LSF: LAYER SELECTION AND FUSION FOR POST-HOC OUT-OF-DISTRIBUTION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distance-based out-of-distribution detection typically relies on the penultimate representation, whose optimization for ID class discrimination may suppress lower-level variation useful for detecting distribution shifts. We introduce LSF (Layer Selection and Fusion for OOD detection), a post-hoc framework that exploits complementary information across network layers. LSF selects an intermediate layer with the largest decrease in entropy density between consecutive layers and concatenates it with the penultimate representation to preserve their distinct feature spaces. Across ViT, Swin, and ConvNeXt architectures, LSF improves OOD detection across a range of evaluated settings, including near- and far-OOD and covariate shifts.

1 INTRODUCTION

Deep neural networks (DNNs) can produce overconfident predictions on inputs that differ from their training distribution. Distinguishing such out-of-distribution (OOD) inputs from in-distribution (ID) inputs is essential for reliable deployment, particularly in safety-critical applications such as medical diagnosis and autonomous driving (Lu et al., 2025; Yang et al., 2024). Post-hoc OOD detection performs this task using an already trained network, making the representations learned by the network central to detection performance.

A common approach measures how closely an input representation matches an ID class distribution. Mahalanobis-based detectors (Lee et al., 2018) quantify this proximity using class means and a shared covariance matrix, giving greater weight to deviations along directions with little ID variability. Mahalanobis++ (Müller and Hein, 2025) further normalizes each feature vector by its ℓ_2 norm before estimating class statistics, reducing sensitivity to feature magnitude. However, distance-based detection also depends on the representation being measured. The penultimate representation is optimized for ID class discrimination and may suppress lower-level variation that remains informative for detecting distribution shifts.

Earlier hidden layers provide complementary views before the final classification space, motivating multilayer extensions of Mahalanobis-based scoring. X-Mahalanobis (Wei et al., 2025) combines layer features through variance-based weighted addition, requiring equal-dimensional representations and therefore being naturally suited to homogeneous architectures such as Vision Transformers. We instead preserve distinct feature spaces by concatenating a selected intermediate representation with the penultimate representation.

Retaining representations from different depths raises the question of which layers to combine. Figure 1 shows that different layers can provide complementary evidence for different shifts: intermediate representations can better separate covariate shifts, whereas penultimate representations remain more discriminative for semantic OOD inputs. The figure also shows high centered kernel alignment (CKA) (Kornblith et al., 2019) among intermediate layers, suggesting representational redundancy and motivating the use of a single intermediate companion. Because the framework is post-hoc, this layer must be selected using ID data alone, without auxiliary OOD examples.

We select the intermediate layer with the largest entropy-density drop between consecutive layers. Specifically, entropy density is defined as spectral entropy normalized by feature dimension, and the resulting drop equals the difference in the dimensionality-normalized log effec-

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.

tive rank (or, for equal-width layers, the normalized log effective-rank ratio). This provides an ID-only heuristic for identifying transitions in dimension-normalized spectral complexity. The selected intermediate and penultimate representations can have different dimensions, particularly in hierarchical networks such as Swin and ConvNeXt, where channel width changes across stages. Concatenation accommodates these unequal widths without explicit dimension alignment or projection. The resulting joint representation can be modeled with either a block-diagonal covariance, which omits cross-layer covariance terms, or a full covariance that models cross-layer dependencies. We evaluate both alternatives.

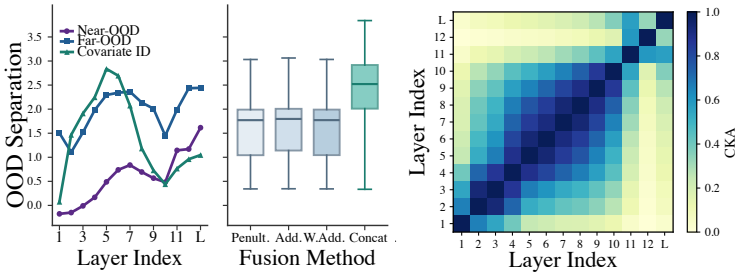


Figure 1: Motivation for Layer Selection and Fusion. **Left:** Intermediate and penultimate features are more sensitive to covariate and semantic shifts, respectively. **Center:** Concatenating intermediate and penultimate features yields higher median OOD separation than penultimate-only scoring, addition, and weighted addition. OOD separation is the difference between the mean Mahalanobis energy of OOD and ID samples (Appendix B). **Right:** High CKA similarity among intermediate features suggests redundancy among them.

We introduce **LSF** (Layer Selection and Fusion for OOD detection), a post-hoc framework that separates layer selection and representation fusion from the choice of distance function. We instantiate the framework with Mahalanobis distance and an intermediate–penultimate pair. Using labeled ID calibration data, LSF selects one intermediate layer through the entropy-density criterion, independently ℓ_2-normalizes the two representations, and concatenates them before estimating class means and covariance. Figure 1 (center) shows that the selected pair achieves higher median separation than penultimate-only modeling and additive fusion. These separation results are retrospective diagnostics; OOD data are never used for layer selection or calibration.

Contribution. LSF is a post-hoc OOD framework comprising (i) entropy-density-based layer selection and (ii) concatenation into a joint representation followed by covariance-based distance estimation, as illustrated in Figure 2. The design accommodates unequal feature widths and different covariance structures without modifying the pretrained encoder. Although we focus on Mahalanobis distance in the main evaluation, the layer-selection and concatenation components are distance-agnostic and also apply to Euclidean distance, as demonstrated in Section 4.

Evaluation across architectures and distribution shifts. We evaluate LSF on ViT, Swin, and ConvNeXt backbones under balanced and long-tailed ID calibration settings, covering semantic OOD and covariate shifts. Across architectures and scoring functions, LSF improves OOD detection in a majority of evaluated settings, including near- and far-OOD as well as covariate shifts. The block-diagonal variant achieves the best performance more frequently than the alternatives, suggesting that modeling the selected representations with separate covariance blocks can be more effective than explicitly modeling their cross-layer covariance.

2 Related Work

OOD detection has evolved from output-based heuristics toward methods exploiting the geometry of deep representations (Lu et al., 2025; Yang et al., 2024). We focus on post-hoc methods and omit training-based approaches (Haas et al., 2024; Lu et al., 2024; Regmi et al., 2024; Tao et al., 2023) that require model retraining.

Output- and Logit-Based Methods. Maximum softmax probability (Hendrycks and Gimpel, 2017) provides a simple OOD score, while energy-based methods (Liu et al., 2020; Regmi, 2026; Xu et al., 2024) derive scores from logits. These methods operate on final outputs rather than intermediate representations.

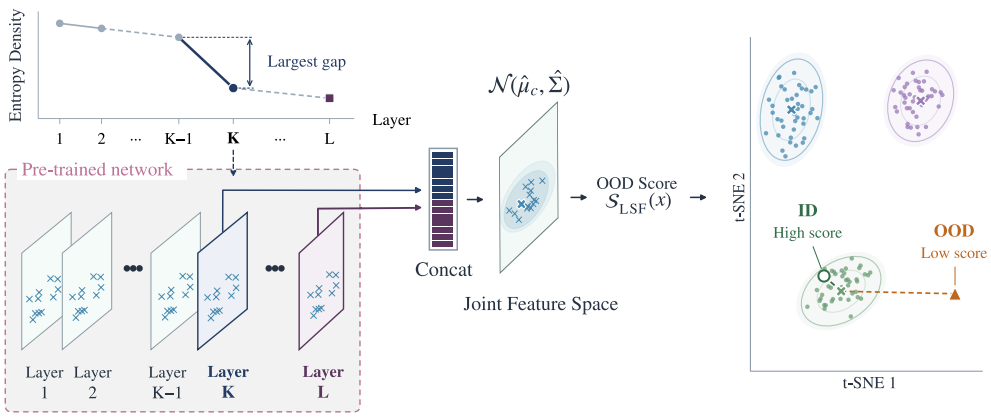


Figure 2: **Overview of the LSF Framework. Left:** Identify the intermediate layer with the largest entropy density drop, anchoring the penultimate layer. **Center**: Concatenate the intermediate and penultimate features into a unified representation. **Right:** In the joint space, ID samples tightly cluster around established class centroids, pushing OOD samples away.

Feature-Based and Geometric Methods. Mahalanobis distance (Lee et al., 2018) models class-conditional features with class means and shared covariance, while Relative Mahalanobis distance (Ren et al., 2021) contrasts class-conditional and background distributions. Non-Mahalanobis alternatives include k-nearest neighbors (Sun et al., 2022), KPCA (Fang et al., 2024), and NNGuide (Park et al., 2023). ReAct (Sun et al., 2021) truncates extreme activations, while Mahalanobis++ (Müller and Hein, 2025) normalizes features to the unit hypersphere. These methods primarily use the penultimate representation and do not address complementary information across layers.

Multilayer and Hierarchical Detection. Multilayer OOD detectors exploit representations at different abstraction levels (Lin et al., 2021; Sun et al., 2022). Trusted (Colombo et al., 2022) and X-Mahalanobis (Wei et al., 2025) fuse multiple layers additively; X-Mahalanobis additionally uses variance-based weighting, while its fine-tuning variant adapts a task-specific linear head. We compare against its post-hoc variant for a fair comparison.

Covariance Entropy and Layer Selection. Information bottleneck and neural collapse studies suggest that deep representations progressively concentrate into lower-dimensional subspaces (Papyan et al., 2020; Tishby and Zaslavsky, 2015). Covariance eigenspectrum entropy measures representational complexity (Ammar et al., 2024; Harun et al., 2025; Janiak et al., 2026). We use the entropy-density drop between consecutive layers as an ID-only criterion for layer selection.

Connection to Our Work. LSF uses a two-layer post-hoc design that retains the penultimate representation, selects one intermediate layer via entropy-density drop, and concatenates the two for OOD detection. Unlike additive multilayer fusion, it preserves distinct feature spaces and accommodates heterogeneous feature widths. The joint representation is modeled using a tied within-class covariance, with full and block-diagonal covariance structures evaluated empirically. LSF requires neither auxiliary OOD data nor classifier fine-tuning.

3 METHOD

Preliminaries. Let $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the ID training set, where x_i is a sample, $y_i \in \{1, \ldots, C\}$ is the label, and N_c is the number of training samples in class c. A pre-trained network $\mathcal{F} = g \circ h$ consists of a feature extractor h and a classifier head g. We consider h to be an L-layer network, with $L \geq 3$, making layer L the terminal feature space of h and hence the penultimate layer of $\mathcal{F}$. During offline calibration, we extract features $h_l(x) \in \mathbb{R}^{D_l}$ from each designated layer $l = 1, \ldots, L$, applying global spatial pooling when necessary.

We characterize the distribution of within-class variation at each layer through its covariance spectrum. To reduce sensitivity to feature magnitude across layers, we independently ℓ_2-normalize the representation for input x at layer l, producing $\tilde{h}_l(x)$.

For each layer l, we estimate the class-conditional mean $\hat{\mu}_{c,l}$ and the tied within-class covariance $\hat{\Sigma}_l$:

$$\hat{\mu}_{c,l} = \frac{1}{N_c} \sum_{i:y_i=c} \tilde{h}_l(x_i), \qquad \hat{\Sigma}_l = \frac{1}{N} \sum_{c=1}^{C} \sum_{i:y_i=c} \big(\tilde{h}_l(x_i) - \hat{\mu}_{c,l}\big)\big(\tilde{h}_l(x_i) - \hat{\mu}_{c,l}\big)^\top. \tag{1}$$

As the pooled class-centered residuals can span at most $N - C$ independent directions, the empirical covariance can be singular or poorly conditioned when the feature dimension is large. We therefore apply Ledoit–Wolf shrinkage (Ledoit and Wolf, 2004):

$$\hat{\Sigma}_{l,\mathrm{shrink}} = (1 - \gamma_l)\hat{\Sigma}_l + \gamma_l \frac{\mathrm{Tr}(\hat{\Sigma}_l)}{D_l}\mathbf{I}, \tag{2}$$

where $\gamma_l \in [0, 1]$ is estimated from the class-centered ID features according to the Ledoit–Wolf rule and is not tuned using OOD data. Consequently, $\hat{\Sigma}_{l,\mathrm{shrink}}$ is positive definite and well-conditioned.

3.1 ID-Only Layer Selection

We use the eigenvalue spectrum of the shrunk covariance to characterize how within-class variation is distributed across feature dimensions. Let $\{\lambda_i^{(l)}\}_{i=1}^{D_l}$ denote the eigenvalues of $\hat{\Sigma}_{l,\mathrm{shrink}}$. For numerical stability, we replace each eigenvalue by $\max(\lambda_i^{(l)}, \epsilon)$, where $\epsilon > 0$ is small. We then define

$$\bar{\lambda}_i^{(l)} = \frac{\lambda_i^{(l)}}{\sum_{j=1}^{D_l} \lambda_j^{(l)}}, \qquad H_l = -\sum_{i=1}^{D_l} \bar{\lambda}_i^{(l)} \ln \bar{\lambda}_i^{(l)}. \tag{3}$$

where H_l is the spectral entropy. We normalize H_l by the feature dimension D_l and define the entropy density $\rho_l = \frac{H_l}{D_l}$. The entropy-density drop between consecutive layers is

$$\Delta_l = \rho_{l-1} - \rho_l = \frac{H_{l-1}}{D_{l-1}} - \frac{H_l}{D_l}, \qquad l = 2, \ldots, L-1. \tag{4}$$

Proposition 1 (Entropy-density transition). The quantity Δ_l is the dimension-normalized difference in log-effective rank between consecutive layers. For equal-width layers, it reduces to the dimensionality-normalized logarithm of the effective-rank ratio. Consequently, a large positive Δ_l indicates a sharp decrease in dimension-normalized effective complexity between consecutive layers. The assumptions and proofs of our propositions are given in Appendix A.

We use Δ_l solely as an ID-based criterion for layer selection; it does not by itself imply that the selected layer maximizes OOD discrimination. Motivated by the commonly observed class-aligned geometry of penultimate representations (Papyan et al., 2020), we always retain the penultimate layer L and select one intermediate layer, forming $\mathcal{K} = \{l^\star, L\}$, where $l^\star = \mathrm{argmax}\, \Delta_l$. This selects exactly one intermediate layer even when all the candidate drops are nonpositive. Hence, the procedure introduces no layer-count hyperparameter.

3.2 Concatenated Fusion

Given the selected layers $\mathcal{K} = \{l^\star, L\}$, we concatenate their normalized representations:

$$\boldsymbol{\phi}(x) = \big[\tilde{h}_{l^\star}(x); \tilde{h}_L(x)\big] \in \mathbb{R}^{D_\mathcal{K}}, \qquad D_\mathcal{K} = D_{l^\star} + D_L. \tag{5}$$

Unlike additive fusion, concatenation does not project the two representations into a common coordinate space. It therefore accommodates different layer widths while retaining the coordinates of each selected representation for subsequent joint covariance estimation.

The concatenated representation permits two covariance constructions. The first one estimates a **full tied covariance** directly in the concatenated space, where $\hat{\boldsymbol{\mu}}_c^\mathcal{K} = \frac{1}{N_c} \sum_{i:y_i=c} \boldsymbol{\phi}(x_i)$ is the class mean.

The corresponding tied and shrinkage regularized covariances are:

$$\hat{\Sigma}_{\mathcal{K}} = \frac{1}{N}\sum_{c=1}^{C}\sum_{i:y_i=c}\big(\boldsymbol{\phi}(x_i)-\hat{\boldsymbol{\mu}}_c^{\mathcal{K}}\big)\big(\boldsymbol{\phi}(x_i)-\hat{\boldsymbol{\mu}}_c^{\mathcal{K}}\big)^\top, \quad \hat{\Sigma}_{\mathcal{K},\text{shrink}} = \begin{bmatrix}\hat{\Sigma}_{L,L} & \hat{\Sigma}_{L,l^\star}\\ \hat{\Sigma}_{l^\star,L} & \hat{\Sigma}_{l^\star,l^\star}\end{bmatrix}. \tag{6}$$

Alternatively, we can estimate a **block-diagonal tied covariance**, by setting $\hat{\Sigma}_{L,l^\star} = \hat{\Sigma}_{l^\star,L} = 0$. This formulation models both representations independently and ignores cross-layer covariance estimation, making it more computationally efficient. The block-diagonal covariance is regularized by construction through Eq. 2.

The two covariance constructions represent different modeling choices rather than universally ordered alternatives. The block-diagonal formulation is more computationally efficient but omits cross-layer covariance, whereas the full covariance explicitly models dependencies between the selected representations. We therefore evaluate both formulations empirically.

3.3 Joint Mahalanobis OOD Scoring

For a test sample x, the OOD confidence score is defined as the negative minimum squared Mahalanobis distance over the ID classes:

$$\mathcal{S}_{LSF}(x) = -\min_{c\in\{1,\dots,C\}}\big(\boldsymbol{\phi}(x)-\hat{\boldsymbol{\mu}}_c^{\mathcal{K}}\big)^\top\hat{\Sigma}_{\mathcal{K},\text{shrink}}^{-1}\big(\boldsymbol{\phi}(x)-\hat{\boldsymbol{\mu}}_c^{\mathcal{K}}\big). \tag{7}$$

Higher values correspond to smaller joint Mahalanobis distance to at least one ID class.

Proposition 2 (Cross-layer decomposition). For **full covariance**, the joint Mahalanobis distance consists of a marginal distance for the penultimate representation and a conditional term measuring how the intermediate representation deviates from its covariance-conditioned prediction given the penultimate representation. For **block-diagonal covariance**, the joint distance instead decomposes into the sum of marginal Mahalanobis distances for the penultimate and intermediate representations, with no cross-layer conditional term.

Proposition 3 (Non-degradation and increased sensitivity to intermediate-layer shifts). Concatenating the intermediate representation cannot reduce the expected quadratic OOD separation provided by the penultimate representation. For **full covariance**, the joint distance retains the penultimate-layer contribution and adds a nonnegative conditional term that captures intermediate-layer shifts not explained by the penultimate-layer shift. For **block-diagonal covariance**, the joint distance instead retains the penultimate-layer contribution and adds a nonnegative marginal term for the intermediate representation. Thus, in both cases, the joint representation is at least as sensitive as the penultimate representation and becomes strictly more sensitive whenever the intermediate representation contains additional shift information that contributes positively to the corresponding quadratic term.

3.4 Calibration and Inference

All statistics required by LSF are estimated once during offline calibration using labeled ID data. Specifically, calibration extracts representations from the designated layers, computes the per-layer covariance spectra and entropy-density drops used for layer selection, and then estimates the class means and covariance in the concatenated space. No OOD samples are required, and the pre-trained network remains frozen.

At inference, a test input is passed through the frozen network, the representations at $l^\star$ and L are independently ℓ_2-normalized and concatenated. The resulting vector is then scored using Eq. 7.

4 Evaluation

4.1 Experimental Setup

Datasets. We follow the OpenOOD benchmark (Zhang et al., 2023), and evaluate across two in-distribution (ID) calibration sets: standard large-scale class-balanced ImageNet-1K (Russakovsky

et al., 2015) and the long-tailed ImageNet-LT (Liu et al., 2019). All evaluations are performed in a post-hoc, training-free setting with frozen backbones. We adopt the OOD categorization defined by the benchmark. Near-OOD datasets include SSB-Hard (Yang et al., 2022) and NINCO (Bitterwolf et al., 2023), which contain samples that are visually similar to ID data but belong to distinct, unseen semantic classes. Far-OOD datasets include iNaturalist (Van Horn et al., 2018), Textures (Cimpoi et al., 2014), and OpenImage-O (Yang et al., 2022), containing samples with obvious domain shift. We also evaluate on ImageNet-C (Hendrycks and Dietterich, 2019), ImageNet-R (Hendrycks et al., 2021), ImageNet-V2 (Recht et al., 2019), and ImageNet-ES (Baek et al., 2024) as covariate-shifted ID datasets and, these datasets retain the semantic classes of the ID data but differ from the calibration set in their input distributions.

Models and Baselines. We primarily use a frozen ViT-B/16 pretrained on ImageNet-21K as the backbone. For comparison, we use confidence-based methods, including MSP (Hendrycks and Gimpel, 2017), Energy (Liu et al., 2020), ReAct (Sun et al., 2021), SCALE (Xu et al., 2024), and AdaSCALE (Regmi, 2026); non-parametric feature-based methods, including KNN (Sun et al., 2022), NNGuide (Park et al., 2023), and KPCA (Fang et al., 2024); and parametric feature-based methods, including Mahalanobis (Lee et al., 2018), Relative Mahalanobis (Ren et al., 2021), Mahalanobis++ (Müller and Hein, 2025), Relative Mahalanobis++, and Radially Scaled Mahalanobis (Janiak et al., 2026). We additionally compare with X-Mahalanobis (Wei et al., 2025) in a post-hoc setting without fine-tuning, denoted as X-Maha w/o FT. We evaluate two variants of LSF: LSF-C uses full covariance, while LSF-B uses block-diagonal covariance; both share the same layer-selection and feature-fusion procedure. For all baselines, we use the official code published by the authors. Appendix D reviews all the baseline methods briefly, and Appendix E provides the reproduction details of LSF.

Metrics. We report FPR95 as our primary metric, while the corresponding AUROC results can be found in the appendix. All performance differences are in *percentage points* (pp). The threshold τ is selected such that $P(\mathcal{S}(x) \geq \tau \mid x \in \mathcal{D}_{\text{ID}}) = 0.95$. An input x is classified as ID if $\mathcal{S}(x) \geq \tau$, and as OOD otherwise.

Table 1: FPR95 (↓, %) using ViT-B/16 under ImageNet-1K ID calibration. SSB: SSB-Hard; NIN: NINCO; iNat: iNaturalist; TxT: Textures; OpO: OpenImage-O. Im-C/R/V2/ES denote ImageNet-C/R/V2/ES, respectively. (**Bold**: best, Underline: second-best)

Method	Near-OOD		Far-OOD			Covariate-shifted ID				Avg.
	SSB	NIN	iNat	TxT	OpO	Im-C	Im-R	Im-V2	Im-ES	
MSP	76.73	62.91	34.35	58.39	45.23	74.13	66.44	90.21	70.15	64.28
Energy	66.05	52.51	29.73	47.07	34.46	67.26	67.86	91.29	70.78	58.56
ReAct	**65.47**	50.15	18.75	44.45	27.38	66.26	60.95	90.27	68.29	54.66
SCALE	84.03	86.22	86.04	88.26	79.67	88.29	97.05	94.80	83.99	87.59
AdaSCALE	69.41	55.69	24.85	52.00	34.09	68.98	62.49	**89.50**	66.80	58.20
KNN	82.71	64.72	16.20	42.93	31.47	72.79	49.61	89.77	65.70	57.32
NNGuide	77.65	56.00	11.72	40.09	25.62	69.27	47.85	89.55	63.89	53.51
KPCA	85.67	75.40	58.68	69.91	66.19	81.34	71.52	91.55	82.42	75.85
Maha	85.04	70.92	13.14	64.57	38.15	84.93	51.94	89.84	84.11	64.74
RS-Maha	67.68	47.32	**5.15**	48.26	24.02	73.25	48.17	89.98	72.21	52.89
rMaha	77.66	59.56	10.62	58.79	29.59	81.33	51.30	89.77	80.63	59.92
Maha++	67.68	47.32	**5.15**	48.26	24.02	73.25	48.17	89.98	72.21	52.89
rMaha++	67.61	47.30	5.16	48.26	24.10	73.30	48.32	90.02	72.11	52.91
X-Maha w/o FT	68.90	48.54	5.30	46.15	23.02	73.57	47.18	90.13	71.93	52.75
LSF-C	65.71	43.55	5.45	15.89	**20.80**	42.63	42.16	90.26	46.38	41.42
LSF-B	66.67	**43.22**	6.32	**12.80**	22.28	**37.44**	**42.13**	91.05	**43.82**	**40.63**

4.2 Comparison to State-of-the-Art

Results under ImageNet-1K ID Calibration. Table 1 shows that LSF-B achieves the lowest mean FPR95 of 40.63%, improving over X-Maha w/o FT by 12.12 pp and Mahalanobis++ by 12.26 pp. LSF-C achieves the second-lowest mean FPR95 of 41.42%, improving over the same baselines by

11.33 pp and 11.47 pp, respectively. Relative to X-Maha w/o FT, LSF-C and LSF-B reduce FPR95 by 4.09 pp and 3.78 pp on Near-OOD, 10.78 pp and 11.02 pp on Far-OOD, and 15.34 pp and 17.09 pp under covariate shift, respectively. The largest improvements are observed on Textures, ImageNet-C, and ImageNet-ES. Together, LSF-B and LSF-C obtain the best results on 6 of the 9 benchmarks, with LSF-B leading on 5 and LSF-C on 1.

Table 2: FPR95 (↓, %) using ViT-B/16 under ImageNet-LT ID calibration.

Method	Near-OOD		Far-OOD			Covariate-shifted ID				Avg.
	SSB	NIN	iNat	TxT	OpO	Im-C	Im-R	Im-V2	Im-ES	
MSP	76.55	62.69	34.04	58.19	44.89	73.94	66.23	90.14	70.01	64.08
Energy	65.22	51.62	28.85	46.28	33.76	66.65	67.04	90.85	70.31	57.84
ReAct	**64.87**	49.24	18.26	43.88	26.74	65.85	60.11	89.91	67.84	54.08
SCALE	83.59	85.84	85.69	87.93	79.34	87.95	96.93	94.52	83.75	87.28
AdaSCALE	68.88	54.79	24.40	51.51	33.59	68.68	62.10	89.30	66.61	57.76
KNN	96.68	96.10	85.01	92.43	95.06	88.45	84.17	94.34	91.43	91.52
NNGuide	91.94	88.69	66.09	70.80	78.00	83.93	73.80	94.42	84.47	81.35
KPCA	88.34	80.38	67.87	75.74	73.29	85.01	77.22	92.42	85.65	80.66
Maha	85.54	71.17	13.11	65.04	38.54	85.00	51.85	89.76	84.53	64.95
RS-Maha	68.98	48.87	4.37	48.05	24.55	73.05	46.92	89.37	73.09	53.03
rMaha	78.02	58.92	10.37	57.68	29.87	80.32	52.00	89.60	80.22	59.67
Maha++	68.98	48.87	4.37	48.05	24.55	73.05	46.92	89.37	73.09	53.03
rMaha++	68.85	48.73	4.36	47.89	24.51	72.88	46.91	89.36	72.91	52.93
X-Maha w/o FT	69.94	49.15	4.53	45.89	23.25	72.87	45.39	89.20	72.32	52.50
LSF-C	67.34	44.64	**4.00**	23.65	**20.98**	53.80	41.83	**88.94**	52.43	44.18
LSF-B	68.01	**42.76**	4.48	**13.00**	22.16	**36.18**	**39.58**	89.11	**41.67**	**39.66**

Results under ImageNet-LT ID Calibration. Table 2 shows that LSF-B achieves the lowest mean FPR95 of 39.66%, improving over X-Maha w/o FT and Relative Mahalanobis++ by 12.84 pp and 13.27 pp, respectively. LSF-C achieves the second-lowest mean FPR95 of 44.18%, improving over the same baselines by 8.32 pp and 8.75 pp, respectively. Consistent with ImageNet-1K results, the largest improvements occur on Textures, ImageNet-C, and ImageNet-ES. Overall, LSF-B and LSF-C achieve the best results on 5 and 3 benchmarks, respectively. Together, they show the lowest FPR95 on 8 of the 9 datasets.

Table 3: FPR95 (↓, %) of non-parametric feature-based detectors using ViT-B/16 under ImageNet-1K ID calibration. *Penultimate Only* uses the detector's penultimate-layer feature; *with LSF* concatenates it with the intermediate-layer chosen by our layer selection criterion.

Method	Feature	Near OOD		Far OOD			Covariate-Shifted ID				Avg.
		SSB	NIN	iNat	TxT	OpO	Im-C	Im-R	Im-V2	Im-ES	
KNN	Penultimate Only	82.71	64.72	**16.20**	42.93	31.47	72.79	49.61	**89.77**	65.70	57.32
	with LSF	**81.69**	**63.77**	16.39	**38.83**	**29.69**	**70.60**	**48.33**	89.85	**65.02**	**56.02**
NNGuide	Penultimate Only	77.65	56.00	11.72	40.09	25.62	69.27	47.85	**89.55**	63.89	53.51
	with LSF	**66.18**	**43.93**	**7.96**	**30.96**	**17.75**	**62.30**	**45.71**	90.01	**62.89**	**47.52**
KPCA	Penultimate Only	85.67	75.40	58.68	69.91	66.19	81.34	71.52	91.55	82.42	75.85
	with LSF	**76.65**	**60.00**	**13.20**	**39.59**	**30.35**	**64.51**	**47.95**	**89.93**	**65.35**	**54.17**

Applicability beyond Mahalanobis Scoring. Table 3 shows that concatenating the penultimate feature with the intermediate representation selected by LSF improves three non-Mahalanobis detectors, KNN, NNGuide, and KPCA, reducing mean FPR95 by 1.30 pp, 5.99 pp, and 21.68 pp, respectively. These results demonstrate that the benefit of selecting a complementary intermediate layer and fusing it with the penultimate representation is not specific to covariance-based Mahalanobis scoring, but extends to OOD detectors that do not require covariance estimation.

Additional Results. In Appendix C.1, LSF-C improves the mean AUROC over X-Maha w/o FT by 3.86 pp and 3.09 pp under ImageNet-1K and ImageNet-LT calibration, respectively. Under ImageNet-LT calibration, LSF-B achieves the highest mean AUROC of 87.56%, improving over X-Maha w/o FT by 4.64 pp.

In Appendix C.2, with ImageNet-LT calibration, both LSF-C and LSF-B achieve lower mean FPR95 and higher mean AUROC than the compared baselines on CLIP ViT-B/16, ConvNeXt-T, and Swin-T. Relative to the strongest compared baselines, LSF-C reduces mean FPR95 by 19.35 pp, 4.66 pp, and 4.08 pp on these backbones, respectively. The corresponding LSF-B's reductions are 21.90 pp, 6.29 pp, and 4.22 pp. Further evaluation results across 26 checkpoints are reported in Appendix C.3.

4.3 Ablation Studies

Layer Selection Criteria. Table 4 compares ID-only criteria for selecting an intermediate companion layer with LSF. All LSF variants in this ablation use the full covariance. The first four criteria following H_l/D_l differ from LSF-C only in how the intermediate layer is selected; they fuse the selected intermediate representation with the penultimate feature and process the resulting representation in the same manner as LSF-C. In contrast, the penultimate-only method uses only the penultimate feature. The random-layer policy randomly selects an intermediate representation and uses it for OOD detection without the penultimate feature; we report its average performance across all intermediate layers, equivalent to uniform random selection in expectation. The OOD-informed oracle uses OOD labels to select the intermediate layer. Since OOD labels are unavailable in practice, this oracle is included only as a reference for comparison.

Table 4: Ablation of layer-selection criteria under ImageNet-1K ID calibration; mean FPR95 (↓, %).

Criterion	ViT-B/16			ConvNeXt-T		
	Near	Far	Covariate	Near	Far	Covariate
H_l/D_l (LSF-C)	**54.63**	**14.04**	**55.36**	**63.60**	**25.20**	**64.45**
$H_l/\log D_l$	54.63	14.04	55.36	63.89	25.63	68.05
H_l	54.63	14.04	55.36	63.89	25.63	68.05
$\exp(H_l)$	54.63	14.04	55.36	78.58	41.52	74.56
$\exp(H_l)/D_l$	54.63	14.04	55.36	63.89	25.63	68.05
Penultimate Layer only	57.56	25.98	71.08	62.95	32.46	71.02
Random Layer	60.05	22.56	65.79	65.60	27.18	66.10
Oracle (OOD-informed)	52.24	12.36	55.05	62.08	20.05	58.68

The first five criteria in Table 4 select the same intermediate layer and therefore yield identical FPR95 on ViT-B/16. On ConvNeXt-T, where the alternative criteria select intermediate layers differently from LSF, H_l/D_l achieves the best performance among the tested ID-only criteria across all three shift groups. Relative to the penultimate-only policy, H_l/D_l in LSF-C reduces FPR95 by 2.93–15.72 pp on ViT-B/16. On ConvNeXt-T, it reduces FPR95 by 6.57–7.26 pp on Far and Covariate shifts, while slightly increasing FPR95 by 0.65 pp on Near-OOD. LSF also outperforms the random-layer policy across all three shift groups on both backbones. On ConvNeXt-T, $H_l/\log D_l$, H_l, and $\exp(H_l)/D_l$ yield identical results, whereas $\exp(H_l)$ performs substantially worse. This suggests that the layer-selection behavior is robust to several alternative criterion formulations, but is sensitive to applying the exponential transformation to entropy without normalization by D_l. Finally, the layer selected by LSF remains within 2.39 pp of the OOD-informed oracle on ViT-B/16, with a larger performance gap of 1.52–5.77 pp on ConvNeXt-T.

Table 5: Ablation of additional layers using ViT-B/16 under ImageNet-1K ID calibration.

# of Layers	CKA ↓	Mean CCA ↓	Residual Variance ↑	Condition No. ↓	FPR95 ↓
2	**0.140**	**0.193**	**0.388**	**5.21**	**41.42**
3	0.941	0.988	0.019	6.01	67.97
4	0.764	0.896	0.112	6.14	67.77
5	0.471	0.742	0.107	6.21	67.12

Impact of Augmenting More Intermediate Layers. Table 5 shows that, for ViT-B/16, incorporating additional intermediate layers increases the representational similarity among the layers in the joint representation, as measured by centered kernel alignment (CKA), canonical correlation analysis (CCA), and residual variance. These results suggest that adding more layers introduces increasing representational redundancy while making covariance estimation less well-conditioned, which can ultimately hurt OOD detection.

Table 6: Ablation of layer aggregation strategy using ViT-B/16; mean FPR95 (↓, %). All rows use ℓ_2 normalization, entropy-density drop for layer selection, and empirical covariance without shrinkage.

Aggregation Strategy	ImageNet-1K			ImageNet-LT		
	Near	Far	Covariate	Near	Far	Covariate
Uniform Feature Mean (Colombo et al., 2022)	57.21	24.37	68.73	58.22	24.00	68.09
Weighted Feature Mean (Wei et al., 2025)	57.46	25.73	70.66	58.88	25.54	70.26
Power Mean (Rücklé et al., 2018)	63.82	29.16	73.57	64.98	29.02	73.24
Concatenation (LSF-C w/o shrinkage)	55.83	14.33	53.88	56.15	13.67	51.92
Concatenation (LSF-B w/o shrinkage)	**55.61**	**13.85**	**53.69**	**56.13**	**13.25**	**51.73**

Comparing Layer Aggregation Strategies. We compare fusion rules using the same H_l/D_l criterion for layer selection, per-layer ℓ_2 normalization, and empirical covariance without shrinkage. In Table 6, concatenation with either full covariance (LSF-C) or block-diagonal covariance (LSF-B) outperforms the tested additive fusion rules across all reported shift groups under both calibration settings. With ImageNet-1K calibration, LSF-C improves over uniform feature mean, the strongest additive fusion baseline, by 1.38 pp, 10.04 pp, and 14.85 pp on near-OOD, far-OOD, and covariate shifts, respectively; the corresponding improvements achieved by LSF-B are 1.60 pp, 10.52 pp, and 15.04 pp. Under empirical covariance without shrinkage, LSF-B achieves the lowest FPR95 in all six comparisons, improving over LSF-C by 0.02–0.48 pp. These results empirically support retaining the two representations as distinct coordinates under both covariance formulations.

Additional Results. Appendix C.4.1 shows that regularized block-diagonal covariance improves upon empirical covariance by 0.21 pp and 0.22 pp with ImageNet-1K and ImageNet-LT calibration, respectively. It regularizes the joint precision analytically without requiring any OOD data. Additionally, Appendix C.4.2 empirically validates the Gaussianity assumption underlying Mahalanobis distance and shows that LSF improves Gaussianity over the existing single- and multi-layer Mahalanobis baselines. Moreover, Appendix C.5 shows that, during online inference, LSF incurs less overhead than X-Maha w/o FT while introducing only marginal additional overhead over Mahalanobis++ (+3 KB memory and +0.0011 ms latency) and substantially reducing FPR95. Finally, statistical tests reported in C.6 show that the standard deviation of LSF ranges from 0.21 pp to 0.70 pp, while its improvements over X-Maha w/o FT range from 0.53 pp to 22.17 pp, indicating that the observed gains generally exceed the variability across multiple evaluations.

5 Conclusion

We introduced LSF, a post-hoc OOD detector that leverages the penultimate representation alongside a single intermediate companion layer, selected via ID covariance entropy density drops. By concatenating the normalized pair, our approach preserves both feature spaces and naturally accommodates unequal layer widths without requiring dimension-alignment projections. A single shrinkage-regularized Mahalanobis model then captures their joint geometry. Extensive experiments across Transformer and convolutional backbones—under both balanced and long-tailed ID calibration—demonstrate the effectiveness of this design for semantic OOD and covariate-shift detection. These findings reveal that selectively pairing and jointly modeling complementary layers significantly improves detection performance without the need for auxiliary OOD data, backbone fine-tuning, or all-layer aggregation.

A limitation of LSF is its reliance on pretrained models that feature complementary information from intermediate layers, as detailed in Appendix F. Overcoming this architectural dependency remains an open challenge for future work.

AI USE STATEMENT

Generative AI tools were used for language editing, and LaTeX restructuring. All scientific claims, equations, experimental results, statistical tests, and bibliographic entries were checked by the authors against the experiment artifacts and cited sources. No AI system selected experimental outcomes or replaced human verification of the reported numbers.

REFERENCES

Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. NECO: Neural collapse based out-of-distribution detection. In *Proceedings of the International Conference on Learning Representations*, 2024.

Eunsu Baek, Keondo Park, Jiyoon Kim, and Hyung-Sin Kim. Unexplored faces of robustness and out-of-distribution: Covariate shifts in environment and sensor domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22294–22303, June 2024.

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? Fixing ImageNet out-of-distribution detection evaluation. In *Proceedings of the International Conference on Machine Learning*, pages 2471–2506, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

Pierre Colombo, Eduardo Dadalto, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. Beyond mahalanobis distance for textual OOD detection. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 17744–17759. Curran Associates, Inc., 2022. doi: 10.52202/068431-1290.

Kun Fang, Qinghua Tao, Kexin Lv, Mingzhen He, Xiaolin Huang, and JIE YANG. Kernel pca for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 37, pages 134317–134344, 2024.

Jarrod Haas, William Yolland, and Bernhard T. Rabus. Exploring simple, high quality out-of-distribution detection with l2 normalization. *Transactions on Machine Learning Research*, 2024.

Md Yousuf Harun, Jhair Gallardo, and Christopher Kanan. Controlling neural collapse enhances out-of-distribution detection and transfer learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 22149–22176, 2025.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the International Conference on Learning Representations*, 2017.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, 2021.

Denis Janiak, Jakub Binkowski, and Tomasz Kajdanowicz. A geometry-based view of mahalanobis OOD detection. In *Forty-third International Conference on Machine Learning*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR, 09–15 Jun 2019.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

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

Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15313–15323, 2021.

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

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2537–2546, 2019.

Haodong Lu, Dong Gong, Shuo Wang, Jason Xue, Lina Yao, and Kristen Moore. Learning with mixture of prototypes for out-of-distribution detection. In *International Conference on Learning Representations*, 2024.

Shuo Lu, Yingsheng Wang, Lijun Sheng, Lingxiao He, Aihua Zheng, and Jian Liang. Out-of-distribution detection: A task-oriented survey of recent advances. *ACM Computing Surveys*, 58(2): 1–39, 2025.

Maximilian Müller and Matthias Hein. Mahalanobis++: Improving OOD detection via feature normalization. In *Proceedings of the International Conference on Machine Learning*, 2025.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

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 (ICCV)*, pages 1686–1695, 2023.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, 2019.

Sudarshan Regmi. AdaSCALE: Adaptive scaling for OOD detection. In *Forty-third International Conference on Machine Learning*, 2026.

Sudarshan Regmi, Bibek Panthi, Sakar Dotel, Prashnna K Gyawali, Danail Stoyanov, and Binod Bhattarai. T2fnorm: Train-time feature normalization for OOD detection in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 153–162, 2024.

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.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv*, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, pages 144–157, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the International Conference on Machine Learning*, pages 20827–20840, 2022.

Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *Proceedings of the IEEE Information Theory Workshop*, 2015.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8769–8778, 2018. doi: 10.1109/cvpr.2018.00914.

Tong Wei, Bo-Lin Wang, Jiang-Xin Shi, Yu-Feng Li, and Min-Ling Zhang. X-Mahalanobis: Transformer feature mixing for reliable OOD detection. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2025.

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, 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.

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

A THEORETICAL JUSTIFICATION OF LSF

Proposition 1. [Layer selection by entropy-density drop] Let the entropy density of layer l be

$$\rho_l = \frac{H_l}{D_l}, \tag{A.1}$$

where H_l is the spectral entropy and D_l is the feature dimension. With effective rank defined as

$$\mathrm{ER}_l = e^{H_l}, \tag{A.2}$$

the entropy-density drop between successive layers is

$$\Delta_l = \rho_{l-1} - \rho_l = \frac{\log \mathrm{ER}_{l-1}}{D_{l-1}} - \frac{\log \mathrm{ER}_l}{D_l}. \tag{A.3}$$

Thus, Δ_l measures the difference in log effective rank per feature dimension between two consecutive layers. In particular, for equal-dimensional layers,

$$\Delta_l = \frac{1}{D} \log \frac{\mathrm{ER}_{l-1}}{\mathrm{ER}_l}. \tag{A.4}$$

Proof. By the definition of spectral entropy,

$$H_l = \log \mathrm{ER}_l \,.$$

Substituting this identity into $\rho_l = H_l/D_l$ gives

$$\rho_l = \frac{\log \mathrm{ER}_l}{D_l}.$$

Taking the difference between two successive layers yields Eq. (A.3). When $D_{l-1} = D_l = D$, the expression reduces to

$$\Delta_l = \frac{1}{D} \left(\log \mathrm{ER}_{l-1} - \log \mathrm{ER}_l \right) = \frac{1}{D} \log \frac{\mathrm{ER}_{l-1}}{\mathrm{ER}_l}.$$

Therefore, the entropy-density drop measures a reduction in log effective rank after accounting for feature dimensionality. □

The criterion is not invariant to changes in feature dimension. In particular, for an isotropic covariance, $H_l = \log D_l$, giving

$$\frac{H_l}{D_l} = \frac{\log D_l}{D_l},$$

which varies with feature width. We therefore interpret Δ_l as a dimension-normalized, ID-only heuristic for identifying candidate intermediate layers, with its effectiveness evaluated empirically against alternative layer-selection criteria.

Setup for Mahalanobis distance. Let $\boldsymbol{\phi}(x)_L$ and $\boldsymbol{\phi}(x)_{l^\star}$ denote the feature representations at the penultimate layer L and the selected intermediate layer $l^\star$, respectively. Let $\boldsymbol{\mu}_{L,y}$ and $\boldsymbol{\mu}_{l^\star,y}$ denote their corresponding ID class means. We define the within-class residuals

$$e_1 = \boldsymbol{\phi}(x)_L - \hat{\boldsymbol{\mu}}_{L,y}, \qquad e_2 = \boldsymbol{\phi}(x)_{l^\star} - \hat{\boldsymbol{\mu}}_{l^\star,y}, \tag{A.5}$$

and the concatenated residual

$$e = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}. \tag{A.6}$$

The joint covariance estimated from the concatenated representation is

$$\hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{bmatrix}, \tag{A.7}$$

where $\hat{\Sigma}_{11}$ and $\hat{\Sigma}_{22}$ are the within-layer covariance matrices and $\hat{\Sigma}_{12} = \hat{\Sigma}_{21}^\top$ captures cross-layer covariance. We use Ledoit–Wolf shrinkage to obtain positive-definite covariance estimates. Hence, $\hat{\Sigma} \succ 0$ and the diagonal blocks $\hat{\Sigma}_{11}$ and $\hat{\Sigma}_{22}$ are positive definite. For the full-covariance formulation, the corresponding Schur complement

$$S_2 = \hat{\Sigma}_{22} - \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12} \tag{A.8}$$

is also positive definite.

Proposition 2. [Joint Mahalanobis decomposition] The joint Mahalanobis distance admits two equivalent decompositions depending on the covariance structure. Under the full-covariance formulation,

$$e^\top \hat{\Sigma}^{-1} e = e_1^\top \hat{\Sigma}_{11}^{-1} e_1 + r_2^\top S_2^{-1} r_2, \tag{A.9}$$

where

$$r_2 = e_2 - \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1} e_1, \qquad S_2 = \hat{\Sigma}_{22} - \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12}. \tag{A.10}$$

Under the block-diagonal formulation,

$$\hat{\Sigma}_{\mathrm{BD}} = \begin{bmatrix} \hat{\Sigma}_{11} & 0 \\ 0 & \hat{\Sigma}_{22} \end{bmatrix}, \tag{A.11}$$

and therefore

$$e^\top \hat{\Sigma}_{\mathrm{BD}}^{-1} e = e_1^\top \hat{\Sigma}_{11}^{-1} e_1 + e_2^\top \hat{\Sigma}_{22}^{-1} e_2. \tag{A.12}$$

Thus, both formulations retain the penultimate-layer Mahalanobis contribution and add a nonnegative contribution from the intermediate representation. The full-covariance formulation measures the intermediate contribution conditionally on the penultimate representation, whereas the block-diagonal formulation uses its marginal Mahalanobis contribution.

Proof. For the full-covariance case, since $\hat{\Sigma}_{11} \succ 0$ and $S_2 \succ 0$, the block inverse identity gives

$$\hat{\Sigma}^{-1} = \begin{bmatrix} \hat{\Sigma}_{11}^{-1} + \hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12} S_2^{-1} \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1} & -\hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12} S_2^{-1} \\ -S_2^{-1}\hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1} & S_2^{-1} \end{bmatrix}. \tag{A.13}$$

Substituting this expression into $e^\top \hat{\Sigma}^{-1} e$ and collecting terms gives

$$e^\top \hat{\Sigma}^{-1} e = e_1^\top \hat{\Sigma}_{11}^{-1} e_1 + \left(e_2 - \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1} e_1\right)^\top S_2^{-1} \left(e_2 - \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1} e_1\right),$$

which is Eq. (A.9). Since $S_2^{-1} \succ 0$, the second term is nonnegative.

For the block-diagonal case, the inverse of Eq. (A.11) is

$$\hat{\Sigma}_{\mathrm{BD}}^{-1} = \begin{bmatrix} \hat{\Sigma}_{11}^{-1} & 0 \\ 0 & \hat{\Sigma}_{22}^{-1} \end{bmatrix}.$$

Direct multiplication then yields Eq. (A.12). Because both $\hat{\Sigma}_{11}^{-1}$ and $\hat{\Sigma}_{22}^{-1}$ are positive definite, the intermediate-layer contribution is nonnegative. □

The two covariance formulations therefore differ only in how they model the relationship between the two representations. The block-diagonal formulation models their covariance structure separately and adds the marginal Mahalanobis contribution of the intermediate layer. The full-covariance formulation additionally models cross-layer dependence and evaluates the intermediate residual after accounting for its linear dependence on the penultimate residual. In both cases, the penultimate-layer contribution is preserved and the intermediate representation contributes a nonnegative additional term.

OOD mean-shift analysis. We next examine how the two formulations respond to a distribution shift. Let the OOD residual be written as

$$e^{\mathrm{OOD}} = e + \delta, \qquad \delta = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}, \tag{A.14}$$

where e is the centered ID residual and δ_1 and δ_2 denote the OOD mean shifts of the penultimate and intermediate representations, respectively, relative to their corresponding ID class means. We assume

$$\mathbb{E}[e] = 0, \tag{A.15}$$

and interpret δ as a fixed OOD mean shift. More generally, the analysis also holds for a random δ that is independent of e.

Proposition 3. [Non-degradation and increased sensitivity to intermediate-layer shifts] Under the above mean-shift assumption, define the expected increase in Mahalanobis distance as the expected OOD score minus the expected ID score. For the full-covariance formulation,

$$\Delta_{\text{full}} = \mathbb{E}\Big[\delta_1^\top \hat{\Sigma}_{11}^{-1}\delta_1\Big] + \mathbb{E}\big[\zeta_2^\top S_2^{-1}\zeta_2\big]\,, \tag{A.16}$$

where

$$\zeta_2 = \delta_2 - \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\delta_1. \tag{A.17}$$

For the block-diagonal formulation,

$$\Delta_{\text{BD}} = \mathbb{E}\Big[\delta_1^\top \hat{\Sigma}_{11}^{-1}\delta_1\Big] + \mathbb{E}\Big[\delta_2^\top \hat{\Sigma}_{22}^{-1}\delta_2\Big]\,. \tag{A.18}$$

Consequently,

$$\Delta_{\text{full}} \geq \mathbb{E}\Big[\delta_1^\top \hat{\Sigma}_{11}^{-1}\delta_1\Big]\,, \qquad \Delta_{\text{BD}} \geq \mathbb{E}\Big[\delta_1^\top \hat{\Sigma}_{11}^{-1}\delta_1\Big]\,. \tag{A.19}$$

The inequalities are strict whenever the corresponding intermediate-layer quadratic contribution is positive with nonzero probability.

Proof. For the full-covariance formulation, applying Proposition 2 to $e^{\text{OOD}} = e + \delta$ gives

$$(e^{\text{OOD}})^\top \hat{\Sigma}^{-1} e^{\text{OOD}} = (e_1+\delta_1)^\top \hat{\Sigma}_{11}^{-1}(e_1+\delta_1) + (r_2+\zeta_2)^\top S_2^{-1}(r_2+\zeta_2), \tag{A.20}$$

where

$$r_2 = e_2 - \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}e_1$$

and ζ_2 is given by Eq. (A.17). Expanding both quadratic forms and subtracting the corresponding ID score gives terms that are linear in e and terms quadratic in δ. Because $\mathbb{E}[e] = 0$, the expected linear terms vanish. Therefore,

$$\Delta_{\text{full}} = \mathbb{E}\Big[\delta_1^\top \hat{\Sigma}_{11}^{-1}\delta_1\Big] + \mathbb{E}\big[\zeta_2^\top S_2^{-1}\zeta_2\big]\,.$$

The second term is nonnegative because $S_2^{-1} \succ 0$, proving Eq. (A.16) and the first inequality in Eq. (A.19).

For the block-diagonal formulation,

$$(e^{\text{OOD}})^\top \hat{\Sigma}_{\text{BD}}^{-1} e^{\text{OOD}} = (e_1+\delta_1)^\top \hat{\Sigma}_{11}^{-1}(e_1+\delta_1) + (e_2+\delta_2)^\top \hat{\Sigma}_{22}^{-1}(e_2+\delta_2). \tag{A.21}$$

Subtracting the ID score and taking expectations eliminates the linear terms because $\mathbb{E}[e_1] = \mathbb{E}[e_2] = 0$, yielding Eq. (A.18). Since $\hat{\Sigma}_{22}^{-1} \succ 0$, the intermediate-layer term is nonnegative, proving the second inequality in Eq. (A.19). Strictness follows whenever the corresponding intermediate-layer quadratic term is positive with nonzero probability. □

Proposition 3 shows that adding the intermediate representation cannot reduce the expected quadratic sensitivity to a mean shift in the penultimate representation under either covariance formulation. The penultimate representation therefore acts as an anchor, while the intermediate representation provides an additional source of shift sensitivity. The nature of this additional contribution differs between the two formulations: the block-diagonal covariance captures the intermediate shift directly through its marginal Mahalanobis distance, whereas the full covariance captures the component of the intermediate shift that remains after accounting for its linear dependence on the penultimate shift. The two formulations therefore capture intermediate-layer shifts differently: the full covariance measures the component of the intermediate shift not explained by the penultimate shift, whereas the block-diagonal covariance measures the intermediate shift in its marginal feature space.

Table 7: AUROC (↑, %) using ViT-B/16 under ImageNet-1K ID calibration.

Method	Near OOD		Far OOD			Covariate-Shifted ID				Avg.
	SSB	NIN	iNat	TxT	OpO	Im-C	Im-R	Im-V2	Im-ES	
MSP	74.25	82.52	92.57	82.72	89.22	71.07	77.54	58.49	69.19	77.51
Energy	76.55	82.69	91.59	83.53	88.83	71.72	72.25	56.45	66.90	76.73
ReAct	77.99	85.11	95.54	85.53	92.50	72.98	77.61	57.85	68.97	79.34
SCALE	64.75	61.59	59.41	61.02	60.53	56.19	37.06	49.34	54.30	56.02
AdaSCALE	75.61	83.63	94.18	80.04	91.66	70.17	79.58	58.48	72.50	78.43
KNN	68.92	84.07	96.57	89.56	94.05	70.89	83.94	56.59	70.90	79.50
NNGuide	74.31	87.57	97.54	89.81	95.33	74.11	85.79	58.57	73.41	81.83
KPCA	72.58	84.64	91.66	85.63	89.73	74.15	84.31	58.75	72.46	79.32
Maha	71.53	85.34	96.98	85.43	93.64	71.00	86.30	57.72	69.75	79.75
RS-Maha	77.18	90.07	98.76	89.26	95.86	74.93	87.20	58.78	73.80	82.87
rMaha	75.02	87.90	97.61	86.89	94.76	72.36	85.68	**59.17**	72.13	81.28
Maha++	77.18	90.07	98.76	89.26	95.86	74.93	87.20	58.78	73.80	82.87
rMaha++	77.26	90.13	98.77	89.31	95.86	74.98	87.21	58.80	73.88	82.91
X-Maha w/o FT	76.79	89.95	**98.78**	89.78	96.04	75.53	87.61	58.62	74.11	83.03
LSF-C	**78.75**	**91.08**	98.47	96.74	**96.34**	88.24	89.14	58.87	84.39	86.89
LSF-B	78.71	91.04	98.11	**97.28**	96.04	**89.82**	**89.40**	58.58	**85.59**	**87.17**

Table 8: AUROC (↑, %) using ViT-B/16 under ImageNet-LT ID calibration.

Method	Near OOD		Far OOD			Covariate-Shifted ID				Avg.
	SSB	NIN	iNat	TxT	OpO	Im-C	Im-R	Im-V2	Im-ES	
MSP	74.37	82.57	92.57	82.77	89.23	71.21	77.63	58.73	69.35	77.60
Energy	77.19	83.19	91.88	83.99	89.16	72.34	72.87	57.29	67.54	77.27
ReAct	**78.56**	85.55	95.71	85.92	92.76	73.55	78.19	58.65	69.57	79.83
SCALE	65.01	61.84	59.64	61.28	60.74	56.42	37.22	49.56	54.50	56.25
AdaSCALE	76.61	84.36	94.44	80.77	92.02	71.00	80.33	59.81	73.40	79.20
KNN	38.16	40.83	62.25	56.40	50.34	57.73	65.61	49.80	55.88	53.00
NNGuide	53.22	63.98	81.98	77.86	77.97	64.62	75.28	52.44	65.89	68.14
KPCA	69.17	81.11	88.61	83.06	86.69	71.57	81.26	58.16	70.55	76.69
Maha	70.63	84.99	97.01	85.29	93.54	70.94	86.23	58.29	69.53	79.61
RS-MD	76.09	89.69	98.89	89.12	95.81	74.86	87.22	59.52	73.51	82.74
rMaha	74.50	87.77	97.69	86.79	94.62	71.73	85.02	59.18	71.30	80.96
Maha++	76.09	89.69	98.89	89.12	95.81	74.86	87.22	59.52	73.51	82.74
rMaha++	76.17	89.75	98.89	89.16	95.82	74.91	87.22	59.55	73.58	82.78
X-Maha w/o FT	75.68	89.60	98.90	89.62	95.97	75.49	87.62	59.55	73.82	82.92
LSF-C	77.26	90.84	**98.97**	94.93	**96.43**	84.35	88.84	60.78	81.70	86.01
LSF-B	77.38	**91.18**	98.42	**97.20**	96.01	**89.89**	**89.77**	**61.75**	**86.45**	**87.56**

B PER-LAYER OOD SEPARATION

The per-layer Mahalanobis distance can be written as:

$$\mathrm{MD}_{c,l}(x) = (\tilde{h}_l(x) - \hat{\mu}_{c,l})^\top \hat{\Sigma}_l^{-1} (\tilde{h}_l(x) - \hat{\mu}_{c,l}). \tag{B.1}$$

Following (Janiak et al., 2026), we define the per-layer OOD separation as the difference between the mean Mahalanobis energy of all the OOD and ID samples:

$$S_l \triangleq \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{OOD}}}[\mathrm{MD}_{c,l}(x)] - \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{ID}}}[\mathrm{MD}_{c,l}(x)]. \tag{B.2}$$

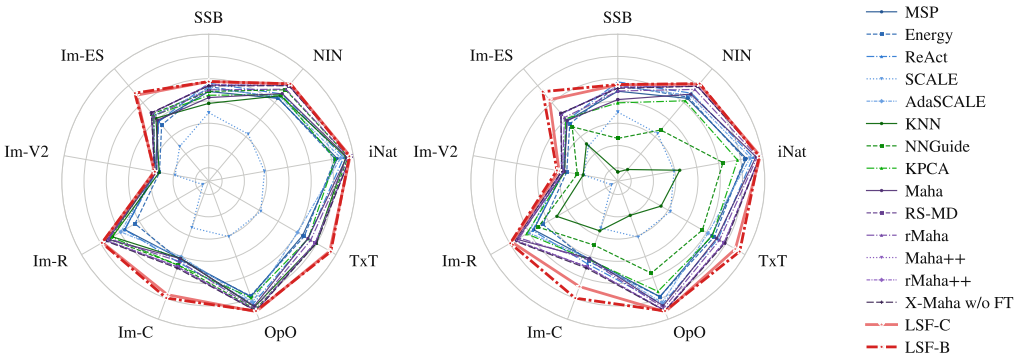


Figure 3: AUROC (%) across OpenOOD benchmark for ViT-B/16 with ImageNet-1K (left) and ImageNet-LT (right) ID calibration.

C Additional Results

C.1 AUROC Comparison

Results under ImageNet-1K ID calibration. Table 7 and the left panel of Figure 3 shows that LSF-B achieves the highest mean AUROC of 87.17%, followed by LSF-C at 86.89%, improving over X-Maha w/o FT, the strongest baseline by mean AUROC, by 4.14 pp and 3.86 pp, respectively. Together, the two variants achieve the best AUROC on seven of the nine datasets: LSF-C leads on SSB-Hard, NINCO, and OpenImage-O, while LSF-B leads on Textures, ImageNet-C, ImageNet-R, and ImageNet-ES. On SSB-Hard, LSF-C achieves the highest AUROC, although ReAct obtains a slightly lower FPR95 in Table 1. Both variants remain competitive on iNaturalist, while all evaluated methods obtain comparatively low AUROC on ImageNet-V2. This difficulty is consistent with ImageNet-V2's design as a close replication of ImageNet-1k dataset, making it challenging to distinguish from the clean ID distribution.

Results under ImageNet-LT ID calibration. Table 8 and the right panel of Figure 3 shows that LSF-B again achieves the highest mean AUROC of 87.56%, followed by LSF-C at 86.01%, exceeding X-Maha w/o FT by 4.64 pp and 3.09 pp, respectively. Together, the two variants achieve the best AUROC on eight of the nine datasets, with LSF-B leading on six and LSF-C on two; ReAct performs best on SSB-Hard. Relative to ImageNet-1K calibration, the mean AUROC of LSF-B increases by 0.39 pp, while that of LSF-C decreases by 0.88 pp. Both variants therefore maintain strong aggregate performance under long-tailed calibration, in contrast to the substantial declines observed for KNN and NNGuide.

These AUROC results complement the FPR95 findings, supporting the benefit of combining selected intermediate features with the penultimate representation. The improvements are particularly pronounced on Textures and several covariate-shift benchmarks, although neither variant dominates every dataset. Figure 4 further illustrates the behavior of LSF-C relative to Mahalanobis++ under ImageNet-LT calibration, showing how changes in the ID and OOD score distributions relate to the observed detection gains.

C.2 Results Across Backbones

We evaluate LSF on CLIP ViT-B/16, ConvNeXt-T, and Swin-T under ImageNet-LT ID calibration. These experiments assess whether the selected intermediate and penultimate representations and fusion remain effective across different backbones and feature hierarchies. LSF achieves the lowest mean FPR95 for each backbone, although the gains vary across shift types.

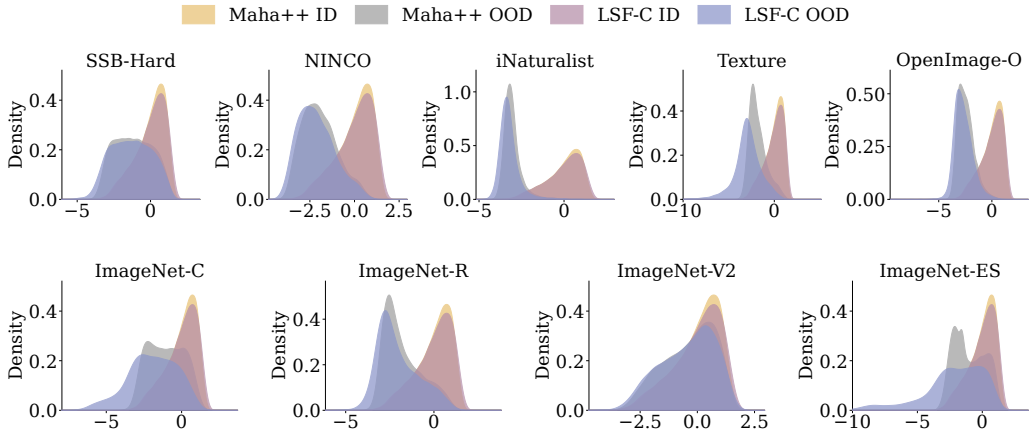


Figure 4: Score distributions for ViT-B/16 under ImageNet-LT ID calibration on semantic-shift (top) and covariate-shift (bottom) datasets.

Table 9: FPR95 (↓, %) using CLIP ViT-B/16 under ImageNet-LT ID calibration.

Method	Near OOD		Far OOD			Covariate-Shifted ID				Avg.
	SSB	NIN	iNat	TxT	OpO	Im-C	Im-R	Im-V2	Im-ES	
MSP	76.32	62.54	36.75	55.16	46.17	74.50	67.26	89.87	71.14	64.41
Energy	75.07	65.81	51.67	59.38	49.66	79.14	62.34	90.64	75.04	67.64
ReAct	74.48	64.74	47.98	57.96	46.32	78.74	60.66	90.33	74.93	66.24
SCALE	72.98	59.48	38.44	50.87	40.53	75.07	58.65	89.50	70.34	61.76
AdaSCALE	68.39	57.88	48.65	55.92	34.09	72.28	58.44	89.04	65.08	61.09
KNN	85.99	77.15	45.56	59.98	50.49	84.10	58.69	90.20	86.88	71.01
NNGuide	82.07	69.02	33.12	43.10	35.98	76.56	54.93	89.82	75.36	62.22
KPCA	85.48	77.21	64.73	70.21	66.20	84.04	76.00	91.33	84.30	77.72
Maha	78.39	62.18	16.01	43.69	26.74	75.95	53.84	88.84	79.66	58.37
RS-Maha	73.65	56.63	13.06	37.02	23.23	66.92	52.51	88.87	65.09	53.00
rMaha	77.02	59.10	17.11	45.05	28.23	73.30	57.70	88.63	76.01	58.02
Maha++	74.74	57.40	13.37	37.85	23.07	68.11	52.45	88.72	66.75	53.61
rMaha++	74.71	57.32	13.36	37.84	23.04	68.10	52.43	88.73	66.71	53.58
X-Maha w/o FT	76.29	59.15	17.86	37.82	26.40	69.43	50.16	88.53	71.17	55.20
LSF-C	33.82	41.71	11.29	14.84	12.17	17.36	37.35	84.22	50.09	33.65
LSF-B	**26.17**	**40.91**	**11.14**	**12.62**	**10.77**	**10.97**	**35.67**	**83.71**	**47.99**	**31.10**

C.2.1 CLIP ViT-B16 Results

Tables 9 and 10 report results for the LAION-2B-pretrained, CLIP ViT-B/16 checkpoint under ImageNet-LT calibration. LSF-B achieves the lowest mean FPR95 of 31.10%, followed by LSF-C at 33.65%, improving over RS-Maha, the strongest baseline by mean FPR95, by 21.90 pp and 19.35 pp, respectively. LSF-B obtains the lowest FPR95 on all nine datasets, with LSF-C ranking second throughout. Relative to the strongest baseline on each dataset, LSF-B reduces FPR95 by 55.95 pp on ImageNet-C, 42.22 pp on SSB-Hard, and 24.40 pp on Textures; the corresponding reductions for LSF-C are 49.56 pp, 34.57 pp, and 22.18 pp.

The AUROC results reinforce these findings. LSF-B achieves the highest mean AUROC of 91.00%, followed by LSF-C at 90.32%, exceeding X-Maha w/o FT, the strongest baseline by mean AUROC, by 8.80 pp and 8.12 pp, respectively. LSF-B leads on eight datasets, while LSF-C performs best on iNaturalist. Both variants outperform all evaluated baselines on every dataset in both metrics, demonstrating gains across near-OOD, far-OOD, and covariate-shift detection. These results provide

Table 10: AUROC (↑, %) using CLIP ViT-B/16 under ImageNet-LT ID calibration.

Method	Near OOD		Far OOD			Covariate-Shifted ID				Avg.
	SSB	NIN	iNat	TxT	OpO	Im-C	Im-R	Im-V2	Im-ES	
MSP	72.77	80.98	89.97	81.48	87.66	68.20	73.90	56.87	68.02	75.54
Energy	66.74	71.49	80.76	73.50	81.39	60.72	70.22	53.76	63.15	69.08
ReAct	68.00	73.68	83.37	75.21	83.88	61.64	72.00	54.32	64.00	70.68
SCALE	72.62	80.63	88.55	82.20	88.21	66.26	77.32	56.26	68.54	75.62
AdaSCALE	79.31	85.36	88.06	82.03	93.54	77.08	81.77	58.29	77.98	80.38
KNN	59.85	72.94	88.76	84.02	88.05	63.31	79.26	55.70	63.87	72.86
NNGuide	60.12	75.15	89.46	85.20	90.23	65.01	78.10	54.83	68.53	74.07
KPCA	71.46	83.24	89.74	85.42	89.34	71.63	80.19	58.25	69.81	77.67
Maha	74.10	86.88	96.79	89.48	95.04	72.83	84.09	58.91	69.90	80.89
RS-Maha	75.61	88.23	97.24	90.86	95.75	75.22	84.50	59.22	72.92	82.17
rMaha	75.02	87.64	96.55	88.64	94.60	73.14	82.38	58.96	69.73	80.74
Maha++	75.29	87.99	97.23	90.65	95.71	74.74	84.47	59.17	72.33	81.95
rMaha++	75.30	87.99	97.22	90.65	95.70	74.74	84.46	59.17	72.34	81.95
X-Maha w/o FT	75.28	87.32	96.29	90.90	95.11	76.22	86.35	59.68	72.62	82.20
LSF-C	93.70	91.21	**97.49**	96.78	97.59	97.26	90.32	64.65	83.89	90.32
LSF-B	**95.38**	**91.57**	97.41	**97.34**	**97.76**	**98.32**	**90.92**	**65.36**	**84.98**	**91.00**

Table 11: FPR95 (↓, %) using ConvNeXt-T under ImageNet-LT ID calibration.

Method	Near OOD		Far OOD			Covariate-Shifted ID				Avg.
	SSB	NIN	iNat	TxT	OpO	Im-C	Im-R	Im-V2	Im-ES	
MSP	79.09	65.43	43.78	61.54	61.63	60.82	67.02	88.49	65.06	65.87
Energy	87.74	87.11	84.88	85.35	92.84	80.46	91.89	89.73	76.08	86.23
ReAct	87.49	85.23	81.77	82.34	90.72	77.46	89.88	89.37	75.12	84.38
SCALE	93.88	96.78	95.26	97.73	97.24	93.91	97.79	94.15	88.41	95.02
AdaSCALE	80.82	68.22	52.74	64.73	64.43	62.25	66.15	**88.24**	**62.33**	67.77
KNN	97.68	99.07	99.46	91.88	97.76	84.87	91.85	96.35	94.43	94.82
NNGuide	95.91	96.50	95.03	83.81	94.44	82.00	87.10	95.05	91.72	91.28
KPCA	92.18	88.51	85.13	84.15	85.24	86.05	83.16	93.16	88.06	87.29
Maha	87.70	74.49	45.41	61.91	48.96	74.86	51.91	91.22	82.78	68.81
RS-Maha	71.88	56.49	13.83	49.04	39.51	64.26	52.26	90.32	78.40	57.33
rMaha	82.83	66.01	28.40	58.30	42.72	69.75	52.63	90.43	79.57	63.40
Maha++	71.88	56.49	13.83	49.04	39.51	64.26	52.26	90.32	78.40	57.33
rMaha++	**71.87**	56.45	13.79	49.10	39.47	64.32	52.38	90.32	78.48	57.35
X-Maha w/o FT	–	–	–	–	–	–	–	–	–	–
LSF-C	72.79	55.61	**13.39**	29.41	**37.68**	56.36	49.11	90.33	69.35	52.67
LSF-B	74.98	**54.23**	13.94	**22.43**	39.45	**51.08**	**48.39**	90.42	64.42	**51.04**

additional evidence that the benefits of the proposed layer selection and fusion extend to a distinct pretraining and fine-tuning configuration.

C.2.2 ConvNeXt-T Results

Tables 11 and 12 report results for ConvNeXt-T under ImageNet-LT calibration. LSF-B achieves the lowest mean FPR95 of 51.04%, followed by LSF-C at 52.67%, improving over RS-Maha and Maha++, the strongest baselines by mean FPR95, by 6.29 pp and 4.66 pp, respectively. Similarly, LSF-B achieves the highest mean AUROC of 84.42%, followed by LSF-C at 83.62%, exceeding rMaha++, the strongest baseline by mean AUROC, by 2.19 pp and 1.39 pp, respectively.

Together, the two variants achieve the lowest FPR95 on six of the nine datasets, including all three far-OOD benchmarks. LSF-B leads on NINCO, Textures, ImageNet-C, and ImageNet-R, while LSF-C leads on iNaturalist and OpenImage-O. The exceptions are SSB-Hard, where rMaha++ performs best,

Table 12: AUROC (↑, %) using ConvNeXt-T under ImageNet-LT ID calibration.

Method	Near OOD		Far OOD			Covariate-Shifted ID				Avg.
	SSB	NIN	iNat	TxT	OpO	Im-C	Im-R	Im-V2	Im-ES	
MSP	73.08	80.68	90.23	79.58	82.83	78.65	77.58	61.53	74.36	77.61
Energy	64.68	63.71	69.70	63.40	57.11	70.11	57.71	58.86	64.98	63.36
ReAct	66.30	68.02	75.22	68.46	64.45	73.33	64.86	60.06	67.80	67.61
SCALE	59.94	59.90	71.32	54.46	60.75	62.13	53.79	54.87	62.55	59.97
AdaSCALE	73.30	82.00	89.00	78.82	86.26	77.93	83.11	**62.29**	80.77	79.28
KNN	35.73	34.64	34.31	64.09	44.81	59.54	65.56	46.21	52.80	48.63
NNGuide	49.16	54.16	64.43	74.76	63.05	69.95	72.25	52.31	63.63	62.63
KPCA	62.96	72.80	78.33	77.82	78.21	73.47	77.79	57.13	70.58	72.12
Maha	68.04	82.11	92.87	87.19	91.57	75.59	87.03	57.61	70.13	79.13
RS-Maha	73.81	86.62	96.80	89.43	93.12	79.27	87.32	59.29	74.33	82.22
rMaha	72.47	85.43	95.23	87.72	92.61	77.74	85.55	59.40	73.51	81.07
Maha++	73.81	86.62	96.80	89.43	93.12	79.27	87.32	59.29	74.33	82.22
rMaha++	**73.86**	86.68	96.82	89.40	93.14	79.24	87.29	59.32	74.35	82.23
X-Maha w/o FT	–	–	–	–	–	–	–	–	–	–
LSF-C	73.38	86.97	**96.83**	93.70	**93.40**	82.52	88.08	59.70	77.98	83.62
LSF-B	72.63	**87.42**	96.69	**95.46**	93.13	**84.62**	**88.58**	60.35	**80.94**	**84.42**

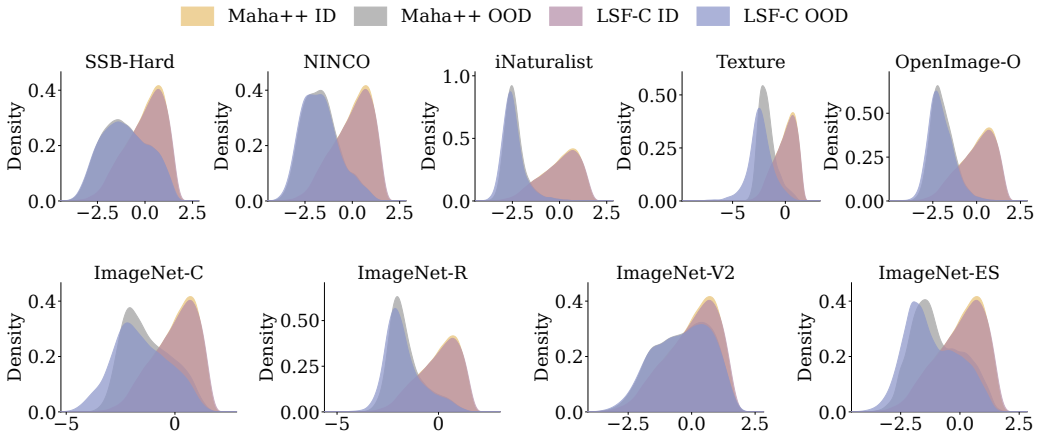


Figure 5: Score distributions for ConvNeXt-T with ImageNet-LT calibration on semantic-shift (top) and covariate-shift (bottom) datasets.

and ImageNet-V2 and ImageNet-ES, where AdaSCALE achieves lower FPR95. For AUROC, the two variants jointly lead on seven datasets, with LSF-B additionally achieving the highest AUROC on ImageNet-ES. These results demonstrate that the gains extend to a convolutional backbone, with LSF-B providing stronger aggregate performance and LSF-C remaining advantageous on some far-OOD datasets.

X-Maha w/o FT is not reported for ConvNeXt-T because its additive feature-fusion formulation requires equal-dimensional representations. Unlike the equal-width representations in the evaluated ViT backbones, ConvNeXt-T produces features with different channel widths across stages, preventing direct addition without an additional dimension-alignment mechanism. The proposed concatenation accommodates these unequal widths directly, allowing both LSF-C and LSF-B to operate on this hierarchical architecture without modifying the pretrained backbone.

Additionally, Figure 5 compares the score distributions of LSF-C and Mahalanobis++. On datasets with pronounced gains, such as Textures, the OOD score distribution shifts further to the left relative to ID, improving separation.

Table 13: FPR95 (↓, %) using Swin-T under ImageNet-LT ID calibration.

Method	Near OOD		Far OOD			Covariate-Shifted ID				Avg.
	SSB	NIN	iNat	TxT	OpO	Im-C	Im-R	Im-V2	Im-ES	
MSP	83.28	70.52	52.35	66.42	60.21	68.36	63.92	90.11	64.88	68.89
Energy	82.59	72.42	68.65	68.69	67.98	65.33	66.59	90.18	59.48	71.32
ReAct	83.57	72.22	65.64	64.02	64.08	64.10	60.12	89.88	59.39	69.23
SCALE	83.52	72.01	63.07	66.26	62.09	65.33	60.00	89.88	61.80	69.33
AdaSCALE	84.02	70.49	58.77	65.51	59.01	67.15	58.71	**89.36**	59.82	68.09
KNN	96.95	95.30	91.19	65.09	89.41	74.15	65.05	93.93	82.33	83.71
NNGuide	94.25	88.92	78.10	58.19	78.93	70.11	58.27	92.62	72.94	76.93
KPCA	92.50	87.66	82.59	82.07	82.84	85.33	80.86	92.45	87.41	85.97
Maha	89.67	79.76	43.97	54.17	54.30	77.28	52.10	91.30	88.58	70.13
RS-Maha	73.92	57.79	11.35	41.13	**36.86**	58.98	45.72	89.92	63.90	53.29
rMaha	84.24	70.50	22.75	55.05	43.74	70.69	52.49	90.49	83.09	63.67
Maha++	73.92	57.79	11.35	41.13	**36.86**	58.98	45.72	89.92	63.90	53.29
rMaha++	**72.78**	**56.21**	**10.95**	40.94	37.03	57.97	46.65	89.95	62.67	52.79
X-Maha w/o FT	–	–	–	–	–	–	–	–	–	–
LSF-C	77.20	60.07	18.20	22.62	43.12	**43.56**	**42.91**	89.61	**41.09**	48.71
LSF-B	77.64	59.48	17.59	**22.13**	42.72	43.97	43.03	89.40	41.12	**48.57**

Table 14: AUROC (↑, %) on Swin-T under ImageNet-LT ID calibration.

Method	Near OOD		Far OOD			Covariate-Shifted ID				Avg.
	SSB	NIN	iNat	TxT	OpO	Im-C	Im-R	Im-V2	Im-ES	
MSP	**72.21**	80.30	88.54	80.68	85.68	76.19	80.14	59.86	76.17	77.75
Energy	69.93	77.10	81.29	75.96	79.08	75.27	76.95	59.77	76.74	74.68
ReAct	70.28	78.62	84.39	79.81	82.93	76.46	80.74	60.40	77.53	76.80
SCALE	71.33	80.04	86.97	80.66	85.30	76.70	81.99	60.47	77.25	77.86
AdaSCALE	70.69	80.78	87.41	78.95	87.16	73.99	84.11	59.78	80.01	78.10
KNN	37.69	43.16	49.34	76.86	54.11	66.16	77.77	50.56	61.98	57.51
NNGuide	51.57	62.28	73.41	83.11	75.68	73.02	82.63	54.18	70.50	69.60
KPCA	61.81	74.11	80.19	79.38	80.17	72.87	79.27	56.87	70.71	72.82
Maha	66.26	80.46	92.85	87.99	90.47	74.54	87.10	57.86	67.36	78.32
RS-Maha	71.33	85.87	96.86	90.43	93.14	79.60	88.32	59.13	76.63	82.37
rMaha	69.36	83.59	95.52	87.23	91.87	76.37	85.45	58.56	73.12	80.12
Maha++	71.33	85.87	96.86	90.43	93.14	79.60	88.32	59.13	76.63	82.37
rMaha++	71.84	**86.32**	**96.94**	90.56	**93.15**	80.14	88.29	59.33	77.38	82.66
X-Maha w/o FT	–	–	–	–	–	–	–	–	–	–
LSF-C	70.37	85.55	95.75	94.94	92.44	**86.26**	89.29	60.67	86.25	84.61
LSF-B	70.26	85.67	95.85	**95.02**	92.39	**86.26**	**89.36**	**60.74**	**86.32**	**84.65**

C.2.3 SWIN-T RESULTS

Tables 13 and 14 report results for Swin-T under ImageNet-LT calibration. LSF-B achieves the lowest mean FPR95 of 48.57%, followed by LSF-C at 48.71%, improving over rMaha++, the strongest baseline by mean FPR95, by 4.22 pp and 4.08 pp, respectively. Similarly, LSF-B achieves the highest mean AUROC of 84.65%, closely followed by LSF-C at 84.61%, exceeding rMaha++ by 1.99 pp and 1.95 pp, respectively.

The gains are concentrated on Textures and the covariate-shift benchmarks. LSF-B achieves the lowest FPR95 on Textures, while LSF-C leads on ImageNet-C, ImageNet-R, and ImageNet-ES. AdaSCALE retains the lowest FPR95 on ImageNet-V2, outperforming LSF-B by only 0.04 pp. For AUROC, LSF-B leads on Textures and all four covariate-shift benchmarks, tying with LSF-C on ImageNet-C. Single-layer Mahalanobis variants retain lower FPR95 on SSB-Hard, NINCO, iNaturalist, and OpenImage-O; rMaha++ also achieves the highest AUROC on the latter three datasets,

while MSP leads on SSB-Hard. Thus, the improvements are not uniform across shift types. The results are consistent with intermediate features providing complementary information for detecting appearance changes, yielding stronger aggregate performance despite reduced performance on some semantic OOD benchmarks.

As with ConvNeXt-T, X-Maha w/o FT is not reported for Swin-T because its additive feature-fusion formulation requires equal-dimensional representations. Swin-T uses patch merging between stages to reduce spatial resolution and increase channel width, preventing direct addition of features from different stages without dimension alignment. The proposed concatenation directly accommodates these unequal feature widths, allowing both LSF-C and LSF-B to be applied post hoc without modifying or adapting the pretrained backbone.

C.3 Extended Results

Table 15: FPR95 (↓, %) on OOD datasets averaged over Near OOD (SSB-Hard, NINCO) and Far OOD (iNaturalist, Textures, OpenImage-O) under ImageNet-LT ID calibration. Shorthands—EN: Energy; RE: ReAct; SC: SCALE; ASC: AdaSCALE; NNG: NNGuide; KPCA: KPCA; M: Mahalanobis; RM: Radially Mahalanobis; rM: Relative Mahalanobis; MP: Mahalanobis++; rMP: Relative Mahalanobis++; XM†: X-Maha w/o FT

Model	MSP	EN	RE	SC	ASC	KNN	NNG	KPCA	M	RM	rM	MP	rMP	XM†	LSF-C	LSF-B
ViT-B16-In21k-augreg	46.64	33.66	36.13	31.87	37.21	91.55	55.78	77.89	27.58	27.58	32.81	30.34	30.36	27.25	**26.74**	33.05
ViT-L16-In21k-augreg	40.51	29.18	24.98	32.30	32.25	90.26	56.25	78.66	22.84	22.84	27.01	26.00	26.07	**20.79**	23.85	35.01
ViT-T16-In21k-augreg	70.72	54.91	48.13	63.04	59.64	80.47	72.30	82.37	56.38	56.38	59.65	49.34	49.51	47.96	**45.51**	46.78
ViT-S16-In21k-augreg	56.91	39.07	42.56	37.83	49.15	86.89	53.98	79.76	38.84	38.84	44.54	33.82	33.77	34.71	**31.44**	35.62
ViT-B16-augreg	63.92	59.44	56.07	61.25	61.40	91.78	70.43	78.08	50.85	50.85	**48.63**	49.36	49.37	51.63	52.50	49.90
ViT-S16-augreg	65.04	59.01	60.65	61.14	63.13	97.27	70.69	84.27	50.76	50.76	49.24	50.90	50.97	47.03	**43.82**	46.61
ViT-B16-In21k-orig	44.49	30.34	30.52	29.56	34.36	65.24	37.91	73.09	31.92	29.38	36.62	28.59	28.10	**27.81**	33.96	30.12
ViT-L16-In21k-orig	40.87	29.30	29.26	31.31	34.67	67.59	34.57	79.20	31.86	29.74	34.80	**27.49**	**27.49**	41.95	33.94	29.74
ViT-B16-In21k-mil	48.33	35.38	34.93	35.77	38.93	92.40	56.97	78.51	47.35	32.23	43.75	32.23	**31.35**	36.31	39.41	37.66
ViT-B16-In21k-augreg2	55.27	45.14	40.60	84.48	46.63	93.06	79.10	77.13	54.68	38.96	46.97	38.96	38.87	38.55	31.83	**30.08**
Swin-T	66.55	72.07	69.91	69.39	67.56	87.59	79.68	85.53	64.37	44.21	55.26	44.21	**43.58**	–	44.09	43.91
SwinV2-S	61.28	66.54	61.14	99.18	62.46	95.01	90.52	77.46	53.75	40.58	50.10	40.58	40.43	–	**39.07**	39.36
SwinV2-B	62.30	64.35	56.95	57.04	59.03	92.12	89.19	75.52	48.95	41.09	45.93	41.09	**39.15**	–	39.84	39.60
SwinV2-L-In21k	41.06	37.35	33.27	71.35	36.96	51.43	42.51	79.08	43.68	27.72	37.99	27.72	**26.27**	–	26.85	27.09
SwinV2-B-In21k	43.83	35.64	31.53	72.77	37.45	66.74	49.22	73.34	42.68	29.27	38.28	29.27	29.03	–	**27.99**	28.12
ConvNeXt-T	62.30	87.58	85.51	96.18	66.19	97.17	93.14	87.04	63.70	46.15	55.65	46.15	46.14	–	41.40	**41.01**
ConvNeXt-B	59.29	86.71	79.59	96.65	64.16	89.27	88.18	77.70	55.86	47.03	50.86	47.03	47.56	–	43.47	**41.47**
ConvNeXt-B-In21k	42.27	40.76	36.47	35.61	41.60	97.23	80.31	78.83	36.75	27.45	33.11	27.45	27.44	–	**25.72**	25.76
ConvNeXtV2-B-In21k	37.35	27.56	27.22	28.14	38.13	95.80	71.17	70.78	28.08	22.83	27.39	23.43	23.43	–	22.65	**21.11**
ConvNeXtV2-T-In21k	45.29	37.63	37.59	41.10	49.52	95.72	74.76	74.91	35.46	30.45	36.76	31.57	31.57	–	29.70	**27.30**
ConvNeXtV2-T	60.34	63.15	56.51	96.01	56.19	92.51	87.32	83.63	57.66	46.93	50.81	46.93	99.32	–	42.04	**39.58**
ConvNeXtV2-B	58.91	68.25	61.54	77.72	55.81	94.50	91.43	74.72	48.22	40.27	44.11	40.27	40.12	–	38.01	**36.21**
ViT-B16-CLIP-L2b	55.39	60.32	58.29	52.46	52.99	63.84	52.66	72.77	45.40	40.72	45.30	41.29	41.25	43.50	22.70	**20.32**
ViT-B16-CLIP-L2b-In12k	43.22	38.81	36.65	42.93	39.90	91.05	65.38	76.15	44.81	**31.39**	39.29	**31.39**	31.40	32.93	54.38	38.86
CLIP-ViT-B16-WIT	56.22	57.96	56.57	50.02	51.38	64.75	51.54	72.52	45.10	**40.24**	45.60	41.54	41.48	42.84	51.10	43.09
DINOv2-B14	59.25	43.48	43.33	44.26	85.15	78.35	43.05	82.77	33.14	33.14	36.42	33.60	33.58	31.74	**26.08**	29.56

We extend the ImageNet-LT evaluation to the 26 checkpoints reported in Tables 15–18, covering ViT, Swin/SwinV2, ConvNeXt/ConvNeXtV2, CLIP, and DINOv2. Each entry reports the mean AUROC or FPR95 over either the four covariate-shift datasets or the five semantic OOD datasets, following the evaluation protocol of (Janiak et al., 2026; Müller and Hein, 2025). All detectors use frozen backbones. Table 23 maps the abbreviated model names to their corresponding checkpoints.

Semantic OOD detection. Tables 15 and 16 show that, LSF-C or LSF-B achieves the highest mean AUROC on 17 of the 26 checkpoints and the lowest mean FPR95 on 16. Both variants outperform all evaluated baselines in both metrics on all seven ConvNeXt/ConvNeXtV2 checkpoints, supporting the usefulness of the selected intermediate representation for semantic OOD detection in convolutional hierarchies. LSF-B achieves the highest AUROC on all seven of these checkpoints and the lowest FPR95 on six, while LSF-C obtains the lowest FPR95 on ConvNeXt-B-In21k.

Table 16: AUROC (↑, %) on OOD datasets averaged over Near OOD (SSB-Hard, NINCO) and Far OOD (iNaturalist, Textures, OpenImage-O) under ImageNet-LT ID calibration

Model	MSP	EN	RE	SC	ASC	KNN	NNG	KPCA	M	RM	rM	MP	rMP	XM†	LSF-C	LSF-B
ViT-B16-In21k-augreg	87.29	91.93	91.32	92.33	90.60	52.94	82.29	79.42	92.86	92.86	91.53	91.84	91.83	92.80	**93.00**	91.69
ViT-L16-In21k-augreg	89.53	93.09	94.12	92.40	91.31	55.86	83.45	80.06	94.55	94.55	93.52	93.55	93.53	**94.83**	94.16	92.12
ViT-T16-In21k-augreg	79.78	85.60	86.90	83.43	83.70	62.51	77.15	74.36	83.69	83.69	83.30	85.95	85.91	86.55	**87.43**	87.22
ViT-S16-In21k-augreg	84.28	90.45	89.94	90.57	86.66	58.47	83.35	77.66	89.77	89.77	88.61	90.80	90.79	90.68	**91.77**	90.96
ViT-B16-augreg	82.64	85.79	86.37	84.86	84.60	48.66	75.19	80.20	86.91	86.91	**87.58**	87.05	87.05	86.53	86.72	87.11
ViT-S16-augreg	82.16	84.93	84.65	84.34	82.75	41.32	73.02	75.31	86.21	86.21	86.99	86.06	86.04	86.78	**88.24**	87.86
ViT-B16-In21k-orig	87.99	93.24	93.16	93.34	91.81	73.70	88.72	84.10	92.47	93.10	91.09	93.28	**93.40**	93.09	92.35	92.91
ViT-L16-In21k-orig	89.48	93.35	93.37	92.97	91.47	73.64	89.83	81.32	92.57	93.16	91.78	**93.65**	**93.65**	88.79	92.20	93.15
ViT-B16-In21k-mil	87.49	90.47	90.85	90.15	90.30	55.42	80.79	81.98	89.32	92.07	89.82	92.07	**92.24**	90.92	90.00	90.25
ViT-B16-In21k-augreg2	84.30	85.08	87.70	61.70	85.64	49.60	71.00	81.73	86.29	89.92	88.27	89.92	89.96	89.96	91.79	**92.04**
Swin-T	81.48	76.67	79.21	80.86	81.00	52.23	69.21	75.13	83.60	87.53	85.51	87.53	87.76	–	**87.85**	87.84
SwinV2-S	82.37	75.69	82.05	29.86	78.36	39.80	63.17	82.31	87.43	89.47	87.42	89.47	89.51	–	**89.86**	89.85
SwinV2-B	82.46	76.89	83.48	83.79	82.51	40.47	60.81	82.39	88.44	89.71	87.87	89.71	**90.20**	–	90.09	90.12
SwinV2-L-In21k	88.32	86.15	90.15	54.08	88.01	77.74	82.94	82.12	90.46	92.95	90.54	92.95	**93.21**	–	93.07	93.03
SwinV2-B-In21k	87.87	86.79	90.22	61.90	87.33	69.97	80.36	84.11	90.15	92.21	90.05	92.21	92.27	–	**92.46**	92.43
ConvNeXt-T	81.28	63.72	68.49	61.27	81.88	42.71	61.11	74.03	84.36	87.96	86.69	87.96	87.98	–	88.97	**89.07**
ConvNeXt-B	82.03	64.43	74.43	51.24	81.89	46.66	63.78	82.01	86.66	88.26	87.16	88.26	88.71	–	89.12	**89.58**
ConvNeXt-B-In21k	88.38	86.47	88.67	89.91	87.35	45.87	73.87	82.04	91.04	92.72	91.24	92.72	92.72	–	93.22	**93.26**
ConvNeXtV2-B-In21k	89.45	90.19	90.78	89.75	87.98	45.62	76.01	86.55	93.38	94.31	92.82	94.18	94.19	–	94.38	**94.85**
ConvNeXtV2-T-In21k	87.10	88.05	88.24	85.07	87.12	49.23	75.35	84.08	91.06	92.31	90.30	92.00	91.99	–	92.48	**93.07**
ConvNeXtV2-T	83.58	79.19	83.04	53.75	85.63	39.90	61.76	78.14	86.27	88.05	87.47	88.05	21.01	–	89.10	**89.66**
ConvNeXtV2-B	82.87	74.26	79.78	68.25	86.19	39.18	61.37	83.67	88.40	89.74	88.78	89.74	89.83	–	90.27	**90.69**
ViT-B16-CLIP-L2b	82.57	74.78	76.83	82.44	85.66	78.72	80.03	83.84	88.46	89.54	88.49	89.37	89.37	88.98	95.40	**95.89**
ViT-B16-CLIP-L2b-In12k	87.29	85.08	87.25	81.47	88.00	57.58	78.10	83.19	89.80	**92.33**	90.47	**92.33**	**92.33**	92.05	89.02	91.02
CLIP-ViT-B16-WIT	82.55	75.00	76.83	86.40	87.26	80.08	81.12	84.28	88.81	**89.86**	88.60	89.58	89.59	89.29	88.35	89.33
DINOv2-B14	81.95	87.89	87.92	87.66	54.96	72.96	86.48	77.22	91.12	91.12	88.76	90.84	90.84	91.47	**93.06**	92.50

On ViT-B16-CLIP-L2b, LSF-B improves AUROC by 6.35 pp and reduces FPR95 by 20.40 pp relative to the strongest baseline for each metric; the corresponding improvements for LSF-C are 5.86 pp and 18.02 pp. On DINOv2-B14, LSF-C leads in both metrics, improving AUROC by 1.59 pp and reducing FPR95 by 5.66 pp over X-Maha w/o FT. The two metrics do not always favor the same detector: on Swin-T, LSF-C and LSF-B exceed rMaha++ in mean AUROC by 0.09 pp and 0.08 pp, respectively, but have higher FPR95 by 0.51 pp and 0.33 pp. Performance varies across checkpoints, with existing baselines remaining stronger on several ViT and CLIP configurations. The results demonstrate broad, though not universal, benefits from the proposed layer selection and fusion.

Covariate-shift detection. Tables 17 and 18 show that, across the two variants, LSF-C or LSF-B achieves the highest mean AUROC on 21 of the 26 checkpoints and the lowest mean FPR95 on 20. At least one variant leads in both metrics on each of the seven AugReg-family ViT checkpoints. On ViT-B16-In21k-augreg, LSF-B improves AUROC by 7.65 pp and reduces FPR95 by 15.14 pp relative to the strongest baseline for each metric; the corresponding improvements for LSF-C are 5.79 pp and 12.43 pp. The benefits also extend to hierarchical architectures: LSF-B leads in both metrics on all seven ConvNeXt/ConvNeXtV2 checkpoints, while the two variants collectively achieve the highest AUROC on all five Swin/SwinV2 checkpoints and the lowest FPR95 on four.

Particularly large improvements occur on ViT-B16-CLIP-L2b, where LSF-B improves AUROC by 11.12 pp and reduces FPR95 by 23.77 pp relative to the strongest baseline for each metric; LSF-C achieves corresponding improvements of 10.25 pp and 21.09 pp. On DINOv2-B14, LSF-B improves AUROC by 12.92 pp and reduces FPR95 by 13.86 pp. These results support the usefulness of selected intermediate representations for detecting distribution changes that are less apparent in the penultimate feature space.

Comparison across checkpoints. The benefits of LSF-C and LSF-B vary across backbones and training configurations. The original and MIIL ViT checkpoints do not show the improvements observed on many AugReg checkpoints, and both variants underperform the strongest baselines on ViT-B16-CLIP-L2b-In12k and CLIP-ViT-B16-WIT. Differences among checkpoints with the

Table 17: FPR95 (↓, %) on Covariate-Shifted ID datasets averaged over ImageNet-V2, ImageNet-C, ImageNet-R, and ImageNet-ES under ImageNet-LT ID calibration

Model	MSP	EN	RE	SC	ASC	KNN	NNG	KPCA	M	RM	rM	MP	rMP	XM†	LSF-C	LSF-B
ViT-B16-In21k-augreg	73.95	66.54	68.52	67.55	68.85	84.36	67.67	82.98	68.36	68.36	69.94	66.97	67.02	66.47	54.04	**51.33**
ViT-L16-In21k-augreg	74.56	66.41	67.40	68.31	69.08	84.32	68.89	85.53	68.39	68.39	71.77	66.91	66.90	63.35	57.10	**53.61**
ViT-T16-In21k-augreg	70.38	60.06	62.91	63.86	65.36	67.40	60.01	76.34	82.10	82.10	80.69	59.46	59.53	69.44	55.79	**55.49**
ViT-S16-In21k-augreg	72.08	63.27	64.28	66.52	66.89	79.68	62.07	82.19	74.78	74.78	74.75	62.10	62.13	66.05	**50.77**	50.97
ViT-B16-augreg	73.45	70.13	70.40	69.21	71.35	90.37	73.11	82.99	71.41	71.41	73.19	69.53	69.56	72.74	**69.03**	70.00
ViT-S16-augreg	73.10	69.50	68.95	72.97	70.82	95.26	73.16	86.88	68.37	68.37	69.94	68.49	68.58	68.52	54.87	**53.46**
ViT-B16-In21k-orig	71.66	67.82	68.05	67.97	68.99	90.86	70.83	84.69	71.11	68.84	70.76	68.51	68.32	**67.60**	69.20	69.62
ViT-L16-In21k-orig	72.44	70.74	70.87	71.30	70.52	93.97	73.18	87.39	74.79	71.88	72.38	70.73	70.76	**70.28**	72.17	72.12
ViT-B16-In21k-mil	70.60	**65.03**	65.21	65.74	67.11	91.83	72.43	85.96	83.64	68.20	77.92	68.20	67.63	75.18	74.67	75.80
ViT-B16-In21k-augreg2	75.08	73.71	70.93	90.79	71.67	89.60	84.16	85.08	77.79	70.61	75.54	70.61	70.51	69.95	59.25	**51.63**
Swin-T	71.82	70.40	68.37	69.25	68.76	78.86	73.49	86.51	77.32	64.63	74.19	64.63	64.31	–	**54.29**	54.38
SwinV2-S	72.20	71.98	67.90	89.92	69.35	91.94	87.61	83.59	73.56	66.71	72.72	66.71	66.64	–	**65.52**	66.12
SwinV2-B	72.32	68.03	**63.71**	65.57	66.04	88.13	86.59	83.03	72.01	67.28	71.30	67.28	66.25	–	66.18	66.18
SwinV2-L-In21k	75.33	79.36	77.94	92.75	75.10	72.17	73.24	87.06	74.94	67.63	73.02	67.63	66.94	–	**64.36**	64.97
SwinV2-B-In21k	74.49	76.13	73.38	92.31	72.90	73.32	73.51	83.30	73.79	67.14	72.54	67.14	67.09	–	**64.48**	65.12
ConvNeXt-T	70.35	84.54	82.96	93.56	69.74	91.88	88.97	87.61	75.19	71.31	73.09	71.31	71.38	–	66.29	**63.58**
ConvNeXt-B	68.98	82.26	76.88	93.53	68.85	83.26	84.45	82.48	72.79	71.60	71.75	71.60	73.80	–	68.65	**65.78**
ConvNeXt-B-In21k	74.43	79.38	76.65	78.83	72.37	92.18	86.62	85.31	68.40	66.66	69.35	66.66	66.66	–	64.28	**61.56**
ConvNeXtV2-B-In21k	76.86	75.50	75.19	75.27	72.58	95.01	90.62	85.46	71.20	69.92	71.49	70.18	70.18	–	69.40	**66.54**
ConvNeXtV2-T-In21k	73.65	74.51	74.45	77.15	65.40	95.29	87.80	84.70	68.87	67.01	69.79	67.30	67.30	–	65.96	**61.92**
ConvNeXtV2-T	69.97	68.13	66.03	90.62	65.98	84.53	77.44	85.01	71.94	66.57	70.73	66.57	94.76	–	63.10	**59.08**
ConvNeXtV2-B	71.60	70.54	67.25	77.56	67.25	90.13	85.40	82.17	73.93	69.47	72.52	69.47	69.36	–	67.35	**63.96**
ViT-B16-CLIP-L2b	75.69	76.79	76.16	73.39	71.21	79.97	74.17	83.92	74.57	68.35	73.91	69.01	68.99	69.82	47.26	**44.58**
ViT-B16-CLIP-L2b-In12k	76.00	77.83	76.71	79.42	74.73	94.15	86.80	86.30	76.74	69.58	74.59	69.58	69.59	**69.05**	78.65	73.39
CLIP-ViT-B16-WIT	75.28	75.27	74.90	73.41	71.30	81.21	73.49	83.89	74.18	**66.56**	73.85	67.99	67.96	68.66	77.56	71.70
DINOv2-B14	83.69	79.00	78.90	78.13	63.21	95.71	82.61	89.29	77.45	77.45	77.10	74.68	74.68	72.12	53.78	**49.35**

same architecture suggest that pretraining and fine-tuning influence the usefulness of intermediate representations for detection. The covariance formulation also matters: LSF-B generally performs better on the ConvNeXt/ConvNeXtV2 checkpoints, whereas LSF-C remains advantageous on several Swin/SwinV2 checkpoints and for semantic OOD detection on DINOv2-B14.

These results indicate that adding an intermediate representation does not guarantee improved detection for every checkpoint. When the selected features provide limited complementary information, their inclusion may fail to improve, or may degrade, detection performance. Relative scoring and radial normalization offer potential extensions within the layer-selection and concatenation framework, but their ability to address these checkpoint-specific limitations requires separate evaluation.

Importantly, this variation does not reflect an inherent limitation of joint cross-layer scoring. In regimes where intermediate features offer little complementary information relative to the penultimate layer, the joint score approximately reduces to the penultimate-layer Mahalanobis baseline, and achieves competitive performance. Furthermore, on checkpoints where LSF underperforms the feature-based baselines, typically Relative-Maha or RS-Maha, the source of gain is orthogonal. Consequently, relative scoring or beta-normalization can be integrated directly into LSF without modifying our selection and fusion framework.

C.4 Additional Ablation Results

C.4.1 Covariance Estimation

While traditional methods utilize pseudo-inverse estimation for the tied precision matrix, LSF adopts Ledoit–Wolf shrinkage (LW) for well-conditioned covariance. Table 19 shows the comparison of empirical covariance with a pseudo-inverse (EC) against Ledoit–Wolf shrinkage (LW) regularized covariance. With ImageNet-1K calibration, LW reduces near- and far-OOD FPR95 by 1.20 pp and 0.29 pp, respectively. LW regularizes the joint precision analytically without requiring any OOD data. This ablation supports LW as a robust estimator for the fused feature space, while attributing the larger detection gains to entropy-based layer selection and aggregation.

Table 18: AUROC (↑, %) on Covariate-Shifted ID datasets averaged over ImageNet-V2, ImageNet-C, ImageNet-R, and ImageNet-ES under ImageNet-LT ID calibration

Model	MSP	EN	RE	SC	ASC	KNN	NNG	KPCA	M	RM	rM	MP	rMP	XM†	LSF-C	LSF-B
ViT-B16-In21k-augreg	69.50	73.39	73.17	72.28	71.39	61.80	70.74	69.62	73.60	73.60	72.84	73.73	73.72	74.36	80.15	**82.01**
ViT-L16-In21k-augreg	69.63	74.29	74.20	73.46	70.59	61.36	70.57	68.97	74.64	74.64	72.46	74.48	74.47	76.80	78.98	**80.86**
ViT-T16-In21k-augreg	73.01	77.46	76.78	75.62	75.43	71.17	76.35	73.13	69.52	69.52	72.96	77.55	77.53	75.34	81.26	**81.57**
ViT-S16-In21k-augreg	70.84	75.33	75.05	73.83	72.89	64.92	74.00	70.33	72.05	72.05	72.78	75.67	75.64	74.70	82.62	**83.17**
ViT-B16-augreg	71.27	74.27	74.10	74.40	73.81	54.89	69.68	71.11	73.77	73.77	73.15	74.31	74.29	73.36	**74.48**	74.04
ViT-S16-augreg	71.74	74.37	74.45	72.46	73.08	47.85	68.88	68.62	74.19	74.19	74.02	74.32	74.30	74.19	80.53	**81.33**
ViT-B16-In21k-orig	71.17	73.89	73.83	73.76	72.94	56.45	69.71	70.71	73.85	74.40	73.76	74.50	74.58	**75.48**	74.24	74.22
ViT-L16-In21k-orig	71.16	73.02	73.08	73.11	72.98	53.77	69.27	68.95	73.18	73.99	73.81	74.44	74.43	**76.15**	73.58	73.95
ViT-B16-In21k-mil	71.87	73.66	73.75	73.22	73.46	53.87	67.95	69.79	70.01	73.68	72.33	73.68	**73.93**	72.40	70.34	70.77
ViT-B16-In21k-augreg2	69.23	67.51	69.99	49.43	71.14	57.26	64.56	70.39	71.25	73.78	71.81	73.78	73.82	74.12	78.92	**81.97**
Swin-T	73.09	72.18	73.78	74.10	74.47	64.12	70.08	69.93	71.71	75.92	73.38	75.92	76.28	–	80.62	**80.67**
SwinV2-S	72.06	69.42	73.30	44.52	70.94	51.35	62.49	70.43	72.27	73.78	72.09	73.78	73.87	–	**74.91**	74.71
SwinV2-B	72.43	71.06	74.87	72.76	74.07	51.91	61.67	70.20	73.37	74.34	72.24	74.34	75.79	–	**75.98**	75.90
SwinV2-L-In21k	70.42	65.54	70.00	40.36	68.50	68.54	68.10	69.21	71.48	73.79	70.92	73.79	74.87	–	**76.48**	76.19
SwinV2-B-In21k	70.21	65.17	69.36	44.93	67.30	66.65	67.27	70.84	71.82	73.96	71.30	73.96	74.14	–	**76.78**	76.43
ConvNeXt-T	73.03	62.91	66.51	58.33	76.02	56.03	64.54	69.74	72.59	75.05	74.05	75.05	75.05	–	77.07	**78.62**
ConvNeXt-B	73.52	63.96	70.50	55.40	75.49	56.61	63.92	73.74	74.05	75.22	73.67	75.22	76.33	–	76.82	**78.29**
ConvNeXt-B-In21k	68.36	64.24	66.05	65.64	71.90	56.76	63.46	70.11	72.76	73.94	71.61	73.94	73.93	–	75.83	**77.66**
ConvNeXtV2-B-In21k	66.43	64.84	65.40	63.55	71.47	48.34	58.40	68.45	71.04	71.83	69.60	71.68	71.68	–	72.63	**75.55**
ConvNeXtV2-T-In21k	68.41	66.36	66.48	63.60	74.96	50.05	61.57	69.52	72.30	73.08	71.42	72.88	72.88	–	74.15	**77.18**
ConvNeXtV2-T	73.48	71.66	74.07	50.91	77.48	59.48	66.96	71.53	73.21	75.17	73.76	75.17	33.12	–	76.69	**78.74**
ConvNeXtV2-B	72.32	69.67	72.63	65.95	75.15	54.15	63.14	71.58	71.74	73.42	71.67	73.42	73.56	–	74.65	**76.75**
ViT-B16-CLIP-L2b	66.74	61.97	62.99	67.09	73.78	65.54	66.62	69.97	71.43	72.96	71.05	72.68	72.68	73.72	84.03	**84.90**
ViT-B16-CLIP-L2b-In12k	66.40	61.91	63.73	59.60	69.63	52.67	61.46	68.76	70.85	72.96	70.88	72.96	72.95	**74.42**	71.68	71.82
CLIP-ViT-B16-WIT	66.74	61.64	62.53	68.72	**73.82**	65.56	66.66	70.25	71.65	73.52	71.34	73.01	73.01	73.72	71.99	72.34
DINOv2-B14	63.78	66.35	66.39	66.38	68.02	49.13	64.26	65.50	70.13	70.13	68.36	70.10	70.10	71.82	81.76	**84.74**

Table 19: FPR95 (% ↓) of Empirical (EC) and Ledoit–Wolf (LW) covariance using ViT-B/16

Method	Feature	Near OOD		Far OOD			Covariate-Shifted ID				Avg.
		SSB	NIN	iNat	TxT	OpO	Im-C	Im-R	Im-V2	Im-ES	
ImageNet-1K											
LSF-C	EC	66.74	44.92	6.37	13.85	22.77	38.03	42.78	90.92	**43.79**	41.13
LSF-C	LW	**65.71**	43.55	**5.45**	15.89	**20.80**	42.63	42.16	**90.26**	46.38	41.42
LSF-B	EC	66.74	44.47	6.36	12.91	22.27	37.56	42.27	91.00	43.93	40.84
LSF-B	LW	66.67	**43.22**	6.32	**12.80**	22.28	**37.44**	**42.13**	91.05	43.82	**40.63**
ImageNet-LT											
LSF-C	EC	68.05	44.25	4.50	13.95	22.55	36.77	40.22	89.10	**41.61**	40.11
LSF-C	LW	**67.34**	44.64	**4.00**	23.65	**20.98**	53.80	41.83	**88.94**	52.43	44.18
LSF-B	EC	68.06	44.20	4.47	13.01	22.25	36.26	39.81	89.11	41.74	39.88
LSF-B	LW	68.01	**42.76**	4.48	**13.00**	22.16	**36.18**	**39.58**	89.11	41.67	**39.66**

C.4.2 Empirical Validation of Gaussianity Assumption

We conducted multivariate normality analyses measuring skewness, excess kurtosis, and Mardia's multivariate kurtosis ratio across features to empirically validate the Gaussianity assumption. As shown in the Table 20, for ViT-B/16 using ImageNet-1K ID calibration, feature representations in LSF lie substantially closer to ideal multivariate Gaussianity than both standard single-layer Maha, normalized single-layer Maha++, and multi-layer X-Maha w/o FT.

C.5 Efficiency & Computational Overhead

We compare the computational overhead of LSF with Mahalanobis++ and X-Maha w/o FT in Table 21. In online deployment, beyond the penultimate layer, LSF-B and LSF-C retain only one

Table 20: Empirical validation of gaussianity assumption of within-class residuals using ViT-B/16 under ImageNet-1K ID calibration.

Metric	Maha	Maha++	X-Maha w/o FT	LSF-B	LSF-C	Ideal Gaussian
Mean absolute skew (↓)	0.034	0.024	0.024	0.023	**0.019**	0
Mean absolute excess kurtosis (↓)	2.116	0.894	0.821	0.633	**0.609**	0
Mardia ratio (→ 1)	1.661	1.261	1.249	1.099	**1.102**	1

Table 21: Inference overhead comparison using ViT-B/16 under ImageNet-LT ID calibration.

Method	No of Layers	Feature Storage	Covariance Storage	GPU Latency	Mean FPR95
Mahalanobis++	1	**3 KB**	**2.4MB**	**0.0009 ms**	53.03
LSF-B	2	6 KB	4.8MB	0.0016 ms	**39.66**
LSF-C	2	6 KB	9.4MB	0.0020 ms	**44.18**
X-Maha w/o FT	12	36 KB	**2.4MB**	0.0464 ms	52.50

additional layer, compared to 11 additional layers for X-Maha w/o FT, keeping its overhead close to Mahalanobis++. Both LSF variants require 6 KB of feature storage, with GPU latencies of 0.0016 ms for LSF-B and 0.0020 ms for LSF-C per sample at batch size 128. These latencies are close to Mahalanobis++ (0.0009 ms) and substantially lower than X-Maha w/o FT (0.0464 ms). Doubling the feature dimension doubles the precision matrix for LSF-B and quadruples for LSF-C using FP32. Meanwhile, LSF-B and LSF-C achieve mean FPR95 scores of 39.66 and 44.18, respectively, compared with 53.03 for Mahalanobis++ and 52.50 for X-Maha w/o FT. Thus, LSF improves OOD detection substantially while adding minimal additional overhead.

C.6 Statistical Evaluation

Table 22 reports FPR95 bootstrap means ± standard deviations using ViT-B/16 under ImageNet-LT ID calibration across all OpenOOD evaluation sets. For each evaluation set, we draw ID and OOD samples separately, with replacement. Within each replicate, every method is evaluated on the same sampled ID indices and the same sampled OOD indices. We use three random sampling seeds, each generating $1{,}000$ resamples, for a total of $B = 3{,}000$ paired bootstrap replicates per evaluation set. We retain the calibrated detectors and its parameters fixed.

LSF variants achieve the lowest bootstrap mean FPR95 on 8 of the 9 evaluation sets. The bootstrap standard deviations range from 0.21 pp to 0.70 pp for LSF-C and from 0.24 pp to 0.55 pp for LSF-B. The gains are 0.25 pp to 22.17 pp for LSF-C and 0.06 pp to 36.53 pp fpr LSF-B.

Table 22: **Bootstrap statistical evaluation.** FPR95 is reported using ViT-B/16 under ImageNet-LT ID calibration. Each cell reports the bootstrap mean ± standard deviation across 3 resampled runs.

Method	Near OOD		Far OOD			Covariate-Shifted ID			
	SSB	NIN	iNat	TxT	OpO	Im-C	Im-R	Im-V2	Im-ES
MSP	76.55 ± 0.35	62.67 ± 0.67	34.07 ± 0.62	58.16 ± 0.79	44.91 ± 0.56	73.96 ± 0.52	66.22 ± 0.43	90.13 ± 0.33*	70.03 ± 0.25
Energy	65.23 ± 0.36	51.67 ± 0.66	28.88 ± 0.54	46.27 ± 0.73	33.75 ± 0.46	66.60 ± 0.55	67.04 ± 0.41	90.83 ± 0.32	70.33 ± 0.24
ReAct	**64.88 ± 0.35**	49.27 ± 0.68	18.26 ± 0.45	43.85 ± 0.71	26.75 ± 0.45	65.85 ± 0.51	60.12 ± 0.43	89.90 ± 0.34	67.84 ± 0.25
SCALE	83.57 ± 0.25	85.82 ± 0.47	85.64 ± 0.40	87.93 ± 0.45	79.31 ± 0.35	87.94 ± 0.36	96.92 ± 0.11	94.51 ± 0.25	83.74 ± 0.18
AdaSCALE	68.89 ± 0.34	54.81 ± 0.72	24.40 ± 0.50	51.47 ± 0.73	33.57 ± 0.47	68.69 ± 0.51	62.08 ± 0.40	89.27 ± 0.36	66.60 ± 0.26
KNN	96.67 ± 0.10	96.08 ± 0.22	84.98 ± 0.41	92.40 ± 0.38	95.06 ± 0.20	88.45 ± 0.36	84.14 ± 0.30	94.32 ± 0.25	91.41 ± 0.18
NNGuide	91.94 ± 0.16	88.72 ± 0.39	66.03 ± 0.59	70.82 ± 0.64	77.96 ± 0.47	83.88 ± 0.40	73.80 ± 0.40	94.40 ± 0.25	84.47 ± 0.25
KPCA	88.31 ± 0.26	80.29 ± 0.61	67.81 ± 0.68	75.62 ± 0.73	73.24 ± 0.58	84.96 ± 0.44	77.17 ± 0.40	92.39 ± 0.29	85.61 ± 0.29
Maha	85.54 ± 0.31	71.12 ± 0.69	13.06 ± 0.53	65.07 ± 0.85	38.54 ± 0.75	84.99 ± 0.45	51.86 ± 0.53	89.75 ± 0.35	84.54 ± 0.30
RS-MD	68.95 ± 0.33	48.86 ± 0.68	4.38 ± 0.23	48.04 ± 0.80	24.47 ± 0.52	73.00 ± 0.51	46.86 ± 0.48	89.34 ± 0.38	73.05 ± 0.42
rMaha	77.94 ± 0.41	58.82 ± 0.76	10.34 ± 0.38	57.63 ± 0.90	29.80 ± 0.59	80.25 ± 0.55	51.92 ± 0.47	89.52 ± 0.36	80.18 ± 0.43
Maha++	68.95 ± 0.33	48.86 ± 0.68	4.38 ± 0.23	48.04 ± 0.80	24.47 ± 0.52	73.00 ± 0.51	46.86 ± 0.48	89.34 ± 0.38	73.05 ± 0.42
rMaha++	68.84 ± 0.34	48.73 ± 0.68	4.37 ± 0.22	47.92 ± 0.80	24.47 ± 0.53	72.92 ± 0.51	46.88 ± 0.48	89.34 ± 0.37	72.88 ± 0.42
X-Maha w/o FT	69.94 ± 0.34	49.16 ± 0.72	4.54 ± 0.23	45.81 ± 0.84	23.21 ± 0.56	72.81 ± 0.54	45.35 ± 0.48	89.19 ± 0.38	72.30 ± 0.43
LSF-C	67.35 ± 0.34	44.70 ± 0.70	**4.01 ± 0.21**	23.64 ± 0.63	**20.96 ± 0.46**	53.76 ± 0.54	41.83 ± 0.42	**88.94 ± 0.37**	52.46 ± 0.29
LSF-B	68.07 ± 0.36	**43.20 ± 0.51**	4.48 ± 0.24	**13.07 ± 0.48**	22.25 ± 0.48	**36.28 ± 0.55**	**39.79 ± 0.49**	89.10 ± 0.36	**41.73 ± 0.26**

D METHODS

We briefly review the OOD detection methods evaluated in this paper.

Let $f(x) \in \mathbb{R}^C$ denote the pre-softmax logits of the network, and $h(x) \in \mathbb{R}^d$ denote the penultimate feature representation (with $h_l(x)$ for layer-wise features).

- **MSP Hendrycks and Gimpel (2017):** Uses the maximum softmax probability as the confidence score:
$$\mathcal{S}_{\text{MSP}}(x) = \max_c \frac{\exp(f_c(x))}{\sum_j \exp(f_j(x))}$$
- **Energy Liu et al. (2020):** Uses the free energy of logits as an unnormalized confidence score:
$$\mathcal{S}_{\text{Energy}}(x) = T \log \sum_c \exp(f_c(x)/T)$$
- **ReAct Sun et al. (2021):** Truncates the penultimate activations $h(x)$ at a threshold v to produce rectified logits $f^{\text{clip}}(x)$, computing the energy score as:
$$\mathcal{S}_{\text{ReAct}}(x) = T \log \sum_{c=1}^{C} \exp(f_c^{\text{clip}}(x)/T)$$
- **SCALE Xu et al. (2024):** Scales all activations by $e^{r_p(x)}$ and computes energy confidence, using a fixed percentile p:
$$r_p(x) = \frac{\sum_{j=1}^{d} h_j(x)}{\sum_{j:\, h_j(x) > P_p(h(x))} h_j(x)}, \quad \mathcal{S}_{\text{SCALE}}(x) = \log \sum_c \exp\Big([W(e^{r_p(x)} h(x)) + b]_c\Big).$$
Here P_p is the activation percentile across all features.
- **AdaSCALE Regmi (2026):** Replaces SCALE's fixed percentile with a sample-dependent value:
$$p(x) = p_{\min} + \big[1 - F_{Q'}(Q'(x))\big](p_{\max} - p_{\min}).$$
Here Q' combines perturbation-induced shifts in top activations with a correction from positive perturbed activations, and $F_{Q'}$ is its ID-validation empirical CDF.
- **KNN Sun et al. (2022):** Uses the negative distance to the k-th nearest neighbor in feature space:
$$\mathcal{S}_{\text{KNN}}(x) = -\|h(x) - h_{(k)}(x)\|_2$$
- **NNGuide Park et al. (2023):** Multiplies the energy confidence by guidance from an ID feature bank. For bank samples x_i, let $s_i = \mathcal{S}_{\text{Energy}}(x_i)$ and $a_i(x) = s_i \tilde{h}(x_i)^\top \tilde{h}(x)$:
$$\mathcal{S}_{\text{NNGuide}}(x) = \mathcal{S}_{\text{Energy}}(x)\, \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} a_i(x).$$
Here $\mathcal{N}_k(x)$ contains the k largest confidence-weighted similarities $a_i(x)$, favoring nearby, high-confidence ID features and suppressing confidence away from the ID bank.
- **KPCA Fang et al. (2024):** Maps normalized features through Gaussian random Fourier features $\Phi(x)$, then scores their PCA reconstruction error:
$$\mathcal{S}_{\text{KPCA}}(x) = -\left\|(I - U_q U_q^\top)\big(\Phi(x) - \mu_\Phi\big)\right\|_2.$$
Here μ_Φ and U_q are the mapped ID training mean and top q principal directions, respectively.
- **Mahalanobis Lee et al. (2018):** Models features as class-conditional Gaussians with shared covariance Σ:
$$\mathcal{S}_{\text{Maha}}(x) = -\min_c (h(x) - \mu_c)^\top \Sigma^{-1} (h(x) - \mu_c)$$
- **Mahalanobis++ Müller and Hein (2025):** Applies a scale-invariant Mahalanobis distance on ℓ_2-normalized features:
$$\mathcal{S}_{\text{Maha++}}(x) = -\min_c (\tilde{h}(x) - \tilde{\mu}_c)^\top \tilde{\Sigma}^{-1} (\tilde{h}(x) - \tilde{\mu}_c)$$

- **Radially Scaled Mahalanobis Janiak et al. (2026):** Preserves feature directions while adjusting their radii through $h_\beta(x) = h(x)/\|h(x)\|_2^\beta$:

$$\mathcal{S}_{\text{RS-Maha}}(x) = -\min_c \big(h_\beta(x) - \mu_{c,\beta}\big)^\top \Sigma_\beta^{-1} \big(h_\beta(x) - \mu_{c,\beta}\big).$$

  The cases $\beta = 0$ and $\beta = 1$ recover Mahalanobis and Mahalanobis++, respectively. β is selected combining local intrinsic dimensionality and the magnitude of the within-class spectral slope from ID calibration set.

- **Relative Mahalanobis Ren et al. (2021):** Refines the Mahalanobis score by subtracting a background distance term based on (μ_0, Σ_0):

$$\mathcal{S}_{\text{rMaha}}(x) = -\min_c \left[d_c(x) - d_0(x)\right], \text{where}$$

$$d_c(x) = (h(x) - \mu_c)^\top \Sigma^{-1}(h(x) - \mu_c), \quad d_0(x) = (h(x) - \mu_0)^\top \Sigma_0^{-1}(h(x) - \mu_0).$$

- **Relative Mahalanobis++:** Extends relative Mahalanobis by using ℓ_2-normalized features $\tilde{h}(x)$:

$$\mathcal{S}_{\text{rMaha++}}(x) = -\min_c \left[(\tilde{h}(x) - \tilde{\mu}_c)^\top \tilde{\Sigma}^{-1}(\tilde{h}(x) - \tilde{\mu}_c) - (\tilde{h}(x) - \tilde{\mu}_0)^\top \Sigma_0^{-1}(\tilde{h}(x) - \tilde{\mu}_0)\right]$$

- **X-Mahalanobis Wei et al. (2025):** Variance-weighted additive fusion of layer-wise features, followed by calculation of Mahalanobis distance. Relies on different parameter-efficient finetuning techniques.

$$h_{\text{f}}(x) = \sum_{l=1}^{L} \alpha_l h_l(x),$$

$$\mathcal{S}_{\text{X-Maha}}(x) = -\min_c \left(h_{\text{f}}(x) - \mu_c\right)^\top \Sigma^{-1} \left(h_{\text{f}}(x) - \mu_c\right).$$

- **LSF (Ours):** Evaluates a single joint Mahalanobis++ distance on a unified representation space $\phi_c(x)$, fusing ℓ_2-normalized features from the terminal (penultimate) layer and the top $K-1$ intermediate layers. The tied precision matrix $\hat{\Sigma}_{\mathcal{K}}^{-1}$ is derived analytically via Ledoit–Wolf shrinkage:

$$\mathcal{S}_{\text{LSF}}(x) = -\min_c \big(\boldsymbol{\phi}(x) - \hat{\boldsymbol{\mu}}_c^{\mathcal{K}}\big)^\top \hat{\Sigma}_{\mathcal{K}}^{-1} \big(\boldsymbol{\phi}(x) - \hat{\boldsymbol{\mu}}_c^{\mathcal{K}}\big)$$

E Reproduction Details

Evaluation Rigor and Stability. Because LSF is a post-hoc framework operating on frozen, standardized model checkpoints, the extracted feature representations and resulting OOD scores are algorithmically deterministic.

Consequently, we empirically validate the robustness of our method across multiple distinct architectural paradigms and diverse distribution shifts, rather than through stochastic repetition. The consistent performance of LSF across multiple, distinct architectural designs (global attention, hierarchical attention, and modernized convolutions) and under highly varied distribution shifts (from long-tailed class imbalance to severe covariate degradation) serves as a comprehensive validation of its algorithmic robustness.

Hyperparameter, Feature Extraction, and Well-Conditioned Covariance Estimation. LSF is ID calibrated, training-free method. We use publicly available pretrained backbones (e.g., ViT-B/16, Swin-T, and ConvNeXt-T) and standard OOD benchmarks with their official splits. Features are extracted from intermediate layers, ℓ_2-normalized, and class-conditional statistics are computed from in-distribution training data.

The joint covariance matrix over the selected layers is estimated using Ledoit–Wolf shrinkage, and OOD scores are computed via the Mahalanobis++ formulation described in Section 3. No additional training or hyperparameter tuning is required. Moreover, our detailed experimental set-up is described in Section 4.

Table 23: Model checkpoint names and the corresponding names used in tables.

Checkpoint	Paper/table name
vit_base_patch16_224.augreg2_in21k_ft_in1k	ViT-B16
vit_base_patch16_224.augreg_in21k_ft_in1k	ViT-B16-In21k-augreg
vit_large_patch16_224.augreg_in21k_ft_in1k	ViT-L16-In21k-augreg
vit_tiny_patch16_224.augreg_in21k_ft_in1k	ViT-T16-In21k-augreg
vit_small_patch16_224.augreg_in21k_ft_in1k	ViT-S16-In21k-augreg
vit_base_patch16_224.augreg_in1k	ViT-B16-augreg
vit_small_patch16_224.augreg_in1k	ViT-S16-augreg
vit_base_patch16_224.orig_in21k_ft_in1k	ViT-B16-In21k-orig
vit_large_patch32_384.orig_in21k_ft_in1k	ViT-L16-In21k-orig
vit_base_patch16_224_miil.in21k_ft_in1k	ViT-B16-In21k-mil
vit_base_patch16_224.augreg2_in21k_ft_in1k	ViT-B16-In21k-augreg2
swin_tiny_patch4_window7_224.ms_in1k	Swin-T
swinv2_small_window16_256.ms_in1k	SwinV2-S
swinv2_base_window16_256.ms_in1k	SwinV2-B
swinv2_large_window12to24_192to384.ms_in22k_ft_in1k	SwinV2-L-In21k
swinv2_base_window12to24_192to384.ms_in22k_ft_in1k	SwinV2-B-In21k
convnext_tiny.fb_in1k	ConvNeXt-T
convnext_base.fb_in1k	ConvNeXt-B
convnext_base.fb_in22k_ft_in1k	ConvNeXt-B-In21k
convnextv2_base.fcmae_ft_in22k_in1k_384	ConvNeXtV2-B-In21k
convnextv2_tiny.fcmae_ft_in22k_in1k_384	ConvNeXtV2-T-In21k
convnextv2_tiny.fcmae_ft_in1k	ConvNeXtV2-T
convnextv2_base.fcmae_ft_in1k	ConvNeXtV2-B
vit_base_patch16_clip_224.laion2b_ft_in1k	ViT-B16-CLIP-L2b
vit_base_patch16_clip_224.laion2b_ft_in12k_in1k	ViT-B16-CLIP-L2b-In12k
clip-ViT-B16	CLIP-ViT-B16-WIT
vit_base_patch14_dinov2.lvd142m	DINOv2-B14

Code Release. An anonymized ZIP archive containing the complete implementation is included with the submission. Upon acceptance, the codebase will be made publicly available. Our implementation ensures full reproducibility using the publicly accessible OOD benchmarks detailed in this work.

Compute Resources. All our calibration and evaluation are performed using on a workstation with an Intel Core i9 processor (16 cores), 32 GB DDR5 SDRAM, and an NVIDIA RTX 4090 GPU with 24GB GDDR6X SDRAM.

F LIMITATIONS

Reliance on Pretrained Backbones. LSF relies on pretrained models with complementary intermediate layers. If the intermediate layers provide limited complementary information, the benefit of joint modeling may be reduced. Reducing this reliance on complementary intermediate representations remains an open challenge for future work.

Computational and memory overhead. LSF incurs additional computational and memory overhead compared to single-layer methods. However, this overhead remains small, as quantified in Appendix C.5. This is because LSF selects only one additional intermediate layer with the sharpest entropy-density drop while anchoring the penultimate layer.

Theoretical assumptions on OOD shifts. The non-degradation and increased-sensitivity result in Proposition A relies on modeling the OOD residual as $e^{\mathrm{OOD}} = e + \delta$, where the OOD perturbation δ is independent of the centered ID residual e. This assumption enables the cross terms in the expected quadratic distance to vanish and yields the nonnegative conditional shift contribution. In practice, OOD shifts may interact with the ID representation in ways that violate this independence assumption. Thus, the theoretical guarantee in Proposition A should be interpreted as a characterization under the stated shift model rather than a guarantee for arbitrary distribution shifts. Extending the analysis to more general OOD shift models remains an open direction.

G BROADER IMPACTS

This work improves the reliability of deep neural networks in open-world settings by enabling post-hoc out-of-distribution (OOD) detection. The proposed approach reduces computational overhead by avoiding retraining and does not use OOD data.

However, LSF shares limitations common to OOD detection methods. It is not guaranteed to detect all anomalous inputs, and over-reliance on its outputs may lead to insufficient oversight. As such, it should be used as part of a broader safety framework that includes complementary safeguards, such as uncertainty estimation or human supervision.

H PUBLIC RESOURCES USED

In this section, we acknowledge the public resources used during the course of this work.

H.1 PUBLIC DATASETS USED

- **OpenOOD** ..... MIT License
- **NINCO** ..... MIT License
- **ImageNet-1K** ..... ImageNet Agreement
- **ImageNet-LT** ..... BSD-3-Clause license
- **Texture** ..... Creative Commons BY-SA 4.0
- **OpenImage-O** ..... Apache License 2.0
- **ImageNet-C** ..... ImageNet Agreement
- **ImageNet-ES** ..... ImageNet Agreement
- **ImageNet-R** ..... ImageNet Agreement
- **ImageNet-V2** ..... ImageNet Agreement
- **iNaturalist** ..... Mixed Creative Commons Licenses

H.2 PUBLIC IMPLEMENTATIONS USED

- **Mahalanobis++** ..... Creative Commons BY 4.0 License
- **timm** ..... Apache License 2.0
- **scikit-learn** ..... BSD-3-Clause License
- **X-Mahalanobis** ..... Creative Commons BY 4.0 License
- **Energy** ..... MIT License
- **ReAct** ..... MIT License
- **Energy** ..... MIT License
- **KPCA** ..... N/A
- **NNGuide** ..... Apache License 2.0
- **SCALE** ..... MIT License
- **AdaSCALE** ..... MIT License