

# On the Importance of Regularisation & Auxiliary Information in OOD Detection<sup>\*</sup>

John Mitros<sup>[0000–0002–0189–2130]</sup> and Brian Mac Namee<sup>[0000–0001–7842–1516]</sup>

University College Dublin, IR  
School of Computer Science  
`ioanni.mitro@ucdconnect.ie`, `brian.macnamee@ucd.ie`

**Abstract.** Neural networks are often utilised in critical domain applications (e.g. self-driving cars, financial markets, and aerospace engineering), even though they exhibit overconfident predictions for ambiguous inputs. This deficiency demonstrates a fundamental flaw indicating that neural networks often overfit on spurious correlations. To address this problem in this work we present two novel objectives that improve the ability of a network to detect out-of-distribution samples and therefore avoid overconfident predictions for ambiguous inputs. We empirically demonstrate that our methods outperform the baseline and perform better than the majority of existing approaches while still maintaining a competitive performance against the rest. Additionally, we empirically demonstrate the robustness of our approach against common corruptions and demonstrate the importance of regularisation and auxiliary information in out-of-distribution detection.

**Keywords:** Out-of-Distribution Detection · Neural Networks · Robust Predictions · Stability · Overconfident Predictions · Anomaly Detection · Open Set Recognition.

## 1 Introduction

Out of distribution (OOD) detection is becoming more important as machine learning solutions are developed for critical applications (e.g. self-driving cars, financial markets, and aerospace engineering), and especially in evaluating the robustness of deployed models. The main goal of OOD is to equip a classifier with the ability to provide stable, consistent and low confidence predictions for data points that might be far away from the in-distribution (ID) training data. This is often referred to as the capacity of the model to generalise.

A central assumption in statistical learning theory [41,42,1,43,26] states that the train and test set  $(x_{i=1}^n, y_{i=1}^n)$ , are generated independently and identically distributed (IID) from a distribution  $P$ , such that data points are usually assigned randomly to either train or test set. Unfortunately, this assumption fails

---

<sup>\*</sup> This research is supported by Sience Foundation Ireland (SFI) under Grant number SFI/12/RC/2289.P2.

to assess whether the model has learned to properly generalise to new unseen data or has simply overfit to irrelevant factors (e.g., backgrounds in image recognition task) that might be spuriously correlated with the correct label due to shortcut learning [8]. Numerous methods have been proposed to mitigate this deficiency and improve OOD detection [21,24,39,22,33,29,20] within supervised, semi-supervised, and unsupervised learning [2], including discriminative [37,31,15] and generative [4,25,23,49] models. In addition, some methods cast the OOD problem as binary classification [38] with alternative approaches relying either only on ID data [7] or both ID and OOD data [14] during training.

Inspired by recent progress in contrastive learning [30,9,11] we propose two novel objectives for OOD detection and demonstrate empirically that our method not only is competitive with existing approaches but it also outperforms some of them in most occasions. Additionally, we empirically study the role of regularisation in OOD detection and robust classification.

In this work we investigate the following questions:

- Can contrastive learning improve OOD detection in neural networks?
- What is the role of explicit regularisation in OOD detection, and does additional regularisation improve or degrade OOD detection?

The main contributions of this work are:

- A novel objective function based on the cosine angle between the ID and OOD data.
- A novel objective function inspired by prior work on margin and ranking objectives utilising the cosine angle between ID and OOD data as well as additional explicit regularisation.

## 2 Related Work

In this section we describe existing work on the OOD problem, and the objectives used in recent approaches based on contrastive learning.

### 2.1 Out of Distribution Detection

Early attempts at OOD detection [13] used the maximum softmax probability as an indicator to identify OOD samples, while alternative approaches such as ODIN [24] use adversarially perturbed samples while computing the softmax with high temperature during training. Furthermore, the Mahalanobis detector [22] fits a Gaussian distribution to the activation of the last layer of a neural network and performs OOD by measuring the Mahalanobis distance from the inputs to the ID data.

In addition, methods explicitly trained to output uniform distribution over ID perturbed samples, usually resemble techniques simulating OOD inputs from a GAN [21], or utilising auxiliary information (e.g. additional datasets) as outliers [14]. Finally, there exist approaches relying on averaging predictions of

randomly initialised, independently trained, neural networks, either in a discriminative [19] or generative [4] approach.

A naturally occurring question is: “*What do these methods have in common?*” To the best of our knowledge the majority of techniques proposed to tackle the OOD problem can be attributed to one of the following categories: *optimisation*, *regularisation*, or *sampling*—or a combination of the three.

## 2.2 Objective Functions in Machine Learning

Most objectives adopted today in machine learning (e.g. cross-entropy, mean squared error, and log-likelihood) have a single goal, to induce a cost in order for the underlying model to directly learn a label, a value, or a set of values from a specific input. In contrast, ranking objectives strive to predict similarities (i.e. relative distances) between inputs, thus the underlying task is often identified and referred to as metric learning. The key idea is to employ a metric function (e.g. Euclidean distance) in order to obtain a similarity score between inputs embedded in a latent feature space, where the score should be small for similar inputs and large otherwise. One such example is SimCLR [3] that maximises the agreement in latent representations via a contrastive objective between pairs of inputs.

Let  $\cos(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} / \|\mathbf{u}\| \|\mathbf{v}\|$  define a similarity score indicated by the cosine angle between vectors  $\mathbf{u}$  and  $\mathbf{v}$ . This is utilised in the SimCLR objective. Given a pair of distinct latent features  $(\mathbf{z}_i, \mathbf{z}_j)$ , such that  $\mathbf{z}_{i,j} = f_\theta(t_{i,j}(\mathbf{x}))$ ,  $\forall \mathbf{x} \in \mathcal{X}$ , with augmentation operations  $t_i, t_j \sim \mathcal{T}$ , such that  $t_i \neq t_j$ , and temperature scaling  $\tau$  then the objective is formulated as:

$$L(\mathbf{z}_i, \mathbf{z}_j) = -\log \frac{\exp(\cos(\mathbf{z}_i, \mathbf{z}_j) / \tau)}{\sum_{j=1}^{2n} \mathbb{1}_{[j \neq i]} \exp(\cos(\mathbf{z}_i, \mathbf{z}_j) / \tau)} \quad (1)$$

Although ranking objectives (e.g. pairwise, triplet, etc.) might differ with regards to the number of inputs they operate on (e.g. pairs or triplets), nevertheless, their main concept is to learn a similarity/dissimilarity metric, (e.g.  $\ell^p$ -norm) on latent representations  $(\mathbf{z}_i, \mathbf{z}_j)$  such that  $d = \|\mathbf{z}_i, \mathbf{z}_j\|_2$  for  $j = \{1, \dots, n\}$ . Given a set of inputs  $\{\mathbf{x}, \mathbf{x}_1^+, \mathbf{x}_2^-, \dots, \mathbf{x}_n^-\}$  with one positive and  $n-1$  negative samples, where  $\mathbf{x}$  represents an anchor sample,  $\mathbf{x}_{i=1}^+$  a positive sample and  $\mathbf{x}_{j=2, \dots, n}^-$  a negative sample, with  $y = \{0, 1\}$  being the labels. Then, a **pairwise ranking objective** strives to learn representations with small distance  $d$  between positive pairs  $(\mathbf{x}, \mathbf{x}_i^+)$  and greater than a margin  $\gamma$  for negative pairs  $(\mathbf{x}, \mathbf{x}_j^-)$  such that.

$$L(\mathbf{x}_i, \mathbf{x}_j, y) = \begin{cases} y d(\mathbf{z}, \mathbf{z}^+) & \text{if } (\mathbf{x}, \mathbf{x}_i) \text{ is a positive pair} \\ (1-y) \max(0, \gamma - d(\mathbf{z}, \mathbf{z}^-)) & \text{if } (\mathbf{x}, \mathbf{x}_j) \text{ is a negative pair} \end{cases} \quad (2)$$

Instead of pairs the **triplet ranking objective** uses triplets  $\{\dots, \mathbf{x}, \mathbf{x}^-, \mathbf{x}^+, \dots\}$ . We have an anchor  $\mathbf{x}$ , a positive  $\mathbf{x}^+$ , and a negative  $\mathbf{x}^-$  sample, instead of pairs of

positive  $(\mathbf{x}, \mathbf{x}^+)$  and negative  $(\mathbf{x}, \mathbf{x}^-)$  samples as illustrated in Equation 2. The goal is to learn representations with greater distance between the anchor and the negative sample  $d(\mathbf{z}, \mathbf{z}^-)$  than between the anchor and the positive sample  $d(\mathbf{z}, \mathbf{z}^+)$ . The final triplet objective is formulated as:

$$L(\mathbf{x}, \mathbf{x}^+, \mathbf{x}^-) = \max(0, \gamma + d(\mathbf{z}, \mathbf{z}^+) - d(\mathbf{z}, \mathbf{z}^-)) \quad (3)$$

$$= \begin{cases} \text{Easy triplets:} & \text{if } d(\mathbf{z}, \mathbf{z}^-) > d(\mathbf{z}, \mathbf{z}^+) + \gamma \\ \text{Semi-hard triplets:} & \text{if } d(\mathbf{z}, \mathbf{z}^+) < d(\mathbf{z}, \mathbf{z}^-) < d(\mathbf{z}, \mathbf{z}^+) + \gamma \\ \text{Hard triplets:} & \text{if } d(\mathbf{z}, \mathbf{z}^-) < d(\mathbf{z}, \mathbf{z}^+) \end{cases}$$

### 3 Objective Functions for OOD Detection

The first contribution that this article makes to OOD detection is a novel objective based on the cosine similarity between ID and OOD predictions. Let  $(\mathbf{x}_{id}, y_{id}) \sim \mathcal{S}_{ID}$  and  $(\mathbf{x}_{ood}, y_{ood}) \sim \mathcal{S}_{OOD}$  represent two data points sampled from the ID data  $\mathcal{S}_{ID}$  and OOD data  $\mathcal{S}_{OOD}$  respectively, and, define  $\mathbf{p}_{id} = \max_{y_{id}} p(y_{id} | f_{\theta}(\mathbf{x}_{id}))$  to be the maximum softmax probability (MSP) for  $\mathbf{x}_{id} \in \mathcal{S}_{ID}$ , and,  $\mathbf{p}_{ood} = \max_{y_{ood}} p(y_{ood} | f_{\theta}(\mathbf{x}_{ood}))$  be the MSP for  $\mathbf{x}_{ood} \in \mathcal{S}_{OOD}$ . Then our objective is formulated as:

$$L(\mathbf{x}_{id}, \mathbf{x}_{ood}, y_{id}) = \underbrace{-\mathbb{E}[\log p(y_{id} | \mathbf{x}_{id})]}_{\text{cross-entropy}} + \underbrace{\lambda \cos(\mathbf{p}_{id}, \mathbf{p}_{ood})}_{\text{cosine-regularisation}} \quad (4)$$

The regularisation strength  $\lambda$  is often obtained using the validation set, and whenever  $\lambda = -1$  then the underlying objective becomes a minimax optimisation formulation similar to adversarial learning paradigms [32], with the advantage that it is faster to train a model with this objective since it avoids computing gradients for worst-case perturbations on the inputs. The goal is to lower the cross-entropy error on  $\mathcal{S}_{ID}$  while at the same time increasing the cosine angle between  $\mathcal{S}_{ID}$  and  $\mathcal{S}_{OOD}$ . This synergy of minimax optimisation can also be found in energy-based models [25,10] where the intention is to lower the energy for similar samples while at the same time increasing the energy on dissimilar inputs. This approach is summarised in Algorithm 1.

---

#### Algorithm 1 Contrastive Regularised Objective

---

```

procedure CONTRREG( $\mathbf{x}_{id}, \mathbf{x}_{ood}, y_{id}$ )
     $f_{\theta} \leftarrow \theta$  ▷ initialise model
     $\mathbf{z}_{id}, \mathbf{z}_{ood} \leftarrow f_{\theta}(\mathbf{x}_{id}, \mathbf{x}_{ood})$  ▷ compute logits for  $\mathbf{x}_{id} \in \mathcal{S}_{ID}, \mathbf{x}_{ood} \in \mathcal{S}_{OOD}$ 
     $\hat{\mathbf{p}}_{id}, \hat{\mathbf{p}}_{ood} \leftarrow \text{softmax}(\mathbf{z}_{id}, \mathbf{z}_{ood})$  ▷ probab. for logits  $\in (\mathcal{S}_{ID}, \mathcal{S}_{OOD})$ 
     $CE \leftarrow -\mathbb{E}[\log p(y_{id} | \mathbf{x}_{id})]$  ▷ compute cross-entropy for  $(\mathbf{x}_{id}, y_{id}) \in \mathcal{S}_{ID}$ 
     $\cos \leftarrow \frac{\mathbf{p}_{id}^T \mathbf{p}_{ood}}{\|\mathbf{p}_{id}\| \|\mathbf{p}_{ood}\|}$  ▷ compute cosine for probabilities  $\hat{\mathbf{p}}_{id}, \hat{\mathbf{p}}_{ood}$ 
     $L \leftarrow CE + \lambda \cos$  ▷ compute final regularised loss
     $\theta_{t+1} = \theta_t - \eta \nabla_{\theta} L$  ▷ compute gradient w.r.t params  $\theta$  and backprop errors
end procedure

```

---

The second contribution is a novel ranking objective for OOD detection. We utilise cosine similarity as a metric learning function in addition to explicit  $\ell^2$  and  $\ell^1$  regularisation for  $\mathcal{S}_{ID}$  and  $\mathcal{S}_{OOD}$  respectively, which essentially constitutes into the following objective:

$$\begin{aligned}
 L(\mathbf{x}_{id}, \mathbf{x}_{ood}, y_{id}) = & \underbrace{\max(0, \gamma + \cos(\mathbf{p}_{id}, \mathbf{p}_{ood}))}_{\text{ranking objective}} + \underbrace{\lambda_1 \sum_n |\mathbf{p}_{ood} - 1/k|}_{\ell_1\text{-regularisation on } \mathcal{S}_{OOD}} \\
 & + \underbrace{\lambda_2 \sum_n \|y_{id} \mathbf{p}_{id} - \alpha\|_2}_{\ell_2\text{-regularisation on } \mathcal{S}_{ID}} \quad (5)
 \end{aligned}$$

Notice that  $k \in \mathbb{Z}$  refers to the number of ID classes in  $\mathcal{S}_{ID}$ , and  $y_{id}$  represents a one-hot encoding of the labels, while  $\alpha \in \mathbb{R}$  is a user defined scalar that indicates the desired ID accuracy. There are a number of hyperparameters  $\{\gamma, \lambda_1, \lambda_2\}$  which can be tuned on the validation set,  $\gamma$  defines the margin and  $\lambda_1, \lambda_2$  refer to the regularisation strength. This approach is depicted in Algorithm 2.

---

**Algorithm 2** Contrastive Ranking Objective
 

---

```

procedure CONTRANK( $\mathbf{x}_{id}, \mathbf{x}_{ood}, y_{id}$ )
     $f_\theta \leftarrow \theta$  ▷ initialise model
     $\mathbf{z}_{id}, \mathbf{z}_{ood} \leftarrow f_\theta(\mathbf{x}_{id}, \mathbf{x}_{ood})$  ▷ compute logits for  $\mathbf{x}_{id} \in \mathcal{S}_{ID}, \mathbf{x}_{ood} \in \mathcal{S}_{OOD}$ 
     $\hat{\mathbf{p}}_{id}, \hat{\mathbf{p}}_{ood} \leftarrow \text{softmax}(\mathbf{z}_{id}, \mathbf{z}_{ood})$  ▷ probab. for logits  $\in (\mathcal{S}_{ID}, \mathcal{S}_{OOD})$ 
     $\ell_1 \leftarrow \lambda_1 \sum_n |\mathbf{p}_{ood} - 1/k|$  ▷ compute  $\ell_1$ -regularisation for  $\hat{\mathbf{p}}_{ood} \in \mathcal{S}_{ood}$ 
     $\ell_2 \leftarrow \lambda_2 \sum_n \|y_{id} \mathbf{p}_{id}, \alpha\|$  ▷ compute  $\ell_2$ -regularisation for  $\hat{\mathbf{p}}_{id} \in \mathcal{S}_{id}$ 
     $\cos \leftarrow \frac{\mathbf{p}_{id}^T \mathbf{p}_{ood}}{\|\mathbf{p}_{id}\| \|\mathbf{p}_{ood}\|}$  ▷ compute cosine for probabilities  $\hat{\mathbf{p}}_{id}, \hat{\mathbf{p}}_{ood}$ 
     $L \leftarrow \max(0, \gamma + \cos(\cdot)) + \ell_1 + \ell_2$  ▷ compute the final ranking loss
     $\theta_{t+1} = \theta_t - \eta \nabla_\theta L$  ▷ compute gradient w.r.t params  $\theta$  and backprop errors
end procedure
  
```

---

## 4 Experiments

In this section we describe a set of experiments designed to evaluate the effectiveness of the proposed objectives defined in the previous section, and to compare them to existing approaches. We first evaluate the objectives using an artificially generated dataset, before using a selection of real image classification datasets to evaluate them.

### 4.1 Artificial Data Experiments

To validate the efficacy of our proposed objectives for OOD detection we designed a controlled experiment utilising synthetic data. The training ID data  $\mathcal{S}_{ID}$  is comprised of 3 Gaussians with standard deviation  $\sigma$  representing different classes

in a multi-class classification setting. The different subset splits  $train \sim \mathcal{S}_{ID}^{train}$ ,  $test \sim \mathcal{S}_{ID}^{test}$  and  $test\ OOD \sim \mathcal{S}_{OOD}$  over the synthetic dataset are depicted in Figure 1.

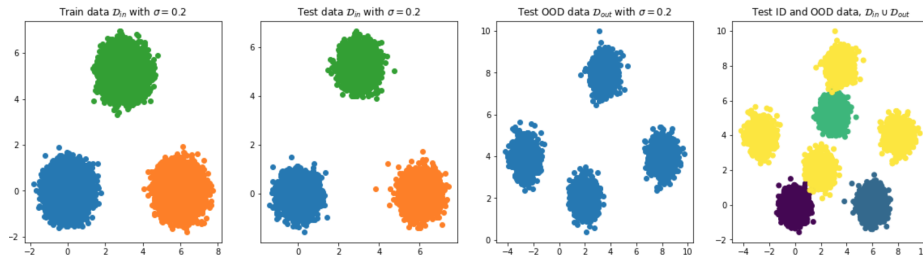


Fig. 1: Synthetic dataset from left to right, comprised of ID train data (1st), ID test data (2nd), OOD test data (3rd), and finally the union of  $ID \cup OOD$  (yellow) test data (4th).

The underlying model is a 3-layer MLP network, trained on the synthetic ID train split. During inference we test each objective on the OOD test data  $\mathcal{S}_{OOD}$  constructed of 4-Gaussians displaced at different locations. To measure performance at OOD detection we measure AUC based on three OOD metrics on the models’ logits: *confidence*, *entropy*, and *mutual information*.

**Results & Discussion** Table 1 presents our findings for OOD detection across objectives, while Figure 2 depicts the different decision boundaries across each objective for ID (1st row) and OOD (2nd row) test data.

Table 1: Accuracy and AUC-ROC-scores across objectives & metrics represented in percentage (%).

Data		Objectives	Accuracy	AUC-ROC scores		
$\mathcal{S}_{ID}$	$\mathcal{S}_{OOD}$			Confidence	Entropy	Mutual Information
3-Gaussians	4-Gaussians	CrossEntropy	100	61.64	61.61	63.62
		CrossEntropy+MC-Dropout	100	75.14	73.63	73.56
		ContReg (ours)	100	99.99	99.99	99.99
		ContRank (ours)	100	99.99	99.99	99.99

As suggested in Table 1 our proposed methods achieve near optimal OOD detection when presented with ambiguous test data. Notice that explicit regularisation (e.g. MC-Dropout) does indeed provide additional benefit in OOD detection. Similar conclusions supporting our claims have been demonstrated in prior works of [36,34,16,44,17] regarding the impact of regularisation. To understand why explicit regularisation improves OOD detection we exhibit the existence of a connection among Dropout [6], Mixup [48] and Randomised smooth-

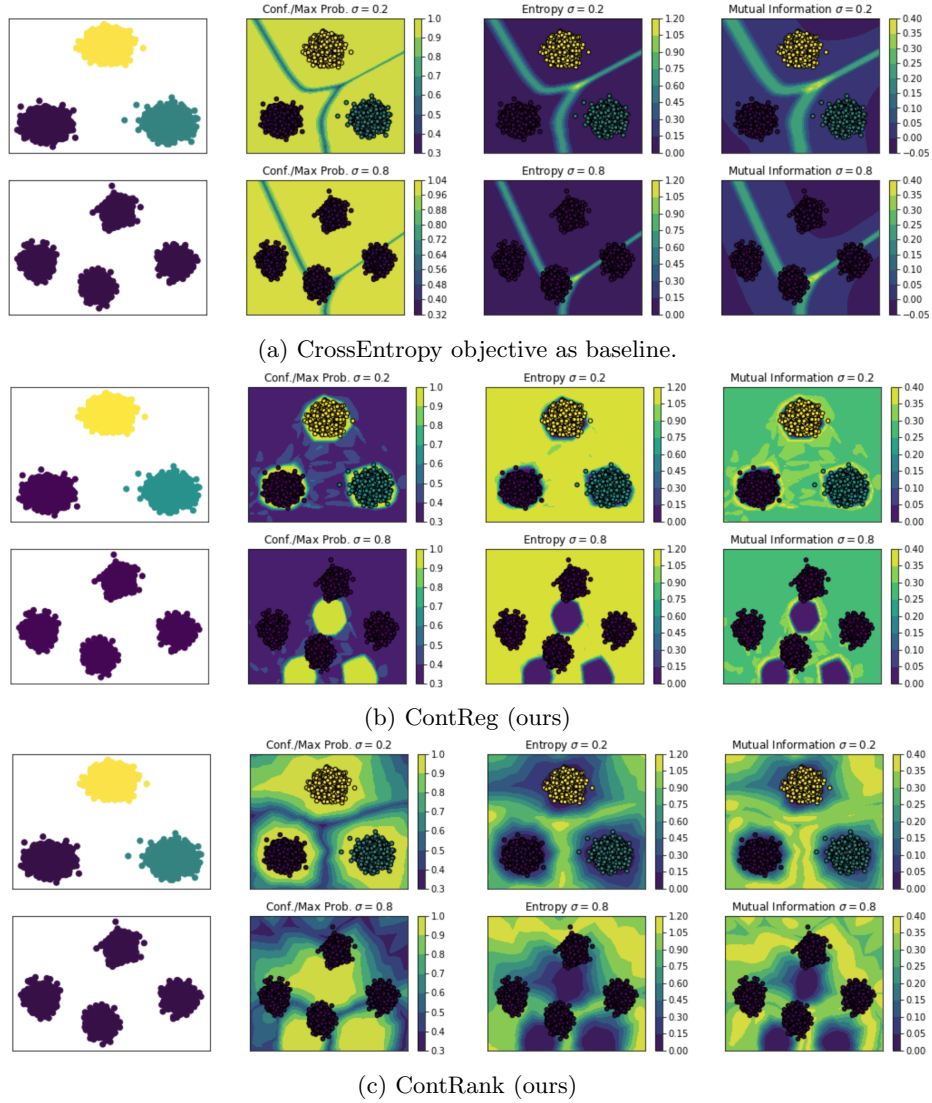


Fig. 2: Decision boundaries across objective functions for ID (1st row) and OOD (2nd row) test data.

ing [5], where these methods act as boundary thickness [46]. It is evident from Figure 2 that a model trained with CrossEntropy on a classification setting acts as a max-margin predictor while our objective act as density estimator. This indicates that the choice of objective and regularisation play a crucial role in the final behaviour of the predictor.

## 4.2 Real Data Experiment

Five well-known image classification datasets are used in this experiment: *CIFAR-10*, *CIFAR-100*, *SVHN*, *FashionMNIST* and *LSUN*. Every dataset was split into three distinct sets  $\{train, validation, test\}$  with random mirroring and cropping augmentations. We utilised WideResNet28x10 [47] as the DNN model trained for 300 epochs using a validation set for hyper-parameter tuning and rolling back to the best checkpoint to avoid overfitting. The optimiser was Stochastic Gradient Descent (SGD) [35,18] with momentum set to 0.9 and weight decay in the range  $[3e^{-4}, 5e^{-4}]$ . Given that all datasets have balanced class distributions we utilised classification accuracy to measure their performance on the clean ID test data. To measure the ability of the models to recognise OOD examples we utilised the predictions of the test portion of three OOD datasets (see Table 5). We measure the separation between ID and OOD data using the area under the curve (AUC-ROC) for each approach.

We also compare the effectiveness of the custom objective described in this paper with the following existing objectives designed to address the OOD problem.

$$\text{Mahalanobis [22]} \quad M(\mathbf{x}) = \max_c -(f(\mathbf{x}) - \hat{\mu}_c)^\top \hat{\Sigma}^{-1} (f(\mathbf{x}) - \hat{\mu}_c)$$

$$\text{ODIN [24]} \quad g(\mathbf{x}; \delta, T, \varepsilon) = \begin{cases} 1 & \text{if } \max_i p(\tilde{\mathbf{x}}; T) \leq \delta \\ 0 & \text{if } \max_i p(\tilde{\mathbf{x}}; T) > \delta \end{cases}$$

$$\text{MSRep [39]} \quad \bar{\ell}(\mathbf{x}, y; \theta) = \sum_{k=1}^K d_{\cos}(e^k(y), f_{\theta^k}^k(\mathbf{x}))$$

$$\text{OutlierExposure [14]} \quad \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{in}}} \left[ \mathcal{L}(f(x), y) + \lambda \mathbb{E}_{x' \sim \mathcal{D}_{\text{out}}^{\text{OE}}} [\mathcal{L}_{\text{OE}}(f(x'), f(x), y)] \right]$$

$$\text{EnergyOOD [25]} \quad \min_{\theta} \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}_{\text{in}}} [-\log F_y(\mathbf{x})] + \lambda \cdot L_{\text{energy}}$$

$$\text{CSI [40]} \quad -\frac{1}{|\{x_+\}|} \log \frac{\sum_{x' \in \{x_+\}} \exp(\text{sim}(z(x), z(x'))/\tau)}{\sum_{x' \in \{x_+\} \cup \{x_-\}} \exp(\text{sim}(z(x), z(x'))/\tau)}$$

$$\text{DoSE [28]} \quad \frac{1}{m} \sum_j^m (\log q(x_j | \{x_i\}_i^n, T, \gamma)^2 + 2\mathbb{H}[p]) \frac{1}{m} \sum_j^m \log q(x_j | \{x_i\}_i^n, T, \gamma)$$

**Results & Discussion** According to Table 3 and Table 5 our objectives outperform the max softmax probability (MSP) baseline by a large margin (see Figure 3), except when  $\{CIFAR-100\} \in \mathcal{S}_{ID}$  while  $\{CIFAR-10, LSUN\} \in \mathcal{S}_{OOD}$ . This observation is interesting since it suggests that the value of auxiliary information from  $\mathcal{S}_{OOD}$  might be degrading when  $\mathcal{S}_{ID} \supseteq \mathcal{S}_{OOD}$ . With the term superset  $\mathcal{S}_{ID} \supseteq \mathcal{S}_{OOD}$  we refer to the fact that the ID data  $\mathcal{S}_{ID}$  might represent a broader set of features compared to OOD data  $\mathcal{S}_{OOD}$ . Thus, training a model with a small subset of the ID data as OOD might not be beneficial since no additional information is presented to the model because the features from  $\mathcal{S}_{ID}$  and  $\mathcal{S}_{OOD}$  conflict with each other. Another factor that impairs OOD detection is the presence of label noise [27], (e.g.  $\{CIFAR-10 \text{ vs. } CIFAR-100\}$ ), which has



been identified with the term *near-OOD vs. far-OOD* in subsequent work [45]. A natural question arising from this observation is whether we can identify the inflection point between label noise and OOD detection?

Table 2: Accuracy of models on ID dataset classification tasks.

Model	CIFAR-10	SVHN	FashionMNIST	CIFAR100
DNN	95.06	96.67	95.27	77.44
DPN	88.10	90.10	93.20	79.34
MC-Dropout	96.22	96.90	95.40	78.39
SWAG	96.53	97.06	93.80	78.61
JEM	92.83	96.13	83.21	77.86
CE+ $\ell_1$	90.66	95.34	93.89	62.30
CE+ $\ell_1$ +MCD	90.33	94.85	91.37	60.35
ContReg (ours)	90.76	95.25	93.68	72.78
ContReg+MCD	90.31	94.75	93.01	64.04
ContRank (ours)	89.01	94.97	93.40	64.32
ContRank+MCD	91.96	82.34	93.13	60.43

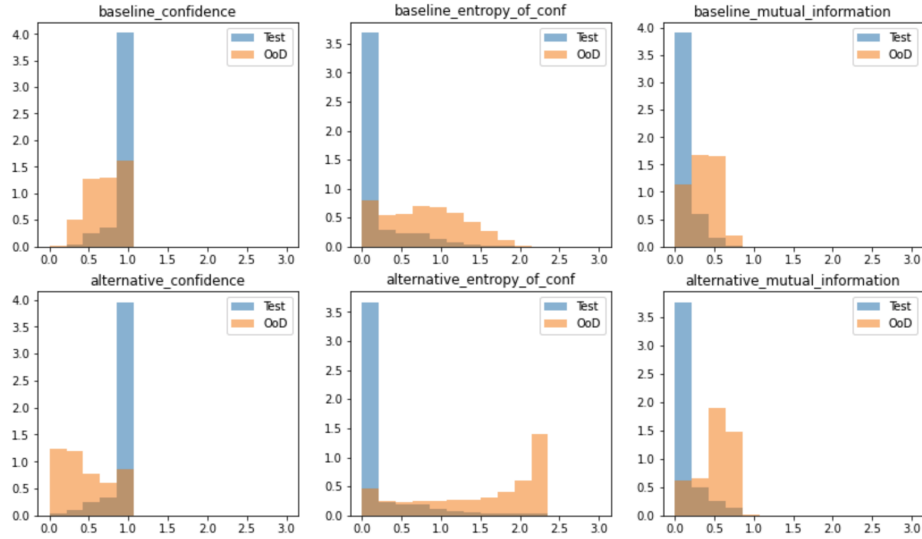


Fig. 3: Comparison of CrossEntropy (baseline) against our proposed objective (i.e. ContReg (see Table 3 and 5) across three metrics  $\{confidence, entropy, mutual\}$  information} with respect to a WideResNet28x10 architecture.

Table 3 compares the performance of the proposed objectives described in this paper with existing OOD detection approaches, where we report confidence-

based AUC-ROC scores matching our experimental setting. Our methods outperform *MSP*, *ODIN*, and *EnergyOOD* (w/o pre-training), and provide comparable results with *Mahalanobis*, *MSRep*, and *OE*.

Table 3: Comparison of our methods with related work based on published results in the literature corresponding with our setting.

Data		AUC-ROC scores							
$S_{ID}$	$S_{OOD}$	MSP	Mahalanobis	ODIN	MSRep	OE	EnergyOOD	ContReg(ours)	ContRank(ours)
CIFAR-10	CIFAR-100*	86.15	93.90	85.59	91.23	93.30	92.60	92.23	94.23
	SVHN	89.60	97.62	91.96	99.48	98.40	90.96	99.18	95.40
	LSUN	88.54	96.30	90.35	96.05	97.60	94.24	92.44	94.77
CIFAR-100	CIFAR-10	73.41	81.34	74.54	81.49	75.70	76.61	72.94	68.89
	SVHN*	71.44	86.01	67.26	87.42	86.66	73.99	99.68	99.95
	LSUN	75.38	93.9	78.94	79.05	79.71	79.23	70.50	62.17

From Table 5 we can observe that even though explicit regularisation overall is beneficial compared to no regularisation, on the contrary, stronger regularisation might deteriorate OOD detection. An ongoing inquiry is to formally characterise and identify the necessary and sufficient conditions of regularisation in order to robustify models against ambiguous and corrupted inputs.

To evaluate whether our method is robust against common corruptions we utilised CIFAR10-C and CIFAR100-C. Similar to [12] we report the mean corruption error (mCE) in Table 4, with the exception that we do not adjust for the varying corruption difficulties by dividing the average corruption error with those of a baseline model. Observe that our objective attains the smallest mCE on CIFAR10-C indicating that is indeed robust against common corruptions while on CIFAR100-C cross-entropy with  $\ell_1$ -regularisation attain the smallest error with ours being second best.

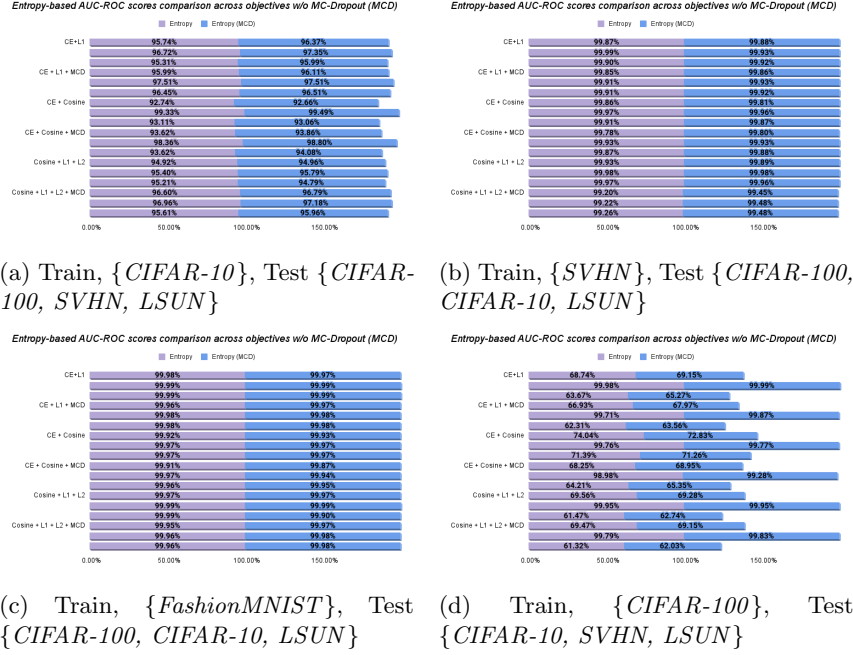


Fig. 4: Comparison of objective functions with and without MC-Dropout during inference. Each histogram corresponds to evaluating a particular loss on a different OOD dataset.

Table 4: Evaluating objective functions across common corruptions against CIFAR10-C and CIFAR100-C measured in average corruption error (mCE).

Objectives	mCE	
	CIFAR10-C	CIFAR100-C
CrossEntropy	161.14	717.04
CrossEntropy+MCD	120.91	536.63
CrossEntropy+ $\ell_1$	144.02	247.78
CrossEntropy+ $\ell_1$ +MCD	140.96	285.04
ContReg (ours)	119.98	337.94
ContReg+MCD	129.20	269.52
ContRank (ours)	149.46	258.73
ContRank+MCD	167.30	306.42

Table 5: Out-of-distribution results. AUC-ROC scores based on entropy. The values in bracket are % improvement of algorithm w.r.t. baseline. An  $\uparrow$  indicates improvement &  $\downarrow$  degradation. The asterisk (\*) indicates auxiliary information during training.

Data	S <sub>UD</sub>	So <sub>CD</sub>	Entropy AUC-ROC score (% gain wrt. baseline)										
			baseline	DNN	DPN	MCD	SWAG	JEM	CE- $t_1$	CE- $t_1$ +MCD	ContReg(ours)	ContReg+MCD	ContRank(ours)
CIFAR-10	SVHN	CIFAR-100*	86.27	85.60 (4.0.78%)	89.92 (4.23%)	91.89 (76.51%)	87.35 (11.25%)	95.74 (10.97%)	95.99 (11.26%)	92.74 (17.50%)	93.62 (18.52%)	94.92 (10.03%)	96.60 (11.97%)
		SVHN	89.72	98.90 (10.23%)	96.25 (17.28%)	98.62 (19.92%)	89.22 (40.56%)	96.72 (17.80%)	97.51 (18.68%)	98.33 (10.71%)	98.36 (19.63%)	95.40 (16.33%)	96.96 (18.07%)
		LSUN	88.83	83.30 (46.23%)	92.04 (13.61%)	95.12 (17.08%)	89.84 (11.14%)	95.31 (17.29%)	96.45 (18.57%)	93.11 (14.82%)	93.62 (15.39%)	95.21 (17.18%)	95.61 (17.63%)
SVHN	CIFAR-10*	CIFAR-100	93.19	99.10 (76.34%)	94.33 (11.22%)	95.97 (12.98%)	92.34 (40.91%)	99.87 (17.17%)	99.85 (17.15%)	99.86 (17.16%)	99.78 (17.07%)	99.93 (17.23%)	99.20 (16.45%)
		CIFAR-10*	94.58	99.60 (10.41%)	94.97 (10.41%)	96.03 (11.53%)	92.85 (11.83%)	99.99 (15.72%)	99.91 (15.64%)	99.97 (15.70%)	99.93 (15.66%)	99.98 (15.71%)	99.22 (14.91%)
		LSUN	92.97	99.70 (17.24%)	93.31 (10.37%)	95.71 (12.95%)	91.82 (11.24%)	99.90 (17.45%)	99.91 (17.46%)	99.91 (17.46%)	99.87 (17.42%)	99.97 (17.53%)	99.26 (16.77%)
FashionMNIST	CIFAR-10*	CIFAR-100	91.20	99.50 (19.10%)	93.75 (12.80%)	96.19 (15.47%)	62.79 (131.15%)	99.98 (19.63%)	99.96 (19.61%)	99.92 (19.56%)	99.91 (19.55%)	99.97 (19.62%)	99.95 (19.59%)
		CIFAR-10*	94.59	99.60 (15.30%)	96.06 (11.55%)	94.28 (10.33%)	64.76 (131.54%)	99.99 (15.71%)	99.98 (15.70%)	99.97 (15.69%)	99.97 (15.69%)	99.99 (15.71%)	99.96 (15.68%)
		LSUN	93.34	99.80 (16.92%)	97.40 (14.35%)	99.05 (16.12%)	65.38 (129.96%)	99.99 (17.12%)	99.98 (17.11%)	99.97 (17.10%)	99.96 (17.09%)	99.99 (17.12%)	99.96 (17.09%)
CIFAR-100	SVHN*	CIFAR-10	78.25	85.15 (18.82%)	80.70 (13.13%)	84.92 (18.52%)	77.64 (40.78%)	68.74 (112.15%)	66.93 (114.46%)	74.04 (15.38%)	68.25 (112.77%)	69.56 (111.10%)	69.47 (111.22%)
		SVHN*	81.52	92.64 (13.64%)	85.59 (14.99%)	94.16 (115.51%)	81.22 (40.37%)	99.98 (122.64%)	99.71 (122.31%)	99.76 (122.37%)	98.98 (121.41%)	99.95 (122.61%)	99.79 (122.41%)
		LSUN	77.22	86.38 (11.86%)	76.58 (40.83%)	87.22 (112.95%)	77.54 (10.41%)	63.67 (117.54%)	62.31 (119.30%)	71.39 (17.54%)	64.21 (116.84%)	61.47 (120.39%)	61.32 (120.59%)
Avg % improvement				(16.48%)	(12.76%)	(16.60%)	(17.96%)	(15.15%)	(14.98%)	(16.26%)	(14.82%)	(14.80%)	(14.90%)

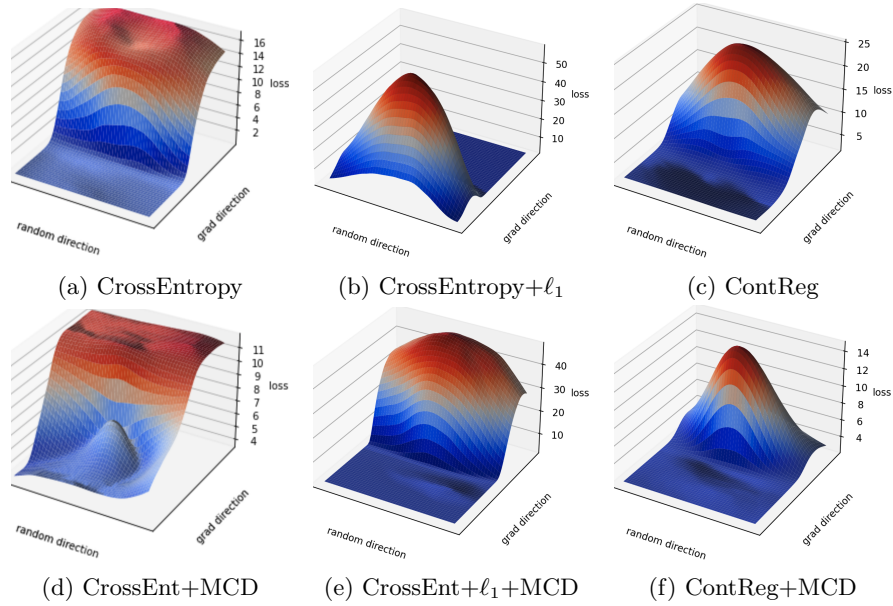


Fig. 5: Comparison of different objectives trained on ID CIFAR-10 and tested on OOD CIFAR-100 with (1st row) and without (2nd row) explicit regularisation (MCD: Monte-Carlo Dropout).

## 5 Conclusion

In this work we presented two novel objective functions with the goal of being utilised in a normal classification setting while at the same time exhibiting some robustness properties against common corruptions and ambiguous inputs when evaluated in OOD detection. We demonstrated that our approach outperforms half of the competitive methods and performs comparably to the remaining ones. Furthermore, we presented the efficacy of our method against common corruptions measured in mCE compared to competitive alternative methods. Finally, we identified the importance of auxiliary information as well as the role of regularisation in OOD detection, followed some important questions in identifying the role of bias in the choice of objective function, family class, and algorithm when considering open set classification problems.

## References

1. Bousquet, O., Boucheron, S., Lugosi, G.: Introduction to statistical learning theory. In: Summer School on Machine Learning. pp. 169–207. Springer (2003)
2. Bulusu, S., Kailkhura, B., Li, B., Varshney, P.K., Song, D.: Anomalous instance detection in deep learning: A survey. In: IEEE Symposium on Security and Privacy. vol. 42 (2020)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (2020)
4. Choi, H., Jang, E., Alemi, A.A.: WAIC, but Why? Generative Ensembles for Robust Anomaly Detection. ArXiv: 1810.01392 (2018)
5. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 1310–1320. PMLR (09–15 Jun 2019)
6. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: International Conference on Machine Learning (2016)
7. Geifman, Y., El-Yaniv, R.: Selectivenet: A deep neural network with an integrated reject option. In: Proceedings of the 36th International Conference on Machine Learning. vol. 97. PMLR (2019)
8. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**(11), 665–673 (2020)
9. Goyal, P., Mahajan, D., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
10. Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Norouzi, M., Swersky, K.: Your classifier is secretly an energy based model and you should treat it like one. arXiv preprint arXiv:1912.03263 (2019)
11. Henaff, O.: Data-efficient image recognition with contrastive predictive coding. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 4182–4192. PMLR (2020)
12. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2019)
13. Hendrycks, D., Gimpel, K.: A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *ICLR* (2017)
14. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. In: International Conference on Learning Representations (2019)
15. Hsu, Y.C., Shen, Y., Jin, H., Kira, Z.: Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
16. Khani, F., Liang, P.: Feature noise induces loss discrepancy across groups. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 5209–5219. PMLR (2020)

17. Khani, F., Liang, P.: Removing spurious features can hurt accuracy and affect groups disproportionately. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, Association for Computing Machinery (2021). <https://doi.org/10.1145/3442188.3445883>
18. Kiefer, J., Wolfowitz, J.: Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics* **23**(3), 462 – 466 (1952). <https://doi.org/10.1214/aoms/1177729392>
19. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *NeurIPS* (2017)
20. Lee, H.B., Lee, H., Na, D., Kim, S., Park, M., Yang, E., Hwang, S.J.: Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks. In: International Conference on Learning Representations (2020)
21. Lee, K., Lee, H., Lee, K., Shin, J.: Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. *arXiv:1711.09325 [cs, stat]* (2017)
22. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: *Advances in Neural Information Processing Systems*. vol. 31. Curran Associates, Inc. (2018)
23. Li, Y., Vasconcelos, N.: Background data resampling for outlier-aware classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
24. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations (2018)
25. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: *Advances in Neural Information Processing Systems*. vol. 33, pp. 21464–21475 (2020)
26. Mendelson, S.: A few notes on statistical learning theory. In: *Advanced lectures on machine learning*, pp. 1–40. Springer (2003)
27. Mitros, J., Pakrashii, A., Mac Namee, B.: Ramifications of Approximate Posterior Inference for Bayesian Deep Learning in Adversarial and Out-of-Distribution Settings, pp. 71–87. *European Conference on Computer Vision*, Springer (01 2020). [https://doi.org/10.1007/978-3-030-66415-2\\_5](https://doi.org/10.1007/978-3-030-66415-2_5)
28. Morningstar, W., Ham, C., Gallagher, A., Lakshminarayanan, B., Alemi, A., Dillon, J.: Density of states estimation for out of distribution detection. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research*, vol. 130. PMLR (2021)
29. Nandy, J., Hsu, W., Lee, M.L.: Towards maximizing the representation gap between in-domain & out-of-distribution examples. In: *Advances in Neural Information Processing Systems*. vol. 33, pp. 9239–9250. Curran Associates, Inc. (2020)
30. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018)
31. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In: *Advances in Neural Information Processing Systems*. vol. 32 (2019)
32. Pang, T., Xu, K., Dong, Y., Du, C., Chen, N., Zhu, J.: Rethinking softmax cross-entropy loss for adversarial robustness. In: International Conference on Learning Representations (2020)
33. Ren, J., Liu, P.J., Fertig, E., Snoek, J., Poplin, R., Deprieto, M., Dillon, J., Lakshminarayanan, B.: Likelihood ratios for out-of-distribution detection. In: *Advances in Neural Information Processing Systems*. pp. 14680–14691 (2019)

34. Rice, L., Wong, E., Kolter, Z.: Overfitting in adversarially robust deep learning. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 8093–8104. PMLR (2020)
35. Robbins, H., Monro, S.: A Stochastic Approximation Method. The Annals of Mathematical Statistics **22**(3), 400 – 407 (1951). <https://doi.org/10.1214/aoms/1177729586>
36. Sagawa\*, S., Koh\*, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks. In: International Conference on Learning Representations (2020)
37. Schulam, P., Saria, S.: Can you trust this prediction? auditing pointwise reliability after learning. In: Proc. of Artificial Intelligence and Statistics (2019)
38. Shafaei, A., Schmidt, M., Little, J.J.: Does Your Model Know the Digit 6 Is Not a Cat? A Less Biased Evaluation of ”Outlier” Detectors. ArXiv: 1809.04729 (2018)
39. Shalev, G., Adi, Y., Keshet, J.: Out-of-distribution detection using multiple semantic label representations. In: Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
40. Tack, J., Mo, S., Jeong, J., Shin, J.: Csi: Novelty detection via contrastive learning on distributionally shifted instances. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33. Curran Associates, Inc. (2020)
41. Vapnik, V.: The nature of statistical learning theory. Springer (2013)
42. Vapnik, V.N.: An overview of statistical learning theory. IEEE transactions on neural networks **10**(5), 988–999 (1999)
43. Von Luxburg, U., Schölkopf, B.: Statistical learning theory: Models, concepts, and results. In: Handbook of the History of Logic, vol. 10, pp. 651–706. Elsevier (2011)
44. Wei, C., Kakade, S., Ma, T.: The implicit and explicit regularization effects of dropout. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 10181–10192. PMLR (2020)
45. Winkens, J., Bunel, R., Roy, A.G., Stanforth, R., Natarajan, V., Ledsam, J.R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., Cemgil, A.T., Eslami, S.M.A., Ronneberger, O.: Contrastive training for improved out-of-distribution detection. CoRR **abs/2007.05566** (2020)
46. Yang, Y., Khanna, R., Yu, Y., Gholami, A., Keutzer, K., Gonzalez, J.E., Ramchandran, K., Mahoney, M.W.: Boundary thickness and robustness in learning models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6223–6234. Curran Associates, Inc. (2020)
47. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
48. Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations. vol. 15 (2017)
49. Zisselman, E., Tamar, A.: Deep residual flow for out of distribution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)