# Rethinking Out-of-Distribution Detection: The Model Perspective

William Yang*, Byron Zhang*, Olga Russakovsky
Princeton University
{williamyang, zishuoz, olgarus}@princeton.edu

## Abstract

*Out-of-Distribution (OOD) detection is the task of identifying inputs on which a given model should not be trusted to make a prediction. OOD detection is a well-studied problem, commonly formulated for image classifiers as detecting test images from classes that did not appear in the model's training data (semantic shift). However, this formulation is limited, as it ignores test images which may come from familiar classes but may look different from the training examples (covariate shift). Incorporating covariate shift into OOD detection poses a challenge because detecting* all *such examples may undervalue the model's robustness and ability to generalize beyond its training distribution.*

*In this work we propose to evaluate OOD detection from a new perspective: approaching the problem with respect to the learned model distribution rather than the training data distribution. This formulation seamlessly incorporates both semantic and covariate shift, and synergizes the task of OOD detection with model generalization. Our empirical analysis reveals a number of interesting findings, the most striking being that the simplest OOD detection baseline, Maximum Softmax Probability (MSP), outperforms all prior state-of-the-art OOD detection methods under the new perspective across semantic and covariate shifts. This suggests that the time is ripe to rethink the task of OOD detection, and to adopt the more generalized framework which encompasses a broader range of real-world situation.*

## 1. Introduction

Modern computer vision systems have achieved remarkable performance in closed-world classification settings. However, the performance of such systems decreases in real-world settings where images encounter different types of distribution shifts, attributed to camera calibrations, sensory corruptions, introduction of novel object classes, etc. Addressing such distribution shifts plays a crucial role in the safety and practicality of deep learning models in real-world applications such as autonomous vehicles.

(a) In-distribution (b) Covariate Shift (c) Semantic Shift

Figure 1: **Different types of distribution shifts.** Assume a model is trained to predict ImageNet [22] dogs (a). The current perspective on out-of-distribution (OOD) detection specifies that images of cats (c) should be considered OOD, but does not consider the case when encountering images of dogs that do not belong to the distribution of ImageNet dogs (b). We offer a new perspective on OOD detection that incorporates such cases for evaluation of OOD detectors.

Out-of-distribution (OOD) detection aims to address such distribution shifts by rejecting problematic examples to promote the safety and reliability of deep learning models in real-world applications. Common literature in OOD detection separates distribution shifts into two categories: semantic and covariate [12, 26, 31] (Figure 1). Under semantic shift, images are drawn from classes that do not belong to the in-distribution (ID) dataset (*i.e.* training dataset). An image classifier will make erroneous predictions on images with semantic shift because the concepts of interest are not present. The appropriate action to handle images with semantic shift is thus to reject these images or to flag them for human intervention. Under covariate shift, images are drawn from the classes in the training dataset but exhibit differences in terms of style, contrast, brightness, domain, etc. An image classifier will not necessarily make erroneous predictions because it may have some robustness towards these shifts. Therefore, the appropriate action for handling images with covariate shifts becomes unclear because it is not ideal to reject correctly classified examples.

Research on detection of covariate shifts remains controversial due to its conflicting objective with OOD gen-

* equal contributors.

eralization [31]. Mainstream research in OOD detection evades this conflict by not considering covariate shifts at all [20, 19, 18, 29, 14, 24, 15]. Some works incorporate covariate shift only from the **data perspective**—by characterizing all the covariate shifted data as only ID [32] or only OOD [12]. However, not every covariate shifted example leads to an erroneous prediction. As a result, under this characterization, an OOD detector can reject too many correctly classified examples, which hurts generalization, or too few incorrectly classified examples, which undermines safety.

To incorporate covariate shifts, we argue that OOD detection should be evaluated from the **model perspective** instead of the data perspective. Similar to [16], the model perspective motivates the task of OOD detection from its first principle—catching model failures before they occur. Concretely, we consider distribution shifts with respect to the distribution of the trained classification model instead of the training data, since the notion of safety and generalization is highly dependent on the model behavior. We reformulate the notion of *in-distribution* as data that the model generalizes to, *i.e*. correctly classifies. As a result, the model perspective accounts for the false rejection of correct predictions, resolving the emerging issues from covariate shifts and unifies OOD detection with OOD generalization. We further show that the model perspective on distribution shift naturally emerges when we consider the notion of failure with respect to the training objective.

To empirically motivate the model perspective on distribution shifts and better understand current OOD detection algorithms, we perform extensive analysis using covariate shift datasets ImageNet-C [10], ImageNet-R [9], and our newly proposed semantic shifted dataset ImageNet-OOD. We demonstrate that current OOD detection algorithms do detect covariate shifts, sometimes to a **greater extent** than semantic shift. We also discover several shortcomings under existing OOD formulation and the mitigation of such issues under the new model perspective. Our analysis reveals that under the model perspective, the simplest baseline (Maximum Softmax Probability-MSP [11]) outperforms many recent OOD detection algorithms across both semantic shift and covariate shift datasets, sometimes to the extent of 5%-7% during evaluation on Area Under Receiver Characteristics Operating Curve (AUROC). Finally, we empirically show that the model perspective better aligns with our intuition on the notion of in-distribution and does not conflict with the classification task performance.

## 2. Related Work

**Covariate shift in OOD detection.** Covariate shift was first considered in OOD detection by Generalized ODIN [12], where OOD detection performance was evaluated considering all covariate shifted examples as out-of-distribution. Tian *et al*. [26] designed a scoring function that disentangles detection of semantic vs. covariate shifts. Later, Yang *et al*. [32] pointed out that models should ideally generalize, instead of detect, in the case of covariate shifts, because generalization is the primary goal of machine learning. Thus, they defined all covariate shifted examples as in-distribution and proposed several benchmarks that include covariate shifted data. Our work share similar motivation as [32], but we do not characterize all covariate shifted examples as in-distribution, as the rejection of covariate shift examples that would otherwise hurt the model's classification performance should not be penalized.

**OOD Detection Methods.** OOD detection methods are scoring functions that assign higher scores to ID examples and lower scores to OOD examples. Logit-based methods [11, 20, 8] derive scoring functions from classification logits with the intuition that OOD examples tend to have lower activations across all classes. Feature-based methods [18, 25, 29] operate on the penultimate layer of the model by comparing the features of OOD data to those of the training data. Gradient-based methods incorporate information in the gradient space to generate more gap between ID and OOD data, either implicitly through adversarial perturbation [19], or explicitly through backpropogation of the gradient norm [14]. Previous works have also explored the modification of the training process, through decomposing the prediction head to incorporate prior knowledge [12] and using group labels to simplify the decision boundary between ID vs. OOD examples [15]. We perform our evaluation using representative logit-based and feature-based detectors, which require minimal computational cost [31].

**Failure Detection.** OOD detection is often categorized as a sub-task of failure detection, as the primary goal of OOD detection is to catch unsafe prediction before models make a mistake. Other tasks that share this common goal of failure detection include misclassification detection [11] and uncertainty calibration [17], both of which differs from OOD detection in that they do not consider examples outside the set of training classes. However, since all failure detection methods can be interpreted as scoring functions and are motivated from a common principle (alerting failure before occurrence), Jaeger *et al*. [16] argues that all failure detection tasks should be evaluated across a common benchmark that takes into account the classification performance of the model. Our work shares similar motivation as [16], but focuses specifically on the behavior of covariate distribution shifts within the framework of OOD detection, which is not explored in the current field of failure detection.

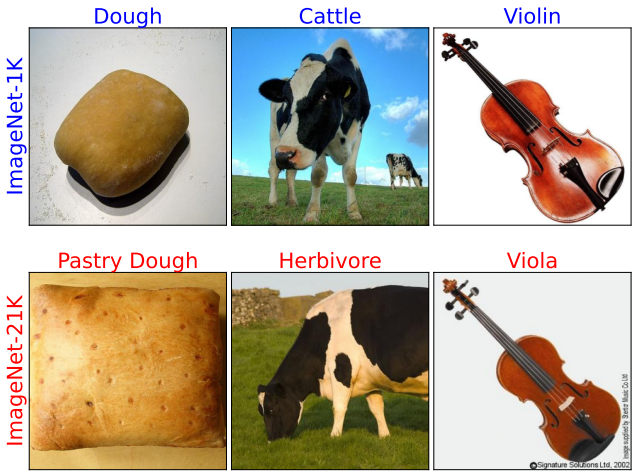



Figure 2: **Example ImageNet-21K [3] classes omitted from ImageNet-OOD.** With ImageNet-1K as ID, The following classes are removed: *Left*. Semantic ancestors and descendants of ImageNet-1K classes (e.g. "Pastry Dough", a semantic child of ID class "Dough"). *Middle*. Classes under the Wordnet [5] subtree "organism" (e.g. "Herbivore", a semantic parent of ID class "cattle", even though this relationship is not captured in the WordNet hierarchy). *Right*. Classes visually ambiguous to labelers (e.g. "viola," as it is visually indistinguishable from ID class "violin".

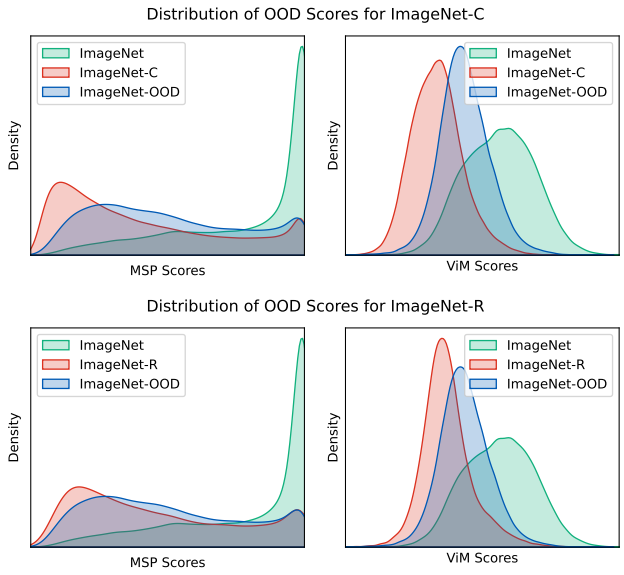



Figure 3: **Distribution of OOD scores.** A kernel density estimator reveals how the baseline OOD detector MSP [11] and the modern OOD detector ViM [29] scores images from different data sources. Results reveal that for both covariate shift datasets ImageNet-C [10] and ImageNet-R [9], the distribution of scores is lower than that of semantic shift.

## 3. Importance of Covariate Shift

Common changes within the data distribution fall under two categories: covariate shift, which is label-preserving (i.e. concerns examples only from training classes), and semantic shift, which is label-altering (concerns examples only from new classes). Prior work formulates the OOD detection problem exclusively around semantic shifts, assuming that only examples with semantic shift can lead to erroneous model behavior. Recent concurrent work points out this problematic assumption and argues that failure prevention tasks such as OOD detection should incorporate covariate shift for more comprehensive evaluation [16]. In this section, we show empirical support for this argument, compare OOD detection behavior on semantic vs. covariate shifts with large-scale datasets, and illustrate additional unintended consequences of neglecting covariate shifts.

### 3.1. Preliminaries

**Problem Setup.** For the task of image classification, a dataset $D_{tr} = \{(x_i, y_i); x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$ sampled from training distribution $P_{tr}(x, y)$ is used to train some classifier $C : \mathcal{X} \to \mathcal{Y}$. In real-world deployments, distribution shift occurs when classifier $C$ receives data from test distribution $P_{te}(x, y)$ where $P_{tr}(x, y) \neq P_{te}(x, y)$ [21]. An OOD detector is a scoring function $s$ that maps an image $x$ to a real number $\mathbb{R}$ such that for some threshold $\tau$, we arrive at a detection rule $f$

$$f(x) = \begin{cases} \text{in-distribution} & \text{if } s(x) \geq \tau \\ \text{out-of-distribution} & \text{if } s(x) < \tau \end{cases} \quad (1)$$

**Covariate Shift.** Covariate shift occurs when the marginal distribution with respect to the image differs: $P_{tr}(x) \neq P_{te}(x)$ [31], while the label distribution remains fixed: $P_{tr}(y|x) = P_{te}(y|x)$.

**Semantic Shift.** Semantic shift occurs when given sets of semantic labels $Y_{tr} \subset \mathcal{Y}$ from the training distribution and semantic labels $Y_{te} \subset \mathcal{Y}$ from a test distribution, the two sets have the following property: $Y_{tr} \cap Y_{te} = \emptyset$, such that $P_{tr}(y) = 0 \; \forall y \in Y_{te}$.

**Datasets.** In this work, we examine the OOD detection behaviors of classifiers pretrained on ImageNet-1K [22]. For covariate shifts, we evaluate on two common datasets used in the field of robustness and OOD generalization: ImageNet-C [10], which consists of 75 corrupted versions (15 corruption types $\times$ 5 severity levels*) of the ImageNet validation set, and ImageNet-R [9], which consists

*Severity level 3 is used throughout this work unless otherwise stated.

of 30,000 images containing various renditions (e.g., paintings, embroidery, etc.) from ImageNet-1K classes [9].

To establish fair comparison between detection behaviors on semantic vs. covariate shifts, a desirable semantic shift dataset should (1) satisfy the condition $Y_{tr} \cap Y_{te} = \emptyset$, and (2) minimizes of unintended covariate shifts that does not affect image semantics. We introduce a new large-scale dataset, ImageNet-OOD, with 50,000 images spanning 1,000 classes from ImageNet-21K [3] based the WordNet [5] semantic tree, providing stronger guarantees to these two properties than prior evaluation practices in OOD detection. Our new dataset satisfies condition (1) by carefully removing a set of classes that have visual and semantic ambiguities with ID classes (see Figure 2), and condition (2) by virtue of the fact that ImageNet-21K comes from the same data collection process as ImageNet-1K. More details on the process of constructing ImageNet-21K can be found in the supplemental materials.

### 3.2. Complications Introduced by Covariate Shift

Covariate shift detection seems ill-motivated at first: why detect covariate shift rather than leave it for the field of OOD generalization? However, OOD detectors cannot choose to avoid covariate shifts during test time, and thus, we cannot guarantee that all examples are generated only from semantic-shifted distributions. Our analysis demonstrates that many OOD detectors are heavily confounded by covariate shifts without being instructed to detect them.

Unlike semantic shift, where the model is guaranteed to fail, the outcome from covariate shift is ambiguous, because models may have learned to generalize to certain examples. In the following experiment, we will show that that OOD detectors cannot *only* reject a large proportion of semantic shift data without simultaneously rejecting a large proportion correctly classified covariate shift data.

Using a ResNet-50 [7] model pretrained on ImageNet-1K, we obtained Kernel Density Estimation (KDE) plots in Figure 3 of OOD detection scores for test data (ImageNet-1k), covariate shift (ImageNet-C and ImageNet-R), and semantic shift (ImageNet-OOD) datasets. The plots reveal that most of the of ImageNet-C and ImageNet-R scores are lower than those of ImageNet-1k and ImageNet-OOD scores. As a consequence, an OOD detector cannot reject data from ImageNet-OOD without rejecting a substantial number of ImageNet-R and ImageNet-C data.

We provide a more detailed breakdown of our analysis on Figure 3 by discarding the scores from test data and partitioning the covariate shift scores by correctly classified vs. incorrectly classified examples. Our findings in Figure 4 reveals that OOD detection algorithm ViM outputs similar scores for correct ImageNet-R examples, and even lower on correct ImageNet-C examples when compared to ImageNet-OOD examples. To put into perspective, correctly rejecting 75% of semantic OOD data on ImageNet-OOD would simultaneously reject 85% of correct examples in ImageNet-C and 69% of correct examples in ImageNet-R. As a result, OOD detectors cannot neglect covariate shift in evaluation because false rejection may potentially hurt model performance and undermine the goal of model generalization. Full quantitative results are included in the supplementary. In the next section, we will address the conflict with model generalization by providing a new perspective on the OOD detection problem.

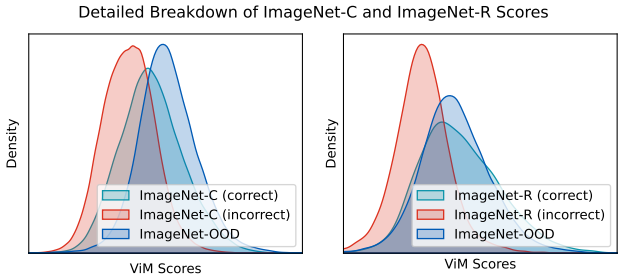



Figure 4: **Score breakdown on covariate shift data.** Comparison of ViM [29] score distributions for correctly classified covariate shift examples (ImageNet-C and ImageNet-R) and semantic shift examples (ImageNet-OOD). Scores of correct ImageNet-R examples are visually inseparable from ImageNet-OOD examples, and scores of correct ImageNet-C examples tend to be even lower than ImageNet-OOD examples. In both cases, rejecting a significant portion of semantic shift data leads to the rejection of a significant portion of correct covariate shift data.

## 4. Integrating Covariate Shift

Dealing with covariate shift creates many new nuances that are difficult to resolve with current formulation of OOD detection. The solution to the complications is rather simple if we revisit the key goal of OOD detection: to detect model faults. In this section, we will present a new perspective on distribution shift through the lens of fault detection and construct a model-centric framework of evaluating OOD detection under the new perspective.

### 4.1. The Model Perspective

**New Detection Rule.** Instead of using the data source as the positive and negative labels for the evaluation of OOD detection, we leverage the correctness of the model prediction as the target label. Rigorously, given an image $x$, a sample $y \sim P_{te}(Y|X = x)$, and $\hat{y} = \text{argmax}_y P_\theta(Y|X = x)$, $x$ is considered in-distribution if $y = \hat{y}$ and out-of-distribution otherwise. Defining our target label with respect to model predictions captures both the intricacies of the model and the data. Additionally, the new evaluation procedure natu-

rally circumvents the issue with covariate shift by directly addressing model failure. Figure 5 demonstrates the workflow of the evaluation pipeline and the key differences of the new perspective across different data scenarios.

**Connection with Failure Detection.** In contrast to previous works on failure detection, our work expand the idea to different types of data distribution shifts (*i.e.* covariate and semantic), which is more reflective of real-world classification settings. Concurrent work from Jaeger *et al.* [16] explores a similar notion of failure detection with adjacent tasks of selective classification and uncertainty quantification. Our work provides a more in-depth exploration of ideas within the landscape of OOD detection, a more comprehensive formulation of distribution shift that includes incorrectly classified in-distribution examples, and clear empirical support motivating the need to revisit the notion of failure detection with datasets of larger scale.

**Connection with Distribution Shift.** Utilizing the model prediction as label still falls under the realm of OOD detection. The differences are not within the task definition itself but rather the distribution we define as "in-distribution." Under the previous data perspective, the data generating distribution for the training data $P_{tr}(x, y)$ is used to categorize ID vs. OOD examples. Hence, the positives and negatives are determined by the data source. Instead, we propose the model perspective where we use the fitted model distribution $P_\theta(y|x)$ to categorize ID vs OOD examples. The usage of model correctness as positive and negative examples naturally emerges under this perspective.

The goal of the image classification task is to use data pairs $(x, y)$ sampled from the training distribution $P_{tr}(x, y)$ to build a classifier $P_\theta(y|x)$ with the objective of matching model distribution with the training distribution $P_\theta(y|x) = P_{tr}(y|x)$. The objective of distribution matching organically manifests as the KL-Divergence between the distributions from the maximum likelihood objective given the common assumption of image classification: only a single object present and no errors in the labeling process.

Current formulation of distribution shift with respect to the training distribution $P_{tr}$ to prevent failures in the fitted model $P_\theta$ implicitly assumes that sufficient minimization of Equation 2 to the point where $P_{tr}$ can adequately describe $P_\theta$. However, this is never true in practice due to the high dimensional nature of images, limits in the model capacity, and noise during the training process. Additionally, we expect our model to generalize beyond $P_{tr}$. If we approach the OOD detection problem from its first principle of catching failures, it becomes clear that the distribution shift of interest is with respect to $P_\theta$ instead of $P_{tr}$.

$$D_{KL}(P_{tr}(y|x) \parallel P_\theta(y|x)) = \underbrace{\mathbb{E}_{P_{tr}(y|x)}[\log P_{tr}(y|x)]}_{\text{entropy close to 0 given clean labels}} - \underbrace{\mathbb{E}_{P_{tr}(y|x)}[\log P_\theta(y|x)]}_{\text{likelihood that is maximized}} \quad (2)$$

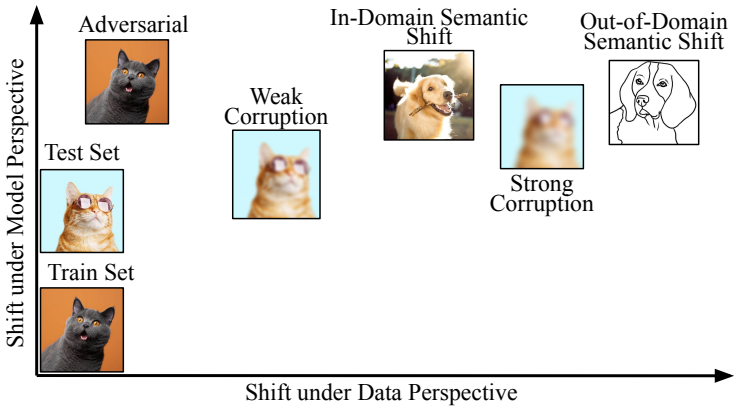
### 4.2. Model-centric Framework

With the introduction of the model perspective, we propose a new framework with metrics that considers two facets of the OOD detection problem: detection performance under the model perspective and generalization ability in the presence of the detection mechanism.

**Detection Metric.** Area under the Receiver Operating Characteristic Curve (AUROC) is an existing OOD detection evaluation metric extended to the model perspective. Formally, AUROC is a threshold free way of estimating detection performance by calculating the area under false positive rate vs. true positive rate curve. This measure has a probabilistic interpretation of $P(s(x_{in}) > s(x_{out}))$ for some OOD detection scoring function $s$, in-distribution image $x_{in}$ and out-of-distribution image $x_{out}$.

**Robustness Metric.** We also include a robustness metric that provides a better connection between the OOD detector's performance and the model's task accuracy. Accuracy vs. Declaration Rate (ADR) is an existing threshold free metric that quantifies the trade-off between model generalization and refusal of predicting uncertain images [34]. ADR calculates the area under the curve between the proportion of images an OOD detector declare as OOD vs. the classification accuracy of the model on the remaining undetected images. ADR metric accounts for model generalization where a higher value indicates better overall OOD detection and classification performance. In contrast with AUROC, ADR metric incorporate the task performance into the evaluation, allowing easy comparison across different models but at a cost of interpretability.

**Sample Reweighting.** The issue of data imbalance is exacerbated under the model perspective because positive and negative instances are retrieved from multiple datasets. Data imbalance was not an issue from the data perspective because it is implicitly assumed that each dataset is either entirely filled with positive instances or negative instances. Therefore, false positive rate (FPR) and true positive rate (TPR) are unaffected by the number of images in each of the datasets. With the new framework, positive and negative examples can come from different data sources. As a result, the metrics favor larger datasets. To counteract this undesired property, each datapoint from dataset $D$ is weighted




| | Correct Test | Incorrect Test | Correct Covariate | Incorrect Covariate | Semantic |
|---|---|---|---|---|---|
| Data Perspective | ID | ID | OOD | OOD | OOD |
| Model Perspective | ID | OOD | ID | OOD | OOD |

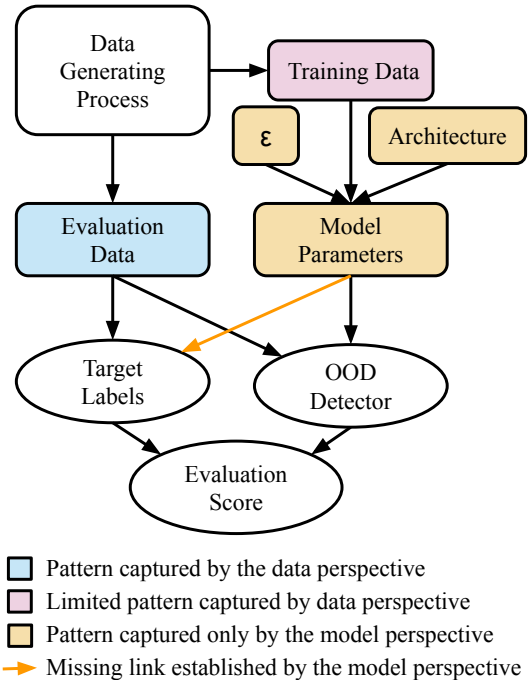



Figure 5: **Data vs. Model Perspective.** *Left Top.* Assuming ImageNet cats as the training distribution, we showcase where various data source lie in terms of shift against the model distribution vs. the data distribution. *Left Bottom.* Classification of distribution shift for correct test data, incorrect test data, correct covariate shift data, incorrect covariate shift data, and semantic shift data. *Right.* The standard workflow on the evaluation of an OOD detector. The orange arrow illustrate the key difference between the model perspective and the data perspective. Evaluation under data perspective forces the detector to pick up on patterns in the evaluation data without considering the learned model parameters. The model perspective adds the necessary link to capture model behavior.

| Dataset | ImageNet-C [10] | | | ImageNet-R [9] | | | ImageNet-OOD | | |
|---|---|---|---|---|---|---|---|---|---|
| Metric | AUROC$_D$ | AUROC$_M$ | ADR | AUROC$_D$ | AUROC$_M$ | ADR | AUROC$_D$ | AUROC$_M$ | ADR |
| MSP [11] | 84.2 | **88.3** | **73.9** | 80.5 | **88.2** | **80.4** | 79.2 | **86.8** | **65.9** |
| Max-logit [8] | 88.1 | 85.1 | 71.4 | 86.7 | 86.5 | 78.8 | **80.4** | 84.6 | 63.8 |
| Energy [20] | 88.2 | 84.0 | 70.8 | 87.0 | 85.5 | 78.3 | 79.9 | 83.6 | 63.3 |
| Mahalanobis [18] | 76.1 | 65.0 | 62.5 | 69.4 | 66.2 | 65.9 | 57.8 | 59.1 | 48.2 |
| ViM [29] | **92.2** | 82.5 | 69.3 | **88.7** | 85.4 | 78.4 | 78.6 | 82.3 | 63.3 |
| KNN [25] | 84.1 | 69.1 | 63.2 | 77.5 | 68.6 | 64.2 | 61.6 | 58.1 | 44.1 |
| ReAct [24] | 87.1 | 81.6 | 69.8 | 84.1 | 81.7 | 76.6 | 75.6 | 79.4 | 61.2 |

Table 1: **Performance differences with the new model-centric framework.** OOD detection performance under the old data perspective is labeled as AUROC$_D$, OOD detection performance under the proposed new model perspective is labeled as AUROC$_M$, and robustness under OOD detection is labeled as ADR. The data indicates MSP outperforms every other method under the model perspective.

by $\frac{1}{|D|}$ during evaluation. This will balance each dataset's contribution to the performance on the metrics.

## 5. Analyses from the Model Perspective

With the introduction of the model perspective, we can proceed with evaluating existing OOD detection algorithms. We observed that under the model perspective, the baseline MSP [11] consistently outperforms other OOD detection algorithms. We discovered that the source of this disparity is the exploitation of detecting incorrect examples as ID rather than detecting OOD examples as OOD. Finally, we show that the model perspective better aligns with our basic conception of "in-distribution" and OOD detection performance under the model perspective does not conflict with

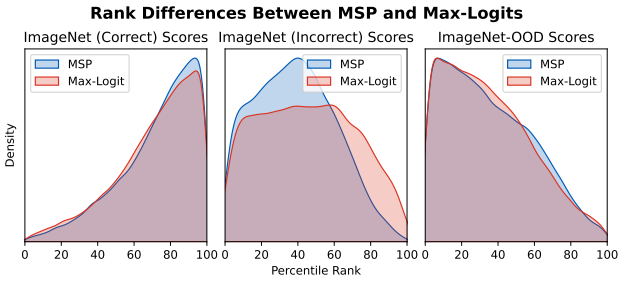



Figure 6: **Comparison on the ranking between MSP [11] and Max-Logit [8].** *Left.* MSP is slightly better at ranking correctly predicted ImageNet images higher. *Center.* Max-Logit ranks more incorrect ImageNet images higher than MSP. *Right.* MSP and Max-Logits have near identical ranks on ImageNet-OOD examples. The comparison indicates that Max-Logit improves in evaluation under data perspective without any improvement on failure prevention.

the classification task performance.

### 5.1. The Baseline is All You Need

To analyze the difference of OOD detection under the new model-centric framework, we evaluated a pretrained ResNet-50 using seven popular OOD detection methods: MSP [11], Max-Logits [8], Energy [20], Mahalahobois [18], ViM [29], KNN [25], and ReAct (with Energy) [24]. Results in Table 1 reveals that the new framework arrives at a different conclusion as to which OOD detector is best. In particular, we see that across all three datasets, the baseline MSP outperforms every other method on AUROC from the model perspective ($\text{AUROC}_M$) and ADR for calibrating model robustness. This result reorders the state-of-the-art under the data perspective where ViM or Max-Logits outperformed other methods. This revelation challenges many established intuitions in the OOD detection community such as superior performance of Energy [20] and Maximum Logits [8] with respect to MSP [11]. Similar patterns can be seen using ViT [4], DenseNet-121 [13], and Wide Resnet50 [33] (details included in supplementary material).

### 5.2. The Curious Case of Semantic Shift

Evaluation of OOD algorithms under the model perspective reaches different conclusions than the data perspective even under semantic shift. The ImageNet-OOD results in Table 1 reveals that MSP outperforms Max-Logit under the model perspective but Max-Logit outperforms MSP under the data perspective. This is particularly an unexpected outcome because under both perspectives, semantic shift (ImageNet-OOD) is considered as OOD. Breaking down the test data by whether or not they are correctly predicted resolves this mystery. Fundamentally, the goal of an OOD detection algorithm is to give lower scores, and hence, lower

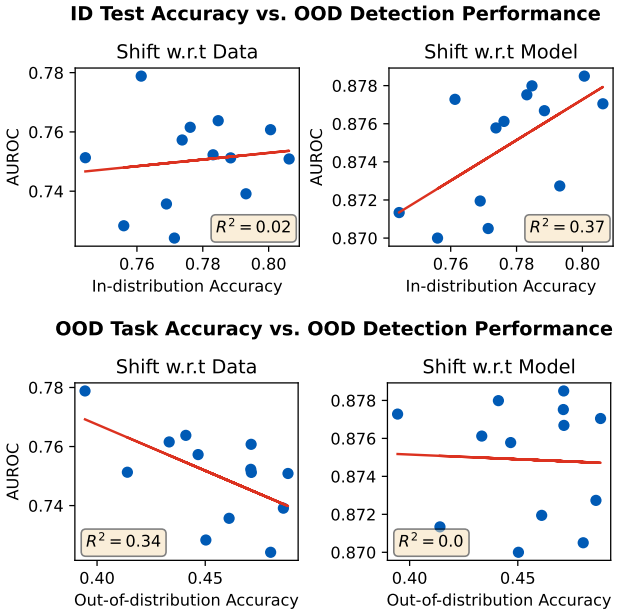



Figure 7: **Model perspective does not conflict with generalization.** *Top.* Larger $R^2$ value between ID task accuracy and OOD detection performance under the model perspective indicates that a better closed-set classifier also leads to better detection performance, aligning OOD detection and the ID task. *Bottom.* In contrast to the data perspective, there exists no negative relationship under the model perspective between OOD task accuracy and OOD detection, indicating the lack of conflict between the objectives.

rank, on semantic shift examples to detect them (Equation 1). However, we observe from Figure 6 that Max-Logit does not rank the semantic shift (ImageNet-OOD) examples lower than MSP. Instead, it ranks the incorrect test (ImageNet-1k) examples higher. In other words, Max-Logit is neither better at detecting semantic shift examples (Figure 6 *right*) and nor better at preserving correct test examples (Figure 6 *left*). Instead, it is better at preserving incorrect test examples (Figure 6 *center*). We argue that this is the *opposite* of the ideal behavior when we are trying to catch model failures. The model perspective accounts for this, and hence, results in a worse Max-Logit performance.

### 5.3. The Model Perspective Better Aligns with Task Performance

We investigate the relationship between OOD detection performance vs. task classification performance across different model architectures. Vaze *et al.* [28] showed that there exhibits a strong correlation between the closed-set classification accuracy of the model and detection performance of MSP [11] on semantic shift. With the introduction of the model perspective and covariate shift, a natural ques-

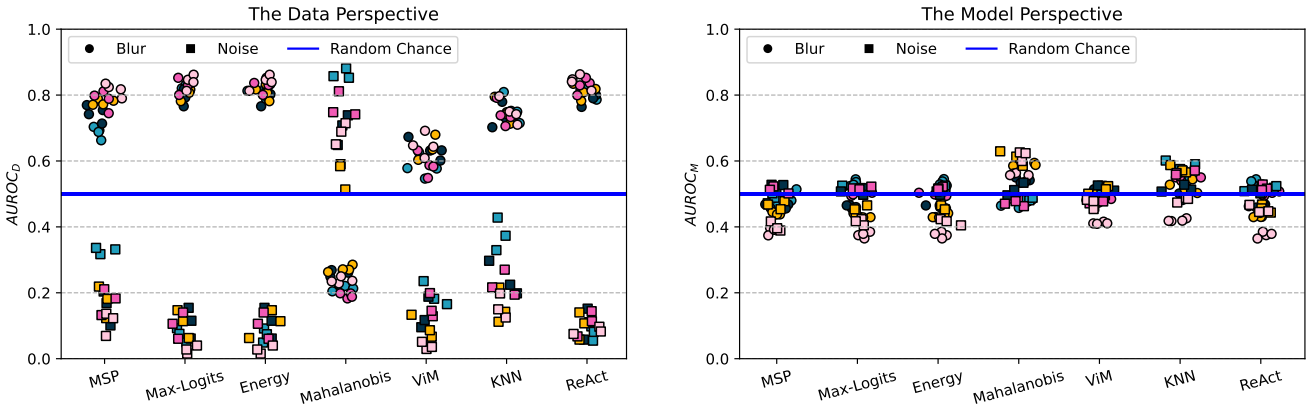



Figure 8: **Performance of OOD detection under random models.** AUROC performance of 5 ResNet-50 [7] models with *random*, untrained parameters on subsets of ImageNet-C [10] under the data perspective (left) vs the model perspective (right). Colors indicate the specific random model and the markers indicate the corruption type. Results reveal that OOD detection methods confidently identify corruptions as being "out-of-distribution" for these models – despite the models not being trained on *any* data! Thus, the existing OOD formulation appears to be misaligned. In contrast, under the model perspective, all OOD detectors perform near random chance for these untrained models, as expected.

tion emerges on whether such pattern holds: i.e. would improving the task accuracy improve the OOD detection accuracy? We performed analysis on ImageNet-C with 14 pretrained convolution-based models from torchvision ranging from ResNet-50 to RegNet-32 [30].

The outcome of our analysis shown in Figure 7 further supports the importance of the model perspective on OOD detection. From the model perspective, we observe a significantly stronger correlation between ID task accuracy and OOD detection performance of MSP, indicated by the higher $R^2$, which also suggests that ID task accuracy is a better predictor of MSP performance, and hence, better alignment between the tasks. Additionally, with the added intricacies of covariate shift, we ponder the relation between OOD task performance and OOD detection performance. Surprisingly, we observe a negative correlation between OOD accuracy and OOD detection under the data perspective, indicating that detection conflicts with generalization: i.e. increase generalization performance decreases detection performance, and vice versa. With the model perspective, we observe no correlation, resolving the conflict.

### 5.4. The Model Perspective Passes Sanity Check

The concept of "in-distribution" still has to be connected with the data that the model is trained on. Therefore, our intuition suggests that given an *untrained* model (i.e., a model with random parameters), the concept of "in-distribution" is not well-defined. Every datapoint should be considered OOD because the model has not seen any data. While previous works have demonstrated the power of random models for other tasks [1, 23, 2, 27], a well-behaved OOD detector should perform around 50% or random chance [11] on a random model. Otherwise, the OOD detector's behavior would challenge the connection between "in-distribution" and the training distribution.

For the experiments, five random models are used for evaluating OOD detection on blurring and noise corruptions with the ImageNet-C dataset against ImageNet-1k. Performance of seven commonly used OOD detection methods is evaluated on both the data perspective and model perspective. This experiment is performed on Resnet-50 [7] initialized with Pytorch's default initialization [6]. We use only the most severe corruptions in ImageNet-C.

Figure 8 reveals that under the existing data perspective, OOD detectors with random models detect blurry images as OOD and noisy images as ID. Energy [20] and Max-Logit [8] detectors illustrate this effect the strongest, where the $\text{AUROC}_D$ is greater than 0.7 and less than 0.2 respectively, a far-cry from the expected random chance. In contrast, given the same experimental setup but evaluated under the model perspective, all the OOD detection algorithm result in near 0.5 $\text{AUROC}_M$ of random chance.

## 6. Conclusion

We provide analysis of OOD detection algorithms across different distribution shifts and identified the key limitations in the existing perspective for OOD detection with the introduction of covariate shift. We introduce a novel model perspective on distribution shift that reformulates OOD detection with regard to model distribution to overcome limitations of the existing formulation and provide a more comprehensive evaluation of existing OOD detec-

tion algorithms. We also designed a new diverse OOD detection dataset *ImageNet-OOD* that accounts for the hierarchical nature of semantic concepts and potential annotation errors from human labelers. Finally, we provide holistic analysis demonstrating the desirable properties of the model perspective: better alignment with our intuition of "in-distribution", conflict removal between OOD generalization and OOD detection across different model designs.

## References


[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 8

[2] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. 8

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 3, 4

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 7

[5] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998. 3, 4

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 8

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4, 8

[8] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *ICML*, 2022. 2, 6, 7, 8

[9] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, Dawn Xiaodong Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, 2020. 2, 3, 4, 6

[10] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 2, 3, 6, 8

[11] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017. 2, 3, 6, 7, 8

[12] Y. C. Hsu, Y. Shen, H. Jin, and Z. Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10948–10957, 2020. 1, 2

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 7

[14] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, 2021. 2

[15] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[16] Paul F Jaeger, Carsten Tim Lüth, Lukas Klein, and Till J. Bungert. A call to reflect on evaluation practices for failure detection in image classification. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 5

[17] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[18] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 2, 6, 7

[19] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 2

[20] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020. 2, 6, 7, 8

[21] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012. 3

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 3

[23] Andrew M Saxe, Pang Wei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, and Andrew Y Ng. On random weights and unsupervised feature learning. In *Icml*, 2011. 8

[24] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 2, 6, 7

[25] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *arXiv preprint arXiv:2204.06507*, 2022. 2, 6, 7

[26] Junjiao Tian, Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Exploring covariate and concept shift for out-of-distribution detection. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. 1, 2

[27] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018. 8

[28] 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. 7

[29] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3, 4, 6, 7

[30] Jing Xu, Yu Pan, Xinglin Pan, Steven Hoi, Zhang Yi, and Zenglin Xu. Regnet: Self-regulated network for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–6, 2022. 8

[31] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 1, 2, 3

[32] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection, 2022. 2

[33] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016. 7

[34] Peng Zhang, Jiuling Wang, Ali Farhadi, Martial Hebert, and Devi Parikh. Predicting failures of vision systems. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3566–3573, 2014. 5