Both arms are trained from scratch with OpenOOD's pinned implementation, ResNet-18 at 32×32 pixels and 400 epochs, on five new seeds per arm with CIFAR-100 and with CIFAR-10 as ID data. The protocol was committed before any training run. Its two claims come from the archived runs of Appendix F and the published CIFAR-10 table, so LogitNorm was selected because it showed the pattern there, and the thresholds were set with that evidence known. The first claim is that on CIFAR-100, with MSP, LogitNorm raises MNIST AUROC by more than one point

Table 15: **The CIFAR-100 per-set loss persists at a 95% in-distribution (ID) true-positive rate.** Difference in false-positive rate (FPR@95), the percentage of out-of-distribution (OOD) inputs accepted as ID at the fifth percentile of ID-test confidence, in points. Lower is better, so a positive entry is a loss for the first arm. ResNet-18, ten runs per training configuration, TinyImageNet-597 auxiliary pool, TinyImageNet-selected scorer settings unless marked. OE is Outlier Exposure. Both arms share the trainer, budget and seeds. *Beyond the two-standard-error threshold. †Beyond the threshold but not surviving Holm correction at 5% under both paired and Welch tests. This read-out was computed after the registered verdicts from saved features. Its per-set area under the receiver operating characteristic curve (AUROC) reproduces the registered rows exactly. *Abbreviations:* maximum softmax probability (MSP); activation shaping (ASH); nearest-neighbour scoring (KNN).

FPR@95 difference	MSP	Energy	ODIN	ReAct	Scale	ASH	KNN
CIFAR-100, OE minus vanilla, seeds 10 to 19							
CIFAR-10 as OOD	$+2.3^{*}$	$+3.9^{*}$	$+4.3^{*}$	$+7.4^{*}$	$+8.6^{*}$	$+10.6^{*}$	$+1.0^{*\dagger}$
Textures-selected	$+2.3^{*}$	$+3.9^{*}$	$+4.3^{*}$	$+4.1^{*}$	$+3.9^{*}$	$+6.3^{*}$	$+0.5^{*\dagger}$
TinyImageNet	-73.3^{*}	-73.7^{*}	-63.9^{*}	-73.7^{*}	-73.6^{*}	-76.4^{*}	-65.4^{*}

and lowers CIFAR-10 by more than one point. The second is that on CIFAR-10 it raises both near sources by more than half a point. Each component is a one-sided Welch test against its threshold, a claim holds when both of its components do, and Holm's correction runs across the two claims. All twenty runs completed 400 epochs and cleared their validation-accuracy floors of 65 percent on CIFAR-100 and 90 percent on CIFAR-10. Evaluation uses OpenOOD's existing image lists, so these are new training runs on existing images.

Both claims hold (Table 16). On CIFAR-100, MNIST rises by 11.13 points (one-sided 95% lower bound 7.62) and CIFAR-10 falls by 4.21 (lower bound on the loss 3.78), with Holm-adjusted $p = 0.00036$. On CIFAR-10, CIFAR-100 rises by 3.04 (lower bound 2.77) and TinyImageNet by 3.64 (lower bound 3.42), with Holm-adjusted $p = 3.1\times10^{-6}$. Paired-by-seed tests give the same verdicts. Every LogitNorm run lies above every cross-entropy run on MNIST and below every one on CIFAR-10 and on Textures, whose loss of 4.81 points the protocol named as a secondary direction. On CIFAR-100 the near-OOD average falls by 2.21 points and the equal six-source average rises by 0.39, and LogitNorm costs 2.41 points of ID accuracy (0.62 on CIFAR-10). On the path from MNIST to CIFAR-10, LogitNorm with MSP leads cross-entropy with MSP until CIFAR-10 carries 0.73 of the weight (seed-bootstrap range $[0.64, 0.78]$). On CIFAR-10 it leads on every source, so no mixture of them reverses it. A separate implementation recomputed every AUROC from the saved per-image scores and reproduced both verdicts.

The study also scored all twenty runs with the six other scorers at OpenOOD's validation settings. Its automated secondary analysis stopped at a check that every scorer of a model reports the same ID accuracy. In three of the twenty models, separate scoring jobs of the same checkpoint differ on one to three of the $9{,}000$ saved classifier predictions, which changes two models' ID accuracy. A repair script, run after the analysis cutoff, completed the aggregation with each scorer's accuracy recorded separately, and it changed no frozen analysis source, threshold, checkpoint or saved score. The separate implementation recomputed all 140 profiles from the saved scores and agrees with the repaired aggregation. With CIFAR-100 as ID data, LogitNorm is below cross-entropy on CIFAR-10 under all seven scorers, by 4.21 to 9.63 points with no overlap between the arms, and above it on MNIST under all seven, by 4.33 to 20.87 points. Its best near-pair AUROC, 79.46 with MSP, is 2.60 points below the best original choice, cross-entropy with Scale at 82.06, so it fails the first condition of Definition 1 at the equal near pair. With CIFAR-10 as ID data, LogitNorm with MSP beats cross-entropy under every scorer at these settings on each of the six sources, by 1.52 to 4.33 points, so no mixture of them reverses its lead. The third condition was not tested for LogitNorm.

D.7 Decision Boundaries of Every Matched Comparison

Table 17 applies the construction of §4.3 to every matched comparison of OE with vanilla, from the saved per-seed AUROCs (a read-out computed after the registered verdicts). With d_{loss} the OE-minus-vanilla difference on the source where OE loses and d_{gain} the difference on the other

Table 16: **Without auxiliary images, LogitNorm gains on MNIST and loses on CIFAR-10 and Textures with CIFAR-100 as ID data, and gains on every source with CIFAR-10.** Five new training runs per arm of ResNet-18, trained for 400 epochs with OpenOOD's recipe and scored with maximum softmax probability (MSP). Entries are mean area under the receiver operating characteristic curve (AUROC) in percent over runs for each out-of-distribution (OOD) source. Differences are LogitNorm minus cross-entropy (CE), with two Welch standard errors in parentheses, and the four differences that the two prespecified claims test are in bold. "Other CIFAR" is CIFAR-10 when CIFAR-100 is the in-distribution (ID) dataset and CIFAR-100 when CIFAR-10 is. The last row is ID accuracy in percent.

	CIFAR-100 ID			CIFAR-10 ID		
OOD source	CE	LogitNorm	Difference	CE	LogitNorm	Difference
Other CIFAR	79.72	75.51	$\mathbf{-4.21}$ (0.43)	89.02	92.06	$\mathbf{+3.04}$ (0.28)
TinyImageNet	83.62	83.41	-0.21 (0.28)	91.09	94.73	$\mathbf{+3.64}$ (0.23)
MNIST	78.31	89.44	$\mathbf{+11.13}$ (3.76)	94.08	99.42	$+5.34$ (0.47)
SVHN	82.43	83.02	$+0.59$ (4.03)	93.65	98.62	$+4.97$ (1.15)
Textures	79.03	74.22	-4.81 (1.62)	92.16	95.62	$+3.46$ (0.60)
Places365	80.65	80.49	-0.15 (0.49)	90.96	95.34	$+4.37$ (0.58)
ID accuracy	78.76	76.34	-2.41	95.63	95.00	-0.62

near-OOD source, vanilla ties OE when the loss source carries weight $d_{\text{gain}}/(d_{\text{gain}} - d_{\text{loss}})$, and OE trails by one point at weight $(d_{\text{gain}} + 1)/(d_{\text{gain}} - d_{\text{loss}})$. Both are computed from seed means with a paired seed bootstrap of $2{,}000$ replicates, and again from each seed's own differences. A comparison has no boundary when the two differences share a sign. On CIFAR-100 as ID every scorer has a boundary with TinyImageNet-597 on both networks, at 0.64 to 0.97 on ResNet-18 and 0.69 to 0.98 on WideResNet-40-2, and the bootstrap intervals stay within 0.03 of the mean for the logit scorers and KNN and reach 0.12 for ReAct and Scale. Per seed the boundaries spread further: for ReAct they range from 0.36 to above 1, and one seed each for MSP, Energy and ReAct shows no CIFAR-10 loss at all. A boundary computed from seed means therefore describes the mean detector, and single training runs can sit on either side of it. With the 300K pool the boundaries fall to 0.57 to 0.91 and KNN has none. With CIFAR-10 as ID, OE leads on both near-OOD sources on every scorer for three of the four comparisons, and on six of seven for ResNet-18 with TinyImageNet-597, where Scale's boundary at 0.93 rests on four of ten seeds. On ImageNet-200, iNaturalist against the equal-weight near pair reproduces the boundaries of §4.1, 0.36 [0.33, 0.38] and 0.54 [0.52, 0.56] with MSP. At equal weight over all five sources vanilla leads by 0.23 points with a tie radius of 0.037, while with KNN OE leads that reference by 0.27.

E Exploratory Text Experiment

Both tested text sources improve in a separate experiment. We fine-tuned `roberta-base` (Liu et al., 2019) on SST-2 sentiment classification (Socher et al., 2013) for 3 epochs over 5 seeds, with WikiText-2 sentences (Merity et al., 2017) as the auxiliary pool and maximum softmax probability as the scorer. The near-OOD source is a class-balanced sample of $1{,}000$ IMDB test reviews (Maas et al., 2011), 500 per label, drawn once with a fixed seed, and the far-OOD source is the RTE validation set of GLUE (Wang et al., 2019) with each sentence pair concatenated before tokenization. OE raises mean AUROC from 65.69 to 84.06 on IMDB ($+18.37$, sample standard deviations 2.40 and 6.92) and from 86.35 to 96.25 on RTE ($+9.90$, standard deviations 0.71 and 2.40), with SST-2 validation accuracy of 94.29 and 94.43 percent. The full IMDB test split gives 65.65 and 84.44, and positive reviews are the harder half for both models (61.53 against 69.77 for vanilla, 82.55 against 86.33 for OE). An earlier version of this experiment scored the first $1{,}000$ reviews of the distributed test split, which are all negative. The same checkpoints give 71.31 and 86.87 on that subset, so the single-polarity subset raised both models' AUROC, by about 6 and 3 points, and understated the gain by about 3. This experiment shows gains on both measured sources. It does not identify what causes the differences between the text and image results, has two measured OOD sources, and does not establish a general modality effect or a geometric cause. Seeds are 10 through 14. The

Table 17: **Decision boundaries of every matched comparison of Outlier Exposure (OE) with vanilla.** Weight on the source where OE loses (CIFAR-10 as out-of-distribution (OOD) data with CIFAR-100 as in-distribution (ID) data, CIFAR-100 with CIFAR-10 as ID, iNaturalist on ImageNet-200), with the remainder on the other near-OOD source (ImageNet-200: split equally over SSB-hard and NINCO), at which vanilla ties OE (tie) or leads it by one point of area under the receiver operating characteristic curve (AUROC). MSP: maximum softmax probability, with a 95% paired seed-bootstrap interval (2,000 replicates). Over scorers: the range of the seven scorers' boundaries, with the number of scorers that have one. Per seed: the range of the boundaries computed from each seed's own differences, with the number of seed-scorer pairs without a boundary. "None" means OE leads or trails on both sources, so no two-source mixture changes the leader. TIN-597: TinyImageNet-597; WRN-40-2: WideResNet-40-2. Ten seeds per arm, five on ImageNet-200. A read-out from the saved per-seed rows, computed after the registered verdicts.

Comparison	Tie, MSP [95%]	Tie, over scorers	Tie, per seed	One point, over scorers
CIFAR-100 ID, ResNet-18, TIN-597	0.97 [0.96, 0.98]	0.64–0.97 (7/7)	0.36–0.98 (3/70 none)	0.68–0.89 (5/7)
CIFAR-100 ID, ResNet-18, 300K	0.86 [0.80, 0.93]	0.57–0.91 (6/7)	0.45–0.99 (13/70 none)	0.84–0.96 (3/7)
CIFAR-100 ID, WRN-40-2, TIN-597	0.90 [0.88, 0.91]	0.69–0.98 (7/7)	0.49–1.00 (2/70 none)	0.73–0.94 (6/7)
CIFAR-100 ID, WRN-40-2, 300K	0.84 [0.79, 0.88]	0.38–0.95 (6/7)	0.02–0.99 (16/70 none)	0.71–0.89 (3/7)
CIFAR-10 ID, ResNet-18, TIN-597	none	0.93–0.93 (1/7)	0.59–0.71 (66/70 none)	none (0/7)
CIFAR-10 ID, ResNet-18, 300K	none	none (0/7)	none (70/70 none)	none (0/7)
CIFAR-10 ID, WRN-40-2, TIN-597	none	none (0/7)	none (70/70 none)	none (0/7)
CIFAR-10 ID, WRN-40-2, 300K	none	none (0/7)	none (70/70 none)	none (0/7)
ImageNet-200, iNaturalist vs near pair	0.36 [0.33, 0.38]	0.18–0.44 (3/7)	0.01–0.99 (15/35 none)	0.30–0.61 (3/7)

ImageNet-200, all five sources at equal weight: vanilla leads with MSP by 0.23 points (tie radius 0.037, one-point radius 0.198) and on 6 of 7 scorers; with KNN, OE leads by 0.27 (tie radius 0.029).

evaluation script, one result file per run and the run logs are in the supplement, separate from the image experiments.

F Archived Evaluations of Published Objectives, per Set

Archived evaluations of published objectives, trained with OpenOOD's own trainers for 400 epochs and evaluated with the same seven scorers, read per set as follows. They have three seeds per method except CIFAR-100 OE, which has ten, hold means and standard deviations only, were computed after the registered results were known, and carry no registered label. On CIFAR-10, PFS is higher than OE on CIFAR-100 on six scorers and lower on TinyImageNet on six, and LogitNorm is higher than OE on CIFAR-100 on four and lower on TinyImageNet on the five interpreted scorers (its Energy and ReAct standard deviations exceed 20 points). On CIFAR-100, OE is below vanilla on CIFAR-10 as OOD on all seven scorers, by 1.2 to 9.6 points, and PFS on all seven, while both gain 14 to 21 points on TinyImageNet. At 100 epochs, OE trained with Places365 or SVHN as the pool is below vanilla on CIFAR-10 on every scorer as well, and its TinyImageNet gain falls from 15.5 to 19.7 points with TinyImageNet-597 to 7.4 to 14.1 with Places365 and -0.6 to -1.8 with SVHN. On ImageNet-200 at 224×224, OE's difference from OpenOOD's vanilla checkpoint is larger on SSB-hard, which shares ImageNet's source, than on NINCO, which was curated to exclude it, on six of seven scorers, at unmatched budgets and with two points less ID accuracy. The separate matched ImageNet-200 confirmation is reported in Appendix C.4. Table 18 decomposes, per near-OOD set, the archived CIFAR-100 evaluations of objectives that OpenOOD v1.5 ranks: vanilla cross-entropy, OE, MixOE, PFS, LogitNorm and VOS, each trained with OpenOOD's own trainer and configuration for 400 epochs. The CIFAR-10 evaluations, the four-pool runs at 100 epochs and the ImageNet-200 runs read above are in the supplement's archived evaluations. All were evaluated in July 2026 with the seven scorers at each run's selected settings. The per-seed rows were not recovered, so the table is descriptive.

G Exploratory Source-Mixture Decision Analysis

The expanded candidate comparisons, direct pool-change contrasts and reference-sensitivity analyses were developed after examining earlier results and feedback. They are exploratory. They reuse saved evaluations and introduce no new training or test-image inference. The earlier registrations

Table 18: **Archived evaluations, CIFAR-100, 400 epochs, per near-OOD set.** Area under the receiver operating characteristic curve (AUROC) on CIFAR-10 as out-of-distribution (OOD) data (C10) and on TinyImageNet (TIN) per scorer, mean and sample standard deviation over n seeds, ResNet-18 trained with OpenOOD v1.5's own trainers and configurations and evaluated with the seven scorers at each run's selected settings. Per-seed rows were not recovered. Three seeds per arm, ten for Outlier Exposure (OE), so differences between sources of a few points lie within retraining variation. Descriptive, computed after the registered results were known, no registered label. The main text cites these runs only as a corroborating pattern for a loss that the registered comparisons establish.

		MSP		Energy		ODIN		ReAct		Scale		ASH		KNN	
Objective	n	C10	TIN	C10	TIN	C10	TIN	C10	TIN	C10	TIN	C10	TIN	C10	TIN
Vanilla	3	$79.65_{\pm 0.20}$	$83.73_{\pm 0.26}$	$80.03_{\pm 0.31}$	$84.04_{\pm 0.26}$	$79.31_{\pm 0.28}$	$82.78_{\pm 0.24}$	$79.86_{\pm 0.28}$	$84.09_{\pm 0.27}$	$80.20_{\pm 0.29}$	$83.91_{\pm 0.26}$	$77.15_{\pm 0.16}$	$78.86_{\pm 0.60}$	$78.20_{\pm 0.10}$	$84.42_{\pm 0.29}$
OE	10	$78.46_{\pm 0.56}$	$99.89_{\pm 0.02}$	$78.51_{\pm 0.65}$	$99.93_{\pm 0.02}$	$76.17_{\pm 0.96}$	$97.26_{\pm 0.43}$	$74.61_{\pm 5.08}$	$99.92_{\pm 0.02}$	$70.59_{\pm 8.66}$	$99.94_{\pm 0.03}$	$70.01_{\pm 3.61}$	$99.95_{\pm 0.01}$	$75.28_{\pm 0.32}$	$98.27_{\pm 0.35}$
MixOE	3	$78.58_{\pm 0.31}$	$83.62_{\pm 0.19}$	$78.32_{\pm 0.42}$	$83.29_{\pm 0.32}$	$76.21_{\pm 0.62}$	$80.70_{\pm 0.06}$	$77.99_{\pm 0.89}$	$82.89_{\pm 0.57}$	$76.63_{\pm 0.22}$	$82.33_{\pm 0.47}$	$57.30_{\pm 1.54}$	$61.24_{\pm 2.29}$	$76.36_{\pm 0.24}$	$84.21_{\pm 0.08}$
PFS	3	$79.35_{\pm 0.43}$	$99.93_{\pm 0.02}$	$79.28_{\pm 0.43}$	$99.94_{\pm 0.01}$	$77.72_{\pm 1.27}$	$98.61_{\pm 0.18}$	$79.15_{\pm 0.35}$	$99.94_{\pm 0.01}$	$66.37_{\pm 12.13}$	$99.93_{\pm 0.05}$	$71.95_{\pm 0.97}$	$99.80_{\pm 0.05}$	$60.51_{\pm 2.22}$	$99.84_{\pm 0.03}$
LogitNorm	3	$74.68_{\pm 1.00}$	$82.51_{\pm 0.28}$	$70.23_{\pm 0.60}$	$78.76_{\pm 0.22}$	$74.22_{\pm 0.27}$	$83.04_{\pm 0.33}$	$69.68_{\pm 0.48}$	$79.15_{\pm 0.53}$	$70.02_{\pm 0.36}$	$79.53_{\pm 0.52}$	$68.34_{\pm 0.56}$	$77.38_{\pm 0.72}$	$73.51_{\pm 0.13}$	$82.01_{\pm 0.24}$
VOS	3	$79.53_{\pm 0.47}$	$83.75_{\pm 0.24}$	$79.94_{\pm 0.45}$	$84.21_{\pm 0.22}$	$79.11_{\pm 0.52}$	$82.81_{\pm 0.23}$	$79.90_{\pm 0.42}$	$84.25_{\pm 0.21}$	$80.03_{\pm 0.45}$	$84.05_{\pm 0.22}$	$76.54_{\pm 0.96}$	$78.55_{\pm 0.46}$	$62.68_{\pm 21.38}$	$67.04_{\pm 24.44}$

and their verdicts remain in Appendix C. The supplemental archive provides a frozen input snapshot, analysis scripts, saved outputs and a manifest.

G.1 Candidates and Estimands

CIFAR-10 uses ten training seeds, 0 to 9. Its original configurations are vanilla, OE with TinyImageNet-597 auxiliaries and OE with the size-matched 300K pool. CIFAR-100 uses ten confirmation seeds, 10 to 19, and additionally includes the mixing-only control for each pool. Each original has seven scorers under TinyImageNet-selected and recorded Textures-selected settings. This gives 42 and 70 nominal original choices, respectively. Duplicate choices remain because MSP and Energy have no selected setting and ODIN is unchanged. Textures reselects ReAct, ASH, Scale and KNN only.

The four new configurations use TinyImageNet auxiliary shares of 25, 50 and 75 percent, and the full union. Each contributes seven default scorers, giving 28 new candidates. Their alternative-settings evaluations are excluded from this exploratory comparison. Giving the originals both settings regimes expands the comparator set. It does not constitute an exhaustive search over hyperparameters.

For each candidate, the source vector averages AUROC over seeds. At each weight, $M(w)$ compares the best new candidate with the best original. $V(w)$ also permits retaining an original and is therefore nonnegative. For a fixed new candidate n, its best margin solves

$$\max_{w,t} t \quad \text{subject to} \quad t \le w \cdot (a_n - a_b) \text{ for every } b \in \mathcal{B}, \quad w \in \mathcal{W}.$$

We solve this linear program for each new candidate and take the largest optimum. The allowed families used here have linear constraints. Source-weight radii use the greedy construction in Appendix A or an independent linear-programming formulation. These are exact optimizations on estimated means, separate from statistical uncertainty.

G.2 Training-Run Uncertainty and Held-Out Choices

The expanded-choice bootstrap uses 5,000 resamples of complete seed records. Every source and candidate in a resample uses the same drawn seed indices, retaining the observed dependence within a seed. We recompute means and optimize over a 0.001-spaced near-pair weight grid. Point optima use the linear program. Nominal 95 percent ranges use the 2.5th and 97.5th percentiles. Bands at individual weights are pointwise. Optimizing over weights and inspecting several interventions gives no general selection-adjusted or simultaneous coverage guarantee.

For each held-out seed, the other nine seeds determine the mixture weight, best original and best candidate from the expanded set. Those choices are evaluated on the held-out run. CIFAR-10 selects the 75 percent blend with MSP in every fold. Its ten differences range from 0.533 to 1.182 points, with mean 0.865. The training folds overlap. All folds reuse the same evaluation images, so they are not independent significance tests and do not test new-source generalization. Fixing the

weight and choosing only the candidates on the other nine seeds gives held-out gains at each weight, and these curves are among the saved outputs.

The five-source extension adds MNIST, SVHN and Places365 to the near pair and excludes Textures because it supplies selection data. Under equal weights, CIFAR-10's new-candidate margin is -0.287 points and held-out selection retains an original in every fold. CIFAR-100's margin is $+0.387$ with range $[-0.628, +1.127]$. A new candidate wins in 76.84 percent of its bootstrap replicates. Its held-out expanded-choice gain averages 0.516, with seven positive and three negative differences ranging from -2.904 to $+2.906$. This spread leaves the positive point estimate unresolved. The selected blend's ID accuracy is 77.422 percent, compared with 77.732 for the original TinyImageNet-pool OE model.

G.3 Pool-Change Contrasts

The direct contrasts jointly resample all three configurations, vanilla and both OE pools, 10,000 times. We compute the change in the source difference and the change in the crossover within each resample. For Scale with original settings on CIFAR-100, both tie boundaries exist in every resample. The source-repair range is $[1.709, 9.553]$ AUROC points and the crossover-change range is $[-0.260, -0.014]$. The one-point-deficit change has range $[-0.050, +0.213]$.

ODIN's source repair is 2.244 points with range $[1.869, 2.566]$, and its tie coordinate falls from 0.836 to 0.763, with change range $[-0.131, -0.012]$. ReAct's crossover change is uncertain, and ASH moves in the opposite direction from Scale at the point estimates. These are exploratory, pointwise comparisons across several scorers. Saved paired and independent-arm bootstraps for individual boundaries assess sensitivity to the seed-pairing assumption. Finite-boundary quantiles are conditional on existence, and saved outputs separately report the fraction with no boundary.

G.4 Published Tables, References and Noise Models

We retain all measured sources for each benchmark, compare near and far subsets, and allocate total near mass of 0.25, 0.50 or 0.75 equally within each group. Unrestricted scenarios allow every source weight to vary. Grouped scenarios hold within-group proportions fixed and vary only total near/far mass. The equal-six-source reference is distinct from equal near/far group mass.

For tables with reported seed standard deviations, we use 2,000 Gaussian perturbations per reference, noise model and error multiplier. Standard errors are the published standard deviations divided by $\sqrt{3}$. The independent model draws a separate perturbation for each detector/source cell. The shared model uses a common standardized perturbation for scorers within the same training method and source, scaled by each cell's standard error. Different sources remain independent in both models. Error multipliers are 1, 1.5 and 2. These assumptions cannot recover missing covariance or uncertainty in standard errors estimated from three seeds.

Uncertainty is evaluated at the equal-source, near-pair, balanced unrestricted and balanced grouped references. Other near-mass settings are point estimates. The original reference leader remains fixed for one radius calculation. A separate calculation reselects the leader in each perturbed table. Retention frequencies and 5th to 95th percentile ranges describe these noise scenarios. When the original leader loses, its fixed-leader tie radius is zero. No uncertainty is manufactured for ImageNet-1K, whose table lacks standard deviations.

The CIFAR-100 equal-source leader's retention is 0.8925 under independent noise and 0.8840 under shared noise at the original scale. At twice that scale the frequencies are 0.7085 and 0.7040. The near-pair leader is retained in every simulated table in both models at every scale. A separate all-source mixture-share calculation uses 20,000 fixed uniform-simplex draws and 1,000 perturbed tables per noise/scale scenario. Under the original error scale, the near leader has the largest winning region in 18.0 and 15.4 percent of CIFAR-100 replicates. Uniform weight volume is an assumed measure and is no estimate of deployment frequency. The earlier 15 percent independent result in Appendix B uses different simulation draws.

G.5 Numerical Validation

Validation compared 630 radius calculations against an independent linear program and checked 210 source-splitting cases. Splitting an identical source preserves the reference's total mass and corresponding feasible family. This invariance does not imply invariance to changing group constraints or assigning a new equal-source reference. Every finite published boundary has a checked rival and source-weight witness. An independent 100,001-weight grid agrees with the expanded-choice point optima within 0.002 AUROC points. Saved records also check the earlier grouping results, ImageNet arithmetic and CPU/GPU winner agreement. The supplemental evidence index lists the scripts and their inputs.

H Prespecified Evaluation of a Frozen Choice on Separate Images

The protocol was committed before computing new predictions. The discovery results were already known. We retained the selected 75 percent TinyImageNet blend with MSP, all 42 nominal original choices, ten existing checkpoints per training configuration, and their recorded scorer settings. The primary mixture places weight 0.65 on CIFAR-100 and 0.35 on TinyImageNet. Neither the candidate nor this weight was reselected on the new outcomes. The mean averages per-checkpoint AUROC and is not the AUROC of a score ensemble.

Images and source scope. CIFAR-10.2 was collected to reproduce the CIFAR-10 task and exhibits distribution shift (Lu et al., 2020). We combine its two released splits as an evaluation pool, and no classifier is trained on either split. Within each released-split/class stratum, a fixed hash ordering assigns 20 percent to threshold calibration and the remainder to confirmation, yielding $2{,}400$ and $9{,}600$ images. This role assignment differs from the dataset authors' train/test use. The changed ID distribution makes this a transfer evaluation, rather than a replication under identical source distributions.

The two OOD samples contain $10{,}000$ images each from the CIFAR-100 and TinyImageNet training splits. These split names refer to their original datasets, and those splits did not train the CIFAR-10 classifiers in this comparison. Category proportions follow the previous OOD test lists, using deterministic largest-remainder allocation and hash ordering within categories. TinyImageNet categories used for scorer tuning are excluded. We screen decoded images at the evaluation resolution against $274{,}360$ distinct previous training, auxiliary, tuning and evaluation paths. The fixed screen excludes exact duplicates and prior-image matches with per-pixel root mean squared error at most 2 on a 0 to 255 scale. Four exact duplicates are excluded: one CIFAR-100 image present in the 300K auxiliary pool, one repeated new CIFAR-100 image, and two TinyImageNet images matching earlier test images. Replacement samples follow the same fixed ordering. No CIFAR-10.2 images are excluded. This pixel-level screen does not establish semantic independence.

An audit of $1{,}725$ project records found no prior CIFAR-10.2 evaluations. The CIFAR-100 training-split references concern other experiments in which CIFAR-100 was ID. Absence from those records cannot exclude unrecorded evaluations. The protocol, input hashes, exclusions and code were frozen before this study's predictions were computed.

Requirements and failure semantics. The primary rule requires a gain exceeding 0.5 AUROC points over every original choice and an ID-accuracy loss below 0.5 percentage points relative to TinyImageNet-pool OE. These requirements were set with the discovery result known. They are evaluation scenarios and do not estimate application utility. Each gain uses its one-sided 95% lower bound, and the accuracy loss uses its one-sided 95% upper bound. Success requires all conditions to hold, an intersection-union rule for these fixed hypotheses. The bootstrap bounds are approximate and provide no exact finite-sample guarantee. The requirement fails, and a secondary result or different mixture cannot replace that verdict.

Image and training-run uncertainty. We use $5{,}000$ stratified bootstrap replicates. Calibration images, ID-confirmation images and the two OOD sources are resampled separately, retaining category counts. The same image indices are used across all candidates and checkpoints. Tie-corrected weighted ranks recompute AUROC exactly for each resample. Accuracy, calibration thresholds and

false-acceptance rates are also recomputed. The primary intervals concern finite image samples for the fixed ten-checkpoint set.

A separate sensitivity analysis also resamples ten checkpoints within each training configuration, jointly across its scorers and sources, while retaining shared image draws across configurations. Independent seed draws across configurations avoid assuming that equal numeric seed indices establish paired replication. The AUROC margin range becomes $[0.170, 0.691]$ and the ID-accuracy-cost range becomes $[0.337, 0.954]$. Same-index seed resampling gives $[0.170, 0.700]$ and $[0.341, 0.942]$, respectively. Both preserve the positive AUROC comparison and the failed primary requirement. These are approximate sensitivity summaries and do not replace training on new seeds.

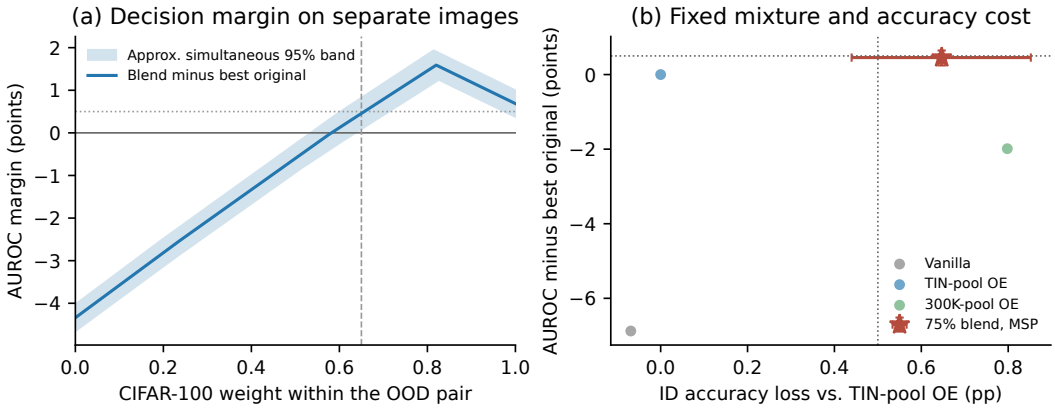


Figure 7: **The frozen blend retains a conditional AUROC advantage but fails the practical requirement.** CIFAR-10.2 is ID, and CIFAR-100 and TinyImageNet supply the OOD pair. (a) Margin of the frozen blend over the best original. The approximate simultaneous band propagates bounds on all source-wise candidate differences. The vertical line marks the prespecified weight 0.65. (b) At that fixed mixture, each original training configuration is represented by its best scorer across both recorded settings regimes. The blend's error bars are pointwise 95% image-bootstrap ranges. Dotted lines mark the 0.5-point gain and 0.5-percentage-point accuracy-cost requirements. These visual error bars are distinct from the one-sided bounds used by the joint primary rule.

Mixture regions and practical alternatives. For the secondary all-weight comparison, we use the bootstrap maximum absolute standardized deviation over all original choices and both source differences. Its 95th percentile gives an approximate simultaneous componentwise band. Linear propagation and minimization over originals give the band in Figure 7a. On the 0.001 grid, the image-only band supports a positive margin from weight 0.637 to 1, and a negative margin from 0 to 0.534. The intervening weights are unresolved. Adding independent training-run resampling moves the positive endpoint to 0.670. These conclusions remain conditional on the measured category mixtures and the bootstrap approximation.

The point-estimate tie radius from weight 0.65 is 0.068, with image-bootstrap range $[0.037, 0.091]$. The one-point-deficit radius is 0.205, with range $[0.180, 0.228]$. All replicates have both boundaries. Under the prespecified ID-cost ceilings of 0, 0.25 and 0.5 percentage points relative to TinyImageNet-pool OE, the point-estimate expanded choice retains an original. The one-point ceiling admits the blend and its 0.46-point AUROC gain. The same eligibility pattern holds with vanilla as the accuracy reference. This secondary ceiling comparison cannot rescue the failed primary requirement.

Operating point. For each checkpoint and scorer, the fifth percentile of confidence on the separate ID-calibration images sets its threshold. This targets 95 percent ID acceptance, and observed confirmation acceptance is 94.358 percent for the discovery-selected original and 94.046 for the blend. The original uses Textures-selected KNN. Its OOD false acceptance is 43.836 percent against 46.866 for the blend. The blend's increase is 3.03 points, with range $[0.859, 6.069]$, including calibration-sample uncertainty. The best original under confirmation false acceptance uses

TinyImageNet-selected KNN and gives 43.817 percent. Neither operating-point comparison favors the blend.

Validation and artifacts. All 40 checkpoint hashes match the earlier extraction records. The saved-feature pipeline reproduces every checked original AUROC to numerical precision. Forward passes on 128 old validation images per checkpoint agree on every class prediction, with a maximum relative logit root mean squared error of 0.000388 across computing stacks. The supplemental notes record the engineering tolerance revisions made using old data before new inference. Synthetic tied-score tests verify weighted AUROC and calibration quantiles against explicit resampling on CPU and GPU. The compact bootstrap arrays reproduce the decision summaries, including the failed rule, without raw images or checkpoints. Raw per-example scores and checkpoints are retained separately and are not included in the anonymous archive.

Table 19: **The blend does not meet the switching requirement.** CIFAR-10.2 is in-distribution (ID), with CIFAR-100/TinyImageNet out-of-distribution (OOD) weights 0.65/0.35. The original uses Outlier Exposure (OE) with TinyImageNet auxiliaries and Textures-selected nearest-neighbour scoring (KNN). It led under the discovery rule and again has the highest area under the receiver operating characteristic curve (AUROC) among originals on the new images. The blend uses maximum softmax probability (MSP). Values average ten fixed checkpoints. Thresholds use separate ID-calibration images to target 95 percent acceptance.

Outcome	Original	75% blend, MSP
AUROC (higher is better)	88.672	89.127
ID classification accuracy (%)	84.681	84.034
OOD false acceptance (%, lower is better)	43.836	46.866
Observed ID acceptance (%)	94.358	94.046

H.1 A Consistent Gain Meets All Three Conditions

This read-out asks whether OE with TinyImageNet-597 auxiliaries and MSP, on CIFAR-10 with ResNet-18, changes detector choice against vanilla under Definition 1. It uses saved outputs only: the ten-seed rows of the pool intervention for the first two conditions and the separate-image evaluation above for the third. Its protocol was committed before any quantity was computed, but the separate-image point estimates at the blend's weight 0.65 were already visible in the saved candidate table, so the result is reported as a read-out of saved outputs. The fourteen original choices are vanilla under all seven scorers and both settings regimes, compared on the five sources the two regimes share, since Textures cannot score settings it selected.

On the original images, OE with MSP beats every original choice on each of the five sources, by at least 1.16 points on CIFAR-100, 8.04 on TinyImageNet, 0.81 on MNIST, 7.10 on SVHN and 5.43 on Places365. On Textures, which only the TinyImageNet-selected settings can score, its smallest lead over those seven choices is 5.27 points. Its margin is therefore positive at every mixture of them, 4.60 points at the equal near pair, and no shift reverses it. On the separate images at the equal near pair, its smallest gain over an original is 8.81 points, against vanilla with KNN, with one-sided 95% lower bound 8.52. Its ID-accuracy cost is 0.07 percentage points with upper bound 0.23. The requirement above is met, and it is also met under the paired seed-and-image bootstrap (lower bound 8.48, upper bound 0.30) and at the weight 0.65 (lower bound 6.54). At independently calibrated thresholds OE with MSP accepts 34.8 percent of the new near-OOD images, against 70.4 percent for the vanilla choice with the lowest false acceptance, ReAct.