ONE SCORE FOR ALL: ROBUST OOD DETECTION ACROSS PRETRAINED MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-hoc out-of-distribution (OOD) detection should remain useful across pretrained classifiers and calibration budgets. In practice, however, changes in the pretrained model can reverse the in-distribution (ID)/OOD ordering of feature statistics, while class-relative references can become unreliable when estimated from few samples. We propose *ONE*, a common scoring and fitting procedure that uses pretrained class information both to combine detection evidence and to estimate its reference. Raw logits and Gaussian-form fit in normalized feature space are combined within corresponding classes before pooling. The resulting score retains each branch's overall evidence together with their soft class overlap, rather than discarding class correspondence through scalar fusion. We estimate the feature reference by shrinking empirical class centers toward classifier-derived priors according to sample counts and estimated mismatch, while regularizing covariance according to sample availability. The reference and separate evidence scales are fitted using labeled ID data, with the network frozen. Across nine benchmark settings, *ONE* achieves 86.3 mean AUROC with full calibration. On 30 public ImageNet checkpoints, it achieves 85.1 mean AUROC and a 6.8-point maximum gap to the best evaluated score on each checkpoint, improving both criteria over the evaluated alternatives. With one labeled ID calibration sample per class, it achieves higher mean AUROC (84.2) than the classifier-only GEN baseline (83.5).

1 INTRODUCTION

Post-hoc out-of-distribution (OOD) detection aims to equip an existing classifier with a rejection rule without retraining the network (Hendrycks & Gimpel, 2016; Liu et al., 2020). In practice, however, changing the pretrained classifier can change which OOD scoring method works best. A method that performs well on one model may fail on another, while selecting a replacement method using OOD validation data requires representative OOD examples and makes the choice depend on that validation set. We instead investigate a common scoring and fitting procedure that uses only labeled in-distribution (ID) data across architectures, training recipes, and calibration budgets.

Two forms of dependence complicate this goal. **Model dependence** concerns both detection performance and the interpretation of the underlying statistics. The relative performance of OOD scores *varies across pretrained classifiers* (Figure 1a,b), and even a simple statistic can reverse its ID/OOD ordering: in our fixed-architecture ImageNet comparison, larger feature norms favor ID inputs for the plain cross-entropy model but OOD inputs for its label-smoothed counterpart (Figure 2). **Sample dependence** arises because class-relative feature methods rely on references estimated from labeled ID data, such as stored features, class centers, or covariance statistics. With few calibration examples, these references can become unreliable, and the resulting detection performance depends on *how many ID examples* are available for estimation (Figure 1c).

To address these two forms of dependence, we propose *ONE*, a scoring and fitting procedure built from two class-indexed measurements. **Classifier support** is the raw logit for a candidate class. **Class-relative feature fit** measures how closely the normalized feature matches that class's estimated reference, using within-class variation to define the distance. Their relative detection performance changes across checkpoints and OOD sets (Figure 3). We use the classifier both to supply input-specific responses and to inform the estimation of class locations.

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.

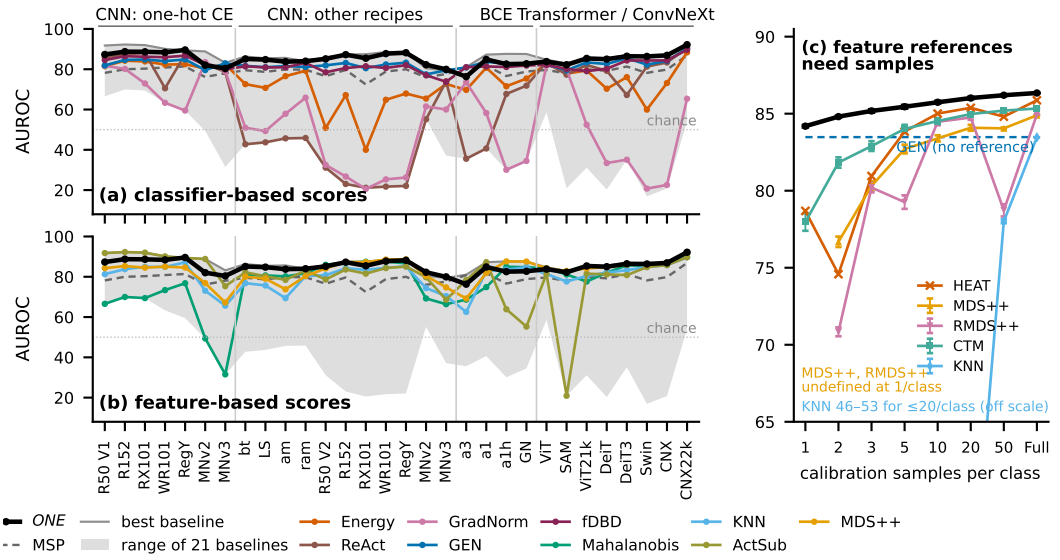


Figure 1: **OOD detection performance depends on the pretrained model and calibration budget.** (a, b) AUROC averaged over eight OOD datasets on 30 public ImageNet checkpoints. The gray band spans the 21 baselines included in these plots; Table 2 reports the expanded comparison. (c) Mean AUROC over nine settings at each balanced per-class budget; error bars show the standard deviation over four random draws of the per-class samples; HEAT uses one draw because it is retrained for each draw.

The first distinction is *where evidence is combined*. Two sources can each produce a strong scalar score while favoring different classes. We therefore combine the classifier's logit and Gaussian-form feature fit within each candidate class before pooling. An exact decomposition separates the resulting score into the two branches' overall evidence and a soft class-overlap term. This retains class correspondence that scalar fusion discards, while log-sum-exp aggregation preserves alternatives to the top prediction.

The second distinction is *how the reference is estimated*. We treat classifier weight directions as informative but imperfect priors for class centers and shrink empirical means toward them. Class counts determine the available sample evidence, while estimated classifier–sample mismatch determines a shared prior strength. Covariance shrinkage separately regulates the distance metric when samples are scarce. Both evidence scales are fitted separately using ID labels, with no network updates or OOD-dependent weight search. The same scoring form and fitting rules are fixed across models, while class references and evidence scales are fitted separately for each model.

With full ID calibration, *ONE* achieves 86.3 mean AUROC across nine benchmark settings. On 30 public ImageNet checkpoints, it reaches 85.1 mean AUROC and a 6.8-point maximum gap to the best evaluated score on each checkpoint, improving both criteria over the evaluated baselines. With balanced calibration using just one sample per class, it exceeds GEN (Liu et al., 2023) on average. Matched controls separate the contribution of class correspondence from that of the fitted branches.

Our contributions are: (1) an analysis of model and sample dependence in OOD evidence; (2) *ONE*, a class-aligned score with classifier-informed reference estimation and ID-only scale fitting; and (3) cross-model and calibration-budget evaluations with matched combination controls.

2 Related Work

We focus on post-hoc OOD detection with frozen classifiers and labeled ID calibration data. Prior work has documented sensitivity to architecture and training methodology (Kim et al., 2024). We organize the relevant methods around the two measurements, *classifier support* and *class-relative feature fit*, and how they are estimated and combined.

Table 1: **Evidence-combination methods by ingredient and by combination.** Class correspondence: whether and how the combination uses which class each source supports.

	Ingredients		Combination	
Method	Classifier evidence	Feature evidence	Class correspondence	Rule
ViM	energy	residual norm	none	difference
NNGuide	energy	kNN similarity	none	product
ActSub	subspace energy	subspace cosine	none	product
HEAT	energy	class Gaussians	none	standardized sum
MME	SCALE, ViM, fDBD	PCA, nearest mean	top-1 agreement, fixed factor	product
COMBOOD	none	kNN, Mahalanobis	none	sum
ONE	per-class logits	per-class Gaussian fit	soft overlap, all classes	sum per class, LSE

Classifier Evidence. MSP, GEN, and TRIM summarize normalized class probabilities (Hendrycks & Gimpel, 2016; Liu et al., 2023; Kim et al., 2024), while MaxLogit, Energy, and LogitGap use unnormalized logits (Hendrycks et al., 2022; Liu et al., 2020; Liang et al., 2025). GradNorm uses gradients from divergence to a uniform output distribution (Huang et al., 2021). ReAct, ASH, SCALE, and DICE modify activations or classifier weights (Sun et al., 2021; Djurisic et al., 2023; Xu et al., 2024; Sun & Li, 2022). *ONE*'s classifier branch retains the full original logit vector so that each response can be combined with feature fit to the corresponding class.

Feature References. Feature references differ in what they keep from the ID data. KNN is nonparametric and keeps the ID features themselves (Sun et al., 2022). The others are parametric. CTM and SHE keep one mean per class and score by cosine similarity or dot product (Ngoc-Hieu et al., 2023; Zhang et al., 2023). The Mahalanobis family adds a shared covariance: Mahalanobis detection fits class centers and covariance on raw features (Lee et al., 2018), relative Mahalanobis subtracts a class-agnostic reference (Ren et al., 2021), Mahalanobis++ normalizes features before fitting (Müller & Hein, 2025), and GEM pools the class Gaussians with log-sum-exp (Morteza & Li, 2022). fDBD uses classifier decision-boundary geometry, while NCI uses classifier-weight geometry as detection evidence (Liu & Qin, 2024; 2025). *ONE* uses classifier directions as imperfect priors, balancing them against empirical means according to sample counts and estimated mismatch while separately regularizing covariance.

Evidence Combination. In Table 1, ViM, NNGuide, COMBOOD, ActSub, and HEAT combine complementary detection signals through different constructions (Wang et al., 2022; Park et al., 2023; Rajasekaran et al., 2024; Zöngür et al., 2025; Lafon et al., 2023). MME provides a close precedent for class correspondence through a hard classifier–nearest-mean consistency factor (Rakotoarivony, 2025). The distinction of *ONE* is its *specific combination* of class-aligned scoring and classifier-informed reference estimation. The score preserves soft class correspondence alongside the overall evidence of both branches, while the reference estimator adjusts reliance on pretrained class geometry according to the available ID observations.

3 Model and Sample Dependence of OOD Evidence

We examine two questions that guide our construction: (1) how the interpretation of detection evidence changes across pretrained models, and (2) how reliably a class reference can be estimated from limited ID data. Each reported AUROC averages eight OOD datasets (Appendix A). When a statistic is treated as an ID score, AUROC above 50 indicates that larger values favor ID inputs, whereas AUROC below 50 indicates the opposite.

3.1 Detection Signals Change with the Pretrained Model

Across 30 public ImageNet checkpoints spanning training recipes and architectures, Figure 2 compares five measurements: maximum logit, maximum softmax probability, feature norm, activation concentration, and maximum cosine similarity to ID class centers. Maximum logit and class similarity exceed 60 AUROC on all 30 public ImageNet checkpoints. Feature norm does so on only four and falls below 40 on 17; activation concentration also changes direction across the panel. Model

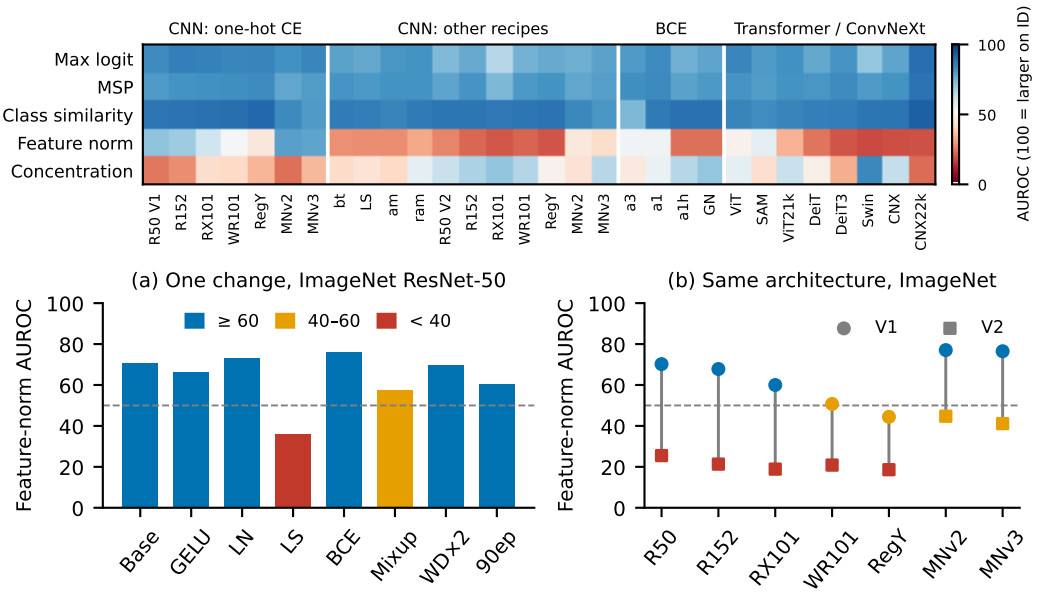


Figure 2: **Feature statistics can change their ID/OOD ordering across models. Top:** AUROC of five statistics (defined in Appendix A) on the 30 public ImageNet checkpoints. Values above 50 (blue) indicate that larger scores favor ID inputs; values below 50 (red) indicate the opposite. **Bottom:** (a) eight ResNet-50 configurations sharing an initialization, including baseline and label-smoothed (LS) runs with opposite feature-norm orderings. (b) Seven same-architecture V1/V2 checkpoint pairs, each with lower feature-norm AUROC in V2. Dashed lines mark chance.

dependence therefore concerns more than the relative performance of scoring methods: a statistic interpreted as ID evidence for one model can favor OOD inputs for another.

Opposite orderings also occur when one ResNet-50 training or architectural factor changes: feature-norm AUROC is 70.5 for the baseline and 35.8 with label smoothing, and Mixup weakens it to 57.6 (Figure 2a). All seven public V1/V2 checkpoint pairs likewise exhibit lower feature-norm AUROC in V2, five of them below 40 (Figure 2b). Appendix A specifies the training configurations and public checkpoint recipes.

3.2 Feature References Depend on the Available ID Samples

A class-relative reference must itself be estimated. Figure 1c varies the number of labeled calibration examples per class across nine settings while keeping the networks fixed. Reducing the budget from full calibration to two samples per class lowers MDS++ mean AUROC from 84.9 to 76.7, and at one sample per class its pooled covariance is undefined, whereas classifier-only GEN, which requires no estimated reference, remains at 83.5. Thus, the usefulness of feature-based evidence depends not only on the representation but also on the available ID observations. Because the classifier already encodes class-specific information, we use it to supplement the available ID observations when estimating normalized class references. Section 4 specifies how these references are estimated and combined with the classifier's responses while retaining class correspondence.

4 *ONE*: Class-Aligned Scoring with Classifier-Informed References

A frozen classifier produces features $z \in \mathbb{R}^D$ and logits $\ell(z) = Wz + \beta \in \mathbb{R}^C$, where $\boldsymbol{\beta}$ denotes the classifier bias. Given access to these quantities and labeled ID calibration data $\mathcal{D} = \{(z_i, y_i)\}_{i=1}^N$, *ONE* estimates class centers, a shared distance metric, and two evidence scales without updating the network. Following Mahalanobis++ (Müller & Hein, 2025), we use normalized features $x = z/\|z\|_2$ for feature fit while retaining the original logits for classifier support (Figure 4, Appendix C). Fitting occurs once per model and calibration budget; larger detection scores indicate ID.

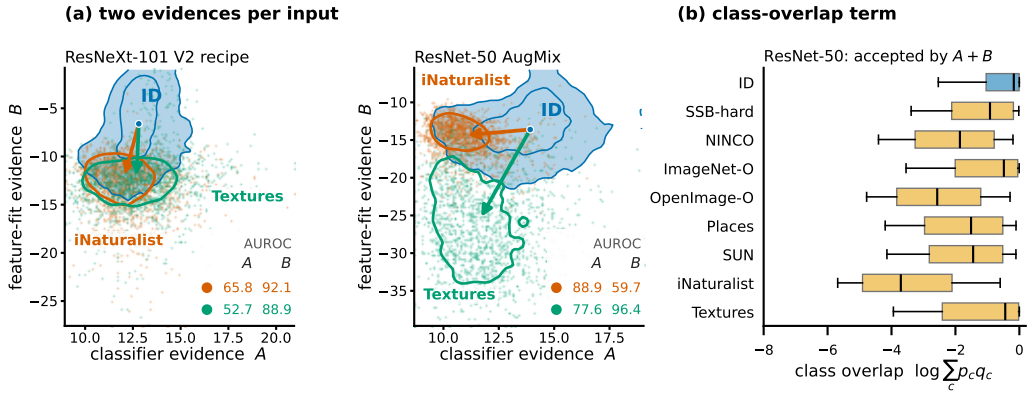


Figure 3: **Complementary branch evidence and class overlap beyond scalar fusion.** (a) Classifier evidence (A) versus feature-fit evidence (B), with full ID calibration. Blue: ID density (50% and 90%); points and contour: an OOD set and its 50% region; arrows: from the ID median to the OOD sets' medians. Insets report standalone branch AUROCs. (b) Soft class overlap among ID and OOD inputs accepted by $A + B$ at a threshold retaining 95% of ID inputs. Boxes span 25–75%; whiskers 10–90%. This conditional diagnostic examines information remaining after scalar acceptance; Section 5.2 evaluates the complete scores. Appendix C.1 repeats Figure 3 on four checkpoints.

4.1 Combine Corresponding Classes Before Pooling

We retain classifier support and feature fit for each candidate class until the final aggregation. Let $\widetilde{\mu}_c$ denote the estimated center of class c, and let P_λ be the precision matrix obtained from the regularized shared covariance $\widehat{\Sigma}_\lambda$. For the normalized feature x, define

$$d_c^2(x) = (x - \widetilde{\mu}_c)^\top P_\lambda (x - \widetilde{\mu}_c), \qquad a_c(z) = \kappa_l \ell_c(z), \qquad b_c(x) = -\kappa_m d_c^2(x), \tag{1}$$

where $\kappa_l, \kappa_m > 0$ are scales fitted separately using ID labels. The classifier term a_c measures support for class c, while the feature term b_c increases as the input approaches that class's reference under the fitted metric. The reference and scale-fitting procedures are specified in Sections 4.2–4.4.

We combine these measurements within each class before pooling:

$$\boxed{S(z) = \mathrm{LSE}_c\big(a_c(z) + b_c(x)\big),} \qquad \mathrm{LSE}_c(u_c) = \log \sum_{c=1}^{C} e^{u_c}. \tag{2}$$

The shared class index pairs each classifier response with the feature fit to the corresponding class reference. Log-sum-exp aggregation retains alternative classes rather than restricting feature fit to the classifier's top prediction. The quadratic feature term has Gaussian log-likelihood form, up to an input- and class-independent normalizer when the covariance is nonsingular. Both terms nevertheless derive from the same representation: the construction does not assume independent observations or identify S with a calibrated probability of being ID.

Why retain both sources? Pooling the branches separately gives their overall evidence: $A = \mathrm{LSE}_c(a_c)$ and $B = \mathrm{LSE}_c(b_c)$. Figure 3a shows that their relative usefulness depends on both the model and the OOD set. Feature fit is stronger for both iNaturalist and Textures on ResNeXt-101 V2. On ResNet-50 AugMix, however, classifier evidence A is stronger for iNaturalist (88.9 versus 59.7 AUROC), while feature fit B is stronger for Textures (96.4 versus 77.6). Thus, even for a fixed model, neither branch is uniformly preferable across the illustrated OOD sets. We therefore retain both measurements and examine what information their combination should preserve.

What scalar fusion discards: class correspondence. Two branches can produce strong overall evidence while favoring different classes. Adding only their scalar scores, $A + B$, cannot recover that distinction. Let $p = \mathrm{softmax}(a)$ and $q = \mathrm{softmax}(b)$ denote the branches' normalized class responses. Eq. (2) decomposes exactly as

$$S = \underbrace{A}_{\text{classifier evidence}} + \underbrace{B}_{\text{feature-fit evidence}} + \underbrace{\log \sum_c p_c q_c}_{\text{class overlap}}. \tag{3}$$

Scalar fusion $A + B$ retains each branch's overall evidence but discards their correspondence across classes. Eq. (3) restores this information through soft class overlap. Equivalently, e^{A+B} includes all classifier–feature class pairs, whereas e^S includes only matching indices. The overlap depends on concentration as well as correspondence: identical uniform responses have overlap $1/C$. Figure 3b examines information remaining among scalar-accepted inputs; Section 5.2 evaluates the complete scores with both fitted branches held fixed.

Raw logits retain the classifier level. Our classifier branch uses the full logit vector rather than reducing it to scalar MaxLogit. Since $a_c = \log p_c + A$, the score satisfies

$$S(z) = A(z) + \log \sum_c p_c(z)\, e^{b_c(x)}. \tag{4}$$

Probability weighting preserves class correspondence but removes the input-dependent classifier level A; scalar fusion instead removes the overlap term. We retain both quantities and evaluate their contributions through the matched controls in Section 5.2.

4.2 Estimate Class Centers from Classifier Priors and ID Samples

We estimate each class center in three steps. First, we compute empirical statistics from the available ID features. Second, we place classifier-derived class directions in the same feature space to obtain prior centers. Third, we blend the empirical and prior centers according to class counts and an estimated shared prior strength.

Empirical statistics. Let $\hat{y}_i = \arg\max_c \ell_c(z_i)$, $\mathcal{I}_c = \{i : y_i = \hat{y}_i = c\}$, $n_c = |\mathcal{I}_c|$, and $C_+ = \{c : n_c > 0\}$. For observed classes, $\bar{x}_c = n_c^{-1} \sum_{i \in \mathcal{I}_c} x_i$, while $\mu_G = N^{-1} \sum_i x_i$ uses all calibration examples. Thus, n_c counts correctly classified examples rather than the nominal budget. The empirical covariance is computed around these class means before shrinkage. Class-specific statistics describe correctly classified ID examples; the nominal budget is measured before this filtering.

Classifier-derived priors. Centered classifier rows provide class-specific directions:

$$\bar{w} = \frac{1}{C} \sum_{c=1}^{C} w_c, \qquad \widehat{w}_c = \mathrm{unit}(w_c - \bar{w}), \qquad m_c^W = \mu_G + \alpha \widehat{w}_c. \tag{5}$$

Centering removes the classifier-row component shared across classes before normalization. We place the resulting directions in feature space using μ_G as an offset and a shared coefficient α, fitted by minimizing $\sum_{c \in C_+} \|\bar{x}_c - \mu_G - \alpha \widehat{w}_c\|_2^2$. The resulting m_c^W is an imperfect prior, which we combine with the empirical mean rather than use unconditionally.

Balancing prior and sample information. For an observed class, the final center is

$$\widetilde{\mu}_c = \mathrm{unit}\left[\rho_c \bar{x}_c + (1 - \rho_c) m_c^W\right], \qquad \rho_c = \frac{n_c}{n_c + n_0}. \tag{6}$$

For an unobserved class, we use $\widetilde{\mu}_c = \mathrm{unit}(m_c^W)$. Eq. (6) is the unit-vector minimizer of a weighted squared-distance objective balancing proximity to the empirical mean and the classifier-derived prior. Thus, n_0 acts as a shared pseudo-count. For fixed n_0, classes with fewer observations rely more on the prior, while well-sampled classes rely more on their empirical means. We blend the mean vectors before normalizing; only the final center is constrained to unit norm.

We set the shared pseudo-count n_0 by comparing two quantities: the variability of ID features within classes and the remaining mismatch between empirical and classifier-derived centers.

$$\widehat{\tau}^{\,2} = \frac{1}{|C_+|} \sum_{c \in C_+} \left[\|\bar{x}_c - m_c^W\|_2^2 - \frac{\mathrm{tr}(\widehat{\Sigma})}{n_c} \right], \qquad n_0 = \max\left\{ \frac{\mathrm{tr}(\widehat{\Sigma})}{\max(\widehat{\tau}^{\,2}, \epsilon)}, \epsilon \right\}, \quad \epsilon = 10^{-6}. \tag{7}$$

The ratio balances within-class variation against noise-adjusted classifier–sample mismatch. Greater remaining mismatch reduces prior reliance. Because the prior offset and coefficient are fitted from the same observations, the correction is a plug-in approximation rather than an exactly unbiased estimate.

4.3 Regularize the Distance Metric According to Sample Availability

Center estimation determines where each class is represented; covariance estimation determines how deviations from that location are weighted. Using the correctly classified calibration examples defined above, let $\nu = \sum_{c \in C_+} n_c - |C_+|$. When $\nu > 0$, the empirical pooled covariance is

$$\widehat{\Sigma} = \frac{1}{\nu} \sum_{c \in C_+} \sum_{i \in \mathcal{I}_c} (x_i - \bar{x}_c)(x_i - \bar{x}_c)^\top. \tag{8}$$

When no within-class degrees of freedom remain, we use the total scatter of the correctly classified examples instead (Appendix B). This covariance, or its fallback, is computed before center shrinkage and supplies the trace used in Equation (7). The effective count used for regularization is $n_{\text{eff}} = \max(\nu, 1)$.

We shrink the empirical covariance toward an isotropic reference:

$$\widehat{\Sigma}_\lambda = (1 - \lambda)\widehat{\Sigma} + \lambda t I, \qquad t = \frac{\operatorname{tr}(\widehat{\Sigma})}{D}, \qquad \lambda = \frac{D}{D + n_{\text{eff}}}. \tag{9}$$

We use the symmetric pseudoinverse $P_\lambda = \widehat{\Sigma}_\lambda^\dagger$ as the precision matrix. Smaller effective counts increase shrinkage toward an isotropic metric. For unit features and centers, that endpoint reduces to cosine-based fit up to a common scale and offset. Center shrinkage determines class location; covariance shrinkage regulates how deviations are weighted.

4.4 Fit Evidence Scales Using ID Labels

After estimating the reference, we fit one positive scale for each branch. Let $E_{ic}^{(l)} = \ell_c(z_i)$ and $E_{ic}^{(m)} = -d_c^2(x_i)$. For either branch $j \in \{l, m\}$, we minimize ID classification cross-entropy:

$$\min_{\kappa_j > 0} \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \left[\operatorname{LSE}_c\left(\kappa_j E_{ic}^{(j)}\right) - \kappa_j E_{iy_i}^{(j)} \right], \tag{10}$$

where $\mathcal{T}$ is the calibration set. Both scales use all available ID calibration examples, including incorrectly classified ones, and are fitted separately with the fixed numerical procedure in Appendix B. We then sum the scaled potentials without an additional mixture coefficient. Class-label fitting does not determine calibrated ID probabilities: an input-dependent shift shared across classes leaves class probabilities unchanged while shifting the detection score. Separate fitting is a fixed design choice, not a resolution of this ambiguity.

5 Experiments

Evaluation Settings. We evaluate benchmark OOD performance and robustness across pretrained models. The benchmark comprises *nine settings* on public checkpoints spanning CNN, transformer, and ConvNeXt families: ResNet-18 on CIFAR-10 and CIFAR-100; WideResNet-40-2 (WRN) on CIFAR-100; and ResNet-50, ViT-B, Swin-B, ConvNeXt-B, DeiT-B, and EfficientNet-B0 on ImageNet. Each setting averages *eight OOD datasets*, including both near- and far-OOD shifts, following the dataset organization of OpenOOD v1.5 (Yang et al., 2022; Zhang et al., 2024); checkpoint sources are listed in Appendix A. The 30-checkpoint ImageNet panel tests robustness across architectures and training recipes within one ID dataset; eight controlled ResNet-50 models vary one training or architectural factor at a time under the same protocol. All comparisons use frozen models and cached features and logits; head access is specified in Appendix A, estimator details in Appendix B, and per-OOD results in Appendix D.

Metrics and Baselines. We report AUROC in this section, and FPR95 in Appendix C. To measure the cost of using one score across models, define

$$R_{m,j} = \max_{b \in \mathcal{B} \cup \{ONE\}} \text{AUROC}_{b,j} - \text{AUROC}_{m,j}, \tag{11}$$

where $\mathcal{B}$ is the set of evaluated baselines, m indexes methods, and j indexes settings or checkpoints. Regret is zero for the best score on each model; mean and maximum regret summarize $R_{m,j}$ across the evaluated models. For each setting or checkpoint, the reported AUROC is the average over its eight OOD datasets, and regret is computed from these averages. The baselines cover output scores, activation modification, class distances, neighbors, and hybrids.

Table 2: **9 benchmark settings and 30 public ImageNet models, full ID calibration.** Left: AUROC averaged over eight OOD datasets per setting, with the regret over the nine settings. Regret is the AUROC gap to the best listed score on each setting or model. (CNX: ConvNeXt-B; EffB0: EfficientNet-B0.) Right: mean AUROC per training-recipe group of Figure 1 (CE: torchvision V1 CNNs with one-hot cross-entropy; Other: their V2 versions and four other ResNet-50 recipes; BCE: four ResNet-50 variants with binary cross-entropy; Trf.: transformers and ConvNeXt) and over all 30 models, with regret. †Needs additional layer features; ‡trains energy networks.

	9 benchmark settings												30 ImageNet models						
	AUROC										Regret		AUROC					Regret	
Score	R18 (C10)	R18 (C100)	WRN	R50	ViT	Swin	CNX	DeiT	EffB0	Mean	Mean	Max	CE	Other	BCE	Trf.	All	Mean	Max
MSP	90.3	78.2	74.3	78.2	80.0	85.0	79.8	79.2	79.6	80.5	7.7	13.6	79.3	78.1	79.5	80.4	79.2	7.7	15.0
GEN	91.2	80.3	77.4	81.9	83.4	88.3	84.1	82.6	82.1	83.5	4.8	9.9	83.3	81.2	**82.4**	83.7	82.5	4.3	9.9
TRIM	90.2	80.2	77.8	80.3	82.1	87.0	85.5	84.1	81.1	83.1	5.1	11.5	81.8	81.2	**82.4**	83.8	82.2	4.7	11.5
MaxLogit	90.9	80.2	77.2	81.6	80.5	85.3	76.3	75.4	77.7	80.6	7.7	11.5	82.6	74.8	77.6	78.8	78.1	8.8	23.2
Energy	91.0	80.3	77.1	81.5	79.4	84.1	73.2	70.3	74.2	79.0	9.2	15.2	82.5	66.2	74.3	76.0	73.7	13.2	47.4
LogitGap	89.8	80.1	77.8	81.8	82.6	87.0	81.8	80.5	80.9	82.5	5.7	10.0	83.2	79.0	80.7	82.3	81.1	5.8	13.8
GradNorm	61.3	70.3	50.3	81.7	52.4	39.7	22.6	33.4	56.0	52.0	36.3	65.3	74.3	43.4	49.1	48.9	52.8	34.0	66.9
ReAct	91.0	80.3	75.9	85.8	82.1	89.5	83.4	79.2	80.2	83.0	5.2	9.0	82.9	38.8	53.9	80.6	62.2	24.6	66.8
DICE	85.4	79.8	70.4	83.1	38.7	71.0	55.7	59.8	44.6	65.4	22.8	47.1	80.7	43.4	64.6	58.8	59.0	27.8	63.2
ASH-S	23.5	64.6	51.9	89.0	31.1	45.5	20.8	20.3	50.8	44.2	44.1	71.6	86.7	62.7	79.7	35.4	63.3	23.6	69.6
SCALE	34.0	61.0	51.7	89.6	57.7	85.2	48.5	66.8	53.8	60.9	27.3	61.1	87.0	63.6	77.3	66.1	71.6	15.3	60.1
KNN	92.3	81.5	80.8	81.3	80.4	87.6	85.8	82.1	79.4	83.5	4.8	10.5	80.1	78.6	78.4	83.1	80.2	6.7	18.5
SHE	86.1	78.4	60.9	83.0	80.7	82.2	76.2	83.0	70.2	77.8	10.4	24.0	77.3	67.1	71.4	76.6	72.6	14.3	33.0
MDS	86.3	66.4	77.0	66.6	77.6	88.2	86.5	82.1	77.6	78.7	9.5	25.2	62.4	81.2	78.3	84.2	77.3	9.6	51.7
MDS++	91.8	81.1	81.0	84.3	84.1	89.9	86.9	84.9	80.0	84.9	3.3	7.5	81.1	82.2	81.6	85.9	82.9	4.0	16.0
RMDS++	92.1	81.1	80.8	82.5	84.8	88.9	86.7	**85.5**	82.1	84.9	3.3	9.3	82.5	81.6	82.3	**86.1**	83.1	3.8	10.9
fDBD	90.4	80.8	80.9	84.5	79.0	88.4	84.4	80.2	81.2	83.3	4.9	7.2	84.6	80.0	81.3	83.0	82.1	4.8	7.9
NCI	90.7	81.4	80.2	84.7	82.0	89.6	84.4	80.9	82.6	84.1	4.2	7.1	83.5	82.2	76.8	83.4	82.1	4.7	11.7
ViM	92.4	80.2	78.4	82.6	77.7	90.3	83.8	81.7	82.2	83.3	5.0	9.2	81.0	79.8	73.4	83.2	80.1	6.7	17.8
NNGuide	90.5	81.5	79.0	84.7	80.1	89.7	85.6	82.5	82.4	84.0	4.2	7.1	84.5	80.1	74.3	83.0	81.1	5.8	16.5
ActSub	86.7	84.1	77.8	**91.8**	81.5	87.6	85.8	81.2	76.7	83.7	4.5	8.4	**88.5**	80.5	71.2	75.8	79.9	7.0	62.0
HEAT‡	92.7	84.0	80.3	87.3	**85.8**	**90.8**	85.4	83.2	83.2	85.9	2.4	4.6	85.5	82.0	80.1	85.3	83.4	3.4	10.1
COMBOOD†	**95.1**	**87.6**	**84.9**	82.2	83.6	89.9	**87.8**	84.4	74.9	85.6	2.6	10.0	82.8	81.9	81.7	85.5	83.0	3.8	10.6
MME	35.7	64.1	55.6	89.7	79.0	90.4	83.1	82.0	59.8	71.0	17.2	59.4	87.1	74.1	77.8	83.5	80.1	6.7	25.7
ONE	92.5	83.8	80.6	87.3	85.4	**90.8**	86.9	85.0	**84.9**	**86.3**	**1.9**	**4.5**	86.4	**84.9**	81.6	86.0	**85.1**	**1.7**	**6.8**

5.1 Standard OOD Benchmark Performance and Cross-Model Robustness

Table 2 shows that with full ID calibration, *ONE* achieves **86.3 mean AUROC (the highest)** across nine settings, compared with 85.9 for HEAT and 85.6 for COMBOOD. COMBOOD performs better on the three CIFAR settings, whereas *ONE* performs better on five of the six ImageNet settings. On the 30-checkpoint ImageNet panel, *ONE* reaches **85.1 mean AUROC (the highest)** versus 83.4 for HEAT, and **a maximum checkpoint-level gap of 6.8 points (the smallest)** versus 7.9 for fDBD. Group-level results reveal different strengths rather than a common ranking. *ONE* is below ActSub on one-hot cross-entropy CNNs and below GEN and TRIM on BCE-trained heads, but has the highest mean AUROC over the complete panel and the smallest mean and maximum checkpoint-level regret. Further checkpoint-level results are described in Appendix C. Table 3 separates near- and far-OOD sets: *ONE* leads both ImageNet splits and stays within 0.4 points of the best CIFAR near-OOD score, whereas the CIFAR leaders, RMDS++ and COMBOOD, score below *ONE* on both ImageNet splits.

Table 3: **Near- and far-OOD AUROC** for the four strongest baselines in Table 2. Boldface/underlining mark first/second among all 25 methods.

	CIFAR		ImageNet	
Score	Near	Far	Near	Far
MDS++	78.5	86.7	81.8	87.0
RMDS++	**82.3**	85.5	81.9	87.0
HEAT	82.1	86.9	82.8	87.8
COMBOOD	79.8	**92.3**	81.0	85.5
ONE	81.9	86.9	**83.2**	**88.8**

Controlled Training Interventions. On the eight controlled ResNet-50 models, *ONE* leads all 24 baselines in mean AUROC (85.6), mean regret (2.4), and maximum regret (4.6), compared with 84.6, 3.5, and 6.3 for HEAT. The per-model results for *ONE* are reported in Table 9 in Appendix C. Its largest gaps are to ActSub on the plain cross-entropy models, and it exceeds every baseline under label smoothing and mixup.

Table 4: **Class correspondence contributes beyond the two scalar scores.** The fitted branches are identical across rows. Full and 2/class report nine-setting mean AUROC. All rows reuse the fitted potentials $a_c = \kappa_l \ell_c$ and $b_c = -\kappa_m d_c^2$. Probability weighting uses $p = \mathrm{softmax}(a)$, so the control is exactly $S - A$.

Combination	Full	2/class	IN mean AUROC	IN max. regret
Classifier only: A	80.37	80.42	77.44	26.09
Feature only: B	84.58	81.42	81.59	15.27
Scalar fusion: $A + B$	<u>86.15</u>	<u>84.51</u>	84.13	<u>8.86</u>
Probability weighting: $\mathrm{LSE}_c(\log p_c + b_c)$	85.54	83.95	<u>84.23</u>	10.56
Class-wise: $\mathrm{LSE}_c(a_c + b_c)$ (***ONE***)	**86.34**	**84.87**	**85.11**	**6.80**

5.2 Matched Combination Controls

Table 4 holds both fitted branches fixed and changes only their combination. On the 30-model panel, preserving class correspondence increases mean AUROC from 84.13 for scalar fusion to 85.11 and improves on 25 checkpoints. Retaining the classifier level also contributes: probability weighting, which removes A but preserves correspondence, reaches 84.23. The matched controls therefore distinguish the value of class overlap from the value of the classifier's overall evidence. The benefit of preserving class correspondence is most pronounced on the heterogeneous checkpoint panel, matching our objective of applying one procedure across pretrained models.

5.3 Calibration Efficiency

Figure 1c shows performance across balanced per-class calibration budgets (nine-setting means over four draws); Table 10 in Appendix C lists every sample-dependent score. With one sample per class, *ONE* reaches 84.2, above GEN (83.5). At this budget, the class-wise blend weights ρ_c are effectively zero, so the centers rely almost entirely on their classifier-derived priors. With two samples per class, *ONE* reaches 84.8, above GEN and well above MDS++ (76.7); from three samples per class. it exceeds GEN in every setting, and with 20 it reaches 86.0, within 0.3 points of full calibration.

Table 5 varies one part of the reference at a time. With our covariance, classifier-derived priors outperform sample centers at one sample per class, whereas sample centers are stronger at every larger budget. At full calibration, classifier-derived centers alone reach only 78.7, whereas sample centers and the adaptive blend reach 86.3. At two samples per class, *ONE*'s adaptive blend outperforms both fixed-center alternatives, reaching 84.8 versus 84.2 and 84.0, and matches the sample-center result at twenty samples per class and full calibration. With our centers, the sample covariance falls to 77.2 at two samples per class, whereas our shrinkage matches the isotropic metric when samples are scarce and exceeds it at full calibration (86.3 versus 85.6). These budget-level results support relying on classifier-derived centers and an isotropic metric when observations are scarce, and on the empirical centers and covariance as class observations accumulate.

Table 5: **Center and covariance estimation across calibration budgets.** Upper rows vary class centers under our covariance; lower rows vary covariance with our centers. All rows share the calibration subsets; mean AUROC over nine settings and four draws.

Estimator \ per class	1	2	5	20	Full
Centers (our covariance)					
Sample means	83.1	84.2	85.3	**86.0**	**86.3**
Classifier priors	**84.2**	84.0	83.8	83.4	78.7
Covariance (our centers)					
Sample covariance	80.4	77.2	80.3	82.9	85.8
Isotropic	**84.2**	**84.8**	85.2	85.5	85.6
ONE	**84.2**	**84.8**	**85.4**	**86.0**	**86.3**
Blend weight $\bar{\rho}$	0.00	0.46	0.66	0.88	1.00
Shrinkage weight $\bar{\lambda}$	1.00	0.65	0.36	0.15	0.00

6 Conclusion

We propose *ONE*, a post-hoc procedure that combines classifier and feature evidence within corresponding classes and estimates the feature reference using classifier-derived priors. The resulting score improves aggregate detection performance and reduces the worst observed checkpoint-level gap across the evaluated models, while remaining effective with limited calibration data. Gains are not uniform, motivating further study of both the reference assumptions and evidence combination.

AI USE STATEMENT

In this work, we used generative AI tools to port the released code of the baselines to our evaluation setting, to give feedback on design choices of the method, and to polish figures, tables, and writing. The authors made all design decisions and validated them experimentally. We did not use them for any other task that requires disclosure under the ICLR 2027 AI policy. We have reviewed all AI-assisted work: we checked the reported numbers against the result files that produced them, checked the baseline implementations against the released code and papers, and verified the cited references. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 4 and Appendix B specify the complete procedure, including the center and covariance estimators, the numerical conventions, and the temperature fitting. Section 5 and Appendix A list the checkpoint sources, the OOD datasets, the calibration protocols, the baseline configurations, and the training recipe of the eight controlled ResNet-50 models; all other checkpoints are publicly released. Budget results average four random draws of the per-class calibration samples. Appendix C reports checkpoint-level AUROC, FPR95, and near- and far-OOD results. We will release the code upon publication.

REFERENCES

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing ImageNet out-of-distribution detection evaluation. In *ICML*, 2023.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, pp. 8759–8773, 2022.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

Byung Chun Kim, Byungro Kim, and Yoonsuk Hyun. Investigation of out-of-distribution detection across various models and training methodologies. *Neural Networks*, 175:106288, 2024.

Marc Lafon, Elias Ramzi, Clément Rambour, and Nicolas Thome. Hybrid energy based model in the feature space for out-of-distribution detection. 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. 18250–18268. PMLR, 23–29 Jul 2023.

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.

Jiachen Liang, Ruibing Hou, Minyang Hu, Hong Chang, Shiguang Shan, and Xilin Chen. Revisiting logit distributions for reliable out-of-distribution detection. *Advances in Neural Information Processing Systems*, 38:20848–20880, 2025.

Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31728–31746. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/liu24ax.html.

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

Xixi Liu, Yaroslava Lochman, and Christopher Zach. GEN: Pushing the limits of softmax-based out-of-distribution detection. In *CVPR*, 2023.

Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7831–7840, 2022.

Maximilian Müller and Matthias Hein. Mahalanobis++: Improving OOD detection via feature normalization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45151–45184. PMLR, 2025. URL https://proceedings.mlr.press/v267/muller25a.html.

Nguyen Ngoc-Hieu, Nguyen Hung-Quang, The-Anh Ta, Thanh Nguyen-Tang, Khoa D Doan, and Hoang Thanh-Tung. A cosine similarity-based method for out-of-distribution detection. *arXiv preprint arXiv:2306.14920*, 2023.

Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1686–1695, 2023.

Magesh Rajasekaran, Md Saiful Islam Sajol, Frej Berglind, Supratik Mukhopadhyay, and Kamalika Das. COMBOOD: A semiparametric approach for detecting out-of-distribution data for image classification. In *Proceedings of the 2024 SIAM International Conference on Data Mining*, pp. 643–651, 2024.

Lucas Rakotoarivony. Multi-method ensemble for out-of-distribution detection. *BMVC*, 2025.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, pp. 691–708, 2022.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in neural information processing systems*, 34:144–157, 2021.

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.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *ICLR*, 2022.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4921–4930, 2022.

Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In *The Twelfth International Conference on Learning Representations*, 2024.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *NeurIPS*, 2022.

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*, 2:1–32, 2024.

Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Xiaoguang Liu, Shi Han, and Dongmei Zhang. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *The Eleventh International Conference on Learning Representations*, 2023.

Barış Zöngür, Robin Hesse, and Stefan Roth. Activation subspaces for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3509–3519, 2025.

A Evaluation Protocol

Datasets. For ImageNet, the near-OOD sets are SSB-hard (Vaze et al., 2022), NINCO (Bitterwolf et al., 2023), and ImageNet-O (Hendrycks et al., 2021); the far-OOD sets are Places, SUN, iNaturalist, Textures, and OpenImage-O. For CIFAR, the near-OOD sets are the other CIFAR dataset and Tiny-ImageNet, and the far-OOD sets are SVHN, Places365, LSUN, iSUN, Textures, and MNIST. Each reported setting-level AUROC or FPR95 is the unweighted mean over its eight OOD sets; aggregate means weight checkpoints or benchmark settings equally. FPR95 is the fraction of OOD examples accepted at a threshold retaining 95% of ID examples.

Checkpoint Panel. The nine benchmark settings use the OpenOOD v1.5 ResNet-18 checkpoints for CIFAR-10 and CIFAR-100, the RobustBench AugMix WRN-40-2 for CIFAR-100, and for ImageNet the torchvision ResNet-50 (V1), the timm ViT-B/16 (AugReg, ImageNet-21k pretraining with ImageNet-1k fine-tuning), ConvNeXt-B, DeiT-B/16, and EfficientNet-B0, and the Hugging Face Swin-B; the ResNet-50, ViT-B/16, ConvNeXt-B, and DeiT-B/16 checkpoints also belong to the panel below. The 30 ImageNet checkpoints consist of ten ResNet-50 configurations, twelve torchvision V1/V2 models from six other architecture families, and eight transformer/ConvNeXt models; Table 6 lists them with their sources. The controlled ImageNet comparison trains ResNet-50 from one random initialization for 30 epochs with SGD (Nesterov momentum 0.9, learning rate 0.1, batch size 256, one warmup epoch then cosine decay, weight decay 10^{-4}, random-resized crops and horizontal flips) and changes one factor per model: GELU activations, an affine LayerNorm before the classifier, label smoothing 0.1, one-hot BCE, input mixup with $\mathrm{Beta}(1, 1)$, doubled weight decay, and a 90-epoch schedule. The nine-setting benchmark measures OOD performance on public checkpoints, and the public ImageNet panel measures robustness across pretrained models. Public ImageNet recipe comparisons change several factors together, whereas each controlled configuration changes one specified factor. Table 9 gives the checkpoint-level values used in the main analysis.

Table 6: The 30 public ImageNet checkpoints of the panel. Figure labels are the axis labels of Figures 1 and 2; table labels are the row names of Table 9.

Figure label	Table label	Source
CE: torchvision V1 CNNs with one-hot cross-entropy		
R50 V1	R50 V1 plain CE	torchvision `resnet50`, `IMAGENET1K_V1`
R152	tv R152 V1	torchvision `resnet152`, `IMAGENET1K_V1`
RX101	tv RX101 V1	torchvision `resnext101_32x8d`, `IMAGENET1K_V1`
WR101	tv WR101 V1	torchvision `wide_resnet101_2`, `IMAGENET1K_V1`
RegY	tv RegY16 V1	torchvision `regnet_y_16gf`, `IMAGENET1K_V1`
MNv2	tv MNv2 V1	torchvision `mobilenet_v2`, `IMAGENET1K_V1`
MNv3	tv MNv3-L V1	torchvision `mobilenet_v3_large`, `IMAGENET1K_V1`
Other: torchvision V2 CNNs and four other ResNet-50 recipes		
bt	R50 bt	timm `resnet50.bt_in1k`
LS	R50 gluon LS	timm `resnet50.gluon_in1k`
am	R50 am AugMix	timm `resnet50.am_in1k`
ram	R50 ram AugMix+RA	timm `resnet50.ram_in1k`
R50 V2	R50 V2 LS+mixup	torchvision `resnet50`, `IMAGENET1K_V2`
R152	tv R152 V2	torchvision `resnet152`, `IMAGENET1K_V2`
RX101	tv RX101 V2	torchvision `resnext101_32x8d`, `IMAGENET1K_V2`
WR101	tv WR101 V2	torchvision `wide_resnet101_2`, `IMAGENET1K_V2`
RegY	tv RegY16 V2	torchvision `regnet_y_16gf`, `IMAGENET1K_V2`
MNv2	tv MNv2 V2	torchvision `mobilenet_v2`, `IMAGENET1K_V2`
MNv3	tv MNv3-L V2	torchvision `mobilenet_v3_large`, `IMAGENET1K_V2`
BCE: ResNet-50 with per-class BCE		
a3	R50 a3 BCE short	timm `resnet50.a3_in1k`
a1	R50 a1 BCE+mixup	timm `resnet50.a1_in1k`
a1h	R50 a1h BCE	timm `resnet50.a1h_in1k`
GN	R50 GN a1h	timm `resnet50_gn.a1h_in1k`
Trf.: transformers and ConvNeXt		
ViT	ViT augreg IN-1k	timm `vit_base_patch16_224.augreg_in1k`
SAM	ViT SAM IN-1k	timm `vit_base_patch16_224.sam_in1k`

Figure label	Table label	Source
ViT21k	ViT augreg2 21k	timm `vit_base_patch16_224.augreg2_in21k_ft_in1k`
DeiT	DeiT-B	timm `deit_base_patch16_224.fb_in1k`
DeiT3	DeiT3-B	timm `deit3_base_patch16_224.fb_in1k`
Swin	Swin timm in1k	timm `swin_base_patch4_window7_224.ms_in1k`
CNX	ConvNeXt IN-1k	timm `convnext_base.fb_in1k`
CNX22k	ConvNeXt 22k	timm `convnext_base.fb_in22k_ft_in1k`

Calibration Budgets. Full calibration uses the ID training set. Balanced budgets take k examples per class; budget results average four random draws.

Classifier Heads. All logit-based quantities use the cached logits of each model. Methods that require the classifier weights, including *ONE* for its center prior, use the stored head of the model. For the twelve torchvision panel models, the head is recovered by least squares from the full set of cached features and logits; the recovered weights match the stored weights to a relative error below 10^{-5}.

Evidence-Direction Statistics. Feature norm is $\|z\|_2$, maximum logit is $\max_c \ell_c$, MSP is $\max_c \mathrm{softmax}(\ell)_c$, and class similarity is $\max_c x^\top \mathrm{unit}(\bar{x}_c)$. For the concentration diagnostic, let $T(z)$ index the largest $\max(1, \lfloor .1D \rfloor)$ coordinates of the raw feature vector, and define

$$r(z) = \min\left(30, \frac{\sum_{d=1}^{D} z_d}{\max(\sum_{d \in T(z)} z_d, 10^{-6})}\right). \tag{12}$$

We report the AUROC of $-r$, equivalently 100 minus the AUROC of r. For nonnegative features this reflects concentration in the top channels; for signed features it is a top-channel statistic rather than a literal count of nonzero activations.

Score Implementations. MDS (raw) and MDS++ use pooled class covariance, with the latter operating on unit features; RMDS++ additionally subtracts a global-reference distance. SHE takes the raw feature dot product with the mean of the classifier's predicted class. KNN uses the 50th neighbor from a bank of at most 200,000 normalized ID features. NNGuide uses ten neighbors from a confidence-weighted bank. ReAct uses the 90th percentile of ID activations. ASH-S keeps the 10% largest activations of each feature vector, SCALE rescales the whole vector by the ratio of its total to its top-10% activation mass, and DICE keeps the 30% of weights per output unit with the largest mean contribution on ImageNet and 10% on CIFAR, the sparsity levels of its paper; on the panel, DICE's activation mean uses the 12,812-image calibration subset. COMBOOD combines a kNN score on penultimate features with a Mahalanobis score on the per-layer minimum and maximum activations of every activation layer (its extreme mode); we extracted these multi-layer statistics for all checkpoints in Table 2. HEAT trains a residual network on a class-conditional Gaussian-mixture energy (HEAT-GMM) with the released settings of each benchmark, using the ImageNet settings for every ImageNet model, and combines it with the Energy score by HEAT's rule: each energy is standardized on 256 calibration inputs, and the energies are summed. We keep the Energy prior of HEAT's logit term without its trained residual and omit its term on standard-deviation-pooled features, which HEAT defines only on convolutional feature maps. MME uses the released hyperparameters of its ImageNet or CIFAR configuration, with statistics computed on the full calibration set. All 24 baseline names and scores are shown in Table 2; the regret reference comprises these 24 baselines and *ONE*.

B Estimator and Score Details

Covariance and Numerical Conventions. The covariance is the within-class scatter of correctly classified calibration examples divided by its degrees of freedom, the number of examples minus the number of observed classes; this count is n_{eff} in Equation (9). When the degrees of freedom fall below one, that is, when each observed class has a single correctly classified example, the covariance is the scatter of the correctly classified examples about their mean, divided by $\max(n-1, 1)$ for n such examples; the same matrix supplies the trace in Equations (7) and (9). At one example per

class this trace exceeds the discrepancy term, so $\widehat{\tau}^2$ is floored, n_0 becomes large, and every center is its classifier-derived prior. *ONE* requires at least one correctly classified calibration example. Covariance accumulation and eigendecomposition use double precision. The precision matrix is implemented by a symmetric pseudoinverse, retaining eigenvalues above 10^{-15} times the largest eigenvalue.

Temperature Fitting. For either branch evidence E_{ic}, we optimize

$$\min_{\kappa>0} -\frac{1}{N}\sum_i \log \frac{\exp(\kappa E_{iy_i})}{\sum_c \exp(\kappa E_{ic})}. \tag{13}$$

We use the calibration examples, including incorrectly classified ones, and optimize log temperature for 300 Adam steps with learning rate 0.1. Before optimization, evidence is divided by its empirical standard deviation; the fitted temperature is converted back to the original scale. This avoids starting large-magnitude quadratic distances in a saturated softmax regime. *ONE* fits both scales separately on the available calibration set.

C Additional Results

Normalization Comparisons. Feature normalization follows Mahalanobis++ (Müller & Hein, 2025). Figure 4 compares raw and normalized Mahalanobis references on all 38 ImageNet checkpoints. MDS++ exceeds raw-feature Mahalanobis on 32 of the 38 checkpoints, with an average gain of 7.0 points.

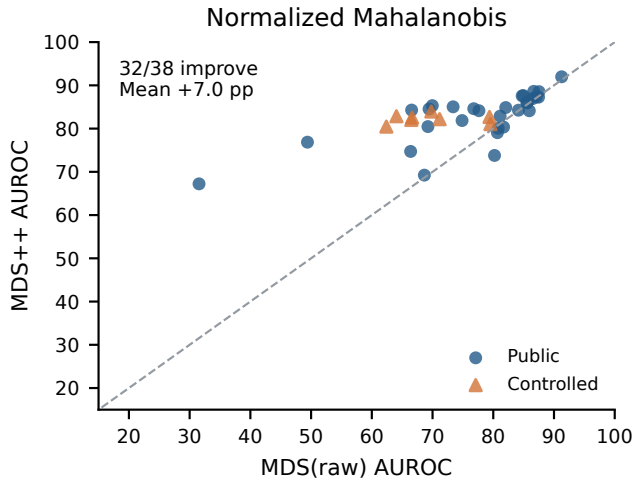


Figure 4: **Normalized class references improve feature-based detection across models.** Each point is a checkpoint; points above the diagonal favor the normalized-feature score. The comparison fits the Mahalanobis reference on raw versus normalized features.

Table 7: Full-budget FPR95 (↓), with the same settings and OOD averaging as Table 2.

Score	C10	C100	WRN	R50	ViT	Swin	CNX	DeiT	EffB0	Mean
MSP	53.6	79.2	83.6	74.5	67.0	51.9	66.0	68.9	71.3	68.5
GEN	39.8	76.5	81.3	70.4	61.3	42.6	55.8	63.6	67.4	62.1
TRIM	39.6	80.1	82.2	79.5	71.0	57.6	58.6	65.6	78.5	68.1
MaxLogit	40.6	77.6	82.2	68.7	60.6	45.0	64.5	68.0	72.6	64.4
Energy	39.7	76.5	81.8	67.9	59.8	45.1	70.1	74.0	80.9	66.2
LogitGap	59.0	78.6	82.3	70.3	61.0	44.5	59.2	66.0	68.9	65.6
GradNorm	79.5	77.3	93.6	55.7	90.2	87.0	98.7	97.6	89.4	85.4
ReAct	40.6	74.5	80.0	51.5	59.0	39.7	60.1	71.6	72.5	61.0
DICE	59.6	77.9	84.7	54.5	98.0	70.5	79.6	80.3	98.5	78.2
ASH-S	99.7	83.0	90.4	40.5	98.7	86.6	99.4	99.3	99.2	88.5
SCALE	99.8	88.5	93.4	38.4	91.8	45.9	98.3	82.0	99.0	81.9
KNN	44.1	72.1	70.8	61.8	78.6	53.7	58.6	76.9	69.8	65.1
SHE	49.7	77.6	87.1	55.8	62.6	48.0	55.4	60.3	72.8	63.2
MDS (raw)	64.0	89.7	75.8	82.6	82.4	52.5	55.9	75.7	77.8	72.9
MDS++	39.9	70.9	72.8	55.8	62.9	42.9	53.2	63.0	65.2	58.5
RMDS++	43.2	73.4	73.3	62.9	59.9	47.0	54.8	**60.0**	64.3	59.8
fDBD	48.4	76.8	74.8	63.0	69.0	48.2	61.8	73.7	75.2	65.7
NCI	47.8	76.5	77.5	61.8	59.7	45.2	61.0	72.5	71.8	63.7
ViM	39.0	76.2	76.5	67.8	76.7	45.3	54.1	73.5	66.8	64.0
NNGuide	43.1	74.6	77.6	52.8	74.1	44.8	59.0	74.5	64.4	62.8
ActSub	56.6	64.7	77.2	**34.4**	76.7	55.4	60.2	81.7	74.0	64.5
HEAT	35.1	66.9	76.6	48.2	**52.8**	**35.1**	52.7	63.3	62.4	54.8
COMBOOD	**24.5**	**47.8**	**50.1**	60.6	60.2	41.1	**49.2**	61.4	78.0	**52.6**
MME	99.2	82.4	89.8	38.3	73.4	42.5	55.9	72.1	86.2	71.1
ONE	36.6	67.8	75.8	48.4	57.1	38.2	52.7	61.0	**55.0**	54.7

Table 8: Near- and far-OOD AUROC at full calibration. CIFAR averages three benchmark models, and ImageNet averages six.

Score	CIFAR near	CIFAR far	IN near	IN far
MSP	80.9	81.0	76.0	82.9
GEN	81.8	83.4	79.2	86.5
TRIM	81.5	83.1	79.2	85.9
MaxLogit	81.6	83.1	75.3	81.9
Energy	81.5	83.3	73.4	79.3
LogitGap	82.0	82.7	78.0	85.2
GradNorm	58.5	61.3	48.5	47.1
ReAct	79.0	83.6	78.1	86.5
DICE	76.9	79.1	58.7	58.9
ASH-S	42.1	48.2	43.9	42.3
SCALE	43.7	50.6	63.9	68.8
KNN	82.0	85.8	79.2	84.9
SHE	73.3	75.7	77.1	80.5
MDS (raw)	70.1	78.7	76.3	81.8
MDS++	78.5	86.7	81.8	87.0
RMDS++	**82.3**	85.5	81.9	87.0
fDBD	82.0	84.7	78.0	86.0
NCI	81.8	84.9	78.7	87.2
ViM	78.3	85.5	79.2	85.4
NNGuide	81.5	84.4	80.3	86.5
ActSub	77.9	84.5	80.5	86.3
HEAT	82.1	86.9	82.8	87.8
COMBOOD	79.8	**92.3**	81.0	85.5
MME	46.1	53.7	76.6	83.1
ONE	81.9	86.9	**83.2**	**88.8**

Table 9: Checkpoint-level AUROC, with the checkpoint groups of Table 2. All models use full calibration data; ctrl rows are the eight controlled ResNet-50 models. Best baseline is the highest AUROC among the 24 baselines of Table 2.

Model	Norm	Concentration	GEN	MDS++	*ONE*	Best baseline
CE: torchvision V1 CNNs with one-hot cross-entropy						
R50 V1 plain CE	70.2	23.6	81.9	84.3	87.3	**91.8**
tv R152 V1	67.8	26.6	84.6	85.3	88.7	**92.3**
tv RX101 V1	60.0	41.4	84.8	84.5	88.7	**92.0**
tv WR101 V1	50.7	40.0	84.0	85.0	88.4	**90.1**
tv RegY16 V1	44.4	33.2	84.9	84.6	**89.6**	89.4
tv MNv2 V1	77.1	22.5	79.6	76.9	82.0	**88.8**
tv MNv3-L V1	76.5	33.7	82.9	67.2	80.4	**83.2**
Other: torchvision V2 CNNs and four other ResNet-50 recipes						
R50 bt	27.0	41.0	81.7	80.1	85.1	**86.3**
R50 gluon LS	26.8	43.1	81.0	79.1	84.8	**85.4**
R50 am AugMix	26.6	40.2	81.6	73.8	83.8	**85.7**
R50 ram AugMix+RA	30.9	54.5	81.1	80.4	**84.0**	**84.0**
R50 V2 LS+mixup	25.5	61.1	81.9	84.1	85.2	**85.9**
tv R152 V2	21.3	65.8	83.1	**87.4**	87.2	**87.4**
tv RX101 V2	18.9	69.7	80.5	87.3	85.4	**87.5**
tv WR101 V2	20.9	64.5	82.4	**88.5**	87.8	**88.5**
tv RegY16 V2	18.6	47.7	83.3	**88.6**	88.2	**88.6**
tv MNv2 V2	44.7	42.7	77.5	80.5	**82.2**	80.5
tv MNv3-L V2	41.1	64.4	79.4	74.7	79.9	**80.3**
BCE: ResNet-50 with per-class BCE						
R50 a3 BCE short	51.9	45.3	80.9	69.2	76.3	**81.0**
R50 a1 BCE+mixup	51.8	53.2	83.7	81.9	84.9	**87.2**
R50 a1h BCE	22.6	64.4	82.7	**87.6**	82.5	**87.6**
R50 GN a1h	22.6	69.2	82.3	**87.6**	82.7	**87.6**
Trf.: transformers and ConvNeXt						
ViT augreg IN-1k	47.9	54.8	82.3	84.3	83.8	**84.7**
ViT SAM IN-1k	54.9	41.8	79.5	**82.9**	82.2	**82.9**
ViT augreg2 21k	32.5	61.0	83.4	84.1	85.4	**85.8**
DeiT-B	23.3	47.1	82.6	84.9	85.0	**85.5**
DeiT3-B	18.5	31.7	85.4	86.1	86.5	**86.9**
Swin timm in1k	17.0	82.1	82.3	85.9	86.4	**86.5**
ConvNeXt IN-1k	17.7	61.9	84.1	86.9	86.9	**87.8**
ConvNeXt 22k	18.1	22.1	90.3	92.0	92.2	**92.5**
Controlled ResNet-50 models						
R50 ctrl base	70.5	28.4	82.3	82.9	86.5	**91.0**
R50 ctrl gelu	66.0	24.2	82.9	80.5	85.4	**90.0**
R50 ctrl ln	73.1	34.6	82.9	82.7	84.8	**85.3**
R50 ctrl ls	35.8	38.0	80.1	81.0	**85.2**	83.8
R50 ctrl bce	76.1	25.0	81.3	82.0	83.8	**85.7**
R50 ctrl mixup	57.6	32.4	78.5	82.2	**85.7**	82.9
R50 ctrl wd2	69.6	34.9	81.6	82.6	85.8	**90.2**
R50 ctrl 90ep	60.4	33.0	83.5	83.9	87.7	**91.1**

Calibration Budgets per Setting. Table 10 lists the nine-setting mean AUROC behind Figure 1c at every balanced budget, for the baselines whose results change with the number of calibration samples. Scores that use only the logits (MSP, MaxLogit, Energy, GEN, TRIM, LogitGap, GradNorm) do not change with the budget, and NCI, fDBD, and ReAct, which use only a global mean or an activation percentile, change by at most 0.2 points; these are omitted.

Table 10: Balanced per-class calibration budgets. Mean AUROC over the nine benchmark settings at k calibration samples per class (mean over four random draws of the per-class samples; HEAT uses one draw) and at full calibration, for the baselines whose results change with the number of calibration samples (scores that change by at most 0.2 points are omitted); bold marks the best score per budget. GEN, which uses no reference, is 83.5 at every budget. The released Mahalanobis estimators have no within-class degrees of freedom at one sample per class, and COMBOOD's released power transform overflows on EfficientNet-B0 in some draws at five or fewer samples per class (–). †Needs additional layer features; ‡trains energy networks.

Score	1	2	3	5	10	20	50	Full
KNN	52.7	50.0	47.5	46.8	45.9	52.2	78.0	83.5
NNGuide	75.8	78.8	80.5	82.3	83.8	84.3	84.2	84.0
MDS (raw)	–	74.6	76.4	77.3	77.5	78.0	77.5	78.7
MDS++	–	76.7	80.3	82.7	83.4	84.1	84.0	84.9
RMDS++	–	70.8	80.2	79.3	84.5	84.8	78.7	84.9
CTM	78.0	81.8	82.9	84.0	84.5	85.0	85.2	85.3
SHE	71.3	74.4	75.5	76.4	76.8	77.4	77.6	77.8
ViM	68.6	73.6	75.9	78.4	79.9	80.9	82.6	83.3
ActSub	68.4	71.9	74.2	77.3	81.1	82.6	83.4	83.7
HEAT‡	78.7	74.6	81.0	83.8	85.0	85.4	84.8	85.9
MME	65.4	70.3	71.6	72.2	72.4	72.5	71.0	71.0
COMBOOD†	–	–	–	–	76.5	77.1	82.0	85.6
ONE	**84.2**	**84.8**	**85.2**	**85.4**	**85.7**	**86.0**	**86.2**	**86.3**

C.1 Branch Evidence per OOD Set and Checkpoint

Figure 5 repeats Figure 3a for all eight OOD sets on four checkpoints: ResNet-50 plain CE, ViT-B/16 (ImageNet-21k pretraining), the ResNeXt-101 V2 recipe, and ResNet-50 AugMix. Where an OOD set leaves the ID cloud changes with the checkpoint and with the set, and the branch with the higher AUROC changes accordingly. Figure 6 repeats Figure 3b on the same four checkpoints: among the inputs that scalar fusion accepts, the class-overlap term stays near zero on ID and is lower on every OOD set, and the two branches favor different top classes for 8–21% of the accepted ID inputs against 20–59% of the accepted OOD inputs. Adding only an agreement bit to scalar fusion, $A + B + \log 2 \cdot \mathbb{1}[\hat{y}_a = \hat{y}_b]$ with the weight of MME, changes mean AUROC by $+0.07$, -0.06, $+0.50$, and $+0.87$ on the four checkpoints, against $+0.26$, -0.25, $+1.95$, and $+4.06$ for class-wise combination.

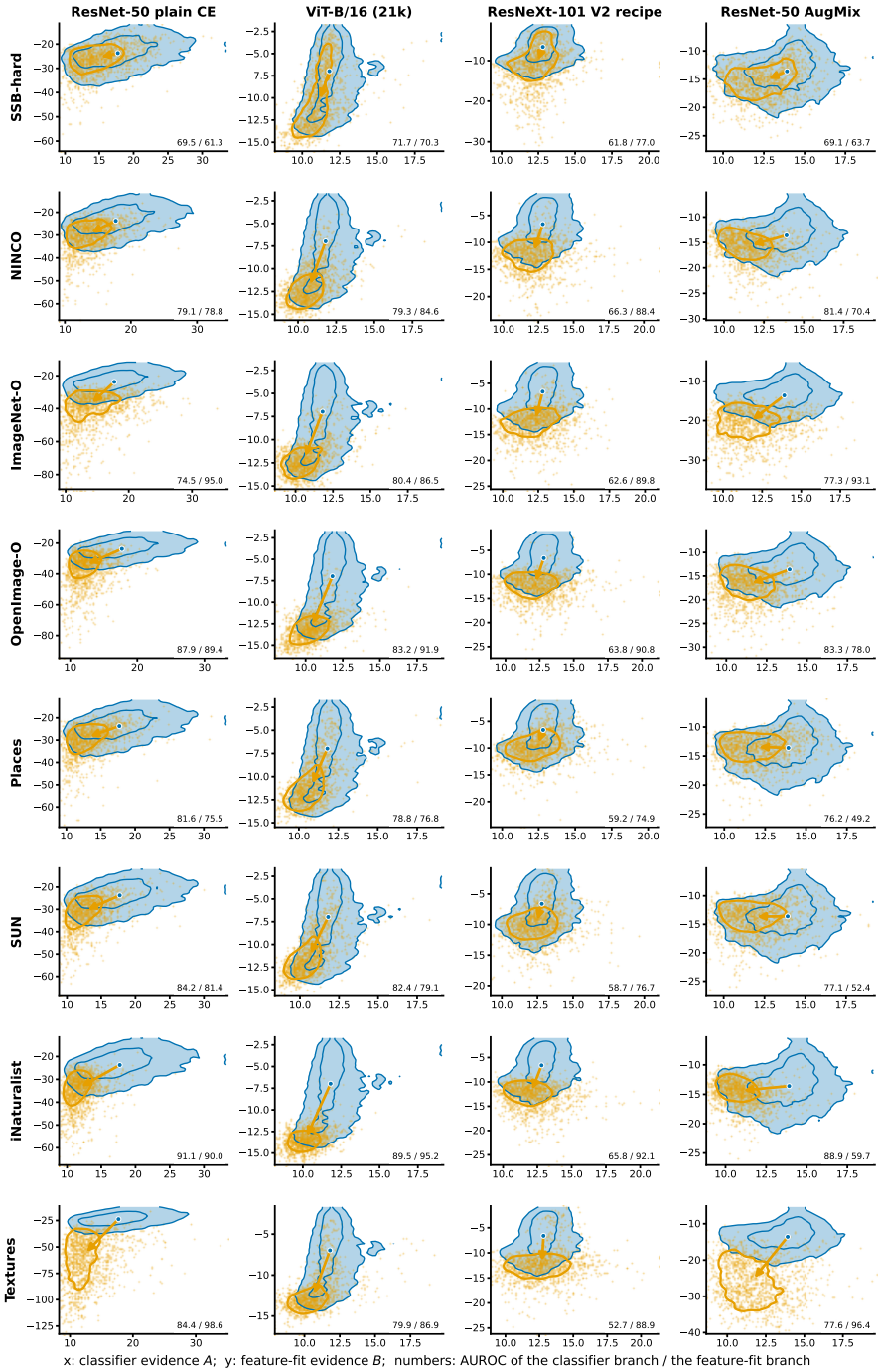


Figure 5: Classifier evidence A (x) against feature-fit evidence B (y) for each OOD set (rows) and checkpoint (columns), full ImageNet-train calibration. Blue: ID density (50% and 90%); points and contour: the OOD set and its 50% region; arrow: from the ID median to the set's median; numbers: AUROC of the classifier branch / the feature-fit branch on that set.

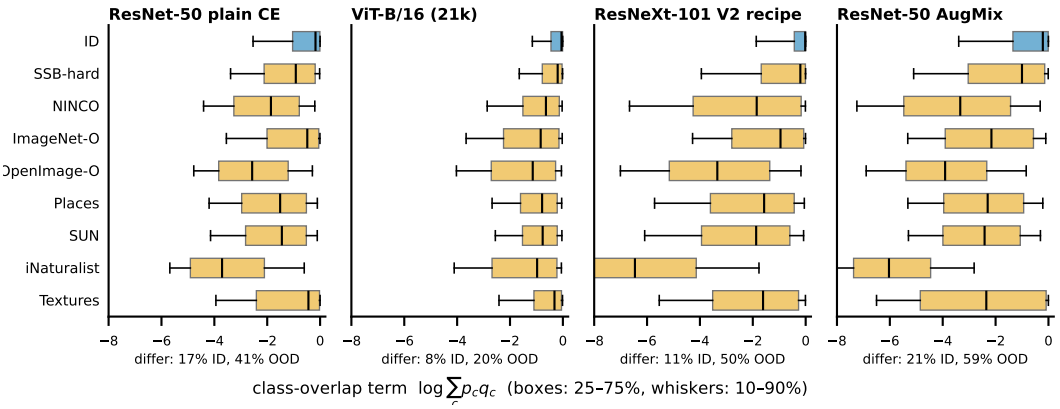


Figure 6: The class-overlap term $\log \sum_c p_c q_c$ of Equation (3) per OOD set on four checkpoints, on the inputs that scalar fusion $A + B$ accepts at a 95% true-positive rate (boxes 25–75%, whiskers 10–90%). Below each panel: share of these inputs whose two top classes, from the classifier and from feature fit, differ.

D Per-OOD Benchmark Results

Table 11: Per-OOD results at full calibration. Each entry is AUROC / FPR95.

OOD dataset	GEN	RMDS++	MDS++	ActSub	*ONE*
CIFAR-10 / ResNet-18					
cifar100	87.3 / **51.6**	**88.4** / 55.6	86.4 / 59.8	83.8 / 61.6	87.8 / 53.7
tiny_imagenet	86.9 / **51.2**	**87.7** / 56.2	85.3 / 61.7	83.3 / 60.7	86.9 / 55.0
svhn	94.2 / 30.8	95.2 / 31.4	**96.5** / **20.7**	90.5 / 53.1	96.1 / 21.9
places365	88.9 / **45.5**	**89.3** / 49.1	86.9 / 57.7	88.5 / 45.6	88.9 / 49.4
lsun	93.6 / 35.3	94.0 / 38.5	93.6 / 38.4	90.1 / 49.0	**94.4** / **33.2**
isun	92.9 / 38.0	93.7 / 40.6	93.5 / 38.0	88.6 / 53.9	**94.2** / **34.0**
textures	89.7 / 44.1	93.6 / 38.9	**94.0** / **31.4**	86.8 / 56.5	93.9 / 32.2
mnist	96.1 / 22.3	95.1 / 34.9	**98.0** / **11.7**	81.9 / 72.1	97.8 / 13.6
CIFAR-100 / ResNet-18					
cifar10	**79.2** / **79.6**	78.4 / 81.6	71.6 / 86.7	77.0 / 84.4	78.1 / 82.5
tiny_imagenet	80.2 / **74.8**	**80.4** / 77.5	74.5 / 85.5	79.9 / 79.7	80.1 / 79.9
svhn	83.5 / 76.5	87.0 / 58.6	89.8 / 51.4	**92.5** / **36.6**	90.0 / 49.7
places365	77.4 / 80.3	**81.0** / **69.8**	75.1 / 81.2	79.9 / 77.5	79.0 / 76.8
lsun	85.4 / 66.2	84.0 / 71.7	86.2 / 62.7	88.4 / 57.5	**89.0** / **56.6**
isun	84.4 / 68.9	84.0 / 70.8	86.3 / 60.1	88.6 / **55.5**	**88.8** / 56.6
textures	79.5 / 77.5	84.0 / 65.6	87.2 / **53.6**	87.6 / 54.0	**87.7** / 55.7
mnist	73.0 / 88.6	70.3 / 91.3	78.4 / 85.9	**79.2** / **72.4**	77.4 / 84.6
CIFAR-100 / WRN-40-2					
cifar10	78.9 / **78.4**	78.0 / 82.5	75.9 / 83.7	71.3 / 86.7	**79.2** / 79.7
tiny_imagenet	78.0 / 77.1	**80.7** / 77.2	77.4 / 81.2	71.9 / 83.0	79.3 / **76.3**
svhn	79.2 / 83.9	89.0 / 56.9	**89.5** / **54.7**	80.9 / 75.3	85.2 / 71.5
places365	76.0 / 80.9	**79.6** / **75.3**	74.9 / 81.9	76.8 / 80.8	77.1 / 79.3
lsun	79.1 / 81.2	**86.2** / **66.3**	80.9 / 81.5	74.7 / 87.6	81.1 / 78.8
isun	75.2 / 85.9	**84.2** / **68.6**	79.8 / 80.0	75.8 / 84.8	78.2 / 82.3
textures	76.4 / 82.7	83.3 / 68.1	**90.4** / 44.1	89.4 / **43.1**	85.8 / 61.8
mnist	76.1 / 80.3	65.1 / 91.4	79.2 / **75.1**	**81.4** / 76.4	79.0 / 76.7

Table 11 continued

OOD dataset	GEN	RMDS++	MDS++	ActSub	*ONE*
ImageNet / ResNet-50					
ssb_hard	69.2 / 88.4	70.1 / 81.0	61.5 / 90.2	**76.0** / **77.2**	67.7 / 86.9
ninco	81.2 / 77.1	83.2 / 69.8	79.7 / 75.5	**86.9** / **58.3**	83.7 / 67.1
imagenet_o	77.2 / 76.2	82.0 / 63.2	**95.1** / **29.6**	93.8 / 32.8	92.8 / 34.0
places	81.1 / 74.0	76.6 / 79.3	76.2 / 78.9	**93.1** / **32.2**	81.9 / 68.3
sun	83.0 / 72.0	77.8 / 79.3	81.8 / 69.3	**95.8** / **20.4**	86.2 / 57.9
inaturalist	92.3 / 44.6	91.9 / 43.0	91.2 / 48.4	**98.7** / **5.7**	94.9 / 29.1
textures	83.2 / 65.9	89.2 / 34.6	**98.7** / **6.3**	96.8 / 15.0	98.2 / 7.5
openimage_o	87.9 / 65.1	88.9 / 52.8	90.1 / 48.0	93.0 / **33.5**	**93.3** / 36.1
ImageNet / ViT-B					
ssb_hard	69.8 / 85.5	**72.2** / **73.3**	71.3 / 75.6	65.2 / 91.0	71.6 / 78.8
ninco	82.3 / 69.1	**86.8** / **58.1**	85.5 / 64.6	80.0 / 83.7	85.5 / 63.2
imagenet_o	85.7 / 57.0	85.9 / 67.2	86.0 / 68.5	84.2 / 73.2	**88.2** / **55.4**
places	**81.7** / **65.7**	79.9 / 78.7	77.3 / 82.4	80.8 / 80.5	80.3 / 72.2
sun	**83.9** / **62.2**	81.5 / 76.7	79.4 / 79.9	82.3 / 80.5	82.6 / 68.2
inaturalist	93.3 / 36.1	**96.0** / **21.1**	95.3 / 27.9	87.8 / 76.9	95.6 / 25.3
textures	81.6 / 63.1	83.9 / 63.1	86.3 / 59.3	85.4 / 58.3	**86.5** / **53.2**
openimage_o	88.9 / 51.3	**92.5** / 40.8	92.0 / 45.1	86.6 / 69.7	92.4 / **40.6**
ImageNet / Swin-B					
ssb_hard	78.3 / 68.0	77.7 / 64.6	80.0 / 62.0	77.0 / 74.4	**81.2** / **61.6**
ninco	89.0 / 45.6	91.6 / 41.6	91.6 / 41.1	87.5 / 62.5	**92.1** / **38.1**
imagenet_o	87.5 / 46.0	90.9 / 48.3	**93.5** / 36.2	93.0 / 35.6	**93.5** / **34.7**
places	**84.9** / **55.0**	83.8 / 71.6	82.4 / 75.4	81.2 / 84.4	84.6 / 61.6
sun	87.1 / **49.0**	85.5 / 68.9	85.2 / 69.1	82.4 / 83.8	**87.3** / 56.0
inaturalist	97.3 / 11.3	98.2 / 7.6	98.9 / 4.5	95.7 / 28.3	**99.2** / **3.6**
textures	87.8 / 43.4	88.3 / 48.0	91.0 / 34.7	90.0 / 36.8	**91.3** / **34.0**
openimage_o	94.6 / 22.5	95.5 / 25.5	96.6 / 19.8	93.9 / 37.4	**97.1** / **15.9**
ImageNet / ConvNeXt-B					
ssb_hard	69.7 / 82.7	**75.5** / **68.5**	75.1 / 69.4	70.5 / 85.5	75.1 / 71.1
ninco	84.8 / 61.9	**88.6** / **50.9**	88.5 / 51.5	85.3 / 65.8	**88.6** / 52.4
imagenet_o	82.7 / 56.5	87.0 / 58.0	88.2 / 55.4	**88.7** / **53.1**	87.6 / 55.2
places	82.4 / **62.5**	83.2 / 73.1	82.5 / 71.8	**83.4** / 72.3	82.4 / 68.8
sun	84.3 / **58.1**	84.0 / 72.7	83.8 / 70.4	**84.7** / 70.8	83.9 / 66.4
inaturalist	94.1 / 28.9	95.9 / 18.5	**96.1** / **17.9**	93.4 / 38.3	96.0 / 21.0
textures	85.6 / 50.4	86.7 / 57.6	88.2 / 52.3	**88.7** / **49.5**	88.6 / 49.6
openimage_o	89.5 / 45.1	92.7 / 38.8	**93.2** / **36.9**	91.9 / 46.0	93.0 / 37.1
ImageNet / DeiT-B					
ssb_hard	67.5 / 86.4	**73.5** / **73.7**	72.7 / 75.9	67.6 / 92.0	72.2 / 80.0
ninco	81.9 / 70.5	**87.2** / **58.1**	86.2 / 63.1	81.2 / 83.9	85.9 / 64.9
imagenet_o	81.8 / 64.9	84.9 / 72.5	85.9 / 70.2	85.1 / 70.2	**86.2** / **63.7**
places	81.1 / **68.0**	**81.6** / 75.2	79.6 / 79.2	77.2 / 89.2	80.6 / 72.6
sun	**82.6** / **65.7**	82.4 / 76.5	80.9 / 81.3	78.3 / 92.0	82.0 / 72.2
inaturalist	92.3 / 41.4	**96.7** / **16.3**	95.9 / 21.7	87.9 / 85.4	94.9 / 29.2
textures	84.9 / **57.8**	85.4 / 62.6	86.0 / 62.7	84.8 / 66.3	**86.7** / 57.9
openimage_o	88.5 / 54.3	**92.2** / **44.6**	91.7 / 49.6	87.4 / 74.5	91.7 / 47.5
ImageNet / EfficientNet-B0					
ssb_hard	70.8 / 86.8	69.7 / **79.6**	64.7 / 88.8	64.5 / 90.3	**71.9** / 80.1
ninco	81.9 / 72.4	83.2 / 69.0	76.8 / 79.3	74.6 / 83.4	**84.7** / **63.5**
imagenet_o	84.1 / 62.7	83.6 / 68.9	**90.0** / 50.5	87.3 / 55.4	89.2 / **46.4**
places	**78.7** / **72.4**	76.1 / 78.2	68.2 / 86.7	65.2 / 90.3	76.6 / 75.3
sun	**79.8** / 71.6	77.1 / 77.9	73.9 / 79.1	71.2 / 85.1	79.6 / **69.8**
inaturalist	90.6 / 48.1	93.5 / 31.6	88.6 / 52.2	81.0 / 75.9	**94.4** / **27.6**
textures	83.7 / 62.4	85.6 / 53.1	**94.4** / **23.1**	90.4 / 38.2	92.9 / 29.6
openimage_o	87.0 / 62.5	88.0 / 56.3	83.6 / 61.6	79.1 / 73.1	**89.6** / **47.9**