When Models Leave the Training Distribution

A Tutorial on Out-of-Distribution (OOD) Detection

Sanjay Chawla[2] Dishanika Denipitiyage[1] Suranga Seneviratne[1] Aditya Krishna Menon[3]

KDD '26 — 32nd ACM SIGKDD Conference
Aug 09–13, 2026, Jeju Island, Republic of Korea

[1]The University of Sydney [2]QCRI, HBKU [3]Google

Tutorial Roadmap

Part I — Foundations

- Introduction & Motivation
- Problem Formulation & Settings
- Benchmarks, Datasets & Metrics
- Detection vs. Generalization

Part II — OOD Method Families

- Post-hoc Methods
- Training *without* Outliers
- Training *with* Outliers
- Foundation-Model Approaches

The core — organised by two axes:
needs retraining? needs auxiliary OOD?

Part III — Frontiers

- Applications
- Open Problems & Future Directions

Introduction & Motivation

An Example from Human Learning Process

- What is this?

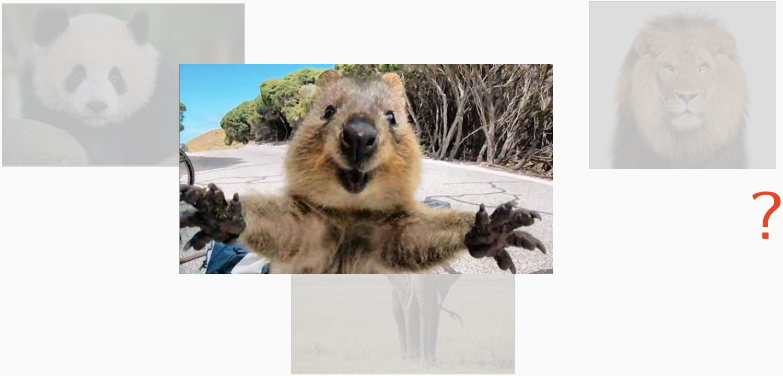
Panda

Lion

Elephant

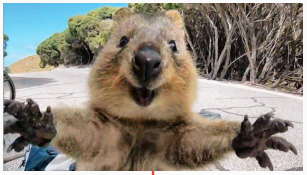
An Example from Human Learning Process

- What is this?



An Example from Human Learning Process

- What is this? Two reasonable answers:



?

Guess the **closest known class**

(with some uncertainty)

Stock DNNs *always* do this
— confidently, and wrongly

Refuse to answer:

"I don't know"

Only an OOD-aware model can do this
— the subject of this tutorial

Why OOD Detection Matters

- **Clinical screening.** A diabetic-retinopathy model validated at specialist-grade accuracy was deployed in Thai clinics; images taken under real conditions — poor lighting, no dedicated dark room — fell outside its training distribution and were silently rejected, delaying referrals.
- **Autonomous driving.** In the 2018 Tempe fatality, the perception stack cycled the pedestrian through *vehicle*, *bicycle* and *other* — it had no class for a person crossing away from a crosswalk, and each re-classification threw away the object's motion history.
- **Cyberecurity.** Malware family classifiers keep labelling never-before-seen families as known ones, at high confidence, as the threat landscape drifts away from the training set.

The Closed-World Assumption (and How It Breaks)

Classifiers are trained assuming $\mathcal{X} \times \mathcal{Y}$ is fixed. At deployment they face everything else. Two fundamentally different situations arise:

Same concept, different appearance

blurry dog, sketch of a cat, rainy road scene

$\Rightarrow$ **generalize**

Completely new concept

car appears in an animal classifier, unknown malware type

$\Rightarrow$ **refuse**

No current system does both reliably.

What This Tutorial Covers (and Does Not)

In scope

- Standard (semantic-shift) OOD
- Four method families
- Benchmarks, apps, open problems

Adjacent / lighter touch

- Full-spectrum / covariate shift
- OOD generalisation
- Test-time adaptation

Problem Formulation & Related Settings

Standard OOD Detection: Formal Setup

Input space $\mathcal{X}$, labels $\mathcal{Y} = \{1, \ldots, C\}$, classifier f, dataset D.

$$g_\lambda(x) = \begin{cases} \text{in} & S(x; f) \geq \lambda \\ \text{out} & S(x; f) < \lambda \end{cases}$$

The Detector's Objective: Open Space Risk

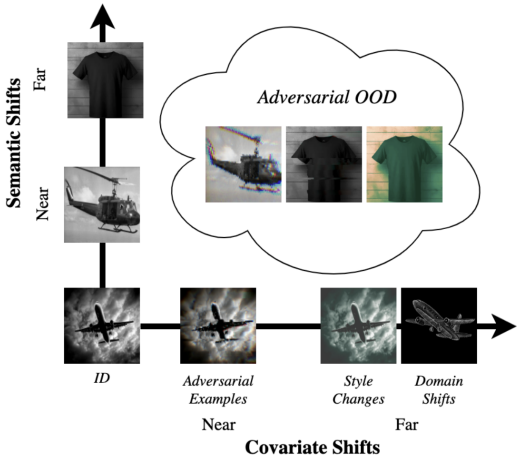
Wrap the score/threshold rule into a detector $G(x; f) \in \{0, 1\}$ (1 =OOD) and ask what it should minimise:

$$R_{\mathcal{O}}(G) = \frac{\iint_{\{G=0\}} p_{\mathcal{D}_{\text{out}}}(x, y)\, dx\, dy}{\iint_{\{G=0\}} \left[p_{\mathcal{D}_{\text{out}}}(x, y) + p_{\mathcal{D}_{\text{in}}}(x, y)\right] dx\, dy}$$

- Numerator and denominator both range over the accepted (G=0, predicted-ID) region.
- $R_{\mathcal{O}}$ is the fraction of accepted mass that is actually OOD — the open-space-risk formulation used by **OpenOOD** v1.5.
- Choosing S and λ *is* choosing G: every method in Part II is a different answer to this one optimisation problem.

Zhang, Jingyang, et al. "OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection." Journal of Data-centric Machine Learning Research (2024).

Semantic vs. Covariate Shift: The Key Distinction



$y \notin \mathcal{Y}_{\text{in}}$

$\Rightarrow$ **reject**

$P(Y \mid X)$ **preserved**

$\Rightarrow$ **generalize**

Semantic Shift: New Classes Appear

New **classes** appear at test time: $y \notin \mathcal{Y}$.

- Classifier trained on animals encounters a car
- Medical AI sees an unknown pathology
- Fraud detector faces a new fraud type

Desired behaviour: detect and reject.

Related problems

- *OOD Detection* — broad spectrum of unseen classes
- *Open-Set Recognition* — known + unknown classes, one classifier
- *Anomaly Detection* — rare, highly deviant instances

Covariate Shift: Same Classes, Different Appearance

Same classes, **different appearance**:
$y \in \mathcal{Y},\ p(x) \neq p_{\text{train}}(x)$.

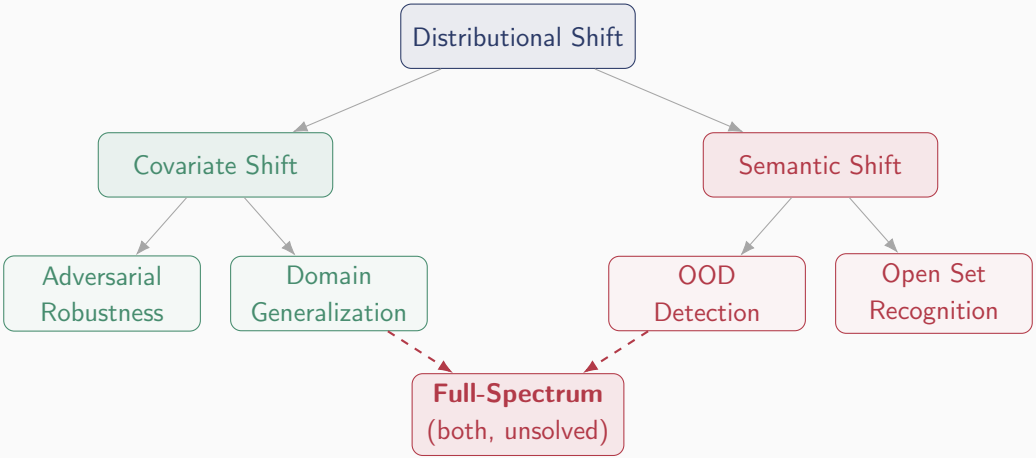
- Corruption: blur, noise, JPEG artefacts
- Style: paintings, sketches, domain shift
- Adversarial: imperceptible perturbations

Desired behaviour: classify correctly (generalize).

Related problems

- *Domain Generalization* — non-adversarial style/domain change
- *Adversarial Robustness* — maliciously crafted perturbations
- *Sensory Anomaly Detection* — unusual readings within an ID class

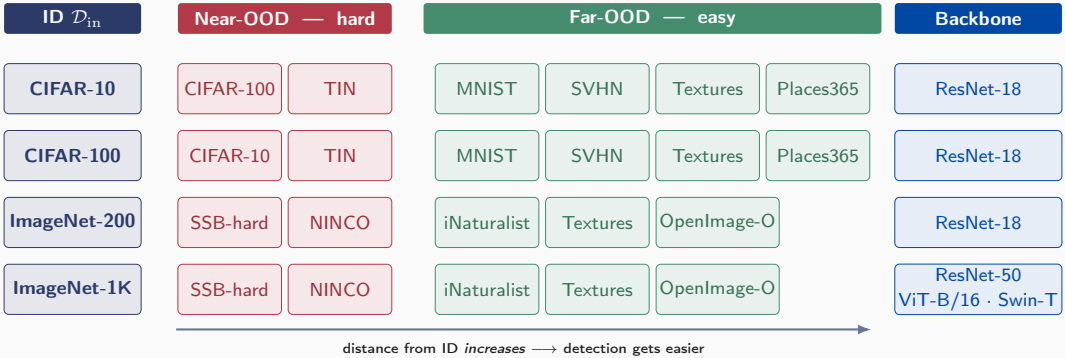
Distributional Shift: The Taxonomy



Neighbouring Problems

Setting	Shift type	Train-time OOD data?	Goal
Anomaly / outlier detection	semantic (usually 1-class)	no	flag rare instances
Open-Set Recognition (OSR)	semantic	no	classify knowns *and* reject novel
OOD Detection (this tutorial)	semantic	optional (OE)	accept ID, reject OOD
Distribution-shift / OOD generalisation	covariate	no	stay accurate under shift
Full-spectrum OOD detection	both	optional	reject semantic, absorb covariate

OpenOOD v1.5: Benchmark Taxonomy

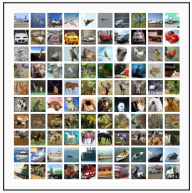


Full-spectrum adds csID to the ID test set: ImageNet-C/-R/-V2. Foundation models (CLIP zero-shot, DINOv2 linear probe) share a ViT-B backbone.

Zhang, Jingyang, et al. "OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection." Journal of Data-centric Machine Learning Research (2024).

Near-OOD vs. Far-OOD

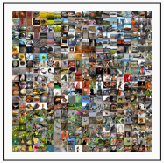
ID (train)



CIFAR-10

Near-OOD — hard

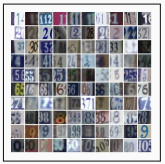
CIFAR-100



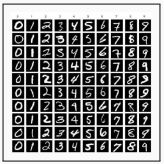
ImageNet

same domain, unseen classes

Far-OOD — easy



SVHN



MNIST

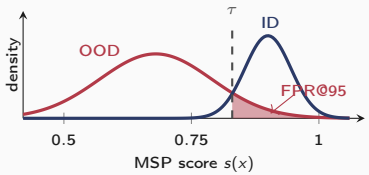
different domain / low-level statistics

Distributional distance from ID *increases* $\longrightarrow$ detection generally gets easier

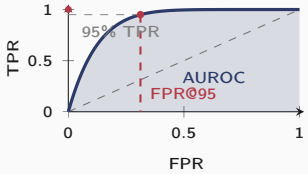
Evaluation Metrics

OOD detection is binary (ID = positive); score each input by MSP $s(x) = \max_c p_c(x)$, then:

- AUROC — area under the ROC curve; equals $\Pr(s_{\text{ID}} > s_{\text{OOD}})$. Threshold-free (0.5 chance, 1.0 perfect). **Higher** ↑.
- FPR@95 — FPR at the threshold with $\text{TPR} = 95\%$: a single operating point. **Lower** ↓.
- AUPR — area under precision–recall; focuses on the positive class under *class imbalance*.



FPR@95: OOD mass past τ once 95% of ID is retained.



AUROC: area under ROC; FPR@95 = FPR at TPR = 95%.

Detection vs. Generalization: An Open Tension

The Literature Imbalance

OOD Detection (semantic)

- >100 papers in recent years
- Well-established benchmarks (**OpenOOD**)
- Rich theory: energy, density, distance scores

OOD Generalization (covariate)

- Active but smaller community
- Separate benchmarks (**DomainBed**, **RobustBench**)
- Methods: data augmentation, self-supervision

Full-Spectrum (both simultaneously)

≈ 0 dedicated methods

Introduced as a benchmark setting in **OpenOOD** v1.5.

No method consistently succeeds.

Zhang, Jingyang, et al. "OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection." Journal of Data-centric Machine Learning Research (2024).

The Tension: One System, Two Desiderata

We want a single system that:

(D) Detects semantic OOD: $S(x) \geq \lambda$ when $y \notin \mathcal{Y}$

(G) Generalizes over covariate shift: $S(x) < \lambda$ when $y \in \mathcal{Y}$, despite x being covariate-shifted

The problem: from the model's perspective, *both* types of shift look like departure from training data.

Low confidence $\approx$ *covariate shift* ***or*** *semantic shift?*

The Mathematical Reason a Single Score Fails

Most OOD scores are monotone functions of distributional distance:

$$S(x) = h\big(d(x, \mathcal{D}_{\text{in}})\big), \qquad h \text{ increasing.}$$

Semantic OOD

$d(x, \mathcal{D}_{\text{in}}) \uparrow \Rightarrow \; S(x) \uparrow$

want: $S(x) \geq \lambda$ ✓

Covariate OOD

$d(x, \mathcal{D}_{\text{in}}) \uparrow$ (also!) $\Rightarrow S(x) \uparrow$

want: $S(x) < \lambda$ ×

A single scalar score cannot separate them.

Score distributions overlap: $p_{\text{sem}}(S) \cap p_{\text{cov}}(S) \neq \emptyset$.

Why Scores Fail in Practice: Energy & Mahalanobis

Energy $E(x) = -\log \sum_k e^{f_k(x)}$

- Semantic: input activates no class strongly $\Rightarrow \sum_k e^{f_k} \downarrow \Rightarrow E \uparrow$.
- Covariate: blur/style destroys the learned texture cues, suppressing *all* logits similarly $\Rightarrow E \uparrow$ too.

Mahalanobis

$M(x) = \min_k (\phi(x) - \mu_k)^\top \Sigma^{-1} (\phi(x) - \mu_k)$

- Semantic: $\phi(x)$ lands far from every class cluster $\Rightarrow M \uparrow$.
- Covariate: a supervised ϕ has no incentive to keep corrupted images near $\mu_k \Rightarrow M \uparrow$ too.

Any score built from "distance to training data" inherits this ambiguity — the root cause is that ϕ (or f) is a representation optimised only to classify clean ID images.

Liu, Weitang, et al. "Energy-based out-of-distribution detection." Advances in neural information processing systems 33 (2020): 21464-21475.

Evidence: OpenOOD v1.5 — Standard vs. Full-Spectrum

Full-spectrum setting: ID test $= \mathcal{D}_{\text{in}}^{\text{test}} \cup \mathcal{D}_{\text{in}}^{\text{cs-test}}$ (ImageNet-C/R/V2 folded in as in-distribution)

Method	Standard (Near/Far AUROC)		Full-Spectrum (Near/Far AUROC)	
	IN-1K	IN-200	IN-1K	IN-200
MSP	76.0 / 85.2	83.3 / 90.1	60.8 / 72.3	62.1 / 74.4
ReAct	77.4 / 93.7	81.9 / 92.3	62.7 / 83.5	65.3 / 81.6
ASH	78.2 / 95.7	82.4 / 93.9	61.2 / 83.0	66.6 / 82.6
MOS	72.9 / 82.8	69.8 / 80.5	66.2 / 76.8	—

Near-OOD AUROC drops >10 points once covariate-shifted ID is folded in, for every method. Best full-spectrum result: AugMix+SHE at 69.7% near-OOD on IN-1K.

Zhang, Jingyang, et al. "OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection." Journal of Data-centric Machine Learning Research (2024).

Evidence: The Unified Robustness Problem

A unified robust system must satisfy simultaneously:

$$\forall \|\delta\|_p \leq \epsilon : \ f(x+\delta) = f(x) \qquad \text{and} \qquad g(x) = \begin{cases} 1 & x \in \mathcal{D}_{\text{in}} \\ 0 & \text{else} \end{cases}$$

Known trade-off from adversarial training:

$$\textit{Robustness} \uparrow \iff \textit{Generalization accuracy} \downarrow$$

- Adversarial training flattens the loss landscape
- Flat landscape reduces overconfidence (helps detection, i.e. balances D/G)
- ...but also reduces in-distribution accuracy (hurts classification)

Karunanayake, Naveen, et al. "Out-of-Distribution Data: An Acquaintance of Adversarial Examples — A Survey." arXiv preprint arXiv:2404.05219 (2024).

Foundation Models: A Promising Direction

Observation (Zhang et al. 2024, **OpenOOD** v1.5):

- **DINOv2** (linear probe, ViT-B): consistently and *remarkably* outperforms the fully-supervised ResNet on all OOD detection tasks.
- **CLIP** (zero-shot, ViT-B): substantial gains on *far-OOD* detection, in both standard and full-spectrum settings.

Why might this be? Foundation models train on internet-scale data with diverse augmentations and objectives (contrastive, masked prediction, language alignment) — exposure that may induce a *richer* internal representation, partially separating semantic content from appearance.

Is the disentanglement implicit in the embedding space?

The Surprising Result: Scale Beats Similarity

Natural worry Mayilvahanan et al., ICLR 2024

LAION simply *contains* near-duplicates of OOD benchmarks — is CLIP's "generalisation" just retrieval from training data?

Intervention: prune LAION so its train-test similarity gap *exactly matches* ImageNet-Train's; retrain CLIP.

	IN-Sketch	**IN-R**
ResNet (ImageNet-Train)	∼14%	∼35%
CLIP baseline (LAION-200M)	48.6%	72.6%
CLIP *sketch-pruned*	**44.8%**	**69.4%**

Identical similarity gap, yet CLIP $\gg$ ResNet by **30+ points** $\Rightarrow$ foundation models generalise from *scale & diversity*, not from memorising near-duplicates.

Mayilvahanan, Prasanna, et al. "Does CLIP's Generalization Performance Mainly Stem from High Train-Test Similarity?" ICLR, 2024.

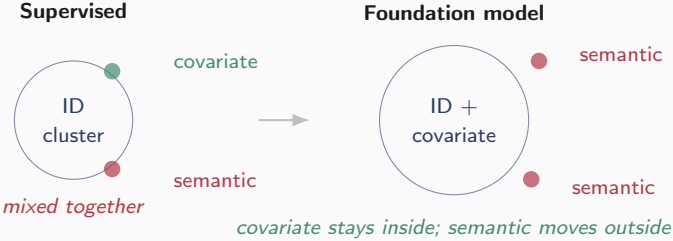
Foundation Models: The Disentanglement Hypothesis

Supervised model

Trained on clean $\mathcal{D}_{\text{in}}$, one task only.
Covariate-shifted *and* semantic-OOD images both scatter outside the training cluster $\Rightarrow$ score conflates the two.

Foundation model

Trained on billions of image–text pairs; sees the same object under many styles.
Covariate-shifted images *may stay* near their class cluster; semantic OOD does not $\Rightarrow$ score **partially separable**.



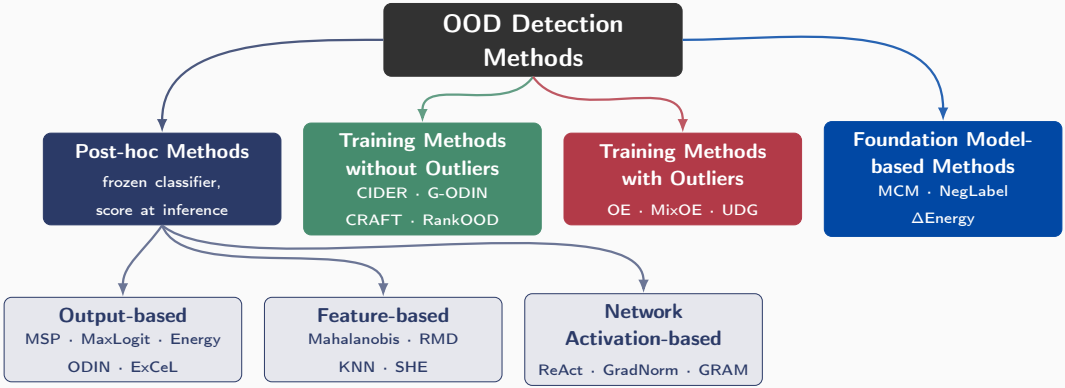
Where This Leaves Us

Robustness against **semantic shift** (detect & reject) **and** robustness against **covariate shift** (generalize) simultaneously is an open problem.

The core obstacle: both shifts increase a model's uncertainty; a single scalar score cannot distinguish them in general. **A glimmer of hope:** foundation models may implicitly disentangle the two by keeping covariate-shifted instances near their class cluster.

- Part II (next) covers the **semantic** branch in depth: post-hoc, training-based, and foundation-model detectors.
- The ΔEnergy+EBM method (Foundation Model-based Approaches) is a concrete step toward *one* objective for both — revisited in **Open Problem 3: full-spectrum detection**.

Taxonomy at a Glance



Post-hoc Methods

Post-Hoc Methods:

(a) Output-based methods

Output-based Methods: Read the Classifier's Head

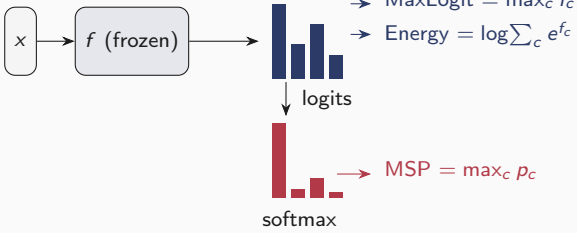
Score OOD from the output layer of a frozen classifier — the cheapest signal, nothing but a forward pass.

- **Logits** $f(x) \in \mathbb{R}^C$ (pre-softmax).
- **Softmax** $p_c = \mathrm{softmax}(f)_c$.

Three naive baselines, by *where* they read:

- MSP — Maximum Softmax Probability.
- MaxLogit — Maximum Raw Logit.
- Energy — Log-sum-exp of logits.

Then enhancements: ODIN, ExCeL.



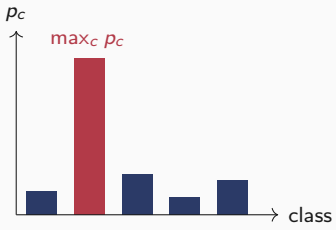
Same forward pass, three ways to read the head.

Baseline: Maximum Softmax Probability (MSP)

The original OOD baseline: use the model's top confidence.

$$S_{\text{MSP}}(x) = \max_c \; p_c = \max_c \frac{e^{f_c(x)}}{\sum_j e^{f_j(x)}}.$$

- **Intuition:** correctly-classified ID inputs get a confident prediction; OOD should be *less* confident.
- Free, single forward pass; the number every method is compared against.
- Works surprisingly well as a floor — but has a fundamental flaw (next slide).

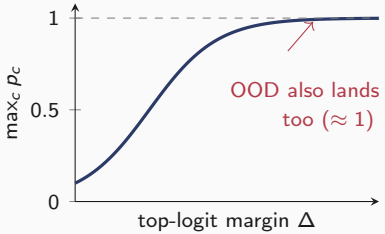


Score = height of the tallest softmax bar.

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

Why Softmax Confidence Fails: Overconfidence

- **Saturation.** exp amplifies gaps: a modest logit lead $\Rightarrow$ softmax ≈ 1. The useful range is squashed near the ceiling.
- **Normalisation discards magnitude.** Softmax depends only on logit *differences*; the overall activation level (a strong OOD cue) is thrown away.
- **ReLU extrapolation.** Piecewise-linear nets give arbitrarily high confidence far from the data (Hein et al. 2019).



Softmax saturates $\rightarrow$ small margins already read as near-certain.

$\Rightarrow$ OOD inputs routinely score $\max_c p_c \approx 1$; ID and OOD overlap near the ceiling.

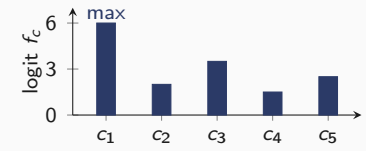
Guo, Chuan, et al. "On calibration of modern neural networks." International conference on machine learning. PMLR, 2017.

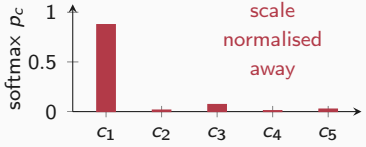
Keep the Magnitude: MaxLogit

Skip the softmax — score with the raw top logit:

$$S_{\text{MaxLogit}}(x) = \max_c f_c(x).$$

- Retains the absolute activation level that softmax normalises away — a direct OOD cue.
- Scales far better than MSP in *large-class* regimes (ImageNet-1K), where softmax saturates hardest.
- **Caveat:** logits carry class-dependent scale/bias; frequent classes can dominate (motivates KL-Matching, per-class centring).





Same vector: logits keep the absolute scale; softmax squeezes it into $[0, 1]$.

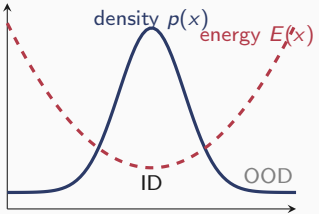
Hendrycks, Dan, et al. "Scaling Out-of-Distribution Detection for Real-World Settings." International Conference on Machine Learning. PMLR, 2022.

The Energy Score: a Density Proxy

Free energy of the logit vector (temperature T):

$$E(x) = -T \log \sum_c e^{f_c(x)/T}, \qquad S_{\mathrm{E}}(x) = -E(x).$$

- **Density link:** $p(x) \propto e^{-E(x)/T}$, so low energy $\Leftrightarrow$ high likelihood $\Leftrightarrow$ ID.
- Aggregates *all* logits (not just the top), and keeps the magnitude MSP discards.
- Same forward pass; a single, theoretically grounded scalar.



Energy is low where data is dense (ID), high in the tails (OOD).

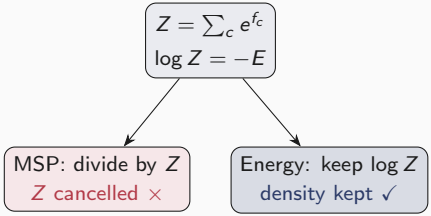
Liu, Weitang, et al. "Energy-based out-of-distribution detection." Advances in neural information processing systems 33 (2020): 21464-21475.

Why Energy Beats MSP: The Partition Function

Write the log of the max-softmax score:

$$\log S_{\mathrm{MSP}} = f_{\max} - \underbrace{\log \sum_c e^{f_c}}_{\text{log-partition} = -E\ (T=1)} = f_{\max} + E(x).$$

- Softmax normalises away the partition function $Z = \sum_c e^{f_c}$ — but $\log Z = -E$ is exactly the *density signal*.
- MSP entangles $f_{\max}$ with E and then *cancels* the very term that tracks likelihood.
- Energy uses $-E = \log Z$ directly $\Rightarrow$ cleaner ID/OOD separation, no saturation ceiling.



The normaliser Z carries the OOD information; softmax throws it out, energy keeps it.

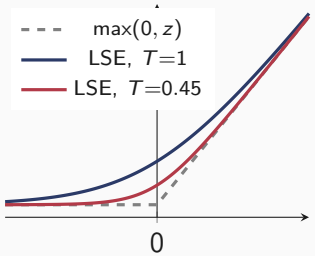
Energy is a *Smooth* Maximum of the Logits

Log-sum-exp is the smooth max (a.k.a. soft-max-value):

$$\max_c f_c \;\le\; \log \sum_c e^{f_c} \;\le\; \max_c f_c + \log C.$$

- Within $\log C$ of the true max $\Rightarrow$ Energy is a smooth surrogate for MaxLogit.
- Temperature: $-E = T \log \sum e^{f_c/T} \to \max_c f_c$ as $T \to 0$; larger T blends in *all* logits.
- Differentiable everywhere (max is not); its gradient is the softmax $\Rightarrow$ robust, uses the full logit vector.

Log-sum-exp bound: $Z \in [e^{f_{\max}},\ C\, e^{f_{\max}}]$.



logits $\{z, 0\}$: sweep z

LSE rounds the corner of max; smaller $T \to$ sharper (closer to max).

Naive Output Scores, Side by Side

	MSP	**MaxLogit**	**Energy**
Score	$\max_c \frac{e^{f_c}}{\sum_j e^{f_j}}$	$\max_c f_c$	$T \log \sum_c e^{f_c/T}$
Reads	softmax	top logit	all logits
Keeps magnitude?	no	yes	yes
Density-aligned?	no	approx.	yes ($\propto \log p$)
Saturates?	yes	no	no

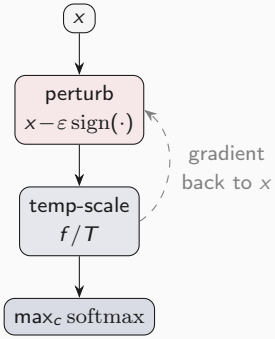
- **Rule of thumb:** Energy $\geq$ MaxLogit $\geq$ MSP for separability; all are free and post-hoc.
- MSP stays useful as a *calibrated confidence* for accepted predictions, not just OOD.
- Next: enhance these scores with temperature + input perturbation (ODIN) and logit structure (ExCeL).

ODIN: Temperature Scaling + Input Perturbation

Two post-hoc tricks that sharpen the ID/OOD gap:

1. **Temperature scaling**
   $S = \max_c \mathrm{softmax}(f/T)_c$ with large T — un-saturates softmax, exposing logit structure.
2. **Input perturbation**
   $\tilde{x} = x - \varepsilon \,\mathrm{sign}\big(-\nabla_x \log S_{\hat{y}}\big)$ — nudges the input to *raise* its confidence.

The confidence gain is larger for ID than OOD $\Rightarrow$ wider separation. Score on $\tilde{x}$.



Perturb $\rightarrow$ temperature-scale $\rightarrow$ max softmax.

Costs. One extra backprop at test; tune T, ε.

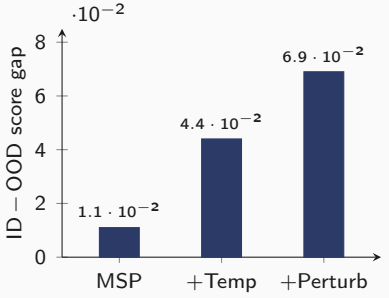
Liang, Shiyu, Yixuan Li, and R. Srikant. "Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks." International Conference on Learning Representations. 2018.

ODIN: A Toy Example

A 3-class net, two inputs with near-identical top logits:

$$f_{\mathrm{ID}} = (9,\ 4,\ 3.5), \qquad f_{\mathrm{OOD}} = (8.3,\ 4,\ 3.5).$$

stage	S_{ID}	S_{OOD}	gap
MSP (T=1)	0.989	0.979	0.011
+ Temp (T=2)	0.873	0.828	0.044
+ Perturb	0.911	0.842	0.069



Each ODIN trick widens the ID/OOD gap.

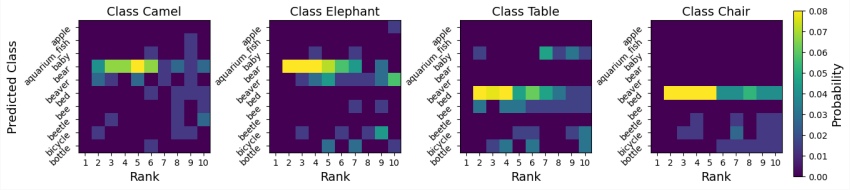
- **MSP fails:** both saturate near 1 (gap 0.011).
- **Temperature** un-saturates → logit structure resurfaces (gap ×4).
- **Perturbation** raises the ID score more than OOD ⇒ gap widens further.

ExCeL — Free Class Ranks

Observation

For an ID input predicted as class c, the *ranking of the remaining* $C-1$ *classes* is much more deterministic than for OOD inputs.

Idea: Encode rank patterns to Class Probability Matrices (CPM) and combine with MaxLogit.



Class rank signatures for top-10 ranks for four base ID classes in CIFAR100. Notably, *Bear* often ranks in the top five for *Camel* and *Elephant* classes, while *Bed* frequently appears in the top-5 for *Chair* and *Table* classes.

Karunanayake, Naveen, Suranga Seneviratne, and Sanjay Chawla. "ExCeL: Combined Extreme and Collective Logit Information for Out-of-Distribution Detection." Transactions on Machine Learning Research (2025)..

Step 1 — Build the Class Probability Matrix

For each ID class c, using ***correctly classified*** training samples:

$$P_c \in \mathbb{R}^{C \times C}, \quad p_{ij}^c = \frac{n_{ij}^c}{N_c}$$

$$P_c = \begin{pmatrix} p_{11}^c & p_{12}^c & \cdots & p_{1C}^c \\ p_{21}^c & p_{22}^c & \cdots & p_{2C}^c \\ \vdots & \vdots & \ddots & \vdots \\ p_{C1}^c & p_{C2}^c & \cdots & p_{CC}^c \end{pmatrix}$$

- **Rows:** classes, **Columns:** rank positions.
- p_{ij}^c: probability that class i appears at rank j when input is classified as c.
- By construction $p_{c1}^c = 1$ — top-1 is always c.

One CPM per class $\Rightarrow$ C matrices learned at calibration time.

Step 2 — Smooth the CPM and Compute the Rank Score

Frequencies are noisy due to DNN overfitting. Replace p_{ij}^{c} with a piecewise reward/penalty:

$$\hat{p}_{ij}^{c} = \begin{cases} \frac{a}{C-1} & p_{ij}^{c} \geq \frac{b}{C-1} \quad \text{(reward dominant)} \\ \frac{1}{C-1} & \frac{1}{C-1} \leq p_{ij}^{c} < \frac{b}{C-1} \\ -\frac{1}{C-1} & 0 < p_{ij}^{c} < \frac{1}{C-1} \\ -\frac{a}{C-1} & p_{ij}^{c} = 0 \quad \text{(penalise absent)} \end{cases}$$

Rank score of a test input x with predicted ranking $c_1, \ldots, c_C$ (top class c_1):

$$\mathcal{RS}(x) = \sum_{i=1}^{C} \hat{p}_{c_i,i}^{c_1} = \operatorname{tr}\big(\hat{P}_{c_1}^{\top} \rho_x\big),$$

where ρ_x is the one-hot rank matrix of x. Higher $\mathcal{RS}$ = better alignment with the signature.

The ExCeL Score

Combine collective $\mathcal{RS}$ with extreme MaxLogit:

ExCeL Score

$$\boxed{\text{ExCeL}(x) = \alpha \cdot \mathcal{RS}(x) + (1 - \alpha) \cdot \text{MaxLogit}(x)}$$

Properties

- Pure post-hoc — no retraining, no architectural change.
- Constant-time inference: one matrix trace per sample.

Toy Walk-Through — A 3-Class Problem (cat, dog, bus)

1. CPM for cat (rows: class, cols: rank)

$$P_{\text{cat}} = \begin{pmatrix} 1.0 & 0 & 0 \\ 0 & 0.8 & 0.2 \\ 0 & 0.2 & 0.8 \end{pmatrix} \begin{matrix} \text{cat} \\ \text{dog} \\ \text{bus} \end{matrix}$$

Dogs typically at 2_{nd} and bus at 3_{rd}.

2. Smooth (a=2, b=1.5)
$\frac{1}{C-1}$=0.5, $\frac{b}{C-1}$=0.75, $\frac{a}{C-1}$=1):

$$\hat{P}_{\text{cat}} = \begin{pmatrix} +1 & -1 & -1 \\ -1 & +1 & -0.5 \\ -1 & -0.5 & +1 \end{pmatrix}$$

Shaded = expected rank pattern for *cat*.

3. Two test inputs, both top-1 = *cat*:

	predicted ranking $\langle c_1, c_2, c_3 \rangle$	MaxLogit
x_{ID}	⟨cat, dog, bus⟩	8.0
x_{OOD}	⟨cat, bus, dog⟩	7.0

4. Rank score = sum of $\hat{P}_{\text{cat}}$ cells along the ranking:

$$\mathcal{RS}(x_{\text{ID}}) = \hat{p}^{\text{cat}}_{\text{cat},1} + \hat{p}^{\text{cat}}_{\text{dog},2} + \hat{p}^{\text{cat}}_{\text{bus},3} = 1 + 1 + 1 = 3.0$$

$$\mathcal{RS}(x_{\text{OOD}}) = \hat{p}^{\text{cat}}_{\text{cat},1} + \hat{p}^{\text{cat}}_{\text{bus},2} + \hat{p}^{\text{cat}}_{\text{dog},3} = 1 - 0.5 - 0.5 = 0.0$$

5. Combine: $\text{ExCeL} = 0.5\,\mathcal{RS} + 0.5\,\text{MaxLogit}$ ($\alpha = 0.5$)

	$\mathcal{RS}$	MaxLogit	ExCeL
x_{ID}	3.0	8.0	5.5
x_{OOD}	0.0	7.0	3.5
gap	3.0	1.0	2.0

Output-based Methods: Takeaways

1. **Softmax hides the signal.** Normalisation discards magnitude and saturates $\Rightarrow$ MSP overlaps ID/OOD; go back to logits.
2. **Energy is the principled scalar.** $-E = \log \sum_c e^{f_c}$ is a *smooth max* of the logits and tracks $\log p(x)$ — better separation, no saturation.
3. **Enhance the score, not the model.** ODIN adds temperature + input perturbation;
4. **Use logit *structure*.** ExCeL shows the *ranking* of all logits carries OOD information beyond any single number.

Cheapest family, strong baselines — but a performance ceiling that motivates feature-based and training-time methods next.

Output-based Methods: Other Notable Works

Method	Venue	Key idea
G-ODIN	CVPR'20	Decomposes confidence as $h(x)/g(x)$ and learns it end-to-end, so ODIN's temperature and perturbation need *no* OOD data to tune.
KL Matching	ICML'22	Scores by the minimum KL divergence between the softmax output and per-class mean posterior distributions — no single scalar needed.
DICE	ECCV'22	Directed sparsification: keeps only the most salient weight contributions to each logit, cutting OOD-driven noise in the output.
LogitNorm	ICML'22	Training-time fix — normalises the logit-vector norm during training to curb overconfidence, sharpening downstream MSP / Energy gaps.
GEN	CVPR'23	A generalized-entropy score over the softmax probabilities; parameter-light and drops onto any pretrained classifier.
ELogitNorm	arXiv'25	Hyperparameter-free extension of LogitNorm that fixes feature collapse, boosting many post-hoc scores (MSP, GEN, KNN, ...).
LogitGap	NeurIPS'25	Scores by the average gap between the max logit and the remaining logits, and auto-selects the most informative logit subset.

All operate on the classifier head — differing in whether they reshape the score, the weights, or the training objective.

Post-Hoc Methods:

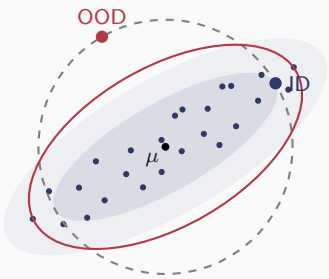
(b) Feature-based methods

Feature Distances: Why Mahalanobis, not Euclidean?

Score OOD by how far a feature $z = f(x)$ lies from the ID cluster.

$$\underbrace{d_{\mathrm{E}}(z) = \|z - \mu\|}_{\text{Euclidean}} \qquad \underbrace{d_{\mathrm{M}}(z) = \sqrt{(z-\mu)^{\top}\Sigma^{-1}(z-\mu)}}_{\text{Mahalanobis}}$$

- Euclidean treats *every direction equally* — ignores correlation and scale.
- Mahalanobis = Euclidean distance after whitening: measured in *std-devs along the data's own axes*; scale- & rotation-invariant.
- **OOD payoff:** a point can be Euclidean-close to μ yet lie off the directions the class occupies — Mahalanobis flags it (•), Euclidean misses it.



Both points are the *same* Euclidean distance from μ; only Mahalanobis separates in-distribution from OOD.

Feature-based: Mahalanobis Distance (MD)

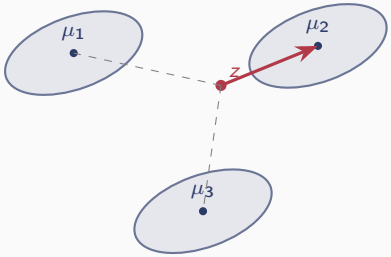
On a frozen model, fit class-conditional Gaussians with a single tied covariance (LDA assumption):

$$z \mid y{=}k \sim \mathcal{N}(\mu_k, \Sigma), \quad \mathrm{MD}_k(z) = (z-\mu_k)^\top \Sigma^{-1}(z-\mu_k).$$

Nearest-centroid score: $S_{\mathrm{MD}}(x) = -\min_k \mathrm{MD}_k\big(f(x)\big)$.

Eigenbasis view ($\Sigma = U\Lambda U^\top$):

$$\mathrm{MD}_k(z) = \sum_{i=1}^{d} \lambda_i^{-1}\big(u_i^\top (z-\mu_k)\big)^2.$$



Tied $\Sigma \Rightarrow$ every class shares the *same* ellipse; assign to nearest centroid.

- Low-variance directions get large inverse weight
- μ_k, Σ from ID train features.

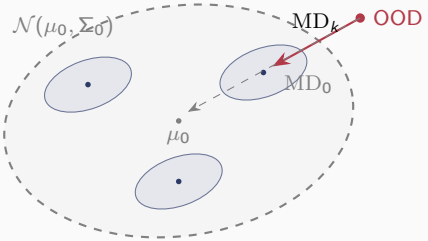
Lee, Kimin, et al. "A simple unified framework for detecting out-of-distribution samples and adversarial attacks." Advances in neural information processing systems 31 (2018).

Variant: Relative Mahalanobis Distance (RMD)

Problem. Plain MD_k is dominated by *background* structure common to all inputs $\Rightarrow$ inflates ID–OOD overlap, hurting near-OOD.

Fix. Fit one class-agnostic Gaussian $\mathcal{N}(\mu_0, \Sigma_0)$ to *all* ID features (distance MD_0) and subtract it as a global reference:

$$\mathrm{RMD}_k(z) = \mathrm{MD}_k(z) - \mathrm{MD}_0(z), \quad S_{\mathrm{RMD}} = -\min_k \mathrm{RMD}_k.$$



RMD = class distance − global-reference distance.

- Cancels the common background $\Rightarrow$ keeps the class-discriminative part; strong near-OOD gains.

features first.

Mahalanobis++ (ICML 2025) ℓ_2-normalise features to the unit sphere before fitting the Gaussians.

Ren, Jie, et al. "A simple fix to mahalanobis distance for improving near-ood detection." arXiv preprint arXiv:2106.09022 (2021).

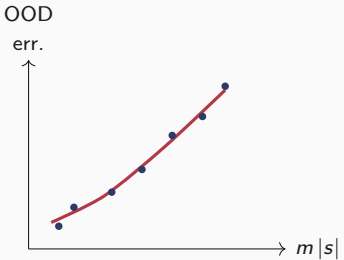
A Geometry-Based View of Mahalanobis OOD (ICML 2026)

Large study across foundation-model backbones (BEiT-v2, ViT, CLIP, EVA, DeiT; SSL + fine-tuned) $\times$ MD variants.

Finding 1 — not universally reliable. The *same* detector is excellent on one representation and fails on another. High ID accuracy does not imply good detection.

Finding 2 — an ID-only predictor. Detector quality is tracked by a two-term, *OOD-free* summary of the ID feature space:

$$m \cdot |s| \;=\; \underbrace{\text{local intrinsic dim.}}_{\#\text{ active directions}} \times \underbrace{\text{within-class slope}}_{\text{how fast variance decays}} .$$

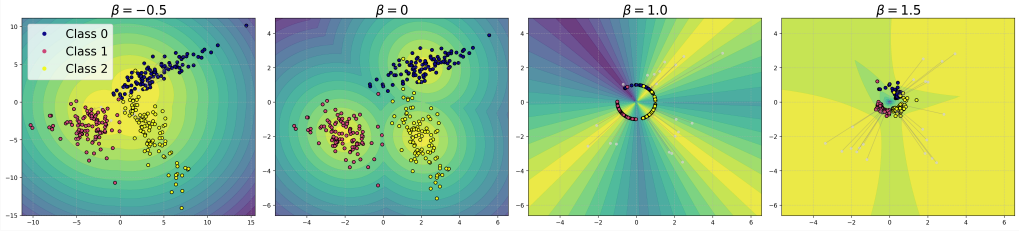


Lower $m|s|$ (compact, well-structured classes) $\Rightarrow$ better detection.

Janiak, Denis, et al. "A Geometry-Based View of Mahalanobis OOD Detection." The International Conference on Machine Learning (ICML), 2026.

Finding 3 — normalisation as *control*. Reshape feature *radii* while preserving *directions*, so the detector form is unchanged:

$$\phi_\beta(z) = \frac{z}{\|z\|^\beta}, \qquad \beta = 0 \Rightarrow \text{standard MD}, \quad \beta = 1 \Rightarrow \text{unit-sphere (MD++)}.$$



Larger β contracts norms (tighter, more localised decision regions); smaller β expands them. Applied to MD / RMD $\Rightarrow$ **RS-MD**, **RS-RMD**.

Choosing β Without OOD Data — and Results (ICML 2026)

ID-only selection. Pick β that optimises the same geometric proxy $P(\beta) = m(\beta)\,|s(\beta)|$ — *no OOD samples needed*. Approaches oracle tuning and beats fixed $\beta \in \{0, 1\}$; the optimum is model- & dataset-specific.

	MSP	MD	MD++	RMD	RS-MD	RS-RMD
Avg FPR@95 ↓	52.7	37.2	36.0	37.2	**35.3**	35.8

FPR@95 averaged over five OpenOOD datasets (lower is better).

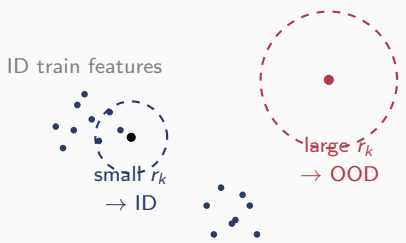
Practitioner takeaways.

- Start with RMD (or MD++) — a strong, free near-OOD upgrade over plain MD.
- Don't trust a detector to transfer across backbones — *re-check per representation*.
- If you can afford an ID-only sweep of β, RS-MD / RS-RMD squeeze out the best FPR@95, esp. on pretrained backbones.

Non-parametric: Deep Nearest Neighbours (KNN)

Idea. Drop the Gaussian assumption of MD — score *directly* by distance to the training data.

- ℓ_2-normalise features $z = f(x)/\|f(x)\|$ (crucial in practice).
- $r_k(z)$ = distance to the k-th nearest ID training embedding; score $S(x) = -r_k(z)$.
- Close ID neighbours $\Rightarrow$ ID; an *empty* neighbourhood $\Rightarrow$ OOD.



Distance to the k-th neighbour: near for ID, far for OOD.

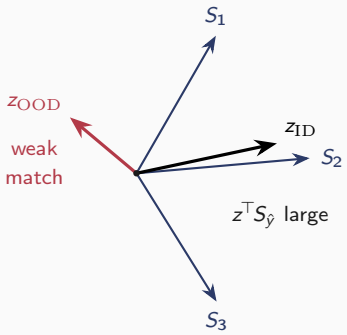
Trade-offs. No distributional assumption $\Rightarrow$ captures multi-modal / non-convex ID (unlike MD); model-agnostic (one k); cost = stored train features + NN search.

Sun, Yiyou, et al. "Out-of-distribution detection with deep nearest neighbors." International conference on machine learning. PMLR, 2022.

Associative Memory: Simplified Hopfield Energy (SHE)

Idea. *Memorise* one ID pattern per class, then score by how strongly a test feature retrieves it.

- Stored pattern (per class c): $S_c = \frac{1}{|X_c|} \sum_{x \in X_c^{\checkmark}} f(x)$ (mean of *correctly-classified* train features).
- Test feature z, predicted class $\hat{y}$; simplified modern-Hopfield energy $S_{\mathrm{SHE}}(x) = z^\top S_{\hat{y}}$.
- ID $\rightarrow$ strong match (high energy); OOD $\rightarrow$ poor match (low energy).



Energy = alignment with the predicted class's stored pattern: high for ID, low for OOD.

Per class prototype, one dot product per test sample — a template-matching view of the penultimate space.

Zhang, Jinsong, et al. "Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy." The Eleventh International Conference on Learning Representations. 2022.

Feature-based Methods: Takeaways

1. **Score the representation, not the head.** The penultimate feature space carries geometry that softmax and logits compress away — the source of feature-based gains over output-based scores.
2. **Pick the right prior on that geometry.** Parametric Gaussians (MD / RMD / MD++), non-parametric neighbours (KNN), and associative-memory templates (SHE) trade assumptions for flexibility.
3. **Near-OOD is where features win.** RMD / MD++ cancel the common *background* structure; KNN captures multi-modal ID that a single Gaussian cannot.
4. **The backbone decides.** Detector quality is set by the representation's geometry (ICML'26) — *re-check per model*, and tune β (RS-MD / RS-RMD) with ID data only.

Versus output-based: no free lunch — feature-based stores ID statistics (μ, Σ or a feature bank) and is representation-sensitive, but breaks the logit-space ceiling, especially on near-OOD.

Feature-based Methods: Other Notable Works

Method	Venue	Key idea
SSD	ICLR'21	Mahalanobis scoring on *contrastively* trained features — a strong detector using ID data only, no class labels.
ReAct	NeurIPS'21	Rectified activations: clip the penultimate features at a threshold to suppress the extreme activations that inflate OOD confidence.
ViM	CVPR'22	Virtual-logit matching: adds a logit built from the feature *residual* to the principal ID subspace, fusing feature- and logit-space evidence.
ASH	ICLR'23	Activation shaping: prune the low-percentile activations and scale the rest at test time — simple, backbone-agnostic.
fDBD	ICML'24	Scores by the feature distance to the class *decision boundaries* rather than to class centroids.
NECO	ICLR'24	Exploits *neural collapse*: measures how well a feature aligns with the collapsed ID principal subspace.
ActSub	ICCV'25	SVD of the classifier-head weights splits activations into *decisive* (near-OOD) and *insignificant* (far-OOD) subspaces, then combines both.
MahaVar	arXiv'26	Augments MD with a *class-wise distance-variance* term — ID shows a sharp-minimum, high-variance profile under neural collapse.

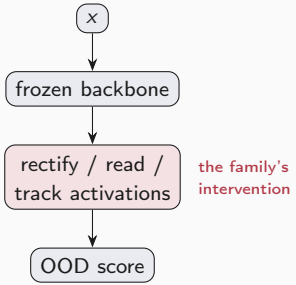
Post-Hoc Methods:

(c) Network Activation-based Methods

Network-activation Methods

Don't just *read* the output (output-based) or *model* the penultimate geometry (feature-based) — intervene on the internal signal path of a frozen network.

- OOD inputs leave internal fingerprints: a few units with runaway activations, unusually weak gradients, or out-of-range feature correlations.
- **Rectify / read / track** those internals, then feed a standard score (Energy, MSP).
- Post-hoc, no retraining; typically one extra hyperparameter or backward pass.

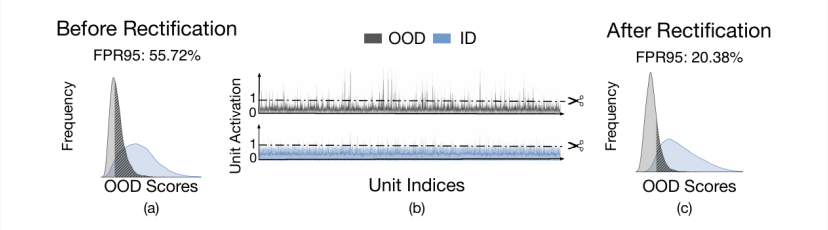


Same frozen model — the family acts *between* backbone and score.

Three representative methods are covered next: ReAct (clip activations), GradNorm (gradient magnitude), GRAM (feature correlations).

ReAct: Rectified Activations

Observation. OOD inputs drive a handful of penultimate units to abnormally high activations that dominate the logits ⇒ over-confident OOD scores. **Fix.** *Clip* the penultimate activations at a threshold c before the classifier:

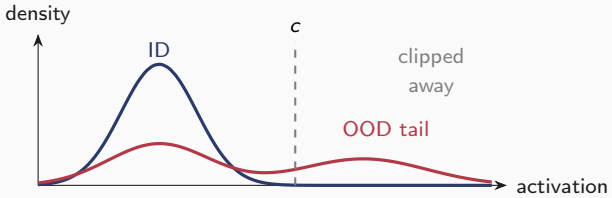


(b) Per-unit activations: the OOD tail (grey) far exceeds ID (blue); clipping at c (scissors) truncates it, turning poorly separated OOD scores (a) into well-separated ones (c) — FPR95 55.7% → 20.4% on ImageNet.

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

ReAct: Why Clipping Helps

- **ID activations are well-behaved** — concentrated, near their mean. **OOD activations have heavy positive tails** in a few units.
- Clipping removes that chaotic tail: it lowers the OOD energy much more than the ID energy $\Rightarrow$ the ID–OOD gap widens.
- Interpreted as truncating the out-of-distribution part of the per-unit activation distribution while leaving the ID bulk essentially untouched.



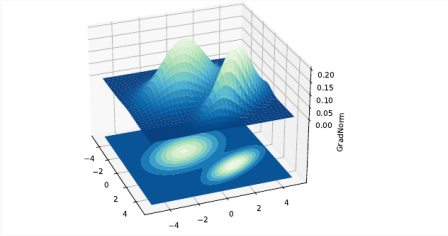
Everything past c (the OOD-heavy tail) is truncated; ID mass is preserved.

GradNorm: OOD in the Gradient Space

Idea. Score by the gradient an input *would* induce, not its activations. Push the output toward uniform and measure how hard the network resists.

$$S_{\text{GradNorm}}(x) = \left\| \frac{\partial \, \mathrm{KL}\big(\mathrm{u} \,\|\, \mathrm{softmax}(f(x))\big)}{\partial W} \right\|_1$$

- u = uniform distribution over classes; W = last-layer weights.
- ID $\Rightarrow$ confident, far from uniform $\Rightarrow$ large gradient norm.
- OOD $\Rightarrow$ near-uniform output $\Rightarrow$ small gradient norm.



Gradient norm over a 2-D input space: the peaks (ID) sit higher than the OOD region.

Fig. 1, Huang et al.

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

GradNorm: What the Score Captures

- **Two signals in one scalar.** The last-layer gradient factorises into the output residual $(\mathrm{softmax} - \mathrm{u})$ *and* the feature vector — so GradNorm blends confidence *and* feature magnitude.
- **Cheap.** A single backward pass to the classifier weights; no OOD data, no retraining.
- **Complementary.** Uses a *gradient*-space view that pure activation- or logit-space scores miss — often stacks well with them.

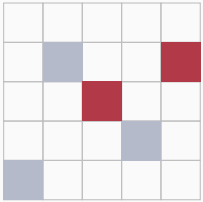
Rule of thumb: confident, feature-rich inputs ⇒ big gradients ⇒ ID; flat, weak responses ⇒ OOD.

GRAM: Feature Correlations Across Layers

Idea. Characterise internal activity by channel correlations, not single activations. At each layer l form the (order-p) Gram matrix

$$G_p^{(l)} = \left(F^{(l)\,p}\right)\left(F^{(l)\,p}\right)^{\top},$$

- On ID *train* data, record the per-class min/max range of every Gram entry.
- A test input's Gram values that fall outside their ID range signal OOD.
- Uses intermediate layers — richer than head- or penultimate-only methods.



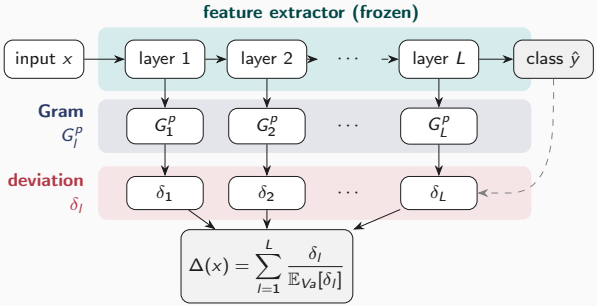
Gram matrix $G_p^{(l)}$

Red = correlations outside their ID range.

Sastry, Chandramouli Shama, and Sageev Oore. "Detecting out-of-distribution examples with Gram matrices." International Conference on Machine Learning. PMLR, 2020.

GRAM: Feature Correlations Across Layers

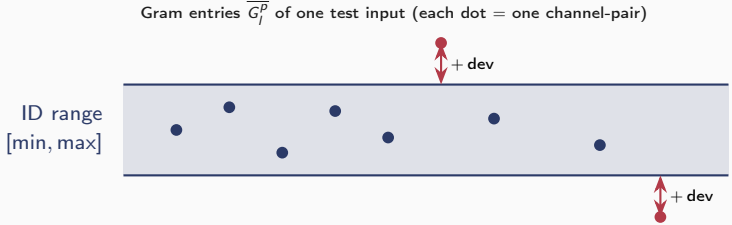
Idea. Characterise internal activity by channel correlations at *every* layer, not single activations: form the (order-p) Gram matrix $G_l^p = \left(F_l^p \, F_l^{p\top}\right)^{1/p}$. On ID *train* data, record each entry's per-class min/max range; a test input whose correlations fall outside their range is flagged OOD.



Per-layer Gram correlations → deviation from the ID range → layer-normalised sum Δ. Uses intermediate layers — richer than head- or penultimate-only scores.

GRAM: Layerwise Deviation Score

- **Deviation.** For each Gram entry, measure how far a test value lies outside its recorded ID $[\min, \max]$ band (zero if inside).
- **Aggregate.** Sum deviations over entries, orders p, and layers $\Rightarrow$ a single OOD score; larger $\Rightarrow$ more OOD.
- **Properties.** No OOD data, no retraining; captures higher-order co-activation patterns that pointwise activation methods (ReAct/ASH) cannot.



$\delta_l = \sum$(amounts outside band); inside $\Rightarrow 0$

Network-activation Methods: Takeaways

1. **Intervene, don't just observe.** Edit activations (ReAct), read gradients (GradNorm), or track correlations (GRAM) to expose OOD on the signal path.
2. **OOD leaves internal fingerprints.** Heavy activation tails, weak gradients, out-of-range correlations — each a *different* handle on the same phenomenon.
3. **Cheap and composable.** Post-hoc, no retraining; they feed a standard score (Energy/MSP) and stack with feature- and output-based methods.
4. **Static → adaptive.** Fixed thresholds (ReAct/ASH) are giving way to per-sample shaping (AdaSCALE).

Output-based reads the head; feature-based models the penultimate geometry; activation-based reshapes / inspects the path between — often the extra points on hard benchmarks, for one knob or one backward pass.

Network-activation Methods: Other Notable Works

Method	Venue	Key idea
DICE	ECCV'22	Directed sparsification: keeps only the most salient weight contributions to each logit, cutting activation-driven OOD noise.
ASH	ICLR'23	Activation shaping: prune the low-percentile activations and scale the rest at test time — simple, backbone-agnostic.
LINe	CVPR'23	Leverages important neurons: Shapley-based neuron / weight pruning plus activation clipping.
FeatureNorm	CVPR'23	Uses the *norm* of a selected block's feature map (chosen via NormRatio) as the OOD indicator, not the last block.
VRA	NeurIPS'23	Variational rectified activation: suppress abnormally low *and* high activations while amplifying the middle — generalises ReAct.
SCALE	ICLR'24	Pure percentile *scaling* of activations (no pruning); doubles as a training-time enhancer (ISH).

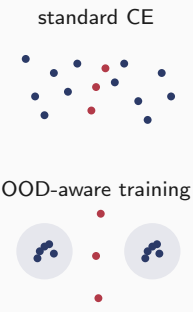
All act on internal activations or the weights that read them — clipping, scaling, sparsifying, or re-normalising the signal path.

Training Methods without Outliers

Training without Outliers: Reshape the Model, not the Score

Post-hoc methods leave the model *frozen*. Here we instead change the training objective / architecture so ID and OOD separate better — using ID data only (no auxiliary outliers).

- **Regularise the geometry** — compact, well-separated ID clusters (CIDER).
- **Restructure the confidence** — decomposed / normalised outputs (G-ODIN, LogitNorm).
- **Exploit class *ranking*** — sharpen the logit order (CRAFT, RankOOD).
- **Generative / self-supervised** — distillation and reconstruction pretexts.



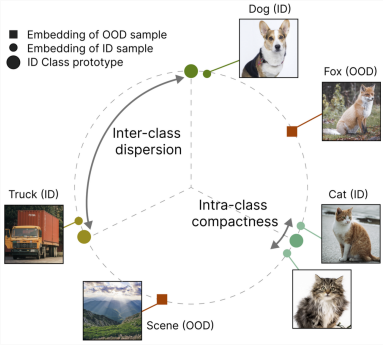
OOD overlaps ID under plain CE; OOD-aware training compacts ID ⇒ OOD stands out.

Note: G-ODIN appears under post-hoc scores too, but its decomposed head must be trained — so it lives here.

CIDER: Hyperspherical Embeddings (ICLR 2023)

Idea. Learn a representation on the unit sphere where OOD is easy to spot by *angle*. Model each class as a von Mises–Fisher distribution around a learned prototype μ_c.

- Dispersion: push class prototypes far apart (large inter-class angles).
- Compactness: pull each sample toward its own prototype (small intra-class angle).
- Large inter-class dispersion is the key driver of ID–OOD separability.



Inter-class *dispersion* + intra-class *compactness*;

OOD (Fox, Scene) lies *between* ID clusters.

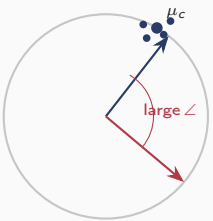
Ming, Yifei, et al. "How to exploit hyperspherical embeddings for out-of-distribution detection?" International Conference on Learning Representations (ICLR), 2023.

CIDER: Two Losses, then Distance-based Scoring

Training (ID only), on ℓ_2-normalised features $z = f(x)/\|f(x)\|$:

$$\mathcal{L}_{\text{CIDER}} = \underbrace{\mathcal{L}_{\text{comp}}}_{\text{sample}\to\text{own } \mu_c} + \lambda \underbrace{\mathcal{L}_{\text{disp}}}_{\mu_c \text{ apart}} .$$

- $\mathcal{L}_{\text{comp}}$: maximise $z^\top \mu_{c(x)}$ (vMF likelihood of the correct prototype).
- $\mathcal{L}_{\text{disp}}$: minimise the average $\mu_i^\top \mu_j$ over $i \neq j$.



Small angle to $\mu_c \Rightarrow$ ID; large $\Rightarrow$ OOD.

Test score. Non-parametric KNN distance in the embedding (or nearest-prototype angle) — no OOD data, no logits.

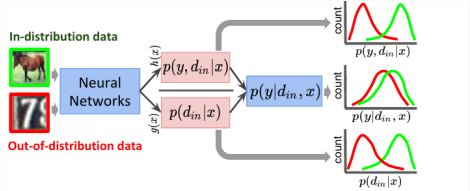
Strong near-OOD gains over plain CE + Mahalanobis; representation-learning cost is one training run.

G-ODIN: Decomposed Confidence (CVPR 2020)

From ODIN (post-hoc) to G-ODIN (trained). ODIN adds temperature + input perturbation to a frozen net but needs OOD data to *tune*. G-ODIN removes that by learning a decomposed output:

$$f_i(x) = \frac{h_i(x)}{g(x)}.$$

- $h_i(x)$ — class-conditional *dividend* ("is this class i?").
- $g(x)$ — a single *divisor* learning a confidence / domain term ("is this ID at all?").



Dividend $h(x)$ and divisor $g(x)$ heads; the OOD (red) and ID (green) score distributions separate.

Hsu, Yen-Chang, et al. "Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data." CVPR, 2020.

G-ODIN: Why the Divisor Detects OOD

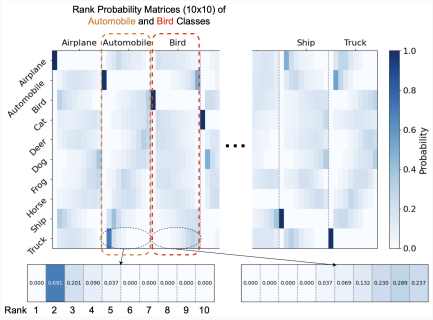
- The network is forced to explain a large logit as high dividend h_i *and* high divisor g — $g(x)$ learns to be large only for inputs that look in-distribution.
- **Score** with either $\max_i h_i(x)$ or the divisor $g(x)$ directly; add ODIN-style input preprocessing — but now no OOD data is needed to set it.
- Variants of the divisor (g via cosine / Euclidean / inner-product) trade off near- vs far-OOD.

Same trick as decomposing $p(y, x) = p(y \mid x)\, p(x)$: h carries the class posterior, g acts as a learned $p(x)$ gate — all from ID data.

CRAFT: Class-Ranking Aware Fine-Tuning (WACV 2025)

Insight (builds on ExCeL). A trained classifier implicitly learns a ranking over ID classes; OOD inputs *do not respect* it.

- **Fine-tune** the pre-trained model (ID only) to sharpen each class's rank order — widen the gaps between successively ranked logits.
- More consistent ranking makes violations by OOD stand out; score by rank consistency.
- No outliers; preserves — even improves — ID accuracy.



Rank-probability matrices: each ID class has a *consistent* rank signature that OOD disrupts.

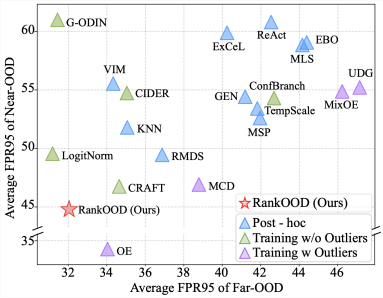
$\sim$3% average near-OOD FPR95 improvement over the next-best, with the best ID accuracy across benchmarks.

Karunanayake, Naveen, Suranga Seneviratne, and Sanjay Chawla. "CRAFT: Class Ranking Aware Fine-Tuning for Enhanced Out-of-Distribution Detection." WACV, 2025.

RankOOD: A Ranking Objective (2025)

Idea. Make the whole logit ordering the learning target, not just the top-1 label.

- Extract a canonical rank permutation π_c per class from a base model (rank-probability matrix $\rightarrow$ ILP).
- Re-train with a Plackett–Luce / ListMLE loss so the full order matches π_c: $\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha\,\mathcal{L}_{\text{ListMLE}}$.
- **Score**: penalise rank violations vs $\pi_{\hat{c}}$ (exponential penalty), compare to per-class reference. No outliers.



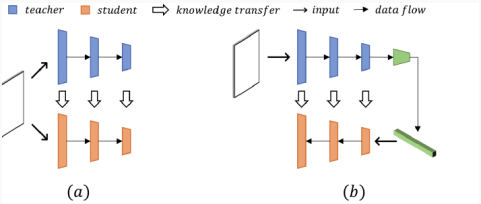
RankOOD (star) beats all post-hoc and training-w/o-outlier methods (incl. CIDER, CRAFT, LogitNorm, G-ODIN); near-OOD TinyImageNet SOTA, -4.3% FPR95.

Denipitiyage, D., Seneviratne, S., Chawla, S. "RankOOD: Class Ranking-based Out-of-Distribution Detection." CVPR, 2025.

Knowledge-Distillation Based Detection

Idea. Train a student to mimic a fixed teacher on *ID data*. On ID they agree; on OOD, they diverge.

- RD4AD (CVPR'22) — *reverse* distillation: student decoder reconstructs teacher features; cosine discrepancy flags anomalies.
- MKD (CVPR'21) — multiresolution feature distillation.
- Multi-Level KD (ACL'23) — ID-only KD for text OOD.
- Progressive Self-KD (ICML'25) — self-distillation boosts OOD sensitivity.



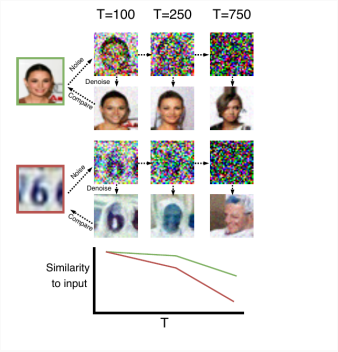
(a) conventional vs. (b) *reverse* distillation: student decoder restores teacher (encoder) features; residual discrepancy = OOD score.

Deng & Li, "Anomaly Detection via Reverse Distillation from One-Class Embedding," CVPR 2022; Salehi et al., CVPR 2021; Wu et al., ACL 2023.

Reconstruction-Based Detection

Idea. Train a generative model (AE / diffusion) on ID; at test, reconstruct the input. ID reconstructs well; OOD reconstructs poorly — reconstruction error is the score.

- DDPM-OOD (CVPRW'23) — diffusion as a denoiser; poor de-noising ⇒ OOD.
- Layer-wise Semantic Reconstruction (NeurIPS'24) — diffuse & reconstruct *multi-layer features*.
- MOODv2 — masked-image-modeling reconstruction pretext; strong features rival complex scores.



ID face reconstructs faithfully; OOD digit drifts toward ID ⇒ lower similarity (red curve).

Graham et al., CVPRW 2023; "Diffusion-based Layer-wise Semantic Reconstruction," NeurIPS 2024; "MOODv2," arXiv:2401.02611.

Training without Outliers: Takeaways

1. **Fix the cause, not the symptom.** Reshaping training with ID data only tackles overconfidence at its source — beyond what any frozen-model score can reach.
2. **Four levers.** Representation geometry (CIDER), confidence structure (G-ODIN, LogitNorm), logit ranking (CRAFT, RankOOD), and generative pretexts (KD, reconstruction).
3. **Ranking is a recurring thread.** The ExCeL insight — *the whole logit order carries OOD signal* — now drives training objectives, not just post-hoc scores.
4. **Cost vs. post-hoc.** You pay one (re)training run and give up plug-and-play, but gain the strongest near-OOD numbers short of using real outliers.

Next: training with auxiliary outliers (Outlier Exposure & synthesis) — trading the no-OOD constraint for another accuracy step.

Training without Outliers: Other Notable Works

Method	Venue	Key idea
ConfBranch	arXiv'18	Adds a learned *confidence* branch; low predicted confidence (calibrated on misclassified ID) flags OOD.
CSI	NeurIPS'20	Self-supervised contrastive learning that contrasts an input with *distribution-shifted* augmentations of itself.
VOS	ICLR'22	Synthesises *virtual* outliers from low-likelihood regions of the ID feature density to regularise the decision boundary.
LogitNorm	ICML'22	Constrains the logit-vector norm during training to curb overconfidence and sharpen downstream scores.

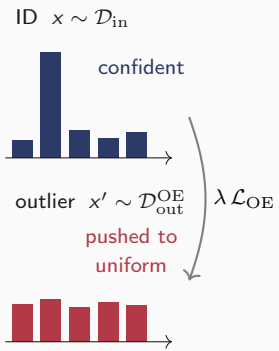
All train on ID data only (VOS synthesises its own outliers) — regularising confidence, representation, or the decision boundary.

Training Methods with Outliers

Training with Outliers: Show the Model What "Unlike ID" Means

The previous families never saw OOD. Here we bring in an auxiliary outlier set $\mathcal{D}_{\text{out}}^{\text{OE}}$ at training time — *disjoint* from the test outliers — so the model learns an explicit cue for "this is unlike my training data."

- **Outlier Exposure (OE)** — drive outputs on x' toward the uniform distribution.
- **Generalises:** heuristics learned on *seen* outliers transfer to unseen OOD at test time.
- Two ways to obtain the outliers:
  - Real outliers — a large, diverse auxiliary dataset (OE, MCD, MixOE).
  - Data generation — synthesise outliers when none are available (VOS, GAN-based).



ID outputs stay peaked; outlier outputs are driven to max-entropy ⇒ unseen OOD inherits the flat response (low $\max_c p_c$).

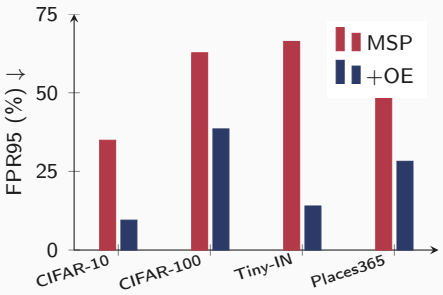
Caveat: gains hinge on $\mathcal{D}_{\text{out}}^{\text{OE}}$ being representative of test OOD — a large surrogate–test gap hurts transfer.

Outlier Exposure (OE): Fine-tune to be Uniform on Outliers

Recipe. Take a *pretrained* classifier and fine-tune (∼10 epochs) with one extra term — cross-entropy to the uniform distribution on auxiliary outliers:

$$\underbrace{\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{in}}}[-\log f_y(x)]}_{\text{keep ID accuracy}} + \lambda \underbrace{\mathbb{E}_{x'\sim\mathcal{D}_{\text{out}}^{\text{OE}}}[H(\mathcal{U}; f(x'))]}_{\text{be unsure on outliers}}$$

- Never tuned to the test OOD —unlike ODIN / GAN methods.
- Generalises to unseen anomalies (e.g. emojis, street-view letters) it never trained on.



MSP baseline vs. after OE fine-tuning. Lower is better — OE roughly halves FPR95.

- **Real & diverse > synthetic:** a diverse real dataset beats GAN outliers; *diversity* matters more than size (∼1% of 80M suffices).
- **Bonus:** also improves calibration and fixes density-estimator OOD scores (PixelCNN++).

MixOE: Mixture Outlier Exposure

Generate outliers by *mixing*. Combine an ID and an outlier sample into a *virtual* outlier that can lie near *or* far from ID:

$$\tilde{x} = \mathrm{mix}(x_{\mathrm{in}}, x_{\mathrm{out}}, \lambda), \quad \lambda \sim \mathrm{Beta}(\alpha, \alpha)$$

- mix = Mixup (linear) or CutMix (cut-paste).

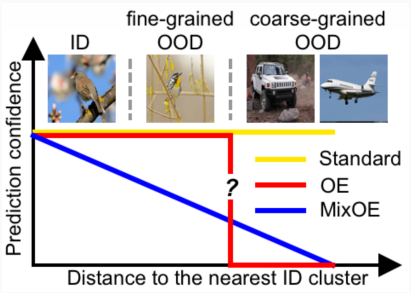
Construct the soft label. Blend the ID one-hot with the uniform dist. by the *same* λ:

$$\tilde{y} = \lambda\, y_{\mathrm{in}} + (1-\lambda)\,\mathcal{U}$$

- encoded confidence $\lambda + (1-\lambda)\frac{1}{K}$ decays linearly ID→OOD.

Loss. Add a β-weighted term on the virtual outliers:

$$\mathbb{E}_{\mathcal{D}_{\mathrm{in}}}\big[\mathcal{L}(f(x), y)\big] + \beta\, \mathbb{E}_{\mathrm{virtual}}\big[\mathcal{L}(f(\tilde{x}), \tilde{y})\big]$$



Virtual outliers span near→far OOD; their soft targets make confidence decay *linearly* along the ID→OOD continuum.

Fixing $\lambda = 0$ makes $\tilde{x} = x_{\mathrm{out}}$ and $\tilde{y} = \mathcal{U}$ — the term collapses to plain Outlier Exposure.

UDG: Outliers That Are *Found*, Not Given

The benchmark first (SC-OOD). Re-split ID/OOD by semantics, not by data source:

- ∼15% of Tiny-ImageNet is semantically CIFAR-10 (six dog breeds → *dog*).
- Methods that looked solved collapse (MCD) — they had learned *source* artefacts, not semantics.
- **UDG's answer:** Cluster labeled and unlabeled data together; let the *group*, not the sample, decide who is ID.

Group — every epoch. k-means over the features of the *whole* training set, labeled and unlabeled together:

$$G^{(t)} \leftarrow \mathrm{kmeans}\big(\{E^{(t)}(x) : x \in \mathcal{D}_L \cup \mathcal{D}_U\}\big)$$

- Stable for $K \gtrsim 500$: *small* groups stop ID and OOD from sharing a cluster.

$\pi_{k,c} > \tau$ promote ∘ to $\mathcal{D}_L$ (label = c)

$\pi_{k,c} \leq \tau$ stay in $\mathcal{D}_U$ (entropy loss)

• labeled $\mathcal{D}_L$ ∘ unlabeled $\mathcal{D}_U$.

The ID/OOD split of $\mathcal{D}_U$ is re-estimated as the representation improves.

Group purity $\pi_{k,c}$ = fraction of group k made up of *labeled* samples of class c

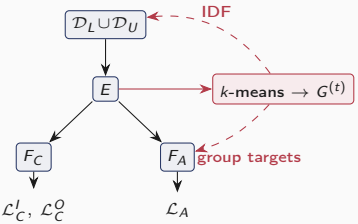
Yang, Jingkang, et al. "Semantically Coherent Out-of-Distribution Detection." ICCV, 2021.

UDG: One Clustering, Two Jobs

Filter groups, not samples (IDF). Purity of group k w.r.t. labeled class c, then promotion:

$$\pi_{k,c}^{(t)} = \frac{\left|\mathcal{D}_k^{(t)} \cap [\mathcal{D}_L]_c\right|}{\left|\mathcal{D}_k^{(t)}\right|}, \quad \mathcal{D}_L^{(t)} = \mathcal{D}_L \cup \{x \in \mathcal{D}_k^{(t)} : \pi_{k,c}^{(t)} > \tau\}$$

- $\tau = 0.8$; promoted samples take the majority label, the remainder $\mathcal{D}_U^{(t)}$ keeps the uniform-output loss.
- Rebuilt from the *original* $\mathcal{D}_L$ each epoch ⇒ error correction; safer than trusting a per-sample softmax score.



The grouping re-writes *which* samples each loss applies to.

Ablation (SC-OOD CIFAR-10):

Train — one grouping, two jobs.

$$\mathcal{L}^{(t)} = [\mathcal{L}_C^I]^{(t)} + \lambda_U [\mathcal{L}_C^O]^{(t)} + \lambda_A \mathcal{L}_A^{(t)}$$

- $[\mathcal{L}_C^I]^{(t)}$: CE on the *grown* ID set, $[\mathcal{L}_C^O]^{(t)}$: uniform outputs on the *purified* outlier set, and $\mathcal{L}_A$: predicting the group index.

Inference: only, backbone encoder E and classification head F_C run, thresholding using MSP: $\max_c p_c$

Training with Outliers: Other Notable Works

Method	Venue	Key idea
MCD	ICCV'19	*Two* classifiers on one shared encoder, both correct on ID but with different decision boundaries; training *maximises their discrepancy* on unlabeled data. The disagreement is the OOD signal.
PASCL	ICML'22	First to tackle a *long-tailed* ID set, a *partial, asymmetric* contrastive loss pulls only within each tail class and pushes tail-ID away from the outliers; an auxiliary BN/classifier branch protects ID accuracy.
DivOE	NeurIPS'23	Drops the assumption that the collected outliers already *cover* the ID/OOD boundary; a multi-step objective *extrapolates beyond* the auxiliary set to synthesise more informative outliers — a plug-in for any OE variant.
RSCL	CVPR'25	Refines *separate class learning* for long-tailed OOD: a *dynamic class-wise temperature* replaces one static value, and outliers are mined by their *affinity to head vs. tail* classes.

One OE skeleton, different pressure points: what the loss asks of outliers (MCD), whether the given outliers suffice at all (DivOE), and what breaks when the ID set is long-tailed (PASCL, RSCL).

Foundation Model-based Approaches

Foundation Model-based Approaches

The shift. Classic detectors (post-hoc, or trained with/without outliers) optimise a boundary *inside a closed ID dataset* $\mathcal{D}_{\text{in}}$. Foundation models (CLIP, LLMs) instead carry *open-world, web-scale* knowledge $\Rightarrow$ the OOD boundary is drawn in a rich *semantic* space, often without touching ID training data.

	Classic Detectors	**Foundation-model Detectors**
Knowledge	closed-world, ID-only, (or limited OE)	**open-world**, web-scale
Supervision	train / fine-tune on ID, (or limited OE)	**zero-shot**, prompt-based
Modality	single (vision)	**multi-modal** (vision+language)
OOD Signal	logits / features / activations	image–text **concept alignment**

A growing family:

MCM NegLabel CLIPN ZOC CLIPScope ΔEnergy

New capabilities & directions

- **Zero-shot, training-free** from class *names* alone (MCM).
- **"Outside" via language**: mine *negative* concepts to bound OOD (NegLabel).
- **One model, many tasks**; Works well when the ID/OOD gap is large.
- **Theory:** CLIP post-hoc scores $\approx$ PMI in an energy density model.

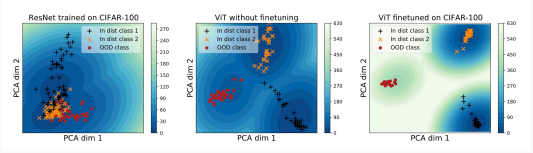
Caveat: *near*-OOD still stays difficult.

The Precursor: Pre-training Transforms OOD Detection

Thesis. Large-scale *pre-trained* transformers dramatically improve the hard regime — *near*-OOD (semantically close outliers, e.g. CIFAR-100 vs CIFAR-10) — where prior SOTA had stalled.

Two simple recipes, big gains:

- **Fine-tune a pre-trained ViT** (ImageNet-21k), then score with plain *MSP* or *Mahalanobis*: CIFAR-100 vs CIFAR-10 AUROC 85% → 96%; genomics 66% → 77% (BERT).
- **Few-shot outlier exposure:** a *handful* of outlier images suffices — 98.7% with 1 image/OOD class, 99.5% with 10.



PCA of embeddings (Mahalanobis contours). ResNet (left) lets OOD (red) overlap ID; a fine-tuned ViT (right) cleanly isolates OOD — the representation does the work.

Why this works. Pre-training yields embeddings less prone to *shortcut learning*: ID classes cluster tightly and OOD points land far away, so even a simple distance/confidence score separates them.

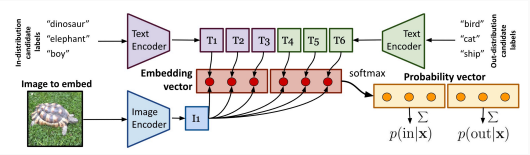
Fort, Stanislav, Jie Ren, and Balaji Lakshminarayanan. "Exploring the limits of out-of-distribution detection." Advances in neural information processing systems 34 (2021): 7068-7081.

Zero-shot OOD with CLIP: outlier *names* as the only signal

First multi-modal OOD detector. For an image x, score it against *two* sets of candidate text labels — ID names and *OOD* names (e.g. CIFAR-100 vs CIFAR-10 class names) — using CLIP image–text similarity z, then softmax:

$$\mathrm{p} = \mathrm{softmax}(\mathrm{z}), \qquad \boxed{\mathrm{score}(\mathrm{x}) = p(\mathrm{in} \mid \mathrm{x}) = \sum_{i \in \mathrm{ID}} p_i}$$

High ID-probability mass ⇒ ID; mass on OOD names ⇒ OOD. No fine-tuning, no images — just class *names*.



ID + OOD candidate names → text encoder; image → image encoder; softmax over all similarities; sum the ID-set probabilities as the score.

Results (zero-shot).

- Near-OOD CIFAR-100 vs CIFAR-10: AUROC 94.7% — beats the prior fully-trained SOTA.
- Far-OOD CIFAR-{100,10} vs SVHN: 99.6/99.9%.

Caveat = "weak outlier exposure": it needs the *OOD class names* up front. Exactly the gap MCM closes (ID-only, OOD-agnostic); and NegLabel later *mines* those negatives automatically.

Maximum Concept Matching (MCM)

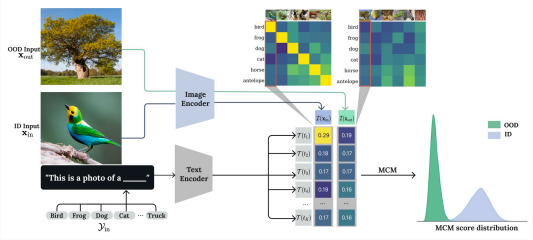
Setup. A vision–language model (CLIP) gives an image encoder $\mathcal{I}\colon \mathcal{X} \to \mathbb{R}^d$ and a text encoder $\mathcal{T}\colon \text{text} \to \mathbb{R}^d$, pre-trained to *align* matching image–text pairs. No training on ID data $\Rightarrow$ zero-shot.

Concepts as prototypes. For ID labels $\mathcal{Y}_{\text{in}} = \{y_1, \ldots, y_K\}$, wrap each in a prompt ("`a photo of a` $\langle y_i \rangle$") and embed it: the text vectors $\mathcal{T}(t_i)$ are the concept prototypes.

Score. Match an image to its nearest concept via cosine similarity $s_i(\mathrm{x}') = \dfrac{\mathcal{I}(\mathrm{x}') \cdot \mathcal{T}(t_i)}{\|\mathcal{I}(\mathrm{x}')\| \, \|\mathcal{T}(t_i)\|}$, then take a temperature-scaled softmax-max:

$$S_{\text{MCM}}(\mathrm{x}') = \max_i \frac{e^{s_i(\mathrm{x}')/\tau}}{\sum_{j=1}^{K} e^{s_j(\mathrm{x}')/\tau}}$$

ID images align strongly with one concept ($\uparrow S$); OOD aligns with none ($\downarrow S$).



Distance of the visual feature to the closest ID concept characterises OOD uncertainty.

Ming, Yifei, et al. "Delving into out-of-distribution detection with vision-language representations." Advances in Neural Information Processing Systems (NeurIPS) 35 (2022): 35087-35102.

MCM: Why the Softmax Scaling Works?

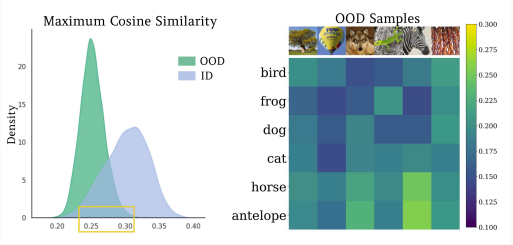
Key observation. CLIP uses a *contrastive* (not cross-entropy) loss $\Rightarrow$ for an OOD input the cosine scores are nearly uniform across concepts, while an ID input has one dominant peak.

Softmax as a magnifier. Softmax *sharpens* this contrast and pushes the ID / OOD score distributions apart — the opposite of CE-trained models, where softmax breeds overconfidence.

Formal guarantee. If OOD inputs are δ-uniform, $\frac{1}{K-1}\sum_{i\neq\hat{y}}[s_{\hat{y}_2}(x) - s_i(x)] < \delta$, then $\exists\, T$ such that for all $\tau > T$:

$$\boxed{\mathrm{FPR}(\tau, \lambda) \;\leq\; \mathrm{FPR}^{\mathrm{wo}}(\lambda^{\mathrm{wo}})}$$

a moderate τ *provably* lowers the false-positive rate.



Left: max cosine similarity — ID (blue) vs OOD (green) overlap (yellow box). **Right:** for OOD inputs the similarity to every ID concept is roughly *uniform*.

Results (ImageNet-1k ID). MCM (CLIP-L) = 91.5% AUROC *zero-shot*, beating fine-tuned baselines; +13.1% over visual-only features on hard OOD.

NegLabel: Guiding OOD Detection with *Negative* Labels

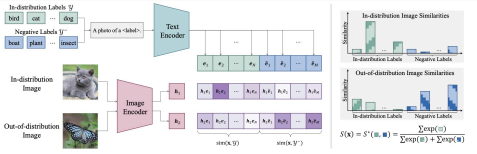
Idea. MCM uses only the K ID concepts. NegLabel also feeds the text encoder a large bank of negative labels $\mathcal{Y}^-$ (words mined from WordNet) naming concepts *far* from any ID class — an OOD image matches these negatives, an ID image does not.

NegMining ($\mathcal{Y}^-$). Score each candidate $\tilde{y}_i$ by its *distance* to the ID label set; keep the farthest $M \approx 10$k:

$$d_i = \text{percentile}_\eta\big(\{-\cos(\tilde{\mathrm{e}}_i, \mathrm{e}_k)\}_{k=1}^{K}\big), \quad \mathcal{Y}^- = \text{top}_M(d_i)$$

NegLabel score over $\mathcal{Y}^{\text{ext}} = \mathcal{Y} \cup \mathcal{Y}^-$ — ID affinity vs. negative affinity:

$$S(\mathrm{x}) = \frac{\sum_{i=1}^{K} e^{\cos(\mathrm{h}, \mathrm{e}_i)/\tau}}{\sum_{i=1}^{K} e^{\cos(\mathrm{h}, \mathrm{e}_i)/\tau} + \sum_{j=1}^{M} e^{\cos(\mathrm{h}, \tilde{\mathrm{e}}_j)/\tau}}$$



ID labels (green) vs mined negatives (blue): OOD images put more similarity mass on the negative labels, lowering $S(\mathrm{x})$.

Jiang, Xue, et al. "Negative Label Guided OOD Detection with Pretrained Vision-Language Models." The Twelfth International Conference on Learning Representations.

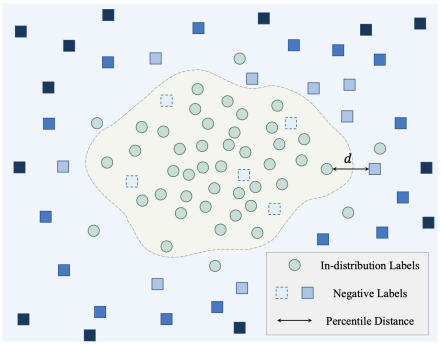
NegLabel: Why Negatives Help, and How Well

NegMining intuition. Pick negatives *far* from ID (large percentile distance d) so ID images stay dissimilar to them while OOD images do not — maximising the affinity gap that drives $S(x)$.

A multi-label view. Treat the M negatives as an M-way classification. The positive count $c = \sum_i \mathbb{1}[s_i \geq \psi]$ is Binomial, $c^{\text{in}} \sim B(M, p_1)$, $c^{\text{out}} \sim B(M, p_2)$ with $p_1 < p_2$ (OOD matches more negatives). Then

$$\frac{\partial \, \text{FPR}_\lambda}{\partial M} < 0.$$

i.e. *adding more negative labels provably lowers FPR* — until the corpus runs out of truly-distant concepts.



NegMining: negatives (squares) with the *largest* percentile distance d from the ID cluster (green) are selected; words too close to ID are discarded.

Results (ImageNet-1k ID, CLIP-B/16). 94.2% AUROC, 25.4% FPR95 (avg. over iNaturalist, SUN, Places, Textures) — new zero-shot SOTA, beating MCM (90.8/43.9) and CLIPN (93.1/31.1), and even some *fine-tuned* methods. Robust across VLM backbones and domain shifts.

ΔEnergy: OOD Detection from the *Energy Change*

Setting. Full-spectrum OOD: separate closed-set ID, *covariate*-shifted ID, and *semantic*-shifted open-set OOD. Start from the energy score over the K ID concepts, $E_0(\mathrm{x}) = -\log \sum_{j=1}^{K} e^{s_j(\mathrm{x})/\tau}$.

Key observation. Re-align vision & language by cropping the top-c cosine similarities to zero ($\tilde{s}_{\hat{y}_j} = 0$). The induced energy change is large for ID but small for open-set OOD.

ΔEnergy score. With the re-aligned energy $E_1(\mathrm{x}) = -\frac{1}{c}\sum_{j=1}^{c} \log\left[e^{\tilde{s}_{\hat{y}_j}/\tau} + \sum_{p\neq\hat{y}_j} e^{s_p/\tau}\right]$,

$$\boxed{\Delta\mathrm{Energy}(\mathrm{x}) = E_1(\mathrm{x}) - E_0(\mathrm{x})}$$

larger ΔEnergy ⇒ ID; smaller ⇒ OOD.

Theory. vs. MCM, ΔEnergy *amplifies* the ID/OOD gap ($d_{\Delta\mathrm{E}} > d_{\mathrm{MCM}}$) and provably lowers FPR.

Reset the top-c similarities one at a time, compute each energy change, and sum into ΔEnergy. Still zero-shot, post-hoc — no training.

Zhu, Lin, et al. "ΔEnergy: Optimizing Energy Change During Vision-Language Alignment Improves both OOD Detection and OOD Generalization." Advances in Neural Information Processing Systems 38 (2026): 80353-80386.

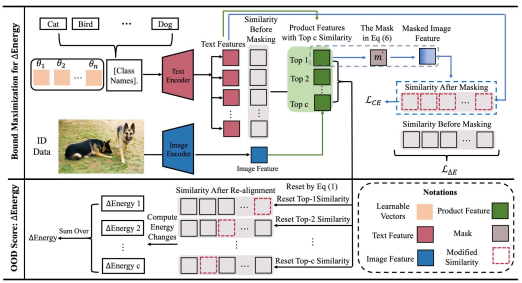
ΔEnergy + EBM: One Objective, Detection *and* Generalization

From score to objective. ΔEnergy can also be *maximized* during few-shot prompt-tuning (encoders frozen, only prompts θ learnable) via $\mathcal{L}_{\Delta E} = \frac{1}{N}\sum_i [E_2(\mathrm{x}_i) - E_0(\mathrm{x}_i)]$ (E_2: top-p *masked* product feature):

$$\boxed{\mathcal{L}_{\text{EBM}} = \mathcal{L}_{\text{CE}} + \lambda_0\, e^{\mathcal{L}_{\Delta E}}}$$

Why one loss helps both:

- **Detection:** raises the *lower bound* of ΔEnergy (Thm 3.4) ⇒ sharper ID/OOD gap.
- **Generalization:** domain-consistent Hessians bound the covariate-shift gap $|\hat{\mathcal{E}}_{\mathcal{T}} - \hat{\mathcal{E}}_{\mathcal{S}}|$ (Thm 3.5).



EBM: mask the top-p product-feature elements, then optimize the energy gap ($\mathcal{L}_{\Delta E}$) jointly with classification ($\mathcal{L}_{\text{CE}}$) over learnable prompts.

Results. Zero-shot ΔEnergy tops FS-OOD on ImageNet-1k (AUROC 87.1/88.8); EBM leads tuning-based methods (ID acc **84.5**, AUROC **81.9**) — +10–25% AUROC over recent work.

Foundation Models for OOD: The Wider Landscape

Beyond MCM, NegLabel, ΔEnergy — a fast-growing family pushing the same idea:

Method (venue)	**Key idea**	**Pro**	**Result highlight**
ZOC AAAI'22	Train a captioner to *generate* candidate OOD labels	First zero-shot VLM detector	Good on CIFAR; weak at scale (∼81.8)
CLIPN ICCV'23	Teach CLIP to "say no": extra *negative* prompt encoder	Explicit "no-match" signal	IN-1k AUROC **93.1** / FPR **31.1**
GL-MCM IJCV'25	Add a *local* patch-level score to global MCM	Training-free; localized OOD	Beats MCM (IN-1k, multi-object)
LoCoOp NeurIPS'23	Few-shot; ID *background* regions as OOD surrogates	Few-shot (1–16), no real OOD	16-shot FPR **29.5** (+GL-MCM)
CLIPScope WACV'25	*Bayesian* posterior; normalize by class likelihood	Training-free; debiases classes	AUROC ↑∼**96**, FPR ↓∼6.5
DualCnst NeurIPS'25	Text–image *dual consistency*: also match synthesized images	Training-free; adds visual cues	FPR ↓3.9 far / 9.9 robust

Two recurring ideas: (1) *enrich the concept space* (negatives, captions, Bayesian priors); (2) *add visual locality/consistency* (local features, synthesized images).

Concluding Theory: From Information-Theoretic Lens

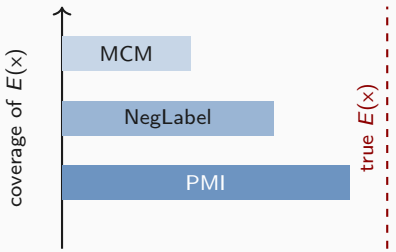
How to theoretically justify the empirical effectiveness of CLIP-based post-hoc OOD detection?

Model ID as an energy-based density built from the point-wise mutual information (PMI) of image x and label y:

$$\mathcal{W}(\mathrm{x}, y) = \log \frac{p(y \mid \mathrm{x})}{p(y)}, \quad E(\mathrm{x}) = \tfrac{1}{\alpha} \log \sum_{y \in \mathcal{Y}_{\mathbf{I}}} e^{\alpha \mathcal{W}(\mathrm{x}, y)}$$

with $p_{\mathrm{ID}}(\mathrm{x}) \propto e^{E(\mathrm{x})}$. OOD *diverges* from ID in density $\Rightarrow$ ID energy is the ideal discriminator.

Popular scores are all MC estimators of this PMI/Energy: a) **MCM** = max over K ID concepts: recovers $E(\mathrm{x})$ up to a constant, b) **NegLabel** = adds L negatives and raises the estimation upper bound, cutting underestimation bias.



Longer reach = tighter upper bound: $\log K < \log(K+L/T) < \log(K+L)$. A *better OOD score* is a *better PMI estimator*.

Peng, Bo, et al. "An information-theoretical framework for understanding out-of-distribution detection with pretrained vision-language models." Advances in Neural Information Processing Systems 38 (2026): 1930-1957.

From Theory to Method: Divide & Conquer the PMI

Divide & conquer. Rather than estimate $\mathcal{W}$ in one shot, decompose it with the chain rule of PMI using a sub-view $\tilde{x} = \mathcal{T}(x)$ (e.g. a random crop):

$$\boxed{\mathcal{W}(x, y) = \mathcal{W}(\tilde{x}, y) + \mathcal{W}(x, y \mid \tilde{x})}$$

Easier sub-tasks that *provably raise the estimation upper bound* ($\log(K+L)$ nats) without adding more negatives; averaging over sub-views also cuts variance (Thm 3–4).

Payoff. New zero-shot SOTA on ImageNet-1k — beating even most *fine-tuned* methods, with *no* ID training.

Zero-shot OOD on ImageNet-1k (avg.)

Method	AUROC↑	FPR95↓
MCM	90.8	43.9
NegLabel	94.2	25.4
PMI	**96.0**	**22.4**

Each step — max → negatives → sub-views — tightens the PMI estimate and the score.

Foundation Models for OOD: Takeaways

- **One lens.** CLIP-based post-hoc OOD detection $\approx$ estimating the ID density via *point-wise mutual information / energy*.
- **Two levers.** *Widen coverage* — concepts, negatives, sub-views $\rightarrow$ raise the estimation bound; and *reduce variance* — grouping, averaging.
- **A single design space.** Fort et al. (outlier *names*), **MCM** (max-concept), **NegLabel** (mined negatives), Δ**Energy** (energy change), and the CLIPN/GL-MCM/LoCoOp/ CLIPScope/DualCnst family are all points within it.
- **What foundation models changed.** Zero-shot, OOD-agnostic, training-free detection that scales — one frozen model across tasks. closing the theory$\leftrightarrow$practice gap.

Foundation models turned OOD detection from a closed-world, per-task problem into open-world reasoning over a shared semantic space.

Applications

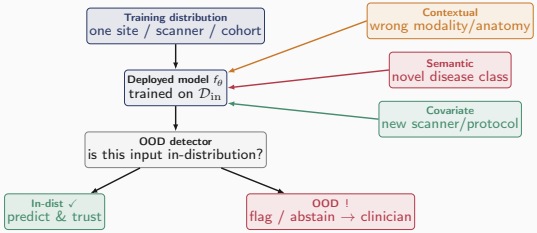
Medical Image Analysis

Medical Imaging and OOD

- Consequence is a misdiagnosis, not a mislabelled photo: errors carry clinical risk.
- Real deployments see a long tail: new scanners, protocols, populations, and rare diseases never in the training set.

Three kinds of shifts the model can meet

- **Contextual** — wrong input (e.g., CT into an X-ray model)
- **Semantic** — a novel class: disease / lesion type unseen in training labels.
- **Covariate** — same task, shifted appearance: scanner, acquisition, quality, patient cohort.



Hong, Zesheng, et al. "Out-of-distribution detection in medical image analysis: A survey." arXiv preprint arXiv:2404.18279 (2024).

Cybersecurity: Adversary Writes the Test Set

Intrusion and malware detection is the original open-world problem: the signal worth catching is precisely the traffic no one has labelled yet.

Why OOD is unavoidable here

- Detection means spotting the never-before-seen — zero-day malware, novel intrusions, fresh evasion — not sorting among known classes.
- The threat distribution drifts adversarially and fast: a model frozen at training time decays within weeks (concept drift).
- Base rates are extreme and one miss is a breach, so a calibrated "*I don't recognise this*" beats raw accuracy.

Key papers

- Sommer & Paxson, *Outside the Closed World* (IEEE S&P 2010) — why ML-based IDS must treat novelty as first-class.
- *Transcend* (Jordaney et al., USENIX Sec. 2017) — statistical concept-drift detection for malware.
- Data: **EMBER** (malware), **CIC-IDS2017** (network).

Autonomous Driving

Here OOD detection is a **safety requirement, not a metric**: an unrecognised obstacle becomes a control decision made with false confidence.

- The road is an open set — lost cargo, debris, wildlife, unusual vehicles, roadworks — the long tail no driving log ever covers.
- This spawned its own subproblem: anomaly segmentation, asking per-pixel "is *anything* here unknown?", distinct from image-level OOD scoring.
- Correct behaviour on an unknown is to *reduce speed or hand back control*, so a good uncertainty estimate is worth more than a confident guess.

Benchmarks that shaped the field: **Fishyscapes** (Blum et al., IJCV 2021) and **SegmentMeIfYouCan** (Chan et al., NeurIPS 2021) for road-anomaly segmentation; **StreetHazards / CAOS** (Hendrycks et al., ICML 2022) for scaling OOD to real driving scenes.

Drug Discovery: Prediction Beyond the Known Chemistry

Virtual screening *deliberately* explores novel chemical space, so the most valuable predictions are the ones furthest from the training data. This is extrapolation, not interpolation — an OOD problem baked into the goal.

- Cheminformatics has framed this for decades as the applicability domain: is a molecule inside the region where the model can be trusted? (An OOD question by another name.)
- A confident but wrong extrapolation is expensive — it costs failed synthesis and assays, so flagging *"outside my domain"* guides where to spend experiments.
- Honest evaluation uses scaffold / assay splits, not random splits, to actually expose OOD generalisation.

Benchmarks: **DrugOOD** (Ji et al., 2022), an OOD benchmark for AI-aided drug discovery; **GOOD** (Gui et al., NeurIPS 2022) for graph OOD; **MoleculeNet** (Wu et al., Chem. Sci. 2018).

Large Language Models: Hallucination as an OOD Symptom

When a prompt falls outside what the model reliably knows, it does not fail loudly — it returns a fluent, confident, wrong answer. Deciding where that "outside" begins is fundamentally an uncertainty / OOD question, and it is where safety and alignment are weakest.

How the "outside" gets estimated in practice

- **Self-assessment** — P(True) / verbalised confidence: *Language Models (Mostly) Know What They Know* (Kadavath et al., 2022).
- **Consistency across samples** — *SelfCheckGPT* (Manakul et al., EMNLP 2023); semantic entropy (Farquhar et al., *Nature* 2024).

Adjacent frontier: jailbreaks are *adversarial OOD prompts* — reformulations engineered to leave the region where alignment training holds. Survey: Ji et al., *Hallucination in NLG*, ACM Comput. Surv. 2023.

Fraud & Adversarial Machine Learning

Fraud models are trained on *yesterday's* fraud; the adversary's entire job is to produce *tomorrow's* — which is out-of-distribution by construction.

What makes it an OOD problem

- Non-stationary + adversarial: schemes mutate continuously, so concept drift is the steady state, not the exception.
- Extreme imbalance (fraud $\ll 1\%$): "flag the unfamiliar" (anomaly / novelty scoring) often beats supervised classification.
- Cost is asymmetric — a missed novel scheme is direct loss, so the system should *hold and escalate*, not silently approve.

Data & work

- ULB **Credit-Card Fraud** dataset + Dal Pozzolo et al. (IEEE TNNLS 2018) — realistic modelling under drift.
- **Elliptic** Bitcoin AML graph (Weber et al., 2019); **IEEE-CIS Fraud** (Kaggle 2019).

Remote Sensing

A model trained on a handful of sensors, regions and seasons is expected to work globally. Here the dominant challenge is not novel classes but covariate shift — it is the normal operating condition, not an edge case.

Why OOD generalisation is the core question

- New satellites, atmospheres, geographies and seasons appear continually; "same class, different look" is the rule.
- The operational question is often "can I trust this model *here*?" — generalisation across space and time.
- Outputs feed policy and disaster response, so a silently wrong map has real-world cost.

Benchmarks

- **WILDS** (Koh et al., ICML 2021), incl. **fMoW-WILDS** — the canonical in-the-wild shift benchmark.
- **Functional Map of the World** (Christie et al., CVPR 2018); **BigEarthNet** (Sumbul et al., 2019).

Predictive Maintenance

Training data is overwhelmingly "healthy"; the faults that matter are rare and often *never seen*. So the task is naturally posed as one-class / novelty detection — model normal, and flag anything that isn't — rather than fault classification.

- A missed novelty is unplanned downtime or a safety incident: the alarm that *should* fire simply doesn't.
- The same "learn only the good parts" framing drives industrial visual inspection — detect the defective unit you have no examples of — one of the most deployed OOD use cases today.

Reference benchmarks: **MVTec AD** (Bergmann et al., CVPR 2019), the standard for industrial anomaly detection, with *PatchCore* (Roth et al., CVPR 2022) as a strong baseline; **NASA C-MAPSS** turbofan data for degradation / remaining-useful-life.

Open Problems & Future Directions

1 Near-OOD Performance

FPR@TPR95 — the fraction of OOD inputs wrongly accepted while keeping 95% of ID data — is what deployment feels; AUROC can look great while it stays bad. A random detector scores $\approx 95\%$.

The gap that has barely closed

- Far-OOD (e.g. Textures, iNaturalist) is nearly a solved eval: FPR95 in the *tens* of percent for good detectors.
- Near-OOD (semantically adjacent classes) is not: for the same method FPR95 is routinely 2–4$\times$ worse.
- On **ImageNet-1K** hard near-OOD (**SSB-hard**), even the best post-hoc scores sit at AUROC $\approx$ 72–79%.

FPR@TPR95 — ImageNet-1K (approx.)

Regime	FPR95 $\downarrow$
Far-OOD (best methods)	$\sim$ **30%**
Near-OOD (SSB-hard)	$\sim$ **65%**
Random detector	$\sim$95%

At a "safe" 95% TPR, the detector still admits $\sim$2 of every 3 genuinely novel inputs.

2 Lack of Unified Theory

We have a zoo of scores — softmax, energy, gradients, Mahalanobis/feature distance, density — and no theory that says which to trust when. The practical question "what *is* the in-distribution?" has no agreed answer.

- Learnability: *Is OOD Detection Learnable?* (Fang et al., NeurIPS 2022) gives PAC-style impossibility and possibility results — a rare formal footing the rest of the field lacks.
- Density is not enough: deep generative models assign *higher* likelihood to some OOD than ID (Nalisnick et al., ICLR 2019) — likelihood $\neq$ typicality.
- Multimodal moves the goalposts: with CLIP, the in-distribution is defined by *text labels* (MCM, Ming et al., NeurIPS 2022), so a unified account must span pixels *and* prompts.

3 Full-spectrum OOD Detection

Real inputs shift in *two* ways at once. A deployed model should keep serving covariate-shifted-but-same-class inputs, yet reject semantically novel ones. Most benchmarks test only the second and quietly penalise the first.

The tension

- Covariate shift (new style, sensor, lighting) makes ID features *look* OOD $\Rightarrow$ detectors over-reject inputs they could still classify correctly.
- Semantic shift (new class) is what we actually want to flag.
- Optimising one hurts the other — they are usually evaluated apart.

Anchor + directions

- *Full-Spectrum OOD Detection* (Yang et al., IJCV 2023) — defines FS-OOD and the SEM benchmark.
- **Pursue:** disentangle covariate from semantic novelty; couple domain-robustness with OOD rejection; report both together.

4 Test-time Adaptation to Drift — Without Forgetting

Distributions drift continuously; models must adapt online, unlabelled, and without erasing what they knew. The current tools are promising but brittle exactly where deployment is hardest.

- The toolkit: entropy minimisation at test time (*TENT*, Wang et al., ICLR 2021); continual TTA that fights error accumulation (*CoTTA*, Wang et al., CVPR 2022).
- The failure mode: under long, mixed, or imbalanced streams TTA *collapses* — work on stable/robust TTA (*EATA*, ICML 2022; *SAR*, ICLR 2023) is a direct response.
- The OOD twist: adapting on inputs that are actually semantic OOD *corrupts* the model — adaptation and detection must be coupled.

5 Multimodal Settings: Hallucinations & Cross-modal Misalignment

Vision-language models describe objects that aren't in the image and assert attributes the pixels don't support — a grounding failure that is OOD in the *joint* image–text distribution, not either modality alone.

- How it's measured: object-hallucination probing with **POPE** (Li et al., EMNLP 2023) and the CHAIR metric; reasoning-level benchmarks (**HallusionBench**, **MMHal-Bench**).
- How it's being fixed: training-free decoding-time correction — visual contrastive decoding (*VCD*, CVPR 2024), *OPERA* (CVPR 2024); post-hoc verification (*Woodpecker*).
- Why it's OOD: the reliable signal is whether generation stays *grounded* as the input drifts from the training pairs.

Thank You!

Questions?

Slides available at: https://ood-kdd2026.nss-research.io/