WHEN OOD CLASSES WERE SEEN UPSTREAM: EFFECTS OF CLASS-SPECIFIC SUPERVISION ON DOWNSTREAM OOD DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) detection is typically evaluated with respect to a downstream task, but the representation used by the detector may have encountered an OOD class during upstream training. We ask whether the upstream supervision history of a class affects its detectability as OOD downstream. To estimate this effect, we introduce a controlled evaluation in which the downstream task, evaluation data, architecture, training objective, and number of upstream classes are held fixed, while future-OOD classes either receive supervised treatment upstream or are withheld from upstream training. We make this comparison through a grouped rotating leave-one-out construction across multiple semantic classes. Experiments on a WordNet-structured ImageNet subset, CIFAR-100, and iNaturalist 2021 show that, under cosine kNN, future-OOD classes supervised upstream are systematically easier to detect downstream than matched classes withheld from upstream training. The mean AUROC difference is 5.9–7.5 points across the three datasets, and the effect persists with from-scratch Vision Transformers, with differences of 2.6–4.9 points. Across nine established OOD detectors, upstream supervision consistently affects OOD performance, but the magnitude and direction vary by detector. These results show that the upstream provenance of evaluated OOD classes is a previously underexamined source of variation in OOD detection, highlighting the importance of accounting for class provenance in OOD evaluation. The source code is available at this anonymous repository.

1 INTRODUCTION

Whether an input is out-of-distribution (OOD) depends on the task in which it is evaluated, but the representation used for OOD detection may have encountered that input's class long before evaluation. A class can therefore be OOD with respect to a downstream classifier while being familiar to the representation because it was explicitly supervised during upstream training. Alternatively, the class may have been present upstream only without labels, or absent from upstream training altogether. These cases are indistinguishable under the usual downstream definition of OOD, yet they give the learned representation very different histories with respect to the classes being evaluated.

This distinction matters because modern image OOD detection is typically performed post hoc, after a classifier or representation has already been trained. Given a fixed model, an OOD detector applies a score to a new input and decides whether to accept it as in-distribution (ID) or reject it as OOD. Common approaches include maximum softmax probability (MSP), ODIN, Energy, Mahalanobis distance, and deep kNN (Hendrycks & Gimpel, 2017; Liang et al., 2018; Liu et al., 2020; Lee et al., 2018; Sun et al., 2022). Benchmarks such as OpenOOD evaluate these methods on fixed models and downstream ID/OOD splits (Yang et al., 2022). In such evaluations, the upstream training history of an OOD class is typically not treated as a separate experimental factor.

This raises a basic question: does the upstream treatment of a class affect how easily that class can be detected as OOD downstream? In particular, if two classes are identical in their downstream role but differ in whether they were supervised during upstream training, will an OOD detector treat them differently? We study this question by systematically varying the upstream treatment of classes that are held out from the downstream task, while controlling the architecture, training objective, and

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.

downstream evaluation setting. This design provides a controlled estimate of how class provenance changes downstream OOD detectability.

Prior work provides evidence that the upstream history of a representation can influence downstream OOD behavior. Miyai et al. (2023) show that classes encountered during pretraining can behave differently at downstream OOD evaluation from classes absent during pretraining, with outcomes also differing between supervised and self-supervised pretraining. Yang et al. (2025) demonstrate that semantic label information can affect the separation of closely related OOD classes, while Krumpl et al. (2026) show that representation-learning choices can substantially alter the behavior of post-hoc OOD detectors. Together, these studies suggest that what a representation learns before downstream evaluation can shape subsequent OOD detection. However, they do not isolate whether *supervised upstream treatment of the same semantic class that later becomes OOD* systematically changes its downstream detectability.

We address this question with a controlled comparison in which the downstream task and evaluation data are held fixed while the upstream treatment of future-OOD classes is varied. For each future-OOD class, we compare matched models in which that class either receives supervised treatment during upstream learning or is withheld from upstream training. A grouped rotating leave-one-out construction repeats this comparison across many semantic classes while matching the architecture, objective, training recipe, and number of upstream classes. This design allows us to attribute differences in downstream OOD detectability specifically to the upstream treatment of the classes being evaluated.

Across CIFAR-100 (Krizhevsky, 2009), a WordNet-structured ImageNet subset (Deng et al., 2009; Russakovsky et al., 2015) (described in Section 4.1), and iNaturalist 2021 (Van Horn et al., 2021), under cosine kNN, future-OOD classes that received supervised treatment upstream are systematically easier to detect as OOD downstream than matched classes that were withheld from upstream training. With cosine kNN, this difference averages 5.9–7.5 AUROC points across the three datasets and is consistent across most classes and semantic groups. The pattern also persists with from-scratch Vision Transformers, with mean differences of 2.6–4.9 AUROC points.

The effect is not specific to cosine kNN. Across nine established OOD detectors, upstream supervision changes downstream OOD performance, but the magnitude and even the direction of the change vary across methods. Thus, upstream class provenance does not simply shift OOD performance by a fixed amount; its consequences depend on how the detector scores the learned representation. This makes the upstream treatment of the evaluated classes a potential source of variation in both absolute detector performance and comparisons between detectors.

Our contributions are:

- We identify upstream treatment of future-OOD classes as a previously underexamined factor in image OOD evaluation and introduce a controlled experimental design that estimates its same-class effect while matching the downstream task, evaluation images, and other training conditions.
- We show across CIFAR-100, WordNet-structured ImageNet, and iNaturalist 2021, under cosine kNN, that classes supervised upstream are systematically more detectable as OOD downstream than matched classes withheld from upstream training, with the effect replicated using both ResNet and from-scratch Vision Transformer representations.
- We establish that the effect generalizes across nine OOD detectors, while varying substantially in magnitude and direction across methods, showing that upstream class provenance can affect not only OOD performance but also comparisons among detectors.
- We discuss implications for OOD evaluation and recommend reporting the upstream provenance of OOD classes and, where feasible, testing detector conclusions under matched alternative provenance conditions.

2 Related Work

Post-hoc OOD detection. Many image OOD methods start from a trained classifier or representation and assign each test input an OOD score. Output-based approaches include MSP, ODIN, and Energy (Hendrycks & Gimpel, 2017; Liang et al., 2018; Liu et al., 2020), while feature-space approaches include Mahalanobis distance and deep kNN (Lee et al., 2018; Sun et al., 2022). Bench-

marks such as OpenOOD compare these methods across common ID/OOD settings (Yang et al., 2022; 2024). We instead study whether existing detectors behave differently when the future-OOD classes have different upstream training histories, rather than introducing a new detector.

Upstream exposure and supervision. Miyai et al. (2023) study OOD data that were represented during pretraining and compare them with OOD data outside the pretraining distribution, including under supervised and self-supervised pretraining. Their results show that pretraining exposure can change downstream OOD behavior, but their comparison varies the OOD data itself rather than holding the semantic class fixed while changing its upstream treatment. Yang et al. (2025) study a related issue and show that semantic label information can be important when OOD classes are close to the downstream ID classes. In contrast, our setting holds the OOD class, downstream role, and evaluation images fixed, and varies only whether that class received supervised treatment during upstream learning.

OOD benchmark construction and contamination. OOD evaluation can also fail when the intended ID/OOD separation is violated by benchmark construction. Bitterwolf et al. (2023) note substantial ID contamination in commonly used ImageNet-scale OOD test sets, while Balachandran (2026) study a class-split failure in which a class designated as OOD was already represented in model training and in the ID set used to fit the detector. Our setting deliberately excludes this form of downstream leakage: every focal class remains outside the downstream ID task and the data used to fit the detector in every condition. Our contrast instead concerns upstream supervision of a class that remains genuinely OOD with respect to the downstream task: the same future-OOD class is either supervised or withheld during upstream learning, while its downstream role and evaluation images remain fixed.

Representation geometry and detector behavior. OOD scores are closely tied to the geometry of learned representations. Recent work has connected supervised feature geometry and neural-collapse structure to OOD detection (Ben Ammar et al., 2024; Harun et al., 2025; Liu & Qin, 2025). Separately, Krumpl et al. (2026) show that changing the overall representation-learning strategy can alter the behavior of post-hoc detectors. Our experiments examine a more specific source of variation: whether supervising the exact class later evaluated as OOD changes its feature geometry relative to the downstream ID classes and thereby affects its downstream detectability.

3 METHOD

3.1 UPSTREAM CLASS PROVENANCE AND DOWNSTREAM OOD MEMBERSHIP

Let f_θ be an encoder learned during upstream training, and let $\mathcal{C}_{\mathrm{ID}}$ denote the set of classes in the downstream ID task. A semantic class c is treated as downstream OOD whenever $c \notin \mathcal{C}_{\mathrm{ID}}$, regardless of whether examples from that class appeared during upstream training. We compare two upstream histories for the same downstream-OOD class: *supervised*, where examples of c are included as an explicitly labeled upstream class, and *withheld*, where examples of c are absent from upstream training. Thus, the primary comparison captures the effect of supervised upstream treatment relative to complete upstream absence; it does not separately identify the contributions of class presence and explicit labels. We additionally examine unlabeled upstream presence with SimCLR in Section 4.1 to test whether presence alone produces a similar downstream pattern.

For detector D, let $A_D(c \mid h)$ denote the downstream AUROC for class c under upstream history h. We define the within-class effect as

$$\Delta_D(c) = A_D(c \mid \text{withheld}) - A_D(c \mid \text{supervised}). \tag{1}$$

The downstream ID task and the test images of c are the same in both conditions. Thus, $\Delta_D(c)$ captures the change in downstream OOD detectability when the upstream treatment of the same semantic class changes.

3.2 GROUPED ROTATING LEAVE-ONE-OUT CONSTRUCTION

We organize the selected classes into $G = 20$ semantic groups. Each group g contains one fixed downstream-ID class d_g and four future-OOD classes,

$$\mathcal{O}_g = \{c_{g,1}, c_{g,2}, c_{g,3}, c_{g,4}\}.$$

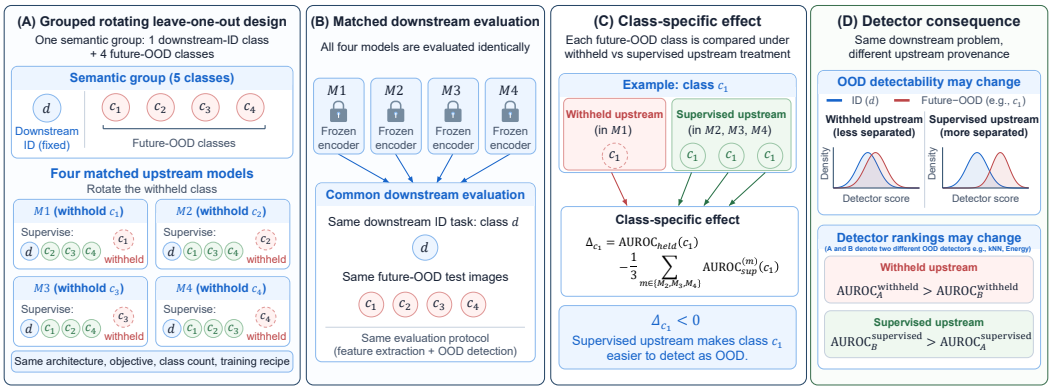


Figure 1: **Grouped rotating leave-one-out design for measuring class-specific upstream provenance effects on downstream OOD detection.** **(A)** Each semantic group contains one fixed downstream-ID class d and four future-OOD classes $c_1, \ldots, c_4$. Across four matched upstream models, each future-OOD class is withheld once and explicitly supervised in the other three models, while architecture, objective, class count, and training recipe are held fixed. **(B)** All four encoders are evaluated using the same downstream ID task, the same future-OOD test images, and the same OOD evaluation protocol. **(C)** For a future-OOD class c, the class-specific effect compares its AUROC in the unique withheld condition with its mean AUROC across the three supervised conditions. A negative Δ_c indicates that the class is easier to detect as OOD when it receives supervised upstream treatment than when it is withheld from upstream training. **(D)** Upstream provenance can affect both absolute OOD detectability and relative detector rankings. The score distributions are illustrative; the direction and magnitude of the provenance effect can depend on the detector.

We train four upstream models $M_1, \ldots, M_4$. In model M_j, the j-th future-OOD class in every group, $c_{g,j}$, is withheld from upstream training, while d_g and the other three future-OOD classes are supervised upstream. Each model therefore contains exactly $4G = 80$ supervised classes. Across the four models, every future-OOD class is withheld once and supervised three times.

The four models use the same architecture, supervised objective, optimizer, augmentation recipe, and number of supervised upstream classes. Within each training seed, the four rotations also use the same random initialization. The downstream ID task is always defined by the 20 classes $\{d_g\}_{g=1}^{20}$, and all 80 classes in $\cup_g \mathcal{O}_g$ remain OOD downstream in every condition. The downstream ID data and the OOD test images are unchanged across the four upstream models.

For each focal class, AUROC treats examples from that single class as OOD and the fixed 20 downstream-ID classes as ID. For a focal class $c_{g,j}$ and detector D, let $A_D^{(m)}(c_{g,j})$ denote its AUROC using the representation learned by model M_m. We define its rotating leave-one-out effect as

$$\Delta_D(c_{g,j}) = A_D^{(j)}(c_{g,j}) - \frac{1}{3}\sum_{m \neq j} A_D^{(m)}(c_{g,j}), \tag{2}$$

where the first term corresponds to the rotation in which the class is withheld, and the second averages the three rotations in which it is supervised. We compute this effect separately for each training seed and then average the two seed-level estimates for each class.

For dataset-level uncertainty, we bootstrap the 20 semantic groups rather than the 80 class-level estimates independently, so that all four future-OOD classes from a sampled group remain together. We use 10,000 percentile-bootstrap draws and report 95% confidence intervals. All 80 future-OOD classes therefore contribute one class-level effect estimate.

3.3 Downstream OOD evaluation

After upstream training, each encoder is frozen and placed in evaluation mode. No parameters or normalization statistics are updated during downstream feature extraction. The downstream ID feature bank contains only the 20 fixed ID classes $\{d_g\}_{g=1}^{20}$.

Our primary detector is cosine kNN with $k = 50$. We ℓ_2-normalize the features and score each test example by its mean cosine distance to the 50 nearest downstream-ID training features. Larger scores indicate more OOD, and we report AUROC with OOD as the positive class.

To examine whether the effect extends beyond cosine kNN, we also evaluate Energy (Liu et al., 2020), Mahalanobis (Lee et al., 2018), ViM (Wang et al., 2022), NECO (Ben Ammar et al., 2024), NCI (Liu & Qin, 2025), GradOrth (Behpour et al., 2023), and ODIN (Liang et al., 2018). MSP is included as an additional baseline. Where a detector requires classifier outputs or parameters, we train a linear classifier using only frozen downstream-ID training features; no future-OOD examples are used for classifier fitting or detector calibration.

All detector hyperparameters and fitting procedures are fixed without using any future-OOD evaluation data. Each detector is evaluated on the same downstream ID data and the same OOD test images for every upstream condition. Scores are oriented so that larger values indicate more OOD, and we apply the same class-level effect in Eq. 2 and the same semantic-group bootstrap to every detector. Full detector definitions and hyperparameter settings are given in Appendix A.5.

3.4 Downstream-ID-relative feature geometry

To examine how upstream class treatment changes the location of a future-OOD class in representation space, we analyze the frozen ResNet representations from the primary experiments. We compute the cosine distance between its feature centroid and the nearest downstream-ID class centroid. All centroids are computed from ℓ_2-normalized frozen features. For each future-OOD class, we measure this distance in the rotation where the class is withheld and average it across the three rotations where the class is supervised. We then compare the supervised-versus-withheld change in centroid distance with the corresponding change in cosine-kNN AUROC. Specifically, we define the geometry change as supervised minus withheld distance and the kNN effect as withheld minus supervised AUROC. Across the 80 future-OOD classes in each dataset, we quantify their association using Pearson and Spearman correlations.

4 Experiments

4.1 Experimental setup

CIFAR-100. CIFAR-100 contains 100 fine-grained classes organized into 20 predefined coarse categories, with five fine classes in each category (Krizhevsky, 2009). We use each coarse category as one semantic group: one fine class is fixed as downstream ID and the remaining four serve as future-OOD classes. We train four ResNet-18 rotations for each of two training seeds.

ImageNet. ImageNet does not provide the same fixed five-class grouping as CIFAR-100, so we construct an analogous 20-group setup using its WordNet hierarchy (Deng et al., 2009; Russakovsky et al., 2015). Each group contains five semantically related ImageNet classes, with one fixed as downstream ID and the other four used as future-OOD classes. We train four ResNet-50 rotations for each of two training seeds.

iNaturalist 2021. For iNaturalist 2021 (Van Horn et al., 2021), we build the same 20-group, five-class design using the dataset's taxonomic hierarchy. We select 20 genera containing five species each, with one species fixed as downstream ID and the other four used as future-OOD classes. We train four ResNet-50 rotations for each of two seeds and retain all available training images for each selected species. Class sizes are therefore fixed across rotations, although they may differ across species.

Architecture robustness. We repeat the supervised rotations with a from-scratch ViT-Tiny on all three datasets. We use a 192-dimensional CLS representation, with 4×4 patches for CIFAR-100 and 16×16 patches for ImageNet and iNaturalist 2021. The same semantic groups, downstream tasks, evaluation images, seeds, estimand, and OOD evaluation protocol are retained. For each dataset, the ViT training recipe was selected using classification generalization rather than OOD performance and then fixed across all rotations and seeds.

Explicit supervision versus unlabeled upstream presence. The main comparison contrasts supervised upstream treatment with complete exclusion of the focal class, and therefore does not by

itself distinguish the effect of explicit labels from the effect of upstream class presence. We therefore repeat the same class rotations under SimCLR, where the focal class is absent from upstream training in one rotation and present without labels in the other three. This provides a matched presence-versus-absence comparison within a contrastive training regime. Because the supervised and SimCLR pipelines differ in objective and augmentation, the comparison does not isolate the causal effect of labels alone.

Class selection procedure, dataset statistics, and other implementation details are provided in Appendix.

4.2 The same-class provenance effect is consistent across datasets

We first evaluate the supervised-versus-withheld rotations using cosine kNN, our primary detector. Across all three datasets, future-OOD classes are easier to detect when they receive supervised upstream treatment than when they are withheld from upstream training.

On CIFAR-100, supervised upstream treatment improves OOD AUROC by 7.5 points on average (95% CI: $[6.1, 9.0]$), with all 80 future-OOD classes and all 20 semantic groups showing the same direction. On the WordNet-structured ImageNet subset, the improvement is 6.3 points (95% CI: $[5.0, 7.8]$), again with all 80 classes and all 20 groups showing the same direction. On iNaturalist 2021, the improvement is 5.9 points (95% CI: $[4.2, 7.7]$), with the same direction for 67/80 species and 19/20 genera.

Thus, the same-class provenance effect is consistently negative across the three datasets under cosine kNN, although its class-level consistency is somewhat weaker on iNaturalist. Table 1 reports the corresponding results for the full detector suite.

Table 1: Provenance effects across OOD detectors for the ResNet experiments. Effects are reported in AUROC percentage points as $\Delta_D = W - S$, so negative values indicate better OOD detection when the future-OOD class was supervised upstream. Confidence intervals are 95% semantic-group bootstrap intervals. Classes < 0 and Groups < 0 report the number of class-level and semantic-group mean effects with negative sign.

	CIFAR-100 / R18				ImageNet / R50				iNaturalist 2021 / R50			
Detector	Effect	95% CI	Classes < 0	Groups < 0	Effect	95% CI	Classes < 0	Groups < 0	Effect	95% CI	Classes < 0	Groups < 0
kNN	-7.5	$[-9.0, -6.1]$	80/80	20/20	-6.3	$[-7.8, -5.0]$	80/80	20/20	-5.9	$[-7.7, -4.2]$	67/80	19/20
Energy	-4.1	$[-5.2, -3.0]$	75/80	20/20	-3.0	$[-4.1, -1.9]$	66/80	19/20	$+1.1$	$[-0.3, +2.9]$	36/80	6/20
MSP	-3.8	$[-4.9, -2.8]$	73/80	20/20	-2.2	$[-3.3, -1.2]$	63/80	17/20	-0.3	$[-1.4, +0.9]$	42/80	13/20
Mahalanobis	-14.0	$[-15.7, -12.3]$	80/80	20/20	-11.8	$[-13.5, -10.2]$	80/80	20/20	-8.1	$[-10.1, -6.3]$	75/80	20/20
ViM	-8.4	$[-9.7, -7.3]$	80/80	20/20	-5.7	$[-7.1, -4.4]$	80/80	20/20	-8.3	$[-10.1, -6.7]$	76/80	20/20
NECO	-5.2	$[-7.2, -3.1]$	54/80	17/20	-0.8	$[-2.9, +1.0]$	37/80	11/20	-3.5	$[-4.8, -2.2]$	60/80	16/20
NCI	-6.5	$[-8.0, -5.1]$	79/80	20/20	-4.9	$[-6.2, -3.7]$	80/80	20/20	$+0.5$	$[-1.4, +2.9]$	40/80	10/20
GradOrth	$+2.0$	$[+1.4, +2.5]$	21/80	1/20	$+1.6$	$[+0.0, +3.1]$	30/80	4/20	$+4.5$	$[+2.1, +7.9]$	22/80	3/20
ODIN	-5.0	$[-6.2, -3.8]$	79/80	20/20	-3.9	$[-5.3, -2.8]$	73/80	20/20	$+0.3$	$[-1.2, +2.1]$	40/80	10/20

4.3 The effect replicates with a Vision Transformer training setup

We next repeat the same supervised-versus-withheld rotations using from-scratch ViT-Tiny models while keeping the semantic groups, downstream tasks, and evaluation images unchanged. The primary cosine-kNN effect remains negative on all three datasets.

On CIFAR-100, supervised upstream treatment improves kNN AUROC by 4.5 points on average (95% CI: $[3.4, 5.8]$), with 76/80 class effects and all 20 group means showing the same direction. On the WordNet-structured ImageNet subset, the improvement is 2.6 points (95% CI: $[1.6, 3.6]$), with 65/80 classes and 18/20 groups showing the same direction. On iNaturalist 2021, the improvement is 4.9 points (95% CI: $[3.1, 6.8]$), with 59/80 species and 19/20 genera showing the same direction.

The same-class provenance pattern therefore also appears under the ViT-Tiny training setup, although the magnitude of the primary kNN effect differs from the corresponding ResNet experiments. Table 2 reports the full detector-suite results.

4.4 The effect extends across OOD detectors

The same supervised-versus-withheld intervention also changes the behavior of other OOD detectors, although the magnitude and direction vary across methods and architectures.

Table 2: Provenance effects across OOD detectors for the ViT-Tiny experiments. Effects are reported in AUROC percentage points as $\Delta_D = W - S$, so negative values indicate better OOD detection when the future-OOD class was supervised upstream. Confidence intervals are 95% semantic-group bootstrap intervals. Classes < 0 and Groups < 0 report the number of class-level and semantic-group mean effects with negative sign.

	CIFAR-100 / ViT-Tiny				ImageNet / ViT-Tiny				iNaturalist 2021 / ViT-Tiny			
Detector	Effect	95% CI	Classes < 0	Groups < 0	Effect	95% CI	Classes < 0	Groups < 0	Effect	95% CI	Classes < 0	Groups < 0
kNN	-4.5	$[-5.8, -3.4]$	76/80	20/20	-2.6	$[-3.6, -1.6]$	65/80	18/20	-4.9	$[-6.8, -3.1]$	59/80	19/20
Energy	-6.8	$[-8.5, -5.4]$	79/80	20/20	-6.9	$[-8.5, -5.4]$	80/80	20/20	-4.7	$[-6.3, -3.1]$	60/80	18/20
MSP	-6.3	$[-7.9, -4.9]$	78/80	20/20	-6.4	$[-8.0, -5.1]$	80/80	20/20	-3.8	$[-5.4, -2.2]$	57/80	17/20
Mahalanobis	-3.8	$[-5.4, -2.5]$	67/80	20/20	-4.4	$[-5.7, -3.1]$	73/80	19/20	-1.9	$[-3.5, -0.5]$	45/80	11/20
ViM	-2.2	$[-3.6, -1.0]$	56/80	16/20	-4.5	$[-5.9, -3.3]$	75/80	19/20	-2.3	$[-3.7, -0.9]$	50/80	14/20
NECO	-1.8	$[-4.0, +0.4]$	44/80	12/20	-2.4	$[-4.4, -0.4]$	51/80	14/20	$+1.2$	$[-0.3, +2.7]$	34/80	7/20
NCI	-6.7	$[-8.3, -5.3]$	78/80	20/20	-6.6	$[-8.2, -5.2]$	80/80	20/20	-4.6	$[-6.2, -3.0]$	61/80	18/20
GradOrth	-6.8	$[-8.4, -5.4]$	73/80	20/20	-8.5	$[-10.3, -7.0]$	79/80	20/20	-2.4	$[-4.2, -0.7]$	53/80	13/20
ODIN	-8.2	$[-9.8, -6.7]$	80/80	20/20	-10.1	$[-12.7, -7.9]$	80/80	20/20	-6.2	$[-8.3, -4.2]$	63/80	18/20

With the ResNet representations, eight of the nine detectors have negative mean effects on both CIFAR-100 and ImageNet. Mahalanobis shows the largest increase in AUROC under supervised upstream treatment, by 14.0 points on CIFAR-100 and 11.8 points on ImageNet, while ViM increases by 8.4 and 5.7 points, respectively. The iNaturalist results are more mixed: kNN, MSP, Mahalanobis, ViM, and NECO have negative mean effects, whereas Energy, NCI, GradOrth, and ODIN have small positive effects. Mahalanobis and ViM remain strongly negative, with improvements of 8.1 and 8.3 points, respectively.

The ViT results show an even more consistent direction. All nine detectors have higher AUROC under supervised upstream treatment on CIFAR-100 and ImageNet, and eight of nine do so on iNaturalist. On ImageNet, for example, the effect ranges from 2.4 AUROC points for NECO to 10.1 points for ODIN. On iNaturalist, NECO is the only detector with a positive mean effect.

Detector responses can also change substantially with architecture. GradOrth, for example, has positive effects under all three ResNet experiments ($+2.0$, $+1.6$, and $+4.5$ points) but negative effects under all three ViT experiments (-6.8, -8.5, and -2.4 points). Thus, the provenance effect extends beyond kNN, but its magnitude and even direction can depend on both the learned representation and the detector applied to it. This detector dependence can also change pairwise conclusions. For kNN versus Energy, the detector ranking reverses for 79/240 future-OOD classes under ResNet and 90/240 under ViT (Appendix A.8).

4.5 Changes in representation geometry track the kNN effect

For the ResNet representations used in our primary experiments, we examine whether changes in kNN detectability are associated with changes in the location of the future-OOD class relative to downstream ID features.

Classes that move farther from the nearest downstream-ID class centroid under supervised upstream treatment tend to achieve higher kNN OOD AUROC (Figure 2). The association appears across all three datasets, with Pearson correlations of -0.73, -0.55, and -0.69 for CIFAR-100, ImageNet, and iNaturalist 2021, respectively. Spearman correlations show the same pattern (-0.77, -0.53, and -0.63). The correlations are computed across the 80 future-OOD classes in each dataset, after averaging each class-level effect across the two training seeds. These results link the provenance effect to changes in downstream-ID-relative feature geometry, but do not establish that centroid displacement is the mechanism producing the effect.

4.6 Unlabeled upstream presence produces a different pattern

Finally, we repeat the grouped presence-versus-absence experiment under SimCLR using the same ResNet backbones. Unlike the supervised setting, the primary cosine-kNN effect does not have a consistent direction across datasets.

Using the same absent-minus-present convention as Table 3, the kNN effect is $+1.1$ AUROC points on CIFAR-100 (95% CI: $[+0.2, +1.9]$), -0.7 points on the WordNet-structured ImageNet subset ($[-2.2, +0.6]$), and -3.2 points on iNaturalist 2021 ($[-4.5, -1.9]$). Thus, unlabeled upstream

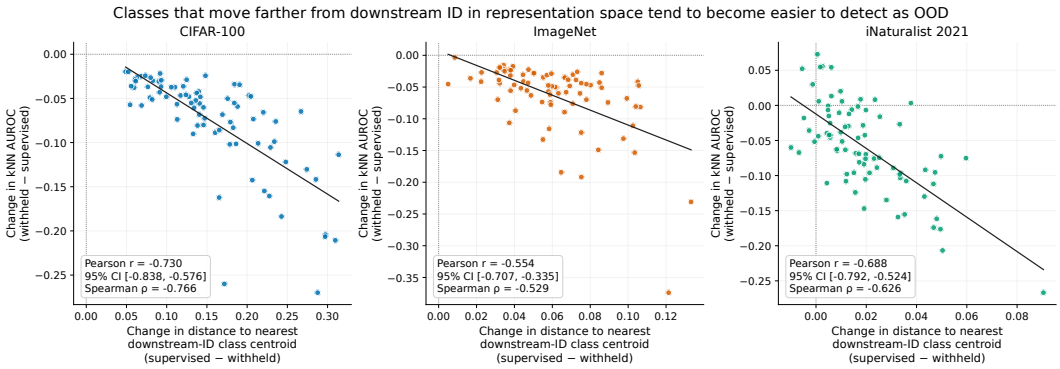


Figure 2: Each point represents one future-OOD class. The horizontal axis shows the change in cosine distance to the nearest downstream-ID class centroid after upstream supervision (supervised − withheld), while the vertical axis shows the corresponding change in cosine-kNN AUROC (withheld − supervised). Points in the lower-right therefore represent classes that move farther from downstream ID and become easier to detect after upstream supervision. The relationship appears across CIFAR-100, the WordNet-structured ImageNet subset, and iNaturalist 2021.

Table 3: SimCLR presence-versus-absence effects across OOD detectors. Effects are reported in AUROC percentage points as $\Delta_{\mathrm{SSL}} = A - P$, where P denotes upstream presence without class labels and A denotes upstream absence. Negative values indicate better OOD detection when the future-OOD class was present upstream. Confidence intervals are 95% semantic-group bootstrap intervals. Classes < 0 and Groups < 0 report the number of class-level and semantic-group mean effects with negative sign.

	CIFAR-100 / R18 SimCLR				WordNet-structured ImageNet subset / R50 SimCLR				iNaturalist 2021 / R50 SimCLR			
Detector	Effect	95% CI	Classes < 0	Groups < 0	Effect	95% CI	Classes < 0	Groups < 0	Effect	95% CI	Classes < 0	Groups < 0
kNN	$+1.1$	$[+0.2, +1.9]$	28/80	4/20	-0.7	$[-2.2, +0.6]$	48/80	12/20	-3.2	$[-4.5, -1.9]$	60/80	16/20
Energy	$+9.0$	$[+7.4, +10.6]$	0/80	0/20	$+2.2$	$[+0.7, +3.6]$	28/80	3/20	$+1.2$	$[-0.5, +3.2]$	34/80	9/20
MSP	$+7.5$	$[+6.0, +8.9]$	1/80	0/20	$+2.3$	$[+0.8, +3.6]$	28/80	4/20	$+2.4$	$[+1.4, +3.4]$	21/80	4/20
Mahalanobis	-15.5	$[-18.2, -12.9]$	79/80	20/20	-10.0	$[-11.6, -8.4]$	79/80	20/20	-6.5	$[-8.8, -4.4]$	68/80	17/20
ViM	-2.0	$[-3.1, -1.0]$	56/80	16/20	-3.9	$[-5.2, -2.6]$	65/80	18/20	-5.9	$[-8.0, -3.9]$	64/80	18/20
NECO	-0.04	$[-1.2, +1.0]$	32/80	8/20	$+3.8$	$[+2.0, +5.4]$	24/80	3/20	$+1.3$	$[+0.4, +2.0]$	34/80	3/20
NCI	$+1.5$	$[+0.6, +2.3]$	31/80	6/20	-1.7	$[-3.0, -0.6]$	56/80	13/20	-0.1	$[-1.0, +0.6]$	39/80	9/20
GradOrth	$+12.4$	$[+10.2, +14.7]$	0/80	0/20	$+4.9$	$[+3.4, +6.2]$	9/80	1/20	$+4.4$	$[+2.9, +6.1]$	12/80	2/20
ODIN	$+7.3$	$[+5.9, +8.8]$	1/80	0/20	$+0.2$	$[-1.3, +1.6]$	42/80	9/20	$+1.3$	$[+0.5, +2.1]$	35/80	4/20

presence makes the future-OOD class harder to detect on CIFAR-100, has a small and uncertain effect on ImageNet, and makes it easier to detect on iNaturalist. This differs from the supervised kNN results, where the effect is negative on all three datasets.

The broader detector suite also shows substantial heterogeneity under SimCLR. Mahalanobis and ViM have negative mean effects on all three datasets, whereas Energy, MSP, and GradOrth have positive mean effects on all three. Other detectors vary in sign across datasets. Unlabeled upstream presence can therefore change downstream OOD detectability, but its effect depends strongly on the detector and dataset and does not reproduce the consistent supervised kNN pattern. Absolute AUROC values corresponding to Tables 1, 2, 3 are reported in the Appendix A.10.

5 Discussion and Implications

5.1 Interpretation and Limitations

Our grouped rotation estimates a class-specific provenance effect by evaluating every selected future-OOD class under both upstream treatments while keeping its downstream role and evaluation images fixed.

The central implication is narrower than the general observation that OOD behavior depends on representation learning, which is already well established. Our results show that the upstream treatment of the particular semantic class later used as OOD can materially alter its downstream

detectability, even when the downstream ID task, detector, and OOD images are unchanged. The effect is strong in aggregate but not universal at the individual-class level: in the primary ResNet kNN experiments, 227/240 class-level kNN effects and 59/60 semantic-group effects have the same direction, while iNaturalist shows meaningful species-level heterogeneity. Thus, class provenance is a systematic but not deterministic source of variation in OOD evaluation.

The detector experiments further show that provenance is not a detector-independent offset: its magnitude, and in some cases its direction, depends on the representation and the OOD scoring rule. Thus, detector rankings and absolute performance should be interpreted conditional on the upstream history of the evaluated classes.

The SimCLR experiments qualify this interpretation. Unlabeled upstream presence alone can also alter downstream OOD behavior, but its direction and magnitude are substantially more heterogeneous than the supervised-versus-withheld effect. Thus, the main finding should not be interpreted as evidence that upstream presence alone is sufficient to produce the observed pattern. Rather, the results indicate that both unlabeled upstream presence and supervised upstream treatment can affect downstream OOD behavior, with supervised treatment producing the more consistent same-class effect in our experiments.

Our study has a few limitations. First, the primary supervised-versus-withheld comparison changes both explicit supervision and upstream presence of the focal class; the SimCLR experiments therefore test whether unlabeled presence alone can reproduce the observed pattern rather than providing a fully controlled three-way decomposition of presence and supervision. Second, our experiments focus on image classification representations and a limited set of architectures and training regimes, so the extent to which the effect transfers to other modalities, pretrained foundation models, or multimodal representations remains to be established. Third, our feature-geometry analysis is correlational and does not establish a causal mechanism linking upstream treatment to downstream OOD detectability. Finally, class-level upstream provenance may be difficult to determine for large pretrained models, which motivates treating provenance uncertainty itself as an explicit part of OOD benchmark documentation.

5.2 Implications and recommendations for OOD evaluation

Class-level upstream provenance should be treated as part of the evaluation specification whenever the evaluated representation has been pretrained or otherwise learned from upstream data. For each OOD class, benchmark reports should indicate whether the class was (i) explicitly supervised upstream, (ii) present upstream without class labels, (iii) absent from upstream training, or (iv) of unknown provenance.

When comparing OOD detectors, researchers should use the same OOD classes and the same upstream provenance conditions across detectors. If provenance cannot be matched, detector comparisons should report the provenance of the evaluated OOD classes and, where feasible, stratify results by provenance.

Benchmark developers should consider provenance sensitivity as an explicit evaluation axis. A practical protocol is to construct matched OOD evaluations in which the downstream ID task and OOD images are fixed while the upstream treatment of the OOD classes is varied. Even a smaller provenance-controlled subset can reveal whether a benchmark's conclusions are sensitive to upstream class history.

For pretrained and foundation-model representations, provenance may be difficult to determine because upstream data and supervision are often only partially documented. This makes class-level provenance an especially important reporting variable: when exact provenance is unknown, benchmark reports should state the uncertainty rather than implicitly treating OOD classes as unseen. Model and dataset documentation should, where possible, provide class-level or semantic-level information about upstream presence and supervision.

The broader significance is that OOD performance is not determined solely by the downstream ID/OOD partition and the detector applied to it. The upstream provenance of the semantic classes used for evaluation can become part of the experimental condition itself. Making this provenance explicit would improve the reproducibility of OOD benchmarks, the interpretability of detector comparisons, and the ability to determine whether reported OOD gains reflect better detection or simply a more favorable upstream training history.

AI USE DISCLOSURE

Generative AI tools were used in this work to provide feedback on experimental design and methodology, support literature exploration, assist with drafting and editing portions of the manuscript, help with code development and editing, and format tables.

All experimental decisions, implementations, analyses, and reported results were reviewed and verified by the authors. The authors take responsibility for the final content of the paper, including text, claims, analyses, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the experimental design and evaluation throughout the paper, appendix, and anonymous code repository. The main text specifies the grouped rotating leave-one-out design, provenance-effect definition, downstream OOD protocol, and bootstrap inference procedure. The appendix provides the complete semantic class assignments, dataset construction details, upstream training recipes for the ResNet, ViT-Tiny, and SimCLR experiments, downstream probe and detector configurations, and absolute AUROC results corresponding to the effect tables in the main paper. Within each upstream seed, encoder initialization is matched across the four rotations, downstream reference and evaluation sets are fixed within each dataset, and detector fitting uses downstream-ID data only.

REFERENCES

Vishnu Bindu Balachandran. Decodable but not detectable: A leakage fingerprint for near-ood benchmarks. *arXiv preprint arXiv:2607.19393*, 2026.

Sima Behpour, Thang Long Doan, Xin Li, Wenbin He, Liang Gou, and Liu Ren. Gradorth: A simple yet efficient out-of-distribution detection with orthogonal projection of gradients. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 38206–38230. Curran Associates, Inc., 2023. doi: 10.52202/075280-1659. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/77cf940349218069bbc230fc2c9c8a21-Paper-Conference.pdf.

Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. Neco: Neural collapse based out-of-distribution detection. In *International Conference on Learning Representations*, volume 2024, pp. 1182–1211, 2024.

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? Fixing ImageNet out-of-distribution detection evaluation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2471–2506. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/bitterwolf23a.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Md Yousuf Harun, Jhair Gallardo, and Christopher Kanan. Controlling neural collapse enhances out-of-distribution detection and transfer learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22149–22176. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/harun25a.html.

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.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009. URL https://cir.nii.ac.jp/crid/1370861707142497920.

Gerhard Krumpl, Henning Avenhaus, and Horst Possegger. One model, many behaviors: Training-induced effects on out-of-distribution detection. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4128–4138. IEEE, 2026.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/abdeb6f575ac5c6676b747bca8d09cc2-Paper.pdf.

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.

Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15424–15433, June 2025.

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

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Can pre-trained networks detect familiar out-of-distribution data? *arXiv preprint arXiv:2310.00847*, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International conference on machine learning*, pp. 20827–20840. PMLR, 2022.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12884–12893, 2021.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4911–4920. IEEE, 2022.

Hong Yang, Qi Yu, and Travis Desell. Can we ignore labels in out of distribution detection? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=falBlwUsIH.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

A Appendix

A.1 Downstream probe and evaluation protocol

After upstream training, encoders are placed in evaluation mode, and their parameters and normalization statistics are not updated during downstream evaluation. The downstream-ID reference bank contains only the 20 d_g classes, and the same reference and evaluation images are used for every

rotation and seed within a dataset. Cosine kNN uses $k = 50$ on ℓ_2-normalized features and scores a test example by its mean cosine distance to the 50 nearest downstream-ID reference features. Larger scores indicate more OOD, and AUROC is computed with OOD as the positive class.

For detectors that require classifier logits or weights, we fit a 20-way linear probe using only downstream-ID reference features. Future-OOD images are never used to train the probe or choose detector settings. On CIFAR-100, each upstream model receives a fresh `nn.Linear(512,20)` probe trained for 50 epochs with SGD, batch size 256, learning rate 0.1, and momentum 0.9; probe initialization is not matched across upstream models. On ImageNet, the 20-way probe is trained only on the d_g classes for 50 epochs with SGD, batch size 256, learning rate 0.1, and momentum 0.9. Within each upstream seed, the four ImageNet probes start from the same initialization. MSP and Energy use the raw probe logits with no additional evaluation temperature.

On iNaturalist 2021, the downstream classifier is a biased 20×2048 linear probe trained on the fixed 1,000-image downstream-ID reference bank. We train it for 50 epochs using cross-entropy, batch size 512, and SGD with learning rate 0.1, momentum 0.9, and zero weight decay, followed by epoch-wise cosine annealing to zero. The probe uses the default PyTorch linear-layer initialization. Initialization and data order are seeded by the upstream seed, giving matched initial probe states across the four rotations within each seed.

For the ViT studies, detectors requiring logits or classifier weights use a biased 20×192 linear probe trained only on the corresponding frozen downstream-ID reference features. Every probe is trained for 50 epochs using cross-entropy and SGD with learning rate 0.1, momentum 0.9, zero weight decay, and epoch-wise cosine annealing to zero. CIFAR-100 uses batch size 256, whereas ImageNet and iNaturalist use batch size 512. Probes use the default PyTorch linear-layer initialization, deterministically matched across the four rotations within each upstream seed. CIFAR-100 and iNaturalist use the upstream seed directly; ImageNet uses $10000 + s$.

Table 4: Downstream reference and evaluation sets. Image counts are unchanged across rotations and training seeds.

Dataset	ID reference	ID evaluation	OOD evaluation	Feature dim. (ResNet)
CIFAR-100	10,000	2,000	8,000	512
ImageNet	25,743	1,000	4,000	2,048
iNaturalist 2021	1,000	200	800	2,048

For CIFAR-100, the reference bank contains all 500 training images from each of the 20 downstream-ID classes and evaluation uses all 100 test images from each selected class. For ImageNet, the reference bank contains 25,743 training images from the 20 downstream-ID classes; evaluation uses 50 validation images from each selected class, giving 1,000 ID and 4,000 future-OOD images. For iNaturalist, the reference bank contains 50 training images from each downstream-ID species and evaluation uses 10 validation images per selected species, giving 200 ID and 800 future-OOD images. No iNaturalist validation image enters upstream training.

A.2 CIFAR-100 CONSTRUCTION AND SUPERVISED TRAINING

CIFAR-100 provides 20 predefined coarse categories with five fine classes in each category. We use each coarse category as one semantic group, assigning one fine class to d_g and the remaining four to $c_{g,1}, \ldots, c_{g,4}$. The complete assignment is shown in Table 5.

The supervised backbone is a CIFAR-style ResNet-18 with a 3×3 stride-1 stem and no max pooling. The downstream representation is the pooled 512-D encoder feature. Each upstream model is trained for 100 epochs with SGD, batch size 256, initial learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, cosine learning-rate decay, and automatic mixed precision. Training uses random crop with padding 4, random horizontal flip, and CIFAR-100 normalization. Within each training seed, the four rotations use matched encoder initialization and the same optimization and data pipeline.

Table 5: Complete CIFAR-100 semantic-group and role assignment.

Coarse category	d_g	$c_{g,1}$	$c_{g,2}$	$c_{g,3}$	$c_{g,4}$
aquatic mammals	seal	beaver	otter	dolphin	whale
fish	trout	flatfish	ray	aquarium fish	shark
flowers	tulip	poppy	rose	sunflower	orchid
food containers	cup	can	bottle	plate	bowl
fruit and vegetables	sweet pepper	pear	orange	apple	mushroom
household electrical devices	clock	television	lamp	keyboard	telephone
household furniture	couch	chair	bed	table	wardrobe
insects	beetle	cockroach	bee	butterfly	caterpillar
large carnivores	wolf	bear	tiger	leopard	lion
large man-made outdoor things	road	castle	skyscraper	bridge	house
large natural outdoor scenes	sea	cloud	mountain	forest	plain
large omnivores and herbivores	chimpanzee	cattle	camel	kangaroo	elephant
medium-sized mammals	possum	fox	porcupine	skunk	raccoon
non-insect invertebrates	snail	spider	worm	crab	lobster
people	baby	boy	woman	man	girl
reptiles	crocodile	snake	lizard	dinosaur	turtle
small mammals	rabbit	squirrel	hamster	shrew	mouse
trees	maple tree	oak tree	palm tree	pine tree	willow tree
vehicles 1	motorcycle	bicycle	pickup truck	bus	train
vehicles 2	streetcar	lawn mower	rocket	tractor	tank

A.3 IMAGENET SEMANTIC CONSTRUCTION AND SUPERVISED TRAINING

ImageNet does not provide a fixed five-class grouping analogous to CIFAR-100, so we construct one from the WordNet hierarchy. We identified 121 WordNet parents with at least five eligible ImageNet descendants. Within each parent, descendants were ordered by relative WordNet depth and then by WNID, and we kept the first five. Our first deterministic rule selected 20 animal-related parents, so before any model training we changed the parent-selection rule to obtain a more semantically varied set. Using the same 121 eligible parents, we selected 20 class-disjoint, non-nested parents that maximized WordNet separation and lexname diversity. The class and role assignments were set before training and were not changed after observing OOD results.

The resulting construction contains 20 downstream-ID classes and 80 rotating future-OOD classes. Native ILSVRC class sizes are retained rather than resampled, giving 102,215–102,990 upstream images across the four rotations, a spread of 775 images (less than 1%). ResNet-50 models are trained from random initialization for 100 epochs with SGD, momentum 0.9, weight decay 10^{-4}, effective batch size 256, initial learning rate 0.1, five epochs of linear warmup followed by cosine decay, automatic mixed precision, `RandomResizedCrop(224)`, random horizontal flip, and ImageNet normalization. Features are the 2,048-D global-average-pooled representation extracted in evaluation mode.

Table 6: Complete ImageNet class assignments. WNIDs are shown beneath class names.

WordNet parent	d_g	$c_{g,1}$	$c_{g,2}$	$c_{g,3}$	$c_{g,4}$
bovid.n.01	impala n02422699	hartebeest n02422106	water buffalo n02408429	ram n02412080	bison n02410509
dish.n.02	burrito n07880968	meat loaf n07871810	potpie n07875152	pizza n07873807	consomme n07584110
ship.n.01	wreck n04606251	submarine n04347754	container ship n03095699	pirate n03947888	aircraft carrier n02687172
butterfly.n.01	admiral n02276258	sulphur butterfly n02281406	monarch n02279972	ringlet n02277742	lycaenid n02281787
fruit.n.01	banana n07753592	custard apple n07760859	fig n07753113	pineapple n07753275	jackfruit n07754684
kitchen appliance.n.01	microwave n03761084	waffle iron n04542943	toaster n04442312	espresso maker n03297495	Dutch oven n03259280
shop.n.01	bookshop n02871525	bakery n02776631	confectionery n03089624	barbershop n02791270	butcher shop n02927161
colubrid snake.n.01	hognose snake n01729322	green snake n01729977	thunder snake n01728572	ringneck snake n01728920	king snake n01734418
retriever.n.01	curly-coated retriever n02099429	Chesapeake Bay retriever n02099849	Labrador retriever n02099712	flat-coated retriever n02099267	golden retriever n02099601
cooking utensil.n.01	coffeepot n03063689	Crock Pot n03133878	caldron n02939185	teapot n04398044	frying pan n03400231

Table 6 continued

WordNet parent	d_g	$c_{g,1}$	$c_{g,2}$	$c_{g,3}$	$c_{g,4}$
vegetable.n.01	cardoon n07730033	cucumber n07718472	mushroom n07734744	mashed potato n07711569	artichoke n07718747
timepiece.n.01	hourglass n03544143	sundial n04355338	digital clock n03196217	digital watch n03197337	analog clock n02708093
finch.n.01	brambling n01530575	indigo bunting n01537544	junco n01534433	goldfinch n01531178	house finch n01532829
teleost fish.n.01	sturgeon n02640242	puffer n02655020	gar n02641379	coho n02536864	eel n02526121
footwear.n.02	sandal n04133789	running shoe n04120489	clog n03047690	cowboy boot n03124043	Loafer n03680355
truck.n.01	minivan n03770679	tow truck n04461696	pickup n03930630	garbage truck n03417042	fire engine n03345487
new world monkey.n.01	titi n02493509	marmoset n02490219	howler monkey n02492660	capuchin n02492035	spider monkey n02493793
ball.n.01	golf ball n03445777	basketball n02802426	croquet ball n03134739	ping-pong ball n03942813	baseball n02799071
seat.n.03	park bench n03891251	folding chair n03376595	barber chair n02791124	rocking chair n04099969	toilet seat n04447861
fungus.n.01	gyromitra n13037406	earthstar n13044778	hen-of-the-woods n13052670	coral fungus n12985857	stinkhorn n13040303

A.4 iNATURALIST 2021 CONSTRUCTION AND SUPERVISED TRAINING

We use genus as the semantic parent for the iNaturalist 2021 study. Among the available classes, 364 genera contained at least five eligible species. We sorted the eligible genera lexicographically and selected the first 20. Within each selected genus, we sorted eligible species by scientific name and selected the first five. This deterministic rule was chosen to avoid manually picking groups based on their expected OOD behavior. The five selected species were then assigned to $d_g, c_{g,1}, \ldots, c_{g,4}$ in that order before model training or OOD evaluation. Table 7 gives the complete 100-species assignment.

For the supervised iNaturalist study, we use every available official training image for the 100 selected species rather than balancing class sizes. The selected species contain 162 to 300 training images each (median 300, mean 277.13), for 27,713 training images in total. Because one future-OOD species per genus is withheld in each rotation, the four upstream training sets contain 22,289, 22,142, 22,035, and 22,058 images. The maximum spread is 254 images, or 1.148% of the mean rotation size.

The supervised backbone is a randomly initialized ResNet-50 trained with the same 224-pixel recipe used for the ImageNet study: 100 epochs, SGD with initial learning rate 0.1, momentum 0.9, weight decay 10^{-4}, effective batch size 256, five epochs of linear warmup followed by cosine decay, automatic mixed precision, `RandomResizedCrop(224)`, random horizontal flip, and ImageNet normalization. The downstream representation is the 2,048-D global-average-pooled feature. Within each seed, the four rotations start from the same model initialization. The downstream-ID reference bank contains the 1,000 images described in Appendix A.1; evaluation uses the same 200 ID and 800 future-OOD validation images for every rotation and seed.

Table 7: Complete iNaturalist 2021 genus/species role assignment.

Group	Genus	d_g	$c_{g,1}$	$c_{g,2}$	$c_{g,3}$	$c_{g,4}$
G001	Abies	*Abies alba*	*Abies balsamea*	*Abies concolor*	*Abies grandis*	*Abies lasiocarpa*
G002	Acacia	*Acacia confusa*	*Acacia cyclops*	*Acacia dealbata*	*Acacia longifolia*	*Acacia mearnsii*
G003	Accipiter	*Accipiter badius*	*Accipiter cooperii*	*Accipiter gentilis*	*Accipiter nisus*	*Accipiter striatus*
G004	Acer	*Acer campestre*	*Acer circinatum*	*Acer floridanum*	*Acer ginnala*	*Acer glabrum*
G005	Acmispon	*Acmispon americanus*	*Acmispon argophyllus*	*Acmispon glaber*	*Acmispon rigidus*	*Acmispon strigosus*
G006	Aconitum	*Aconitum columbianum*	*Aconitum delphinifolium*	*Aconitum lycoctonum*	*Aconitum napellus*	*Aconitum septentrionale*
G007	Acrocephalus	*Acrocephalus arundinaceus*	*Acrocephalus australis*	*Acrocephalus dumetorum*	*Acrocephalus schoenobaenus*	*Acrocephalus scirpaceus*
G008	Acronicta	*Acronicta aceris*	*Acronicta americana*	*Acronicta fallax*	*Acronicta impleta*	*Acronicta innotata*
G009	Adiantum	*Adiantum aleuticum*	*Adiantum capillus-veneris*	*Adiantum cunninghamii*	*Adiantum hispidulum*	*Adiantum jordanii*
G010	Aeshna	*Aeshna canadensis*	*Aeshna constricta*	*Aeshna cyanea*	*Aeshna eremita*	*Aeshna grandis*
G011	Agave	*Agave americana*	*Agave deserti*	*Agave lechuguilla*	*Agave parryi*	*Agave salmiana*
G012	Agrotis	*Agrotis exclamationis*	*Agrotis infusa*	*Agrotis ipsilon*	*Agrotis munda*	*Agrotis porphyricollis*

Table 7 continued

Group	Genus	d_g	$c_{g,1}$	$c_{g,2}$	$c_{g,3}$	$c_{g,4}$
G013	Allium	*Allium acuminatum*	*Allium canadense*	*Allium cernuum*	*Allium drummondii*	*Allium neapolitanum*
G014	Alnus	*Alnus alnobetula*	*Alnus glutinosa*	*Alnus incana*	*Alnus rhombifolia*	*Alnus rubra*
G015	Amanita	*Amanita augusta*	*Amanita bisporigera*	*Amanita citrina*	*Amanita flavoconia*	*Amanita fulva*
G016	Amazilia	*Amazilia beryllina*	*Amazilia cyanocephala*	*Amazilia rutila*	*Amazilia tzacatl*	*Amazilia violiceps*
G017	Ambrosia	*Ambrosia ambrosioides*	*Ambrosia artemisiifolia*	*Ambrosia chamissonis*	*Ambrosia deltoidea*	*Ambrosia dumosa*
G018	Ambystoma	*Ambystoma gracile*	*Ambystoma jeffersonianum*	*Ambystoma laterale*	*Ambystoma macrodactylum*	*Ambystoma maculatum*
G019	Anas	*Anas acuta*	*Anas bahamensis*	*Anas castanea*	*Anas crecca*	*Anas diazi*
G020	Anaxyrus	*Anaxyrus americanus*	*Anaxyrus boreas*	*Anaxyrus cognatus*	*Anaxyrus fowleri*	*Anaxyrus punctatus*

A.5 OOD DETECTOR CONFIGURATIONS

We choose all detector settings without using the 80 future-OOD classes or their images. Any quantities that must be estimated from data are fit using downstream-ID reference data only. Scores are oriented so that larger values are more OOD-like.

kNN. Cosine kNN uses the protocol in Appendix A.1 with $k = 50$.

Energy and MSP. Energy and MSP use logits from the 20-way downstream probe trained on encoder features. We use $-\log \sum_c \exp z_c$ for Energy and $1 - \max_c \mathrm{softmax}(z)_c$ for MSP, with no additional evaluation temperature.

Mahalanobis. We use only the final encoder feature for Mahalanobis scoring (Lee et al., 2018). Class means and a tied empirical covariance are fit from downstream-ID references, and the Moore–Penrose precision is used for the minimum class-conditional quadratic distance. We use no input preprocessing, covariance shrinkage, learned layer ensemble, or OOD calibration.

ViM. We use final-feature Virtual-logit Matching (Wang et al., 2022) with the downstream probe and ID reference features. The virtual origin is $u = -W^+ b$. We use a 256-D principal subspace for the 512-D ResNet-18 representation and a 1,000-D subspace for the 2,048-D ResNet-50 representation. For the 192-D ViT representation, we use the same $D/2$ rule, giving a 96-D principal subspace. The residual-to-logit scale is fit from ID references only. The iNaturalist ResNet-50 fit is numerically degenerate because the 1,000-reference fit and 1,000-dimensional principal subspace leave residual norms near machine precision; we retain the finite result without retuning.

NECO. For NECO (Ben Ammar et al., 2024), a standardizer and PCA are fit using downstream-ID reference features. We retain the smallest principal dimension explaining at least 90% of ID variance and use the standalone feature score without MaxLogit multiplication. The same rule is used for ResNet and ViT features.

NCI. For NCI (Liu & Qin, 2025), we use the downstream probe and global ID-reference feature mean. We set $\alpha = 10^{-2}$ for CIFAR-100/ResNet-18, $\alpha = 10^{-3}$ for the 224-pixel ResNet-50 studies, and $\alpha = 10^{-2}$ for the 192-D ViT studies. We chose the ViT value before OOD evaluation and did not tune it on the future-OOD classes. We report the sign-reversed score so that larger values indicate more OOD.

GradOrth. For GradOrth (Behpour et al., 2023), we use all downstream-ID reference features at the final layer rather than the random efficiency subsample used in the original paper. We fit an uncentered SVD and keep the smallest subspace that explains 97% of the squared singular-value energy. The downstream-probe weight-gradient projection norm is then computed in closed form.

ODIN. ODIN (Liang et al., 2018) uses the trained encoder and downstream probe with their weights held fixed during scoring and $T = 1000$. We use $\epsilon = 0.002$ for CIFAR-100 and $\epsilon = 0.0014$ for the 224-pixel ImageNet and iNaturalist studies, including their ViT counterparts. Scoring uses FP32 input

gradients, one forward–backward–forward pass, and no input clipping. We do not tune temperature or perturbation magnitude on the future-OOD classes.

A.6 Vision Transformer architecture robustness

All ViT models are trained from random initialization with no external pretrained weights. We use ViT-Tiny with depth 12, hidden dimension 192, three attention heads, MLP dimension 768, and the final LayerNorm CLS token as the 192-D downstream representation. CIFAR-100 uses 4×4 patches for 32×32 inputs; ImageNet and iNaturalist use 16×16 patches for 224×224 inputs. The semantic groups, rotations, seeds, downstream roles, reference images, and evaluation images are identical to the corresponding ResNet study.

We chose the ViT training recipe using classification generalization rather than OOD performance, and then used the same recipe for every rotation and seed. It uses AdamW with learning rate 3×10^{-4} and weight decay 0.05, batch size 256, 300 epochs, five warmup epochs followed by cosine decay, label smoothing 0.1, Mixup 0.8, CutMix 1.0, RandAugment with two operations at magnitude 9, random erasing 0.25, and drop path 0.1. CIFAR-100 uses random crop and horizontal flip; the 224-pixel studies use random resized crop and horizontal flip. Mean classification accuracy on the selected-class test split is 73.13% on CIFAR-100, 81.07% on ImageNet, and 59.50% on iNaturalist 2021.

The ViT detector-specific settings that differ because of representation dimension are given explicitly in Appendix A.5: ViM uses a 96-D principal/residual split, NCI uses fixed $\alpha = 0.01$, and Mahalanobis, NECO, and GradOrth operate directly on the 192-D CLS representation. All detector fitting uses downstream-ID data only.

A.7 SimCLR control protocol

The SimCLR controls use the same semantic groups, four rotations, two training seeds, downstream tasks, reference images, and evaluation images as the corresponding supervised ResNet experiments. Class labels are used only to decide which images belong to each rotation; they are not passed to the contrastive objective. After training, we discard the projection head, place the encoder in evaluation mode, and use the pre-projection representation for downstream evaluation without updating the encoder.

CIFAR-100. The encoder is trained for 200 epochs with NT-Xent/InfoNCE, two views per image, batch size 256, temperature 0.5, SGD learning rate 0.5, momentum 0.9, weight decay 10^{-4}, cosine decay, and automatic mixed precision. The projector is $512 \to 512 \to 128$. Augmentations are $\texttt{RandomResizedCrop}(32, \text{scale} = (0.2, 1.0))$, random horizontal flip, $\text{ColorJitter}(0.4, 0.4, 0.4, 0.1)$ applied with probability 0.8, RandomGrayscale with probability 0.2, and CIFAR-100 normalization. The downstream representation is the pooled 512-D encoder feature.

ImageNet. The ImageNet control uses a randomly initialized ResNet-50, 200 epochs, batch size 256 (512 augmented representations in each NT-Xent loss), temperature 0.5, SGD learning rate 0.5, momentum 0.9, weight decay 10^{-4}, cosine decay without warmup, automatic mixed precision, channels-last execution, and activation checkpointing. The projector is $2048 \to 512 \to 128$ with BatchNorm.

Each of the two views independently applies $\texttt{RandomResizedCrop}(224, \text{scale} = (0.2, 1.0))$, random horizontal flip with probability 0.5, $\text{ColorJitter}(0.4, 0.4, 0.4, 0.1)$ with probability 0.8, RandomGrayscale with probability 0.2, and ImageNet normalization; Gaussian blur is not used. Downstream evaluation uses the 2,048-D pre-projector representation.

iNaturalist 2021. The iNaturalist control reuses the same 20-genus/100-species construction and downstream task as the supervised study. We train randomly initialized ResNet-50 encoders for 200 epochs with NT-Xent, batch size 256 (512 augmented representations per loss), temperature 0.5, SGD learning rate 0.5, momentum 0.9, weight decay 10^{-4}, cosine decay without warmup, automatic mixed precision, and channels-last execution. The projector is $2048 \to 512 \to 128$ with BatchNorm. Augmentations are $\texttt{RandomResizedCrop}(224, \text{scale} = (0.2, 1.0))$, random horizontal flip, $\text{ColorJitter}(0.4, 0.4, 0.4, 0.1)$ applied with probability 0.8, RandomGrayscale with

probability 0.2, and ImageNet normalization. Downstream evaluation uses the frozen 2,048-D pre-projector representation.

Nine-detector evaluation after SimCLR. After discarding the projection head, we apply the same downstream probe and detector definitions used for the corresponding supervised ResNet study. All detector fitting uses downstream-ID data only, and future-OOD images are used only for final scoring and AUROC computation. Thus, the SimCLR comparison changes the upstream training regime while keeping the downstream evaluation procedure the same.

The primary kNN presence-versus-absence effects are $+1.1$ AUROC points on CIFAR-100, -0.7 on ImageNet, and -3.2 on iNaturalist. We also compute a paired supervised-minus-SimCLR contrast for each class, $\Delta_{\mathrm{SUP}}(c) - \Delta_{\mathrm{SSL}}(c)$, comparing the provenance effect under the two training regimes for the same future-OOD class. The dataset means are -8.6, -5.6, and -2.8 points, respectively. We use this only as a descriptive comparison between the two training setups. Because supervised training and SimCLR differ in objective, optimization, augmentation, and the use of a projection head, this difference should not be interpreted as the causal effect of labels alone.

A.8 Detector-ordering reversal definition

For detectors D_1 and D_2, define their class-level AUROC gap in condition h as

$$G_h(c) = A_{D_1}(c \mid h) - A_{D_2}(c \mid h).$$

The withheld gap uses the class's single withheld rotation, while the supervised gap uses the mean AUROC of each detector across the same three supervised rotations used in Δ_D. We first compute these gaps separately for each upstream training seed and then average them across the two seeds. We count an ordering reversal when

$$G_{\mathrm{withheld}}(c)\, G_{\mathrm{supervised}}(c) < 0.$$

Exact-zero gaps are treated as ties and are not counted as reversals. For the kNN–Energy comparison reported in the main paper, the observed reversal counts are 24/80, 31/80, and 24/80 for the ResNet studies on CIFAR-100, ImageNet, and iNaturalist, and 36/80, 28/80, and 26/80 for the corresponding ViT studies.

A.9 Representation-geometry protocol

The geometry analysis uses the ResNet representations from the primary supervised experiments. Features are ℓ_2-normalized before centroid construction. Each downstream-ID class centroid is formed from that class's reference features, and each future-OOD class centroid is formed from the same evaluation images used for OOD scoring. For a future-OOD class, we compute cosine distance from its centroid to the nearest of the 20 downstream-ID class centroids.

For each class, the withheld distance is taken from its single withheld rotation, the supervised distance is averaged across its three supervised rotations, and the resulting contrast is then averaged across the two seeds in the same way as in the primary OOD analysis. The geometry contrast shown in the main figure is supervised minus withheld, while the kNN AUROC contrast is withheld minus supervised. Pearson and Spearman correlations are computed across the 80 future-OOD class effects within each dataset. The reported Pearson intervals use the same bootstrap over the 20 semantic groups as the main analysis. The geometry analysis is descriptive and associative; it does not establish mediation or a causal mechanism.

A.10 Additional detector results

Tables 8–10 report the corresponding absolute AUROC values for the supervised ResNet, ViT-Tiny, and SimCLR experiments. Tables 11–15 summarize uncertainty and sign consistency for the detector suite. Effects and confidence intervals are reported in AUROC percentage points.

Table 8: Absolute OOD AUROC for the supervised ResNet experiments. S and W denote supervised and withheld upstream treatment, respectively. Values are mean AUROC across the two training seeds, in percentage points.

	CIFAR-100 / R18		WordNet-structured ImageNet subset / R50		iNaturalist 2021 / R50	
Detector	S	W	S	W	S	W
kNN	94.60	87.11	93.41	87.09	69.37	63.46
Energy	90.46	86.40	88.70	85.74	66.68	67.76
MSP	88.80	84.99	88.08	85.87	68.50	68.24
Mahalanobis	87.96	74.00	83.68	71.87	71.21	63.13
ViM	91.60	83.19	92.39	86.72	70.13	61.84
NECO	85.13	79.96	78.23	77.39	70.74	67.23
NCI	93.23	86.71	92.39	87.51	62.98	63.44
GradOrth	80.60	82.57	72.32	73.95	52.23	56.69
ODIN	91.51	86.52	89.35	85.41	66.65	66.95

Table 9: Absolute OOD AUROC for the supervised ViT-Tiny experiments. S and W denote supervised and withheld upstream treatment, respectively. Values are mean AUROC across the two training seeds, in percentage points.

	CIFAR-100 / ViT-Tiny		WordNet-structured ImageNet subset / ViT-Tiny		iNaturalist 2021 / ViT-Tiny	
Detector	S	W	S	W	S	W
kNN	88.23	83.72	88.65	86.06	79.30	74.40
Energy	90.79	83.95	93.70	86.84	79.97	75.26
MSP	88.46	82.16	92.21	85.78	76.18	72.40
Mahalanobis	87.46	83.63	91.97	87.61	74.53	72.61
ViM	84.46	82.28	91.40	86.88	75.08	72.80
NECO	81.87	80.09	85.96	83.57	66.23	67.41
NCI	90.67	83.96	93.51	86.90	80.06	75.46
GradOrth	86.14	79.31	88.24	79.77	73.76	71.35
ODIN	91.28	83.12	91.58	81.44	80.12	73.90

Table 10: Absolute OOD AUROC for the SimCLR experiments. P and A denote unlabeled upstream presence and upstream absence, respectively. Values are mean AUROC across the two training seeds, in percentage points.

	CIFAR-100 / R18 SimCLR		WordNet-structured ImageNet subset / R50 SimCLR		iNaturalist 2021 / R50 SimCLR	
Detector	P	A	P	A	P	A
kNN	72.61	73.72	83.72	82.99	65.90	62.75
Energy	64.47	73.47	79.58	81.79	54.52	55.75
MSP	65.04	72.51	79.66	81.91	57.24	59.61
Mahalanobis	69.30	53.83	85.45	75.48	64.65	58.11
ViM	76.81	74.79	87.58	83.66	64.97	59.03
NECO	71.70	71.65	70.23	73.99	60.42	61.70
NCI	71.36	72.83	84.95	83.23	60.77	60.64
GradOrth	59.22	71.65	75.07	79.98	54.56	59.00
ODIN	65.81	73.12	81.61	81.84	57.36	58.66

Table 11: 95% bootstrap intervals for supervised ResNet provenance effects, resampling the 20 semantic groups. Values are in AUROC points.

Detector	CIFAR-100	ImageNet	iNaturalist 2021
kNN	$[-9.0, -6.1]$	$[-7.8, -5.0]$	$[-7.7, -4.2]$
Energy	$[-5.2, -3.0]$	$[-4.1, -1.9]$	$[-0.3, +2.9]$
MSP	$[-4.9, -2.8]$	$[-3.3, -1.2]$	$[-1.4, +0.9]$
Mahalanobis	$[-15.7, -12.3]$	$[-13.5, -10.2]$	$[-10.1, -6.3]$
ViM	$[-9.7, -7.3]$	$[-7.1, -4.4]$	$[-10.1, -6.7]$
NECO	$[-7.2, -3.1]$	$[-2.9, +1.0]$	$[-4.8, -2.2]$
NCI	$[-8.0, -5.1]$	$[-6.2, -3.7]$	$[-1.4, +2.9]$
GradOrth	$[+1.4, +2.5]$	$[+0.0, +3.1]$	$[+2.1, +7.9]$
ODIN	$[-6.2, -3.8]$	$[-5.3, -2.8]$	$[-1.2, +2.1]$

Table 12: Negative/positive class and group counts for the supervised ResNet detector suite.

Detector	CIFAR-100 Cls	CIFAR-100 Grp	ImageNet Cls	ImageNet Grp	iNaturalist Cls	iNaturalist Grp
kNN	80/0	20/0	80/0	20/0	67/13	19/1
Energy	75/5	20/0	66/14	19/1	36/44	6/14
MSP	73/7	20/0	63/17	17/3	42/38	13/7
Mahalanobis	80/0	20/0	80/0	20/0	75/5	20/0
ViM	80/0	20/0	80/0	20/0	76/4	20/0
NECO	54/26	17/3	37/43	11/9	60/20	16/4
NCI	79/1	20/0	80/0	20/0	40/40	10/10
GradOrth	21/59	1/19	30/50	4/16	22/58	3/17
ODIN	79/1	20/0	73/7	20/0	40/40	10/10

Table 13: Mean supervised provenance effects across detectors and architectures, in AUROC points. Negative values indicate better OOD detection when the future-OOD class was supervised upstream.

	CIFAR-100		ImageNet		iNaturalist 2021	
Detector	R18	ViT	R50	ViT	R50	ViT
kNN	-7.5	-4.5	-6.3	-2.6	-5.9	-4.9
Energy	-4.1	-6.8	-3.0	-6.9	$+1.1$	-4.7
MSP	-3.8	-6.3	-2.2	-6.4	-0.3	-3.8
Mahalanobis	-14.0	-3.8	-11.8	-4.4	-8.1	-1.9
ViM	-8.4	-2.2	-5.7	-4.5	-8.3	-2.3
NECO	-5.2	-1.8	-0.8	-2.4	-3.5	$+1.2$
NCI	-6.5	-6.7	-4.9	-6.6	$+0.5$	-4.6
GradOrth	$+2.0$	-6.8	$+1.6$	-8.5	$+4.5$	-2.4
ODIN	-5.0	-8.2	-3.9	-10.1	$+0.3$	-6.2

Table 14: 95% bootstrap intervals for ViT provenance effects, resampling the 20 semantic groups. Values are in AUROC points.

Detector	CIFAR-100	ImageNet	iNaturalist 2021
kNN	$[-5.8, -3.4]$	$[-3.6, -1.6]$	$[-6.8, -3.1]$
Energy	$[-8.5, -5.4]$	$[-8.5, -5.4]$	$[-6.3, -3.1]$
MSP	$[-7.9, -4.9]$	$[-8.0, -5.1]$	$[-5.4, -2.2]$
Mahalanobis	$[-5.4, -2.5]$	$[-5.7, -3.1]$	$[-3.5, -0.5]$
ViM	$[-3.6, -1.0]$	$[-5.9, -3.3]$	$[-3.7, -0.9]$
NECO	$[-4.0, +0.4]$	$[-4.4, -0.4]$	$[-0.3, +2.7]$
NCI	$[-8.3, -5.3]$	$[-8.2, -5.2]$	$[-6.2, -3.0]$
GradOrth	$[-8.4, -5.4]$	$[-10.3, -7.0]$	$[-4.2, -0.7]$
ODIN	$[-9.8, -6.7]$	$[-12.7, -7.9]$	$[-8.3, -4.2]$

Table 15: ViT sign consistency. Each cell reports negative classes / 80, followed by negative semantic groups / 20.

Detector	CIFAR-100	ImageNet	iNaturalist 2021
kNN	76/80, 20/20	65/80, 18/20	59/80, 19/20
Energy	79/80, 20/20	80/80, 20/20	60/80, 18/20
MSP	78/80, 20/20	80/80, 20/20	57/80, 17/20
Mahalanobis	67/80, 20/20	73/80, 19/20	45/80, 11/20
ViM	56/80, 16/20	75/80, 19/20	50/80, 14/20
NECO	44/80, 12/20	51/80, 14/20	34/80, 7/20
NCI	78/80, 20/20	80/80, 20/20	61/80, 18/20
GradOrth	73/80, 20/20	79/80, 20/20	53/80, 13/20
ODIN	80/80, 20/20	80/80, 20/20	63/80, 18/20