Detecting Semantic Anomalies

Faruk Ahmed,[1] **Aaron Courville**[1,2]
[1]Mila, Université de Montréal
[2]CIFAR Fellow
{faruk.ahmed, aaron.courville}@umontreal.ca

Abstract

We critically appraise the recent interest in out-of-distribution (OOD) detection and question the practical relevance of existing benchmarks. While the currently prevalent trend is to consider different datasets as OOD, we argue that out-distributions of practical interest are ones where the distinction is semantic in nature for a specified context, and that evaluative tasks should reflect this more closely. Assuming a context of object recognition, we recommend a set of benchmarks, motivated by practical applications. We make progress on these benchmarks by exploring a multi-task learning based approach, showing that auxiliary objectives for improved semantic awareness result in improved semantic anomaly detection, with accompanying generalization benefits.

1 Introduction

In recent years, concerns have been raised about modern neural network based classification systems providing incorrect predictions with high confidence (Guo et al., 2017). A possibly-related finding is that classification-trained CNNs find it much easier to "overfit" to low-level properties such as texture (Geirhos et al., 2019), canonical pose (Alcorn et al., 2019), or contextual cues (Beery, Horn, and Perona, 2018) rather than learning globally coherent characteristics of objects. A subsequent worry is that such classifiers, trained on data sampled from a particular distribution, are likely to be misleading when encountering novel situations in deployment. For example, silent failure might occur due to equally confident categorization of unknown objects into known categories (due to shared texture, for example). This last concern is one of the primary motivating reasons for wanting to be able to detect when test data comes from a different distribution than that of the training data. This problem has been recently dubbed *out-of-distribution (OOD) detection* (Amodei et al.; Hendrycks and Gimpel, 2016; 2017), but is also referred to as anomaly/novelty/outlier detection in the contemporary machine learning context. Evaluation is typically carried out with benchmarks of the style proposed in Hendrycks and Gimpel (2017), where different datasets are treated as OOD after training on a particular in-distribution dataset. This area of research has been steadily developing, with some additions of new OOD datasets to the evaluation setup (Liang, Li, and Srikant, 2018), and improved results.

Current benchmarks are ill-motivated Despite such tasks rapidly becoming the standard benchmark for OOD detection in the community, we suggest that, taken as a whole, they are not very well-motivated. For example, the object recognition dataset CIFAR-10 (consisting of images of objects placed in the foreground), is typically trained and tested against noise, or different datasets such as downsampled LSUN (a dataset of scenes), or SVHN (a dataset of house numbers), or TINY-IMAGENET (a different dataset of objects). For the simpler cases of noise, or datasets with scenes or numbers, low-level image statistics are sufficient to tell them apart. While choices like TINY-IMAGENET might seem more reasonable, it has been noted that particular datasets have particular biases related to specific data collection and curation quirks (Torralba and Efros; Tommasi et al., 2011; 2017), which renders the problem of treating different datasets for OOD detection questionable. It is possible we are only getting better at distinguishing such idiosyncrasies. As an empirical illustration, we show in Appendix C that very trivial baselines can perform reasonably well at existing benchmarks.

***Semantic* distributional shift is relevant** We call into question the practical relevance of these evaluative tasks which are currently treated as standard by the community. While they might have some value as very preliminary reliability certification or as a testbed for diagnosing peculiar pathologies (for example, undesired behaviours of unsupervised density models, as in Nalisnick et al. (2019)), their significance as benchmarks for practical OOD detection is less clear. The implicit goal for the current style of benchmarks is that of detecting one or more of a wide variety of distributional shifts, which mostly consist of irrelevant factors when high-dimensional data has low-dimensional semantics. We argue that this is misguided; in a realistic setting, distributional shift across non-semantic factors (for example, camera and image-compression artefacts) is something we want to be robust to, while shift in semantic factors (for example, object identity) should be flagged down as anomalous or novel. Therefore, OOD detection is well-motivated only when the distributional shift is semantic in nature.

Context determines semantic factors In practical settings, OOD detection becomes meaningful only after acknowledging context, which identifies relevant semantic factors of interest. These are the factors of variation whose unnatural deviation are of concern to us in our assumed context. For example, in the context of scene classification, a kitchen with a bed in the middle is an anomalous observation. However, in the context of object recognition, the primary semantic factor is not the composition of scene-components anymore, but the identity of the foreground object. Now the unusual context should not prevent correct object recognition. If we claim that our object recognition models should be less certain of identifying an object in a novel context, it amounts to saying that we would prefer our models to be biased. In fact, we would like our models to systematically generalize (Fodor and Pylyshyn, 1988) in order to be trustworthy and useful. We would like them to form predictions from a globally coherent assimilation of the relevant semantic factors for the task, while being robust to their composition with non-semantic factors.

Without context, OOD detection is too broad to be meaningful The problem of OOD detection then, as currently treated by the community, suffers from imprecision due to context-free presumption and evaluation. Even though most works assume an underlying classification task, the benchmark OOD datasets include significant variation over non-semantic factors. OOD detection with density models are typically presented as being unaware of a downstream module, but we argue that such a context must be specified in order to determine what shifts are of concern to us, since we typically do not care about all possible variations. Being agnostic of context when discussing OOD detection leads to a corresponding lack of clarity about the implications of underlying methodologies in proposed approaches. The current benchmarks and methods therefore carry a risk of potential misalignment between evaluative performance and field performance in practical OOD detection problems. Henceforth, we shall refer to such realistic OOD detection problems, where the concerned distributional shift is a semantic variation for a specified context, by the term *anomaly detection*.

Contributions and overview Our contributions in this paper are summarized as follows.

1. *Semantic shifts are interesting, and benchmarks should reflect this more closely:* We provided a grounded discussion about the relevance of semanticity in the context of a task for realistic OOD (anomaly) detection. Under the view of regarding distributional shifts as being either semantic or non-semantic for a specified context, we concluded that semantic shifts are of practical interest. If we want to deploy reliable models in the real world, we typically wish to achieve robustness against non-semantic shift.

2. *More practical benchmarks for anomaly detection:* Although our discussion applies generally, in this paper we assume the common context of object recognition. In this context, unseen object categories may be considered anomalous at the "highest level" of semanticity. Anomalies corresponding to intermediate levels of semantic decomposition can also be relevant; for example, a liger should result in 50-50 uncertainties if the training data contains only lions and tigers. However, such anomalies are significantly harder to curate, requiring careful interventions at collection-time. Since detection of novel categories is a compelling anomaly detection task in itself, we recommend benchmarks that reflect such applications in section 2.

3. *Auxiliary objectives for improved semantic representation improves anomaly detection:* Following our discussion about the relevance of semanticity, in sections 4 and 5 we investigate the effectiveness of multi-task learning with auxiliary self-supervised objectives. These have been shown to result in semantic representations, measured through linear separability by object categories. Our experimental results are indicative that such augmented objectives lead to improved anomaly detection, with accompanying improvements in generalization.

2 Motivation and Proposed Tasks

In order to develop meaningful benchmarks, we begin by considering some practical applications where being able to detect anomalies, in the context of classification tasks, would find use.

Nature studies and monitoring: Biodiversity scientists want to keep track of variety and statistics of species across the world. Online tools such as *iNaturalist* (2019) enable photo-based classification and subsequent cataloguing in data repositories from pictures uploaded by naturalists. In such automated detection tools, a potentially novel species should result in a request for expert help rather than misclassification into a known species, and detection of undiscovered species is in fact a task of interest. A similar practical application is camera-trap monitoring of members in an ecosystem, notifying caretakers upon detection of invasive species (Fedor et al.; Willi et al., 2009; 2019). Taxonomy of collected specimens is often backlogged due to the human labour involved. Automating digitization and identification can help catch up, and often new species are brought to light through the process (Carranza-Rojas et al., 2017), which obviously depends on effective detection of novel specimens.

Medical diagnosis and clinical microbiology: Online medical diagnosis tools such as *Chester* (Cohen, Bertin, and Frappier, 2019) can be impactful at improving health-care levels worldwide. Such tools should be especially adept at being able to know when faced with a novel pathology rather than categorizing into a known subtype. Similar desiderata applies to being able to quickly detect new strains of pathogens when using machine learning systems to automate clinical identification in the microbiology lab (Zieliński et al., 2017).

AI safety: Amodei et al. (2016) discuss the problem of distributional shift in the context of autonomous agents operat-

Table 1: Sizes of proposed benchmark subsets from ILSVRC2012. The training set consists of roughly 1300 images per member, and 50 images per member in the test set (which come from the validation set images in the ILSVRC2012 dataset).

Subset	Number of members	Total training images	Total test images
Dog (*hound dog*)	12	14864	600
Car	10	13000	500
Snake (*colubrid snake*)	9	11700	450
Spider	6	7800	300
Fungus	6	7800	300

ing in our midst, with examples of actions that do not translate well across domains. A similar example in that vein, grounded in a computer vision classification task, is the contrived scenario of encountering a novel vehicle (that follows different dynamics of motion), which might lead to a dangerous decision by a self-driving car which fails to recognize unfamiliarity.

Having compiled the examples above, we can now try to come up with an evaluative setting more aligned with realistic applications. The basic assumptions we make about possible evaluative tasks are: (i) that anomalies of practical interest are semantic in nature; (ii) that they are relatively rare events whose detection is of more primary relevance than minimizing false positives; and (iii) that we do not have access to examples of anomalies. These assumptions guide our choice of benchmarks and evaluation.

Recommended benchmarks A very small number of recent works (Akcay, Atapour-Abarghouei, and Breckon; Zenati et al., 2018; 2018) have considered a case that is more aligned with the goals stated above. Namely, for a choice of dataset, for example MNIST, train as many versions of classifiers as there are classes, holding out one class every time. At evaluation time, score the ability of being able to detect the held out class as anomalous. This is a setup more clearly related to the task of being able to detect semantic anomalies, holding dataset-bias factors invariant to a significantly greater extent. In this paper, we shall explore this setting with CIFAR-10 and STL-10, and recommend this as the default benchmark for evaluating anomaly detection in the context of object recognition. Similar setups apply to different contexts. We discourage the recently-adopted practice of treating one category as in-distribution and many other categories as out-distributions (as in Pidhorskyi, Almohsen, and Doretto (2018) and Golan and El-Yaniv (2018), for example). While this setting is not aligned with the context of multi-object classification, it relies on a dataset constructed for such a purpose. Moreover, practical situations calling for one-class modelling typically consider anomalies of interest to be (often subtle) variations of the same object, and not a set of very distinct categories.

While the hold-out-class setting for CIFAR-10 and STL-10 is a good setup for testing anomaly detection of disparate objects, a lot of applications, including some of the ones we described earlier, require detection of more fine-grained anomalies. For such situations, we propose a suite of tasks comprised of subsets of ILSVRC2012 (Russakovsky et al., 2015), with fine-grained subcategories. For example, the SPIDER subset consists of members *tarantula*, *Argiope aurantia*, *barn spider*, *black widow*, *garden spider*, *wolf spider*. We also propose FUNGUS, DOG, SNAKE, and CAR subsets. These subsets have varied sizes, with some of them being fairly small (see table 1). Although this is a significantly harder task, we believe this setting aligns with the practical situations we described above, where sometimes large quantities of labelled data are not always available, and a particular fine-grained selection of categories is of interest. See Appendix A for more details about our construction.

Evaluation Current works tend to mainly use both Area under the Receiver-Operator-Characteristics (AUROC) and Area under Precision-Recall curve (AUPRC) to evaluate performance on anomaly detection. In situations where positive examples are not only much rarer, but also of primary interest for detection, AUROC scores are a poor reflection of detection performance; *precision* is more relevant than the false positive rate (Fawcett; Davis and Goadrich; Avati et al., 2006; 2006; 2018). We shall not inspect AUROC scores because in all of our settings, normal examples significantly outnumber anomalous examples, and AUROC scores are insensitive to skew, thus resulting in optimistic scores (Davis and Goadrich, 2006). Precision and recall are calculated as

$$\text{precision} = \frac{\text{true positives}}{\text{true positives + false positives}}, \tag{1}$$

$$\text{recall} = \frac{\text{true positives}}{\text{true positives + false negatives}}, \tag{2}$$

and a precision-recall curve is then defined as a set of precision-recall points, for a varying threshold, t,

$$\text{PR curve} \triangleq \{\text{recall}(t), \text{precision}(t), -\infty < t < \infty\}. \tag{3}$$

The area under the precision-recall curve is calculated by varying the threshold t over a range spanning the data, and creating a finite set of points for the PR curve. One alternative is to interpolate these points, producing a continuous curve as an approximation to the true curve, and computing the area under the interpolation by, for example, the trapezoid rule. Interpolation in a precision-recall curve can sometimes be misleading, as studied in Boyd, Eng, and Page (2013), who recommend a number of more robust estimators. Here we use the standard approximation to average precision as the weighted mean of precisions at thresholds, weighted by the increase in recall from the previous threshold.

$$\text{average precision} = \sum_k \text{precision}_k(\text{recall}_k - \text{recall}_{k-1}). \quad (4)$$

3 Related Work

Evaluative tasks As discussed earlier, the style of benchmarks widely adopted today follows the recommendation in Hendrycks and Gimpel (2017). Among follow-ups, the most significant successor has been Liang, Li, and Srikant (2018) which augmented the suite of tests with slightly more reasonable choices: for example, TINY-IMAGENET is considered as out-of-distribution for in-distrbution datasets such as CIFAR-10. However, on closer inspection, we find that TINY-IMAGENET shares semantic categories with CIFAR-10, such as species of {dogs, cats, frogs, birds}, so it is unclear how such choices of evaluative tasks correspond to realistic anomaly detection problems. Work in the area of *open-set recognition* is closer to a realistic setup in terms of evaluation; in Bendale and Boult (2016), detection of novel categories is tested with a set of images corresponding to different classes that were discontinued in subsequent versions of Imagenet, but later work (Dhamija, Günther, and Boult, 2018) relapsed into treating very different datasets as novel. We do not encourage using one particular split of a collection of unseen classes as anomalous. This is because such a one-time split might favour implicit biases in the predefined split, and the chances of this happening is reduced with multiple hold-out trials. As mentioned earlier, a small number of works have already used the hold-out-class style of tasks for evaluation. Unfortunately, due to a lack of a motivating discussion, the community at large continues to adopt the tasks in Hendrycks and Gimpel (2017) and Liang, Li, and Srikant (2018).

Approaches to OOD detection In Hendrycks and Gimpel (2017), the most natural baseline for a trained classifier is presented, where the detection score is simply given by the predictive confidence of the classifier (MSP). Follow-up work in Liang, Li, and Srikant (2018) proposed adding a small amount of adversarial perturbation, followed by temperature scaling of the softmax (ODIN). Methodologically, the approach suffers from having to pick a temperature and perturbation weight per anomaly-dataset. Complementary methods such as confidence calibration of DeVries and Taylor (2018), have been shown to improve performance of MSP and ODIN.

Using auxiliary datasets as surrogate anomalies has been shown to improve performance on existing benchmarks in Hendrycks, Mazeika, and Dietterich (2019). This approach is limited, due to its reliance on other datasets, but a more practical variant in Lee et al. (2018) uses a GAN to generate negative samples. However, Lee et al. (2018) suffers from the methodological issue of hyperparameters being optimized per anomaly-dataset. We believe that such contentious practices arise from a lack of a clear discussion of the nature of the tasks we should be concerned with, and a lack of grounding in practical applications which would dictate proper methodology. The primary goal of our paper is to help fill this gap.

Table 2: Multi-task augmentation with the self-supervised objective of predicting rotation improves generalization.

	CIFAR-10	STL-10
Classification only	95.87 ± 0.05	85.51 ± 0.17
Classification+rotation	96.54 ± 0.08	88.98 ± 0.30

Shalev, Adi, and Keshet (2018) augment the training set with semantically similar labels, but it is not always practical to assume access to a corpora providing such labels. In the next part of the paper, we explore a way to potentially induce more semantic representation, with the hope that this would lead to corresponding improvements in semantic anomaly detection and generalization.

4 Encouraging Semantic Representations with Auxiliary Self-supervised Objectives

We hypothesize that classifiers that learn representations which are more oriented toward capturing semantic properties would naturally lead to better performance at detecting semantic anomalies. “Overfitting” to low-level features such as colour or texture without consideration of global coherence might result in potential confusions in situations where the training data is biased and not representative. For a lot of existing datasets, it is quite possible to achieve good generalization performance without learning semantic distinctions, a possibility that spurs the search for removing algorithmic bias (Zemel et al., 2013), and which is often exposed in embarrassing ways. As a contrived example, if the training and testing data consists of only one kind of animal which is furry, the classifier only needs to learn about fur-texture, and can ignore other meaningful characteristics such as the shape. Such a system would fail to recognize another furry, but differently shaped creature as novel, while achieving good test performance. Motivated by this line of thinking, we ask the question of how we might encourage classifiers to learn more meaningful representations.

Multi-task learning with auxiliary objectives Caruana (1993) describes how sharing parameters for learning multiple tasks, which are related in the sense of requiring similar features, can be a powerful tool for inducing domain-specific inductive biases in a learner. Hand-design of inductive biases requires complicated engineering, while using the training signal from a related task can be a much easier way to achieve similar goals. Even when related tasks are not explicitly available, it is often possible to construct one. We explore such a framework for augmenting object recognition classifiers with auxiliary tasks. Expressed in notation, given the primary loss function, ℓ_{primary}, which is the categorical cross-entropy loss in the case of classification, and the auxiliary loss $\ell_{\text{auxiliary}}$ corresponding to the auxiliary task, we aim to optimize the combined loss

$$\ell_{\text{combined}}(\theta; \mathcal{D}) = \ell_{\text{primary}}(\theta; \mathcal{D}) + \lambda \ell_{\text{auxiliary}}(\theta; \mathcal{D}), \quad (5)$$

where θ are the shared parameters across both tasks, $\mathcal{D}$ is the dataset, λ is a hyper-parameter we learn by optimizing

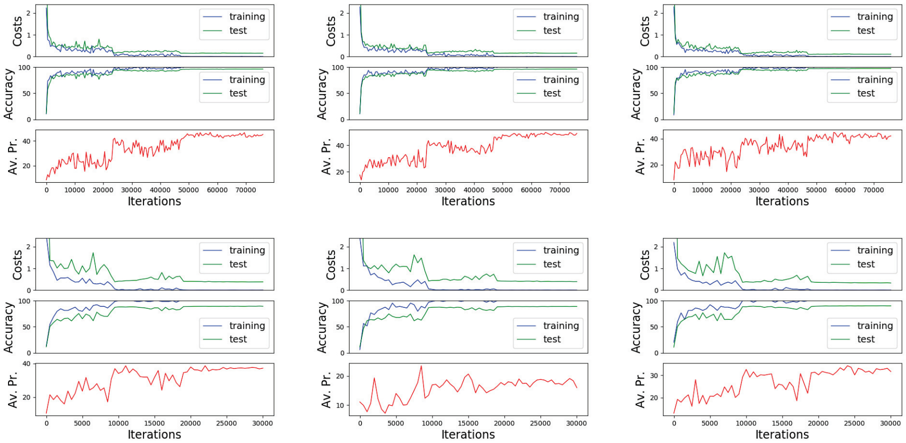

Figure 1: Plots of costs, accuracies, and average precision for hold-out-class experiments with 3 categories each from CIFAR-10 (top) and STL-10 (bottom), using the MSP method (Hendrycks and Gimpel, 2017). While classification performance is not correlated with performance at anomaly detection (compare test accuracy numbers with average precision scores), the "pattern" of improvement at anomaly detection appears roughly related to generalization (compare the coarse shape of test accuracy curves with that of average precision curves).

for classification accuracy on the validation set. In practice, we alternate between the two updates rather than taking one global step; this balances the *training rates* of the two tasks.

Auxiliary tasks Recently, there has been strong interest in self-supervision applied to vision (Doersch, Gupta, and Efros; Pathak et al.; Noroozi and Favaro; Zhang, Isola, and Efros; van den Oord, Li, and Vinyals; Gidaris, Singh, and Komodakis; Caron et al., 2015; 2016; 2016; 2017; 2018; 2018; 2018), exploring tasks that induce representations which are linearly separable by object categories. These objectives naturally lend themselves as auxiliary tasks for encouraging inductive biases towards semantic representations. First, we experiment with the recently introduced task in Gidaris, Singh, and Komodakis (2018), which asks the learner to predict the orientation of a rotated image. In table 2, we show significantly improved generalization performance of classifiers on CIFAR-10 and STL-10 when augmented with the auxiliary task of predicting rotation. Details of experimental settings, and performance on anomaly detection, are in the next section. We also perform experiments on anomaly detection with contrastive predictive coding (van den Oord, Li, and Vinyals, 2018) as the auxiliary task and find that similar trends continue to hold.

The addition of such auxiliary objectives is complementary to the choice of scoring anomalies. Additionally, it enables further augmentation with more auxiliary tasks (Doersch and Zisserman, 2017).

5 Evaluation

We study the two existing representative baselines of maximum softmax probability (MSP) (Hendrycks and Gimpel, 2017), and ODIN (Liang, Li, and Srikant, 2018) on the proposed benchmarks. For ODIN, it is unclear how to choose the hyperparameters for temperature scaling and the weight for adversarial perturbation without assuming access to anomalous examples, an assumption we consider unrealistic in most practical settings. We fix $T = 1000, \epsilon = 5\text{e-}5$ for all experiments, following the most common setting.

Experimental Settings

Settings for CIFAR-10 and STL-10 Our base network for all CIFAR-10 experiments is a Wide ResNet (Zagoruyko and Komodakis, 2016) with 28 convolutional layers and a widening factor of 10 (WRN-28-10) with the recommended dropout rate of 0.3. Following Zagoruyko and Komodakis (2016), we train for 200 epochs, with an initial learning rate of 0.1 which is scaled down by 5 at the 60th, 120th, and 160th epochs, using stochastic gradient descent with Nesterov's momentum at 0.9. We train in parallel on 4 Pascal V100 GPUs with batches of size 128 on each. For STL-10, we use the same architecture but append an extra group of 4 residual blocks with the same layer widths as in the previous group. We use a widening factor of 4 instead of 10, and batches of size 64 on each of the 4 GPUs, and train for twice as long. We use the same optimizer settings as with CIFAR-10. In both cases, we apply standard data augmentation of random crops (after padding) and random horizontal reflections.

Settings for IMAGENET For experiments with the proposed subsets of IMAGENET, we replicate the architecture we use for STL-10, but add a downsampling average pooling layer after the first convolution on the images. We do not use dropout, and use a batch size of 64, train for 200 epochs; otherwise all other details follow the settings for STL-10. The standard data augmentation steps of random crops to a size of 224 × 224 and random horizontal reflections are applied.

Table 3: We train ResNet classifiers on CIFAR-10 holding out each class per run, and score detection with average precision for the maximum softmax probability (MSP) baseline in (Hendrycks and Gimpel, 2017) and ODIN (Liang, Li, and Srikant, 2018). We find that augmenting with rotation results in improved anomaly detection as well as generalization (contrast columns in the right half with the left).

CIFAR-10	Classification-only			Rotation-augmented		
Anomaly	MSP	ODIN	Accuracy	MSP	ODIN	Accuracy
airplane	43.30 ± 1.13	48.23 ± 1.90	96.00 ± 0.16	46.87 ± 2.10	49.75 ± 2.30	96.91 ± 0.02
automobile	14.13 ± 1.33	13.47 ± 1.50	95.78 ± 0.12	17.39 ± 1.26	17.35 ± 1.12	96.66 ± 0.03
bird	46.55 ± 1.27	50.59 ± 0.95	95.90 ± 0.17	51.49 ± 1.07	54.62 ± 1.10	96.79 ± 0.06
cat	38.06 ± 1.31	38.97 ± 1.43	97.05 ± 0.12	53.12 ± 0.92	55.80 ± 0.76	97.46 ± 0.07
deer	49.11 ± 0.53	53.03 ± 0.50	95.87 ± 0.12	50.35 ± 2.57	52.82 ± 2.96	96.76 ± 0.09
dog	25.39 ± 1.17	24.41 ± 1.05	96.64 ± 0.13	32.11 ± 0.82	32.46 ± 1.39	97.36 ± 0.06
frog	40.91 ± 0.81	42.21 ± 0.48	95.65 ± 0.09	52.39 ± 4.58	54.44 ± 5.80	96.51 ± 0.12
horse	36.18 ± 0.77	36.78 ± 0.82	95.64 ± 0.08	39.93 ± 2.30	39.65 ± 4.31	96.27 ± 0.07
ship	28.35 ± 0.81	30.61 ± 1.46	95.70 ± 0.15	29.36 ± 3.16	28.82 ± 4.63	96.66 ± 0.17
truck	27.17 ± 0.73	28.01 ± 1.06	96.04 ± 0.24	29.22 ± 2.87	29.93 ± 3.86	96.91 ± 0.12
Average	34.92 ± 0.41	36.63 ± 0.61	96.03 ± 0.00	40.22 ± 0.16	41.56 ± 0.15	96.83 ± 0.02

Table 4: Average precision scores for hold-out-class experiments with STL-10. We observe that the same trends in improvements hold as with the previous experiments on CIFAR-10.

STL-10	Classification-only			Rotation-augmented		
Anomaly	MSP	ODIN	Accuracy	MSP	ODIN	Accuracy
airplane	19.21 ± 1.05	23.46 ± 1.65	85.18 ± 0.20	22.21 ± 0.76	23.37 ± 1.71	89.24 ± 0.12
bird	29.05 ± 0.69	33.51 ± 0.36	85.91 ± 0.36	36.12 ± 2.08	40.08 ± 3.30	89.91 ± 0.29
car	14.52 ± 0.37	16.14 ± 0.83	84.32 ± 0.55	15.95 ± 2.20	16.87 ± 2.94	89.52 ± 0.44
cat	25.21 ± 0.93	27.92 ± 0.84	86.95 ± 0.36	29.34 ± 1.30	31.35 ± 1.88	90.89 ± 0.26
deer	24.29 ± 0.53	25.94 ± 0.49	85.34 ± 0.35	27.60 ± 2.22	29.71 ± 2.55	89.20 ± 0.17
dog	23.42 ± 0.60	23.44 ± 1.18	87.78 ± 0.45	26.78 ± 0.71	26.14 ± 0.62	91.37 ± 0.33
horse	21.31 ± 1.01	22.19 ± 0.75	85.52 ± 0.21	23.79 ± 1.46	23.59 ± 1.63	89.60 ± 0.11
monkey	23.67 ± 0.83	21.98 ± 0.91	86.66 ± 0.31	28.43 ± 1.67	28.32 ± 1.20	90.07 ± 0.23
ship	14.61 ± 0.12	13.78 ± 0.63	84.65 ± 0.21	16.79 ± 1.20	15.37 ± 1.22	89.33 ± 0.15
truck	15.43 ± 0.17	14.35 ± 0.12	85.34 ± 0.17	17.05 ± 0.50	16.59 ± 0.60	90.08 ± 0.38
Average	21.07 ± 0.25	22.27 ± 0.29	85.77 ± 0.13	24.41 ± 0.23	25.14 ± 0.45	89.92 ± 0.08

Predicting rotation as an auxiliary task For adding rotation-prediction as an auxiliary task, all we do is append an extra linear layer alongside the one that is responsible for object recognition. λ is tuned to 0.5 for CIFAR-10, 1.0 for STL-10, and a mix of 0.5 and 1.0 for IMAGENET. The optimizer and regularizer settings are kept the same, with the learning rate decayed along with the learning rate for the classifier at the same scales.

We emphasize that this procedure is not equivalent to data augmentation, since we do not optimize the linear classification layer for rotated images. Only the rotation prediction linear layer gets updated for inputs corresponding to the rotation task, and only the linear classification layer gets updated for non-rotated, object-labelled images. Asking the classifier to be rotation-invariant would require the auxiliary task to develop a disjoint subset in the shared representation that is not rotation-invariant, so that it can succeed at predicting rotations. This encourages an internally split representation, thus diminishing the potential advantage we hope to achieve from a shared, mutually beneficial space.

CPC as an auxiliary task We also experimented with contrastive predictive coding (van den Oord, Li, and Vinyals 2018) as an auxiliary task. Since this is a patch-based method, the input spaces are different across the two tasks: that of predicting encodings of patches in the image, and that of predicting object category from the entire image. We found that two tricks are very useful for fostering co-operation: (i) replacing the normalization layers with their conditional variants (de Vries et al. 2017) (conditioning on the task at hand), and (ii) using symmetric-padding instead of zero-padding. Since CPC induces significant computational overhead, we resorted to a lighter-weight base network. While this comes at the cost of a drop in performance, we still find, in table 6, that similar patterns of improvements continue to hold. We provide further details in Appendix B.

Discussion

Self-supervised multi-task learning is effective In tables 3 and 4 we report average precision scores on CIFAR-10 and STL-10 for the baseline scoring methods MSP (Hendrycks and Gimpel, 2017) and ODIN (Liang, Li, and Srikant, 2018). We note that ODIN, with fixed hyperparameter settings across all experiments, continues to outperform MSP most of the time. When we augment our classifiers with the auxiliary rotation-prediction task, we find that anomaly detection as well as test set accuracy are markedly improved for both scoring methods. As we have remarked

Table 5: Averaged average precisions for the proposed subsets of Imagenet, with rotation-prediction as the auxiliary task. Each row shows averaged performance across all members of the subset. A random detector would score at the skew rate.

		Classification-only			Rotation-augmented		
Subset	Skew	MSP	ODIN	Accuracy	MSP	ODIN	Accuracy
dog	8.33	23.92 ± 0.49	25.85 ± 0.09	85.09 ± 0.14	24.66 ± 0.58	25.73 ± 0.87	85.25 ± 0.17
car	10.00	21.54 ± 0.62	22.49 ± 0.54	77.17 ± 0.10	21.66 ± 0.19	22.38 ± 0.46	76.72 ± 0.19
snake	11.11	18.62 ± 0.93	19.18 ± 0.79	69.74 ± 1.63	20.23 ± 0.18	21.17 ± 0.12	70.51 ± 0.48
spider	16.67	21.20 ± 0.56	24.15 ± 0.72	68.40 ± 0.21	22.90 ± 1.29	25.10 ± 1.78	68.68 ± 0.77
fungus	16.67	42.56 ± 0.49	44.59 ± 1.46	88.23 ± 0.45	44.19 ± 1.86	46.86 ± 1.13	88.47 ± 0.43

Table 6: Averaged average precisions for the proposed subsets of Imagenet where CPC is the auxiliary task.

		Classification-only			CPC-augmented		
Subset	Skew	MSP	ODIN	Accuracy	MSP	ODIN	Accuracy
dog	8.33	20.84 ± 0.50	22.77 ± 0.74	83.12 ± 0.26	21.43 ± 0.63	24.08 ± 0.63	84.16 ± 0.07
car	10.00	19.86 ± 0.21	21.42 ± 0.48	75.42 ± 0.11	22.21 ± 0.44	23.61 ± 0.57	78.88 ± 0.15
snake	11.11	18.20 ± 0.76	18.67 ± 1.07	66.15 ± 1.89	18.78 ± 0.40	20.39 ± 0.60	68.02 ± 0.85
spider	16.67	22.03 ± 0.68	24.08 ± 0.70	66.65 ± 0.42	22.28 ± 0.60	23.37 ± 0.68	68.67 ± 0.36
fungus	16.67	39.19 ± 1.26	41.71 ± 1.94	87.05 ± 0.06	42.08 ± 0.57	45.05 ± 1.11	88.91 ± 0.46

Table 7: Improving test set performance might not help

Method	Accuracy	Av. Prec. with MSP
Base model	96.03 ± 0.00	34.92 ± 0.41
Random-center-masked	96.27 ± 0.05	34.41 ± 0.74
Rotation-augmented	96.83 ± 0.02	40.22 ± 0.16

earlier, a representation space with greater semanticity should be expected to bring improvements on both fronts. All results report mean ± standard deviation over 3 trials. In table 5, we repeat the same process for the much harder Imagenet subsets. Taken together, our results indicate that multi-task learning with self-supervised auxiliary tasks can be an effective approach for improving anomaly detection, with accompanying improvements in generalization.

Improved test set accuracy is not enough Training methods developed solely to improve generalization, without consideration of the affect on semantic understanding, might perform worse at detecting semantic anomalies. This is because it is often possible to pick up on low-level or contextual discriminatory patterns, which are almost surely biased in relatively small datasets for complex domains such as natural images, and perform reasonably well on the test set. To illustrate this, we run an experiment where we randomly mask out a 16×16 region in CIFAR-10 images from within the central 21×21 region. In table 7, we show that while this leads to improved test accuracies, anomaly detection suffers (numbers are averages across hold-out-class trials). This suggests that while the masking strategy is effective as a regularizer, it might come at the cost of less semantic representation. Certain choices can therefore result in models with seemingly improved generalization but which have poorer representation for tasks that require a more coherent understanding. For comparison, the rotation-augmented network achieves both a higher test set accuracy as well as an improved average precision. This example serves as a caution toward developing techniques that might achieve reassuring test set performance, while inadvertently following an internal *modus operandi* that is misaligned with the pattern of reasoning we hope they discover. This can have unexpected consequences when such models are deployed in the real world.

6 Conclusion

We provided a critical review of the current interest in OOD detection, concluding that realistic applications involve detecting semantic distributional shift for a specified context, which we regard as anomaly detection. While there is significant recent interest in the area, current research suffers from questionable benchmarks and methodology. In light of these considerations, we suggested a set of benchmarks which are better aligned with realistic anomaly detection applications in the context of object classification systems.

We also explored the effectiveness of a multi-task learning framework with auxiliary objectives. Our results demonstrate improved anomaly detection along with improved generalization under such augmented objectives. This suggests that inductive biases induced through such auxiliary tasks could have an important role to play in developing more trustworthy neural networks.

We note that the ability to detect semantic anomalies also provides us with an indirect view of semanticity in the representations learned by our mostly opaque deep models.

Acknowledgements

We thank Rachel Rolland, Ishaan Gulrajani, Tim Cooijmans, and anonymous reviewers for useful discussions and feedback.

This work was enabled by the computational resources provided by Compute Canada and funding support from the Canadian CIFAR AI chair and NSERC Discovery Grant.

Table 8: Imagenet subset members

Dog (hound)	Car	Snake (colubrid)	Spider	Fungus
Ibizan hound	Model T	ringneck snake	tarantula	stinkhorn
bluetick	race car	vine snake	Argiope aurantia	bolete
beagle	sports car	hognose snake	barn spider	hen-of-the-woods
Afghan hound	minivan	thunder snake	black widow	earthstar
Weimaraner	ambulance	garter snake	garden spider	gyromitra
Saluki	cab	king snake	wolf spider	coral fungus
redbone	beach wagon	night snake		
otterhound	jeep	green snake		
Norweigian elkhound	convertible	water snake		
basset hound	limo			
Scottish deerhound				
bloodhound				

A Imagenet benchmarks

We first sorted all candidate subsets by the number of members. We then picked from among the list of top twenty subsets, with a preference for subsets that are more closely aligned with the theme of motivating practical applications we provided. We also manually inspected the data, to check for inconsistencies, and performed some pruning. For example, in the *beetle* subset, *leaf beetle* and *ladybug* appear to overlap sometimes. Finally, we settled on our choice of 5 subsets. In table 8, we list the members under every proposed subset. The sets are collected by first resizing such that the shorter side is of length 256 pixels, followed by a center crop. For tuning λ, we treat 20% of the data in the training sets as validation, and the remaining 80% for training.

B Experiments with CPC

CPC involves performing predictions for encodings of patches of an image from those above them. To avoid learning trivial codes, a contrastive loss is used which essentially trains the model to distinguish between correct codes and “noisy” ones. These negative samples are taken from patches within and across images in the batch.

We use the same network architecture as we used for the Imagenet experiments with rotation-prediction as the auxiliary task, but modify the first convolution layer to have a stride of 2. This reduces the computation overhead sufficiently for concurrent training with CPC at reasonable batch-sizes (CPC training batch-sizes are 32), but at a minor expense of classifier performance. We use the first three blocks of the network for the patch encoder as in (van den Oord, Li, and Vinyals, 2018), and append the final layers for the classification task. Unlike with rotation, the auxiliary task works on patches while the primary classifier works on the entire image. This leads to differences in the operating receptive-fields, and differing proportions of boundary effects. To facilitate easier parameter sharing across the two tasks, we make the following changes. First, we replace all default zero-padding with reflected, symmetric padding. This removes the effect of having a different ratio of border-zeros to pixels when the spatial dimensions of the input changes. Second, we replace all normalization layers with conditional normalization variants (this means separate sets of scale and shift parameters are used depending on the current prediction task). Since batch-normalization allows trivial solutions to CPC for patches sampled from different images, we only use patches from within the same image, and find that we can continue using it to our advantage. We keep the same optimizer settings from the rotation experiments, but it is possible that different choices might lead to further improvements. λ is tuned to 10.0 for all experiments, following a coarse hyperparameter search for best validation-set classification accuracy over a range of $\{0.1, 1.0, 10.0, 20.0, 50.0\}$.

C Trivial baseline for existing benchmarks

To demonstrate that the current benchmarks are trivial with very low-level information, we experiment with CIFAR-10 as in-distribution by simply looking at likelihoods under a mixture of 3 pixel-level Gaussians, trained channel-wise. We find that this simple baseline compares very well with recent approaches at all but one of the benchmark OOD tasks in (Liang, Li, and Srikant, 2018) for CIFAR-10.

OOD dataset	Average precision
TinyImagenet (crop)	96.84
TinyImagenet (resize)	99.03
LSUN	58.06
LSUN (resize)	99.77
iSUN	99.21

We see that this underperforms on LSUN. When we inspect LSUN, we find that the images are cropped patches from scene-images, and a majority of them are of uniform colour and texture, with little variation and structure in them. While this dataset is most obviously different from CIFAR-10, we believe that the appearance of the images results in the phenomenon reported in Nalisnick et al. (2019), where one distribution that “sits inside” the other because of a similar mean but lower variance ends up being more likely under the wider distribution. In fact, thresholding on simply the “energy” of the edge-detection map gives us an average precision of around 87.5% for LSUN, thus indicating that the extremely trivial feature of a lower edge-count is already a strong indicator for telling apart such an obvious difference.

We found that the Gaussian baseline underperforms severely on the hold-out-class experiments on CIFAR-10, achieving an average precision of a mere 11.17% across the 10 experiments.

References

Akcay, S.; Atapour-Abarghouei, A.; and Breckon, T. P. 2018. Ganomaly: Semi-supervised anomaly detection via adversarial training. *ACCV*.

Alcorn, M.; Li, Q.; Gong, Z.; Wang, C.; Mai, L.; shinn Ku, W.; and Nguyen, A. 2019. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. *CVPR*.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P. F.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *CoRR* abs/1606.06565.

Avati, A.; Jung, K.; Harman, S.; Downing, L.; Ng, A.; and Shah, N. H. 2018. Improving palliative care with deep learning. *BMC Medical Informatics and Decision Making* 18(4):122.

Beery, S.; Horn, G. V.; and Perona, P. 2018. Recognition in terra incognita. *CoRR*.

Bendale, A., and Boult, T. E. 2016. Towards open set deep networks. *ICCV*.

Boyd, K.; Eng, K. H.; and Page, C. D. 2013. Area under the precision-recall curve: Point estimates and confidence intervals. *ECML-PKDD*.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *ECCV*.

Carranza-Rojas, J.; Goeau, H.; Bonnet, P.; Mata-Montero, E.; and Joly, A. 2017. Going deeper in the automated identification of herbarium specimens. *BMC Evolutionary Biology* 17(1):181.

Caruana, R. 1993. Multitask learning: A knowledge-based source of inductive bias. In *ICML*.

Cohen, J. P.; Bertin, P.; and Frappier, V. 2019. Chester: A web delivered locally computed chest x-ray disease prediction system. *CoRR* abs/1901.11210.

Davis, J., and Goadrich, M. 2006. The relationship between precision-recall and roc curves. 233–240.

de Vries, H.; Strub, F.; Mary, J.; Larochelle, H.; Pietquin, O.; and Courville, A. C. 2017. Modulating early visual processing by language. *NIPS*.

DeVries, T., and Taylor, G. W. 2018. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*.

Dhamija, A. R.; Günther, M.; and Boult, T. 2018. Reducing network agnostophobia. *NIPS*.

Doersch, C., and Zisserman, A. 2017. Multi-task self-supervised visual learning. In *ICCV*.

Doersch, C.; Gupta, A.; and Efros, A. A. 2015. Unsupervised visual representation learning by context prediction. *ICCV*.

Fawcett, T. 2006. An introduction to roc analysis. *Pattern Recognition Letters* 27(8):861 – 874.

Fedor, P.; Vanhara, J.; Havel, J.; Malenovsky, I.; and Spellerberg, I. 2009. Artificial intelligence in pest insect monitoring. *Systematic Entomology* 34(2):398–400.

Fodor, J. A., and Pylyshyn, Z. W. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28(1):3 – 71.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2019. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *ICLR*.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised representation learning by predicting image rotations. *ICLR*.

Golan, I., and El-Yaniv, R. 2018. Deep anomaly detection using geometric transformations. *NeuRIPS*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. *ICML* 1321–1330.

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

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2019. Deep anomaly detection with outlier exposure. *ICLR*.

iNaturalist. 2019. https://news.developer.nvidia.com/ai-app-identifies-plants-and-animals-in-seconds, accessed on 17 may 2019.

Lee, K.; Lee, H.; Lee, K.; and Shin, J. 2018. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *ICLR*.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing the reliability of out-of-distribution detection in neural networks. *ICLR*.

Nalisnick, E.; Matsukawa, A.; Teh, Y. W.; Gorur, D.; and Lakshminarayanan, B. 2019. Do deep generative models know what they don't know? *ICLR*.

Noroozi, M., and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. *ECCV*.

Pathak, D.; Krähenbühl, P.; Donahue, J.; Darrell, T.; and Efros, A. 2016. Context encoders: Feature learning by inpainting. *CVPR*.

Pidhorskyi, S.; Almohsen, R.; and Doretto, G. 2018. Generative probabilistic novelty detection with adversarial autoencoders. *NIPS*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV*.

Shalev, G.; Adi, Y.; and Keshet, J. 2018. Out-of-distribution detection using multiple semantic label representations. *NeuRIPS*.

Tommasi, T.; Patricia, N.; Caputo, B.; and Tuytelaars, T. 2017. *A Deeper Look at Dataset Bias*. 37–55.

Torralba, A., and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR*, 1521–1528.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *CoRR* abs/1807.03748.

Willi, M.; Pitman, R. T.; Cardoso, A. W.; Locke, C.; Swanson, A.; Boyer, A.; Veldthuis, M.; and Fortson, L. 2019. Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution* 10(1):80–91.

Zagoruyko, S., and Komodakis, N. 2016. Wide residual networks. In *BMVC*.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. *ICML*.

Zenati, H.; Foo, C. S.; Lecouat, B.; Manek, G.; and Chandrasekhar, V. R. 2018. Efficient gan-based anomaly detection. *CoRR* abs/1802.06222.

Zhang, R.; Isola, P.; and Efros, A. A. 2017. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. *CVPR*.

Zieliński, B.; Plichta, A.; Misztal, K.; Spurek, P.; Brzychczy-Włoch, M.; and Ochońska, D. 2017. Deep learning approach to bacterial colony classification. *PLoS One* 12(9).