HOW AND WHY WE DETECT DISTRIBUTION SHIFT: CRITICAL ANALYSIS OF METHODS AND BENCHMARKS

Anonymous authors
Paper under double-blind review

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: open-set recognition (OSR) and out-of-distribution detection (OOD). 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) For the first time, we perform rigorous cross-evaluation between state-of-the-art methods in the OOD 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 and OSR; (iii) We thoroughly examine SOTA methods for OOD and OSR on our large-scale benchmark; and (iv) Finally, we find that the best performing method on previous benchmarks struggles on our large-scale benchmark, while magnitude-aware scoring rules consistently show promise.

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 detection* (OOD) and *open-set recognition* (OSR) 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 unfamiliar with respect to their training set, with a variety of methods proposed within each field. OSR methods are developed for detecting test images which come from different semantic categories to the training set, while OOD methods are developed for detecting images which come from a different data distribution to the training images. Research efforts in both directions largely occur independently (with little cross-pollination of ideas). Though prior work has recognized the similarity of the two sub-fields (Vaze et al., 2022; Tran et al., 2022; Yang et al., 2021; Salehi et al., 2021), OOD and OSR, there have been no rigorous benchmarking to understand the underlining principles of methods for both.

In this work, for the first time, we perform rigorous cross-evaluation between methods developed for OOD and OSR on current standard benchmarks, suggesting that methods which perform well for one are likely to perform well for the other in Sec. 3. We experiment both with methods which require alternate 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.*, Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2017), Maximum Logit Score (MLS) (Vaze et al., 2022) and Energy Scoring (Liu et al., 2020)). We thoroughly evaluate all methods on both standard OOD and OSR benchmarks, after which we find that OE achieves almost saturating performance on the OOD task and also obtains the state-of-the-art (SOTA) results on the OSR task. Meanwhile, we also find that the magnitude-aware scoring rules like MLS (Vaze et al., 2022) and Energy Scoring (Liu et al., 2020) show steady good performance across different methods and datasets.

Next, we propose a reconciling perspective on the tasks tackled by the two fields, and propose a new benchmark to assess this in Sec. 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 and OSR. For example, to isolate semantic shift, we leverage the recently introduced Semantic Shift Benchmark (SSB) (Vaze et al., 2022) containing ImageNet-scale datasets, in which the original ImageNet-1K (Russakovsky et al., 2015b) is regarded as 'seen' closed-set data while 'unseen' data is carefully drawn from the disjoint set of ImageNet-21K-P (Ridnik et al., 2021b). For covariate shift, we leverage ImageNet-C (Hendrycks & Dietterich, 2019) and ImageNet-R (Hendrycks et al., 2020) to demonstrate distribution shift with respect to the standard ImageNet dataset.

Finally, we examine SOTA methods developed for OOD and OSR on this large-scale benchmark to validate whether the findings through the standard (small-scale) datasets still hold on our consolidated large-scale benchmarking. Through the large-scale analysis, we surprisingly find that OE struggles to scale to larger benchmarks, while the magnitude-aware scoring rules, MLS (Vaze et al., 2022) and Energy Scoring (Liu et al., 2020), still show promise. We further provide empirical insights by analysing the representations extracted by different models on data under different distribution shifts, which suggests that the strong performance of OE on the standard benchmark is partially attributed to the fact that the auxiliary OOD data used for training has sufficient distribution overlap with the OOD testing data, while it is not straightforward to come up with an auxiliary OOD data to reflect the actual distribution shift on large-scale datasets. We believe there are still many more open questions to be answered in the shared space of OOD and OSR, and hope the findings in our work can serve as a starting point to have a deeper look into them.

2 Preliminaries and Related Work

Open-set recognition. Previous work (Scheirer et al., 2012) coined "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., 2018b; Kong & Ramanan, 2021) show that the generated data from synthetic distribution would be helpful to improve OSR. OSRCI (Neal et al., 2018b) generates images belonging to the unknown classes but similar to the training data to train an open-set classifier. (Kong & Ramanan, 2021) adversarially trained discriminator to distinguish closed from open-set images and introduced real open-set samples for model selection. Prototype-based methods (Chen et al., 2020; 2021) adjust the boundaries of different classes and identify open-set images based on distances to the learned prototypes of known classes.

Out of Distribution Detection. (Hendrycks & Gimpel, 2017) formalized the task of out-of-distribution detection and provided 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) added adversarial perturbations to both ID and OOD samples and employed 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) rectified the distribution of per-unit activations in the penultimate layer for ID and OOD data. Outlier Exposure (OE) (Hendrycks et al., 2019) and (Huang et al., 2021) both designed 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 leveraged real OOD data for training while the latter directly employed the vector norm of gradients to perform uncertainty estimation.

3 Analysis of SOTA Baselines on Standard Benchmarks

In this section, we perform cross-evaluation of methods from the OOD and OSR literature.

3.1 Experimental setup

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

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., 2022). We further experiment with ReAct (Sun et al., 2021), an activation pruning technique which can be employed in tangent with any scoring rule. While MLS was developed for OSR (Vaze et al., 2022), other scoring rules were developed for OOD detection. We provide descriptions of each scoring rule in the appendix.

For *specialised 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 literature, where real outlier examples are used during training as examples of OOD. In this case, the model is encouraged to predict a uniform softmax output.

Datasets. For the OOD setting, we train models on CIFAR10 (Krizhevsky et al., 2009). 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). We also perform OOD experiments training with CIFAR100 as ID in Appendix E.

For the OSR benchmark, following the standard protocols in (Neal et al., 2018a), 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 CIFAR100. Note that, for a given method, benchmarking on OOD involves training a single model and evaluating on multiple downstream datasets. In contrast, OSR benchmarks involve training a different model for each evaluation.

Training configurations. Due to limited space, we give a detailed description and experimental results of each configuration in Appendix A. Broadly speaking, we train a ResNet-18 on the ID data, with an SGD optimizer and cosine annealing schedule. We train ARPL + CS and OE largely based on the official public implementation. For the auxiliary outlier dataset in the OE loss, we follow (Hendrycks et al., 2019) and use a subset of 80 Million Tiny Images (Torralba et al., 2008) with 300K images, removing all examples that appear in CIFAR10/100, Places or LSUN classes.

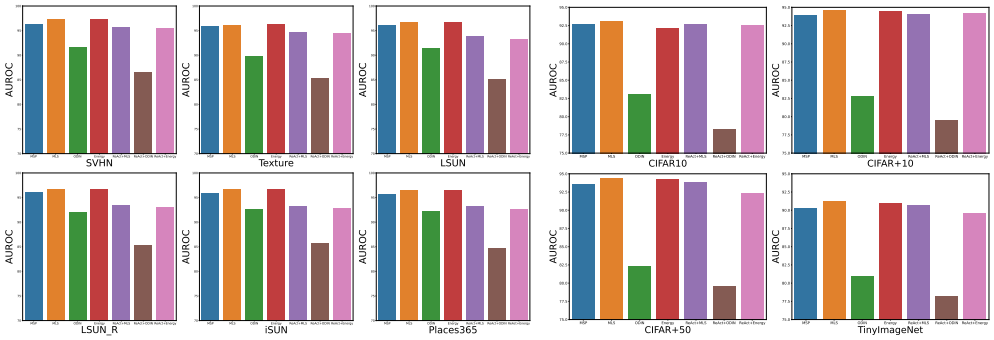
Metrics. Following standard practise in both OOD and OSR tasks, we use the Area Under the Receiver Operating characteristic Curve (AUROC) as an evaluation metric throughout this paper. We found that other metrics, such as the FPR95 (Hendrycks & Gimpel, 2017) (also known as the false alarm rate), were correlated strongly with the AUROC.

3.2 Quantitative results

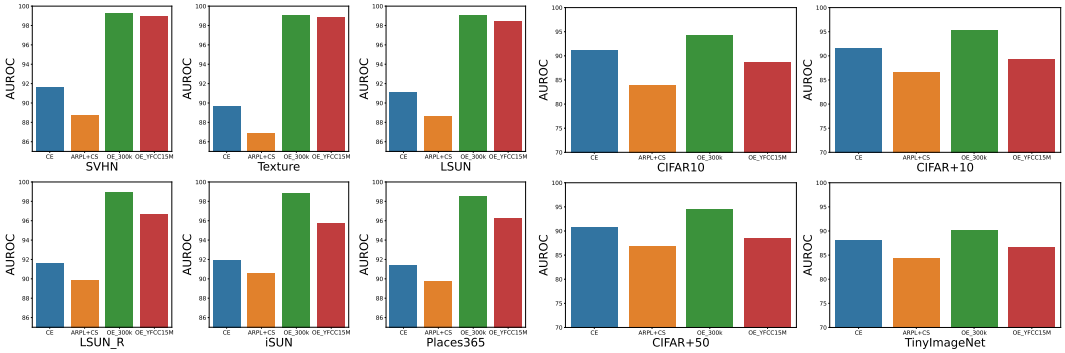
We present results from our benchmarking in Table 1. Although there is not always one clear winner when it comes to methodology, we observe two main takeaways.

Firstly, MLS and Energy tend to perform best across OOD and OSR datasets (Fig. 1(a)). We hypothesize this is because both are sensitive to the magnitude of the feature vector before the networks' linear layer. This phenomenon was observed in (Vaze et al., 2022), as unfamiliar examples tend to have lower feature norms than ID samples, providing a strong signal for the distribution shift decision. Interestingly, we also find that, ReAct, which has been shown to be effective in the literature, does not seem to bring performance gain in the well trained models with a high in-distribution accuracy. Here, we follow (Vaze et al., 2022) to obtain as much high in-distribution accuracy as possible for all models. It appears that when the classifier is strong enough, it is difficult for ReAct to bring extra improvement.

Secondly, we observe that Outlier Exposure (Hendrycks et al., 2019) provides excellent performance on the OOD benchmarks (Fig. 1(b)), often nearly saturating performance. It also often boosts OSR performance, though to a lesser degree, a phenomenon which we explore in the next section.



(a) Performance of different scoring rules across OOD and OSR tasks.



(b) Performance of different training methods across OOD and OSR tasks.

Figure 1: Analysis of different scoring rules and training methods on standard benchmarks. (a) Among various scoring rules, MLS and Energy show their reliability across OOD and OSR datasets (b) For different training methods, Outlier Exposure using different auxiliary data obtains an obvious performance boost compared with others on OOD and have slight gains on OSR.

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 benchmarks. Specifically, we use the value of the maximally activated neuron at various layers to analyze how the networks respond to distribution shift. We pass every related sample through the network, and plot the histogram of maximum activations at every layer in Fig. 2.

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 & 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. Fig. 2 shows histograms of the maximum activations at the outputs from `layer_1` to `layer_4` of a ResNet-18 (He et al., 2016) trained on CIFAR10 when evaluated on data and data with different shifts. Note that here we use 'layer' to refer to ResNet block.

For OSR 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 discover 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. Interestingly, some datasets (like SVHN) which showed markedly different early layer activations actually display *more similar* activations at later layers.

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

Table 1: Evaluation on small-scale OOD and OSR benchmarks with various methods, using CIFAR10 as ID. We report the in-distribution accuracy as 'ID' and denote intractable results as '-', resulting from unaffordable computational cost. Different methods have their optimal scope but MLS/Energy show their stability and models trained with OE dominate almost all OOD datasets.

(a) Evaluation based on ResNet-18 trained with the CE loss.

Method	OOD 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	
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
CE+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**
CE+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
CE+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
CE+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
CE+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**
CE+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
CE+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
CE+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

(b) Evaluation based on ResNet-18 trained with the ARPL+CS loss.

Method	OOD benchmarks							OSR benchmarks					Overall
	SVHN	Textures	LSUN	LSUN-R	iSUN	Places365	AVG ID=91.02	CIFAR10 ID=96.96	CIFAR+10 ID=96.77	CIFAR+50 ID=96.69	TinyImageNet ID=86.91	AVG	
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
ARPL+CS+MLS	**96.36**	90.20	**96.59**	**96.95**	**96.88**	**93.29**	**95.05**	**93.16**	96.58	94.67	**84.79**	**92.3**	**93.95**
ARPL+CS+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
ARPL+CS+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
ARPL+CS+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
ARPL+CS+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**
ARPL+CS+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
ARPL+CS+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
ARPL+CS+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**

(c) Evaluation based on ResNet-18 trained with the OE loss.

Method	OOD benchmarks							OSR benchmarks					Overall
	SVHN	Textures	LSUN	LSUN-R	iSUN	Places365	AVG ID=94.16	CIFAR10 ID=97.8	CIFAR+10 ID=98.3	CIFAR+50 ID=97.92	TinyImageNet ID=83.4	AVG	
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**
OE+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**
OE+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
OE+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
OE+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
OE+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**
OE+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
OE+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
OE+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
OE+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

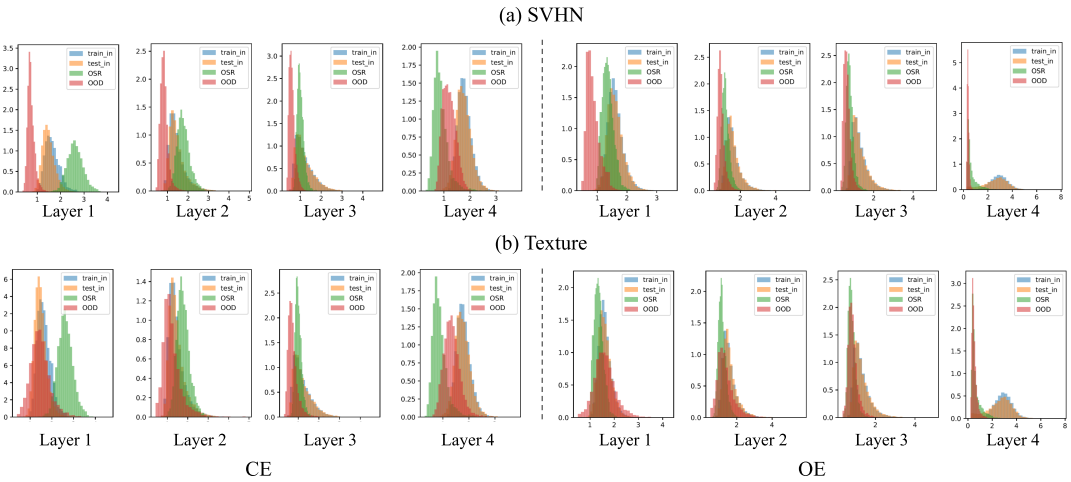


Figure 2: Histogram of activations for ResNet-18 pretrained on a subset of CIFAR10 with four training classes and evaluated on: training and ID testing data; OSR data (disjoint six classes in CIFAR10) and OOD data (from SVHN and Textures). Specifically, each subplot shows the maximum activation (along channel, width and height dimension) at the outputs from `layer_1` to `layer_4` of a ResNet-18, displayed from left to right in the figures. Results on more OOD datasets are shown in Appendix C. The behavior of OE is different from CE, whose activation maps become more separable in the deeper rather than the shallower layers.

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 A CONSOLIDATED BENCHMARKING OF DISTRIBUTION SHIFT

Having analyzed methodologies for detecting distribution shift across the OOD 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 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, orthognal, 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:

$$z \sim p(z) \qquad y^i \sim p(y^i|z) \quad i \in \{1...K\} \qquad \mathbf{x} \sim p(\mathbf{x}|z) \tag{1}$$

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 'shape', 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 car, which has been trained to recognize 'cars' and 'pedestrians' 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., 2015a) 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., 2021a). Meanwhile, CUB-SSB splits the 200 bird classes 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. We further report figures on SCars-SSB (Krause et al., 2013) and FGVC-Aircraft-SSB (Maji et al., 2013) in the appendix.

For *covariate shift*, we propose ImageNet-C (Hendrycks & Dietterich, 2019) and ImageNet-R (Hendrycks et al., 2020) 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 scenery.

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 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.

4.2 QUANTITATIVE ANALYSIS

In Tables 2 and 3, we evaluate a selection of previously discussed methods on our large-scale benchmark for both OOD and OSR. Through this large-scale evaluation, we find that *in terms of training methods, among CE, ARPL (+CS), and OE, there is no clear winners across the board.* It is surprising that the best performer on the previous small scale benchmarks (see Table 1), OE, appears to be struggling (last two rows in Table 2) on the large-scale benchmark. This is contradicting with the finding on the small scale benchmarks, which we will analyse in the next subsection. *In terms of scoring rules, the maganitude-awere 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).*

Table 2: Evaluation on large-scale OOD and OSR benchmarks using ResNet-50 model trained with different losses and scoring rules. Models trained with the CE loss outperforms the ones with ARPL on both covariate shift and semantic shift.

(a) Evaluation based on ResNet-50 trained with the CE loss.

Method	Covariate Shift					Semantic Shift					Overall
	ImageNet-C ID=63.05	ImageNet-R ID=76.13	Waterbird (Easy/Hard)		AVG	ImageNet-SSB (Easy/Hard)		CUB (Easy/Hard)		AVG	
CE+MSP	64.63	80.53	81.65	**75.33**	75.54	**80.16**	**75.01**	88.11	**79.43**	**80.68**	78.11
CE+MLS	**67.92**	**86.71**	81.87	75.18	77.92	**80.28**	**75.05**	88.29	**79.33**	**80.74**	**79.33**
CE+ODIN	63.69	85.62	79.51	71.54	75.09	74.56	75.27	86.24	73.88	77.49	76.29
CE+Energy	**68.05**	**87.04**	**82.49**	74.60	**78.05**	**79.76**	**74.96**	**88.81**	**79.06**	**80.65**	**79.35**
CE+MLS+ReAct	66.64	84.82	81.69	75.12	77.07	**80.28**	75.07	88.29	**79.33**	**80.74**	78.91
CE+ODIN+ReAct	61.69	83.25	79.48	71.50	73.98	74.56	**75.29**	86.24	73.88	77.49	75.74
CE+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.

Method	Covariate Shift					Semantic Shift					Overall
	ImageNet-C ID=63.05	ImageNet-R ID=76.13	Waterbird (Easy/Hard)		AVG	ImageNet-SSB (Easy/Hard)		CUB (Easy/Hard)		AVG	
ARPL+MSP	61.85	78.68	79.42	**72.30**	73.06	**79.90**	**74.67**	**83.53**	**75.64**	**78.44**	75.75
ARPL+MLS	**63.94**	**82.77**	79.48	72.09	**74.57**	**79.92**	**74.60**	**83.50**	**75.49**	**78.38**	**76.47**
ARPL+ODIN	61.88	77.03	73.76	69.26	70.48	68.72	71.23	73.87	69.77	70.90	70.69
ARPL+Energy	**64.13**	**83.25**	**79.64**	71.86	**74.72**	**79.87**	**74.49**	**83.70**	**75.46**	**78.38**	**76.55**
ARPL+MLS+ReAct	62.69	80.69	79.44	72.07	73.72	**79.92**	**74.60**	83.44	**75.43**	**78.35**	76.04
ARPL+ODIN+ReAct	62.23	76.08	73.75	69.23	70.32	68.72	71.23	67.42	63.91	67.82	69.07
ARPL+Energy+ReAct	62.89	81.17	79.60	71.83	73.87	**79.87**	74.49	**83.70**	**75.41**	**78.37**	76.12

Table 3: Results of OOD and OSR benchmarks on large-scale datasets, using ResNet-50 model trained with the OE loss compared with CE and ARPL baselines. We separately introduce outlier data from different data sources including Places and YFCC15M to feed OE.

Method	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+MLS	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 (with Places)+MLS	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 (with YFCC15M)+MLS	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

4.3 WHY DOES OE UNDERPERFORM ON LARGE-SCALE DATASETS?

Here, we investigate why OE underperforms other methods on large-scale benchmark. One critical difference between OE and other methods is that OE requires auxiliary OOD data for training. Intuitively, if the distribution of the auxiliary OOD training data can reflect the distribution of the actual OOD testing data, we would expect a better performance on detecting, while incomplete or biased outlier data may hurt the learning. In fig. 3, we visualize the t-SNE projection of the representations for in-distribution (ID) data (*i.e.*, CIFAR10), auxiliary training OOD data (300K (Hendrycks et al., 2019) vs YFCC15M), and different test-time OOD datasets. As can be seen, using the 300K images generally shows a better overlap with the test-time OOD data. Hence, OE trained with 300K as the auxiliary OOD data achieves better performance than the counterpart trained with YFCC15M (table 1 vs table 7).

For the experiments on standard (small-scale) benchmark, we experiment using 300K images v.s YFCC15M as the auxiliary training data. While for the experimentes on our large-scale benchmark,

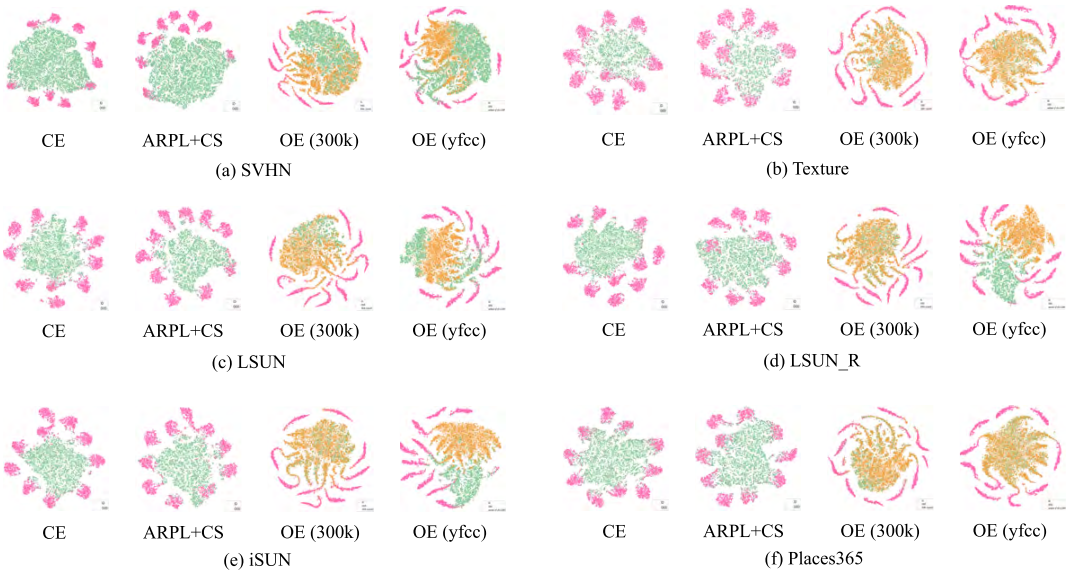


Figure 3: t-SNE visualization of representations extracted by models with CE/ARPL+CS/OE loss. Each point denotes a sample and its color denotes which distribution it comes from. The pink/green/brown dots stand for ID/OOD/auxiliary data respectively. Together with quantitative results shown in Table 1 and those in Table 7, we can observe that the performance boost can be achieved only when the auxiliary data distribution has sufficient overlap with the test-time OOD distribution (e.g. Texture and Places365).

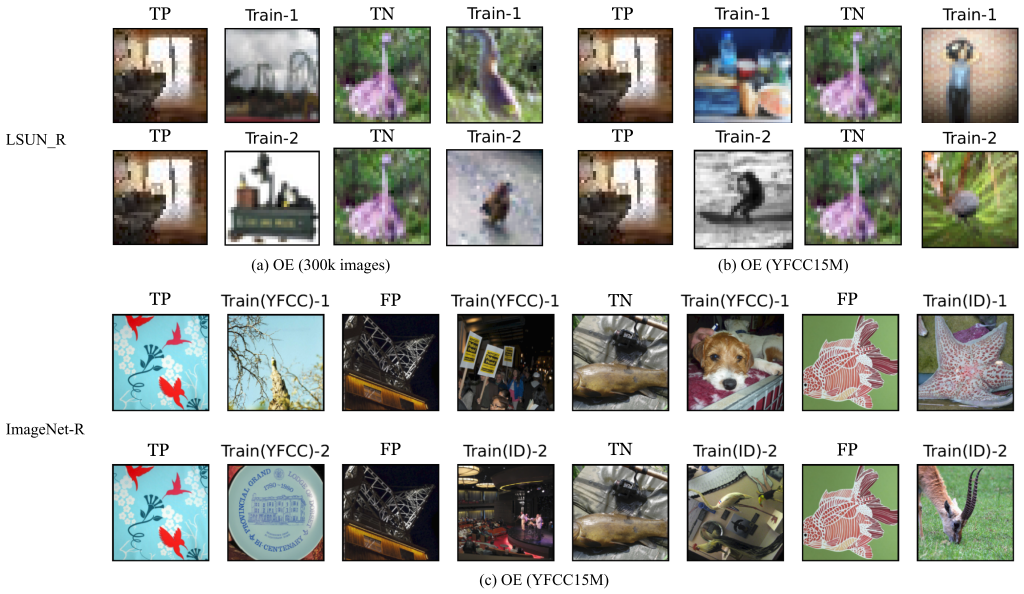
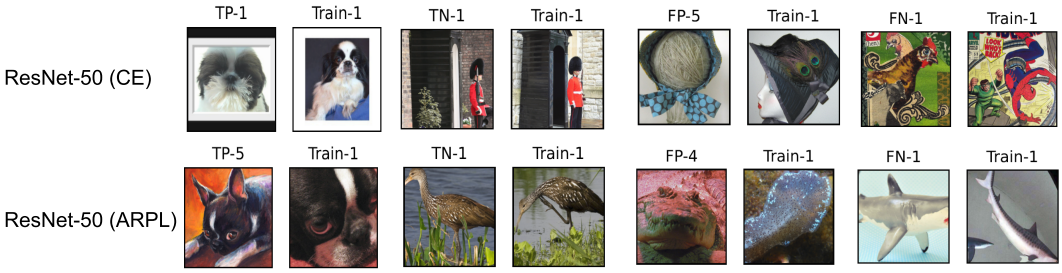
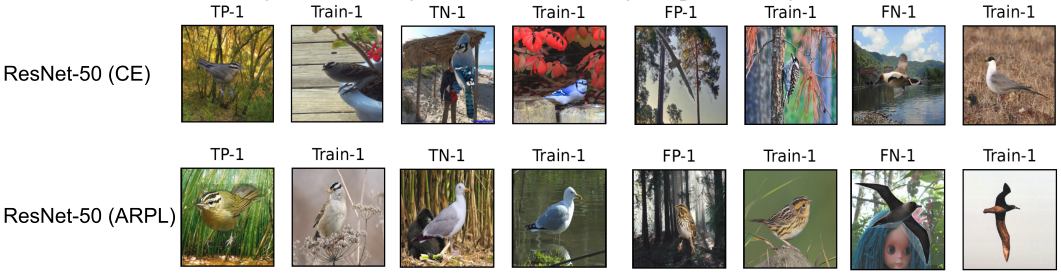


Figure 4: Nearest neighbours of samples in small-scale (e.g. LSUN_R) and large-scale (e.g. ImageNet-R) OOD datasets.

we use YFCC15M because 300K images are not competent in large-scale setting. In fig. 4 (first macro row), as a more fine-grained analysis on standard benchmark, we identify the most confident true positive (TP, *i.e.*, correctly predicted OD sample) and the most confident true negative (TN, correctly predicted ID sample) with MLS scoring, and then find their top-k nearest samples from the auxiliary training OOD data according to feature similarities. As can be seen, for a TP sample,



(a) Nearest Neighbours in ImageNet-1K for confusing samples in ImageNet-R dataset.



(b) Nearest Neighbours in CUB training set for confusing samples in Waterbirds dataset.

Figure 5: We show confusion samples under both covariate shift and semantic shift with testing examples for ResNet-50. Posture/viewpoint, semantic similarity, color and object in the background may cause the confusion.

the NNs retrieved from 300K are more similar than those retrieved from YFCC15M in to the OOD testing sample. This is consistent with the finding in fig. 3. Further, we carry out a similar experiment on our large-scale benchmark (see fig. 4, second macro row), by retrieving the NNs from the union of ID training set and YFCC15. We observe that the retrieved NN for TPs are less similar to the TPs, suggesting less similarities between the test-time OOD data and the auxiliary training OOD data.

Finally, in fig. 5, for CE and ARPL models, in which no auxiliary OOD training data are used, we also identify the most confident predictions (TP and TN) and most confusing predictions (FN and FP) on our benchmarking datasets, ImageNet-R and Waterbirds, and retrieve their NNs from the training ID datasets (*i.e.*, ImageNet and CUB). *We observe that the current decision paradigm is susceptible to posture, viewpoint, background color/object and even semantic similarity.* FNs and FPs are often mislead by posture (*e.g.*, FN-1 from the ARPL model in Waterbirds), viewpoint (*e.g.*, FP-5 and FP-4 in ImageNet-1K), background (*e.g.*, FN-1 from the CE model in Waterbirds) and semantic similarity (*e.g.*, cartoon style in FN-1 from the CE model in ImageNet-1K and the shark from the ARPL one). More results and analysis can be found in the appendix.

5 Conclusion

In this paper, we have provided a consolidated exploration of Out-of-Distribution detection (OOD) and Open-set Recognition (OSR). We performed rigorous cross-evaluation between methods developed for OOD and OSR and identified a strong correlation between their performances. We also proposed a new, large-scale benchmark setting, to disentangle the OOD and OSR, by breaking the distribution shift problem down into *covariate* shift and *semantic* shift, suggesting large-scale evaluation protocols for the task. We also showed that the best performing method on both OSR and OOD does not generalize well to our challenging large-scale benchmark, and found that magnitude-aware scoring rules are generally more reliable than the others. We believe our new benchmark can serve as a better testbed to measure progresses in OSR and OOD, and hope our findings in this work can shed light for further exploration on the shared space of OSR and OOD and foster new development for this field.

REFERENCES

Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *ECCV*, 2020.

Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE TPAMI*, 2021.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Thomas G. Dietterich and Alexander Guyer. The familiarity hypothesis: Explaining the behavior of deep open set methods. *ArXiv e-prints*, 2022.

Zongyuan Ge, Sergey Demyanov, and Rahil Garnavi. Generative openmax for multi-class open set classification. In *BMVC*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, F. Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *CVPR*, pp. 10951–10960, 2020.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *NeurIPS*, 2021.

Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *CVPR*, pp. 813–822, 2021.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *International IEEE Workshop on 3D Representation and Recognition (3dRR)*, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 2015.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *NeurIPS*, 2020.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *ECCV*, 2018a.

Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *ECCV*, 2018b.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *NeurIPS*, 32, 2019.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021a.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *ArXiv e-prints*, 2021b.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015a.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015b.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *ICLR*, 2019.

Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv*, 2021.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE TPAMI*, 2012.

Walter J. Scheirer, Anderson Rocha, Archana Sapkota, and Terrance E. Boult. Towards open set recognition. *IEEE TPAMI*, 2013.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *NeurIPS*, 2021.

Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE TPAMI*, 2008.

Dustin Tran, Jeremiah Liu, Michael W. Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, Neil Band, Tim G. J. Rudner, Karan Singhal, Zachary Nado, Joost van Amersfoort, Andreas Kirsch, Rodolphe Jenatton, Nithum Thain, Honglin Yuan, Kelly Buchanan, Kevin Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, and Balaji Lakshminarayanan. Plex: Towards reliability using pretrained large model extensions. In *ArXiv e-prints*, 2022.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: a good closed-set classifier is all you need? In *International Conference on Learning Representations*, 2022.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *ICLR*, 2022.

Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *ArXiv e-prints*, 2021.

Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *ArXiv e-prints*, 2022.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Robin: A benchmark for robustness to individual nuisances in real-world out-of-distribution shifts. *ECCV*, 2022.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. In *IEEE TPAMI*, 2017.

A DETAILED TRAINING CONFIGURATION FOR CIFAR BENCHMARKS

For training ResNet-18 on CIFAR10, we set the initial learning rate 0.1 and apply cosine annealing schedule, using SGD with 0.9 momentum. The weight decay factor is set to $5e^{-4}$. The total training epochs are 200 and the batch size is 128.

For CIFAR100, we also set the initial learning rate 0.1, but divided by 5 at 60th, 120th, 160th epochs, train for 200 epochs with a batch size of 128 and weight decay 5e-4, Nesterov momentum of 0.9, following (DeVries & Taylor, 2017)

For *ReAct config*, the models are trained with a batch size of 128 for 100 epochs. The start learning rate is 0.1 and decays by a factor of 10 at epochs 50, 75, and 90. For *MLS config*, we train the models with a batch size of 128 for 600 epochs with a cosine annealed learning rate, restarting the learning rate to the initial value at epochs 200 and 400. Besides, we linearly increase the learning rate from 0 to the initial value at the beginning. The initial learning rate is 0.1 for CIFAR but 0.01 for TinyImageNet.

B EXPERIMENTAL RESULTS ON LARGE-SCALE BENCHMARKS

Table 4: Results of OOD and OSR benchmarks on large-scale datasets, using ResNet-50 model trained with the CE loss.

Method	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
CE+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**
CE+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
CE+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**
CE+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
CE+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
CE+Energy+ReAct	66.88	83.92	75.4	79.76	74.99	88.81	79.06	82.48	74.55	93.89	82.00	90.80	82.79	82.91	81.66

Table 5: Results of OOD and OSR benchmarks on large-scale datasets, using ResNet-50 model trained with the ARPL loss.

Method	Covariate Shift			Semantic Shift											Overall
	ImageNet-C ID=65.33	ImageNet-R ID=78.39	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
ARPL+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**
ARPL+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
ARPL+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**
ARPL+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
ARPL+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
ARPL+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

Table 6: Results of OOD and OSR benchmarks on large-scale datasets, using ResNet-50 model trained with the OE loss combined with auxiliary data from Places.

Method	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
OE+MLS	61.77	80.53	**79.31**	**73.88**	**73.87**	82.42	**75.58**	**79.16**	**73.83**	**77.75**	**75.81**
OE+ODIN	57.74	**82.31**	71.28	69.30	70.16	81.75	70.87	73.71	66.05	73.10	71.63
OE+Energy	**64.10**	81.11	76.39	70.86	73.12	**83.47**	**75.61**	78.56	73.01	**77.66**	**75.39**
OE+MLS+ReAct	62.39	79.76	77.00	71.93	72.77	81.23	73.07	72.09	70.16	74.14	73.45
OE+ODIN+ReAct	58.28	77.94	69.74	70.06	69.01	80.27	70.54	74.4	68.00	73.30	71.15
OE+Energy+ReAct	62.26	80.91	75.32	69.71	72.05	82.10	73.79	77.30	70.74	75.98	74.02

Table 7: Results of OOD and OSR benchmarks on small-scale datasets, using ResNet-18 model trained with the OE loss combined with auxiliary data from YFCC-15M.

Method	OOD benchmarks							OSR benchmarks					Overall
	SVHN	Textures	LSUN	LSUN-R	iSUN	Places365	AVG ACC=95.47	CIFAR10 ACC=97.52	CIFAR+10 ACC=97.76	CIFAR+50 ACC=97.29	TinyImageNet ACC=87.3	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
OE+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
OE+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
OE+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
OE+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
OE+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
OE+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

Table 8: Results of OOD and OSR benchmarks on large-scale datasets, using ResNet-50 model trained with the OE loss combined with auxiliary data from YFCC-15M.

Method	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
OE+MLS	**64.12**	**82.01**	**79.72**	**74.08**	**74.98**	79.37	72.55	75.19	**70.28**	74.35	74.67
OE+ODIN	60.80	74.36	66.80	68.94	67.73	72.01	66.87	71.48	67.91	69.57	68.65
OE+Energy	**64.31**	81.50	77.95	72.76	74.13	**81.50**	**74.33**	**77.78**	**70.09**	**75.93**	**75.03**
OE+MLS+ReAct	60.15	79.42	76.99	73.58	72.54	72.30	72.74	73.46	69.67	72.04	72.29
OE+ODIN+ReAct	60.98	74.05	66.10	68.89	67.51	64.67	61.79	72.26	67.76	66.62	67.06
OE+Energy+ReAct	61.87	79.79	78.83	71.98	73.12	71.24	73.26	75.78	69.85	72.53	72.83

C ACTIVATIONS OF OOD AND OSR AT DIFFERENT LAYERS

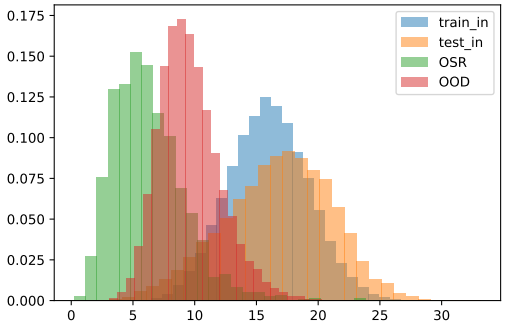


Figure 6: Histogram of the maximum logit scores for a ResNet-18 pretrained on CIFAR10. This is evaluated on the training data (`train_in`), ID testing data (`test_in`), OOD data (`OOD` from SVHN) and OSR data (`OSR` from CIFAR100).

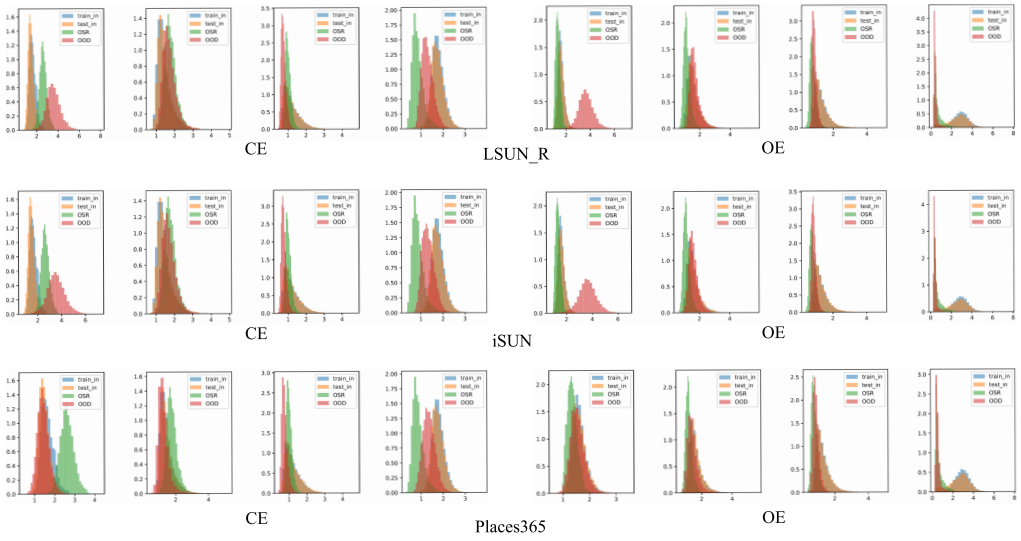


Figure 7: Histogram of activations for ResNet-18 pretrained on CIFAR10, evaluated on training and ID testing data; OSR data (from CIFAR100) and OOD data (from LSUN, LSUN_R and Places365). Specifically, each subplot shows the maximum activation (along channel, width and height dimension) at the outputs from `layer_1` to `layer_4` of a ResNet-18 trained on CIFAR10, displayed from left to right in the figures.

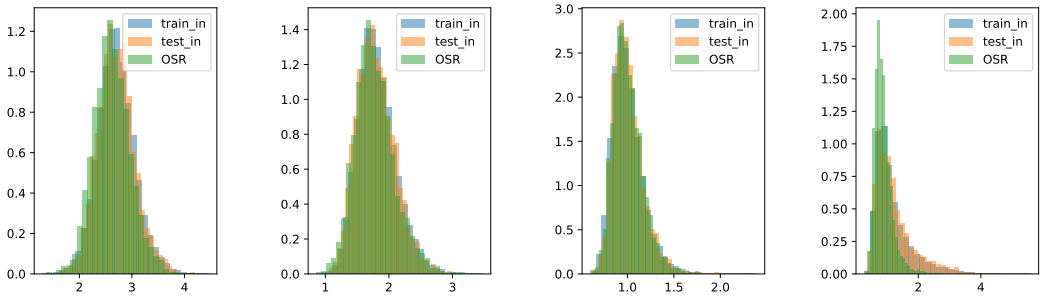


Figure 8: Histogram of activations for ResNet-18 pretrained on CIFAR10. This is evaluated on the training data (`train_in`), ID testing data (`test_in`), OOD data (`OOD` from SVHN) and OSR data (`OSR` from CIFAR100).

D INFLUENCE OF TRAINING CONFIGURATIONS FOR OOD PERFORMANCE

Table 9: Results on OOD and OSR benchmarks, using ResNet-18 with the ReAct config supervised by the CE loss.

Methods	OOD benchmarks							OSR benchmarks					
	SVHN	Textures	LSUN-C	LSUN-R	iSUN	Places365	AVG ID=92.27	CIFAR10 ID=97.13	CIFAR+10 ID=96.6	CIFAR+50 ID=96.8	TinyImageNet ID=83.4	AVG	Overall
CE+MLS	**85.42**	**82.20**	86.56	86.40	85.40	84.62	85.1	**92.54**	**95.62**	**91.81**	81.31	90.32	87.19
CE+ODIN	70.62	65.32	72.49	71.68	70.43	69.10	69.94	59.77	51.37	50.22	80.96	60.58	66.20
CE+Energy	**85.45**	**82.20**	86.62	86.45	85.44	84.65	85.14	**92.52**	**95.68**	**91.86**	81.28	90.34	87.22
CE+MLS+ReAct	83.94	79.67	**98.18**	**93.87**	**92.31**	**88.42**	**89.40**	**92.57**	94.92	90.88	**81.65**	90.01	**89.64**
CE+ODIN+ReAct	83.84	79.62	**98.31**	**94.00**	**92.44**	**88.48**	**89.45**	56.65	47.76	48.40	81.30	58.53	77.08
CE+Energy+ReAct	83.08	76.63	95.06	94.17	92.18	85.46	87.76	92.58	95.02	90.99	**81.67**	90.07	88.14

Table 10: Results on OOD and OSR benchmarks, using ResNet-18 with the ReAct config supervised by the ARPL+CS loss.

Methods	OOD benchmarks							OSR benchmarks					
	SVHN	Textures	LSUN-C	LSUN-R	iSUN	Places365	AVG ID=92.99	CIFAR10 ID=97.13	CIFAR+10 ID=97.75	CIFAR+50 ID=97.83	TinyImageNet ID=85.1	AVG	Overall
ARPL+CS+MLS	79.18	86.85	**95.59**	95.00	94.47	89.98	90.18	**93.23**	**97.93**	**96.27**	**82.5**	**92.48**	91.1
ARPL+CS+ODIN	50.14	38.35	74.25	42.96	43.46	53.99	50.53	92.57	97.37	95.23	53.20	84.59	64.15
ARPL+CS+Energy	79.08	86.82	**95.68**	95.07	94.54	90.03	90.20	**93.19**	**97.96**	**96.30**	80.45	91.98	90.91
ARPL+CS+MLS+ReAct	**84.19**	**89.11**	94.98	**95.52**	**94.98**	**90.23**	**91.50**	92.63	**97.96**	**96.30**	79.78	91.67	**91.57**
ARPL+CS+ODIN+ReAct	49.22	37.19	65.89	37.80	38.77	47.55	46.07	91.00	94.74	91.84	47.47	81.26	60.15
ARPL+CS+Energy+ReAct	**84.13**	**89.13**	95.10	**95.63**	**95.09**	**90.32**	**91.57**	92.59	**98.01**	**96.33**	79.90	91.71	**91.62**

Table 11: Results on OOD and OSR benchmarks, using ResNet-18 with the MLS config supervised by the ARPL+CS loss.

Methods	OOD benchmarks							OSR benchmarks					
	SVHN	Textures	LSUN-C	LSUN-R	iSUN	Places365	AVG Acc=91.02	CIFAR10 ID=96.96	CIFAR+10 ID=96.77	CIFAR+50 ID=96.69	TinyImageNet ID=86.91	AVG	Overall
ARPL+CS+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**
ARPL+CS+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
ARPL+CS+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**
ARPL+CS+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
ARPL+CS+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
ARPL+CS+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**

Table 12: Results on OOD and OSR benchmarks, using ResNet-18 with the official configuration supervised by the OE loss.

Method	OOD benchmarks							OSR benchmarks					
	SVHN	Textures	LSUN-C	LSUN-R	iSUN	Places365	AVG ID=94.77	CIFAR10 ID=97.59	CIFAR+10 ID=97.38	CIFAR+50 ID=97.34	TinyImageNet ID=77.96	AVG	Overall
OE+MLS	**95.94**	**94.57**	76.58	85.20	87.16	88.46	87.99	**96.26**	**98.95**	**98.20**	**77.88**	**92.82**	89.92
OE+ODIN	93.32	92.84	65.85	84.50	87.24	82.29	84.34	93.44	96.18	93.69	77.49	90.20	86.68
OE+Energy	**95.84**	**94.45**	75.50	84.55	86.64	88.19	87.54	**96.33**	**98.95**	98.04	77.73	**92.76**	89.62
OE+MLS+ReAct	95.46	94.32	93.57	90.66	90.75	88.97	**92.29**	**96.20**	**98.93**	**98.18**	77.60	**92.73**	**92.46**
OE+ODIN+ReAct	87.64	88.52	**89.56**	86.19	86.57	76.02	85.75	93.03	95.45	92.75	76.90	89.53	87.26
OE+Energy+ReAct	87.84	89.29	88.66	**90.14**	**93.69**	**94.67**	90.72	91.04	98.93	98.02	77.48	91.37	90.98

E EXPERIMENTS OF OOD ON CIFAR-100

Table 13: Results of various methods on OOD benchmarks, using CIFAR100 as ID. We train a ResNet-18 with the CE loss and report the in distribution accuracy as 'ID'.

Method	OOD benchmarks						
	SVHN	Textures	LSUN	LSUN-R	iSUN	Places365	AVG ID=78.69
CE+MSP	83.56	78.38	78.21	79.60	79.07	73.56	78.73
CE+MLS	85.21	79.33	77.75	**81.63**	**81.12**	73.48	79.75
CE+ODIN	**97.47**	77.05	**90.49**	77.59	79.38	71.96	**82.32**
CE+GODIN	52.70	58.40	59.64	76.34	76.59	62.09	64.29
CE+SEM	65.54	31.85	83.58	35.81	37.38	51.01	50.86
CE+Energy	85.71	79.50	77.12	**82.19**	**81.72**	73.26	79.92
CE+Gradnorm	65.73	62.82	61.13	64.41	64.60	61.08	63.30
CE+MLS+ReAct	84.51	**83.94**	84.96	80.30	79.98	**78.32**	**82.00**
CE+ODIN+ReAct	96.39	78.74	89.62	75.88	77.28	71.78	81.62
CE+Energy+ReAct	84.36	**83.48**	77.93	77.09	77.26	75.55	79.28

Table 14: Results of various methods on OOD benchmarks, using CIFAR100 as ID. We train a ResNet-18 with the ARPL+CS loss and report the in distribution accuracy as 'ID'.

Method	OOD benchmarks						
	SVHN	Textures	LSUN	LSUN-R	iSUN	Places365	AVG ID=78.86
ARPL+CS+MSP	79.95	79.00	80.28	88.16	87.71	74.69	81.63
ARPL+CS+MLS	81.27	79.73	79.93	89.63	89.23	74.77	82.43
ARPL+CS+ODIN	**87.72**	64.90	78.21	82.03	82.60	65.12	76.76
ARPL+CS+GODIN	73.81	62.55	59.30	66.42	75.00	51.41	64.75
ARPL+CS+SEM	55.94	33.81	**86.74**	32.23	34.34	57.19	50.04
ARPL+CS+Energy	81.69	79.84	78.91	**90.57**	**90.21**	74.58	**82.63**
ARPL+CS+Gradnorm	50.79	52.91	49.03	54.68	56.20	50.14	52.29
ARPL+CS+MLS+ReAct	77.71	**80.32**	85.93	88.44	87.78	**76.41**	**82.77**
ARPL+CS+ODIN+ReAct	20.02	33.07	20.25	36.67	36.74	37.82	30.76
ARPL+CS+Energy+ReAct	62.27	63.78	56.36	80.73	81.33	60.05	67.42

Table 15: Results of various methods on OOD benchmarks, using CIFAR100 as ID. We train a ResNet-18 with the OE loss and report the in distribution accuracy as 'ID'.

Method	OOD benchmarks						
	SVHN	Textures	LSUN	LSUN-R	iSUN	Places365	AVG ID=77.16
OE+MSP	93.88	87.76	73.66	73.10	74.72	74.49	79.60
OE+MLS	94.32	88.44	72.53	73.22	75.21	74.10	79.64
OE+ODIN	**97.20**	83.94	**85.96**	71.61	74.60	71.90	80.87
OE+GODIN	74.18	83.35	67.85	74.71	77.22	66.61	73.99
OE+SEM	68.48	47.58	80.55	48.33	46.99	49.15	56.85
OE+Energy	94.26	88.47	71.19	72.56	74.67	73.70	79.14
OE+Gradnorm	86.54	79.75	53.73	55.55	57.59	63.90	66.18
OE+MLS+ReAct	94.50	**89.04**	77.72	80.31	80.88	**76.73**	**83.20**
OE+ODIN+ReAct	96.76	82.66	**85.65**	**76.91**	**78.48**	70.50	81.83
OE+Energy+ReAct	94.07	**89.32**	68.41	75.63	77.25	73.94	79.77

F QUALITATIVE SAMPLES FROM THE ID AND OOD DATASETS

Figure 9: Examples on ID data (CIFAR10).

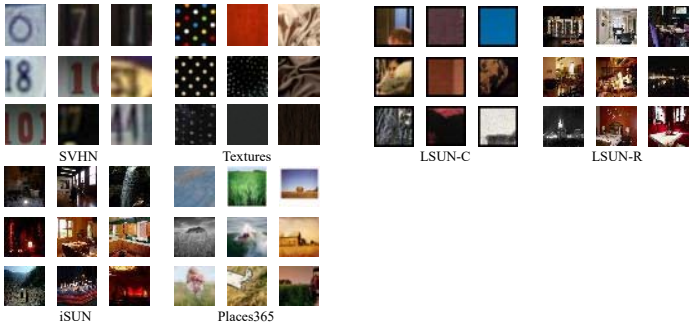


Figure 10: Typical success cases on different OOD datasets.

Figure 11: Typical failure cases on different OOD datasets.

G CORRESPONDENCE POINTS FOR IMAGENET-SSB

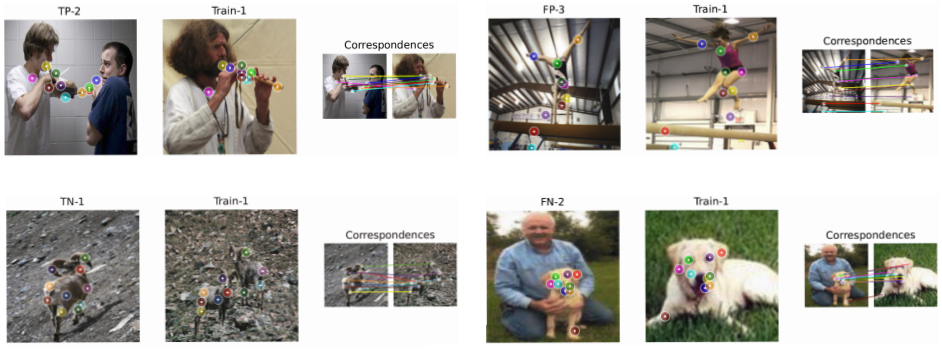


Figure 12: Correspondence Points for ImageNet-SSB using DINO-ViT.