Out-Of-Distribution Detection From a Model-Specific Perspective

Anonymous Authors[1]

Abstract

Out-of-distribution (OOD) detection aims to identify test examples that do not belong to the training distribution and are thus unlikely to be classified reliably. Despite a plethora of existing works, most of them focused only on the scenario where OOD examples come from semantic shift (e.g., unseen classes), ignoring other possible causes (e.g., covariate shift). In this paper, we present a novel framework to study OOD detection in a broader scope. Instead of detecting OOD examples from a particular cause, we propose to detect examples that *a deployed machine learning model is unable to classify correctly*. That is, whether a test example should be detected or not depends on the deployed classifier. We show that this framework enables us to study OOD examples with semantic shift and covariate shift in a unified way, and more closely addresses the concern of applying a machine learning model to unconstrained environments. We provide an extensive analysis that involves a variety of classifiers (e.g., different model architectures and training strategies), sources of OOD examples, and detection approaches and reveal several insights for improving OOD detection in uncontrolled environments.

1. Introduction

The staggering success of deep learning gives rise to the prospect of intelligent systems entering our everyday lives. However, there is a huge difference between achieving impressive results in a laboratory under best-case conditions and applying the technology reliably to uncontrolled settings. Take image classification for instance. While neural network models could perform fairly well on "in-distribution (ID)" data that belong to the training distribution, their reliability often degrades drastically when facing data with covariate shift (e.g., different image domains or styles) (Wilson & Cook, 2020) or semantic shift (e.g., novel classes) (Liang et al., 2017). What is even worse is that neural networks tend to be overconfident in their predictions (Guo et al., 2017), failing to identify these potential error cases. Out-of-distribution (OOD) detection (Yang et al., 2021b; Salehi et al., 2021) thus emerges as a critical paradigm to tackle this problem — "rejecting" examples that the models cannot perform well on.

By definition, "out-of-distribution (OOD)" refers to test examples drawn from a distribution that is different from the training distribution, including both semantic shift and covariate shift. Yet, in the literature on OOD detection (Yang et al., 2021b), the focus is mostly on semantic shift. Covariate shift, in contrast, is more commonly studied in model generalization and robustness (Wiles et al., 2022; Shen et al., 2021). That is, instead of rejecting test examples with covariate shift, the community focuses more on improving the robustness of a neural network model so that the model could classify them correctly. However, given the notable accuracy gap between classifying ID examples and examples with covariate shift (Taori et al., 2020), we argue that it is desirable to also consider covariate shift in OOD detection. This is particularly the case when the neural network model is deployed in any unconstrained environment where different kinds of OOD examples may appear.

How should we bring covariate shift into the study of OOD detection? Such a seemingly naive question surprisingly leads to the key insight of this paper. Unlike examples with semantic shift, which one would like to detect as many as possible given that the classifier trained with ID data can never classify them correctly, we argue that *whether an example with covariate shift should be detected or not is "ill-posed" without taking into account the kind of covariate shift and deployed classifier.* For example, if the covariate shift is not severe (e.g., ImageNet (Russakovsky et al., 2015) as ID; ImageNetV2 as covariate shift (Recht et al., 2019)) and the classifier is considered robust (e.g., with the CLIP-pre-trained backbone (Radford et al., 2021b)), many of the examples with covariate shift will likely be correctly classified by the classifier. In this case, one perhaps should not reject them. In contrast, if the covariate shift is severe (e.g., ImageNet-A as covariate shift (Hendrycks et al., 2021b)) and the classifier is not robust (e.g., a neural network trained

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

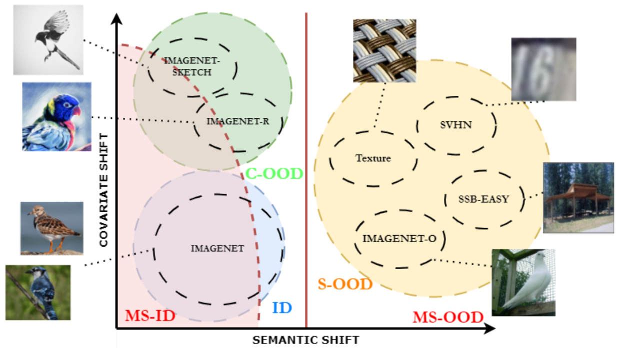

Figure 1. **MS-OOD Framework** using ImageNet as an example. Blue, green, and yellow regions denote in-distribution (ID), covariate shift (C-OOD), and semantic shift (S-OOD) data, respectively, each with their datasets and representative images. Given a classifier, the shaded red region denotes the *correctly classified images* (called the *classified region*). A more robust classifier would shift the red dashed boundary towards the solid red line, whose classified region covers in-distribution and covariate shift data. Using this framework, we can separate data into model-specific ID/ODD cases (MS-ID/MS-OOD), corresponding to the red and white regions. The goal of MS-OOD is thus to detect the MS-OOD examples that are misclassified by the classifier.

from scratch), many of the examples with covariate shift are likely to be misclassified and should be rejected.

Building upon this insight, we propose a novel framework, MS-OOD, to study ODD detection from a "model-specific" perspective. In MS-OOD, *whether a test example should be detected as OOD and rejected from being classified (denote by a ground-truth label* $+1$*) depends on whether the deployed classifier misclassifies it.* With this definition, every test example can be *deterministically* assigned a ground-truth label for OOD detection based on the deployed classifier: -1 for correctly classified examples, which should not be detected; $+1$ for misclassified examples, which should be detected. This enables us to study different causes of OOD examples in a unified way. It is worth noting that while test examples with covariate shift could be assigned different ground-truth labels, all the test examples with semantic shift are assigned ground-truth labels $+1$. In other words, similar to conventional OOD detection, all the test examples with semantic shift should be detected in MS-OOD.

However, unlike conventional OOD detection which treats all the test examples associated with the ID training data[1] as ID, MS-OOD *treats those misclassified ones as OOD and aims to detect them as well.* We argue that this definition does not deviate from the goal of OOD detection. Instead, this definition could better reflect real-world application scenarios. In essence, in most machine learning tasks, we are never provided with the training distribution but the training data drawn from it. For end-users who seek to reliably apply the machine learning model at hand, they may not even be able to access the training data and learning algorithm. MS-OOD therefore can be interpreted as using the trained classifier to model the training distribution: misclassified test (e.g., hard) examples are viewed as OOD.

We conduct an extensive empirical study and analysis under our MS-OOD framework. We consider three dimensions: **1) sources of OOD examples**, which include both semantic and covariate shift; **2) deployed classifiers**, which include different neural network architectures and training strategies; **3) OOD detection methods**, which include representative and state-of-the-art approaches such as Maximum Softmax Probabilities (MSP) (Hendrycks & Gimpel, 2016), Energy Score (Liu et al., 2020), Maximum Logit Score (MLS) (Vaze et al., 2021), Virtual-logit Matching (ViM) (Wang et al., 2022), and GradNorm (Huang et al., 2021). New classifiers, OOD methods, and datasets can easily be incorporated to extend the scope. This experimental framework not only offers a platform to unify the community but also provides a manual to end-users for selecting the appropriate OOD methods in their respective use cases.

Along with this study is a list of novel insights and take-home messages. For instance, we find that the best OOD detection methods under the MS-OOD framework are not consistent across different OOD cases, but somehow consistent across classifiers. For detecting misclassified ID and covariate shift data, MSP performs fairly well in general. For detecting semantic shift, we see notable but consistent differences between the MS-OOD setting and the conventional OOD setting. Specifically, we find that many ID examples that are wrongly detected as OOD in the conventional setting cannot be correctly classified by the classifier. In other words, they actually should be rejected from being classified. More findings can be found in section 5.

Contributions. Our contributions are two-folded:

- We propose a novel framework, MS-OOD, which enables us to study different sources of OOD examples (e.g., covariate shift and semantic shift) in a unified way.
- We conduct an extensive empirical study and analysis under the MS-OOD framework.

2. Related Works

Out-of-distribution (OOD) detection settings. OOD detection is highly related to anomaly detection, novelty detection, open-set recognition, and outlier detection (Yang et al., 2021b). The difference lies in 1) the scope of OOD examples; 2) whether one has to classify ID examples.

In conventional OOD detection, the focus is on detecting

[1]For example, if ID training data is ImageNet (Russakovsky et al., 2015), then the validation or test set of ImageNet will be treated as ID data in conventional OOD detection: these examples are supposed drawn from the training distribution.

OOD examples with semantic shift, ignoring the existence of covariate shift (Yang et al., 2021b; Salehi et al., 2021). Very few works include examples with covariate shift into their studies (Yang et al., 2021a; Ming et al., 2021; Yang et al., 2022; Hsu et al., 2020), but most of them *treat these examples as ID*, aiming to classify them robustly instead of detecting them as OOD.

In anomaly detection and outlier detection, where the focus is to differentiate ID and OOD examples without the need to correctly classify ID examples (i.e., they treat all ID examples as a single class), several works also consider examples with covariate shift (Yang et al., 2021b). In contrast to the above, these works aim to detect covariate shift as OOD.

In our paper, we argue that whether an example with covariate shift should be detected as OOD or ID depends on whether the deployed classifier misclassifies it or not. By taking a model-specific perspective, our MS-OOD framework resolves the dilemma between *OOD detection* (Yang et al., 2021b) and *OOD generalization* (Shen et al., 2021): a robust model should *generalizes* on covariate shift data, while a weak model should *detect* them.

Selective Classification. Equipping a classifier with the option to reject has been studied in another sub-field named selective classification (Geifman & El-Yaniv, 2017). Different from OOD detection, selective classification focuses on rejecting hard or uncertain ID examples. Recently, Xia & Bouganis (2022) proposed to integrate selective classification with OOD detection, aiming to detect both semantic shift and misclassified ID data. In this context, our work can be seen as a generalized version, further taking covariate shift into account. Compared to (Xia & Bouganis, 2022), we provide a more comprehensive study, further emphasizing the role of classifiers in evaluation.

OOD detection methods can roughly be categorized into post-hoc and training-based approaches (Salehi et al., 2021). The difference lies in if one could specifically train a model to detect OOD examples. While training-based approaches like outlier exposure (Hendrycks et al., 2018) have shown a much higher detection rate, they may be prohibitive for end-users who cannot access the original training data. In this paper, we thus focus on post-hoc approaches. The baseline is to use the softmax output as confidence (Hendrycks & Gimpel, 2016). Other approaches consider scaling the temperature and adding input perturbations (Liang et al., 2017); using logits (Vaze et al., 2021), energy (Liu et al., 2020), or gradients (Huang et al., 2021) as the score; combining intermediate features with logits (Wang et al., 2022).

3. Background

We consider the problem of out-of-distribution (OOD) detection in the context of classification. Given a neural network f classifier that is trained on data sampled from a training distribution $P(X, Y)$, the objective is to construct a selection function g such that:

$$g(x; f) = \begin{cases} 1, & g(x; f) > \tau; \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

During test time, inputs that produce $g(x; f) = 1$ are forwarded for classification, while the rest are either rejected or redirected for further investigation. Ideally, one would like to reject a test example x if $f(x) \neq y$, where y is the ground-truth label.

In real-world scenarios, sources of OOD can come in various ways. We categorize them into two major groups: covariate shift and semantic shift. Using $P(X, Y)$ as in-distribution, these shifts occur either on marginal distribution $P(X)$, or both $P(X)$ $P(Y)$, respectively. In a practical setting, semantic shift data contains classes outside the neural network semantic space, while covariant shift includes data from different domains but within the same label space. The focal point of existing OOD detection research has been on semantic shift.

4. Framework

The issue with existing OOD detection studies can be described with more clarity in Figure 1. Let's define a *classified region* (denoted by the red shaded area in the picture) as the input space where a neural network makes correct predictions, and the *potential region* (the boundary is denoted by the solid red line) as the maximum area a model can potentially classify correctly given its semantic space. For instance, the *classified region* boundary of a robust model would be close to the *potential region* border, while a weak model would probably only cover around the in-distribution area. The goal of model generalization and robustness is then to fit the *classified region* into the *potential region*, while the goal of detection is to identify samples outside the *classified region*. If we have a highly robust model, OOD detection would simply become semantic shift detection, which aligns with the existing OOD detection framework. However, given the current progress of OOD generalization on standard models (Taori et al., 2020), we advocate to consider the covariate shift cases as a model-dependent problem, i.e., to detect those that cannot be classified correctly.

Before we go to our proposed framework, we would first redefine OOD as S-OOD and C-OOD for semantic shift and covariate shift OOD respectively. Note that our definition of S-OOD *is the same* as the definition of OOD in existing studies that only focus on semantic shift. We use this new term to further distinguish between the two OOD causes.

Based upon our previous reasoning, we propose a unified view of out-of-distribution, termed as MS-ID for *model-*

Table 1. **C-OOD Detection under existing framework** across datasets and models

MODEL	METHODS	C-OOD								
		IN-V2		IN-Sketch		IN-R		IN-A		AVG
		ACC↑	FPR95↓	ACC↑	FPR95↓	ACC↑	FPR95↓	ACC↑	FPR95↓	
DEPTH										
ResNet18	MSP	66.52	94.40	20.23	68.87	33.06	84.54	1.15	87.36	83.79
	MaxLogit		**94.13**		**59.68**		**37.41**		**44.13**	**58.84**
ResNet50	MSP	72.37	**93.93**	24.09	65.70	36.17	71.56	0.00	81.48	78.17
	MaxLogit		94.01		**53.34**		**30.21**		**42.59**	**55.04**
ResNet152	MSP	75.10	93.67	28.53	66.55	41.34	70.40	6.03	75.33	76.49
	MaxLogit		**93.59**		**55.59**		**32.30**		**34.47**	**53.99**
TRAINING										
ResNet50	MSP	72.37	**93.93**	24.09	65.70	36.17	71.56	0.00	81.48	78.17
	MaxLogit		94.01		**53.34**		**30.21**		**42.59**	**55.04**
Robust ResNet50	MSP	77.71	**93.42**	29.86	**68.79**	42.82	63.66	14.55	74.89	75.19
	MaxLogit		93.57		71.67		**32.82**		**28.64**	**56.68**
PRETRAINING										
ResNet50	MSP	72.37	**93.93**	24.09	65.70	36.17	71.56	0.00	81.48	78.17
	MaxLogit		94.01		**53.34**		**30.21**		**42.59**	**55.04**
CLIP-ResNet50	MSP	59.53	95.36	35.50	78.39	60.60	92.28	22.76	89.55	88.90
	MaxLogit		94.50		97.96		80.87		75.17	**87.13**
ARCHITECTURE										
ResNet50	MSP	72.37	**93.93**	24.09	65.70	36.17	71.56	0.00	81.48	78.17
	MaxLogit		94.01		**53.34**		**30.21**		**42.59**	**55.04**
ViT-B-16	MSP	77.38	**94.19**	29.40	60.11	44.00	54.24	20.84	60.35	67.22
	MaxLogit		94.20		**56.29**		**32.54**		**24.85**	**51.97**

specific in-distribution and MS-OOD for *model-specific out-of-distribution*. We define a classifier as $f : \mathcal{X} \mapsto \mathcal{Y}$, where $\mathcal{Y}$ is the model's semantic space. Given a (test) dataset $D = \{(x_i \in \mathcal{X}, y_i \in \mathcal{Y})\}_{i=1}^{N}$, where x_i and y_i denote the input (e.g. image) and ground truth label, we formally define MS-ID as data points on which the classifier outputs the correct label $f_\theta(x_i) = y_i$, and MS-OOD when the classifier outputs the wrong label $f_\theta(x_i) \neq y_i$. Since the ground truth label for semantic shift data is outside the model semantic space $\mathcal{Y}$, they will *always be labeled as* MS-OOD.

There are several key properties in this framework:

1. The *potential region* of a classifier is defined by the model's semantic space. Notice that this does not correlate to the number of labels. For instance, a dog classifier might have the same semantic space as a classifier that specifies the many different breeds of dogs.

2. We define the three possible sources for MS-OOD as either in-distribution, covariate shift, and semantic shift. Using MS-OOD@S notation, where S denotes the source, we have MS-OOD@ID, MS-OOD@C-OOD, and S-OOD. To further clarify the meaning of these terms: MS-OOD@ID corresponds to *misclassified in-distribution data* and MS-OOD@C-OOD denotes *misclassified covariate shift data*. Similarly, the sources of MS-ID are MS-ID@ID and MS@C-OOD, denoted as *correctly classified in-distribution data* and *correctly classified covariate shift data*.

3. Existing metrics in OOD detection, while can be used, would encompass a novel meaning under this framework. For instance, False Positive Rate 95 (FPR95) denotes the number of OOD samples detected as ID when 95% of ID data passes. In our framework, the positive now becomes the *correctly classified ID* data, providing better control for the trade-off between the model's accuracy and detection. We denote this new metric as FPRN@S+, where N shows the percentage of potentially correctly classified images forwarded to the model and S the source for the true positive (either ID or C-OOD). Note that when setting the threshold τ, users can only access their existing dataset, making FPRN@ID+ useful for practical cases and FPRN@C-OOD+ useful for experiments. The '+' sign means *correctly classified*.

5. Experiments

5.1. Dataset

We choose our datasets based on distribution shifts.

In-distribution. We decide to use ImageNet (Deng et al., 2009), the standard benchmark for image classification which contains images that closely mimic real-world cases.

Covariate-shift. ImageNetV2 (Recht et al., 2019) provides a practical case of a testing dataset on distribution that mimics (albeit different) training data. For images with different styles and domains, we use ImageNet-R (Hendrycks et al., 2021a) and ImageNet-S (Wang et al., 2019). We also test on ImageNet-A (Hendrycks et al., 2021b), a natural adversarial dataset curated by collecting wrongly predicted examples with high confidence. Since ImageNet-R and ImageNet-A use only a subset of 200 classes from ImageNet, we follow the same setting, using only the same subset classes for in-distribution for a fair comparison.

Semantic-shift. We use the common benchmark datasets for OOD detection: SVHN (Netzer et al., 2011), Texture (Cimpoi et al., 2014), Places365 (Zhou et al., 2014), iNaturalist (Van Horn et al., 2018) and SUN (Xiao et al., 2010).

Table 2. **C-OOD Detection under MS-OOD Framework** across datasets models

MODEL	METHODS	MS-OOD@ID		MS-OOD@C-OOD								
		IN		IN-V2		IN-Sketch		IN-R		IN-A		AVG
		ACC↑	FPR95@ID+↓	ACC↑	F95↑	ACC↑	F95↑	ACC↑	F95↑	ACC↑	F95↑	
DEPTH												
ResNet18	MSP	69.76	**63.38**	66.52	**0.835**	20.23	0.515	33.06	0.623	1.15	0.017	0.497
	MaxLogit		71.39		0.817		**0.525**		**0.646**		0.017	**0.502**
	Energy		75.07		0.811		0.506		0.606		0.014	0.484
	ViM		75.07		0.811		0.508		0.606		0.014	0.485
	GradNorm		87.48		0.790		0.429		0.540		**0.024**	0.446
ResNet50	MSP	76.13	**61.66**	72.37	**0.866**	24.09	0.557	36.17	**0.698**	0.00	X	**0.707**
	MaxLogit		70.67		0.853		**0.584**		0.638			0.692
	Energy		74.86		0.849		0.570		0.604			0.674
	ViM		78.72		0.838		0.578		0.614			0.677
	GradNorm		89.60		0.825		0.485		0.575			0.628
ResNet152	MSP	78.31	**60.92**	75.10	**0.879**	28.53	0.601	41.34	**0.737**	6.03	0.129	**0.587**
	MaxLogit		70.56		0.864		**0.617**		0.652		**0.130**	0.566
	Energy		73.45		0.859		0.609		0.626		0.130	0.556
	ViM		77.78		0.859		0.615		0.653		0.098	0.556
	GradNorm		91.67		0.840		0.512		0.608		0.122	0.520
TRAINING												
ResNet50	MSP	76.13	**61.66**	72.37	**0.866**	24.09	0.557	36.17	**0.698**	0.00	X	**0.707**
	MaxLogit		70.67		0.853		**0.584**		0.638			0.692
	Energy		74.86		0.849		0.570		0.604			0.674
	ViM		78.72		0.838		0.578		0.614			0.677
	GradNorm		89.60		0.825		0.485		0.575			0.628
Robust ResNet50	MSP	80.86	**66.52**	77.71	**0.884**	29.86	**0.594**	42.82	**0.754**	14.55	**0.282**	**0.628**
	MaxLogit		72.24		0.878		0.562		0.686		0.235	0.590
	Energy		94.41		0.856		0.452		0.207		0.072	0.397
	ViM		85.10		0.865		0.593		0.633		0.206	0.574
	GradNorm		99.15		0.850		0.458		0.598		0.254	0.540
PRETRAINING												
ResNet50	MSP	76.13	**61.66**	72.37	**0.866**	24.09	0.557	36.17	**0.698**	0.00	X	**0.707**
	MaxLogit		70.67		0.853		**0.584**		0.638			0.692
	Energy		74.86		0.849		0.570		0.604			0.674
	ViM		78.72		0.838		0.578		0.614			0.677
	GradNorm		89.60		0.825		0.485		0.575			0.628
CLIP-ResNet50	MSP	59.82	**72.11**	59.53	**0.775**	35.50	**0.631**	60.60	**0.803**	22.76	**0.413**	**0.656**
	MaxLogit		86.62		0.745		0.531		0.760		0.364	0.600
	Energy		90.67		0.739		0.523		0.736		0.349	0.587
	ViM		90.74		0.738		0.523		0.733		0.348	0.586
	GradNorm		93.94		0.734		0.515		0.749		0.375	0.593
ARCHITECTURE												
ResNet50	MSP	76.13	**61.66**	72.37	**0.866**	24.09	0.557	36.17	**0.698**	0.00	X	**0.707**
	MaxLogit		70.67		0.853		**0.584**		0.638			0.692
	Energy		74.86		0.849		0.570		0.604			0.674
	ViM		78.72		0.838		0.578		0.614			0.677
	GradNorm		89.60		0.825		0.485		0.575			0.628
ViT-B-16	MSP	81.07	**59.43**	77.38	**0.889**	29.40	0.637	44.00	**0.789**	20.84	**0.388**	**0.676**
	MaxLogit		63.89		0.883		**0.640**		0.698		0.298	0.630
	Energy		75.81		0.870		0.623		0.611		0.255	0.590
	ViM		76.26		0.870		0.601		0.596		0.143	0.553
	GradNorm		98.08		0.846		0.469		0.613		0.347	0.569

To ensure these datasets' classes don't intersect with in-distribution data, we use the filtered datasets in (Huang & Li, 2021). We also use the 'Easy" version of ImageNet-21K-P (Ridnik et al., 2021) from Semantic Shift Benchmark (Vaze et al., 2021), following the common setting of Open Set Recognition which primarily focus on *semantically* different dataset.

5.2. Models

We choose our models based on robustness techniques and its *classified region* (e.g. accuracy) using ResNet50 (He et al., 2016) as our basic block: ResNet152 for depth; Robust ResNet50 (Paszke et al., 2019) which use robust intervention training (data augmentations, label smoothing, longer training, etc.); CLIP-ResNet50 (Radford et al., 2021a) for pretraining; and ViT-B-16(Dosovitskiy et al., 2020) for architecture. We use the zero-shot capability of CLIP using 80 prompts and modified ImageNet class names from the official GitHub. We note that despite CLIP-ResNet50 employing a different strategy, the robustness comes from its diverse training data (Fang et al., 2022). Hence, we put the model under the umbrella of pretraining. We use the official PyTorch (Paszke et al., 2019) pretrained models for all models except CLIP-ResNet50 which is based on the original paper (Radford et al., 2021a). The robust ResNet50 refers to pretrained ResNet50 model trained using TorchVision new training recipe.

5.3. Algorithms

We focus on post-hoc methods under the assumption of a fixed classifier. We pick five representative algorithms categorized into output-based, feature-based and mixed.

Table 3. Difference between the previous and our MS-OOD framework. Please see a full version in Table 5 with additional datasets.

MODEL	METHODS	ID	S-OOD									
		IN	SVHN		DTD		SUN		SSB-IN-Easy		AVG	
		ACC↑	FPR95↓	FPR95 @ID+↓	FPR95↓	FPR95 @ID+↓	FPR95↓	FPR95 @ID+↓	FPR95↓	FPR95 @ID+↓	FPR95↓	FPR95 @ID+↓
DEPTH												
ResNet18	MSP	69.758	10.67	2.52	70.16	46.33	73.45	47.99	79.00	**54.52**	58.32	37.84
	MaxLogit		5.80	1.10	57.18	36.12	62.92	39.54	75.36	54.98	50.31	32.93
	Energy		6.53	1.53	52.82	36.06	59.75	39.83	74.97	57.87	48.52	33.82
	ViM		1.61	0.59	38.83	26.97	93.08	87.19	80.09	68.88	53.40	45.91
	GradNorm		**0.29**	**0.12**	**29.31**	**23.35**	**34.66**	**28.56**	**71.69**	65.14	**33.99**	**29.29**
ResNet50	MSP	76.13	12.89	4.21	66.01	45.11	68.58	45.41	72.57	50.81	55.01	36.39
	MaxLogit		7.46	2.33	54.36	36.28	59.90	38.76	**69.02**	**50.00**	47.68	**31.84**
	Energy		8.18	2.94	52.13	37.77	58.28	41.19	69.19	53.38	46.94	33.82
	ViM		**0.79**	**0.16**	**15.74**	**9.10**	82.06	69.65	76.16	63.29	43.69	35.55
	GradNorm		1.14	0.86	32.39	29.31	**37.25**	**33.78**	69.82	66.34	**35.15**	32.57
ResNet152	MSP	78.312	25.55	10.45	59.84	42.13	66.01	45.89	71.31	51.74	55.68	37.55
	MaxLogit		18.00	5.44	45.59	29.63	51.87	**34.06**	**66.05**	**48.39**	45.38	**29.38**
	Energy		21.51	7.24	43.83	29.52	50.25	34.73	66.37	50.58	45.49	30.52
	ViM		**0.27**	**0.08**	**12.98**	**7.71**	77.76	63.40	73.54	60.18	41.14	32.84
	GradNorm		3.49	2.94	31.91	30.32	**43.46**	41.73	73.99	72.26	**38.21**	36.81
TRAINING												
ResNet50	MSP	76.13	12.89	4.21	66.01	45.11	68.58	45.41	72.57	50.81	55.01	36.39
	MaxLogit		7.46	2.33	54.36	36.28	59.90	38.76	**69.02**	**50.00**	47.68	**31.84**
	Energy		8.18	2.94	52.13	37.77	58.28	41.19	69.19	53.38	46.94	33.82
	ViM		**0.79**	**0.16**	**15.74**	**9.10**	82.06	69.65	76.16	63.29	43.69	35.55
	GradNorm		1.14	0.86	32.39	29.31	**37.25**	**33.78**	69.82	66.34	**35.15**	32.57
Robust ResNet50	MSP	80.856	38.41	20.77	71.91	56.22	**70.86**	**53.12**	**72.80**	**56.58**	63.50	46.68
	MaxLogit		54.31	37.01	75.43	64.26	75.04	62.21	75.75	63.82	70.13	56.82
	Energy		99.98	99.97	95.74	95.64	97.72	97.62	95.59	95.46	97.26	97.17
	ViM		**0.07**	**0.03**	**21.91**	**18.72**	73.30	67.25	73.87	67.45	**42.29**	**38.36**
	GradNorm		100.00	100.00	97.45	97.77	99.76	99.81	99.49	99.58	99.17	99.29
CLIP												
ResNet50	MSP	76.13	12.89	4.21	66.01	45.11	68.58	45.41	72.57	50.81	55.01	36.39
	MaxLogit		7.46	2.33	54.36	36.28	59.90	38.76	**69.02**	**50.00**	47.68	**31.84**
	Energy		8.18	2.94	52.13	37.77	58.28	41.19	69.19	53.38	46.94	33.82
	ViM		**0.79**	**0.16**	**15.74**	**9.10**	82.06	69.65	76.16	63.29	43.69	35.55
	GradNorm		1.14	0.86	32.39	29.31	**37.25**	**33.78**	69.82	66.34	**35.15**	32.57
CLIP-ResNet50	MSP	59.818	10.29	4.24	66.76	**46.65**	70.83	**48.72**	80.08	**60.29**	56.99	**39.98**
	MaxLogit		61.22	29.34	89.52	82.55	69.12	53.47	**77.91**	68.30	74.44	58.42
	Energy		97.11	94.22	94.36	92.34	79.59	72.20	81.20	76.74	88.07	83.87
	ViM		96.72	93.43	94.68	92.71	81.75	74.12	81.70	77.03	88.71	84.32
	GradNorm		**0.01**	**0.01**	**48.30**	46.70	**60.59**	58.38	84.04	83.02	**48.23**	47.03
ARCHITECTURE												
ResNet50	MSP	76.13	12.89	4.21	66.01	45.11	68.58	45.41	72.57	50.81	55.01	36.39
	MaxLogit		7.46	2.33	54.36	36.28	59.90	38.76	**69.02**	**50.00**	47.68	**31.84**
	Energy		8.18	2.94	52.13	37.77	58.28	41.19	69.19	53.38	46.94	33.82
	ViM		**0.79**	**0.16**	**15.74**	**9.10**	82.06	69.65	76.16	63.29	43.69	35.55
	GradNorm		1.14	0.86	32.39	29.31	**37.25**	**33.78**	69.82	66.34	**35.15**	32.57
ViT-B-16	MSP	81.068	21.54	10.09	58.30	41.28	66.56	**48.08**	71.31	**53.54**	54.43	38.25
	MaxLogit		18.52	8.23	54.84	38.09	66.89	49.91	**70.17**	54.13	52.60	37.59
	Energy		24.15	13.50	57.39	46.01	72.77	62.80	73.77	63.99	57.02	46.58
	ViM		**2.34**	**0.85**	**43.94**	**31.49**	**59.46**	49.05	71.58	58.70	**44.33**	**35.02**
	GradNorm		96.13	96.43	92.87	93.14	96.85	97.05	97.86	98.04	95.93	96.16

Output-based. Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2016) relies on the output of softmax as confidence and serves as the baseline in most OOD detection literature. Despite its seemingly simple approach, we provide reasons why this algorithm is worth exploring in this framework: (Vaze et al., 2021; Fort et al., 2021) shows the superior performance when the method is deployed in a strong classifier; (Xia & Bouganis, 2022) shows the best performance compared to other states of the art models on *rejecting misclassified samples*; and (Ming et al., 2022) shows relatively good performance on the visual-language model (e.g. CLIP). We also consider MaxLogit (Vaze et al., 2021) and Energy (Liu et al., 2020).

Feature-based. We employ GradNorm (Huang et al., 2021) that relies on gradients and the penultimate layer.

Mixed. ViM (Wang et al., 2022) residual features and logits to produce the confidence scores. It is also the only algorithm in our experiments that use training data.

We provide several reasonings why we don't put other post-hoc methods in our experiments: ReAct (Sun et al., 2021) uses an extra hyperparameter that can impact the accuracy of the model, changing the area of the *classified region* and creating an unfair comparison. ODIN (Liang et al., 2017) requires hyperparameter tuning for input perturbations, hence the knowledge of OOD data. Although Mahalanobis (Lee et al., 2018) can be promising, we use ViM(Wang et al., 2022) as the better representative given its performance and similarity of using training data and intermediate features.

Table 4. C-OOD detection under MS-OOD Framework only on C-OOD dataset

MODEL	METHODS	MS-ID@COOD vs MS-OOD@COOD								
		IN-V2		IN-Sketch		IN-R		IN-A		AVG
		ACC↑	FPR95 @COOD+↓	ACC↑	FPR95 @COOD+↓	ACC↑	FPR95 @COOD+↓	ACC↑	FPR95 @COOD+↓	
DEPTH										
ResNet18	MSP	66.52	**64.85**	20.23	**66.86**	33.06	69.20	1.15	95.04	**73.99**
	MaxLogit		74.61		68.59		**68.97**		95.44	76.90
	Energy		78.32		72.78		71.65		96.98	79.93
	ViM		83.75		87.15		81.97		93.35	86.56
	GradNorm		88.17		76.41		83.62		**92.03**	85.06
ResNet50	MSP	72.37	**64.42**	24.09	69.66	36.17	63.36	0.00	X	**65.81**
	MaxLogit		72.74		**69.47**		**62.85**			68.35
	Energy		76.37		72.97		65.57			71.64
	ViM		81.83		78.94		76.74			79.17
	GradNorm		90.30		74.10		82.56			82.32
ResNet152	MSP	75.10	**62.93**	28.53	**66.45**	41.34	**60.35**	6.03	**87.42**	**69.29**
	MaxLogit		73.37		67.12		60.69		88.76	72.49
	Energy		77.27		70.71		63.12		88.69	74.95
	ViM		79.04		75.61		72.11		94.65	80.35
	GradNorm		92.09		76.05		84.43		90.57	85.79
TRAINING										
ResNet50	MSP	72.37	**64.42**	24.09	69.66	36.17	63.36	0.00	X	**65.81**
	MaxLogit		72.74		**69.47**		**62.85**			68.35
	Energy		76.37		72.97		65.57			71.64
	ViM		81.83		78.94		76.74			79.17
	GradNorm		90.30		74.10		82.56			82.32
Robust ResNet50	MSP	77.71	**69.00**	29.86	**71.23**	42.82	**61.37**	14.55	**88.75**	**72.59**
	MaxLogit		72.77		75.30		61.42		89.89	74.85
	Energy		94.03		95.84		66.21		91.76	86.96
	ViM		85.42		75.19		74.19		91.42	81.56
	GradNorm		99.19		98.87		99.64		96.52	98.56
PRETRAINING										
ResNet50	MSP	72.37	**64.42**	24.09	69.66	36.17	63.36	0.00	X	**65.81**
	MaxLogit		72.74		**69.47**		**62.85**			68.35
	Energy		76.37		72.97		65.57			71.64
	ViM		81.83		78.94		76.74			79.17
	GradNorm		90.30		74.10		82.56			82.32
CLIP-ResNet50	MSP	59.53	**73.61**	35.50	**72.81**	60.60	**65.92**	22.76	**84.64**	**74.25**
	MaxLogit		87.82		87.22		81.98		92.94	87.49
	Energy		91.35		93.58		88.05		94.34	91.83
	ViM		91.52		93.67		88.29		94.67	92.04
	GradNorm		93.65		91.77		88.82		92.91	91.79
ARCHITECTURE										
ResNet50	MSP	72.37	**64.42**	24.09	69.66	36.17	63.36	0.00	X	**65.81**
	MaxLogit		72.74		**69.47**		**62.85**			68.35
	Energy		76.37		72.97		65.57			71.64
	ViM		81.83		78.94		76.74			79.17
	GradNorm		90.30		74.10		82.56			82.32
ViT-B-16	MSP	77.38	**63.97**	29.40	63.00	44.00	**57.39**	20.84	**87.35**	**67.93**
	MaxLogit		67.73		**61.90**		57.81		88.46	68.98
	Energy		77.37		63.84		59.18		90.13	72.63
	ViM		76.88		63.73		58.80		90.23	72.41
	GradNorm		98.59		83.71		91.45		92.44	91.55

6. Discussion and Conclusion

6.1. C-OOD detection is an ill-posed problem

A quick glance between a standard ResNet50 model and a robust CLIP-RN50 on Table 1 shows a large gap performance, indicating that CLIP-RN50 performs worst on all datasets. However, given that the model has higher accuracy, these false positives actually contain potentially classified images, and rejecting them would reduce the overall performance quality of the model. In contrast, a ResNet50 model with 0% accuracy on ImageNet-A would benefit more when these images are detected. Even if we compare across different architecture or training strategies (e.g., Robust ResNet50 and ViT-B-16 having better accuracies with similar or better FPR95), there's no notion of whether this performance comes with the cost of losing potential classified images.

We further discussed this problem based on the score distributions across different models in Figure 2 using the existing and the MS-OOD Framework. Notice how the red area (denoted MS-OOD@COOD) that covers C-OOD increases when the model is weak (since there will be more misclassified data). Hence, once the red region covers the entire C-OOD area, the notion of FPR95 then aligns with our classification objective. However, a robust model's false positive will constitute both misclassified and classified regions of covariate shift data.

Remark. We are aware of the logits uniform distribution characteristics on CLIP models (Ming et al., 2022), causing MaxLogit to perform poorly across all datasets. Hence,

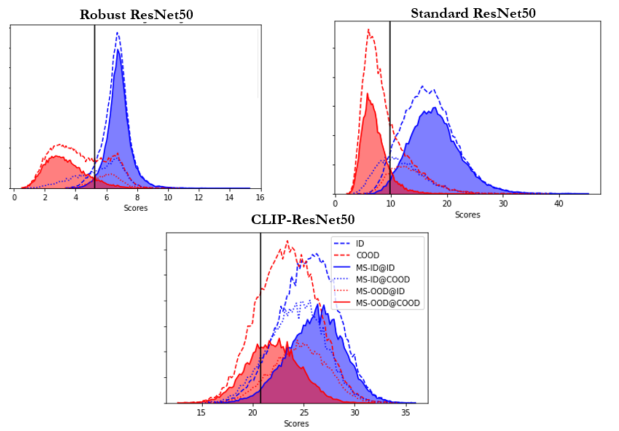

Figure 2. **The score distributions on ImageNet-R** across ResNet50 using standard training, robust and CLIP. Blue denotes MS-ID and red denotes MS-OOD respectively. For C-OOD detection, using the 95% threshold of ID, the false positives contain potential classified images of C-OOD. The shaded regions denote MS-ID@ID vs MS-OOD@COOD to distinguish classified ID data and misclassified COOD data.

we also include softmax output and still observe a similar pattern (lower FPR across datasets).

6.2. COOD detection under MS-OOD Framework

Table 2 depicts algorithms performance across models and datasets using MS-OOD Framework from two sources: MS-OOD@ID and MS-OOD@COOD. We use accuracy to further distinguish the model's *robustness*. We employ FPR95@ID+ for MS-OOD@ID and F95+ for MS-OOD@COOD, using MS-ID@ID as the source to better mimic practical scenarios. The F-Score is a metric that considers both precision and recall (hence both MS-OOD@COOD+ and MS-OOD@COOD-) for calculation, serving as the appropriate metric for the COOD setting.

Looking at models across datasets, we can see that higher accuracy leads to higher F-Score, indicating that employing robustness techniques (either by increasing depth or using pretraining) serves as a good way to improve COOD detection. A similar trend occurs in MS-OOD@ID scenario (except for Robust ResNet50), which is not previously observed in (Xia & Bouganis, 2022). This draws our attention that a careful training strategy is needed for a model to know the difference between *classification* task and *knowing what it doesn't know*.

Furthermore, we show that the baseline method and MaxLogit perform the best across all datasets, models, and sources. Notice that algorithms relying on features perform worse overall compared to output-based methods. We hypothesize the usage of training data backfires for both MS-OOD@ID and MS-OOD@C-OOD, given that they are still within the same semantic space with no clear distinction between covariate shift and in-distribution. The obvious implication can be seen on MS-OOD@ID which shows an overall worse performance, following similar results to (Xia & Bouganis, 2022).

6.3. Existing Framework vs MS-OOD Framework on S-OOD detection

The clear distinction between these two frameworks for S-OOD detection is whether we use the entire in-distribution (e.g. training data) or only the classified images MS-ID@ID. We can observe in Table 3 that the current approach in dealing with OOD detection underestimates the performance of these detection methods. Another key point we found is the consistency of the algorithms' performance between the two frameworks. For instance, a model that performs the best under the previous framework would still maintain its rank under the MS-OOD framework. However, when looking at the average across datasets, we found ResNet152 and CLIP-ResNet50 to shift their best-performing algorithm from feature-based to output-based. Furthermore, MSP along with MaxLogit and Energy observes a higher overall improvement in the false positive rate, indicating that these scores are better representatives when trying to consider detection from classified images. Another implication is that these methods are the better scores for distinguishing the classified images inside the in-distribution data.

6.4. MS-OOD Framework under only C-OOD data

For ensuring that the performance advantage of output-based methods is not due to setting the threshold based on MS-ID@ID, we compare results only using C-OOD datasets. That is, we consider in-distribution vs out-of-distribution for MS-ID@COOD and MS-OOD@COOD. We then calculate the FPR95@COOD+ to see which one performs the best. Based on Table 4, we can see how output-based methods fare similarly to Table 2. Furthermore, our results expand from (Xia & Bouganis, 2022), that not only MSP fare well on in-distribution datasets, but also on covariate shift datasets and across models.

6.5. Conclusion

We can see from previous tables that *MSP shows promising results for detection under the model's semantic space*, although state-of-the-art approaches for OOD detection still reign in superior performance for S-OOD under MS-OOD framework, hence the need to consider possible OOD scenarios encountered in the wild. Furthermore, we show there's consistency when we look model from classifier and algorithms performance on certain datasets, making it critical to consider detection problem from both perspective.

References

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., and Schmidt, L. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, pp. 6216–6234. PMLR, 2022.

Fort, S., Ren, J., and Lakshminarayanan, B. Exploring the limits of out-of-distribution detection. *CoRR*, abs/2106.03004, 2021. URL https://arxiv.org/abs/2106.03004.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *CVPR*, 2021b.

Hsu, Y.-C., Shen, Y., Jin, H., and Kira, Z. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10951–10960, 2020.

Huang, R. and Li, Y. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8710–8719, 2021.

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

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

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

Ming, Y., Yin, H., and Li, Y. On the impact of spurious correlation for out-of-distribution detection. *CoRR*, abs/2109.05642, 2021. URL https://arxiv.org/abs/2109.05642.

Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., and Li, Y. Delving into out-of-distribution detection with vision-language representations. *arXiv preprint arXiv:2211.13445*, 2022.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021a. URL https://arxiv.org/abs/2103.00020.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021b.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. Imagenet-21k pretraining for the masses, 2021.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

Salehi, M., Mirzaei, H., Hendrycks, D., Li, Y., Rohban, M. H., and Sabokrou, M. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*, 2021.

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

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

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Vaze, S., Han, K., Vedaldi, A., and Zisserman, A. Open-set recognition: A good closed-set classifier is all you need. *CoRR*, abs/2110.06207, 2021. URL https://arxiv.org/abs/2110.06207.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019.

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

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

Wilson, G. and Cook, D. J. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

Xia, G. and Bouganis, C.-S. Augmenting softmax information for selective classification with out-of-distribution data. In *Proceedings of the Asian Conference on Computer Vision*, pp. 1995–2012, 2022.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Yang, J., Wang, H., Feng, L., Yan, X., Zheng, H., Zhang, W., and Liu, Z. Semantically coherent out-of-distribution detection. *CoRR*, abs/2108.11941, 2021a. URL https://arxiv.org/abs/2108.11941.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. *CoRR*, abs/2110.11334, 2021b. URL https://arxiv.org/abs/2110.11334.

Yang, J., Zhou, K., and Liu, Z. Full-spectrum out-of-distribution detection, 2022. URL https://arxiv.org/abs/2204.05306.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014.

Table 5. Appendix: Complete table of Table 3.

MODEL	METHODS	ID	S-OOD															
		IN	SVHN		DTD		Places365		iNaturalist		SUN		IN-O		SSB-IN-Easy		AVG	
		ACC↑	FPR95↓	FPR95 @ID+↓	FPR95↓	FPR95 @ID+↓	FPR95↓	FPR95 @ID+↓	FPR95↓	FPR95 @ID+↓	FPR95↓	FPR95 @ID+↓	FPR95↓	FPR95 @ID+↓	FPR95↓	FPR95 @ID+↓	FPR95↓	FPR95 @ID+↓
DEPTH																		
ResNet18	**MSP**	69.76	10.67	2.52	70.16	46.33	75.93	50.95	58.22	30.67	73.45	47.99	99.65	97.80	79.00	**54.52**	66.73	47.25
	MaxLogit		5.80	1.10	57.18	36.12	66.94	45.25	54.51	29.47	62.92	39.54	96.80	89.90	75.36	54.98	59.93	42.34
	Energy		6.53	1.53	52.82	36.06	64.97	46.23	56.49	33.70	59.75	39.83	93.15	85.75	74.97	57.87	58.38	43.00
	ViM		1.61	0.59	38.83	26.97	92.47	85.94	91.15	83.67	93.08	87.19	**72.70**	**64.00**	80.09	68.88	67.13	59.61
	GradNorm		**0.29**	**0.12**	**29.31**	**23.35**	**45.48**	**38.93**	**26.76**	**21.22**	**34.66**	**28.56**	89.25	85.50	**71.69**	65.14	**42.49**	**37.55**
ResNet50	**MSP**	76.13	12.89	4.21	66.01	45.11	71.57	48.78	52.77	29.58	68.58	45.41	100.00	100.00	72.57	50.81	63.48	46.27
	MaxLogit		7.46	2.33	54.36	36.28	65.68	46.18	50.87	26.75	59.90	38.76	100.00	100.00	**69.02**	**50.00**	58.18	42.90
	Energy		8.18	2.94	52.13	37.77	65.40	48.91	53.95	32.48	58.28	41.19	100.00	99.90	69.19	53.38	58.16	45.22
	ViM		**0.79**	**0.16**	**15.74**	**9.10**	83.52	72.38	71.78	55.45	82.06	69.65	**84.90**	**79.20**	76.16	63.29	59.28	49.89
	GradNorm		1.14	0.86	32.39	29.31	**48.69**	**44.80**	**26.95**	**23.71**	**37.25**	**33.78**	95.60	94.35	69.82	66.34	**44.55**	**41.88**
ResNet152	**MSP**	78.31	25.55	10.45	59.84	42.13	68.93	48.38	49.26	30.33	66.01	45.89	99.30	96.95	71.31	51.74	62.89	46.55
	MaxLogit		18.00	5.44	45.59	29.63	58.72	**40.23**	46.31	**26.85**	51.87	**34.06**	96.15	89.55	**66.05**	**48.39**	54.67	**39.16**
	Energy		21.51	7.24	43.83	29.52	57.71	41.65	49.36	30.65	50.25	34.73	93.70	88.05	66.37	50.58	54.68	40.35
	ViM		**0.27**	**0.08**	**12.98**	**7.71**	78.58	64.66	63.14	43.97	77.76	63.40	**71.45**	**61.20**	73.54	60.18	53.96	43.03
	GradNorm		3.49	2.94	31.91	30.32	**55.87**	53.69	**37.53**	35.50	**43.46**	41.73	91.30	89.90	73.99	72.26	**48.22**	46.62
TRAINING																		
ResNet50	**MSP**	76.13	12.89	4.21	66.01	45.11	71.57	48.78	52.77	29.58	68.58	45.41	100.00	100.00	72.57	50.81	63.48	46.27
	MaxLogit		7.46	2.33	54.36	36.28	65.68	46.18	50.87	26.75	59.90	38.76	100.00	100.00	**69.02**	**50.00**	58.18	42.90
	Energy		8.18	2.94	52.13	37.77	65.40	48.91	53.95	32.48	58.28	41.19	100.00	99.90	69.19	53.38	58.16	45.22
	ViM		**0.79**	**0.16**	**15.74**	**9.10**	83.52	72.38	71.78	55.45	82.06	69.65	**84.90**	**79.20**	76.16	63.29	59.28	49.89
	GradNorm		1.14	0.86	32.39	29.31	**48.69**	**44.80**	**26.95**	**23.71**	**37.25**	**33.78**	95.60	94.35	69.82	66.34	**44.55**	**41.88**
Robust ResNet50	**MSP**	80.86	38.41	20.77	71.91	56.22	**74.06**	**57.29**	59.51	40.19	**70.86**	**53.12**	99.55	98.80	**72.80**	**56.58**	69.59	54.71
	MaxLogit		54.31	37.01	75.43	64.26	77.84	65.53	67.35	51.53	75.04	62.21	91.00	85.00	75.75	63.82	73.82	61.34
	Energy		99.98	99.97	95.74	95.64	97.27	97.22	98.07	98.00	97.72	97.62	**29.50**	**29.40**	95.59	95.46	87.70	87.62
	ViM		**0.07**	**0.03**	**21.91**	**18.72**	77.56	72.29	**31.42**	**25.51**	73.30	67.25	64.65	57.55	73.87	67.45	**48.97**	**44.11**
	GradNorm		100.00	100.00	97.45	97.77	99.64	99.77	99.98	99.98	99.76	99.81	98.60	98.80	99.49	99.58	99.27	99.39
CLIP																		
ResNet50	**MSP**	76.13	12.89	4.21	66.01	45.11	71.57	48.78	52.77	29.58	68.58	45.41	100.00	100.00	72.57	50.81	63.48	46.27
	MaxLogit		7.46	2.33	54.36	36.28	65.68	46.18	50.87	26.75	59.90	38.76	100.00	100.00	**69.02**	**50.00**	58.18	42.90
	Energy		8.18	2.94	52.13	37.77	65.40	48.91	53.95	32.48	58.28	41.19	100.00	99.90	69.19	53.38	58.16	45.22
	ViM		**0.79**	**0.16**	**15.74**	**9.10**	83.52	72.38	71.78	55.45	82.06	69.65	**84.90**	**79.20**	76.16	63.29	59.28	49.89
	GradNorm		1.14	0.86	32.39	29.31	**48.69**	**44.80**	**26.95**	**23.71**	**37.25**	**33.78**	95.60	94.35	69.82	66.34	**44.55**	**41.88**
CLIP-ResNet50	**MSP**	59.82	10.29	4.24	66.76	**46.65**	75.88	**54.61**	**62.65**	**39.30**	70.83	**48.72**	95.80	84.95	80.08	**60.29**	66.04	**48.40**
	MaxLogit		61.22	29.34	89.52	82.55	**72.28**	59.11	79.53	63.79	69.12	53.47	75.85	67.40	**77.91**	68.30	75.06	60.57
	Energy		97.11	94.22	94.36	92.34	79.50	72.79	90.14	84.56	79.59	72.20	**72.65**	**67.10**	81.20	76.74	84.94	79.99
	ViM		96.72	93.43	94.68	92.71	80.91	74.33	91.62	86.34	81.75	74.12	73.25	68.10	81.70	77.03	85.80	80.87
	GradNorm		**0.01**	**0.01**	**48.30**	46.70	74.95	73.15	77.54	75.89	**60.59**	58.38	89.20	88.50	84.04	83.02	**62.09**	60.81
ARCHITECTURE																		
ResNet50	**MSP**	76.13	12.89	4.21	66.01	45.11	71.57	48.78	52.77	29.58	68.58	45.41	100.00	100.00	72.57	50.81	63.48	46.27
	MaxLogit		7.46	2.33	54.36	36.28	65.68	46.18	50.87	26.75	59.90	38.76	100.00	100.00	**69.02**	**50.00**	58.18	42.90
	Energy		8.18	2.94	52.13	37.77	65.40	48.91	53.95	32.48	58.28	41.19	100.00	99.90	69.19	53.38	58.16	45.22
	ViM		**0.79**	**0.16**	**15.74**	**9.10**	83.52	72.38	71.78	55.45	82.06	69.65	**84.90**	**79.20**	76.16	63.29	59.28	49.89
	GradNorm		1.14	0.86	32.39	29.31	**48.69**	**44.80**	**26.95**	**23.71**	**37.25**	**33.78**	95.60	94.35	69.82	66.34	**44.55**	**41.88**
ViT-B-16	**MSP**	81.07	21.54	10.09	58.30	41.28	68.68	**49.68**	51.52	32.03	66.56	**48.08**	95.65	91.30	71.31	**53.54**	61.94	46.57
	MaxLogit		18.52	8.23	54.84	38.09	69.14	52.49	52.26	33.03	66.89	49.91	82.90	72.60	**70.17**	54.13	59.25	44.07
	Energy		24.15	13.50	57.39	46.01	74.31	65.24	64.08	50.90	72.77	62.80	**68.65**	62.20	73.77	63.99	62.16	52.09
	ViM		**2.34**	**0.85**	**43.94**	**31.49**	**61.12**	50.35	**17.77**	**10.66**	**59.46**	49.05	72.90	**61.80**	71.58	58.70	**47.01**	**37.56**
	GradNorm		96.13	96.43	92.87	93.14	96.70	96.91	95.56	95.93	96.85	97.05	98.65	98.75	97.86	98.04	96.38	96.61