# Unified Out-Of-Distribution Detection: A Model-Specific Perspective

Reza Averly
The Ohio State University
averly.1@osu.edu

Wei-Lun Chao
The Ohio State University
chao.209@osu.edu

## 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 predicted 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 categories), ignoring other possible causes (*e.g.*, covariate shift). In this paper, we present a novel, unifying 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 (*e.g.*, an image classifier) is unable to predict correctly. That is, whether a test example should be detected and rejected or not is "model-specific". We show that this framework unifies the detection of OOD examples caused by semantic shift and covariate shift, and closely addresses the concern of applying a machine learning model to uncontrolled environments. We provide an extensive analysis that involves a variety of models (*e.g.*, different architectures and training strategies), sources of OOD examples, and OOD detection approaches, and reveal several insights into improving and understanding OOD detection in uncontrolled environments.*

## 1. Introduction

Equipping a model with the capability to identify "what it does not know" is critical for reliable machine learning. Take image classification as an example. While state-of-the-art neural network models [11, 6, 17, 24] could perform fairly well on "in-distribution (ID)" data that belong to the training distribution, their accuracy often degrades drastically when facing data with covariate shift (*e.g.*, different image domains or styles) [46] or semantic shift (*e.g.*, novel categories) [22]. It is thus crucial to detect data on which the models cannot perform well and reject them from being classified.

Out-of-distribution (OOD) detection [13], which aims to identify test examples drawn from a distribution different from the training distribution, is a promising paradigm toward such a goal. OOD detection has attracted significant attention lately, with a plethora of methods being developed [50, 35]. However, most of them focus solely on detecting examples with semantic shift, which arguably limits their applicability in uncontrolled environments where other kinds of OOD examples (*e.g.*, with covariate shift) may appear.

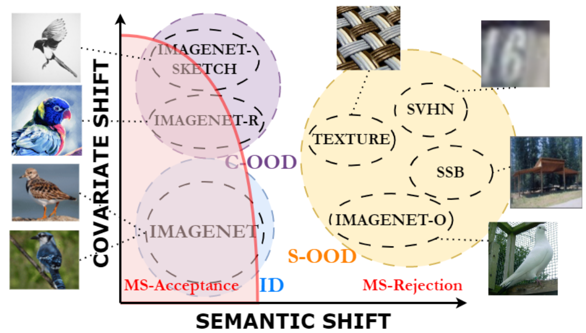


Figure 1: **Model-Specific Out-of-Distribution (MS-OOD) Detection**, using ImageNet [34, 4] as an example. The blue, purple, and yellow regions denote in-distribution (ID), covariate shift (C-OOD), and semantic shift (S-OOD) data, respectively, each with its datasets and representative images. Given an ImageNet classifier, the shaded red region denotes the *correctly classified images* (called the *acceptance region*). A robust classifier would have its acceptance region cover the ID and C-OOD data as much as possible. Using this framework, we can separate test data into **model-specific acceptance (MS-A)** and **rejection (MS-R)** cases, corresponding to the shaded red region and its complement. The goal of MS-OOD DETECTION is to detect the MS-R cases (*i.e.*, examples misclassified by the classifier).

In this paper, we thus attempt to expand the scope of OOD detection to further include covariate shift. At first glance, one may treat OOD examples with covariate shift (denoted as C-OOD) the same as OOD examples with semantic shift (denoted as S-OOD) — *i.e.*, to detect and reject as many of them as possible. However, unlike S-OOD examples, which the deployed classifier can never classify correctly, C-OOD examples have a chance to be correctly classified, as their label space is covered by the classifier. For instance, if the covariate shift is small (*e.g.*, ImageNet [34] as ID; ImageNetV2 as C-OOD [32]), many of the C-OOD examples will likely be correctly classified. Even if the covariate shift is more pronounced (*e.g.*, ImageNet-Sketch [42]), a robust classifier (*e.g.*, with the CLIP-pre-trained backbone [30]) is likely to still classify a decent amount of C-OOD examples correctly. Blindly rejecting these correctly classified C-OOD examples would adversely reduce the efficacy of the classifier.

Taking this insight into account, we propose a novel, unifying framework — **MS-OOD DETECTION** — to study OOD detection from a "**M**odel-**S**pecific" perspective. In MS-OOD DETECTION, *whether an example should be detected and rejected from being classified (denoted by a ground-truth label* $-1$*) depends on whether the deployed classifier would misclassify it.* With this definition, every test example can be *deterministically* assigned a ground-truth label based on the deployed classifier: $+1$ for correctly classified examples, which should not be rejected; $-1$ for misclassified examples, which should be rejected. This enables us to study different causes of OOD examples in a unifying way. It is worth noting that while C-OOD examples could be assigned different ground-truth labels, all the S-OOD examples are assigned ground-truth labels $-1$. In other words, similar to conventional OOD detection, all the S-OOD examples should be detected and rejected in MS-OOD DETECTION.

Nevertheless, unlike conventional OOD detection, which aims to accept all the ID examples drawn from the training distribution, MS-OOD DETECTION, according to how it assigns ground-truth labels, aims to detect and reject misclassified ID examples as well. That is, what MS-OOD DETECTION aims to accept are correctly classified ID and C-OOD examples; what MS-OOD DETECTION aims to reject are misclassified ID and C-OOD examples, and all the S-OOD examples (which are always misclassified). Such a definition seamlessly unifies and generalizes the two related problems studied in the seminal work by Hendrycks and Gimpel [13] — detecting misclassified and OOD examples — and could better reflect real-world application scenarios. For instance, for end-users who seek to reliably apply the machine learning model in uncontrolled environments, misclassification reveals the limitation of the model or the inherent difficulty of the examples (*e.g.*, hard or ambiguous examples), implying the need for end-user intervention.

We conduct an extensive empirical study of MS-OOD DETECTION. 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 approaches such as Maximum Softmax Probabilities (MSP) [13], Energy Score [23], Maximum Logit Score (MLS) [41], Virtual-logit Matching (ViM) [43], and GradNorm [18]. New datasets, classifiers, and OOD methods can easily be incorporated to broaden 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 are 1) a list of novel insights into OOD detection and 2) a unifying re-validation of several existing but seemingly isolated insights found in different contexts. For instance, we find that the best detection methods for S-OOD, misclassified C-OOD, and misclassified ID data are not consistent; their effectiveness could be influenced by the paired model, hence "model-specific". Specifically for C-OOD examples, we find that the more robust the classifier is (*i.e.*, having a higher accuracy in classifying C-OOD examples), the easier the misclassified C-OOD examples can be detected. For S-OOD examples, while in general, we see the same trend as in [41] — the stronger the classifier is (*i.e.*, having a higher accuracy in classifying ID examples), the easier the S-OOD examples can be detected — there are exceptions when we apply particular detection methods and classifiers. This suggests the need for a more thorough study. For misclassified ID examples, we find that they normally have lower scores (*e.g.*, softmax probabilities) than correctly classified ID examples. In other words, one could set a higher threshold to reject more S-OOD examples without sacrificing the true positive rate of accepting ID examples that are correctly classified. Last but not least, we find that the baseline OOD detection method MSP [13] performs favorably in detecting misclassified ID and C-OOD examples, often outperforming other more advanced methods.

**Contributions.** Our contributions are two-folded:

- We propose a novel framework, MS-OOD DETECTION, which enables us to study different OOD examples (*e.g.*, covariate shift and semantic shift) in a unifying way.
- We conduct an extensive study of MS-OOD DETECTION, which reveals novel insights into OOD detection and unifies existing insights gained from different contexts.

## 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 [50]. The main differences lie in 1) the scope of OOD examples; 2) whether one has to perform multi-class classification on ID examples. Specifically, novelty detection and open-set recognition mainly consider OOD examples with semantic shift.

In conventional OOD detection, the focus is on detecting OOD examples with semantic shift (*i.e.*, S-OOD), assuming the absence of covariate shift (*i.e.*, C-OOD) [50, 35]. Please see Figure 2 (a) for an illustration. Very few works include examples with covariate shift in their studies; most of them *treat these examples as ID*, aiming to classify them robustly instead of detecting them as OOD [49, 27, 51] (see Figure 2 (b), the "SEM framework" [51]). Some exceptions are [16, 39, 44], which aim to detect all examples with covariate shift as OOD (see Figure 2 (c), "G-ODIN framework" [16]).

In anomaly detection and outlier detection, the focus is on differentiating ID and OOD examples without the need to classify ID examples (*i.e.*, they treat ID examples as a single

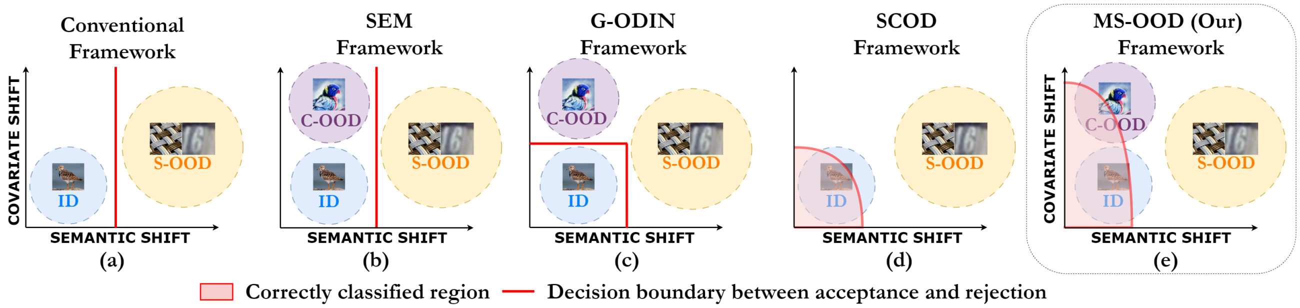


Figure 2: **Comparison of OOD detection frameworks**. The blue, purple, and yellow regions denote in-distribution (ID), covariate shift (C-OOD), and semantic shift (S-OOD) test data, respectively. The red boundary separates the test data into **acceptance (A)** and **rejection (R)** cases; the goal of OOD detection is to detect the R examples while accepting the A examples. Based on the test data involved and how they are separated, OOD detection frameworks can be categorized into (a) conventional OOD detection; (b) SEM [51]; (c) G-ODIN [16]; (d) SCOD [47]; (e) our MS-OOD DETECTION. It is worth noting that the red boundaries in (d) and (e) depend on the deployed multi-class classifiers trained for the ID classes — the boundaries separate data that are correctly and wrongly classified by the classifier.

class). Several works also consider examples with covariate shift [50]. Similar to the G-ODIN framework, these works aim to detect all examples with covariate shift as OOD.

We argue that whether an example with covariate shift should be detected or not depends on whether the deployed classifier would misclassify it or not. By taking a model-specific perspective, our MS-OOD DETECTION resolves the dilemma between *OOD detection* [50] and *OOD generalization* [36]: a robust model should *generalizes* to examples with covariate shift; a weak model should *reject* them.

**OOD generalization.** Covariate shift is commonly studied in model generalization and robustness [45, 36]. Instead of rejecting examples with covariate shift, the community aims to improve 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 covariate-shift examples [38], we argue that it is desirable to also consider covariate shift in OOD detection.

**Selective classification.** Equipping a model with the option to reject has also been studied in selective classification [9]. Different from OOD detection, selective classification focuses on rejecting uncertain "ID" examples. Recently, [47] proposed to integrate selective classification with OOD detection (SCOD), aiming to detect both misclassified ID and semantic shift data (see Figure 2 (d)). In this context, our work can be seen as a generalized version, further taking covariate shifts into account (see Figure 2 (e)). Compared to [47], we provide a more comprehensive study, further emphasizing the role of models in evaluation.

**OOD detection methods** have roughly two categories: post-hoc and training-based [35]. The difference is whether one is allowed to specifically train the model (*e.g.*, the classifier for ID data) to detect OOD examples. While training-based methods like outlier exposure [14] have shown a much higher detection rate, they need access to ID training data and OOD examples that one would encounter, making them prohibitive for end-users and less effective in uncontrolled settings. We therefore focus on post-hoc methods. The baseline is to use the softmax output as confidence [13]. Other approaches consider scaling the temperature and adding input perturbations [22]; using logits [41], energy [23], or gradients [18] as the score; combining intermediate features with logits [43].

## 3. Background

We study the problem of out-of-distribution (OOD) detection in the context of multi-class classification. Given a neural network classifier $f$ that is trained on data sampled from a training distribution $P(X, Y \in \mathcal{S})$, OOD detection aims to construct a scoring function $g(x, f)$ which gives in-distribution (ID) data a higher score; OOD data, a lower score. $\mathcal{S}$ here denotes the label space of the training data.

During test time, examples that produce $g(x, f) > \tau$ are *accepted* and 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 class label.

In real-world applications, sources of OOD examples can roughly be categorized into two groups: covariate shift and semantic shift. Using $P(X, Y)$ as in-distribution, these shifts occur either solely on the marginal distribution $P(X)$, or on both the class distribution $P(Y)$ and $P(X)$, respectively. In conventional experimental setups, semantic shift data consist of novel-class examples that are outside the classifier's label space $\mathcal{S}$; covariate shift data consist of examples from the classifier's label space $\mathcal{S}$ but from different domains.

## 4. Model-Specific Out-of-Distribution Detection

As introduced in section 1 and section 2, the focal point of existing OOD detection has been on semantic shift alone, assuming the absence of covariate shift during the test time. Very few works consider data with covariate shift in their studies, and there is a dilemma if one should detect those examples as OOD [16, 39, 44] or accept them as ID [49, 27, 51]. Figure 2 (a-c) illustrate these three frameworks.

In this paper, we argue that such a dilemma is more fundamental than previously believed. In essence, one ultimate goal of OOD detection is to detect test examples that the deployed classifier is unable to classify correctly and preclude them from being classified. To achieve such a goal, one must detect as many semantic shift examples (denoted as S-OOD) as possible, as they belong to novel classes and cannot be correctly classified by the deployed classifier. Covariate shift examples (denoted as C-OOD), in contrast, belong to the seen classes on which the deployed classifier was trained. Namely, some of the C-OOD examples are likely to be correctly classified; the amount of them depends on the degree of covariate shift and the robustness of the classifier. Taking this fact into account, we argue that the current dilemma in how to handle C-OOD examples can actually be resolved smoothly — *detecting and rejecting the misclassified C-OOD examples while accepting the correctly classified ones.*

## 4.1. Problem definition

In this subsection, we introduce MS-OOD DETECTION, a novel framework that realizes the aforementioned goal and unifies the detection of OOD examples from different causes. Figure 1 gives an illustration.

We consider a test environment that may contain ID, S-OOD, or C-OOD examples. Given a deployed classifier that was trained to classify the seen classes $\mathcal{S}$ in the ID data, we separate the test examples into two cases:

- **Model-specific acceptance (MS-A)**, which contains the ID and C-OOD examples that are correctly classified by the classifier, denoted by ID+ and C-OOD+, respectively. The red region in Figure 1 shows this case.
- **Model-specific rejection (MS-R)**, which contains the ID and C-OOD examples that are misclassified by the classifier as well as all the S-OOD examples, denoted by ID$-$, C-OOD$-$, and S-OOD, respectively. The complement of the red region in Figure 1 shows this case.

The goal of MS-OOD DETECTION is to differentiate MS-A examples from MS-R examples, accepting the former while rejecting the latter.

**Remark.** Building upon MS-OOD DETECTION, we can view the G-ODIN framework [16] and the SEM framework [51] as special cases under two model-specific assumptions. (Please see Figure 2 (c) and (b), respectively, for their illustrations.) The G-ODIN framework aims to reject all the C-OOD examples while accepting all the ID examples, equivalent to assuming that the deployed classifier performs perfectly on the ID data but cannot classify any C-OOD data correctly. The SEM framework aims to accept all the ID and C-OOD examples, equivalent to assuming that the deployed classifier can classify all these examples correctly. In essence, constructing such a classifier is the goal of model generalization and robustness, which is to maximize the red region in Figure 1 to cover all the ID and C-OOD examples.

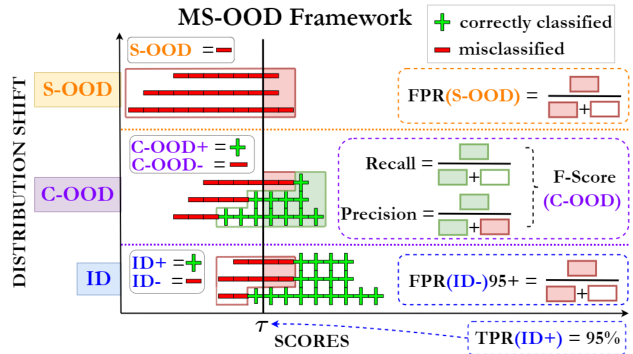


Figure 3: **Evaluation metrics for our MS-OOD DETECTION.** $+$ and $-$: correctly classified (MS-A) and misclassified (MS-R) examples by the deployed classifier $f$; X-axis: the $g(x, f)$ score for MS-OOD DETECTION; $\tau$: the threshold at TPR(ID+)= 95%. For MS-A examples ($+$) that should be accepted, *i.e.*, $g(x, f) > \tau$, the portion of them to the right of the threshold indicates the TPR. For MS-R examples ($-$) that should be rejected, *i.e.*, $g(x, f) \leq \tau$, such a portion indicates the FPR. For C-OOD examples, we further depict the Precision and Recall calculated at the threshold.

## 4.2. Ground-truth label and detection mechanism

We provide a more formal definition using notations introduced in section 3. Let $f : \mathcal{X} \mapsto \mathcal{S}$ be the classifier, where $\mathcal{S}$ is the model's label space. Given a labeled test example $(x \in \mathcal{X}, y \in \mathcal{S} \cup \mathcal{U})$, where $\mathcal{U}$ denotes the novel-class space, we assigns it a ground-truth MS-OOD DETECTION label $z$:

- $z = +1$ **for MS-A**, if $f(x) = y$;
- $z = -1$ **for MS-R**, if $f(x) \neq y$.

MS-OOD DETECTION then aims to construct a scoring function $g(x, f)$ to predict $z$:

- $\hat{z} = +1$, if $g(x, f) > \tau$;
- $\hat{z} = -1$, if $g(x, f) \leq \tau$,

where $\tau$ is the threshold (hyper-parameter) to be selected.

In this paper, instead of developing new scoring functions, we conduct an extensive study using existing OOD detection methods. More details are provided in section 5.

## 4.3. Evaluation metric

To evaluate MS-OOD DETECTION in detail, we separately report the result on each ingredient of the MS-A (*i.e.*, ID+, C-OOD+) and MS-R (*i.e.*, ID$-$, C-OOD$-$, S-OOD) cases. Specifically, for MS-A examples that should be accepted by the scoring function (*i.e.*, $g(x, f) > \tau$; $\hat{z} = +1$), we report the **True Positive Rate (TPR)**: the probability that examples with ground-truth labels $z = +1$ are predicted as $\hat{z} = +1$. For MS-R examples that should be rejected by the scoring function (*i.e.*, $g(x, f) \leq \tau$; $\hat{z} = -1$), we report the **False Positive Rate (FPR)**: the probability that examples with ground-truth labels $z = -1$ are predicted as $\hat{z} = +1$. We use $(\cdot)$ following these metrics to indicate the ingredient. For instance, **TPR(ID+)** is the TPR of ID+ examples; **FPR(S-OOD)** is the FPR of S-OOD examples.

As these metrics depend on the threshold $\tau$, we follow existing works to use one of them as the reference. Throughout the paper, we select $\tau$ such that it leads to **TPR(ID+)**$= 95\%$ unless stated otherwise. **FPR(S-OOD)@TPR(ID+)=95** thus means the **FPR(S-OOD)** at such a $\tau$ value.

We now list the metrics used in the main paper; Figure 3 gives an explanation. We include other metrics in the Suppl.

- **FPR(S-OOD)@TPR(ID+)=95**: FPR for accepting S-OOD data, at TPR$= 95\%$ for accepting ID+ data.
- **FPR(ID−)@TPR(ID+)=95**: FPR for accepting ID− data, at TPR$= 95\%$ for accepting ID+ data. This is the same metric used in [47] for selective classification.
- **$F_1$-Score(C-OOD)@TPR(ID+)=95**: $F_1$-Score for identifying C-OOD+ data from C-OOD data, at TPR$= 95\%$ for accepting ID+ data.

We report $F_1$-Score as it simultaneously quantifies how well the C-OOD+ data are accepted and C-OOD− data are rejected. We calculate Precision (P) and Recall (R) for identifying C-OOD+ data from C-OOD data, using $\tau$ that is selected to obtain TPR(ID+)$= 95\%$. The $F_1$-Score is

$$\frac{2 \times P \times R}{P + R}. \tag{1}$$

**Remark.** **FPR(S-OOD)@TPR(ID+)=95**, the metric we use to evaluate the detection of S-OOD examples, is very much the same as the one used in conventional OOD detection. The only difference is the threshold $\tau$: conventional OOD detection aims to accept all the ID examples and uses TPR(ID)=95 as the reference.

### 4.4. Comparison to SCOD [47]

Despite the superficial similarity to the SCOD framework [47] shown in Figure 2 (d), our MS-OOD DETECTION shown in Figure 2 (e) has several unique significance and contributions. First, as described in section 1, MS-OOD DETECTION seeks a unifying framework to study both S-OOD and C-OOD examples, while SCOD combines selective classification with S-OOD detection. Second, we argue that the inclusion of C-OOD data is by no means trivial. As discussed in [50], existing OOD works do not agree on how to tackle C-OOD examples. Our "model-specific" perspective seamlessly glues the two extremes in the current dilemma (*i.e*., accepting or rejecting all C-OOD examples). Third, as depicted in Figure 2, MS-OOD DETECTION covers all the existing OOD detection frameworks. Some of them seem irrelevant or contrasting to each other at first glance, and MS-OOD DETECTION unifies them: each of them can be viewed as a special case of MS-OOD DETECTION under a certain assumption of the deployed model or the test environment (see subsection 4.1). We consider this important as it provides a platform to connect different frameworks.

## 5. Experimental Setup

Besides proposing MS-OOD DETECTION, the other key contribution of the paper is a comprehensive empirical study and benchmarking. Specifically, we investigate three dimensions that could fundamentally influence the performance of MS-OOD DETECTION: 1) sources of OOD examples, 2) deployed classifiers, and 3) OOD detection methods.

### 5.1. Sources of ID/OOD examples

**In-distribution (ID).** We use the ImageNet-1K dataset [4] as the ID data. It is the standard benchmark for image classification and has been widely used as ID data in OOD detection. Many works in neural network architectures and training strategies also release their pre-trained classifiers.
**Semantic-shift (S-OOD).** We use the common benchmark datasets: SVHN [28], Texture (DTD) [3], Places365 [53], iNaturalist [40] and SUN [48]. We use the filtered versions proposed in [19] to ensure that classes in these datasets do not overlap with classes in ImageNet-1K. We also use the ImageNet-O dataset proposed in [15], which contains adversarially chosen classes to ImageNet-1K. Following [41], we also consider the "Easy" and "Hard" subsets of ImageNet-21K [33, 4] proposed in the semantic-shift benchmark (SSB) for open-set recognition. These S-OOD data extracted from ImageNet-21K share similar styles with the ID data, more faithfully featuring semantic shift.
**Covariate-shift (C-OOD)**. We consider four datasets with different degrees and types of covariate shift; their label spaces are covered by ImageNet-1K. ImageNetV2 [32] is collected via a similar criterion as ImageNet-1K but after a decade, featuring a natural covariate shift by time. For images with different styles and domains, we use ImageNet-R [12] and ImageNet-S [42] ("S" stands for "Sketch"). We also use ImageNet-A [15], a natural adversarial dataset curated by collecting wrongly predicted examples with high confidence. Since ImageNet-R and ImageNet-A contain only a subset of 200 classes from ImageNet-1K, when experimenting with them, we use the same subset of classes from ImageNet-1K as the ID data for a fair comparison.

### 5.2. Neural network models

Since MS-OOD DETECTION is model-specific, we consider different neural network models for ImageNet-1K classification. We treat ResNet50 [11] trained with the standard practice as the reference and consider factors that could influence 1) the classifier's strength in classifying the ID data and 2) its robustness in classifying the C-OOD data. Specifically, we consider **network depths** (*e.g*., ResNet18, ResNet152), **training strategies** (*e.g*., robust ResNet50 [29] trained with the TorchVision new recipe[1]), **pre-training**

[1]This involves data augmentation, label smoothing, longer training, etc. The resulting model has a higher ID and C-OOD classification accuracy.

(*e.g.*, CLIP-ResNet50 [31]), and **network architectures** (*e.g.*, ViT-B-16 [6]). We obtain all these models from PyTorch [29], except for CLIP-ResNet50 which is released on its GitHub [31]. We use the officially released ImageNet-1K version of CLIP, which leverages the zero-shot capability of CLIP without fine-tuning it on ImageNet-1K data: the final fully-connected layer is constructed by prompting with modified ImageNet class names. We note that despite employing a different training strategy, the robustness of CLIP largely comes from its diverse training data [7]. Hence, we put the model under the umbrella of pre-training.

**Evaluation metric.** Besides the metrics introduced in subsection 4.3 for MS-OOD DETECTION, we also report the multi-class classification accuracy of these models on the ID and C-OOD data to quantify their strength and robustness.

### 5.3. Detection methods

We focus on post-hoc methods, assuming that the classifier is pre-trained and fixed. We argue that this better reflects the scenario of how end-users obtain and apply machine learning models. We consider five representative methods categorized into output-based, feature-based, and hybrid. *We use them to detect S-OOD, misclassified C-OOD, and misclassified ID examples.* Details are in the Suppl.

**Output-based** methods. **Maximum Softmax Probabilities (MSP)** [13] sets the scoring function $g(x, f)$ as the largest softmax probability outputted by the classifier $f$. MSP is treated as the baseline in most OOD detection literature. Despite its simplicity, we provide reasons why this algorithm is worth exploring in MS-OOD DETECTION: [41, 8] showed the superior performance when MSP is paired with a strong classifier; [47] showed the best performance compared to other state-of-the-art methods on rejecting misclassified ID examples; [26] showed relatively good performance when MSP is paired with CLIP. Besides MSP, we also consider **Maximum Logit Score (MLS)** [41] and **Energy Score** [23] as $g(x, f)$: MLS uses the logit value before softmax; Energy uses the denominator of the softmax calculation.

**Feature-based** methods. We employ **GradNorm** [18] which relies on gradients and the penultimate layer.

**Hybrid** methods. **Virtual-logit Matching (ViM)** [43] uses residual features and logits to produce the $g(x, f)$ score. Its inner workings are quite similar to Mahalanobis [21] but ViM has a better performance. It is worth noting that ViM requires access to the training data to calculate the score.

**Remark.** We provide reasons why we exclude other post-hoc methods: ReAct [37] has an extra hyperparameter that can impact the accuracy of the classifier, changing the *acceptance region* in Figure 1 and leading to an unfair comparison. ODIN [22] needs hyperparameter tuning for input perturbations (hence the knowledge of OOD data). We include ODIN without tuning in Suppl.

We also provide reasons why we apply the scores developed for S-OOD detection to detect misclassified C-OOD and ID examples. First, C-OOD examples can essentially be seen as S-OOD examples if one treats "style/domain" + "class" as the semantic label. Second, scores like MSP are commonly used in domain adaptation [54, 2, 52, 20] to detect unconfident predictions on out-of-domain data. Third, as studied in [13, 47], scores like MSP are quite useful for identifying misclassified ID data.

Table 1: The accuracy (ACC) to classify C-OOD data and the false positive rate (FPR) to reject C-OOD data. For the former, we apply ResNet50 and CLIP-ResNet50. For the latter, MSP is used; $\tau$ is chosen for TPR(ID)=95. IN: ImageNet (see subsection 5.1).

| MODEL | IN-V2 | | IN-S | | IN-R | | IN-A | |
|---|---|---|---|---|---|---|---|---|
| | ACC↑ | FPR↓ (C-OOD) | ACC↑ | FPR↓ (C-OOD) | ACC↑ | FPR↓ (C-OOD) | ACC↑ | FPR↓ (C-OOD) |
| RN50 | 72.4 | 93.9 | 24.1 | 65.7 | 36.2 | 71.6 | 0.0 | 81.5 |
| CLIP-RN50 | 59.5 | 95.4 | 35.5 | 78.4 | 60.6 | 92.3 | 22.8 | 89.5 |

## 6. Experimental Results

We present the main results. *For clarity, we mainly use figures (*e.g.*, scatter plots) and leave detailed tables in Suppl.*

### 6.1. Model-agnostic C-OOD detection is "ill-posed"

We first justify the necessity to introduce MS-OOD DETECTION. We investigate the two extremes: accepting or rejecting all the C-OOD examples. We consider two classifiers: baseline ResNet50 and CLIP-ResNet50. The CLIP model has been shown surprisingly robust to C-OOD examples [7, 25]. Table 1 summarizes the results, in which we report the Accuracy (ACC) of classifying C-OOD examples, and the False Positive Rate (FPR) if we apply MSP to reject all C-OOD examples. For simplicity, we select $\tau$ at TPR(ID)=95 without separating ID data into ID+ and ID−.

On the three C-OOD datasets (ImageNet-R, ImageNet-S, and ImageNet-A) that have notable covariate shift from ImageNet-1K, CLIP-ResNet50 achieves higher ACC than ResNet50, demonstrating its robustness. With that being said, the ACC is still far from 100%, implying the risk to accept all C-OOD examples. At the other extreme, except for baseline ResNet50 on the challenging ImageNet-A dataset, we see non-zero ACC for the other combinations of models and datasets, implying a waste if we reject all C-OOD examples. Perhaps more interestingly, for CLIP-ResNet50 which achieves higher ACC on C-OOD data, rejecting C-OOD data becomes harder, as evidenced by the much higher FPR(C-OOD) values. These results suggest the necessity to take a model-specific perspective in handling C-OOD data: accepting the correctly classified ones (*i.e.*, C-OOD+) while rejecting the misclassified ones (*i.e.*, C-OOD−).

### 6.2. ID examples in MS-OOD DETECTION

We now separately evaluate MS-OOD DETECTION on different test data. We start with ID examples. We consider the four factors of models and the five detection methods

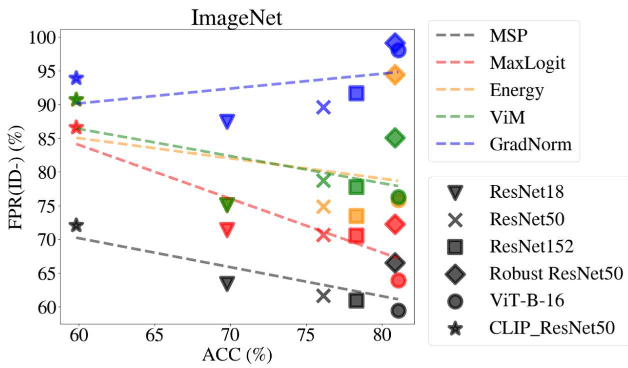


Figure 4: **ID in MS-OOD Detection.** X-axis: ACC of classifying the ID examples; Y-axis: FPR(ID$-$)@TPR(ID+)=95 for wrongly accepting ID$-$ examples. We denote different neural network models by marker shapes; different detection methods by colors. The dashed line is the trend for each detection method.

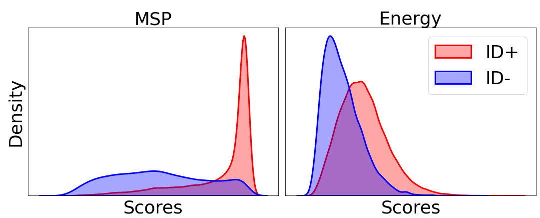


Figure 5: **Histogram of ID+ and ID$-$ data at different $g(x, f)$.** We use ResNet50 as the model $f$, and MSP and Energy for $g(x, f)$.

introduced in section 5. The goal is to differentiate correctly classified and misclassified ID (*i.e.*, ImageNet-1K validation) examples. We report **FPR(ID$-$)@TPR(ID+)=95**.

Figure 4 summarizes the results, from which we have three key observations. First, generally speaking, the higher the ACC is in classifying ID data, the lower the FPR is in wrongly accepting ID$-$ examples. In other words, the stronger the classification model is, the easier it is to differentiate the correctly classified ID+ data from the remaining ID$-$ data. Second, we find two exceptions. In particular, robust ResNet [29] (♦) has a higher ACC than many other models but also a higher FPR (hence poor ID$-$ detection). Such an opposite trend also shows up when we apply GradNorm for ID$-$ detection: using a stronger model leads to a higher FPR. These exceptions suggest the need for a deeper look at 1) what model training strategies might hurt ID$-$ detection[2], and 2) what underlying mechanisms of detection methods could not benefit from a stronger classifier. Specifically, we hypothesize that some of the training tricks may aggravate the overconfidence phenomenon of neural network predictions [10]. Third, no matter which model is used, **MSP** outperforms other detection methods, achieving the lowest FPR. This aligns with what was reported in [47].

We further show the histogram of the ID+ and ID$-$ examples at different $g(x, f)$ values in Figure 5. ID+ data usually have higher $g(x, f)$ than ID$-$ data. In other words,

[2] [41] is trained with data augmentation, label smoothing, etc.

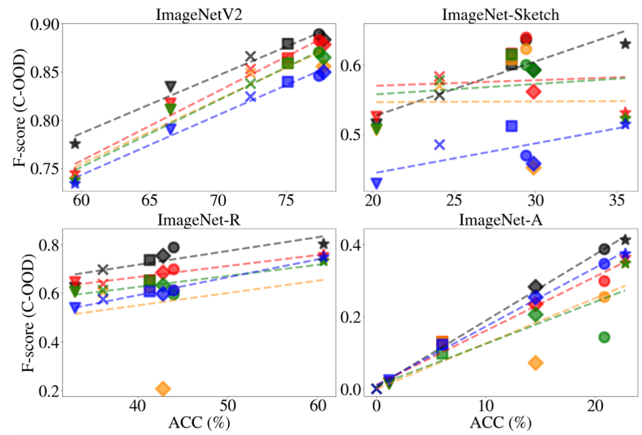


Figure 6: **C-OOD in MS-OOD Detection.** Each sub-figure corresponds to each C-OOD dataset. X-axis: ACC of classifying the C-OOD examples; Y-axis: $F_1$-Score(C-OOD)@TPR(ID+)=95 for identifying C-OOD+ examples from C-OOD data. (The higher, the better.) Other denotations follow Figure 4.

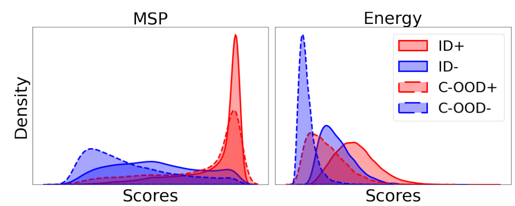


Figure 7: **Histogram of ID+, ID$-$, C-OOD+, and C-OOD$-$ data at different $g(x, f)$.** We use the ImageNet-R as C-OOD. We use ResNet50 as the model $f$, and MSP and Energy for $g(x, f)$.

the $\tau$ value at TPR(ID+)=95 should be higher than the $\tau$ value at TPR(ID)=95. This implies that one could pick a higher threshold $\tau$ than what is normally set, to reject more S-OOD and C-OOD$-$ examples without sacrificing the TPR of accepting ID examples that are correctly classified.

### 6.3. C-OOD examples in MS-OOD Detection

We now move on to C-OOD examples. We consider the four C-OOD datasets introduced in section 5. We report **$F_1$-Score(C-OOD)@TPR(ID+)=95** defined in subsection 4.3.

Figure 6 summarizes the results, in which we draw a scatter plot for each C-OOD dataset. Different detection methods are marked by colors; different models are marked by shapes. The dashed line shows the trend for each detection method.

We have two key observations. First, in general, the higher the ACC is in classifying C-OOD data, the high the $F_1$-Score is in identifying C-OOD+ examples from C-OOD data. This can be seen across detection methods and across datasets. We view this a critical insight into bridging the dilemma of C-OOD detection and generalization — *improving the model's generalizability to C-OOD data would simultaneously benefit its ability to differentiate C-OOD+ data from C-OOD$-$ data.* Second, like rejecting ID$-$ data, MSP generally performs the best (*i.e.*, obtaining the highest $F_1$-Score) in differentiating C-OOD+ and C-OOD$-$ data.

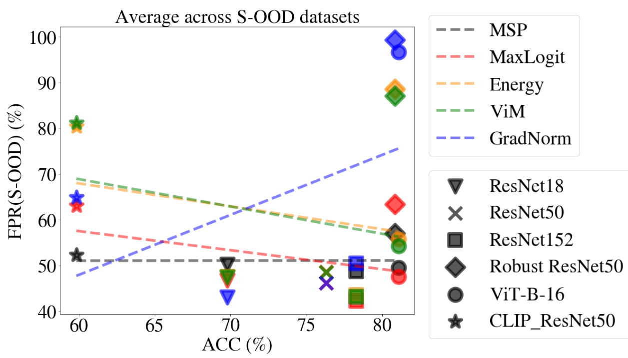


Figure 8: **S-OOD in MS-OOD DETECTION.** The results are averaged over eight S-OOD datasets. X-axis: ACC of classifying the ID examples; Y-axis: FPR(S-OOD)@TPR(ID+)=95 for wrongly accepting S-OOD examples.

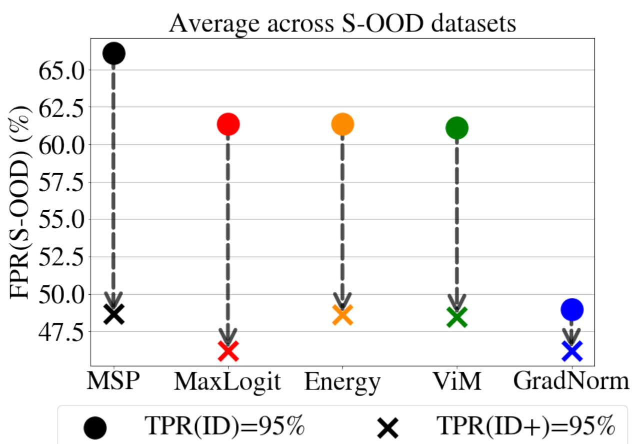


Figure 9: **Comparison of FPR(S-OOD) at TPR(ID+)=95 (×) and TPR(ID)=95 (○).** The main difference is the threshold $\tau$. We use the ResNet50 model; the results are averaged over eight S-OOD datasets. All detection methods improve at TPR(ID+)=95.

We further show the histogram of the C-OOD+ and C-OOD− examples, together with the ID+ and ID− examples, at different $g(x, f)$ values in Figure 7. Overall, we find that it is typically simple to differentiate ID+ and C-OOD− data. The challenge resides in how to differentiate C-OOD+ and ID−: accepting the former while rejecting the latter. We find that MSP could much better separate these two cases.

### 6.4. S-OOD examples in MS-OOD DETECTION

We now move on to S-OOD examples. We consider the eight S-OOD datasets introduced in section 5. We report **FPR(S-OOD)@TPR(ID+)=95**. Figure 8 summarizes the results, in which we average the results over all datasets; the detailed results of individual datasets can be found in Suppl. Other setups and denotations exactly follow Figure 4.

We have three observations. First, in general, the higher the ACC is in classifying ID data, the lower the FPR is in wrongly accepting S-OOD data. This is aligned with what was reported in [41], suggesting that one could improve S-OOD detection by improving the deployed model. Second, we find similar exceptions as in subsection 6.2. When we apply GradNorm for S-OOD detection, using stronger models (hence higher ACC on ID data) leads to higher FPR on S-OOD detection; when we apply robust ResNet [29] (⧫) as the model, it has higher ACC on ID data but higher FPR on S-OOD detection. Similarly to subsection 6.2, we suggest a deeper look at these opposite trends. Finally, comparing different detection methods, we see no definite winner. For instance, the baseline MSP seems to fall behind in the high ACC regime, but it achieves the lowest FPR when paired with CLIP-ResNet50 ($\star$), as also pointed out in [26]. The multiple intersections among dashed lines further suggest the "model-specific" nature in picking the appropriate detection methods for S-OOD examples.

Table 2: The accuracy (ACC) to classify C-OOD data and the false positive rate (FPR) to reject C-OOD data. For the former, we apply CLIP-ResNet50, CLIP-ViT-B/16, CLIP-ViT-B/32, and CLIP-ViT-L/14. For the latter, MSP is used; $\tau$ is chosen for TPR(ID)=95. IN: ImageNet.

| MODEL | IN-V2 | | IN-S | | IN-R | | IN-A | |
|---|---|---|---|---|---|---|---|---|
| | ACC↑ | FPR↓ (C-OOD) | ACC↑ | FPR↓ (C-OOD) | ACC↑ | FPR↓ (C-OOD) | ACC↑ | FPR↓ (C-OOD) |
| CLIP-RN50 | 59.5 | 95.4 | 35.5 | 78.4 | 60.6 | 92.3 | 22.8 | 89.5 |
| CLIP-ViT-B/16 | 68.8 | 94.8 | 48.2 | 83.7 | 77.6 | 94.1 | 50.1 | 89.7 |
| CLIP-ViT-L/14 | 75.8 | 94.6 | 59.6 | 87.4 | 87.7 | 96.1 | 70.7 | 91.2 |

To draw a comparison to conventional S-OOD detection, we report its standard metric **FPR(S-OOD)@TPR(ID)=95**. This metric features accepting all ID examples even if some are misclassified. Overall, we have quite similar findings to those mentioned above (please see Suppl.). Further comparing the two metrics in Figure 9, we see a consistent improvement in S-OOD detection using our new metric: FPR drops for all detection methods. In essence, since ID+ data normally have higher $g(x, f)$ scores than ID− data (cf. Figure 5), replacing TPR(ID)=95 with TPR(ID+)=95 allows us to use a larger $\tau$ to reject S-OOD data. While this comes with the cost of rejecting more ID data, most of them are indeed ID− and do not hurt to be rejected. Perhaps more interesting, among all detection methods, MSP improves the most when setting the threshold by TPR(ID+)=95, arriving at the same level of FPR as other advanced methods.

### 6.5. Additional experiments on CLIP [31]

CLIP [31] has shown superior robustness in many image classification tasks [7]. We, therefore, further analyze whether our results generalize to CLIP with different backbone architectures. For this experiment, we include other pre-trained backbones such as ViT-B/16, ViT-B/32, and ViT-L/14 [6]. First, following subsection 6.1, we investigate C-OOD detection in the extreme cases of accepting and rejecting all C-OOD examples. We see a similar trend in Table 2: the higher the accuracy is in classifying C-OOD examples (*i.e.*, the more robust the model is), the harder it is to reject these examples. These results suggest the necessity to

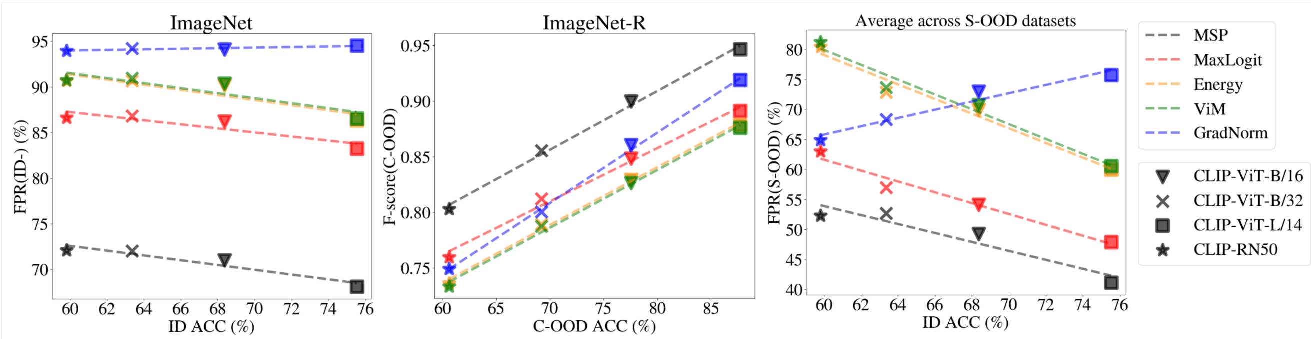

Figure 10: **Results using different CLIP backbones for ID−, C-OOD, and S-OOD in MS-OOD DETECTION**. We use ImageNet-R for C-OOD and averaged over S-OOD datasets for S-OOD. Denotations follow Figure 4, Figure 6, and Figure 8

take a model-specific perspective in handling C-OOD data.

Second, we investigate the performance in ID−, C-OOD (ImageNet-R), and S-OOD examples. Figure 10 summarizes the results: we observe similar trends across all of them as in subsection 6.2, subsection 6.3, and subsection 6.4. These results suggest that our results apply to CLIP with different architectures.

### 6.6. More results in Suppl.

Given the page limit, we leave additional results in Suppl., including the full tables that generate figures, results based on different metrics, and qualitative visualizations.

## 7. Discussion

### 7.1. MSP performs well in ID− & C-OOD detection

Jointly comparing the results in subsection 6.2, subsection 6.3, and subsection 6.4, we find no winning detection method across all tasks. Nevertheless, the baseline MSP [13] performs quite well in detecting the ID− examples and differentiating the C-OOD+ and C-OOD− examples, outperforming other more advanced methods. High-levelly, it makes sense as MSP is optimized to avoid misclassification and has been widely used to detect unconfident data in domain adaptation. We provide more analyses in Suppl. For example, Figure 16 and Figure 17 show the score distributions: except for MSP, existing methods conflate misclassified data with correctly classified data on both ID and C-OOD. Specifically on C-OOD, Table 6 shows that MSP balances well between accepting correctly classified samples and rejecting misclassified examples.

### 7.2. Improving performance in MS-OOD DETECTION

We discuss several directions. First, building on the results in subsection 6.2, subsection 6.3, and subsection 6.4, we suggest that one could improve the deployed classification model in terms of its accuracy and generalizability, as these mostly have positive correlations with the ID−, C-OOD, and S-OOD detection performance. Second, based on subsection 7.1, we hypothesize that properly combining MSP and other OOD detection methods (*e.g.*, through ensemble or cascade) or developing a new method that integrates the advantages of both would excel in MS-OOD DETECTION.

## 8. Conclusion

Out-of-distribution (OOD) detection has emerged as a promising topic for reliable machine learning. It has great potential to serve as a plug-and-play module to deployed machine learning models, helping them detect and reject potential misprediction. Nevertheless, looking at the literature, we found the topic seemingly overly fragmented into multiple scenarios with isolatedly developed algorithms [50, 35], despite that many of these algorithms share some common underlying principles (*e.g.*, using logits or softmax probabilities). This could hinder the development of the topic and make it hard for end-users to apply the algorithms. In this paper, we make an attempt to unify the topic. We propose MS-OOD DETECTION, a framework that aims to accept correctly predicted examples while rejecting wrongly predicted examples. We conduct an extensive study, considering in-distribution, semantic-shift, and covariate-shift test examples and different deployed models. We apply representative OOD detection methods and benchmark their performance in detail. Our study reveals new insights into OOD detection and unifies the existing ones.

Overall, our main contribution is a unifying framework to study OOD detection that involves semantic and covariate shifts. The novelty is in the framework, understanding, and insights. We deliberately do not propose a new OOD detection method but dedicate our paper to comprehensively evaluating existing methods, especially the post-hoc ones that are widely applicable. We hope that our paper serves as a useful manual for OOD detection in practice.

## Acknowledgments

This research is supported in part by NSF (IIS-2107077, OAC-2118240, and OAC-2112606) and Cisco Research. We are thankful for the computational resources of the Ohio Supercomputer Center.

## References


[1] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2021. 12

[2] Hong-You Chen and Wei-Lun Chao. Gradual domain adaptation without indexed intermediate domains. *Advances in Neural Information Processing Systems*, 34:8201–8214, 2021. 6

[3] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5, 13

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 5, 13

[5] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. *Advances in Neural Information Processing Systems*, 31, 2018. 12

[6] 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. 1, 6, 8

[7] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, pages 6216–6234. PMLR, 2022. 6, 8

[8] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *CoRR*, abs/2106.03004, 2021. 6

[9] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017. 3

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 7

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 5

[12] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 5, 13

[13] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 1, 2, 3, 6, 9, 14

[14] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 3

[15] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 5, 13, 14, 19

[16] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960, 2020. 2, 3, 4

[17] 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. 1

[18] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021. 2, 3, 6, 15

[19] 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*, pages 8710–8719, 2021. 5, 13, 14

[20] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pages 5468–5479. PMLR, 2020. 6

[21] 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. 6

[22] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 1, 3, 6, 21

[23] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020. 2, 3, 6, 14

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[25] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021. 6

[26] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *arXiv preprint arXiv:2211.13445*, 2022. 6, 8

[27] Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. *CoRR*, abs/2109.05642, 2021. 2, 3

[28] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5, 13

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5, 6, 7, 8

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 6, 8

[32] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 1, 5, 13

[33] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021. 5, 14

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1

[35] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*, 2021. 1, 2, 3, 9

[36] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021. 3

[37] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021. 6

[38] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020. 3

[39] Junjiao Tian, Yen-Change Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Exploring covariate and concept shift for detection and calibration of out-of-distribution data. *arXiv preprint arXiv:2110.15231*, 2021. 2, 3

[40] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 5, 13

[41] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. *CoRR*, abs/2110.06207, 2021. 2, 3, 5, 6, 7, 8, 14

[42] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 1, 5, 13

[43] 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*, pages 4921–4930, 2022. 2, 3, 6, 14

[44] Hongjun Wang, Sagar Vaze, and Kai Han. How and why we detect distribution shift: Critical analysis of methods and benchmarks, 2023. 2, 3

[45] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Taylan Cemgil. A fine-grained analysis on distribution shift. In *ICLR*, 2022. 3

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

[47] Guoxuan Xia and Christos-Savvas Bouganis. Augmenting softmax information for selective classification with out-of-distribution data. In *Proceedings of the Asian Conference on Computer Vision*, pages 1995–2012, 2022. 3, 5, 6, 7

[48] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 5, 14

[49] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. *CoRR*, abs/2108.11941, 2021. 2, 3

[50] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *CoRR*, abs/2110.11334, 2021. 1, 2, 3, 5, 9, 12

[51] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection, 2022. 2, 3, 4

[52] Luyu Yang, Yan Wang, Mingfei Gao, Abhinav Shrivastava, Kilian Q Weinberger, Wei-Lun Chao, and Ser-Nam Lim. Deep co-training with task decomposition for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8906–8916, 2021. 6

[53] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014. 5, 13

[54] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 6