International Journal of Computer Vision (2025) 133:1326–1351
https://doi.org/10.1007/s11263-024-02222-4

Dissecting Out-of-Distribution Detection and Open-Set Recognition: A Critical Analysis of Methods and Benchmarks

Hongjun Wang[1] · **Sagar Vaze**[2] · **Kai Han**[1]

Received: 16 October 2023 / Accepted: 12 August 2024 / Published online: 4 October 2024

Abstract

Detecting test-time distribution shift has emerged as a key capability for safely deployed machine learning models, with the question being tackled under various guises in recent years. In this paper, we aim to provide a consolidated view of the two largest sub-fields within the community: out-of-distribution (OOD) detection and open-set recognition (OSR). In particular, we aim to provide rigorous empirical analysis of different methods across settings and provide actionable takeaways for practitioners and researchers. Concretely, we make the following contributions: (i) We perform rigorous cross-evaluation between state-of-the-art methods in the OOD detection and OSR settings and identify a strong correlation between the performances of methods for them; (ii) We propose a new, large-scale benchmark setting which we suggest better disentangles the problem tackled by OOD detection and OSR, re-evaluating state-of-the-art OOD detection and OSR methods in this setting; (iii) We surprisingly find that the best performing method on standard benchmarks (Outlier Exposure) struggles when tested at scale, while scoring rules which are sensitive to the deep feature magnitude consistently show promise; and (iv) We conduct empirical analysis to explain these phenomena and highlight directions for future research. Code: https://github.com/Visual-AI/Dissect-OOD-OSR

Keywords Out-of-Distribution Detection · Open-set Recognition

1 Introduction

Any practical machine learning model is likely to encounter test-time samples which differ substantially from its training set; *i.e.*, models are likely to encounter test-time *distribution shift*. As such, *detecting* distribution shift has emerged as a key research problem in the community (Scheirer et al., 2013; Hendrycks & Gimpel, 2017; Liu et al., 2020). Specifically, *out-of-distribution* (OOD) detection (Hendrycks et al., 2019; Sun et al., 2021) and *open-set recognition* (OSR) (Chen et al., 2020, 2021) have emerged as two rich sub-fields to tackle this task. In fact, both tasks explicitly tackle the setting in which multi-way classifiers must detect if test samples are 'unseen' with respect to their training set, with a variety of methods and benchmarks proposed within each field. OOD detection methods test on images which come from different *datasets* to the training set, while OSR methods are evaluated on the ability to detect test images which come from different *semantic categories* to the training set. Research efforts in both of these fields largely occur independently, with little cross-pollination of ideas. Though many prior works have recognized the similarity of the two sub-fields (Vaze et al., 2022; Tran et al., 2022; Yang et al., 2024; Salehi et al., 2021), there has been little benchmarking to understand the underlying similarities and differences between them.

In this study, we investigate the detection of distribution shifts, with a focus on exploring and analyzing OOD detection and OSR methods and benchmarks. Our aim is to gain a comprehensive understanding of the underlying similarities and differences between these two tasks. We perform rigorous cross-evaluation between methods developed for OOD detection and OSR on current standard benchmarks, finding that methods which perform well for one are likely to per-

Communicated by Hong Liu.

✉ Kai Han
kaihanx@hku.hk

Hongjun Wang
hjwang@connect.hku.hk

Sagar Vaze
sagar@robots.ox.ac.uk

[1] The University of Hong Kong, Hong Kong, China

[2] University of Oxford, Oxford, UK

form well for the other (Sect. 3). We experiment both with methods which require specialized training strategies (*e.g.*, Outlier Exposure (Hendrycks et al., 2019) (OE) and ARPL (Chen et al., 2021)) as well as different post-hoc scoring rules (*e.g.*, MSP (Hendrycks & Gimpel, 2017), MLS (Vaze et al., 2022) and Energy (Liu et al., 2020)). We thoroughly evaluate all methods on both standard OOD detection and OSR benchmarks, after which we find that OE achieves almost saturating performance on the OOD detection task and also obtains state-of-the-art results on the OSR task. We further find that the scoring rules which are sensitive to the magnitude of the deep image embeddings (like MLS (Vaze et al., 2022) and Energy Scoring (Liu et al., 2020)) show the best performance across tasks and datasets.

Next, we propose a reconciling perspective on the tasks tackled by the two fields, and propose a new benchmark to assess this (Sect. 4). Specifically, we propose a new, large-scale benchmark setting, in which we disentangle different distribution shifts, namely *semantic* shift and *covariate* shift, that occur in OOD detection and OSR (see Fig. 1). Though these concepts have been discussed before (Hendrycks et al., 2021; Tian et al., 2021), the standard large-scale benchmarks in OOD detection have not adequately separated them. For example, semantic shift and covariate shift simultaneously occur when detecting OOD samples from Places using pre-trained ImageNet models. We propose a conceptual framework to understand them and further propose large-scale evaluation settings, including for pre-trained ImageNet models. For example, to isolate *semantic shift* on ImageNet, we leverage the recently introduced Semantic Shift Benchmark (SSB) (Vaze et al., 2022), in which the original ImageNet-1K (Russakovsky et al., 2015) is regarded as 'seen' closed-set data while 'unseen' data is carefully drawn from the disjoint set of ImageNet-21K-P (Ridnik et al., 2021). For *covariate shift*, we leverage ImageNet-C (Hendrycks & Dietterich, 2019) and ImageNet-R (Hendrycks et al., 2021) to demonstrate distribution shift with respect to the standard ImageNet dataset. Furthermore, to account for the tension between being *robust* to covariate shift (also known as OOD-generalisation (Sagawa et al., 2019; Ye et al., 2022)) and being able to *detect the presence of it*, we further introduce a new metric 'Outlier-Aware Accuracy' (OAA).

Finally, we examine SoTA OOD detection and OSR methods on our large-scale benchmark to validate whether the findings on the standard (small-scale) datasets still hold on our consolidated large-scale evaluation. Through large-scale analysis, we surprisingly find that OE struggles to scale to larger benchmarks, while the magnitude-aware scoring rules, especially MLS (Vaze et al., 2022), still show promise. We further provide empirical insights by analysing the representations extracted by different models under different distribution shifts. Our analysis suggest that the strong performance of OE on existing benchmarks is largely attributed to the fact that the auxiliary OOD data used for training has a distribution overlap with the OOD testing data (as measured by distances in feature space). Meanwhile, we find that it is not straightforward to find auxiliary OOD data which reflects the range of possible distribution shifts with respect to large-scale datasets. We believe there are still many open questions to be answered in the shared space of OOD detection and OSR, and hope the findings in our work can serve as a platform for future investigation.

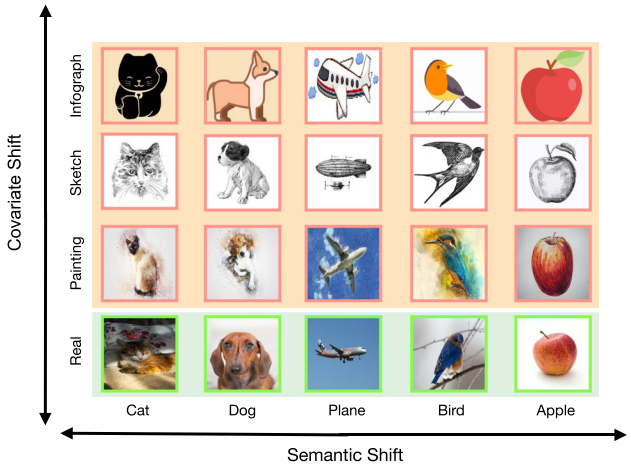

Fig. 1 Semantic shift versus covariate shift. We systematically perform cross-evaluation between SOTA methods for OSR and OOD detection and propose a large-scale benchmark setting in which we disentangle the tasks tackled in the two fields, proposing that they tackle *semantic shift* (x-axis) and *covariate shift* (y-axis) respectively

2 Related Work

2.1 Open-Set Recognition

Previous work (Scheirer et al., 2012) coins 'open-set recognition', the objective of which is to identify unknown classes while classifying the known ones. OpenMax resorts to Activation Vector (AV) and models the distribution of AVs based on the Extreme Value Theorem (EVT). Recent works (Ge et al., 2017; Neal et al., 2018; Kong & Ramanan, 2021) show that the generated data from synthetic distribution would be helpful to improve OSR. OSRCI (Neal et al., 2018) generates images belonging to the unknown classes but similar to the training data to train an open-set classifier. Kong and Ramanan (2021) adversarially trains discriminator to distinguish closed from open-set images and introduces real open-set samples for model selection. Prototype-based methods (Chen et al., 2020, 2021) (*i.e.*, ARPL / ARPL+CS) adjust the boundaries of different classes and identify open-set images based on distances to the learned prototypes of known classes. MLS (Vaze et al., 2022) uses maximum logit

scores rather than softmax scores to maintain the magnitude information.

2.2 Out of Distribution Detection

The goal of OOD detection is generally specified as identifying test-time samples coming from a 'different distribution' from the training data. Hendrycks and Gimpel (2017) formalizes the task of out-of-distribution detection and provides a paradigm to evaluate deep learning out-of-distribution detectors using the maximum softmax probability (MSP). A test sample with a large MSP score is detected as an in-distribution (ID) example rather than out-of-distribution (OOD) example. ODIN (Liang et al., 2018) and its learnable variant G-ODIN (Hsu et al., 2020) add adversarial perturbations to both ID and OOD samples and employ temperature scaling strategy on the softmax output to separate them. Liu et al. (2020) proposes the energy score derived from the logit outputs for OOD uncertainty estimation. Sun et al. (2021) rectifies the distribution of per-unit activations in the penultimate layer for ID and OOD data. GradNorm (Huang et al., 2021) calculates gradients by backpropagating the KL divergence between the softmax output and a uniform distribution, assuming that the magnitude of gradients is higher for ID data than that for OOD data. ASH (Djurisic et al., 2023) removes a large portion of the activations based on the pth-percentile of the entire representation at a late layer. The remaining activations are utilized to calculate an energy score for OOD detection. SHE (Zhang et al., 2023) quantifies the dissimilarity between the ID training samples from each category and the testing samples based on the features extracted from the penultimate layer of the model. This dissimilarity is then used as the score to judge whether a testing sample is OOD or not. Outlier Exposure (OE) (Hendrycks et al., 2019) and GradNorm (Huang et al., 2021) both design a loss based on the KL divergence between the softmax output and a uniform probability distribution to encourage models to output a uniform softmax distribution on outliers. The former leverages real OOD data for training while the latter directly employs the vector norm of gradients to perform uncertainty estimation.

2.3 Relations Between OOD Detection and OSR

Prior works discuss the separation between covariate and semantic distributional shift (Tian et al., 2021; Ahmed et al., 2020; Deecke et al., 2021; Yang et al., 2021). Tian et al. (2021) discusses separately detecting covariate and concept distributional shift on small-scale datasets (*i.e.*, CIFAR-10/100). However, similarly to Vaze et al. (2022), we suggest that small-scale datasets with no explicit taxonomies (like CIFAR) are not well suited for defining semantic shift. As such, we aim to build a large-scale benchmark with a clear underlying taxonomy. Hendrycks et al. (2021) introduces ImageNet-A (*i.e.*, collections of natural adversarial examples) and ImageNet-O (*i.e.*, samples of held-out classes from ImageNet-21K) for robustness evaluation and unseen classes recognition, while (Ahmed et al., 2020) curates a set of artificial datasets to disentangle the evaluation of non-semantic distributional shift and semantic-shift. However, they focus more on achieving robustness to non-semantic distributional shift and do not develop cross-evaluation between state-of-the-art methods in the OOD detection and OSR settings. Deecke et al. (2021) treats both semantic and non-semantic tasks in an anomaly detection (AD) paradigm and applies popular AD methods to them on CIFAR10. Our work explicitly explores the relation between OSR and OOD detection tasks, and verify the effectiveness of respective popular methods in each field. Two surveys (Yang et al., 2024; Salehi et al., 2021) summarize a number of approaches within the OOD detection and OSR settings, along with anomaly detection and Novelty Detection. Kim et al. (2021) constructs a unified benchmark to verify existing OOD detection methods, delineating 'far-OOD' and 'near-OOD'. Meanwhile, Xia and Bouganis (2022) rethinks the importance of ID misclassifications in the OOD context and examines different approaches on selective classification in the presence of OOD datasets. In our work, we not only discuss the link between robustness and OOD detection, but also propose a new metric to reconcile the tasks. Concurrent work (Yang et al., 2022) provides a codebase for representative methods within OSR and OOD detection. In this paper, we categorize shift detection methods into two types: scoring rules (*e.g.*, MSP, MLS, etc), which operate post-hoc on pre-trained networks, and specialized training, which modifies the networks' optimization procedures (*e.g.*, ARPL/ARPL+CS, OE, etc).

2.4 Auxiliary Data in OOD Detection

Inspired by OE (Hendrycks et al., 2019), recent work (Ming et al., 2022; Chen et al., 2021; Wang et al., 2023, 2024) leverage auxiliary data in some form to enhance the model's ability to detect OOD data. This could be through posterior sampling (Ming et al., 2022), adversarial training (Chen et al., 2021), augmenting distributions (Wang et al., 2023) or model perturbation (Wang et al., 2024). POEM (Ming et al., 2022) focuses on posterior sampling to learn a decision boundary between ID and OOD data. ATOM (Chen et al., 2021) introduces an adversarial training method with informative outlier mining, which is specifically designed to improve the robustness against adversarial attacks, in which the adversarial data is considered as a special type of OOD data. DAL (Wang et al., 2023) addresses the distribution discrepancy between auxiliary and unseen real OOD data, by training predictors over the worst OOD data in a Wasserstein ball. DOE (Wang et al., 2024) leverages implicit data transformation through the

Table 1 Summary of representative OOD detection and OSR techniques

	Developed for OOD or OSR	Scoring rule	Training strategy
MSP (Hendrycks & Gimpel, 2017)	OOD	MSP	CE
MLS (Vaze et al., 2022)	OSR	MLS	CE
ODIN (Liang et al., 2017)	OOD	MSP	CE
GODIN (Hsu et al., 2020)	OOD	MSP	CE
GradNorm (Huang et al., 2021)	OOD	MSP	CE
SEM (Yang et al., 2023)	OOD	MSP	CE
Energy (Liu et al., 2020)	OOD	Energy	CE
ReAct (Sun et al., 2021)	OOD	MSP/Energy/ODIN	CE
ASH (Djurisic et al., 2023)	OOD	MSP/Energy/ODIN/ReAct	CE
SHE (Zhang et al., 2023)	OOD	MSP/Energy/ODIN	CE
ARPL+CS (Chen et al., 2021)	OSR	MSP	ARPL+CS
OE (Hendrycks et al., 2019)	OOD	MSP	OE

The works are categorized based on the task that they are developed for (*i.e.*, OOD detection and OSR) and the methodology employed (*i.e.*, scoring rules and training strategies)

embedding features' perturbation to minimize a distribution discrepancy measurement called worst OOD regret, aiming to enhance the model's robustness to distribution shifts. Our work offers unique insights into the selection of auxiliary data to optimize OOD detection performance. By uncovering the relationship between the auxiliary data and the model's OOD detection performance, our work has the potential to inform strategies for auxiliary data selection and manipulation toward more reliable OOD detection solutions.

2.5 Key Similarities and Distinctions with Prior Work

While several papers (Deecke et al., 2021; Yang et al., 2022) have jointly considered methods in the OOD detection and OSR tasks, few works have clearly distinguished the academic and practical differences (or similarities) between them. In this work, we not only provide empirical analysis but also propose a *conceptual framework* and *large-scale benchmark* to better reconcile these problems.

3 Cross-Benchmarking of OOD Detection and OSR Methods

Despite the growing popularity of OOD detection and OSR studies, these two tasks have largely evolved independently and in isolation from each other, as shown in Table 1. Indeed, methods designed for OSR can be seamlessly adopted to address the OOD detection problem, and vice versa. Recent generalized OOD detection frameworks (Deecke et al., 2021; Yang et al., 2022) unify tasks relevant to OOD detection and OSR. However, there is still a lack of cross-evaluation between methods developed for OOD detection and OSR on current standard benchmarks. Given the strong inherent connections between OOD detection and OSR, a comprehensive cross-benchmarking comparison is crucial to shed light on the future development of the broader distribution shift detection problem.

As a starting point to reconcile OOD detection and OSR, in this section we perform cross-evaluation of methods from both sub-fields.

3.1 Experimental Setup

3.1.1 Problem Setting

Let $X \in \mathcal{R}^D$ denote an input sample and $C \in \mathcal{R}$ denote the label of interest. Test-time *distribution shift* occurs when the testing joint distribution is not equal to the training joint distribution, *i.e.*, $P_{test}(X, C) \neq P_{train}(X, C)$. This shift can be further divided into two types: *covariate shift* and *semantic shift*. *Covariate shift* occurs when $P_{test}(C|X) = P_{train}(C|X)$ but $P_{test}(X) \neq P_{train}(X)$. *Semantic shift* occurs when $P_{test}(C|X) \neq P_{train}(C|X)$ but $P_{test}(X) = P_{train}(X)$. In OOD detection and OSR for multi-class classification, the label space contains multiple semantic categories $\{c_1, \cdots, c_L\}$, where L is the total number of categories in the testing data. The model needs to identify the distribution from which test-time samples originate and conduct classification based on the posterior probability, represented as $p(C = c_i \mid X)$.

3.1.2 Methods

We distinguish two categories of shift detection methods: *scoring rules* (which operate post-hoc on top of pre-trained networks); and *specialized training* (which change the optimization procedure of the networks).

For *scoring rules*, we compare the maximum softmax probability (MSP) (Hendrycks & Gimpel, 2017), the Maximum Logit Score (MLS) (Vaze et al., 2022), ODIN (Liang et al., 2018), GODIN (Hsu et al., 2020), Energy scoring (Liu et al., 2020), GradNorm (Huang et al., 2021) and SEM (Yang et al., 2023). We further experiment with ReAct (Sun et al., 2021), an activation pruning technique which can be employed in conjunction with any scoring rule. While MLS was developed for OSR (Vaze et al., 2022), the other scoring rules were developed for OOD detection. For now, we note that MLS, Energy and GradNorm are all sensitive to the *magnitude of the feature norm* of the network, while the others are not. We refer to the former scoring rules as 'magnitude aware'.

For *specialized training*, we first experiment with the standard cross-entropy (CE) loss. We also use ARPL + CS (Chen et al., 2021) from the OSR literature. This method learns a set of 'reciprocal points' which are trained to be far away from all training category embeddings. We note that the reciprocal points can be treated as a linear classification layer, allowing us to use any of the scoring rules mentioned above on top of this representation. Finally, we train models with Outlier Exposure (OE) (Hendrycks et al., 2019) from the OOD detection literature, where real outlier examples are used during training as examples of OOD detection. In this case, the model is encouraged to predict a uniform softmax output.

3.1.3 Datasets

For the OOD detection setting, we treat CIFAR10 (Krizhevsky et al., 2009) as in-distribution data and train models on it. As OOD data, we use six common datasets: SVHN (Cimpoi et al., 2014), Textures (Ovadia et al., 2019), LSUN-Crop (Yu et al., 2015), LSUN-Resize (Yu et al., 2015), iSUN (Xu et al., 2015) and Places365 (Zhou et al., 2017), all of which have mutually exclusive classes on CIFAR10. In the supplementary, we also provide experiments using CIFAR-100 as the ID training data (see Tables S1–S3 in Section S2) and using different training configurations (see Tables S4–S7 in Section S3). The supplementary experiments yield consistent results, reinforcing the main findings to be discussed in this section.

For the OSR benchmark, following the standard protocols in Neal et al. (2018), we set up four sub-tasks containing CIFAR10, CIFAR+10, CIFAR+50 and TinyImageNet (Le & Yang, 2015). In all cases, models are trained on a subset of categories with remaining used as 'unseen' at test time. The CIFAR+N settings involve training on four classes from CIFAR10 and evaluating on N classes from CIFAR-100. Note that, for a given method, benchmarking on OOD detection involves training a single model and evaluating on multiple downstream datasets. In contrast, OSR benchmarks involve training a different model for each evaluation.

3.1.4 Training Configurations

We train the ResNet18 from scratch on all benchmarks. For CIFAR10, we always set the initial learning rate to 0.1 and apply the cosine annealing schedule, using SGD with the momentum of 0.9. The weight decay factor is set to $5e^{-4}$. The number of total training epochs is 200 and the batch size is 128. For CIFAR-100, we also set the initial learning rate to 0.1, which is then divided by 5 at 60th, 120th, 160th epochs. The model is trained for 200 epochs with a batch size of 128, a weight decay of $5e^{-4}$, and Nesterov momentum of 0.9, following (DeVries & Taylor, 2017). Additional results using other network architectures and training setups can be found in Section D in Appendix.

3.1.5 Metrics

Following standard practise in both OOD and OSR tasks, we use the Area Under the Receiver Operating characteristic Curve (AUROC) as an evaluation metric throughout this paper as we find that other metrics were correlated strongly with the AUROC. For results on other metrics, please refer to the supplementary (Section S5).

3.2 Quantitative Results

In Table 2, we benchmark OOD detection and OSR tasks across nine common datasets, with different training strategies and scoring rules. The results are averaged from five independent runs. Although there is not always one clear winner regarding methodology, we have three main observations.

Firstly, MLS (Vaze et al., 2022) and Energy (Liu et al., 2020) tend to perform best across OOD and OSR datasets. We hypothesize that this is because both are sensitive to the magnitude of the feature vector before the networks' classification layer. To verify our conjecture, we investigate the magnitude of features by projecting the features of both ID and OOD/open-set samples into a two-dimensional space in Fig. 2. We experiment on generic and fine-grained datasets, namely, CIFAR-10 and CUB. This projection is achieved by training a linear layer with an output dimension of two, after the penultimate layer of the model. The feature magnitude of ID data is larger than that of open-set/OOD data. This is consistent with the finding in Vaze et al. (2022) that 'unfamiliar' examples tend to have lower feature magnitude than

Table 2 Evaluation on small-scale OOD detection and OSR benchmarks with various methods, using CIFAR10 as ID

Training method	Scoring rule	OOD detection benchmarks							OSR benchmarks					Overall
		SVHN	Textures	LSUN	LSUN-R	iSUN	Places365	AVG ID=95.45	CIFAR10 ID=97.13	CIFAR+10 ID=96.6	CIFAR+50 ID=96.8	TinyImageNet ID=83.4	AVG	
(a) Evaluation based on ResNet-18 trained with the CE loss														
CE	MSP	93.65	91.35	95.49	94.88	94.33	90.77	93.41	91.78	93.81	90.20	79.82	88.90	91.61
	MLS	94.49	91.54	96.94	96.13	95.52	91.64	94.38	*92.54*	*95.62*	*91.81*	81.31	*90.32*	*92.53*
	ODIN	92.23	83.76	94.96	96.16	95.31	90.88	90.88	89.77	81.37	80.22	80.96	83.08	88.56
	GODIN	**97.60**	**96.21**	**99.59**	**97.81**	**97.74**	**94.33**	**97.21**	90.22	91.17	87.38	76.05	86.21	92.21
	SEM	75.65	72.02	75.18	70.93	72.52	76.14	73.74	40.21	43.87	42.70	-	42.26	61.15
	Energy	*94.64*	*91.64*	*97.14*	*96.29*	*95.68*	*91.78*	*94.53*	92.52	**95.68**	**91.86**	81.28	**90.34**	**92.62**
	MLS+ReAct	92.56	89.97	95.39	95.78	95.17	90.69	93.26	**92.57**	94.92	90.88	*81.65*	90.01	91.78
	ODIN+ReAct	91.29	83.50	94.70	96.05	95.19	82.55	90.55	86.65	87.76	88.40	81.30	86.03	88.74
	Energy+ReAct	92.68	90.05	95.67	96.03	95.42	90.89	93.46	92.58	95.02	90.99	**81.67**	90.07	91.92
	MLS+ASH	95.50	88.87	90.06	92.84	85.84	82.44	89.26	89.19	90.15	82.11	78.76	85.05	87.58
	MLS+SHE	86.30	76.40	84.72	81.12	80.56	81.39	81.75	79.19	74.35	78.02	78.78	77.59	80.09
(b) Evaluation based on ResNet-18 trained with the ARPL+CS loss														
ARPL+CS	MSP	93.41	91.64	94.29	94.02	94.28	90.77	93.07	92.53	95.71	94.03	*82.80*	91.27	92.41
	MLS	*96.36*	90.20	*96.59*	*96.95*	*96.88*	*93.29*	95.05	*93.16*	96.58	94.67	**84.79**	**92.30**	**93.95**
	ODIN	75.92	71.64	86.25	95.14	95.19	75.97	83.35	58.04	74.80	71.52	63.13	66.87	76.76
	GODIN	95.78	89.61	95.41	96.88	96.17	92.59	94.41	91.99	95.73	93.76	81.25	90.68	92.92
	SEM	76.42	74.26	84.45	76.08	77.73	71.23	76.70	35.01	38.27	44.15	-	39.14	64.18
	Energy	**96.52**	90.11	**96.76**	**97.16**	**97.07**	**93.45**	*95.18*	**93.22**	**96.74**	**94.82**	82.10	*91.72*	*93.80*
	MLS+ReAct	95.87	**92.37**	96.37	96.34	96.30	92.97	95.04	92.70	96.42	94.53	82.05	91.43	93.59
	ODIN+ReAct	71.87	73.36	83.19	92.34	92.36	69.10	80.37	55.71	62.88	61.85	54.29	58.68	71.70
	Energy+ReAct	96.06	*92.35*	*96.59*	96.58	96.53	93.17	**95.21**	92.80	*96.61*	*94.70*	82.14	91.56	93.75
	MLS+ASH	94.85	91.57	91.14	96.43	88.35	89.11	91.91	91.89	93.26	91.81	79.15	89.03	90.76
	MLS+SHE	83.20	81.54	84.36	88.18	83.50	82.13	83.82	77.10	74.25	74.85	75.44	75.41	80.46
(c) Evaluation based on ResNet-18 trained with the OE loss														
OE	MSP	99.21	98.81	99.02	98.52	98.55	97.29	98.57	96.29	99.29	98.70	78.67	93.24	*96.44*
	MLS	99.21	*98.82*	99.02	98.53	98.57	*97.32*	*98.58*	96.28	*99.32*	*98.72*	**80.19**	**93.63**	**96.60**
	ODIN	*99.43*	98.73	*99.14*	*98.78*	*98.75*	96.41	98.54	*96.29*	95.27	94.30	79.97	91.46	95.71
	GODIN	97.25	95.17	89.05	83.42	84.63	89.51	89.84	93.64	92.01	91.63	78.21	88.87	89.45
	SEM	98.13	97.04	98.77	97.01	97.16	94.86	97.16	30.19	33.73	33.91	-	32.61	73.69
	Energy	99.20	98.78	99.02	98.55	98.58	97.31	98.57	93.12	**99.33**	**98.74**	*80.16*	*92.84*	96.28
	GradNorm	**99.95**	**99.71**	**99.83**	**99.46**	**99.42**	**97.93**	**99.38**	**96.57**	99.26	98.51	60.56	88.73	95.12
	MLS+ReAct	95.18	92.22	79.46	83.34	83.68	87.46	86.89	95.43	98.73	97.93	79.92	93.00	89.34
	ODIN+ReAct	84.16	82.92	64.00	73.90	75.45	71.65	75.35	87.52	87.78	85.62	79.47	85.10	79.25
	Energy+ReAct	94.41	91.36	73.88	80.03	81.16	86.19	84.51	95.43	98.74	78.67	79.84	88.17	85.97
	MLS+ASH	99.10	98.55	98.88	98.52	98.53	97.31	98.49	95.40	89.21	91.54	75.30	87.86	94.24
	MLS+SHE	97.79	93.68	94.16	89.34	88.77	90.26	92.33	82.01	76.45	87.05	70.20	78.93	86.97

The results are averaged from five independent runs. We report the in-distribution accuracy as 'ID' and denote intractable results as '-', resulting from unaffordable computational cost. Bold values represent the best results, while italic values represent the second best results. Different methods have their optimal scope but MLS and Energy demonstrate their stability and models trained with OE dominate on almost all OOD datasets

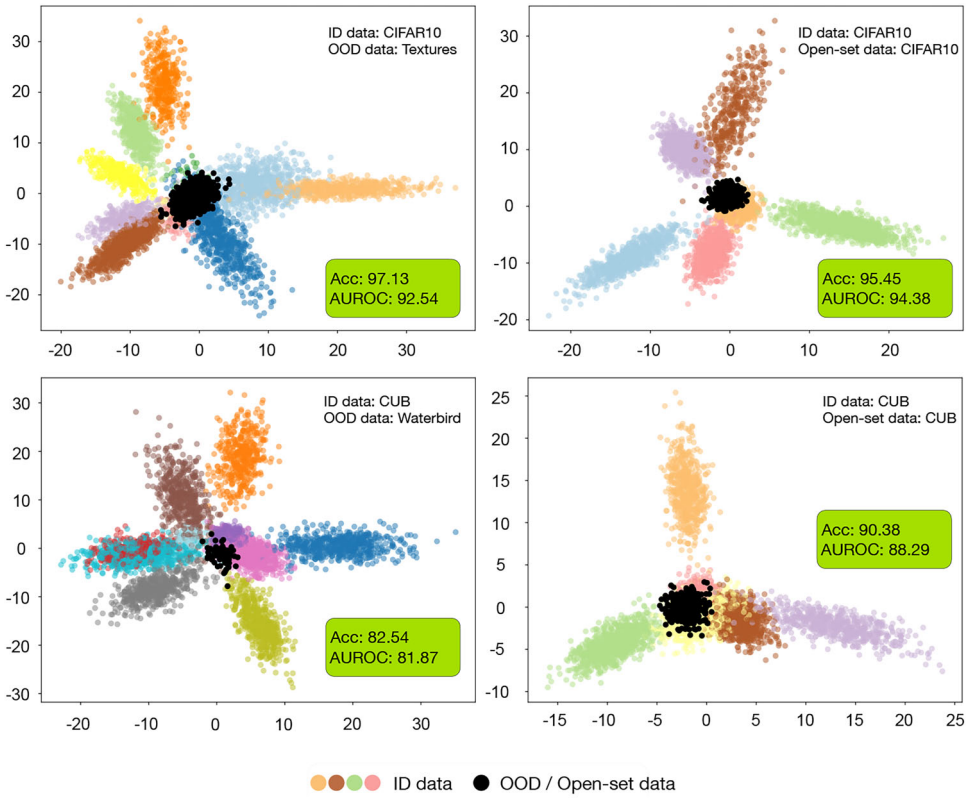

Fig. 2 Visualization of feature projections for images from ID and open-set / OOD datasets. We project the features into a two-dimensional space using an additional linear layer with an output dimension of two after the penultimate layer. We conduct OOD detection and OSR experiments using ResNet-18 on CIFAR-10 (first row) and ResNet-50 on CUB (second row) datasets. For CIFAR-10, the OOD experiment uses the full CIFAR-10 dataset as ID data and Textures as OOD data, while the OSR experiment utilizes the first six classes in CIFAR-10 as ID data and the remaining four as open-set data. For CUB, the OOD experiment employs the full CUB dataset as ID and Waterbird as OOD data, while the OSR experiment uses six classes in CUB as ID data and four CUB classes as open-set data. These classes are randomly selected from the ID and open-set splits introduced in SSB. Notably, these visualizations reveal that the feature magnitudes of ID data exceed those of OOD or OSR data

ID samples, providing a strong signal for distribution shift detection.

Secondly, we observe that Outlier Exposure (Hendrycks et al., 2019) provides excellent performance on the OOD detection benchmarks, often nearly saturating performance. More results can be found in Section A in Appendix. It also often boosts OSR performance, though to a lesser degree, a phenomenon which we explore next in Sect. 4.

Thirdly, for small-scale datasets, OOD detection accuracy is positively related to ID accuracy, while an inverse correlation is observed for large-scale datasets. In Fig. 3, we further include the results using the recent vision-language model, CLIP (Radford et al., 2021), for reference, which are not included for fitting the lines. We experiment with three variants, namely, zero-shot (zs), finetuning (ft), and linear probing (lp). We find that only the linear probing CLIP falls into the correlation fitted for other methods, while the zero-shot and finetuning counterparts are not well aligned with the trend.

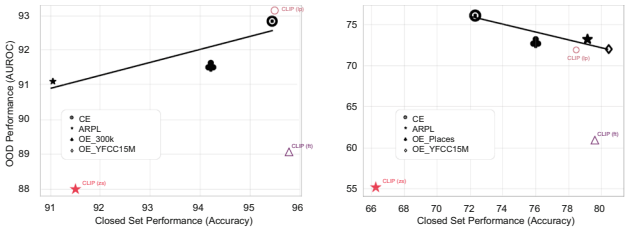

Fig. 3 OSR performance versus OOD detection performance of different training methods averaged across various scoring rules. CLIP variants are included here for reference and are not used to fit the correlation

3.2.1 Additional Analysis

(1) *Mixup hurts magnitude-aware methods' performance.* To investigate the impact of different training setups, we also adopt Mixup (Zhang et al., 2017) to our training procedure. As shown in Table 3, magnitude-aware techniques (*i.e.*, MLS and Energy) demonstrate stability. Interestingly,

Table 3 Evaluation on small-scale OOD detection and OSR benchmarks with various scoring rules on ResNet-18, using CIFAR10 as ID training data

Training method	Scoring rule	OOD detection benchmarks							OSR benchmarks				Overall
		SVHN	Textures	LSUN	LSUN-R	iSUN	Places365	AVG ID=95.81	CIFAR10 ID=97.40	CIFAR+10 ID=96.82	CIFAR+50 ID=96.91	TinyImageNet AVG ID=83.75	
CE+Mixup	MSP	87.63	*81.48*	**95.01**	88.22	*88.58*	**86.02**	**87.82**	92.01	93.96	90.44	80.02	89.11	**88.34**
	MLS	86.83	79.62	*94.95*	88.05	87.99	81.99	86.57	*92.88*	**95.99**	*92.16*	**81.55**	**90.65**	*88.17*
	ODIN	89.43	75.56	87.31	55.94	61.36	83.63	75.54	85.27	79.24	83.51	78.81	81.71	78.01
	GODIN	**94.28**	**89.71**	91.52	*88.50*	87.45	81.28	88.79	90.13	91.01	86.88	72.48	85.13	87.33
	Energy	78.87	76.46	93.21	88.09	85.69	78.53	83.48	92.79	*95.92*	92.10	*81.53*	*90.59*	86.32
	MLS+ReAct	*91.47*	77.25	93.89	**88.73**	**88.84**	*84.32*	*87.42*	90.33	95.02	**92.18**	79.42	89.24	88.15
	ODIN+ReAct	72.19	57.63	69.15	54.50	54.62	72.75	63.47	83.30	81.02	85.82	77.21	81.84	70.82
	Energy+ReAct	89.80	75.02	93.62	87.11	87.40	81.76	85.79	**93.00**	91.69	86.45	81.28	88.11	86.66

The results are averaged from five independent runs. We adopt Mixup (Zhang et al., 2017) to the training procedure and report the in-distribution accuracy as 'ID'. Bold values represent the best results, while italic values represent the second best results. Magnitude-aware techniques (*i.e.*, MLS and Energy) demonstrate stability

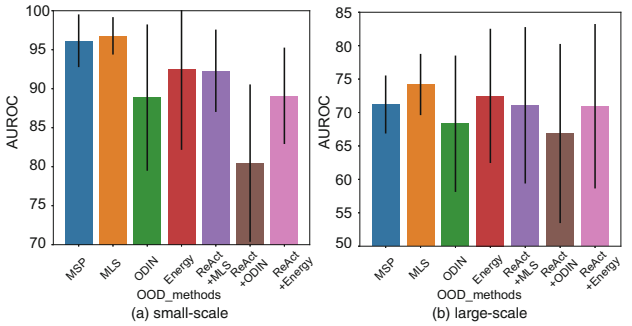

Fig. 4 OOD detection performance of various scoring rules averaged across different models. Magnitude-aware scoring rules, particularly MLS, are the most efficient and stable techniques

we also find that the MSP method achieves the best performance compared to other methods, but it is still inferior to magnitude-aware methods' performance in Table 2a. It appears that the Mixup mechanism, which distributes confidence to both involved categories, decreases the maximum magnitude value. Therefore, we highlight that the training setup (*e.g.*, Mixup) that neglects magnitude information occurs poorer performance in both OOD detection and OSR tasks. (2) *Sensitivity of methods to hyper-parameters.* In Fig. 4, we present the results of different scoring rules averaged across different training methods. For methods that are sensitive to the selection of hyperparameters (*e.g.*, the perturbation magnitude of ODIN, thresholds of ReAct), we average the results among five different selections (instead of using five different random seeds like others). Therefore, in Fig. 4, the result for each scoring rule is the average of 15 independent runs (*i.e.*, 3 training methods $\times$ 5 runs). We find that magnitude-aware scoring rules (*i.e.*, MLS and Energy) offer obvious advantages for evaluating model performance. Considering the error bars of MLS and Energy, MLS is the optimal choice for all the scenarios. Furthermore, we observe that ODIN and ReAct, two techniques that are not magnitude-aware, exhibit instability. This instability can be attributed to the reliance on a carefully tuned noise value for stochastic predictions for ODIN and threshold value to truncate activation for ReAct.

We can also find from Table 2 that ReAct, which has been shown to be effective in the literature, does not seem to bring performance gain in well-trained models with a high in-distribution accuracy. Here, we follow the techniques from (Vaze et al., 2022) to obtain the highest ID accuracy possible. It appears that when the classifier is strong enough, it is difficult for ReAct to bring extra improvement. Besides, ReAct is sensitive to the choice of the activation pruning percentile. The optimal percentile values are different for different open-set/OOD datasets (see Section S4 in the supplementary). For identifying out-of-distribution inputs, we recommend more stable and deterministic magnitude-aware scoring rules.

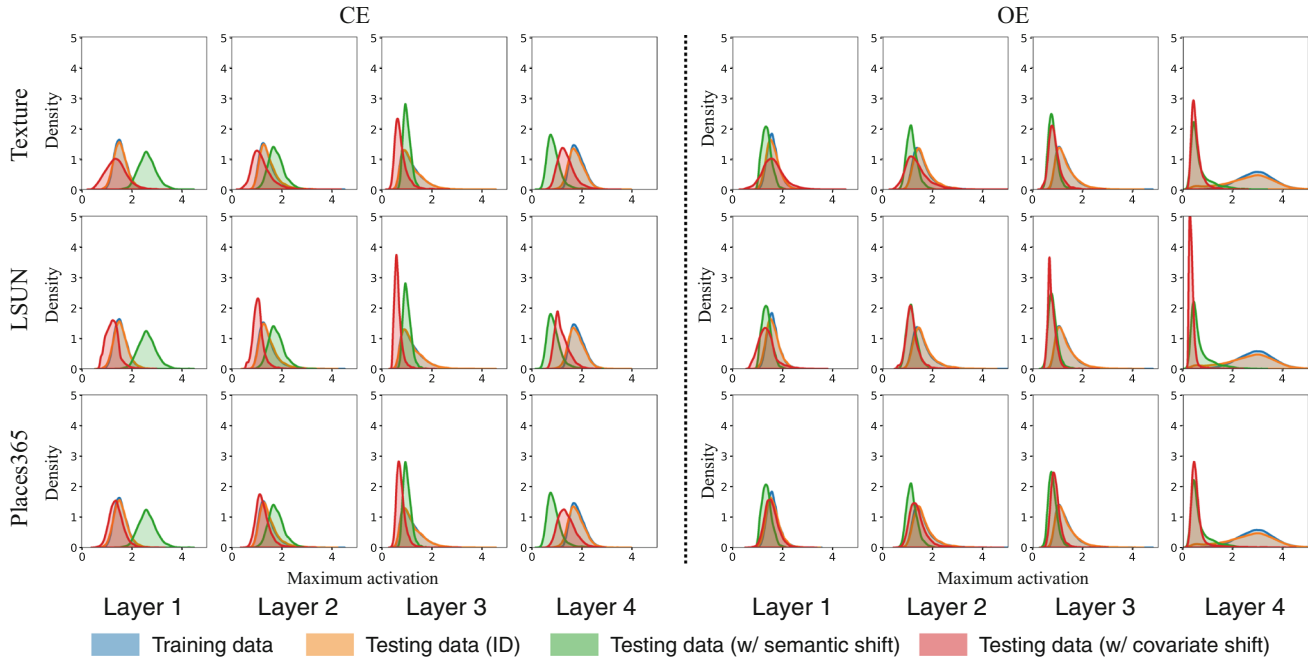

Fig. 5 Histogram of activations for ResNet-18 pretrained on a subset of CIFAR10 with four training classes and evaluated on: training and ID testing data; open-set data (disjoint six classes in CIFAR10) and OOD data (from Textures, LSUN and Places365). Specifically, each subplot shows the maximum activation (along channel, width and height dimension) at the outputs from `layer_1` to `layer_4` of ResNet-18, displayed from left to right in the figures. The behavior of OE is different from CE, whose activation maps become more separable in the deeper rather than the shallower layers. See Section B in Appendix for results on more datasets

3.3 Qualitative Analysis

In this section, we qualitatively interrogate the learned representations of Cross-Entropy and Outlier Exposure networks in order to explain the stark performance boost of OE on existing OOD detection benchmarks. Specifically, we use the value of the maximally activated neuron at various layers to analyze how the networks respond to distribution shifts. We pass every sample through the network, and plot the histogram of maximum activations at every layer in Fig. 5 (see Fig. 13 for the analogous results by training with the ARPL+CS method).

This is inspired by Vaze et al. (2022), who show the 'maximum logit score' (MLS, the maximum activation at a network's output layer) can achieve SOTA for OSR. Furthermore, Dietterich and Guyer (2022) propose that networks respond to a 'lack of familiarity' under distribution shift by failing to light in-distribution activation pathways. We investigate how activations at various stages of a deep network vary under different 'unseen' datasets. Figure 5 shows histograms of the maximum activations at the outputs from `layer_1` to `layer_4` of ResNet-18 (He et al., 2016) trained on CIFAR10 when evaluated on data with different shifts (here we use 'layer' to refer to ResNet block).

For open-set data, we find that early layer activations are largely the same as for the ID test data. It is only later in the network that the activation patterns begin to differ. This is intuitive as the low-level textures and statistics of the open-set data do not vary too much from the training images. Furthermore, it has long been known that early filters in CNNs tend to focus on textural details such as edges (Krizhevsky et al., 2012). In contrast, we find that some OOD datasets, such as SVHN, induce very different activations in the early layers. Our explanation for this phenomenon is analogous: SVHN contains very different image statistics and low-level features to the training dataset of CIFAR10, and hence induces different activations in early layers. Most interestingly, however, some datasets which show markedly different early layer activations actually display *more similar* activations at later layers (like SVHN, see Fig. 12).

Meanwhile, OE displays show substantially different intermediate activations. Interestingly, the maximum activation in early layers look very similar to the ID testing data, but tend to be less so later on in the network. It is clear that activations in later layers are more discriminative after using OE loss when compared with using CE loss.

4 Disentangling Distribution Shifts

Having analysed methodologies for detecting distribution shift across the OOD detection and OSR settings, we turn our attention to the benchmarks. While it is clear that OSR specifically aims to detect unseen categories, there is no specification of the type of distribution shift which OOD detection benchmarks aim to capture, or how they would relate to a real-world scenario. In this section, we propose a lens through which to consolidate types of distribution shift. Specifically, we propose that 'distribution shift' can be parameterised along two broad, orthogonal, axes: *semantic* shift and *covariate* shift. Pure semantic shift is when new categories are encountered, and is the explicit focus of OSR, while covariate shift refers to the setting when the semantics of test images remain constant, but other features change.

Formally, similarly to Wiles et al. (2022), we consider a latent variable model of the data generation process, with latent z:

$$z \sim p(z) \qquad y^i \sim p(y^i|z) \quad i \in \{1 \dots L\} \qquad \mathbf{x} \sim p(\mathbf{x}|z) \quad (1)$$

Here, $\mathbf{x}$ is an image and y^i represents an image attribute. The set of attributes could include traditional features such as 'color' or 'texture', or refer to more abstract features such as 'beak shape' of a bird. We define a set of semantic attributes, Y_S, such that the category label of an image is a function of these attributes. Furthermore, we define covariate attributes, Y_C, which can be freely varied without the category label changing. In this framing, given marginal training distributions $p_{train}(Y_S)$ and $p_{train}(Y_C)$, detecting semantic shift is the task of flagging when $p_{test}(Y_S) \neq p_{train}(Y_S)$. Analogously, we wish to flag covariate shift if $p_{test}(Y_C) \neq p_{train}(Y_C)$.

To motivate this setting, consider the perceptual system in an autonomous vehicle, which has been trained to recognize *cars* during the *day*. A *semantic shift* detector is necessary for when the system encounters a new category, *e.g.*, to flag that *bicycle* is an unknown concept. Meanwhile, a *covariate shift* detector is necessary for when the system is deployed at *night-time*, where the categories may be familiar, but the performance of the system could be expected to degrade.

4.1 Datasets

As a starting point, we note that (Vaze et al., 2022) introduced the Semantic Shift Benchmark (SSB), a distribution shift benchmark with isolates *semantic shift*. We mainly focus on ImageNet-SSB (Russakovsky et al., 2015) and CUB-SSB (Wah et al., 2011) datasets. 'Seen' classes in ImageNet-SSB are the original ImageNet-1K classes, while 'unseen' classes selected from the disjoint set of ImageNet-21K-P (Ridnik et al., 2021). Meanwhile, CUB-SSB splits the 200 bird classes

Table 4 Results of OOD detection and OSR benchmarks on large-scale datasets, using ResNet-50 model trained with the OE loss compared with CE and ARPL baselines

Training method	Scoring rule	Covariate shift			Semantic Shift						Overall
		ImageNet-C	ImageNet-R	AVG	ImageNet-SSB (easy/hard)		CUB (easy/hard)		Scars (easy/hard)		FGVC (easy/hard)		AVG
CE	MLS	**67.92**	**86.71**	**77.32**	80.28	75.05	**88.29**	**79.33**	94.03	82.24	**90.65**	**82.55**	**84.05**	**82.71**
ARPL+CS		63.94	82.77	73.36	79.92	74.60	83.50	75.49	**94.78**	**83.63**	87.04	77.71	82.08	80.34
OE (w/ Places)		61.77	80.53	71.15	**82.42**	**75.58**	79.16	73.83	91.02	78.69	88.38	79.19	80.81	78.88
OE (w/ YFCC15M)		64.12	82.01	73.07	79.37	72.55	75.19	70.28	84.03	71.34	74.20	66.63	71.12	71.51

The results are averaged from five independent runs. We separately introduce outlier data from different data sources including Places and YFCC15M to feed OE

Table 5 Evaluation on large-scale OOD detection and OSR benchmarks using ResNet-50 model trained with different losses and scoring rules

Training method	Scoring rule	Covariate shift					Semantic shift					Overall
		ImageNet-C	ImageNet-R	Waterbird (easy/hard)		AVG	ImageNet-SSB (easy/hard)		CUB (easy/hard)		AVG	
(a) Evaluation based on ResNet-50 trained with the CE loss												
CE	MSP	64.63	80.53	81.65	**75.33**	75.54	80.16	75.01	88.11	**79.43**	80.68	78.11
	MLS	*67.92*	*86.71*	81.87	*75.18*	*77.92*	**80.28**	75.05	88.29	*79.33*	**80.74**	*79.33*
	ODIN	63.69	85.62	79.51	71.54	75.09	74.56	*75.27*	86.24	73.88	77.49	76.29
	Energy	**68.05**	**87.04**	**82.49**	74.60	**78.05**	79.76	74.96	**88.81**	79.06	80.65	**79.35**
	MLS+ReAct	66.64	84.82	81.69	75.12	77.07	**80.28**	75.07	88.29	*79.33*	**80.74**	78.91
	ODIN+ReAct	61.69	83.25	79.48	71.50	73.98	74.56	**75.29**	86.24	73.88	77.49	75.74
	Energy+ReAct	66.88	83.92	*82.48*	74.55	76.96	79.76	74.99	**88.81**	79.06	80.66	78.81
(b) Evaluation based on ResNet-50 trained with the ARPL loss												
ARPL+CS	MSP	61.85	78.68	79.42	**72.30**	73.06	79.90	**74.67**	83.53	**75.64**	**78.44**	75.75
	MLS	*63.94*	*82.77*	79.48	72.09	*74.57*	**79.92**	*74.60*	83.50	*75.49*	*78.38*	*76.47*
	ODIN	61.88	77.03	73.76	69.26	70.48	68.72	71.23	73.87	69.77	70.90	70.69
	Energy	**64.13**	**83.25**	**79.64**	71.86	**74.72**	79.87	74.49	**83.70**	75.46	*78.38*	**76.55**
	MLS+ReAct	62.69	80.69	79.44	72.07	73.72	**79.92**	*74.60*	83.44	75.43	78.35	76.04
	ODIN+ReAct	62.23	76.08	73.75	69.23	70.32	68.72	71.23	67.42	63.91	67.82	69.07
	Energy+ReAct	62.89	81.17	*79.60*	71.83	73.87	79.87	74.49	**83.70**	75.41	78.37	76.12

The results are averaged from five independent runs. Bold values represent the best results, while italic values represent the second best results. Models trained with the CE loss outperforms the ones with ARPL on both covariate shift and semantic shift

in CUB into 'seen' and 'unseen' categories. Furthermore, the unseen categories are split into Easy and Hard classes by their attributes, and the splitting rule depends on semantic similarity of every pair of visual attributes in the unknown classes and the training classes. For all the above datasets, categories appearing in the training set would not be included in the evaluation set.

For *covariate shift*, we propose ImageNet-C (Hendrycks & Dietterich, 2019) and ImageNet-R (Hendrycks et al., 2021) to demonstrate distribution shift with respect to the standard ImageNet dataset. Both datasets contain images from a subset of the ImageNet-1K categories, but with different low-level image statistics. ImageNet-C applies four main corruptions (*e.g.*, noise, blur, weather, and digital) with varying intensities to the validation images of ImageNet-1K, while ImageNet-R collects various artistic renditions of foreground classes from the ImageNet-1K dataset. We also choose Waterbirds (Sagawa et al., 2019) to test the model trained on the CUB-SSB 'Seen' classes. Waterbirds inserts bird photographs from the CUB dataset into backgrounds picked from the Places dataset (Zhou et al., 2017), meaning it has the same semantic categories to CUB but in different contexts.

4.1.1 Discussion

We note that there is no uniquely optimal framing for discussing distribution shift, and here briefly discuss alternate proposals. For instance, Zhao et al. (2022) propose a fine-grained analysis of the shifts, where the test time distribution is controlled for specific attributes such as shape and pose. Also related, Tran et al. (2022) discuss that indications of 'unfamiliarity' in a neural network could refer to many things, including confusing classes and sub-population shift. We propose our simple framing as a way to fill the 'negative space' left by the semantic shift detection task of OSR. Furthermore, we suggest it is important to study distribution shift in this way, as classifiers are specifically optimized to differentiate between one set features (Y_S) while in fact being invariant to others (Y_C). As such, we would expect models to react differently to changes their distributions.

Finally, we note that for the covariately shifted samples, we ideally wish to develop classifiers which are robust and can perform well despite the presence of distribution shift. However, given that machine learning models performance degrades under distribution shift, we wish to be able to *detect* when the shift is present. To measure whether there is a trade-off between robust models and those which can detect covariate shift, we introduce a new metric in Section 4.5.

4.2 Quantitative Analysis

In Tables 5 and 4, we evaluate a selection of previously discussed methods on our large-scale benchmark for both OOD detection and OSR. Through this large-scale evaluation, we find that *in terms of training methods, among CE, ARPL (+CS), and OE, there are no clear winners across the board*. It is surprising that the best performer on the previous small scale benchmarks (see Table 2), OE, appears to struggle when scaled up (last two rows in Table 5). We analyse this contradiction in the next section. In terms of scoring rules, we again find that *the magnitude-aware scoring rules* (MLS and Energy), consistently produce the best performance regardless of the methods and benchmarks (both standard small-scale ones and our large-scale ones).

4.3 OE on Large-Scale Datasets

Here, we investigate why OE performs worse than other methods on a large-scale benchmark. One critical difference between OE and other methods is that OE uses auxiliary OOD data for training. Intuitively, if the distribution of the auxiliary OOD training data accurately reflects the distribution of the actual OOD testing data, we would expect better detection performance. Otherwise, incomplete or biased outlier data can negatively impact learning. To analyze this further, we plot the distribution of maximum activation of the output feature (from the last layer) for samples from different data sources: ID data, OOD data, and auxiliary data. The results are shown in Fig. 6. It is worth noting that the OOD detection performance strongly (negatively) correlates with the overlapped region of the ID and OOD curve. Additionally, when using 300K random images as auxiliary OOD data (as shown in the 2^{nd} row), there is a high correlation with actual OOD data, resulting in excellent performance (see Table 2). We also provide results on a large-scale dataset in Fig. 7 and further qualitative investigation in Section C in Appendix.

We further retrieve nearest neighbors for the given samples on both small-scale (*e.g.*, Textures and Places365) and large-scale (*e.g.*, ImageNet-C and ImageNet-R) benchmarks using models trained by OE (see Fig. 8). By retrieving the nearest neighbors from the union of ID data and auxiliary data, we observe that for small-scale scenarios, the retrieved nearest neighbors are found in the auxiliary data. However, such phenomenon does not occur consistently in the large-scale datasets. This observation aligns with the correlation between the distance from OOD data to auxiliary data and the OOD detection performance of models trained using OE in Table 4.

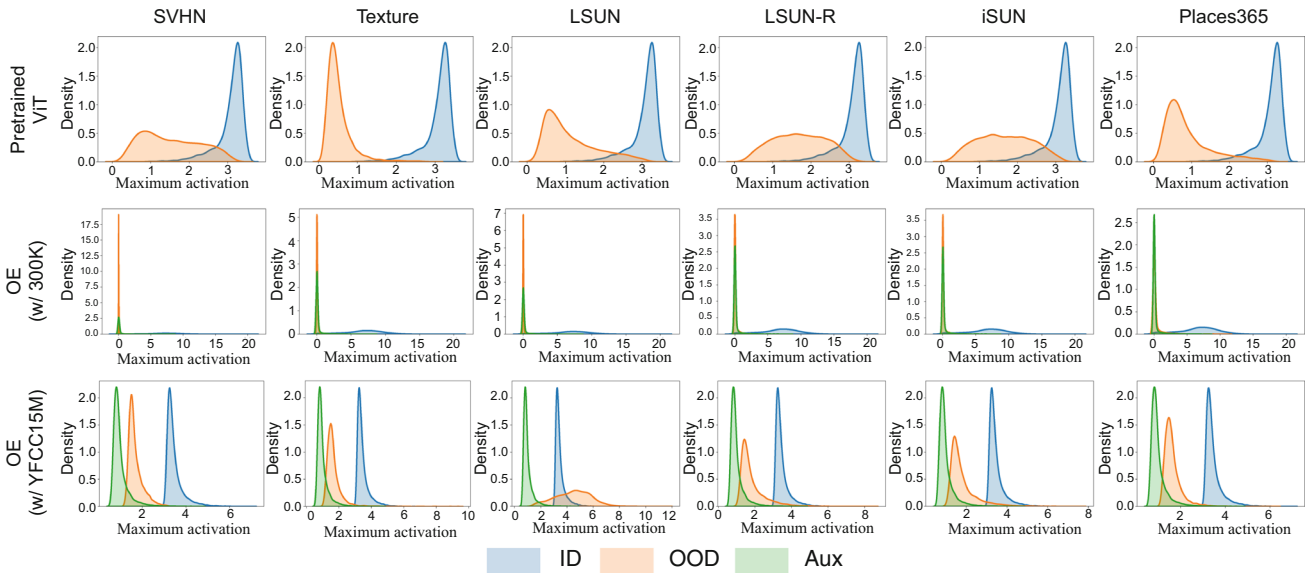

Fig. 6 The distribution of the maximum activation of the output feature from the last block is investigated for ID training data, OOD testing data, and auxiliary training data on small-scale datasets. When equipped with OE loss, the highly correlated auxiliary data (300K images) can greatly enhance the OOD detection performance. Therefore, careful selection of auxiliary data is crucial

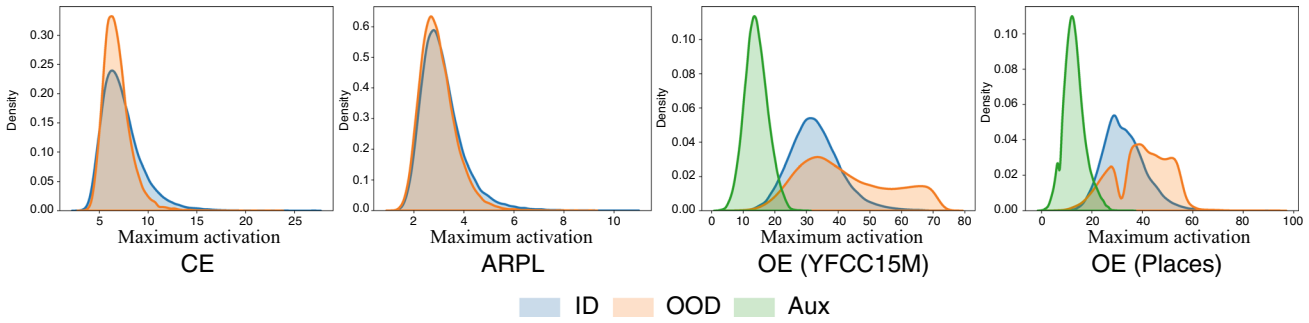

Fig. 7 Analysis on the distribution of the maximum activation of the output feature from the last layer, for ID training data, OOD testing data, and auxiliary training data on large-scale datasets. Compared to the 300 K images for small-scale datasets, there is less similarity between the auxiliary data (*i.e.*, YFCC15M and Places) and the OOD data. This finding aligns with the results in Table 4

4.4 Dataset Proximity Versus OOD Detection Performance

To verify that the dataset proximity between auxiliary data and OOD data correlates to the OOD detection performance, we measure the correlation between OOD detection performance and dataset proximity. We quantify this proximity by calculating the distance between OOD data and auxiliary data. For a specific OOD dataset, we compute the distance via Top-K nearest neighbors and a deep kernel method (Liu et al., 2020), respectively. For Top-K nearest neighbors, we compute the average of all the distances between the normalized feature of each OOD sample and its Top-K nearest neighbors in the auxiliary dataset. It can be formulated by:

$$Dist_{nn}\left(\mathcal{D}^{ood}, \mathcal{D}^{aux}\right) = \frac{\sum_{i=1}^{|\mathcal{D}^{ood}|}\sum_{k=1}^{K} d\left(Z_i^{ood}, Z_k^{aux}\right)}{K|\mathcal{D}^{ood}|}, \tag{2}$$

where $\mathcal{D}^{ood}$ and $\mathcal{D}^{aux}$ represent the OOD and auxiliary datasets. Z^{ood} and Z^{aux} are the l_2-normalized extracted feature from the pretrained model given OOD and auxiliary samples. $d(\cdot,\cdot)$ is the distance measure for nearest neighbors retrieval. We also calculate the deep kernel distance between the OOD and auxiliary datasets following (Liu et al., 2020):

$$Dist_{dk}\left(\mathcal{D}^{ood}, \mathcal{D}^{aux}\right) = \widehat{\mathrm{MMD}}_u^2\left(\mathcal{D}^{ood}, \mathcal{D}^{aux}; \phi\right), \tag{3}$$

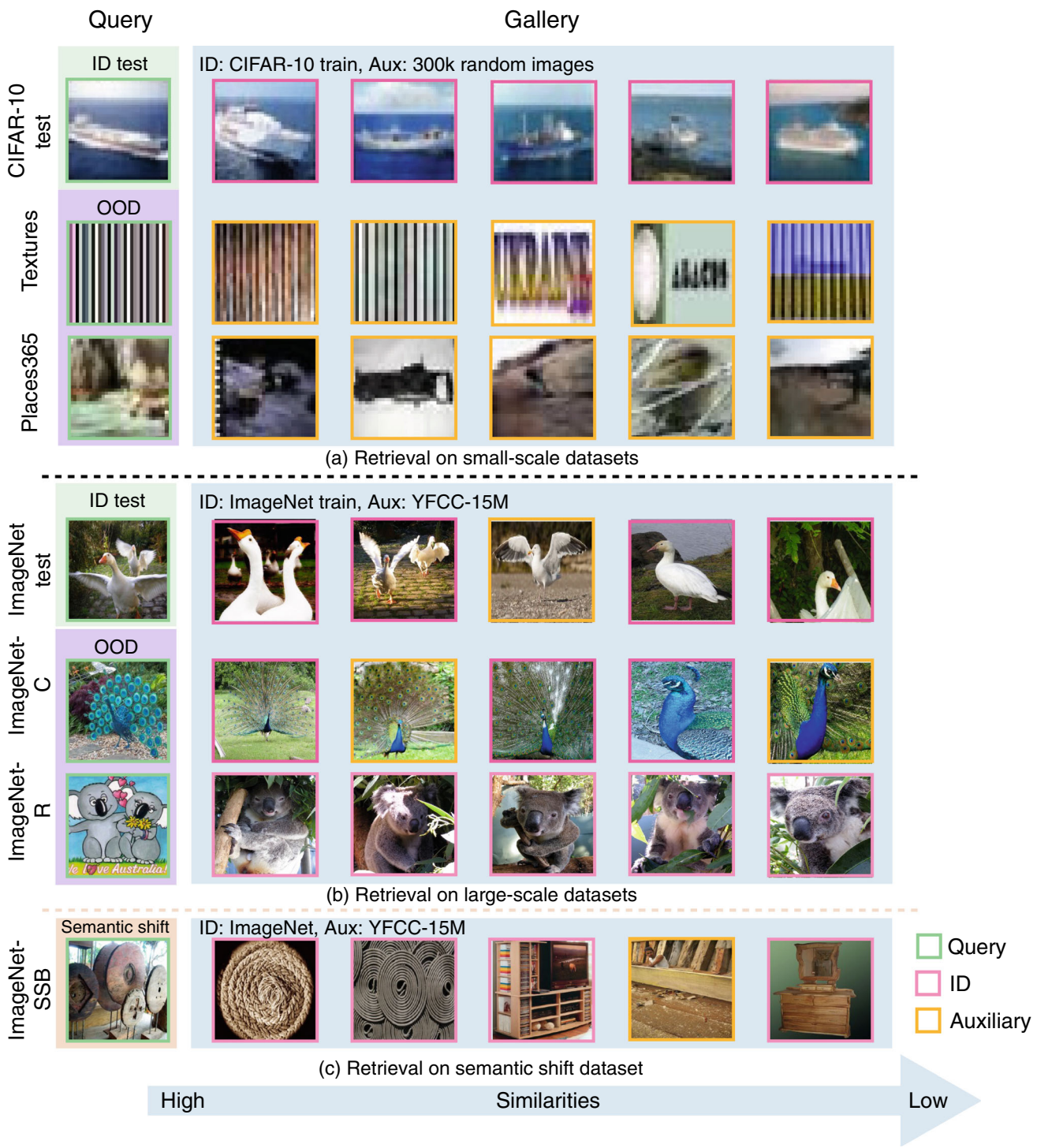

Fig. 8 Visualization of nearest neighbors of test samples retrieved from the union of ID and auxiliary training data. We search for the nearest neighbors of samples in both small-scale (*e.g.*, Textures and Places365) and large-scale (*e.g.*, ImageNet-C and ImageNet-R) OOD datasets. We also investigate the nearest neighbours for the sample with semantic shift using ImageNet-SSB. It is clear that the improvement in OOD detection performance with OE is closely tied to the similarity between OOD and auxiliary data

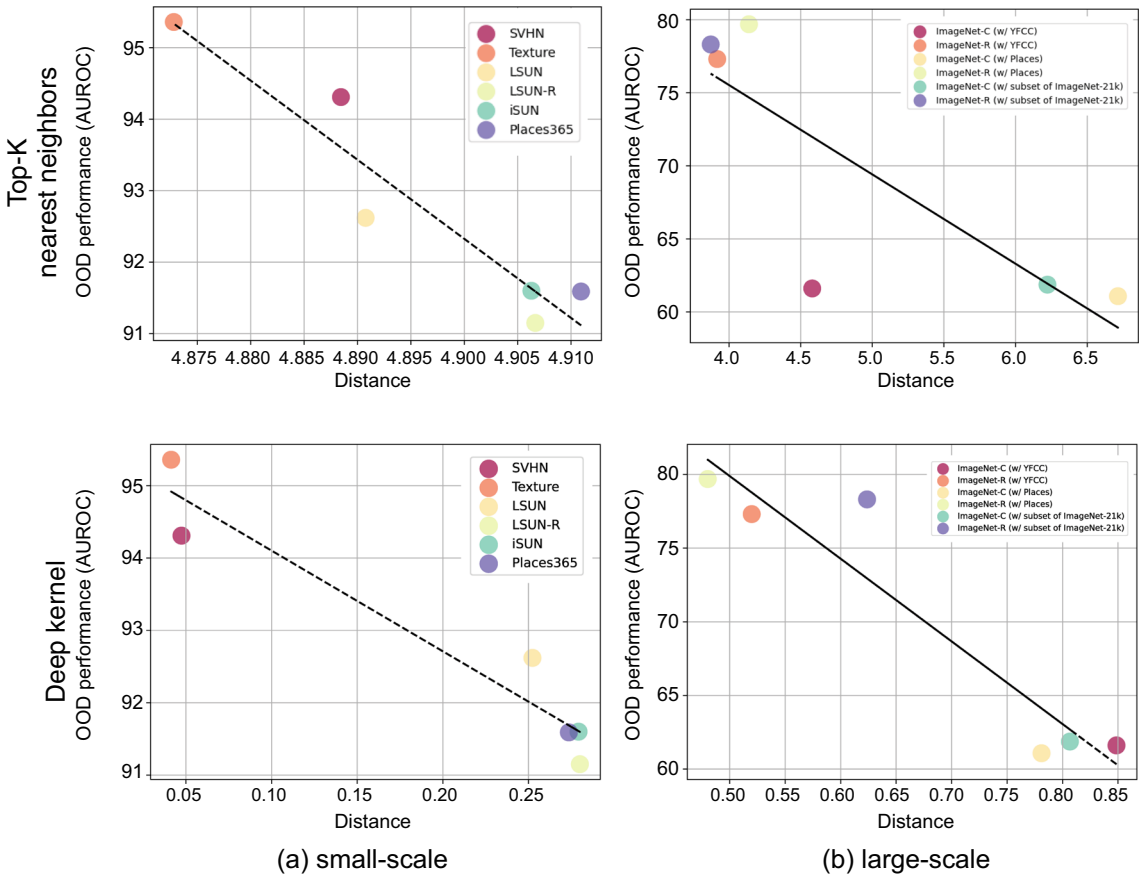

Fig. 9 OOD detection performance versus OOD-AUX data distance of different auxiliary training data for both (**a**) small-scale and (**b**) large-scale OOD data. The OOD detection performance decreases as the distance between OOD data and auxiliary data increases, for both measurements based on Top-K nearest neighbors and deep kernel distance

where $\widehat{\mathrm{MMD}}_u^2(\cdot,\cdot)$ is the U-static estimator of the maximum mean discrepancy (MMD) (Gretton et al., 2012) based on i.i.dlet@tokeneonedotsamples from $\mathcal{D}^{ood}$ and $\mathcal{D}^{aux}$, which is considered as an unbiased estimator having nearly minimal variance (Gretton et al., 2009). $\phi(x, y) = \left[(1-\epsilon)\kappa\left(Z^{ood}, Z^{aux}\right) + \epsilon\right] q(x, y)$, where $\kappa(\cdot,\cdot)$ and $q(\cdot,\cdot)$ are Gaussian kernels. ϵ is sampled from $(0, 1)$. The intuition behind the deep kernel method is to effectively distinguish data distributions via representative deep features of samples from each distribution. We compute the distance (denoted as OOD-AUX data distance) for both small-scale and large-scale OOD data in Fig. 9, which further validates that closer proximity between auxiliary data and OOD data leads to better performance by OE models.

4.5 Outlier-Aware Accuracy

Finally, we introduce a new metric to reconcile the problems of *detecting* covariate shift and being *robust* to it. Although AUROC is commonly used to compare different techniques for distinguishing out-of-distribution samples, it does not capture the model's ability to reliably classify testing samples in the presence of distribution shifts. To analyze the relationship of performance between covariate shift and robustness, we introduce a novel measure, which we term *Outlier-Aware Accuracy* (OAA). At a given threshold, and a given set of predictions (both ID *vs.* OOD predictions, and predictions within the closed-set categories), we compute the aggregate frequency of 'correct' predictions. The definition of 'correct' varies depending on the prediction. Specifically, as shown in Fig. 10, all instances predicted as ID samples should have accurate class predictions, and all OOD samples that are not already categorized should be detected. This is because we expect a good model to correctly classify all samples with any covariate shift and identify any remaining OOD samples.

The number of the counted instances is then divided by the total number of testing instances to produce the OAA, which is robust to non-semantic shift. This measure is computed under different thresholds based on the scoring rules and aggregated:

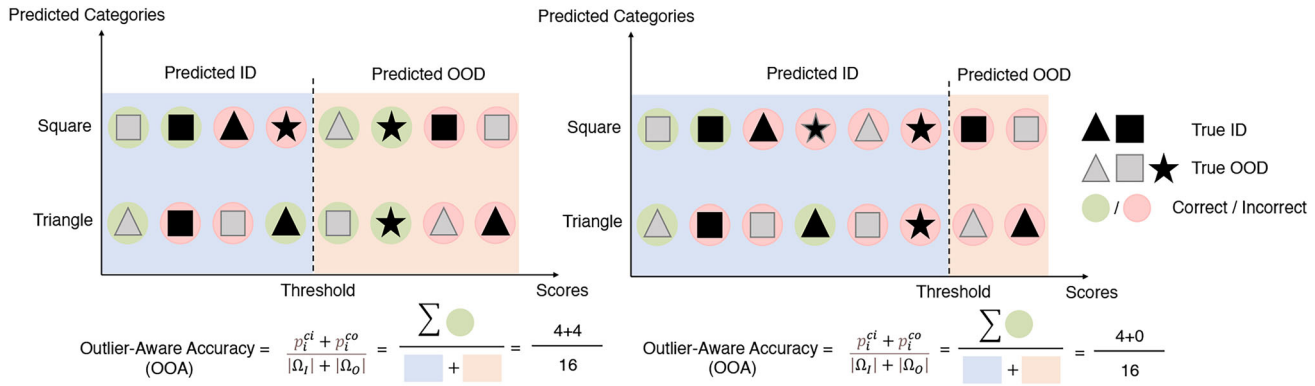

Fig. 10 Demonstration for Outlier-Aware Accuracy (OAA). OAA measures the frequency of 'correct' predictions made by the model at a given threshold. The 'correct' predictions are defined as: (1) Among the testing samples predicted as ID, those whose semantic class labels are *correctly* predicted, denoted as p_i^{ci}; (2) True OOD samples that have *incorrect* semantic class predictions, denoted as p_i^{co}

Table 6 We compute the mean OAA (mOAA) rate across all thresholds to compare different approaches

Methods	MSP	MLS	Energy	MLS+ReAct	Energy+ReAct
CE	0.661	0.658	0.655	0.655	0.641
ARPL	**0.664**	**0.660**	**0.657**	**0.659**	**0.646**
OE (w/ YFCC15M)	0.623	0.611	0.611	0.583	0.597

Bold values reflect the best methods for each dataset

4.5.1 mOAA Metric

The OAA values across all thresholds can also be aggregated into a single value within [0, 1] as an overall measure, mean OOA (mOAA). To compute the mOAA, we consider testing images including both ID data from $\mathcal{D}^{id}$ and OOD data from $\mathcal{D}^{ood}$. The mOAA is defined as follows:

$$mOAA = \frac{1}{N}\sum_{i=1}^{N} OAA_i = \frac{1}{N}\sum_{i=1}^{N} \frac{p_i^{ci} + p_i^{co}}{|\mathcal{D}^{id}| + |\mathcal{D}^{ood}|}, \quad (4)$$

where p_i^{ci} denotes the correct predictions among the testing samples predicted as ID, p_i^{co} denotes the correct predictions among the testing samples predicted as OOD (see Fig. 10), $|\cdot|$ represents the size of the testing set, and N is the number of different thresholds. The mOAA score ranges from 0 to 1. A higher value indicates better performance, with a score of 1 representing perfect detection and recognition, while a score of 0 represents the worst performance in terms of separation and recognition.

Based on the numerical results in Fig. 11 and Table 6, we have observed a turning point for the threshold that achieves the optimal balance between model robustness and OOD detection.

It is worth noting that this metric has a connection to AURC (Geifman et al., 2019). While both AURC and OAA consider classifier performance, our proposed OAA specifically measures the 'correct prediction rate', providing an interpretable value between 0 and 1. Therefore, we believe that this metric can be effectively used to study the trade-off between OOD detection and generalization with greater precision.

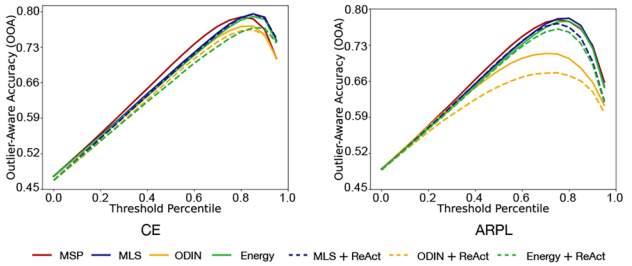

Fig. 11 Outlier-aware accuracy (OAA) of a classifier trained with CE/ARPL loss computed by different thresholds

5 Summary of Empirical Phenomena

In the previous sections, we thoroughly evaluated methods for OOD detection and OSR in terms of scoring rules, training methods, and auxiliary data. To summarize these phenomena, we briefly highlight the key observations as follows: (i) Magnitude-aware scoring rules (*i.e.*, MLS and energy) offer obvious advantages for both OOD detection and OSR. Compared with non-magnitude-aware techniques (*i.e.*, ODIN and ReAct), we recommend more stable and determin-

istic magnitude-aware scoring rules. (ii) Outlier Exposure (OE) is by far the most effective training method for improving performance on OOD detection and OSR benchmarks when the auxiliary data is highly correlated to the actual OOD data. However, while finding such auxiliary data is possible for small-scale benchmarks, it is highly non-trivial to find such data in (more realistic) large-scale settings.

6 Conclusion

In this study, we explore Out-of-Distribution (OOD) detection and Open-set Recognition (OSR). We conducted a thorough cross-evaluation of methods for OOD detection and OSR. Additionally, we introduced a new benchmark setting that separates the distribution shift problem into *covariate* shift and *semantic* shift, proposing large-scale evaluation protocols for both settings. Our study revealed that the best performing method current OSR and OOD datasets (Outlier Exposure) does not generalize well to our challenging large-scale benchmark. We also discovered that magnitude-aware scoring rules are generally more reliable than others. Overall, our new benchmark can serve as an improved testbed for measuring progress in OSR and OOD detection while providing insights into these two problems. We hope that our thorough empirical investigation on the OOD detection and OSR methods and benchmarks can shed light for future study and applications on the broader data distribution shift detection problem.

Appendix A More Experimental Results on Benchmarks

We additionally evaluate different methods on Scars-SSB and FGCV-Aircraft-SSB datasets from the Semantic-Shift Benchmark (Vaze et al., 2022) to further investigate the performance of scoring rules against semantic shifts. From Tables 7 to 10, we can observe that magnitude-aware scoring rules perform well among different training methods.

We also investigate the effect of OE using different auxiliary data (*e.g.*, Places and YFCC-15M). As shown in Tables 9 and 10, we can see that OE performance heavily depends on the auxiliary training data.

The degeneration of OE when applied to large-scale datasets drives us to think about the core contribution behind the OE method. To find out the reason, we apply OE to the small-scale datasets with different auxiliary data (*i.e.*, YFCC-15M) in Table 11. Compared with results using 300K random images, the one using YFCC-15M cannot exceed the performance of the OE baseline. This indicates that the selection of auxiliary data is essential to the OE method and the success of OE may come from the similarity of auxiliary data distribution and test-time outliers.

Table 7 Results of OOD detection and OSR benchmarks on large-scale datasets, using ResNet-50 model trained with the CE loss

Training method	Scoring rule	Covariate shift			Semantic shift							Overall
		ImageNet-C ID=63.05	ImageNet-R ID=76.13	AVG	ImageNet-SSB (easy/hard)		CUB (easy/hard)	Waterbirds (easy/hard)		Scars (easy/hard)		FGVC (easy/hard)		AVG	
CE	MSP	64.63	80.53	72.58	*80.16*	*75.01*	88.11	**79.43**	81.65	**75.33**	**94.15**	**82.34**	90.63	82.55	*82.94*	81.21
	MLS	*67.92*	*86.71*	*77.32*	**80.28**	**75.05**	88.29	*79.33*	81.87	*75.18*	*94.03*	*82.24*	90.65	82.55	**82.95**	*82.01*
	ODIN	63.69	85.62	74.66	74.56	75.27	86.24	73.88	79.51	71.54	92.87	80.88	**90.97**	80.97	80.67	79.67
	Energy	**68.05**	**87.04**	**77.55**	79.76	74.96	**88.81**	79.06	**82.49**	74.60	93.92	**82.03**	90.86	**82.82**	82.93	**82.03**
	MLS+ReAct	66.64	84.82	75.73	80.28	75.07	88.29	79.33	81.69	75.12	94.01	82.23	90.61	82.57	82.92	81.72
	ODIN+ReAct	61.69	83.25	72.47	74.56	75.29	86.24	73.88	79.48	71.50	92.80	80.85	*90.88*	80.92	80.64	79.28
	Energy+ReAct	66.88	83.92	75.40	79.76	74.99	**88.81**	79.06	*82.48*	74.55	93.89	82.00	90.80	*82.79*	82.91	81.66

The results are averaged from five independent runs
Bold values reflect the best methods for each dataset

Table 8 Results of OOD detection and OSR benchmarks on large-scale datasets, using ResNet-50 model trained with the ARPL loss

Training method	Scoring rule	Covariate shift			Semantic shift										Overall
		ImageNet-C ID=63.05	ImageNet-R ID=76.13	AVG	ImageNet-SSB (easy/hard)		CUB (easy/hard)	Waterbirds (easy/hard)		Scars (easy/hard)		FGVC (easy/hard)		AVG	
ARPL	MSP	61.85	78.68	70.27	*79.90*	**74.67**	83.53	**75.64**	79.42	**72.30**	**94.83**	**83.96**	86.81	**78.01**	**80.91**	79.13
	MLS	*63.94*	*82.77*	*73.36*	**79.92**	*74.60*	83.50	*75.49*	79.48	*72.09*	*94.78*	83.63	87.04	77.71	80.82	*79.58*
	ODIN	61.88	77.03	69.46	68.72	71.23	73.87	69.77	73.76	69.26	82.08	69.10	70.24	73.47	72.15	71.70
	Energy	**64.13**	**83.25**	**73.69**	79.87	74.49	**83.70**	75.46	**79.64**	71.86	94.70	83.56	**87.28**	*77.74*	*80.83*	**79.64**
	MLS+ReAct	62.69	80.69	71.69	79.92	74.60	83.44	75.43	79.44	72.07	94.77	*83.66*	87.01	77.69	80.80	79.28
	ODIN+ReAct	62.23	76.08	69.16	68.72	71.23	67.42	63.91	73.75	69.23	82.07	69.09	70.20	73.49	70.91	70.62
	Energy+ReAct	62.89	81.17	72.03	79.87	74.49	**83.70**	75.41	*79.60*	71.83	94.69	83.56	*87.27*	77.71	80.81	79.35

The results are averaged from five independent runs
Bold values reflect the best methods for each dataset

Table 9 Results of OOD detection and OSR benchmarks on large-scale datasets, using ResNet-50 model trained with the OE loss combined with auxiliary data from Places

Training method	Scoring rule	Covariate shift					Semantic shift					Overall
		ImageNet-C	ImageNet-R	Waterbird (easy/hard)		AVG	ImageNet-SSB (easy/hard)		CUB (easy/hard)		AVG	
OE	MSP	61.02	75.30	*79.11*	**73.88**	72.33	82.20	73.45	75.91	69.18	75.19	73.76
	MLS	61.77	80.53	**79.31**	**73.88**	**73.87**	*82.42*	*75.58*	**79.16**	**73.83**	**77.75**	**75.81**
	ODIN	57.74	**82.31**	71.28	69.30	70.16	81.75	70.87	73.71	66.05	73.10	71.63
	Energy	**64.10**	*81.11*	76.39	70.86	*73.12*	**83.47**	**75.61**	*78.56*	*73.01*	*77.66*	*75.39*
	MLS+ReAct	*62.39*	79.76	77.00	71.93	72.77	81.23	73.07	72.09	70.16	74.14	73.45
	ODIN+ReAct	58.28	77.94	69.74	70.06	69.01	80.27	70.54	74.40	68.00	73.30	71.15
	Energy+ReAct	62.26	80.91	75.32	69.71	72.05	82.10	73.79	77.30	70.74	75.98	74.02

The results are averaged from five independent runs
Bold values reflect the best methods for each dataset

Table 10 Results of OOD detection and OSR benchmarks on large-scale datasets, using ResNet-50 model trained with the OE loss combined with auxiliary data from YFCC-15M

Training method	Scoring rule	Covariate shift					Semantic shift					Overall
		ImageNet-C	ImageNet-R	Waterbird (easy/hard)		AVG	ImageNet-SSB (easy/hard)		CUB (easy/hard)		AVG	
OE	MSP	59.02	70.01	73.67	68.71	67.85	68.44	71.60	71.11	65.27	69.11	68.48
	MLS	*64.12*	**82.01**	**79.72**	**74.08**	**74.98**	*79.37*	72.55	75.19	**70.28**	*74.35*	*74.67*
	ODIN	60.80	74.36	66.80	68.94	67.73	72.01	66.87	71.48	67.91	69.57	68.65
	Energy	**64.31**	*81.50*	77.95	72.76	*74.13*	**81.50**	**74.33**	**77.78**	*70.09*	**75.93**	**75.03**
	MLS+ReAct	60.15	79.42	76.99	*73.58*	72.54	72.30	72.74	73.46	69.67	72.04	72.29
	ODIN+ReAct	60.98	74.05	66.10	68.89	67.51	64.67	61.79	72.26	67.76	66.62	67.06
	Energy+ReAct	61.87	79.79	*78.83*	71.98	73.12	71.24	*73.26*	*75.78*	69.85	72.53	72.83

The results are averaged from five independent runs
Bold values reflect the best methods for each dataset

Table 11 Results of OOD detection and OSR benchmarks on small-scale datasets, using ResNet-18 model trained with the OE loss combined with auxiliary data from YFCC-15M.

Training method	Scoring rule	OOD detection benchmarks							OSR benchmarks					Overall	
		SVHN	Textures	LSUN	LSUN-R	iSUN	Places365	AVG ID=95.47	CIFAR10 ID=97.52	CIFAR+10 ID=97.76	CIFAR+50 ID=97.29	TinyImageNet ID=87.30	AVG		
OE	MSP	98.96	**99.50**	*98.17*	94.44	94.50	**99.61**	*97.53*	*90.17*	91.21	88.17	80.54	87.52	93.53	
	MLS	*98.97*	**99.50**	**98.19**	94.46	94.55	**99.61**	**97.55**	**90.36**	*93.47*	89.25	*81.44*	*88.63*	*93.98*	
	ODIN	**99.02**	97.40	97.12	87.21	87.84	97.18	94.30	87.93	79.37	79.47	**81.58**	82.09	89.41	
	Energy	98.93	99.48	98.12	94.19	94.33	**99.61**	97.44	89.90	**94.91**	*90.20*	81.43	**89.11**	**94.11**	
	MLS+ReAct	98.86	99.49	97.76	**94.69**	**94.78**	99.60	*97.53*	89.77	91.99	89.92	81.32	88.25	93.82	
	ODIN+ReAct	98.89	96.66	95.48	82.50	83.60	96.37	92.25	83.31	84.32	82.37	81.43	82.86	88.49	
	Energy+ReAct	98.83	99.47	97.70	*94.51*	*94.66*	**99.61**	97.46	89.45	93.00	**90.21**	81.40	88.52	93.88	

The results are averaged from five independent runs
Bold values reflect the best methods for each dataset

Appendix B Activations of OOD and Open-Set Data at Different Layers

We provide the maximum activations for intermediate layers of ResNet-18 trained on CIFAR10 when evaluated on in-distribution data and data with different shifts. Activations in later layers are more discriminative between ID and OOD/open-set samples. After using the OE loss, we can easily notice that the OOD samples are more separable than the model trained with CE loss in Fig. 5. We also show the maximum activation of the remaining datasets in Fig. 12.

We also visualize the histogram of maximum activations of the model trained using ARPL+CS at every layer in Fig. 13. Our findings align with the observations in Section 3.3 in the main paper, indicating that the early layer activations closely resemble those of the ID test data while the activation patterns begin to differ in the deeper layers. Notably, the ARPL+CS method demonstrates superior separation compared to CE, but it lags behind OE, as illustrated in Figs. 5 and 12. These findings align with results in Table 2.

Appendix C Correlation of OOD and Auxiliary Training Data

In Fig. 14, we also visualize t-SNE projections of representations for various datasets: ID data (CIFAR10), auxiliary training OOD data (300K (Hendrycks et al., 2019) *vs.* YFCC15M), and different test-time OOD datasets. As seen in Figs. 14 and 15, using 300K images generally leads to better overlap with test-time OOD data. Consequently, OE trained with 300K as auxiliary ODD achieves superior performance compared to its counterpart trained with YFCC15M (Table 2 *vs.* Table 11 in Section A) because it shows a better overlap with the test-time OOD data.

Appendix D Different Architectures and Training Setups

Apart from ResNet, we also conduct experiments using DenseNet121 on small-scale datasets and using DinoViT-S/8 on large-scale datasets. As shown in Tables 12 and 13, magnitude-aware approaches still perform others on several datasets.

Table 12 Evaluation on small-scale OOD detection and OSR benchmarks with various methods, using CIFAR10 as ID. The results are averaged from five independent runs. We report the in-distribution accuracy as 'ID' and denote intractable results as '-', resulting from unaffordable computational cost

Training method	Scoring rule	OOD detection benchmarks							OSR benchmarks					Overall
		SVHN	Textures	LSUN	LSUN-R	iSUN	Places365	AVG ID=95.43	CIFAR10 ID=97.04	CIFAR+10 ID=96.73	CIFAR+50 ID=96.91	TinyImageNet ID=83.52	AVG	
(a) Evaluation based on DenseNet-121 trained with the CE loss														
CE	MSP	95.17	91.56	93.39	94.61	94.40	89.39	93.09	91.63	93.77	90.47	79.76	88.91	91.42
	MLS	*97.05*	*91.85*	*94.53*	96.11	*95.84*	*89.46*	*94.14*	*92.63*	*95.70*	*91.88*	*81.52*	**93.59**	**93.92**
	ODIN	95.01	84.09	86.57	**96.31**	95.70	77.41	89.18	87.62	79.42	83.57	78.92	82.38	86.46
	GODIN	96.31	90.85	94.04	96.23	95.37	88.99	93.63	90.30	91.34	87.76	76.23	86.41	90.74
	SEM	75.65	72.02	75.18	70.93	72.52	76.14	73.74	51.26	47.60	45.36	-	48.07	58.67
	Energy	**97.25**	**91.97**	**94.68**	*96.28*	**96.02**	**89.57**	**94.30**	**92.73**	**95.89**	**92.01**	81.04	**90.42**	*92.75*
	MLS+ReAct	78.39	78.31	89.42	91.55	91.20	85.23	85.68	92.40	94.36	90.28	81.04	89.52	87.22
	ODIN+ReAct	83.47	78.74	89.23	94.05	93.79	83.64	87.15	83.27	84.25	80.62	78.27	82.27	85.20
	Energy+ReAct	85.55	82.68	87.81	90.92	89.61	77.07	85.61	*92.63*	95.28	91.03	**81.53**	90.12	87.41
	MLS+ASH	97.01	91.40	93.72	95.56	90.39	86.30	92.40	92.21	94.35	90.56	80.88	89.50	91.24
	MLS+SHE	89.80	75.02	91.62	87.11	87.40	81.76	85.45	78.80	81.92	77.27	79.92	79.48	83.06
(b) Evaluation based on ResNet-18 trained with the ARPL+CS loss														
ARPL+CS	MSP	94.94	91.81	95.12	94.25	94.39	89.57	93.25	92.60	95.77	94.15	*82.71*	91.31	92.47
	MLS	**98.80**	90.24	94.48	*96.88*	*97.01*	**95.41**	*95.05*	*93.28*	96.65	94.80	**84.91**	**92.41**	*93.99*
	ODIN	78.02	73.46	82.30	93.26	91.40	73.77	82.04	67.81	80.82	73.40	72.52	73.64	78.68
	GODIN	94.83	90.86	*96.01*	96.28	95.88	92.12	94.33	91.11	95.32	93.12	80.89	90.11	92.64
	SEM	78.30	76.35	85.52	76.88	78.03	71.93	77.84	41.00	40.35	42.61	-	41.32	65.67
	Energy	*97.82*	*92.05*	**97.24**	**97.85**	**97.72**	*94.12*	**96.15**	**93.63**	**96.87**	**94.99**	82.12	*91.90*	**94.45**
	MLS+ReAct	85.27	**94.50**	82.44	87.50	82.98	89.43	87.02	92.12	95.33	92.91	80.50	90.22	88.97
	ODIN+ReAct	83.29	79.08	85.12	88.62	83.47	82.05	83.61	68.20	73.12	70.24	55.10	66.67	76.83
	Energy+ReAct	92.37	89.45	91.88	87.07	94.13	91.52	91.07	92.96	*96.78*	*94.91*	82.04	91.67	91.27
	MLS+ASH	96.75	92.01	92.87	97.58	90.79	91.80	93.63	93.19	94.15	93.62	81.28	90.56	92.40
	MLS+SHE	85.75	79.92	82.80	83.80	81.27	86.27	83.30	77.52	74.02	75.25	76.53	75.83	80.31
(c) Evaluation based on DenseNet-121 trained with the OE loss														
OE	MSP	*99.27*	98.86	99.38	98.54	98.58	98.01	98.77	*96.42*	99.32	**98.88**	78.90	*93.38*	*96.61*
	MLS	*99.27*	*98.87*	99.38	98.54	98.58	98.02	*98.78*	96.40	**99.36**	*98.83*	**80.37**	**93.74**	**96.76**
	ODIN	98.84	97.95	99.13	*99.00*	*98.99*	96.45	98.39	94.47	93.61	94.55	76.30	89.73	94.93
	GODIN	97.01	95.11	88.98	83.36	84.57	89.39	89.74	93.77	92.63	92.31	78.20	89.23	89.54
	SEM	97.43	96.74	98.62	96.99	97.11	94.80	96.95	41.25	47.11	42.75	-	43.70	79.20
	Energy	99.26	*98.87*	*99.41*	98.52	98.56	**98.04**	*98.78*	93.32	**99.36**	98.81	80.21	92.93	96.4
	Gradnorm	**99.96**	**99.72**	**99.83**	**99.47**	**99.42**	*97.94*	**99.39**	**96.63**	99.27	98.73	61.53	89.04	95.25
	MLS+ReAct	85.25	87.12	89.13	88.12	88.74	83.28	86.94	92.43	95.56	97.06	76.63	90.42	88.33
	ODIN+ReAct	84.27	85.71	86.13	88.12	85.35	88.07	86.78	85.83	84.21	90.46	75.25	83.94	85.64
	Energy+ReAct	81.40	87.30	83.98	82.49	84.35	83.72	83.87	95.89	98.80	82.73	*80.31*	89.43	86.09
	MLS+ASH	99.08	98.57	98.82	98.50	98.51	97.25	98.46	95.62	91.63	93.89	79.21	90.09	95.11
	MLS+SHE	98.22	95.08	97.96	91.08	91.36	90.85	94.09	83.19	84.53	82.75	78.80	82.32	89.38

Bold values represent the best results, while italic values represent the second best results. Different methods have their optimal scope but MLS and Energy demonstrate stability and models trained with OE dominate on almost all OOD datasets

Table 13 Results of various methods on OOD detection benchmarks using CIFAR10 as ID training data. The results are averaged from five independent runs

Training method	Scoring rule	Covariate shift			Semantic shift			
		ImageNet-C ID=63.05	ImageNet-R ID=76.13	AVG	ImageNet-SSB (easy/hard)		AVG	Overall
DenseNet121+CE	MSP	64.63	80.53	72.58	80.16	75.01	77.59	75.08
	MLS	**67.92**	86.71	**77.32**	80.28	75.05	77.67	77.49
DenseNet121+ARPL	MSP	61.85	78.68	70.27	74.56	75.27	74.92	72.59
	MLS	63.94	82.77	73.36	79.76	74.96	77.36	75.36
DinoViT-S/8	MSP	59.23	82.47	70.85	80.28	75.07	77.68	74.26
	MLS	64.00	**89.22**	76.61	**81.24**	**75.99**	78.62	**77.61**

We train a DenseNet121 with the CE loss and report the in distribution accuracy as 'ID'
Bold values reflect the best methods for each dataset

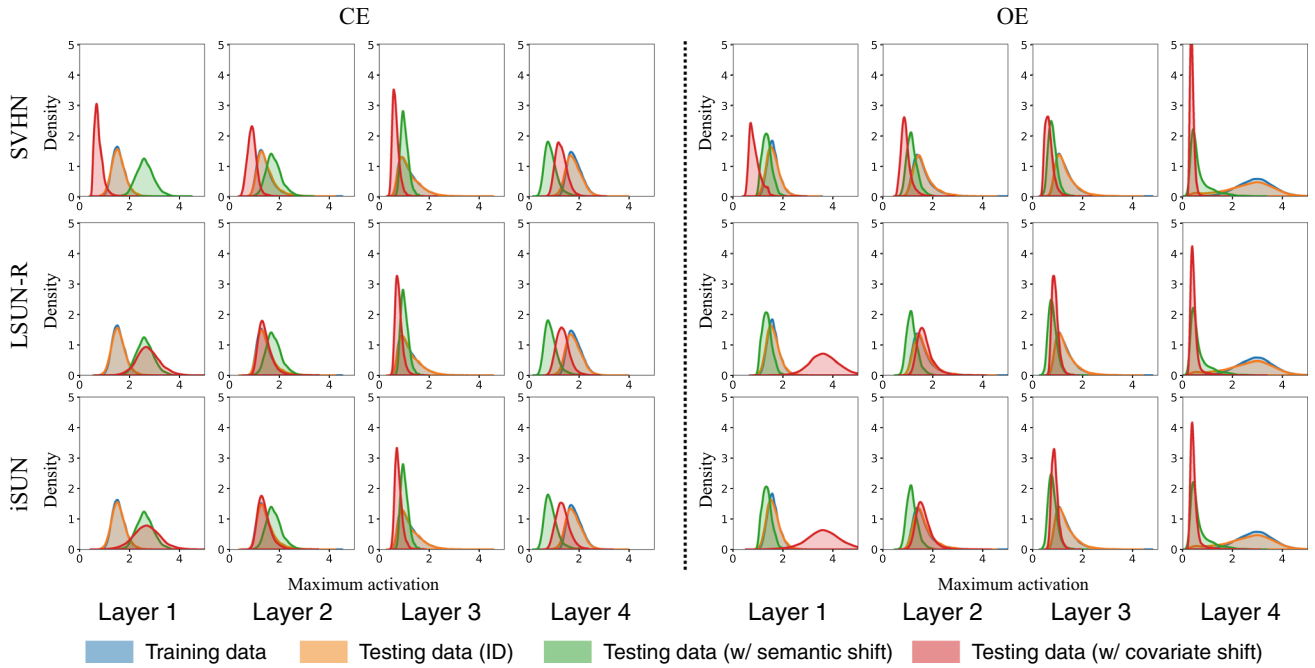

Fig. 12 Distribution of pre-unit activations for ResNet-18 pretrained on CIFAR10, evaluated on training and ID testing data; open-set data (from CIFAR-100) and OOD data (from SVHN, LSUN, iSUN). Specifically, each subplot shows the maximum activation (along the channel, width and height dimension) at the outputs from `layer_1` to `layer_4` of the ResNet-18 trained on CIFAR10, displayed from left to right in the figures

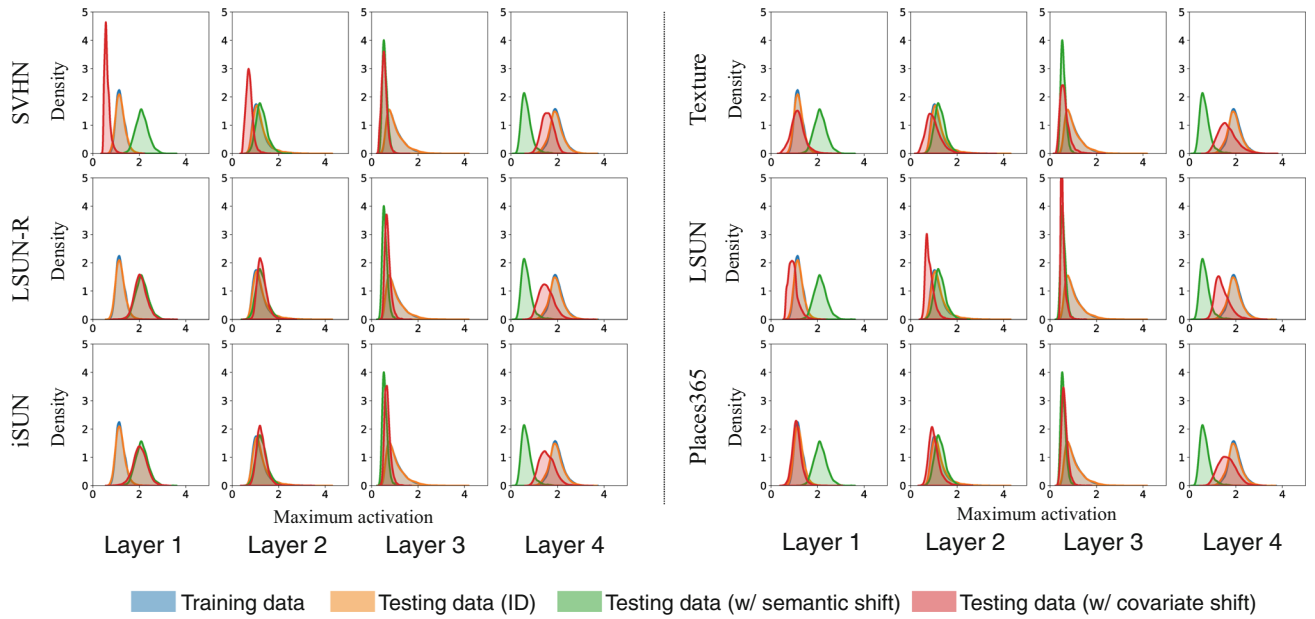

Fig. 13 Histogram of activations for ResNet-18 pretrained using ARPL+CS on a subset of CIFAR10 (with four training classes) and evaluated on: training and testing ID data; open-set data (disjoint six classes in CIFAR10) and OOD data. Specifically, each subplot shows the maximum activation (along channel, width and height dimension) at the outputs from layer 1 to layer 4 of the ResNet-18, displayed from left to right in the figures. Activation maps become notably separable in the last layer between ID and open-set data

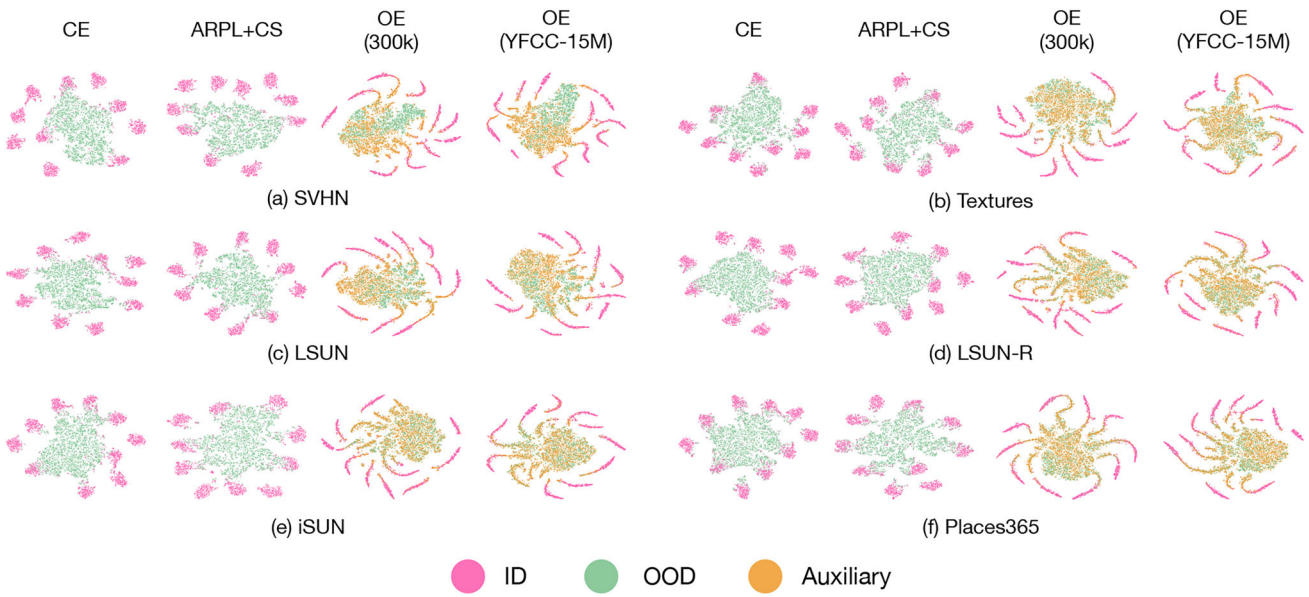

Fig. 14 t-SNEn extracted by models with CE/ARPL+CS/OE loss. Each point denotes a sample and its color denotes which distribution it comes from. The pink/green/brown dots stand for ID/OOD/auxiliary data respectively. Together with quantitative results shown in Table 2, we can observe that the performance boost can be achieved only when the auxiliary data distribution has sufficient overlap with the test-time OOD data distribution (*e.g.*, Texture and Places365) (Color figure online)

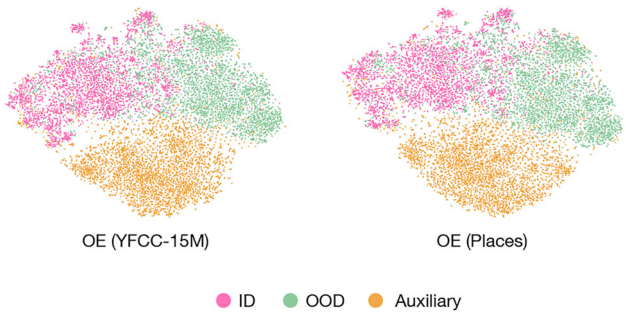

Fig. 15 t-SNE visualization of representations extracted by models with OE loss using different auxiliary data (*e.g.*, Places and YFCC-15M), tested on ImageNet-R. Each point denotes a sample and its color denotes which distribution it comes from. The pink/green/brown dots stand for ID/OOD/auxiliary data respectively. Together with quantitative results shown in Table 5, the dispersion of OOD and auxiliary data may lead to the unsatisfying performance boost on large-scale datasets

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-024-02222-4.

Acknowledgements This work is supported by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27208022); National Natural Science Foundation of China (Grant No. 62306251); a Facebook AI Research Scholarship; and HKU Seed Fund for Basic Research.

References

Ahmed, F., Bengio, Y., Seijen, H., & Courville, A. (2020). Systematic generalisation with group invariant predictions. In *ICLR*

Chen, J., Li, Y., Wu, X., Liang, Y., & Jha, S. (2021). Atom: Robustifying out-of-distribution detection using outlier mining. In *ECML PKDD*

Chen, G., Qiao, L., Shi, Y., Peng, P., Li, J., Huang, T., Pu, S., & Tian, Y. (2020). Learning open set network with discriminative reciprocal points. In *ECCV*

Chen, G., Peng, P., Wang, X., & Tian, Y. (2021). Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(11), 8065–8081.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., & Vedaldi, A. (2014). Describing textures in the wild. In *CVPR*

Deecke, L., Ruff, L., Vandermeulen, R.A., & Bilen, H. (2021). Transfer-based semantic anomaly detection. In *ICML*

DeVries, T., & Taylor, G.W. (2017). Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552

Dietterich, T. G., & Guyer, A. (2022). The familiarity hypothesis: Explaining the behavior of deep open set methods. *Pattern Recognition, 132*, 108931.

Djurisic, A., Bozanic, N., Ashok, A., & Liu, R. (2023). Extremely simple activation shaping for out-of-distribution detection. In *ICLR*

Ge, Z., Demyanov, S., & Garnavi, R. (2017). Generative openmax for multi-class open set classification. In *BMVC*

Geifman, Y., Uziel, G., & El-Yaniv, R. (2019). Bias-reduced uncertainty estimation for deep neural classifiers. In *ICLR*

Gretton, A., Fukumizu, K., Harchaoui, Z., & Sriperumbudur, B.K. (2009). A fast, consistent kernel two-sample test. In *NeurIPS*

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research, 13*(1), 723–773.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., & Guo, M. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. In *CVPR*

Hendrycks, D., Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*

Hendrycks, D., Mazeika, M., & Dietterich, T. (2019). Deep anomaly detection with outlier exposure. In *ICLR*

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D. (2021). Natural adversarial examples. In *CVPR*

Hsu, Y.-C., Shen, Y., Jin, H., & Kira, Z. (2020). Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *CVPR*

Huang, R., Geng, A., & Li, Y. (2021). On the importance of gradients for detecting distributional shifts in the wild. In *NeurIPS*

Kim, J., Koo, J., & Hwang, S. (2021). A unified benchmark for the unknown detection capability of deep neural networks. *Expert Systems with Applications, 229*, 120461.

Kong, S., & Ramanan, D. (2021). Opengan: Open-set recognition via open data generation. In *CVPR*

Krizhevsky, A., Hinton, G. (2009). Learning multiple layers of features from tiny images

Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In *NeurIPS*

Le, Y., & Yang, X. (2015). Tiny imagenet visual recognition challenge. In *CS 231N*

Liang, S., Li, Y., & Srikant, R. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*

Liang, S., Li, Y., & Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., & Sutherland, D.J. (2020). Learning deep kernels for non-parametric two-sample tests. In *ICML*

Liu, W., Wang, X., Owens, J., & Li, Y. (2020). Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems, 33*, 21464–21475.

Ming, Y., Fan, Y., & Li, Y. (2022). Poem: Out-of-distribution detection with posterior sampling. In *ICML*

Neal, L., Olson, M., Fern, X., Wong, W.-K., & Li, F. (2018). Open set learning with counterfactual images. In *ECCV*

Neal, L., Olson, M., Fern, X., Wong, W.-K., & Li, F. (2018). Open set learning with counterfactual images. In *ECCV*

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *ICML*

Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*, 211–252.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision, 115*, 211–252.

Sagawa, S., Koh, P.W., Hashimoto, T.B., & Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*

Salehi, M., Mirzaei, H., Hendrycks, D., Li, Y., Rohban, M.H., Sabokrou, M. (2021). A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. In *TMLR*

Scheirer, W. J., Rocha, A., Sapkota, A., Boult, T. E. (2013). Towards open set recognition. In *IEEE TPAMI*

Scheirer, W. J., Rezende Rocha, A., Sapkota, A., & Boult, T. E. (2012). Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(7), 1757–1772.

Sun, Y., Guo, C., & Li, Y. (2021). React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems, 34*, 144–157.

Tian, J., Hsu, Y.-C., Shen, Y., Jin, H., & Kira, Z. (2021). Exploring covariate and concept shift for out-of-distribution detection. In *NeurIPS Workshop*

Tran, D., Liu, J., Dusenberry, M.W., Phan, D., Collier, M., Ren, J., Han, K., Wang, Z., Mariet, Z., Hu, H., et al. (2022). Plex: Towards reliability using pretrained large model extensions. arXiv preprint arXiv:2207.07411

Vaze, S., Han, K., Vedaldi, A., & Zisserman, A. (2022). Open-set recognition: a good closed-set classifier is all you need? In *ICLR*

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset. In *Technical Report CNS-TR-2011-001*, California Institute of Technology

Wang, Q., Fang, Z., Zhang, Y., Liu, F., Li, Y., & Han, B. (2024). Learning to augment distributions for out-of-distribution detection. In *NeurIPS*

Wang, Q., Ye, J., Liu, F., Dai, Q., Kalander, M., Liu, T., Hao, J., & Han, B. (2023). Out-of-distribution detection with implicit outlier transformation. In *ICLR*

Wiles, O., Gowal, S., Stimberg, F., Rebuffi, S.-A., Ktena, I., Dvijotham, K.D., & Cemgil, A.T. (2022). A fine-grained analysis on distribution shift. In *ICLR*

Xia, G., & Bouganis, C.-S. (2022). Augmenting softmax information for selective classification with out-of-distribution data. In *ACCV*

Xu, P., Ehinger, K.A., Zhang, Y., Finkelstein, A., Kulkarni, S.R., & Xiao, J. (2015). Turkergaze: Crowdsourcing saliency with webcam based eye tracking. arXiv preprint arXiv:1504.06755

Yang, J., Wang, H., Feng, L., Yan, X., Zheng, H., Zhang, W., & Liu, Z. (2021). Semantically coherent out-of-distribution detection. In *ICCV*

Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., & Sun, Y. (2022). Openood: Benchmarking generalized out-of-distribution detection. In *NeurIPS*

Yang, J., Zhou, K., Li, Y., & Liu, Z. (2024). Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*. https://doi.org/10.1007/s11263-024-02117-4

Yang, J., Zhou, K., & Liu, Z. (2023). Full-spectrum out-of-distribution detection. *International Journal of Computer Vision, 131*(10), 2607–2622.

Ye, N., Li, K., Bai, H., Yu, R., Hong, L., Zhou, F., Li, Z., & Zhu, J. (2022). Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *CVPR*

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365

Zhang, H., CissÃ, M., Dauphin, Y., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. In *ICLR*

Zhang, J., Fu, Q., Chen, X., Du, L., Li, Z., Wang, G., Han, S., & Zhang, D. (2023). Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *ICLR*

Zhao, B., Yu, S., Ma, W., Yu, M., Mei, S., Wang, A., He, J., Yuille, A., & Kortylewski, A. (2022). Robin: A benchmark for robustness to individual nuisances in real-world out-of-distribution shifts. In *ECCV*

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(6), 1452–1464.