International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02029-3

Augmenting the Softmax with Additional Confidence Scores for Improved Selective Classification with Out-of-Distribution Data

Guoxuan Xia[1] · **Christos-Savvas Bouganis**[1]

Received: 6 May 2023 / Accepted: 8 February 2024

Abstract

Detecting out-of-distribution (OOD) data is a task that is receiving an increasing amount of research attention in the domain of deep learning for computer vision. However, the performance of detection methods is generally evaluated on the task in isolation, rather than also considering potential downstream tasks in tandem. In this work, we examine selective classification in the presence of OOD data (SCOD). That is to say, the motivation for detecting OOD samples is to reject them so their impact on the quality of predictions is reduced. We show under this task specification, that existing post-hoc methods perform quite differently compared to when evaluated only on OOD detection. This is because it is no longer an issue to conflate in-distribution (ID) data with OOD data *if the ID data is going to be misclassified*. However, the conflation within ID data of correct and incorrect predictions becomes undesirable. We also propose a novel method for SCOD, Softmax Information Retaining Combination (SIRC), that augments a softmax-based confidence score with a secondary class-agnostic feature-based score. Thus, the ability to identify OOD samples is improved without sacrificing separation between correct and incorrect ID predictions. Experiments on a wide variety of ImageNet-scale datasets and convolutional neural network architectures show that SIRC is able to consistently match or outperform the baseline for SCOD, whilst existing OOD detection methods fail to do so. Interestingly, we find that the secondary scores investigated for SIRC do not consistently improve performance on all tested OOD datasets. To address this issue, we further extend SIRC to incorporate multiple secondary scores (SIRC+). This further improves SCOD performance, both generally, and in terms of consistency over diverse distribution shifts. Code is available at https://github.com/Guoxoug/SIRC.

Keywords Deep learning · Uncertainty estimation · Out-of-distribution data · Selective classification

1 Introduction

Out-of-distribution (OOD) detection (Yang et al., 2021), i.e. identifying input data samples that do not belong to the distribution that a model was trained on, is a task that is receiving an increasing amount of attention in the domain of deep learning (Liang et al., 2018; Liu et al., 2020b; Du et al., 2022; Hendrycks & Gimpel, 2017; Hendrycks & Dietterich, 2019; Fort et al., 2021; Hsu et al., 2020; Techapanurak et al., 2020; Sun et al., 2021; Sun et al., 2022; Wang et al., 2022; Huang & Li, 2021; Lee et al., 2018; Pearce et al., 2021; Yang et al., 2021; Zhang et al., 2021; Nalisnick et al., 2019). The task is often motivated by safety-critical applications of deep learning, such as healthcare and autonomous driving. For these scenarios, there may be a large cost associated with sending a prediction on OOD data downstream. For example, it could be potentially dangerous for a self-driving car to unknowingly classify a grizzly bear as one of the classes in its training set.[1]

However, in spite of a plethora of existing research, there is generally a lack of focus with regards to the specific motivation behind OOD detection in the literature, other than it is often performed as part of the pipeline of another primary task, e.g. image classification. As such OOD detection tends to be evaluated in isolation, formulated as binary classification between in-distribution (ID) and OODdata.

Communicated by Lei Wang.

✉ Guoxuan Xia
g.xia21@imperial.ac.uk

Christos-Savvas Bouganis
christos-savvas.bouganis@imperial.ac.uk

[1] Department of Electrical and Electronic Engineering, Imperial College London, Exhibition Road, London SW7 2BX, UK

[1] In this work OOD data is defined as being disjoint from the label space of the training distribution (Yang et al., 2021).

Published online: 23 April 2024

In this work, we consider the question *why exactly do we want to do OOD detection during deployment?* We focus on the problem setting where the primary objective is classification, and we are motivated to detect and then reject OOD data, as predictions on those samples will incur a cost. That is to say, the task is selective classification (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017) where OOD data is present within the input samples. Kim et al. (2021) term this problem setting *unknown detection*. However, we prefer to use Selective Classification in the presence of Out-of-Distribution data (SCOD) as we would like to emphasise the downstream classification task as the primary objective and will refer to the task as such in the remainder of this paper.

The *key difference* between this problem setting and OOD detection is that *both* OOD data *and* incorrect predictions on ID data will incur a cost (Kim et al., 2021). It does not matter if we reject an ID sample if it would be incorrectly classified anyway. As such we can view the task as separating correctly predicted ID samples (ID✓) from misclassified ID samples (ID✗) and OOD samples. This reveals a potential blind spot in designing approaches solely for OOD detection, as the cost of ID misclassifications is ignored if the aim is only to separate OOD|ID.

The *key contributions* of this work are:

1. Building on initial results reported by Kim et al. (2021) that show poor SCOD performance for existing methods designed for OOD detection, we show novel insight into the behaviour of different post-hoc (after-training) detection methods for the task of SCOD. Improved OOD detection often comes directly at the expense of SCOD performance, through the conflation of ID✗ and ID✓. Moreover, the relative SCOD performance of different methods varies with the proportion of OOD data found in the test distribution, the relative cost of accepting ID✗ vs OOD, as well as the distribution from which the OOD data samples are drawn.
2. We propose a novel method, targeting SCOD, Softmax Information Retaining Combination (SIRC). Our approach aims to improve the OOD|ID✓ separation of softmax-based confidence scores, by combining them with a secondary, class-agnostic confidence score, whilst retaining their ability to identify ID✗. It consistently outperforms or matches the baseline maximum softmax probability (MSP) approach over a wide variety of OOD datasets and convolutional neural network (CNN) architectures. On the other hand, existing OOD detection methods fail to achieve this.
3. We find that the secondary scores investigated for SIRC perform inconsistently over different OOD datasets. That is to say, a given secondary score may improve SCOD for some OOD datasets, but won't help on other datasets. Also, different scores appear to be better suited to detecting different distribution shifts. Thus, we extend SIRC to incorporate a combination of *multiple* secondary scores (SIRC+). This results in generally even better SCOD performance, as well as more *consistent* performance gains over a wider range of OOD data.

A preliminary version of this work has been published in ACCV 2022 (Xia & Bouganis, 2022a), which covers points {1, 2}. In this work, we extend the aforementioned preliminary version through:

- more detailed discussion of {1, 2},
- the inclusion of an additional secondary confidence score—KNN (Sun et al., 2022),
- evaluation on an additional OOD dataset–SpaceNet (Etten et al., 2018),
- the novel developments described in 3 (SIRC+).

2 Preliminaries

Neural Network Classifier For a K-class classification problem we learn the parameters $\boldsymbol{\theta}$ of a discriminative model $P(y \mid \boldsymbol{x}; \boldsymbol{\theta})$ over labels $y \in \mathcal{Y} = \{\omega_k\}_{k=1}^{K}$ given inputs $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^D$, using finite training dataset $\mathcal{D}_{\text{tr}} = \{y^{(n)}, \boldsymbol{x}^{(n)}\}_{n=1}^{N}$ sampled independently from true joint data distribution $p_{\text{tr}}(y, \boldsymbol{x})$. This is done in order to make predictions $\hat{y}$ given new inputs $\boldsymbol{x}^* \sim p_{\text{tr}}(\boldsymbol{x})$ with unknown labels,

$$\hat{y} = f(\boldsymbol{x}^*) = \arg\max_{\omega} P(\omega \mid \boldsymbol{x}^*; \boldsymbol{\theta}) \,, \tag{1}$$

where f refers to the classifier function. In our case, the parameters $\boldsymbol{\theta}$ belong to a deep neural network with categorical softmax output $\boldsymbol{\pi} \in [0, 1]^K$,

$$P(\omega_i \mid \boldsymbol{x}; \boldsymbol{\theta}) = \pi_i(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{\exp v_i(\boldsymbol{x})}{\sum_{k=1}^{K} \exp v_k(\boldsymbol{x})} \,, \tag{2}$$

where the logits $\boldsymbol{v} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}$ $(\in \mathbb{R}^K)$ are the output of the final fully-connected layer with weights $\boldsymbol{W} \in \mathbb{R}^{K \times L}$, bias $\boldsymbol{b} \in \mathbb{R}^K$, and final hidden layer features $\boldsymbol{z} \in \mathbb{R}^L$ as inputs. Typically $\boldsymbol{\theta}$ are learnt by minimising the cross entropy loss, such that the model approximates the true conditional distribution $P_{\text{tr}}(y \mid \boldsymbol{x})$,

$$\begin{aligned}
\mathcal{L}_{\text{CE}}(\boldsymbol{\theta}) &= -\frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K} \delta(y^{(n)}, \omega_k) \log P(\omega_k \mid \boldsymbol{x}^{(n)}; \boldsymbol{\theta}) \\
&\approx -\mathbb{E}_{p_{\text{tr}}(\boldsymbol{x})} \left[\sum_{k=1}^{K} P_{\text{tr}}(\omega_k \mid \boldsymbol{x}) \log P(\omega_k \mid \boldsymbol{x}; \boldsymbol{\theta}) \right] \\
&= \mathbb{E}_{p_{\text{tr}}(\boldsymbol{x})} \left[\text{KL} \left[P_{\text{tr}}(\omega_k \mid \boldsymbol{x}) \,||\, P(\omega_k \mid \boldsymbol{x}; \boldsymbol{\theta}) \right] \right] + A,
\end{aligned} \tag{3}$$

where $\delta(\cdot,\cdot)$ is the Kronecker delta, A is a constant with respect to $\boldsymbol{\theta}$ and $\mathrm{KL}[\cdot\|\cdot]$ is the Kullback–Leibler divergence.

Selective Classification A selective classifier (El-Yaniv & Wiener, 2010) can be formulated as a pair of functions, the aforementioned classifier $f(\boldsymbol{x})$ [in our case given by Eq. (1)] that produces a prediction $\hat{y}$, and a binary rejection function

$$g(\boldsymbol{x};t) = \begin{cases} 0 \text{ (reject prediction)}, & \text{if } S(\boldsymbol{x}) < t \\ 1 \text{ (accept prediction)}, & \text{if } S(\boldsymbol{x}) \geq t \end{cases}, \tag{4}$$

where t is an operating threshold and S is a scoring function which is typically a measure of predictive confidence (or $-S$ measures uncertainty). Intuitively, a selective classifier chooses to reject if it is uncertain about a prediction.

2.1 Problem Setting Selective Classification with OOD Data (SCOD)

We consider a scenario where, during deployment, classifier inputs $\boldsymbol{x}^*$ may be drawn from either the training distribution $p_{\text{tr}}(\boldsymbol{x})$ (ID) or another distribution $p_{\text{OOD}}(\boldsymbol{x})$ (OOD). That is to say,

$$\begin{aligned} &\boldsymbol{x}^* \sim p_{\text{mix}}(\boldsymbol{x}) \\ &p_{\text{mix}}(\boldsymbol{x}) = \alpha\, p_{\text{tr}}(\boldsymbol{x}) + (1-\alpha)\, p_{\text{OOD}}(\boldsymbol{x}) \,, \end{aligned} \tag{5}$$

where $\alpha \in [0,1]$ reflects the proportion of ID to OOD data found in the wild. Here "Out-of-Distribution" inputs are defined as those drawn from a distribution with label space that does not intersect with the training label space $\mathcal{Y}$ (Yang et al., 2021). For example, an image of a car is considered OOD for a CNN classifier trained to discriminate between different types of pets. We use this definition as it means that OOD samples are fundamentally incompatible with the primary classifier, and any classification predictions made on them will be automatically invalid. Note that in our case we assume no knowledge of p_{OOD} before deployment.

We now define the predictive loss on an accepted sample as

$$\mathcal{L}_{\text{pred}}(f(\boldsymbol{x})) = \begin{cases} 0, & \text{if } f(\boldsymbol{x}) = y,\ (y,\boldsymbol{x}) \sim p_{\text{tr}} \\ \beta, & \text{if } f(\boldsymbol{x}) \neq y,\ (y,\boldsymbol{x}) \sim p_{\text{tr}} \\ 1-\beta, & \text{if } \boldsymbol{x} \sim p_{\text{OOD}} \end{cases} \tag{6}$$

for classifier $f(\boldsymbol{x})$ [Eq. (1)], where $\beta \in [0,1]$. We define the selective risk as in (Geifman & El-Yaniv, 2017),

$$R(f,g;t) = \frac{\mathbb{E}_{p_{\text{mix}}(\boldsymbol{x})}[g(\boldsymbol{x};t)\mathcal{L}_{\text{pred}}(f(\boldsymbol{x}))]}{\mathbb{E}_{p_{\text{mix}}(\boldsymbol{x})}[g(\boldsymbol{x};t)]}\,, \tag{7}$$

which can be intuitively understood as the average loss of only the accepted samples, when using rejection function $g(\boldsymbol{x};t)$ [Eq. (4)]. We are only concerned with the relative cost of ID✗ and OOD samples, so we use a single parameter β.

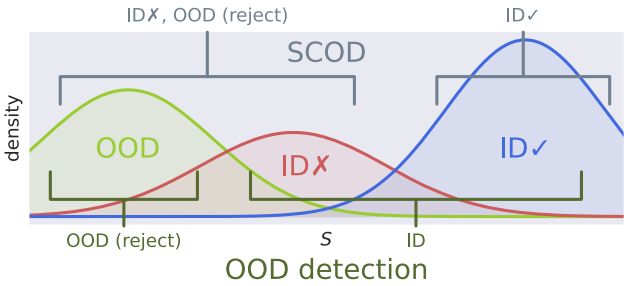

Fig. 1 Illustrative sketch showing how SCOD differs to OOD detection. Densities of OOD samples, misclassifications (ID✗) and correct predictions (ID✓) are shown with respect to confidence score S. For OOD detection the aim is to separate OOD|ID✗ID✓, whilst for SCOD the data is grouped as OODID✗|ID✓

The objective is to find a classifier and rejection function (f,g) that minimise $R(f,g;t)$ for some given setting of t. We focus on comparing post-hoc (after-training) methods in this work, where g (or equivalently S) is varied with f fixed. This removes confounding factors that may arise from the interactions of different training-based and post-hoc methods, as they can often be freely combined. In practice, both α and β will depend on the deployment scenario. However, whilst β can be set freely by the practitioner depending on their own evaluation of costs, α is outside of the practitioner's control and their knowledge of it is likely to be very limited.

It is worth contrasting the SCOD problem setting with OOD detection. SCOD aims to separate OOD, ID✗ |ID✓, whilst for OOD detection the data is grouped as OOD|ID✗, ID✓ (see Fig. 1). The key difference is in the categorisation of ID✗.

SCOD and Types of Uncertainty We note that previous work (Kendall & Gal, 2017; Malinin & Gales, 2018; Malinin et al., 2020; Mukhoti et al., 2021; Pearce et al., 2021) refer to different types of predictive uncertainty, namely *aleatoric* and *epistemic*. The former arises from uncertainty inherent in the data (i.e. the true conditional distribution $P_{\text{tr}}(y \mid \boldsymbol{x})$) and as such is irreducible, whilst the latter can be reduced by having the model learn from additional data. Typically, it is argued that it is useful to distinguish these types of uncertainty at prediction time. Epistemic uncertainty estimates should indicate distributional shift away from the training distribution, i.e. whether a test input $\boldsymbol{x}^*$ is OOD. On the other hand, aleatoric uncertainty estimates should reflect the level of class ambiguity of an ID input. An interesting result within our problem setting is that the conflation of these different types of uncertainties may not be an issue, as there is no need to separate ID✗ from OOD, as both should be rejected.

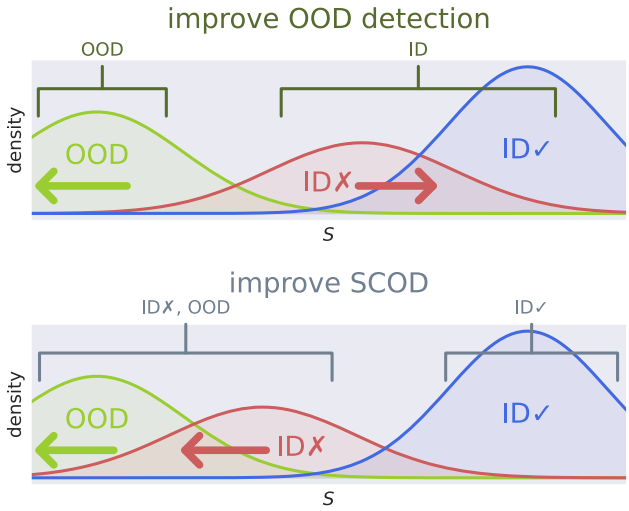

Fig. 2 Illustrations of how a detection method can improve over a baseline. Top: for OOD detection we can either have OOD further away from ID✓ or ID✗ closer to ID✓. Bottom: for SCOD we want both OOD and ID✗ to be further away from ID✓. Thus, we can see how improving OOD detection may in fact be at odds with SCOD

3 Existing OOD Detectors Applied to SCOD

As the explicit objective of OOD detection is different to SCOD, it is of interest to understand how existing detection methods behave for SCOD. Previous work (Kim et al., 2021) has empirically shown that some existing OOD detection approaches don't perform very well, and in this section we shed additional light as to why this is the case.

Improving Performance: OOD Detection vs SCOD In order to build an intuition, we can consider, qualitatively, how detection methods can improve performance over a baseline, with respect to the distributions of OOD and ID✗ relative to ID✓. This is illustrated in Fig. 2.

- For OOD detection the objective is to better separate the distributions of ID and OOD data. Thus, we can either find a confidence score S that, compared to the baseline, has OOD distributed further away from ID✓, and/or has ID✗ distributed closer to ID✓.
- For improving SCOD, we want *both* OOD and ID✗ to be distributed further away from ID✓ than the baseline.

Thus there is a *conflict between the two tasks*. For the distribution of ID✗, the desired behaviour of confidence score S will be different.

Existing Approaches Sacrifice SCOD by Conflating ID✗ and ID✓

Considering post-hoc methods, the generally accepted baseline approach for both selective classification and OOD detection is the Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2017; Geifman & El-Yaniv, 2017) confidence score. Improvements in OOD detection are often achieved by moving away from the softmax $\boldsymbol{\pi}$ in order to better capture the differences between ID and OOD data. Confidence scores such as Energy (Liu et al., 2020b) and Max Logit (Hendrycks et al., 2022) consider the logits $\boldsymbol{v}$ directly, whereas the Mahalanobis detector (Lee et al., 2018) and DDU (Mukhoti et al., 2021) build generative models using Gaussians over the features $\boldsymbol{z}$. ViM (Wang et al., 2022) and Gradnorm (Huang et al., 2021) incorporate class-agnostic, feature-based information into their scores.

Recall that typically a neural network classifier learns a model $P(y \mid \boldsymbol{x}; \boldsymbol{\theta})$ to approximate the true conditional distribution $P_{\text{tr}}(y \mid \boldsymbol{x})$ of the training data [Eqs. (2) and (3), Sect. 2]. As such, scores S extracted from the softmax outputs $\boldsymbol{\pi}$ should best reflect how likely a classifier prediction on ID data is going to be correct or not (and this is indeed the case in our experiments in Sect. 5). As the above (post-hoc) OOD detection approaches all involve moving away from the modelled $P(y \mid \boldsymbol{x}; \boldsymbol{\theta})$, we would expect worse separation between ID✗ and ID✓ even if overall OOD is better distinguished from ID.

Figure 3 shows empirically how well different types of data are separated using MSP ($\pi_{\max}$) and Energy ($\log \sum_k \exp v_k$), by plotting false positive rate (FPR) against true positive rate (TPR). Lower FPR indicates better separation of the negative class away from the positive class.

Although Energy has better OOD detection performance compared to MSP, this is actually because the separation between ID✗ and ID✓ is much less for Energy, so ID as a whole is better separated from OOD. On the other hand the behaviour of OOD relative to ID✓ is not meaningfully different to the MSP baseline. Therefore, SCOD performance for Energy is worse in this case. Another way of looking at it would be that for OOD detection, MSP does worse as it conflates ID with OOD. However, this doesn't harm SCOD performance as much, as those ID samples that are confused with OOD are mostly incorrect anyway. The ID dataset is ImageNet-200 (Kim et al., 2021), OOD dataset is iNaturalist (Huang & Li, 2021) and the model is ResNet-50 (He et al., 2016).

4 Targeting SCOD—Retaining Softmax Information

We would now like to develop an approach that is tailored to the task of SCOD. We have discussed how we expect softmax-based methods, such as MSP, to perform best for distinguishing ID✗ from ID✓, and how existing approaches for OOD detection improve over the baseline, in part, by sacrificing this. As such, to improve over the baseline for SCOD, we will aim to *retain* the ability to separate ID✗ from ID✓ whilst *increasing* the separation between OOD and ID✓.

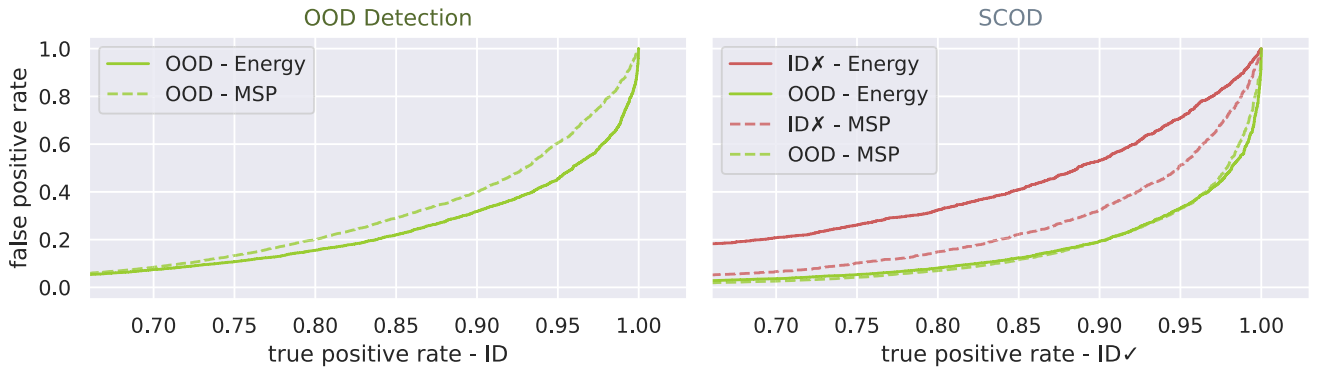

Fig. 3 Left: false positive rate (FPR↓) of OOD samples (negative class) plotted against true positive rate (TPR) of ID (✓+✗) samples (positive class), i.e. how well each confidence score distinguishes OOD from ID. Energy performs better (lower) for OOD detection relative to the MSP baseline. Right: FPR↓ of ID✗ and OOD samples (negative classes) against TPR of ID✓ (positive class). Energy is worse than the baseline at separating ID✗|ID✓ and no better for OOD|ID✓, meaning it is worse for SCOD. Energy's improved OOD detection arises from pushing ID✗ closer to ID✓ (Fig. 2). The ID dataset is ImageNet-200, OOD data is iNaturalist and the model is ResNet-50

Combining Confidence Scores Inspired by Gradnorm (Huang et al., 2021) and ViM (Wang et al., 2022) we consider the combination of two different confidence scores S_1, S_2. We shall consider S_1 our primary score, which we wish to augment by incorporating S_2. For S_1 we investigate scores that are strong for selective classification on ID data, but are also capable of detecting OOD data—MSP and (the negative of) softmax entropy, $(-)\mathcal{H}[\boldsymbol{\pi}]$. For S_2, the score should be useful *in addition* to S_1 in determining whether data is OOD or not. We should consider *scores that capture different information about OOD data* to the post-softmax S_1 if we want to improve OOD|ID✓. We choose to examine the l_1-norm of the feature vector $\|\boldsymbol{z}\|_1$ (Huang et al., 2021), the negative of the Residual[2] score $-\|\boldsymbol{z}^{P^\perp}\|_2$ (Wang et al., 2022) and the negative of the k-th nearest neighbour distance[3] (KNN) (Sun et al., 2022). These scores were chosen as they capture *class-agnostic* information at the feature level. Note that although $\|\boldsymbol{z}\|_1$, Residual and KNN have previously been shown to be useful for OOD detection (Huang et al., 2021; Wang et al., 2022; Sun et al., 2022), we do not expect them to be useful for identifying misclassifications. They are separate from the classification layer defined by $(\boldsymbol{W}, \boldsymbol{b})$, so they are far removed from the categorical $P(y \mid \boldsymbol{x}; \boldsymbol{\theta})$ explicitly modelled by the softmax.

Softmax Information Retaining Combination (SIRC) We want to create a combined confidence score $C(S_1, S_2)$ that retains S_1's ability to distinguish ID✗ |ID✓ but is also able to incorporate S_2 in order to augment OOD|ID✓. We develop our approach based on the following set of *assumptions* about the behaviour of S_1 and S_2:

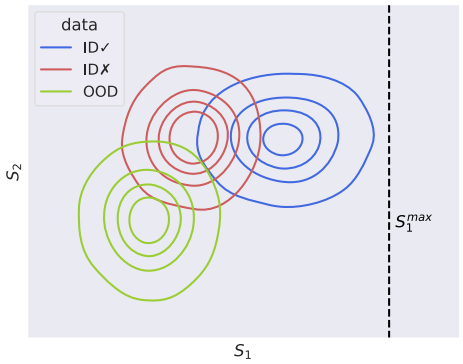

Fig. 4 Illustration on the (S_1, S_2)-plane that satisfies the assumptions behind SIRC. (1) S_1 is higher for ID✓ and lower for ID✗ and OOD. (2) S_1 has maximum value $S_1^{\max}$. (3) S_2 is not useful for ID✗ |ID✓ but is lower for OOD. (4) S_2 is useful in addition to S_1 for detecting OOD

- S_1 will be higher for ID✓ and lower for ID✗ and OOD.
- S_1 is bounded by maximum value $S_1^{\max}$.[4]
- S_2 is unable to distinguish ID✗ |ID✓ well, but is lower for OOD compared to ID.
- S_2 is useful *in addition* to S_1 for separating OOD|ID.

These assumptions are illustrated roughly in Fig. 4. We expect our choices of S_1 (MSP, $-\mathcal{H}$) and S_2 ($\|\boldsymbol{z}\|_1$, Res., KNN) to conform to these assumptions for the reasons stated earlier. Moreover, future choices of confidence score should conform as well.

[2] $\boldsymbol{z}^{P^\perp}$ is the component of the feature vector that lies outside of a principle subspace calculated using ID data. For more details see Wang et al. (2022)'s paper.

[3] This is the Euclidean distance between a test feature vector and its k-th nearest neighbour from an ID dataset. Both features are l_2-normalised. For details see Sun et al. (2022)'s paper.

[4] This holds for our chosen S_1 of $\pi_{\max}$ and $-\mathcal{H}$.

Given the aforementioned assumptions, we propose to combine S_1 and S_2 using

$$C(S_1, S_2) = -(S_1^{\max} - S_1)\,(1 + \exp(-b[S_2 - a]))\,, \tag{8}$$

or equivalently taking logs,[5]

$$C(S_1, S_2) = -\log(S_1^{\max} - S_1) - \log\,(1 + \exp(-b[S_2 - a]))\ , \tag{9}$$

where a, b are parameters chosen by the practitioner. The idea is for the accept/reject decision boundary of C to be in the shape of a sigmoid on the (S_1, S_2)-plane (see Figs. 5, 6). As such the behaviour of only using the softmax-based S_1 is recovered for ID✗ |ID✓ for high S_2, as the decision boundary tends to a vertical line. However, C becomes increasingly sensitive to S_2 as S_2 decreases, and less sensitive to S_1 as S_1 decreases (Fig. 5). This allows for improved OOD|ID✓ as S_2 is "activated" towards the bottom left of the (S_1, S_2)-plane. We term this approach *Softmax Information Retaining Combination (SIRC)*.

The parameters a, b allow the method to be adjusted to different distributional properties of S_2. Rearranging Eq. (8),

$$S_1 = S_1^{\max} + C/[1 + \exp(-b[S_2 - a])]\,, \tag{10}$$

we see that a controls the placement of the sigmoid with respect to S_2, and b the sensitivity of the sigmoid to S_2. Figure 5 shows that the sensitivity of SIRC to S_2 (gradient) increases from zero as S_2 approaches a from above, and then tends to a linear relationship (constant sensitivity proportional to b).

We use the empirical mean and standard deviation of S_2, μ_{S_2}, σ_{S_2} on ID data (training or validation) to set the parameters. We choose $a = \mu_{S_2} - 3\sigma_{S_2}$ so the centre of the sigmoid is below the ID distribution of S_2, and we set $b = 1/\sigma_{S_2}$, to match the ID variations of S_2. We find the above approach to be empirically effective, however, other parameter settings are of course possible. Practitioners are free to tune a, b however they see fit. This may be done using only ID data (training or validation) as we have, or by additionally using synthetic validation OOD data (Hendrycks et al., 2019; Sun et al., 2022).

SIRC Compared to Other Combination Approaches Fig. 6 compares different methods of combination by plotting ID✓, ID✗ and OOD data densities on the (S_1, S_2)-plane. Other than SIRC we consider the combination methods used in ViM, $C = S_1 + cS_2$, where c is a user set parameter, and in Gradnorm, $C = S_1 S_2$. The overlayed contours of C represent decision boundaries for values of t [Eq. (4)].

We see that the linear decision boundary of $C = S_1 + cS_2$ must trade-off significant performance in ID✗ |ID✓ in order to gain OOD|ID✓ (through varying c), whilst $C = S_1 S_2$ sacrifices the ability to separate ID✗ |ID✓ well for higher values of S_1. We also note that $C = S_1 S_2$ is not robust to different ID means of S_2. For example, arbitrarily adding a constant D to S_2 will completely change the behaviour of the combined score. On the other hand, SIRC is designed to be robust to this sort of variation between different S_2. Figure 6 also shows an alternative parameter setting for SIRC, where a is lower and b is higher. The sigmoid is shifted down and steeper. Here more of the behaviour of only using S_1 is preserved, but S_2 contributes less. It is also empirically observable that the assumption that S_2 (in this case $\|z\|_1$) is not useful for distinguishing ID✓ from ID✗ holds, and in practice this can be verified on ID validation data when selecting S_2.

We also note that although we have chosen specific S_1, S_2 in this work, SIRC can be applied to any S that satisfy the above assumptions. It is a combination method, rather than a specific confidence score. As such it has the potential to improve beyond the results we present, especially given the rapid pace of development of new confidence scores for uncertainty estimation.

Limitations We note that one limitation of SIRC is that it does not aim to improve ID✗ |ID✓, only OOD|ID✓. Moreoever, although the approach aims to limit this effect, we expect inevitable minor degradation in ID✗ |ID✓ as a result of the inclusion of S_2.

5 Experimental Results—SIRC

We present experiments across a range of CNN architectures and ImageNet-scale OOD datasets. Extended results can be found in Appendix B.

Data For our ID dataset we use ImageNet-200 (Kim et al., 2021), which contains a subset of 200 ImageNet-1k (Russakovsky et al., 2015) classes. It has separate training, validation and test sets. We use a variety of OOD datasets for our evaluation that display a wide range of semantics and difficulty in being identified. Near-ImageNet-200 (Near-IN-200) (Kim et al., 2021) is constructed from remaining ImageNet-1k classes semantically similar to ImageNet-200, so it is especially challenging to detect. Caltech-45 (Kim et al., 2021) is a subset of the Caltech-256 (Griffin et al., 2007) dataset with non-overlapping classes to ImageNet-200. Openimage-O (Wang et al., 2022) is a subset of the Open Images V3 (Krasin et al., 2017) dataset selected to be OOD with respect to ImageNet-1k. iNaturalist (Huang & Li, 2021) and Textures (Wang et al., 2022) are the same

[5] For SCOD, we are only concerned with the rank ordering of confidence scores, and log is a monotonic function. This version is more numerically stable. We implemented it using the `logaddexp` function in PyTorch (Paszke et al., 2019).

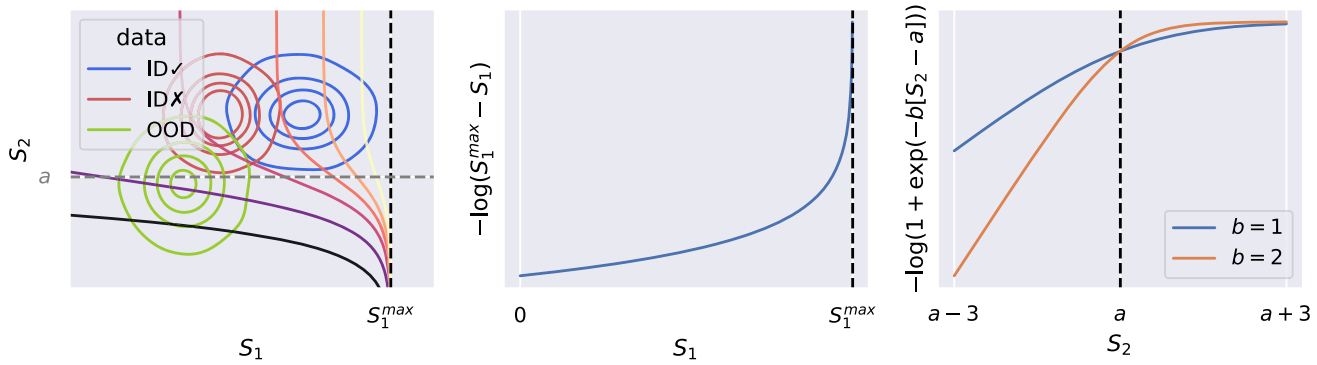

Fig. 5 Left: SIRC isocontours on the (S_1, S_2)-plane—they are sigmoids. Centre: Plot of how the first term in SIRC [Eq. (9)] varies with S_1 – its sensitivity to S_1 (gradient) is high close to $S_1^{\max}$ and gradually decreases with S_1. Right: Plot of how the second term in SIRC varies with S_2—its sensitivity increases from zero as S_2 approaches a from above, eventually tending to a linear relationship proportional to b

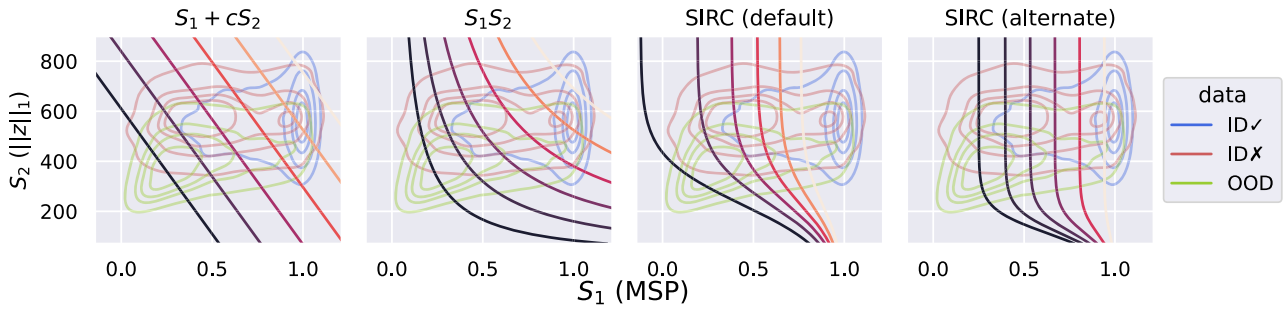

Fig. 6 Comparison of different methods of combining confidence scores S_1, S_2 for SCOD. OOD, ID✗ and ID✓ distributions are displayed using kernel density estimate contours. Graded contours for the different combination methods are then overlayed (lighter means higher combined score). We see that our method, SIRC (centre right) is able to better retain ID✗|ID✓ whilst improving OOD|ID✓. An alternate parameter setting for SIRC, with a stricter adherence to S_1, is also shown (far right). The ID dataset is ImageNet-200, the OOD dataset iNaturalist and the model ResNet-50. SIRC parameters are found using ID training data; the plotted distributions are test data

for their respective datasets (Van Horn et al., 2017; Cimpoi et al., 2014). SpaceNet (Etten et al., 2018) contains satellite images of Rio De Janeiro. Colorectal (Kather et al., 2016) is a collection of histological images of human colorectal cancer, whilst Colonoscopy is a dataset of frames taken from colonoscopic video of gastrointestinal lesions (Mesejo et al., 2016). Noise is a dataset of square images where the resolution, contrast and pixel values are randomly generated (for details see Appendix A.2). Finally, ImageNet-O (Hendrycks et al., 2021) is a dataset OOD to ImageNet-1k that is adversarially constructed using a trained ResNet. Note that we exclude a number of OOD datasets from Kim et al. (2021) and Huang and Li (2021) as a result of discovering samples within said datasets that match ID labels.

Models and Training We train ResNet-50 (He et al., 2016), DenseNet-121 (Huang et al., 2017) and MobileNetV2 (Sandler et al., 2018) using hyperparameters based around standard ImageNet settings.[6] Full training details can be found in Appendix A.1. For each architecture, we train 5 models independently using random seeds $\{1, \ldots, 5\}$ and report the mean result over the runs. Appendix B additionally contains results on single pre-trained ImageNet-1k models, BiT ResNetV2-101 (Kolesnikov et al., 2020) and PyTorch DenseNet-121.

Detection Methods for SCOD We consider six variations of SIRC using the components $\{\text{MSP}, \mathcal{H}\} \times \{\|z\|_1, \text{Residual}, \text{KNN}\}$, as well as the components individually. We additionally evaluate various existing post-hoc methods: MSP (Hendrycks & Gimpel, 2017), Energy (Liu et al., 2020b), ViM (Wang et al., 2022) and Gradnorm (Huang et al., 2021). For the Residual score (used in SIRC and ViM) we use the full ID ImageNet-200 train set to determine parameters. For KNN we sample 12,500 feature vectors from the training set and use $k = 10$. Results for additional approaches, as well as further details pertaining to the methods, can be found in Appendices B and A.3.

[6] https://github.com/pytorch/examples/blob/main/imagenet/main.py.

5.1 Evaluation Metrics

For evaluating different scoring functions S for the SCOD problem setting we consider a number of metrics. Arrows (↑↓) indicate whether higher/lower is better (For graphical illustrations and additional metrics see Appendix A.4).

Area Under the Risk-Recall curve (AURR)↓ We consider how empirical risk [Eq. (7)] varies with recall of ID✓, and aggregate performance over different t by calculating the area under the curve. As recall is only measured over ID✓, the base accuracy of f is not properly taken into account. Thus, this metric is only suitable for comparing different g with f fixed. To give an illustrative example, a f, g pair where the classifier f is only able to produce a single correct prediction will have perfect AURR as long as S assigns that correct prediction the highest confidence (lowest uncertainty) score. Note that results for the AURC metric (Kim et al., 2021; Geifman et al., 2019) can be found in Appendix B, although we omit them from the main paper as they are not notably different to AURR.

Risk@Recall=0.95 (Risk@95)↓ Since a rejection threshold t must be selected at deployment, we also consider a particular setting of t such that 95% of ID✓ is recalled. In practice, the corresponding value of t could be found on a labelled ID validation set before deployment, without the use of any OOD data. It is worth noting that differences tend to be greater for this metric between different S as it operates around the tail of the positive class.

Area Under the ROC Curve (AUROC)↑ Since we are interested in rejecting both ID✗ and OOD, we can consider ID✓ as the positive class, and ID✗, OOD as separate negative classes. Then we can evaluate the AUROC of OOD|ID✓ and ID✗ |ID✓ independently. The AUROC for a specific value of α would then be a weighted average of the two different AUROCs. This is not a direct measure of risk, but does measure the separation between different empirical distributions. Note that due to similar reasons to AURR this method is only valid for fixed f.

False Positive Rate@Recall=0.95 (FPR@95)↓ FPR@95 is similar to AUROC, but is taken at a specific t. It measures the proportion of the negative class accepted when the recall of the positive class (or true positive rate) is 0.95.

5.2 Separation of ID✗ |ID✓ and OOD|ID✓ Independently

Table 1 shows %AUROC and %FPR@0.95 with ID✓ as the positive class and ID✗, OOD independently as different negative classes (see Sect. 5.1). It is important for a confidence score to have strong ID✗ |ID✓ performance as ID✗ *will always be present*[7] regardless of the volume or type of OOD data. It is also important for a confidence score to perform consistently over different OOD data, as we assume no knowledge at the time of deployment of what distribution shifts may occur.

In general, we see that SIRC, compared to S_1, is able to improve OOD|ID✓ whilst incurring only a small ($<$ 0.2%AUROC) reduction in the ability to distinguish ID✗ |ID✓, across all 3 architectures. On the other hand, non-softmax methods designed for OOD detection show poor ability to identify ID✗, with performance ranging from $\sim$ 8 worse %AUROC than MSP to $\sim$ 50% AUROC (random guessing). Furthermore, they cannot consistently outperform the baseline when separating OOD|ID✓, in line with the discussion in Sect. 3.

We note that in some cases SIRC slightly improves ID✗ |ID✓, however, the impact is minimal and inconsistent over model architectures and S_2. We provide some additional empirical analysis in Appendix B.1.1.

SIRC is Robust to Weak S_2 Although for the majority of OOD datasets in Table 1 SIRC is able to outperform S_1, this is not always the case. When SIRC does not provide a boost over S_1, we can see that S_2 individually is not useful for OOD|ID✓. For example, for ResNet-50 on Colonoscopy, Residual performs *worse* than random guessing. However, in cases like this the performance is still close to that of S_1. As S_2 will tend to be higher for these OOD datasets, the behaviour of SIRC is similar to that of for ID✗ |ID✓, with the decision boundaries close to vertical (see Figs. 5, 8). As such SIRC is *robust* to S_2 performing poorly, but is able to improve on S_1 when S_2 is of use. In comparison, ViM, which linearly combines Energy and Residual, is more sensitive to when the latter stumbles. This is shown in Fig. 8. On iNaturalist ViM has $\sim$ 25 worse %FPR@95 compared to Energy, whereas SIRC ($-\mathcal{H}$, Res.) loses $<$ 0.5% compared to $-\mathcal{H}$. Note that the issue of S_2 being inconsistent is addressed in Sect. 6, where we further extend SIRC.

We additionally remark that regardless of the choice of S_2, there is little to no improvement for Near-ImageNet-200. This suggests that softmax-based scores are best suited to capturing this type of distributional shift. For Near-ImageNet-200 the semantic shift from ImageNet-200 is purposely very small (e.g. "cricket" vs "grasshopper"), and there is no higher level overarching shift (e.g. photographs vs cartoons).

OOD Detection Methods are Inconsistent Over Different Data In Table 1 the performance of existing methods for OOD detection relative to the MSP baseline varies considerably from dataset to dataset. This is directly illustrated in Fig. 7. Even though ViM is able to perform very well on Textures, Noise and ImageNet-O (>50 better %FPR@95 on Noise), it does worse than the baseline on many other OOD datasets (>20 worse %FPR@95 for Near-ImageNet-200 and iNaturalist). This suggests that the inductive biases incor-

[7] Assuming $<$ 100% test accuracy of course.

Table 1 %AUROC and %FPR@95 with ID✓ as the positive class, considering ID✗ and each OOD dataset separately. Full results are for ResNet-50 trained on ImageNet-200

Model	Method	ID✗		OOD mean		Near-IN-200		Caltech-45		Openimage-O		iNaturalist	
		AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓
ResNet-50 ID %Error: 19.01 SIRC	(MSP, $\|z\|_1$)	90.34	52.70	92.00	38.36	85.56	59.76	91.36	41.44	92.28	41.36	94.80	29.60
	(MSP, Res.)	**90.43**	**52.10**	92.94	33.57	85.52	60.03	91.19	42.27	92.57	39.95	94.10	33.55
	(MSP, KNN)	90.39	52.27	93.03	32.63	85.61	59.39	91.64	39.37	92.64	39.14	94.53	30.99
	($-\mathcal{H}$, $\|z\|_1$)	90.00	54.26	92.74	33.73	85.88	58.50	92.19	36.08	92.87	37.83	**95.38**	**25.09**
	($-\mathcal{H}$, Res.)	90.13	54.01	**93.76**	28.42	85.85	58.93	92.11	36.76	**93.25**	**36.36**	94.82	28.51
	($-\mathcal{H}$, KNN)	90.03	54.21	93.74	**28.30**	85.90	58.56	**92.44**	**34.67**	93.19	36.56	95.11	26.91
	MSP	90.41	52.13	91.50	41.30	85.59	59.74	91.13	42.72	91.95	43.55	94.23	33.21
	$-\mathcal{H}$	90.07	54.05	92.33	36.06	**85.91**	**58.47**	92.01	37.20	92.59	40.10	94.90	28.01
	$\|z\|_1$	48.06	94.70	79.19	57.98	52.27	94.58	70.28	77.83	72.23	71.51	85.65	49.50
	Residual	47.59	96.45	58.43	79.71	44.30	96.79	47.76	94.83	59.65	86.85	40.07	97.32
	KNN	68.60	88.28	90.22	39.36	68.27	88.57	86.40	60.71	86.50	57.16	84.85	66.11
	Energy	82.05	69.79	92.70	32.64	81.96	68.70	92.15	38.62	90.92	46.28	94.13	31.70
	Gradnorm	60.17	87.88	86.17	42.57	62.90	86.89	81.11	59.23	81.09	57.80	91.00	34.46
	ViM	80.62	78.13	92.87	35.65	78.90	80.30	90.54	54.70	91.87	43.84	90.13	56.97

Model	Method	Textures		SpaceNet		Colonoscopy		Colorectal		Noise		ImageNet-O	
		AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓
ResNet-50 ID %Error: 19.01 SIRC	(MSP, $\|z\|_1$)	93.64	32.02	96.41	21.19	95.93	25.33	95.84	24.39	90.72	49.63	83.44	58.91
	(MSP, Res.)	96.00	19.81	96.35	20.80	95.52	27.31	95.32	26.97	98.21	10.97	84.62	53.99
	(MSP, KNN)	95.72	20.32	97.11	16.65	96.33	22.44	96.84	17.87	95.63	25.80	84.28	54.32
	($-\mathcal{H}$, $\|z\|_1$)	94.38	27.38	97.26	14.69	96.97	16.87	96.71	18.71	91.74	45.84	84.01	56.34
	($-\mathcal{H}$, Res.)	96.68	15.70	97.36	13.70	96.72	18.10	96.41	20.42	99.02	4.89	85.33	50.81
	($-\mathcal{H}$, KNN)	96.30	16.87	97.84	11.39	97.27	14.85	97.54	13.72	96.92	17.47	84.88	51.99
	MSP	92.88	36.61	95.94	23.74	95.75	26.52	94.86	30.28	89.33	56.83	83.29	59.78
	$-\mathcal{H}$	93.77	30.79	96.95	16.43	96.87	17.55	95.93	23.43	90.47	51.63	83.89	57.02
	$\|z\|_1$	88.90	39.67	87.93	51.46	76.97	82.24	97.28	14.64	97.36	13.51	63.00	84.82
	Residual	82.84	46.63	58.32	86.43	38.09	99.64	53.93	88.78	91.31	20.92	68.04	78.98
	KNN	98.13	9.23	97.02	17.94	94.80	38.33	**99.11**	**2.85**	99.68	1.18	87.40	51.50
	Energy	95.37	22.50	**98.45**	**8.49**	**97.51**	**14.19**	99.07	5.00	94.93	29.05	82.52	61.86

Table 1 continued

Model	Method	Textures		SpaceNet		Colonoscopy		Colorectal		Noise		ImageNet-O	
		AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓
	Gradnorm	93.00	26.57	94.76	26.01	90.54	42.85	98.98	4.98	97.59	13.05	70.78	73.88
	ViM	**98.46**	**7.62**	97.63	13.19	94.42	44.55	98.04	8.84	**99.82**	**0.31**	**88.85**	**46.15**

Model	Method	ID✗		OOD mean		Model	Method	ID✗		OOD mean	
		AUROC↑	FPR@95↓	AUROC↑	FPR@95↓			AUROC↑	FPR@95↓	AUROC↑	FPR@95↓
MobileNetV2 ID %Error: 21.35 SIRC	(MSP, $\|z\|_1$)	89.52	55.52	92.57	33.77	DenseNet-121 ID %Error: 17.20 SIRC	(MSP, $\|z\|_1$)	90.22	52.33	92.15	37.19
	(MSP, Res.)	**89.67**	**55.04**	92.05	37.54		(MSP, Res.)	90.20	52.44	93.20	31.38
	(MSP, KNN)	89.60	55.36	92.85	32.35		(MSP, KNN)	90.31	52.10	92.98	32.21
	($-\mathcal{H}$, $\|z\|_1$)	88.89	58.75	93.26	30.89		($-\mathcal{H}$, $\|z\|_1$)	89.95	53.90	92.90	31.20
	($-\mathcal{H}$, Res.)	89.11	57.94	93.02	32.95		($-\mathcal{H}$, Res.)	89.92	54.15	**93.85**	**26.51**
	($-\mathcal{H}$, KNN)	88.94	58.62	**93.49**	**30.11**		($-\mathcal{H}$, KNN)	90.01	53.82	93.62	27.56
	MSP	89.63	55.10	91.85	38.60		MSP	**90.31**	**51.83**	91.91	38.79
	$-\mathcal{H}$	89.01	58.48	92.72	34.62		$-\mathcal{H}$	90.04	53.35	92.72	32.66
	$\|z\|_1$	53.57	93.40	82.20	51.92		$\|z\|_1$	36.87	98.70	63.78	79.71
	Residual	42.00	97.30	40.43	94.42		Residual	46.07	95.47	70.49	71.40
	KNN	68.15	88.89	87.63	45.55		KNN	71.83	86.41	90.10	43.62
	Energy	81.86	68.01	92.29	34.31		Energy	82.12	66.57	91.54	36.74
	Gradnorm	65.27	85.73	88.13	38.66		Gradnorm	50.18	95.19	76.84	61.85
	ViM	80.20	74.37	89.81	51.43		ViM	76.63	84.78	91.14	42.06

We show abridged results for MobileNetV2 and DenseNet-121. Bold indicates best performance, underline 2nd or 3rd best and we show the mean over models from 5 independent training runs. Variants of SIRC are shown as tuples of their components (S_1, S_2). We also show error rate on ID data. SIRC is able to consistently match or improve over S_1 for OOD|ID✓, at a negligible cost to ID✗ |ID✓. Existing OOD detection methods are significantly worse for ID✗ |ID✓ and inconsistent at improving OOD|ID✓

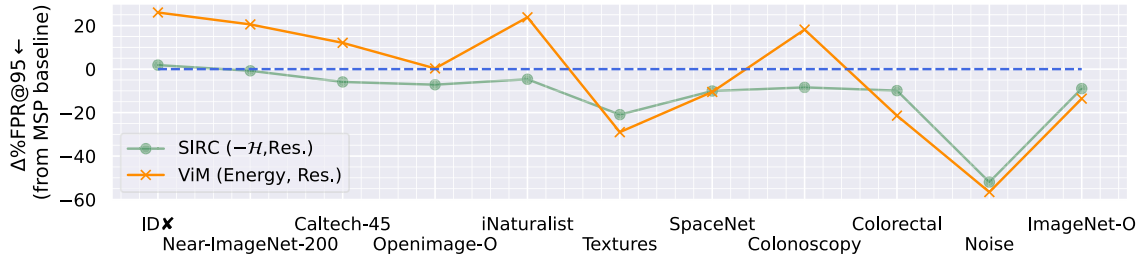

Fig. 7 The change in %FPR@95↓ relative to the MSP baseline across different OOD datasets. SIRC is able to *consistently* match or improve over the baseline, whilst ViM is inconsistent depending on the OOD dataset

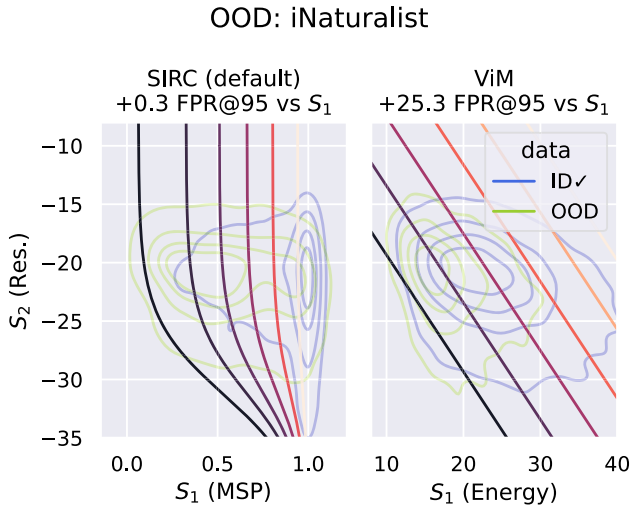

Fig. 8 Comparison (similar to Fig. 5) between SIRC and ViM with OOD data iNaturalist. For this OOD dataset S_2 =Res. cannot distinguish OOD|ID✓. SIRC mostly ignores S_2 in this case (close-to-vertical decision boundaries) leading to performance very close to S_1. On the other hand, ViM incurs a large penalty in FPR@95 from relying on S_2. The ID dataset is ImageNet-200 and the model is ResNet-50

porated, and assumptions made, when designing existing OOD detection methods may prevent them from generalising across a wider variety of OOD data. This behaviour is problematic as we assume no knowledge of the OOD data prior to deployment. In this case, a practitioner may be "unlucky" with the OOD data encountered and incur significant additional loss for choosing ViM over MSP.

In contrast, SIRC more *consistently*, albeit modestly, improves over the baseline (Fig. 7), due to its aforementioned *robustness*. These results suggest that methods designed to deal with OOD data should be evaluated on benchmarks that represent a wider range of distributional shifts than what is currently commonly found in the literature.

5.3 Varying the Importance of OOD Data Through α and β

At deployment, there will be a specific ratio of ID:OOD data exposed to the model. Thus, it is of interest to investigate the risk over different values of α (Eq. 5). Similarly, an incorrect ID prediction may or may not be more costly than a prediction on OOD data so we investigate different values of β (Eq. 6). Figure 9 shows how AURR and Risk@95 are affected as α and β are varied independently (with the other fixed to 0.5). We use the full test set of ImageNet-200, and pool OOD datasets together and uniformly sample different quantities of data randomly in order to achieve different values of α. We use 3 different groupings of OOD data: All, "Close" {Near-ImageNet-200, Caltech-45, Openimage-O, iNaturalist} and "Far" {Textures, SpaceNet, Colonoscopy, Colorectal, Noise}. These groupings are based on relative qualitative semantic difference to the ID dataset (see Appendix A.2 for example images from each dataset). Although the grouping is not formal, it serves to illustrate OOD-data-dependent differences in SCOD performance.

Relative Performance of Methods Changes with α and β At high α and β, where ID✘ dominates the risk, the MSP baseline performs best. However, as α and β are decreased, and OOD data is introduced, we see that other methods improve relative to the baseline. There may be a *crossover* after which the ability to better distinguish OOD|ID✓ allows a method to surpass the baseline. Thus, which method to choose for deployment will depend on the practitioner's setting of β and (if they have any knowledge of it at all) of α.

SIRC Most Consistently Improves Over the Baseline SIRC ($-\mathcal{H}$, Res.) is able to outperform the baseline most consistently over the different scenarios and settings of α, β, only doing worse for ID✘ dominated cases (α, β close to 1). This is because SIRC has close to baseline ID✘ |ID✓ performance and is superior for OOD|ID✓ (Table 1). In comparison, ViM and Energy, which conflate ID✘ and ID✓, are often worse than the baseline for most (if not all) values of α, β. Their behaviour on the different groupings of data illustrates how these methods may be biased towards different OOD datasets, as they significantly outperform the baseline at lower α for the "Far" grouping, but always do worse on "Close" OOD data.

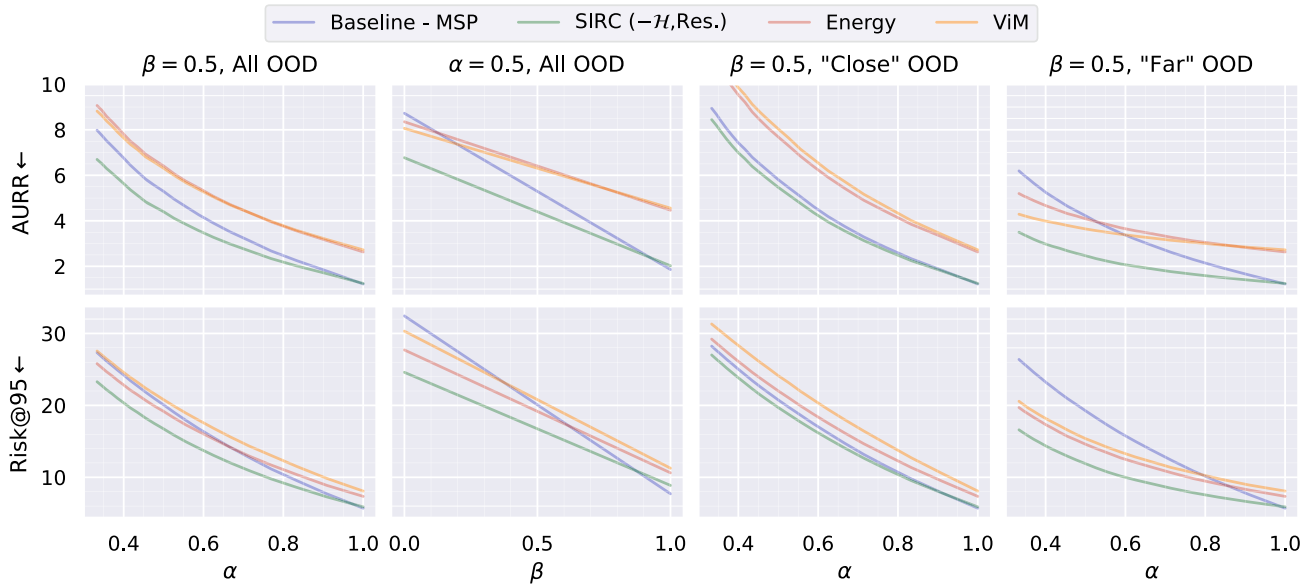

Fig. 9 AURR↓ and Risk@95↓ ($\times 10^2$) for different methods as α and β vary [Eqs. (5), (6)] on a mixture of all the OOD data. We also split the OOD data into qualitatively "Close" and "Far" subsets (Sect. 5.3). For high α, β, where ID✗ dominates in the risk, the MSP baseline is the best. As α, β decrease, increasing the effect of OOD data, other methods improve relative to the baseline. SIRC is able to *most consistently* improve over the baseline. OOD detection methods perform better on "Far" OOD. The ID dataset is ImageNet-200, and the model is ResNet-50. We show the mean over 5 independent training runs. We multiply all values by 10^2 for readability

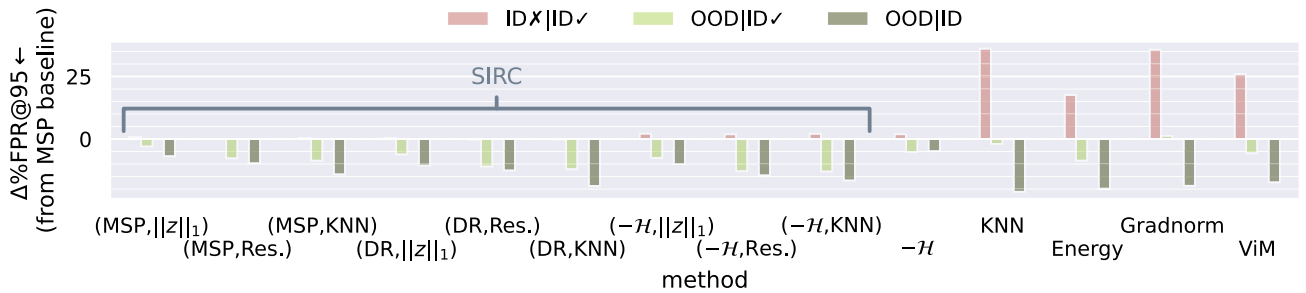

Fig. 10 The change in %FPR@95↓ relative to the MSP baseline of different methods. Different data classes are shown negative|positive. Although OOD detection methods are able to improve OOD|ID, they do so mainly at the expense of ID✗|ID✓ rather than improving OOD|ID✓. SIRC is able to improve OOD|ID✓ with minimal loss to ID✗|ID✓, alongside modest improvements for OOD|ID. Results for OOD are averaged over all OOD datasets. The ID dataset is ImageNet-200 and the model is ResNet-50

5.4 Comparison Between SCOD and OOD Detection

Figure 10 shows the difference in %FPR@95 relative to the MSP baseline for different combinations of negative|positive data classes (ID✗|ID✓, OOD|ID✓, OOD|ID), where OOD results are averaged over all datasets and training runs. In line with the discussion in Sect. 3, we observe that the non-softmax OOD detection methods are able to improve over the baseline for OOD|ID. However, this comes at the cost of significantly degraded ID✗|ID✓, with only small improvements in OOD|ID✓. Thus their SCOD performance is poor compared to the MSP baseline. SIRC on the other hand is able to retain much more ID✗|ID✓ performance whilst improving on OOD|ID✓, allowing it to have better OOD detection *and* SCOD performance compared to the baseline.

6 Extending SIRC—Improving Performance over Diverse Distribution Shifts

A salient result from the previous section is that for certain OOD datasets, certain S_2 fail to improve the OOD|ID✓ performance of SIRC compared to S_1 by itself (e.g. Residual on iNaturalist in Table 1). SIRC is robust to scenarios where S_2 fails, as its behaviour defaults to being similar to only using S_1 (Sect. 5.2). However, ideally we want performance

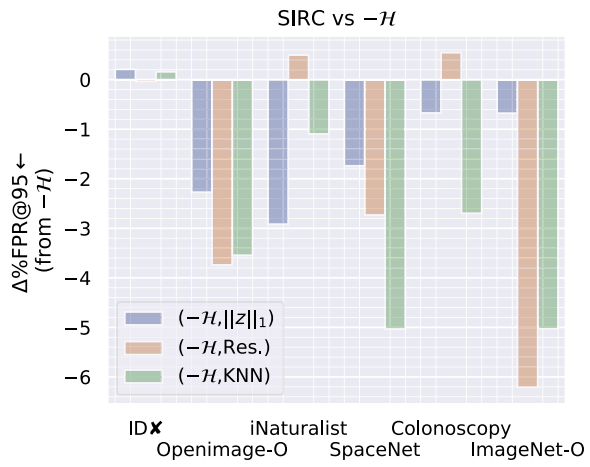

Fig. 11 Comparison of SIRC performance (ΔFPR@95$\downarrow$) compared to only using $S_1(-\mathcal{H})$, over a range of different OOD datasets and S_2. Performance improvements are inconsistent over different distributional shifts (e.g. Residual does not contribute at all for iNaturalist). Moreover, different S_2 seem better suited to different OOD datasets, with no single score being best in all cases. The ID dataset is ImageNet-200, and the model is ResNet-50

improvements over as wide a range of distribution shifts as possible. Furthermore, it appears that different S_2 are better suited for different OOD datasets, so there is not necessarily a "best overall choice" for S_2. This is further illustrated in Fig. 11, which shows the improvement of SIRC vs only S_1 for different S_2 and OOD datasets.

Additionally, each choice of secondary score $\{\|z\|_1$, Residual, KNN$\}$ captures information about distributional shift in a different way. This suggest that by choosing only one, we are leaving information that could be used to further improve SCOD performance on the table. Consequently, we suggest an extension to SIRC, in order to:

1. improve the *consistency* of performance over a wider range of distribution shifts
2. generally boost SCOD performance.

Using Multiple Secondary Scores Given we have access to a selection of options to use as S_2, a natural question to ask is, *can we combine the information from multiple secondary scores*, in order to achieve the above aims? We propose to extend Eq. 8,

$$C(S_1,\dots,S_M) = -(S_1^{\max} - S_1)\prod_{m=2}^{M}[1+\exp(-b_m[S_m - a_m])]\,, \quad (11)$$

and the log version Eq. 9,

$$C(S_1,\dots,S_M) = -\log(S_1^{\max} - S_1) + \sum_{m=2}^{M} -\log\left(1+\exp(-b_m[S_m - a_m])\right)\,, \quad (12)$$

to include $M-1$ secondary scores.[8] Fig. 5 can help with intuition for how the different components contribute in Eq. (12). Multiple secondary scores (righthand plot) contribute additively. We refer to this extended version of SIRC as *SIRC+*.

More Consistent Improvements over Different Distribution Shifts By incorporating multiple secondary scores as in Eq. (12), the idea is that only a single secondary score in SIRC+ needs to contribute usefully in order for OOD|ID✓ to improve. As long as a *single* score moves into the "sensitive zone" past or around a (Fig. 5) for OOD data samples, then SCOD should improve compared to only using S_1.

Thus, different secondary scores may be able to compensate for each other's failures, resulting in more consistent improvements in SCOD over different OOD data. We aim to increase the likelihood of SIRC responding to an unknown distribution shift. In a sense, this approach is an attempt to "safeguard" against as wide a range of distribution shifts as possible, where we do not trust any single secondary score to be able to detect all shifts. This is illustrated for the Colonoscopy OOD dataset in Fig. 13. It shows how the additional useful information from the KNN score can be exploited to improve SCOD even if the Residual score fails to distinguish OOD from ID.

Generally Improved OOD|ID✓ Additionally, when multiple secondary scores react to a distribution shift, we intuitively expect the OOD|ID✓ performance of SIRC+ to be better than using the scores individually. If the different secondary scores provide different information about the distribution shift, then they should contribute in a complementary manner, further improving detection. This is illustrated for the Openimage-O OOD dataset in Fig. 12. KNN is lower for OOD *given* the value of Residual is known, meaning it is additionally useful for detection. Figure 13 then shows how SIRC+ is able to utilise the information in both scores together.

Note that by including more secondary scores in SIRC+, we do expect increased degradation in ID✘ |ID✓. Although SIRC is insensitive to secondary scores for ID✘ |ID✓ (for which we do not expect them to contribute useful information), we still expect the (slight) negative effects to add up as M [Eq. (12)] increases.

[8] Note that the parameters a_m, b_m are found on a per-score basis in the same way as described in Sect. 4.

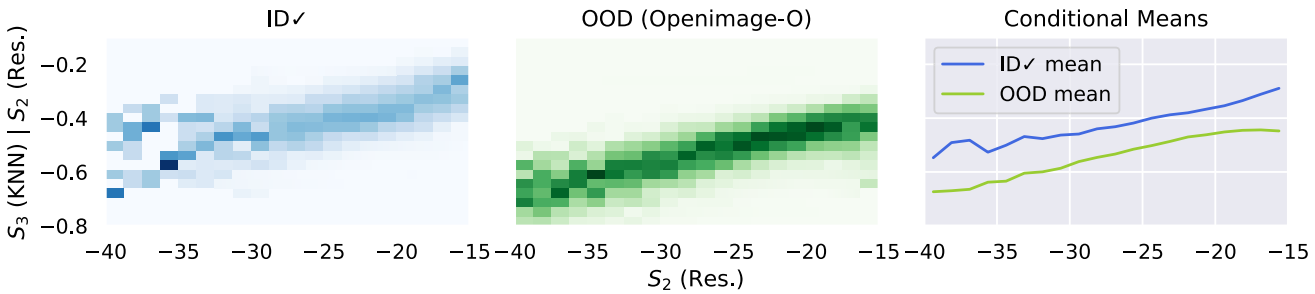

Fig. 12 Conditional plots of KNN given Residual. Left and centre: conditional histograms showing empirical distributions. Right: conditional means. KNN is useful *in addition* to Residual for detecting OOD Openimage-O

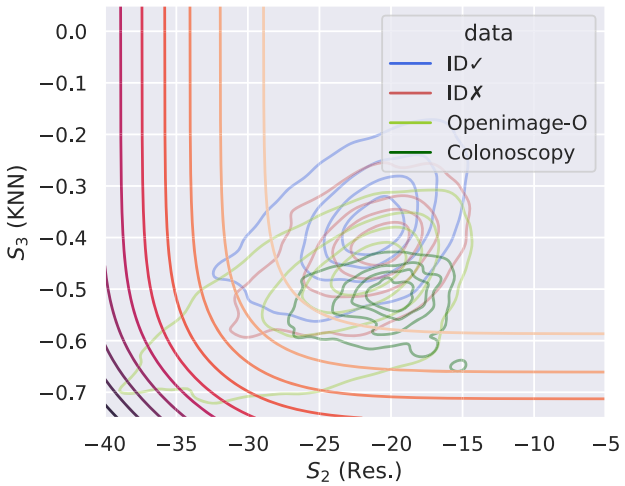

Fig. 13 Visualisation of the combination of multiple secondary scores in SIRC+. OOD, ID✗ and ID✓ distributions are displayed using kernel density estimate contours. Graded contours reflect equidistant values of the second term in Eq. (12). We show two different OOD datasets that illustrate different scenarios. Colonoscopy: Residual is not useful for detecting OOD, but KNN is. By considering *both* scores we are more likely to improve SCOD performance for an unknown distributional shift. Openimage-O: both scores are useful and intuitively capture different information about OOD data. We expect to improve OOD|ID✓ vs using either score individually. The ID dataset is ImageNet-200 and the model is ResNet-50. SIRC parameters are found using ID training data; the plotted distributions are test data

7 Experimental Results—SIRC+

We extend the evaluation in Sect. 5.2, where we consider ID✗|ID✓ and OOD|ID✓ separately, to include SIRC+ where all 3 sary scores are used together ($-\mathcal{H}$, KNN, Res., $\|z\|_1$). Figure 14 shows, for ResNet-50, the difference in SCOD performance between $-\mathcal{H}$ (only using S_1) and different variants of SIRC over the full range of OOD datasets. Full results for other architectures can be found in Appendix B, as well as tables in the format of Table 1 including SIRC+.

SIRC+ Improves over S_1 More Consistently than SIRC Fig. 14 shows that, compared to SIRC with each individual S_2, SIRC+ is able to more consistently boost SCOD performance over the whole range of OOD datasets. For example, for the two OOD datasets iNaturalist and Colonoscopy, SIRC with a single score ($-\mathcal{H}$, Res.) is unable to improve over $-H$. This is because the Residual score fails to recognise samples from these two datasets as OOD. On the other hand, SIRC+ is able to leverage the information in the other two scores KNN and $\|z\|_1$, leading to better SCOD performance, even if the Residual score fails.

SIRC+ Generally Improves SCOD Compared to SIRC For a number of OOD datasets (e.g. Openimage-O), Fig. 14 also shows that SIRC+ is able to achieve better SCOD performance compared to using any of the secondary scores by themselves. This is in line with the discussion in Sect. 6, supporting the idea that even better OOD|ID✓ performance can be achieved by combining multiple secondary scores.

We note that we also observe a slight increase in the degradation of ID✗|ID✓ as expected. However, it is small compared to the improvements in OOD|ID✓, which we believe justifies this trade-off. This is shown in Fig. 15, which reproduces part of Fig. 9 and shows that SIRC+ is able to further improve SCOD over SIRC for the scenarios considered in Sect. 5.3.

8 Related Work

OOD Detection There is extensive existing research into OOD detection, a survey of which can be found in Yang et al. (2021). To improve over the MSP baseline in Hendrycks and Gimpel (2017), early post-hoc approaches, primarily experimenting on CIFAR-scale data, such as ODIN (Liang et al., 2018), Mahalanobis (Lee et al., 2018) and Energy (Liu et al., 2020b) explore how to extract non-softmax information from a trained network. They investigate the use of logits and features, as well as the idea of using input perturbations (inspired by the adversarial attacks literature (Goodfellow et al., 2015)).

More recent work has moved to larger-scale, higher-resolution image datasets (Huang & Li, 2021; Hendrycks et al., 2022; Wang et al., 2022), designed to reflect more real-

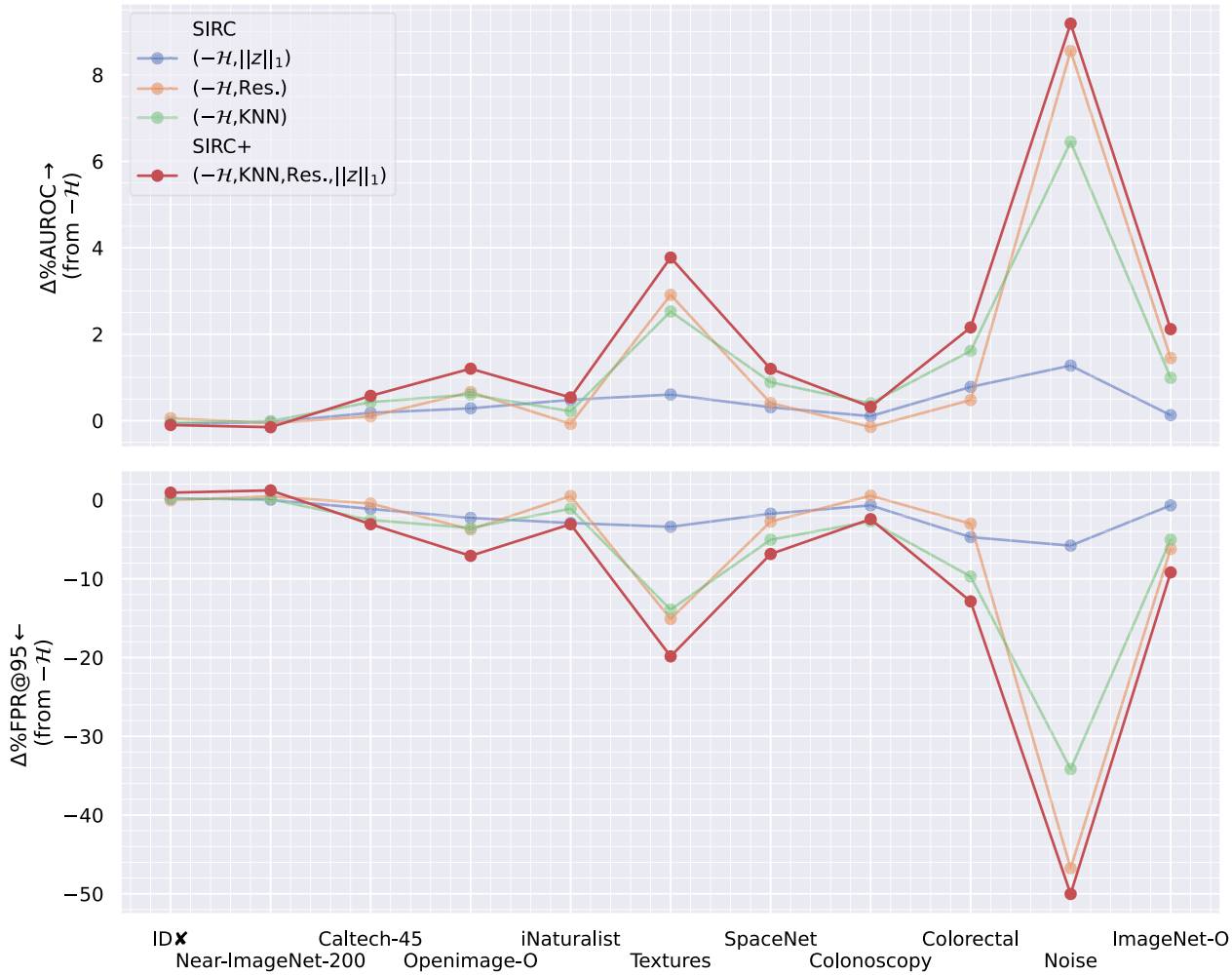

Fig. 14 Δ%AUROC and Δ%FPR@95 (where ID✓ is the positive class) with respect to $-\mathcal{H}$ (S_1 only). Results are for ResNet-50 trained on ImageNet-200. We show the mean over models from 5 independent training runs. SIRC+ is able to provide more consistent improvements over $-\mathcal{H}$ over the different OOD datasets compared to SIRC with a single secondary score. Additionally, on a number of OOD datasets, SIRC+ is able to further improve SCOD performance compared to SIRC

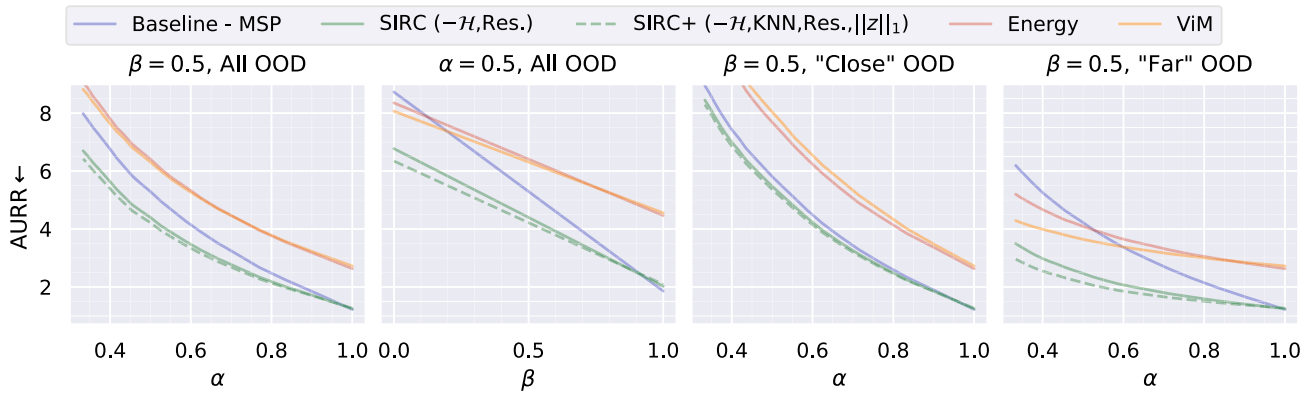

Fig. 15 Part of Fig. 9 reproduced to include SIRC+. SIRC+ is able to further improve SCOD performance compared to SIRC, especially on the "Far" OOD data

istic computer vision applications. Gradnorm (Huang et al., 2021), although motivated by the information in gradients, at its core combines information from the softmax and features together. Similarly, ViM (Wang et al., 2022) linearly combines Energy with the class-agnostic Residual score. ReAct (Sun et al., 2021) aims to improve logit/softmax-based scores by clamping the magnitude of final layer features. KNN (Sun et al., 2022) takes a non-parametric approach, using the distance to the k-th nearest ID neighbour of a test feature vector.

There are also many training-based approaches. Outlier Exposure (Hendrycks et al., 2019) explores training networks to be uncertain on "known" existing OOD data, so that this behaviour generalises to unseen test OOD data. On the other hand VOS (Du et al., 2022) instead generates virtual outliers during training for this purpose. Hsu et al. (2020) and Techapanurak et al. (2020) propose the network explicitly learn a scaling factor for the logits to improve softmax behaviour. There also exists a line of research that explores the use of generative models, $p(\boldsymbol{x};\boldsymbol{\theta})$, for OOD detection (Caterini & Loaiza-Ganem, 2021; Zhang et al., 2021; Ren et al., 2019; Nalisnick et al., 2019). These approaches are separate from classification, however, so are less relevant to this work.

Selective Classification Selective classification, or misclassification detection, has also been investigated for deep learning scenarios. Initially examined in Geifman and El-Yaniv (2017) and Hendrycks and Gimpel (2017), there are a number of approaches to the task that target the classifier f through novel training losses and/or architectural adjustments (Moon et al., 2020; Corbière et al., 2019; Geifman & El-Yaniv, 2019). Post-hoc approaches are fewer. DOCTOR (Granese et al., 2021) provides theoretical justification for using the l_2-norm of the softmax output $\|\boldsymbol{\pi}\|_2$ as a confidence score for detecting misclassifications, however, we find its behaviour similar to MSP and $\mathcal{H}$ (See Appendix B). The comparatively smaller advancement in the selective classification literature, compared to OOD detection, suggests that improving performance on this task is much more challenging. This makes sense given the discussion in Sect. 3. The MSP baseline works well for detecting ID✗ as the softmax directly models $P(y \mid \boldsymbol{x})$, but is inherently ill-suited to OOD detection as it tends to conflate ID✗ with OOD.

General Methods for Uncertainty Estimation There also exist general approaches for uncertainty estimation. These approaches are typically more broadly motivated and aim to improve the quality of uncertainties over a wider range of potential downstream objectives. Earlier methods place neural networks in a Bayesian framework (MacKay, 1995; Jospin et al., 2022), of which a popular and simple-to-implement approach is MC-Dropout (Gal & Ghahramani, 2016). Deep Ensembles (Lakshminarayanan et al., 2017), where multiple models are trained independently using different random seeds, can also be viewed as Bayesian (Wilson, 2020). They offer consistent, and therefore compelling improvements in downstream tasks (Ovadia et al., 2019; Xia & Bouganis, 2022b, 2023; Malinin & Gales, 2021), however, their costs scale linearly with the number of ensemble members. Dirichlet Networks (Malinin & Gales, 2018; Malinin et al., 2020; Ulmer et al., 2023) model a distribution over categorical distributions in order to capture different types of uncertainty. SNGP (Liu et al., 2020a) and DDU (Mukhoti et al., 2021) use spectral normalisation so that shifts in the input space better correspond to shifts in the output space.

Selective Classification with Distribution Shift Here we discuss work that is most closely related to this work (some of which was published *after* the preliminary version of this paper (Xia & Bouganis, 2022a)). Kamath et al. (2020) investigate selective classification under covariate shift for the natural language processing task of question and answering. In the case of *covariate* shift, valid predictions can still be produced on the shifted data, which by our definition is not possible for OOD data (see Sect. 2). Thus the problem setting here is different to our work. They propose that g be a random forest classifier trained on a mixture of ID and covariate-shifted data, after f is fully trained.

Kim et al. (2021) introduce the idea that ID✗ and OOD data should be rejected together and investigate the performance of a range of existing approaches on an image-classification-based benchmark. They examine both training and post-hoc methods (comparing different f and g) on SCOD (which they term unknown detection). They also evaluate performance on misclassification detection and OOD detection independently. They find that Deep Ensembles (Lakshminarayanan et al., 2017) perform best overall. They do not provide a novel approach targeting SCOD, and consider a single setting of (α, β), where the α is not specified and $\beta = 0.5$.

Jaeger et al. (2023) echo a similar sentiment to Kim et al. (2021), presenting a unified evaluation of selective classification with both OOD data and covariate-shifted data for image classification, without presenting a novel approach.

Cen et al. (2023) evaluate the SCOD performance of many approaches under different training regimes. They also propose a SIRC-inspired approach for a "few-shot" problem scenario, where a few OOD samples are available before deployment. They infact benchmark SIRC and report strong results (see their Table 5). We note that whilst both (Cen et al., 2023; Jaeger et al., 2023) are concurrent work to ours, they do not propose any methods that directly compete with SIRC(+), and perform similar classification-based experiments to those in our work [and (Kim et al., 2021)].

9 Future Work

In the future, it would be valuable to explore the ideas in SCOD in problem settings such as Object Detection and Semantic Segmentation that include classification as a sub-

task. These scenarios are more complex compared to our definition of SCOD in Sect. 2 for vanilla classification. For example, in the case of Object Detection with OOD objects (Dhamija et al., 2020; Du et al., 2022), one can imagine a scenario where it is desirable to reject OOD objects as non-objects alongside low-confidence class predictions (just like SCOD), for which a SIRC-like approach may be suitable. However, it may alternatively be desirable to specifically detect OOD objects as unknown objects with a corresponding bounding box, which would require a different style of approach. In the case of semantic segmentation with OOD objects (Hendrycks et al., 2022), there are complications arising from the need to separate uncertainty relating to the edges of objects and uncertainty relating to the overall class of an object. One can easily imagine a SCOD-like problem setting where incorrect pixel predictions on edges would be irrelevant, whereas object-level misclassifications/OOD samples need to be detected.

Additionally, selective prediction for *regression* problems under distributional shift (Malinin & Gales, 2021) is an underexplored problem setting currently. It could also be possible in this case to leverage methods similar to SIRC, that combine multiple confidence scores together.

10 Concluding Remarks

In this work, we consider the performance of existing methods for OOD detection on selective classification in the presence of out-of-distribution data (SCOD). We show how their improved OOD detection vs the MSP baseline often comes at the cost of inferior SCOD performance. Furthermore, we find their performance is inconsistent over different OOD datasets.

In order to improve SCOD performance over the baseline, we develop SIRC. Our approach aims to retain information useful for detecting misclassifications from a softmax-based confidence score, whilst incorporating additional information useful for identifying OOD samples from a secondary score. Experiments show that SIRC is able to consistently match or improve over the baseline approach for a wide range of datasets, CNN architectures and problem scenarios. Moreover, by extending SIRC to include information from multiple secondary scores, we are able to further improve overall SCOD performance, as well as the consistency of SIRC over different distribution shifts.

We hope this work encourages the further investigation of SCOD or other new problem settings that involve detecting or distinguishing distributional shifts during deployment.

Acknowledgements Guoxuan Xia is jointly funded by UK Research and Innovation and Arm ltd.

Data Availability Statement The data analysed in the study are available from the referenced authors. Instructions for obtaining all datasets can be found here: https://github.com/Guoxoug/SIRC.

A Experimental Details

We present detailed information about our experimental setup. Our code is available at https://github.com/Guoxoug/SIRC.

A.1 Models and Training

For the main results we train ResNet-50 (He et al., 2016) using the default hyperparameters found in PyTorch's examples.[9] We train on ImageNet-200 for 90 epochs with a batch size of 256. Stochastic gradient descent is used with a weight decay of 10^{-4}, a momentum of 0.9 and an initial learning rate of 0.1 that steps down by a factor of 10 at epochs 30 and 60. Images are augmented using `RandomResizedCrop` and `RandomHorizontalFlip`. MobileNetV2 (Sandler et al., 2018) uses the same setting, but with an initial learning rate of 0.05. DenseNet-121 is trained with the same settings are ResNet-50 but with Nesterov momentum as per (Huang et al., 2017). We perform 5 independent training runs for each architecture, with random seeds {1, ..., 5}.

Additionally, we also test on two pre-trained ImageNet-1k models. We use ResNetV2-101 from Google's Big Transfer[10] (Kolesnikov et al., 2020), specifically `BiT-S-R101x1`, and DenseNet-121 provided by PyTorch.[11] Note that the BiT model takes 480×480 images as input, whereas all other models take standard ImageNet-scale 224×224 images. Note that for evaluating these models we exclude Near-ImageNet-200 and Caltech-45 due to class overlap with ImageNet-1k.

[9] https://github.com/pytorch/examples/tree/main/imagenet.

[10] https://github.com/google-research/big_transfer.

[11] https://pytorch.org/vision/stable/models.html.

A.2 ImageNet-Scale Datasets

Figure 16 shows a number of random examples from each dataset introduced in Sect. 5, alongside the number of samples in said dataset. Below we describe the methodology for constructing Colonoscopy and Noise. For the remaining datasets please refer to their original papers for details (Huang & Li, 2021; Wang et al., 2022; Kim et al., 2021; Hendrycks et al., 2021; Kather et al., 2016; Etten et al., 2018). We note that there is a slight discrepancy between the number of samples reported in Kim et al. (2021) for ImageNet-200 and in the authors' provided datasets,[12] but we do not believe this affects the validity of our results.

Noise We randomly generate 10,000 square images. All samples are generated independently. Within each image, each value (in space and RGB) is sampled from the same gaussian distribution, with mean 0.5. The standard deviation of said gaussian differs between images. These in turn are generated by sampling from a unit gaussian and squaring the samples. Pixel values are then clipped to be in [0, 1] and mapped to 8-bit integers. The widths of each image are sampled uniformly from {2, ..., 256}, and the images are all scaled to 256×256 using the lanczos interpolation method in PIL.[13] The resulting data thus varies in both scale and contrast (see Fig. 16).

Colonoscopy We separate out frames as individual images from videos provided in Mesejo et al. (2016).[14] We download the first 10 narrow band imaging (NBI) videos in each class of lesion (hyperplasic, serrated, adenoma) and extract each frame as an individual image. Although the data is not independent in this case, we treat it as such for the purposes of our investigation.

A.3 Confidence Scores

Below we detail all confidence scores S implemented and evaluated in our investigation. There are additional approaches that were omitted from the main paper for the sake of brevity.

- SIRC(+): for a description of the score see Sects. 4 and 6 in the main paper. We use the whole of the ImageNet-200 *training* set to determine the values of μ_{S_2}, σ_{S_2}. For ImageNet-1k we randomly sample 250, 000 images from the training set. Note that for all following methods that require ID data to find parameters, we use the same ID data as for SIRC. We investigate combinations of S_1, S_2 from the cartesian product {MSP, DOCTOR, $\mathcal{H}$}$\times${$\|z\|_1$,Residual, KNN}, as well as the use of all secondary scores together for SIRC+.
- Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2017): a baseline score that takes the max value from the softmax $\pi_{\max} = \max_k \pi_k$.
- DOCTOR (Granese et al., 2021): the original paper does not directly present it as such, but the confidence score is equivalent to $\|\boldsymbol{\pi}\|_2$.
- Softmax entropy ($\mathcal{H}$): measures softmax uncertainty, $\mathcal{H}[\boldsymbol{\pi}] = -\sum_k \pi_k \log \pi_k$. We use $S = -\mathcal{H}[\boldsymbol{\pi}]$ to change it to a measure of confidence.
- l_1-norm of the features: used in Gradnorm (Huang et al., 2021), $\|z\|_1$.
- Residual: used in ViM (Huang et al., 2021), this score measures the component of the feature vector that is outside of a principal subspace defined using ID data, $\|z^{P^\perp}\|_2$. We follow Wang et al. (2022) in setting the dimensionality of the subspace to 1000 if the dimensionality of z, $L > 1500$ and 512 otherwise. Like Entropy, we use the negative of the score $S = -\|z^{P^\perp}\|_2$ as this score is meant to be higher for OOD data. Please refer to Wang et al. (2022)'s paper for full details.
- KNN (Sun et al., 2022): a non-parametric approach that uses the Euclidean distance between a test feature vector z^* and its kth nearest neighbour in the training set. Both vectors are L_2-normalised, so this is equivalent to cosine similarity. Sun et al. (2022) subsample the training dataset to reduce search costs at inference, and proportionally scale k. We use a similar-sized training subset of 12, 500 to them for ImageNet-scale data and a value of $k = 10$.
- Max Logit (Hendrycks et al., 2022): Max Logit is similar to MSP, but the score is taken from the logits before the softmax $v_{\max} = \max_k v_k$.
- Energy (Liu et al., 2020b): this score aggregates over all logit values as $\log \sum_k \exp v_k$.
- Gradnorm (Huang et al., 2021): although this score was originally motivated by gradients, we can view it simply as the combination of two scores, $C = \|\boldsymbol{\pi} - \mathbf{1}/K\|_1 \|z\|_1$.
- ViM (Wang et al., 2022): this linearly combines Energy and Residual, $C = \log \sum_k \exp v_k - c\|z^{P^\perp}\|_2$. The parameter c is given by the average value of Max Logit divided by the average value of Residual on ID data, which scales the importance of Residual to be similar to that of Energy in the combination.
- Mahalanobis (Lee et al., 2018): this score involves building a classwise gaussian mixture model over the features with tied covariance matrix. The confidence is then calculated as $-\min_k (z-\boldsymbol{\mu}_k)^T \tilde{\boldsymbol{\Sigma}} (z-\boldsymbol{\mu}_k)$. We use the approach in Wang et al. (2022) and Fort et al. (2021) where only the final layer features are considered.

[12] https://github.com/daintlab/unknown-detection-benchmarks.

[13] https://pillow.readthedocs.io/en/stable/_modules/PIL/Image.html#Image.resize.

[14] http://www.depeca.uah.es/colonoscopy_dataset/.

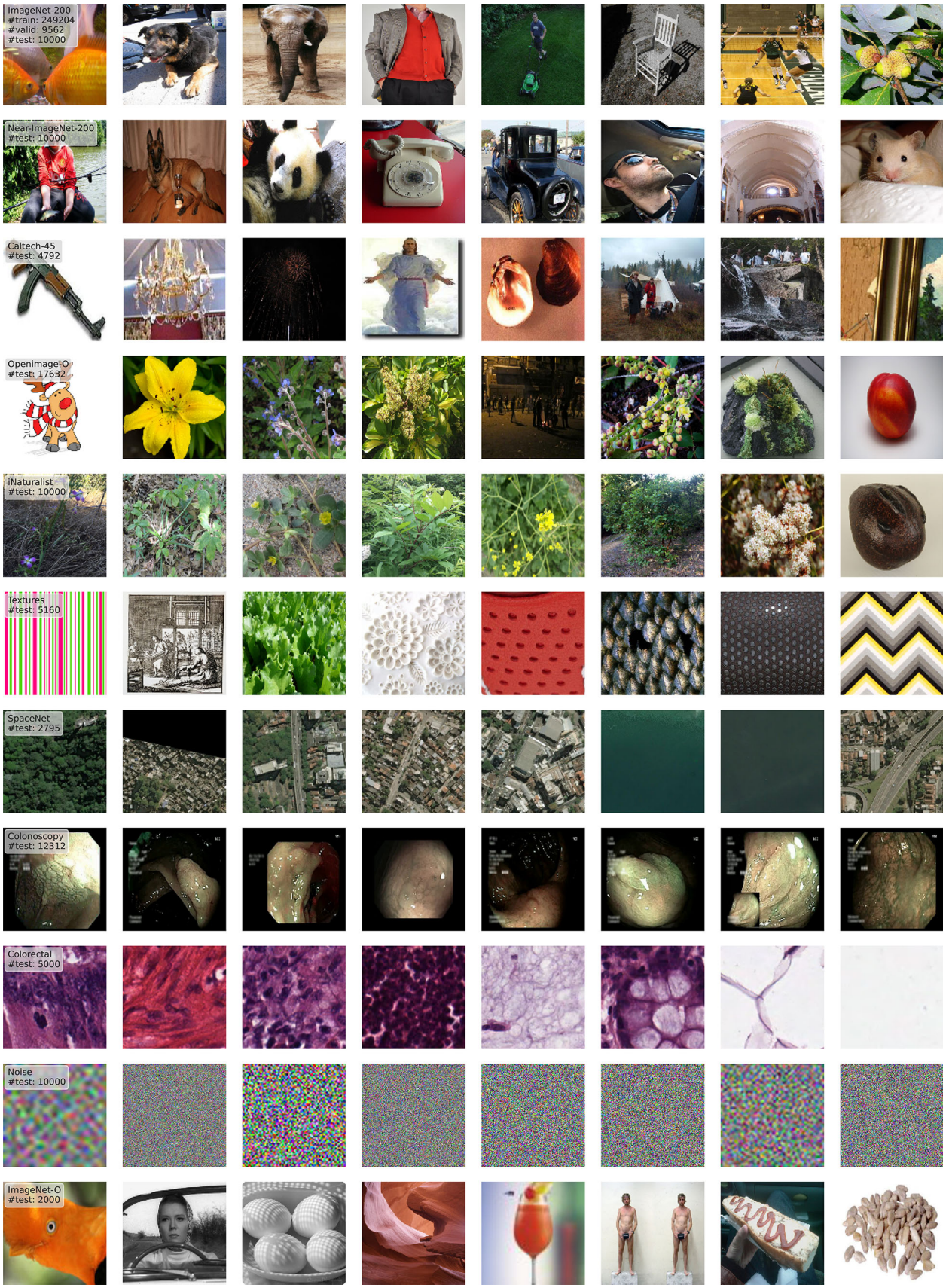

Fig. 16 Random examples from each ImageNet-scale dataset, with the #samples in each

A.4 Evaluation Metrics

Other than the metrics specified in Sect. 5.1, we additionally use Area Under the Risk-Coverage Curve (AURC)↓, from Kim et al. (2021) and Geifman and El-Yaniv (2017). It aggregates risk over all values of *coverage*, which is the proportion of all input data accepted. For AURC their exists an oracle curve, where OOD and ID✗ are perfectly disjoint from ID✓. AURC can be reduced either by lowering the oracle curve by reducing the number of ID✗ (increasing baseline accuracy of f) or by better separating OOD, ID✗ |ID✓ (better choice of g) and so bringing the curve closer to the oracle. Thus the metric is suitable for both training based, and post-hoc approaches. Figure 17 illustrates graphically some of the metrics we use to evaluate SCOD.

B Additional Results

We provide more complete versions of the results presented in Sect. 5 of the main work across all architectures and datasets.

B.1 AUROC and FPR@95

We present results across all post-hoc confidence scores in Appendix A.3 for all architectures in Tables 2, 3, 4 and 5. We also include mean ±2 SD. for experiments with multiple training runs. SIRC performs as expected in all cases – a small reduction in ID✗ |ID✓ in exchange for a meaningful uplift in OOD|ID✓ compared to only using S_1. SIRC+ is able to offer further improvements in OOD|ID✓ over SIRC.

DOCTOR in general performs somewhere in between MSP and $-\mathcal{H}$, both individually and when used in SIRC, so we relegate it to the appendix. We note that Residual and Mahalanobis perform much better only for ResNetV2-101 [these results are inline with Wang et al. (2022)]. This may be due to the fact that BiT uses Weight Standardisation and Group Normalisation when training, rather than standard Batch Normalisation. Mukhoti et al. (2021) show that limiting the Lipschitz constant of the network during training improves the OOD detection performance of gaussian mixture models, which may be also what is occurring in this example. The Mahalanobis detector performs poorly outside ResNetV2-101 otherwise. There is non-negligible variance between training runs on a number of OOD datasets, highlighting the need to perform multiple training runs. Some datasets (e.g. Noise, Colorectal), have especially high variation.

B.1.1 Additional Analysis for SIRC on ID✗ |ID✓

We note that in some cases for ID✗ |ID✓ SIRC is able to slightly outperform S_1 by itself, even when S_2 has $\leq 50\%$ AUROC by itself (e.g. S_2 =Res. in Tables 1 and 2). This is counter-intuitive as S_2 should be harmful to performance. We provide some analysis to show that in some cases S_2 is indeed useful for ID✗ |ID✓. We train a series of linear logistic classifiers[15] with (MSP, Res.) as the input with different class weightings on the test set of ImageNet-200. Figure 18 shows that for ResNet-50, better ID✗ |ID✓ can be achieved by considering (slightly) the value of Res. alongside MSP. However, it also shows that for high MSP, Res. ID✓ has a significant tail of low confidence values. This tail doesn't have much effect when Res. is considered together with MSP, since MSP ID✓ is high and heavily weighted, but will reduce AUROC when Res. is considered by itself for ID✗ |ID✓.

Figure 18 and Tables 1 and 4 show that for DenseNet-121 Res. provides no benefit at all for ID✗ |ID✓. Generally over different architectures and S_2, no secondary score is able to consistently help for ID✗ |ID✓. Moreover, any benefit is minimal and within the range of ±2 SD. Thus we believe that S_2 should be only considered for its contribution to OOD|ID✓ when deploying SIRC.

B.2 Varying α and β

We plot versions of Fig. 9 for all 3 ImageNet-200 architectures (Figs. 19, 20, 21). We also present the mean ± SD. The ability of SIRC to perform consistently better than the baseline generalises across the 3 different CNN architectures. We note that differences in AURC are harder to distinguish, due to the metric considering the proportion of all input data accepted, rather than just the recall of ID✓. The behaviour, however, is similar to AURR in terms of relative performance to the baseline, so we omit AURC from the main results.

B.3 SCOD vs OOD Detection

Similar to the previous section we include versions of Fig. 10 for all architectures and confidence scores (Figs. 22, 23, 24, 25, 26). The behaviour is as discussed in Sect. 5.4, with methods designed for OOD detection achieving gains over the baseline for OOD detection by sacrificing their ability to separate ID✗ |ID✓.

B.4 Plotting S_2 against S_1

In a similar vein to Fig. 6, we plot different SIRC combinations on the S_1, S_2-plane for different experimental

[15] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

Table 2 Full %AUROC and %FPR@95 results for ResNet-50 trained on ImageNet-200

Model		Method	ID✗		OOD mean		Near-IN-200		Caltech-45		Openimage-O		iNaturalist	
			%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓
ResNet-50 ID %Error: 19.01	SIRC	(MSP, $\|z\|_1$)	90.34 ± 0.2	52.70 ± 3.2	92.00	38.36	85.56 ± 0.6	59.76 ± 2.9	91.36 ± 0.6	41.44 ± 3.0	92.28 ± 0.5	41.36 ± 2.8	94.80 ± 0.3	29.60 ± 1.3
		(MSP, Res.)	**90.43 ± 0.3**	52.10 ± 3.0	92.94	33.57	85.52 ± 0.6	60.03 ± 2.4	91.19 ± 0.6	42.27 ± 3.2	92.57 ± 0.6	39.95 ± 3.3	94.10 ± 0.3	33.55 ± 2.7
		(MSP, KNN)	90.39 ± 0.3	52.27 ± 1.9	93.03	32.63	85.61 ± 0.6	59.39 ± 2.0	91.64 ± 0.5	39.37 ± 2.3	92.64 ± 0.4	39.14 ± 1.6	94.53 ± 0.3	30.99 ± 1.7
		(DR, $\|z\|_1$)	90.29 ± 0.3	52.54 ± 3.4	92.32	35.13	85.68 ± 0.6	58.05 ± 2.9	91.67 ± 0.5	37.81 ± 2.7	92.59 ± 0.4	37.82 ± 2.6	95.18 ± 0.3	25.78 ± 1.7
		(DR, Res.)	90.40 ± 0.4	**51.81 ± 2.9**	93.22	30.29	85.62 ± 0.6	58.49 ± 3.2	91.44 ± 0.6	38.92 ± 3.2	92.87 ± 0.6	36.36 ± 3.6	94.32 ± 0.4	30.05 ± 3.2
		(DR, KNN)	90.34 ± 0.3	51.90 ± 2.7	93.40	29.24	85.72 ± 0.6	57.90 ± 3.2	91.96 ± 0.5	35.92 ± 2.9	92.96 ± 0.4	35.86 ± 2.7	94.81 ± 0.3	27.57 ± 2.5
		($-\mathcal{H}$, $\|z\|_1$)	90.00 ± 0.4	54.26 ± 2.7	92.74	33.73	85.88 ± 0.6	58.50 ± 3.3	92.19 ± 0.5	36.08 ± 2.5	92.87 ± 0.4	37.83 ± 3.3	95.38 ± 0.2	25.09 ± 1.9
		($-\mathcal{H}$, Res.)	90.13 ± 0.4	54.01 ± 3.2	93.76	28.42	85.85 ± 0.6	58.93 ± 3.3	92.11 ± 0.5	36.76 ± 3.0	93.25 ± 0.6	36.36 ± 4.5	94.82 ± 0.3	28.51 ± 3.4
		($-\mathcal{H}$, KNN)	90.03 ± 0.4	54.21 ± 2.7	93.74	28.30	85.90 ± 0.6	58.56 ± 3.5	92.44 ± 0.4	34.67 ± 3.1	93.19 ± 0.4	36.56 ± 3.5	95.11 ± 0.3	26.91 ± 2.9
	SIRC+	(MSP, KNN, Res., $\|z\|_1$)	90.32 ± 0.3	52.81 ± 3.4	93.84	27.82	85.47 ± 0.7	59.96 ± 3.1	91.79 ± 0.6	38.23 ± 3.0	93.23 ± 0.5	35.46 ± 2.5	94.88 ± 0.2	28.35 ± 1.5
		(DR, KNN, Res., $\|z\|_1$)	90.21 ± 0.4	53.24 ± 3.9	94.10	25.25	85.53 ± 0.6	59.03 ± 3.6	92.08 ± 0.5	35.18 ± 3.1	93.55 ± 0.5	**32.19 ± 2.7**	95.18 ± 0.2	24.97 ± 1.7
		($-\mathcal{H}$, KNN, Res., $\|z\|_1$)	89.97 ± 0.4	54.98 ± 2.8	**94.42**	**24.74**	85.76 ± 0.7	59.69 ± 3.7	**92.58 ± 0.5**	**34.13 ± 3.8**	**93.79 ± 0.5**	33.02 ± 4.1	**95.43 ± 0.2**	**24.94 ± 2.5**
		MSP	90.41 ± 0.3	52.13 ± 2.0	91.50	41.30	85.59 ± 0.6	59.74 ± 2.0	91.13 ± 0.6	42.72 ± 2.8	91.95 ± 0.5	43.55 ± 2.4	94.23 ± 0.3	33.21 ± 1.2
		DOCTOR	90.39 ± 0.3	51.87 ± 1.6	91.76	38.19	85.73 ± 0.6	**57.89 ± 2.3**	91.41 ± 0.5	39.22 ± 2.2	92.20 ± 0.5	40.22 ± 2.7	94.51 ± 0.3	29.41 ± 1.8
		$-\mathcal{H}$	90.07 ± 0.4	54.05 ± 2.9	92.33	36.06	**85.91 ± 0.6**	58.47 ± 3.3	92.01 ± 0.5	37.20 ± 2.6	92.59 ± 0.5	40.10 ± 3.9	94.90 ± 0.3	28.01 ± 3.2
		$\|z\|_1$	48.06 ± 1.1	94.70 ± 1.4	79.19	57.98	52.27 ± 0.7	94.58 ± 0.5	70.28 ± 1.6	77.83 ± 1.8	72.23 ± 2.4	71.51 ± 2.6	85.65 ± 2.7	49.50 ± 5.8
		Residual	47.59 ± 1.8	96.45 ± 1.1	58.43	79.71	44.30 ± 1.1	96.79 ± 0.4	47.76 ± 1.4	94.83 ± 0.9	59.65 ± 4.0	86.85 ± 2.2	40.07 ± 6.3	97.32 ± 0.7
		KNN	68.60 ± 1.2	88.28 ± 2.2	90.22	39.36	68.27 ± 1.5	88.57 ± 1.3	86.40 ± 1.6	60.71 ± 4.3	86.50 ± 0.8	57.16 ± 3.3	84.85 ± 2.5	66.11 ± 7.6
		Max Logit	83.21 ± 0.6	65.16 ± 3.4	92.94	31.61	82.68 ± 0.7	65.37 ± 3.6	92.48 ± 0.6	36.50 ± 4.1	91.49 ± 0.4	43.27 ± 3.1	94.57 ± 0.3	29.17 ± 3.0
		Energy	82.05 ± 0.6	69.79 ± 3.9	92.70	32.64	81.96 ± 0.7	68.70 ± 4.2	92.15 ± 0.6	38.62 ± 4.9	90.92 ± 0.4	46.28 ± 3.3	94.13 ± 0.4	31.70 ± 2.8
		Gradnorm	60.17 ± 1.5	87.88 ± 2.5	86.17	42.57	62.90 ± 0.5	86.89 ± 0.8	81.11 ± 1.7	59.23 ± 3.3	81.09 ± 1.8	57.80 ± 2.7	91.00 ± 1.8	34.46 ± 3.9
		ViM	80.62 ± 0.7	78.13 ± 2.3	92.87	35.65	78.90 ± 0.8	80.30 ± 2.2	90.54 ± 0.7	54.70 ± 5.0	91.87 ± 1.2	43.84 ± 5.6	90.13 ± 1.8	56.97 ± 8.5
		Mahal	49.96 ± 2.0	96.36 ± 0.9	61.50	79.84	46.57 ± 1.5	96.83 ± 0.5	50.34 ± 1.6	95.01 ± 0.6	63.66 ± 3.7	86.30 ± 2.0	47.42 ± 6.5	96.99 ± 0.8

Model		Method	Textures		SpaceNet		Colonoscopy		Colorectal		Noise		ImageNet-O	
			%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓
ResNet-50 ID %Error: 19.01	SIRC	(MSP, $\|z\|_1$)	93.64 ± 0.7	32.02 ± 3.3	96.41 ± 0.9	21.19 ± 4.2	95.93 ± 1.0	25.33 ± 6.4	95.84 ± 3.3	24.39 ± 13.7	90.72 ± 6.0	49.63 ± 20.8	83.44 ± 0.9	58.91 ± 1.9
		(MSP, Res.)	96.00 ± 0.5	19.81 ± 2.1	96.35 ± 1.0	20.80 ± 3.9	95.52 ± 0.7	27.31 ± 5.3	95.32 ± 4.0	26.97 ± 17.5	98.21 ± 2.3	10.97 ± 16.7	84.62 ± 0.9	53.99 ± 1.4
		(MSP, KNN)	95.72 ± 0.4	20.32 ± 1.4	97.11 ± 0.8	16.65 ± 3.9	96.33 ± 0.8	22.44 ± 5.8	96.84 ± 3.2	17.87 ± 13.9	95.63 ± 5.1	25.80 ± 27.1	84.28 ± 0.9	54.32 ± 1.3
		(DR, $\|z\|_1$)	94.01 ± 0.7	28.62 ± 2.6	96.76 ± 0.9	17.52 ± 4.0	96.34 ± 1.0	20.94 ± 6.0	96.28 ± 3.2	20.30 ± 13.5	91.08 ± 5.8	47.75 ± 16.6	83.64 ± 1.0	56.68 ± 1.5
		(DR, Res.)	96.28 ± 0.5	17.29 ± 2.0	96.67 ± 1.0	17.08 ± 4.1	95.82 ± 0.6	23.07 ± 4.7	95.62 ± 4.1	23.40 ± 18.5	98.63 ± 1.9	7.23 ± 10.4	84.90 ± 0.9	51.05 ± 1.7
		(DR, KNN)	96.07 ± 0.4	17.62 ± 1.0	97.49 ± 0.7	13.48 ± 3.5	96.78 ± 0.7	18.10 ± 4.9	97.24 ± 3.0	14.64 ± 13.0	96.37 ± 4.7	19.97 ± 23.7	84.60 ± 0.9	51.34 ± 1.8
		($-\mathcal{H}$, $\|z\|_1$)	94.38 ± 0.7	27.38 ± 2.7	97.26 ± 0.8	14.69 ± 3.9	96.97 ± 0.8	16.87 ± 4.4	96.71 ± 2.8	18.71 ± 13.5	91.74 ± 4.3	45.84 ± 17.6	84.01 ± 0.9	56.34 ± 2.4

Table 2 continued

Model	Method	Textures		SpaceNet		Colonoscopy		Colorectal		Noise		ImageNet-O	
		%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓
	($-\mathcal{H}$, Res.)	96.68 ± 0.5	15.70 ± 2.1	97.36 ± 0.8	13.70 ± 3.5	96.72 ± 0.6	18.10 ± 3.7	96.41 ± 3.6	20.42 ± 16.7	99.02 ± 1.5	4.89 ± 5.5	85.33 ± 0.9	50.81 ± 2.9
	($-\mathcal{H}$, KNN)	96.30 ± 0.4	16.87 ± 1.5	97.84 ± 0.7	11.39 ± 3.7	97.27 ± 0.6	14.85 ± 4.3	97.54 ± 2.6	13.72 ± 12.9	96.92 ± 3.5	17.47 ± 19.1	84.88 ± 0.9	51.99 ± 2.2
SIRC+	(MSP, KNN, Res., $\|z\|_1$)	97.10 ± 0.3	13.46 ± 1.2	97.49 ± 0.7	13.97 ± 3.7	96.24 ± 0.7	22.12 ± 5.1	97.43 ± 3.0	14.09 ± 14.2	99.45 ± 1.6	2.44 ± 6.7	85.33 ± 0.8	50.15 ± 1.4
	(DR, KNN, Res., $\|z\|_1$)	97.34 ± 0.3	11.39 ± 0.9	97.79 ± 0.7	11.17 ± 2.7	96.62 ± 0.6	17.92 ± 4.8	97.78 ± 2.8	11.28 ± 12.8	99.52 ± 1.5	1.99 ± 6.3	85.64 ± 0.8	47.39 ± 1.4
	($-\mathcal{H}$, KNN, Res., $\|z\|_1$)	97.54 ± 0.3	10.93 ± 1.0	98.15 ± 0.6	9.57 ± 3.1	97.19 ± 0.5	15.12 ± 3.4	98.09 ± 2.4	10.56 ± 12.9	99.65 ± 1.1	1.61 ± 5.0	86.00 ± 0.8	47.83 ± 3.0
	MSP	92.88 ± 0.8	36.61 ± 3.1	95.94 ± 0.9	23.74 ± 3.7	95.75 ± 0.8	26.52 ± 6.2	94.86 ± 3.5	30.28 ± 13.6	89.33 ± 5.7	56.83 ± 20.2	83.29 ± 0.9	59.78 ± 2.1
	DOCTOR	93.16 ± 0.8	33.46 ± 3.6	96.28 ± 0.9	19.95 ± 3.8	96.14 ± 0.8	22.07 ± 5.3	95.16 ± 3.5	27.21 ± 14.2	89.51 ± 5.5	54.83 ± 20.4	83.47 ± 0.9	57.64 ± 1.9
	$-\mathcal{H}$	93.77 ± 0.8	30.79 ± 3.7	96.95 ± 0.8	16.43 ± 3.9	96.87 ± 0.7	17.55 ± 4.6	95.93 ± 3.2	23.43 ± 14.6	90.47 ± 4.2	51.63 ± 19.7	83.89 ± 0.9	57.02 ± 1.7
	$\|z\|_1$	88.90 ± 1.5	39.67 ± 2.6	87.93 ± 5.5	51.46 ± 15.0	76.97 ± 9.7	82.24 ± 14.3	97.28 ± 2.3	14.64 ± 13.9	97.36 ± 4.6	13.51 ± 33.1	63.00 ± 1.7	84.82 ± 1.6
	Residual	82.84 ± 2.4	46.63 ± 3.8	58.32 ± 9.5	86.43 ± 5.1	38.09 ± 13.9	99.64 ± 0.4	53.93 ± 13.2	88.78 ± 10.1	91.31 ± 6.4	20.92 ± 12.5	68.04 ± 2.7	78.98 ± 2.2
	KNN	98.13 ± 0.2	9.23 ± 1.2	97.02 ± 1.1	17.94 ± 6.5	94.80 ± 1.3	38.33 ± 11.1	**99.11 ± 0.5**	**2.85 ± 2.3**	99.68 ± 0.8	1.18 ± 4.2	87.40 ± 0.9	51.50 ± 2.8
	Max Logit	95.44 ± 0.8	22.04 ± 2.8	98.45 ± 0.9	8.74 ± 5.2	**97.65 ± 0.7**	**13.56 ± 6.0**	98.93 ± 1.0	5.83 ± 6.4	94.73 ± 5.3	31.53 ± 28.8	82.98 ± 0.9	60.09 ± 2.6
	Energy	95.37 ± 0.8	22.50 ± 2.8	**98.45 ± 1.0**	**8.49 ± 6.1**	97.51 ± 0.8	14.19 ± 6.5	99.07 ± 1.0	5.00 ± 6.5	94.93 ± 5.4	29.05 ± 30.8	82.52 ± 0.9	61.86 ± 2.7
	Gradnorm	93.00 ± 1.1	26.57 ± 3.0	94.76 ± 3.2	26.01 ± 13.0	90.54 ± 4.6	42.85 ± 16.2	98.98 ± 1.1	4.98 ± 7.2	97.59 ± 4.2	13.05 ± 31.9	70.78 ± 1.8	73.88 ± 3.1
	ViM	**98.46 ± 0.4**	**7.62 ± 2.1**	97.63 ± 0.4	13.19 ± 2.8	94.42 ± 1.4	44.55 ± 14.5	98.04 ± 1.3	8.84 ± 10.2	**99.82 ± 0.1**	**0.31 ± 0.3**	**88.85 ± 0.9**	**46.15 ± 3.9**
	Mahal	84.64 ± 2.1	46.98 ± 3.2	60.06 ± 8.7	88.13 ± 4.9	41.02 ± 14.2	99.70 ± 0.3	57.88 ± 12.5	88.37 ± 8.1	94.08 ± 5.4	20.45 ± 13.0	69.29 ± 2.5	79.65 ± 1.9

We show the mean ±2 SD. over 5 independent training runs. Bold indicates best performance, underline 2nd or 3rd best

Table 3 Full %AUROC and %FPR@95 results for MobileNet-V2 trained on ImageNet-200

Model		Method	ID✗		OOD mean		Near-IN-200		Caltech-45		Openimage-O		iNaturalist	
			%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓
MobileNetV2 ID %Error: 21.35	SIRC	(MSP, $\|z\|_1$)	89.52 ± 0.3	55.52 ± 1.1	92.57	33.77	84.78 ± 0.3	61.31 ± 1.2	90.45 ± 0.3	43.01 ± 0.9	91.26 ± 0.4	44.05 ± 1.5	94.20 ± 0.8	31.89 ± 3.7
		(MSP, Res.)	**89.67 ± 0.3**	**55.04 ± 1.5**	92.05	37.54	84.84 ± 0.4	61.18 ± 0.2	90.24 ± 0.4	44.42 ± 2.3	91.19 ± 0.5	44.82 ± 1.8	93.22 ± 0.9	37.94 ± 3.7
		(MSP, KNN)	89.60 ± 0.3	55.36 ± 1.8	92.85	32.35	84.89 ± 0.3	60.99 ± 1.0	90.78 ± 0.3	41.19 ± 1.7	91.45 ± 0.5	43.28 ± 1.8	93.56 ± 0.9	36.00 ± 3.9
		(DR, $\|z\|_1$)	89.40 ± 0.3	56.56 ± 2.1	92.97	31.24	84.89 ± 0.3	61.32 ± 0.7	90.82 ± 0.3	40.56 ± 2.2	91.60 ± 0.4	42.40 ± 0.9	94.63 ± 0.7	**29.19 ± 2.8**
		(DR, Res.)	89.59 ± 0.3	55.71 ± 2.1	92.36	35.19	84.97 ± 0.3	60.98 ± 1.0	90.58 ± 0.4	42.00 ± 2.3	91.50 ± 0.5	43.26 ± 1.9	93.40 ± 0.9	36.35 ± 4.2
		(DR, KNN)	89.47 ± 0.3	56.40 ± 1.8	93.24	30.01	85.00 ± 0.3	61.20 ± 0.6	91.19 ± 0.4	38.80 ± 2.3	91.77 ± 0.5	42.04 ± 1.6	93.78 ± 0.9	34.71 ± 3.7
		($-\mathcal{H}$, $\|z\|_1$)	88.89 ± 0.3	58.75 ± 2.3	93.26	30.89	84.95 ± 0.2	62.80 ± 0.9	91.35 ± 0.3	39.95 ± 2.1	91.82 ± 0.4	44.05 ± 1.4	**94.73 ± 0.7**	30.52 ± 3.8
		($-\mathcal{H}$, Res.)	89.11 ± 0.3	57.94 ± 3.4	93.02	32.95	**85.08 ± 0.3**	62.14 ± 1.0	91.33 ± 0.3	39.98 ± 2.2	91.92 ± 0.4	43.89 ± 1.4	94.00 ± 0.8	35.32 ± 3.7
		($-\mathcal{H}$, KNN)	88.94 ± 0.2	58.62 ± 2.2	93.49	30.11	85.02 ± 0.2	62.70 ± 0.8	91.63 ± 0.3	38.75 ± 2.1	91.95 ± 0.4	43.86 ± 2.0	94.15 ± 0.8	35.07 ± 4.2
	SIRC+	(MSP, KNN, Res., $\|z\|_1$)	89.52 ± 0.3	55.90 ± 1.0	93.33	29.64	84.82 ± 0.3	61.50 ± 1.2	91.05 ± 0.3	39.81 ± 1.5	91.94 ± 0.5	40.56 ± 2.3	93.92 ± 0.9	33.69 ± 4.1
		(DR, KNN, Res., $\|z\|_1$)	89.34 ± 0.3	57.40 ± 1.8	93.68	**27.20**	84.88 ± 0.3	61.94 ± 0.4	91.45 ± 0.3	**36.91 ± 1.7**	92.29 ± 0.5	**38.61 ± 1.5**	94.20 ± 0.8	31.86 ± 3.7
		($-\mathcal{H}$, KNN, Res., $\|z\|_1$)	88.90 ± 0.3	58.89 ± 2.3	**93.95**	27.44	84.98 ± 0.2	62.95 ± 1.0	**91.93 ± 0.3**	36.99 ± 1.4	**92.47 ± 0.4**	40.85 ± 1.5	94.45 ± 0.7	32.96 ± 4.2
		MSP	89.63 ± 0.3	55.10 ± 1.5	91.85	38.60	84.84 ± 0.3	61.07 ± 1.0	90.17 ± 0.3	44.81 ± 1.8	90.91 ± 0.5	46.36 ± 1.8	93.56 ± 0.9	35.89 ± 3.8
		DOCTOR	89.56 ± 0.3	55.51 ± 2.1	92.17	36.22	84.98 ± 0.3	**60.63 ± 0.3**	90.51 ± 0.3	42.46 ± 2.1	91.19 ± 0.5	44.96 ± 1.0	93.90 ± 0.8	33.41 ± 3.4
		$-\mathcal{H}$	89.01 ± 0.2	58.48 ± 2.3	92.72	34.62	85.03 ± 0.2	62.55 ± 0.5	91.15 ± 0.3	41.31 ± 1.9	91.54 ± 0.4	46.14 ± 1.4	94.23 ± 0.7	33.87 ± 3.9
		$\|z\|_1$	53.57 ± 0.7	93.40 ± 0.5	82.20	51.92	56.05 ± 0.7	92.65 ± 0.5	75.15 ± 1.4	73.17 ± 2.2	74.05 ± 1.4	68.93 ± 1.7	86.04 ± 1.7	48.34 ± 5.0
		Residual	42.00 ± 0.8	97.30 ± 0.3	40.43	94.42	42.46 ± 0.7	97.37 ± 0.3	40.90 ± 1.2	96.70 ± 0.9	44.63 ± 1.1	94.39 ± 0.6	22.87 ± 4.7	99.18 ± 0.4
		KNN	68.15 ± 0.6	88.89 ± 0.9	87.63	45.55	68.63 ± 0.4	87.61 ± 1.1	85.90 ± 0.9	59.92 ± 1.6	81.56 ± 2.0	65.04 ± 2.7	73.05 ± 2.9	83.98 ± 4.0
		Max Logit	83.13 ± 0.6	63.89 ± 1.8	92.64	32.54	81.75 ± 0.4	67.38 ± 1.3	91.39 ± 0.2	42.47 ± 2.2	89.69 ± 0.8	50.68 ± 3.1	92.62 ± 1.0	39.79 ± 5.0
		Energy	81.86 ± 0.7	68.01 ± 2.0	92.29	34.31	80.86 ± 0.4	70.82 ± 1.2	90.92 ± 0.3	45.79 ± 2.0	88.86 ± 0.8	54.54 ± 3.1	91.76 ± 1.0	44.25 ± 5.7
		Gradnorm	65.27 ± 1.1	85.73 ± 1.1	88.13	38.66	66.06 ± 0.7	85.14 ± 1.0	83.94 ± 1.0	56.59 ± 2.9	81.94 ± 1.2	58.21 ± 1.7	90.73 ± 1.3	36.81 ± 3.4
		ViM	80.20 ± 0.4	74.37 ± 2.1	89.81	51.43	79.14 ± 0.3	75.78 ± 1.4	89.16 ± 0.4	58.46 ± 0.4	87.66 ± 1.0	59.55 ± 2.1	81.93 ± 3.0	81.71 ± 6.4
		Mahal	44.44 ± 1.0	97.14 ± 0.6	42.45	94.56	44.57 ± 0.7	97.23 ± 0.4	42.82 ± 1.1	96.65 ± 0.8	48.03 ± 1.2	94.11 ± 0.8	27.31 ± 4.8	99.07 ± 0.3

Model		Method	Textures		SpaceNet		Colonoscopy		Colorectal		Noise		ImageNet-O	
			%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓
MobileNetV2 ID %Error: 21.35	SIRC	(MSP, $\|z\|_1$)	94.05 ± 0.4	28.69 ± 0.5	95.29 ± 1.7	24.36 ± 6.8	96.63 ± 0.8	19.39 ± 4.9	96.97 ± 1.4	19.40 ± 8.8	98.77 ± 1.1	7.62 ± 7.7	83.30 ± 0.3	58.00 ± 1.2
		(MSP, Res.)	94.25 ± 0.2	28.36 ± 1.2	94.58 ± 1.6	28.39 ± 6.2	96.25 ± 0.9	21.58 ± 4.3	95.45 ± 1.7	29.24 ± 9.3	96.44 ± 2.5	24.48 ± 24.6	84.08 ± 0.4	55.03 ± 1.6
		(MSP, KNN)	95.02 ± 0.4	23.45 ± 1.5	95.85 ± 1.5	21.58 ± 5.4	97.01 ± 0.8	17.14 ± 3.9	97.13 ± 1.4	18.23 ± 9.0	98.82 ± 1.4	7.59 ± 9.3	84.05 ± 0.3	54.01 ± 1.5

Table 3 continued

Model	Method	Textures		SpaceNet		Colonoscopy		Colorectal		Noise		ImageNet-O	
		%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓
	(DR, $\|z\|_1$)	94.53 ± 0.4	25.73 ± 1.0	95.74 ± 1.6	21.39 ± 6.2	97.07 ± 0.7	15.68 ± 3.2	97.61 ± 1.3	14.91 ± 7.3	99.25 ± 0.9	4.39 ± 5.6	83.55 ± 0.2	56.86 ± 1.5
	(DR, Res.)	94.68 ± 0.2	25.39 ± 0.6	94.91 ± 1.6	25.67 ± 6.3	96.61 ± 0.8	17.93 ± 3.8	95.77 ± 1.6	26.34 ± 9.2	96.67 ± 2.7	21.43 ± 30.1	84.49 ± 0.3	52.55 ± 2.2
	(DR, KNN)	95.48 ± 0.4	20.59 ± 1.2	96.33 ± 1.4	18.08 ± 5.1	97.49 ± 0.7	13.29 ± 3.6	97.71 ± 1.2	14.10 ± 7.3	99.19 ± 1.1	4.95 ± 6.0	84.44 ± 0.3	52.36 ± 1.1
	($-\mathcal{H}$, $\|z\|_1$)	94.87 ± 0.4	25.29 ± 1.0	96.35 ± 1.5	19.05 ± 6.9	97.71 ± 0.5	11.71 ± 3.4	97.81 ± 1.1	13.81 ± 7.7	99.06 ± 1.2	4.30 ± 6.7	83.92 ± 0.3	57.41 ± 1.5
	($-\mathcal{H}$, Res.)	95.37 ± 0.2	22.66 ± 0.8	96.01 ± 1.4	21.04 ± 5.8	97.57 ± 0.6	12.50 ± 3.9	96.76 ± 1.4	21.17 ± 10.9	97.28 ± 2.8	17.62 ± 30.2	84.90 ± 0.3	53.13 ± 1.2
	($-\mathcal{H}$, KNN)	95.71 ± 0.3	20.48 ± 1.0	96.84 ± 1.3	16.47 ± 5.8	97.96 ± 0.5	10.55 ± 3.5	97.90 ± 1.0	13.41 ± 7.0	99.07 ± 1.6	5.65 ± 11.4	84.64 ± 0.2	54.19 ± 0.8
	SIRC+ (MSP, KNN, Res., $\|z\|_1$)	96.10 ± 0.2	17.84 ± 1.2	96.28 ± 1.4	19.07 ± 5.6	96.88 ± 0.8	17.42 ± 3.9	97.82 ± 1.3	13.41 ± 7.9	99.54 ± 0.5	2.88 ± 3.8	84.98 ± 0.3	50.18 ± 1.3
	(DR, KNN, Res., $\|z\|_1$)	96.49 ± 0.1	**15.26 ± 1.1**	96.74 ± 1.3	15.48 ± 5.1	97.26 ± 0.7	13.47 ± 2.9	98.34 ± 1.1	9.46 ± 5.7	99.72 ± 0.4	1.62 ± 2.8	85.43 ± 0.2	**47.41 ± 0.8**
	($-\mathcal{H}$, KNN, Res., $\|z\|_1$)	**96.70 ± 0.1**	15.45 ± 0.6	97.20 ± 1.2	14.38 ± 5.5	97.87 ± 0.5	10.56 ± 3.1	98.46 ± 0.9	9.32 ± 5.6	99.73 ± 0.4	1.44 ± 2.7	85.69 ± 0.3	49.55 ± 0.5
	MSP	92.93 ± 0.4	35.31 ± 0.8	94.62 ± 1.7	28.20 ± 6.3	96.51 ± 0.8	20.36 ± 4.6	95.62 ± 1.6	28.53 ± 9.4	96.21 ± 2.7	26.38 ± 25.6	83.10 ± 0.3	59.06 ± 1.3
	DOCTOR	93.31 ± 0.4	32.60 ± 1.6	95.03 ± 1.7	25.30 ± 6.5	96.99 ± 0.8	16.46 ± 3.7	96.04 ± 1.5	25.24 ± 8.8	96.45 ± 2.9	23.32 ± 31.4	83.34 ± 0.3	57.86 ± 1.3
	$-\mathcal{H}$	94.05 ± 0.4	29.75 ± 0.9	95.91 ± 1.5	21.60 ± 6.9	97.68 ± 0.6	12.15 ± 3.5	96.81 ± 1.4	21.18 ± 10.2	97.07 ± 2.8	19.35 ± 30.8	83.76 ± 0.3	58.33 ± 1.2
	$\|z\|_1$	92.88 ± 0.3	27.55 ± 1.5	92.43 ± 2.9	37.72 ± 12.3	79.91 ± 5.7	80.69 ± 10.1	98.20 ± 1.1	9.40 ± 6.4	99.93 ± 0.0	0.01 ± 0.0	67.33 ± 2.0	80.74 ± 3.0
	Residual	56.86 ± 1.4	78.82 ± 2.1	31.49 ± 7.8	97.15 ± 4.7	27.32 ± 6.6	99.68 ± 1.1	28.41 ± 8.1	99.38 ± 1.2	49.60 ± 18.3	93.66 ± 5.7	59.77 ± 1.2	87.86 ± 0.6
	KNN	95.87 ± 0.3	17.81 ± 1.5	93.87 ± 1.3	27.53 ± 6.5	93.22 ± 1.7	49.42 ± 9.9	98.63 ± 0.6	6.81 ± 5.1	99.69 ± 0.5	1.50 ± 3.9	85.92 ± 0.7	55.85 ± 1.6
	Max Logit	95.12 ± 0.4	25.47 ± 1.1	97.69 ± 1.0	13.42 ± 6.2	**98.12 ± 0.5**	**9.66 ± 3.8**	98.27 ± 1.0	9.74 ± 7.5	98.74 ± 0.9	3.43 ± 5.2	83.01 ± 0.6	63.30 ± 1.4
	Energy	95.01 ± 0.4	26.48 ± 2.4	**97.82 ± 1.1**	**12.84 ± 7.2**	97.99 ± 0.5	10.64 ± 4.3	98.43 ± 1.1	8.66 ± 8.4	98.93 ± 0.8	2.55 ± 4.4	82.36 ± 0.7	66.55 ± 1.2
	Gradnorm	95.47 ± 0.2	19.24 ± 1.2	96.13 ± 2.0	20.47 ± 8.8	93.36 ± 2.2	36.23 ± 10.0	**99.25 ± 0.6**	**2.98 ± 3.2**	**99.95 ± 0.0**	**0.00 ± 0.0**	74.50 ± 1.6	70.91 ± 2.2
	ViM	95.45 ± 0.2	26.02 ± 1.6	92.99 ± 1.3	46.51 ± 4.7	93.61 ± 1.5	50.42 ± 13.5	94.04 ± 2.1	43.82 ± 16.7	96.71 ± 0.5	19.97 ± 9.9	**87.44 ± 0.6**	52.09 ± 2.0
	Mahal	58.12 ± 1.6	79.52 ± 2.3	31.64 ± 7.6	97.83 ± 3.9	29.12 ± 7.4	99.67 ± 1.0	30.35 ± 8.1	99.15 ± 2.2	51.22 ± 15.5	94.63 ± 5.4	61.28 ± 1.1	87.73 ± 1.1

We show the mean ±2 SD. over 5 independent training runs. Bold indicates best performance, underline 2nd or 3rd best

Table 4 Full %AUROC and %FPR@95 results for DenseNet-121 trained on ImageNet-200

Model		Method	ID✗		OOD mean		Near-IN-200		Caltech-45		Openimage-O		iNaturalist	
			%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓
DenseNet-121 ID %Error: 17.20	SIRC	(MSP, $\|z\|_1$)	90.22 ± 0.8	52.33 ± 2.7	92.15	37.19	85.28 ± 0.1	59.34 ± 0.9	91.33 ± 0.5	40.78 ± 2.1	91.88 ± 0.4	43.31 ± 1.4	93.52 ± 0.8	35.86 ± 2.7
		(MSP, Res.)	90.20 ± 0.8	52.44 ± 4.3	93.20	31.38	85.18 ± 0.1	59.66 ± 1.7	91.29 ± 0.5	40.53 ± 1.6	92.55 ± 0.3	39.25 ± 1.0	93.42 ± 0.7	36.00 ± 1.9
		(MSP, KNN)	90.31 ± 0.8	52.10 ± 3.1	92.98	32.21	85.39 ± 0.1	58.72 ± 1.2	91.68 ± 0.5	38.48 ± 1.8	92.37 ± 0.4	40.30 ± 1.5	93.54 ± 0.7	35.55 ± 2.5
		(DR, $\|z\|_1$)	90.22 ± 0.8	52.31 ± 3.0	92.46	33.67	85.42 ± 0.1	57.48 ± 2.0	91.63 ± 0.5	37.13 ± 1.1	92.14 ± 0.4	39.87 ± 1.1	93.85 ± 0.7	31.95 ± 1.9
		(DR, Res.)	90.18 ± 0.8	52.46 ± 4.3	93.44	28.28	85.28 ± 0.1	57.99 ± 2.7	91.53 ± 0.5	36.84 ± 1.4	92.83 ± 0.3	35.56 ± 1.1	93.69 ± 0.7	32.10 ± 1.0
		(DR, KNN)	90.29 ± 0.8	**51.73 ± 2.8**	93.28	28.91	85.50 ± 0.1	**56.88 ± 2.1**	91.98 ± 0.5	34.95 ± 0.5	92.66 ± 0.4	36.64 ± 0.6	93.80 ± 0.7	31.79 ± 1.2
		($-\mathcal{H}$, $\|z\|_1$)	89.95 ± 0.9	53.90 ± 2.4	92.90	31.20	85.64 ± 0.1	56.98 ± 1.5	92.18 ± 0.4	34.18 ± 1.7	92.50 ± 0.5	38.42 ± 1.9	94.16 ± 0.7	30.23 ± 2.6
		($-\mathcal{H}$, Res.)	89.92 ± 0.9	54.15 ± 3.2	93.85	26.51	85.50 ± 0.1	57.87 ± 1.8	92.13 ± 0.4	34.46 ± 1.2	93.16 ± 0.4	34.86 ± 1.5	94.07 ± 0.6	30.59 ± 1.9
		($-\mathcal{H}$, KNN)	90.01 ± 0.8	53.82 ± 2.1	93.62	27.56	**85.71 ± 0.1**	57.20 ± 1.5	**92.47 ± 0.4**	**32.84 ± 0.9**	92.94 ± 0.4	36.56 ± 1.3	94.15 ± 0.6	30.84 ± 2.3
	SIRC+	(MSP, KNN, Res., $\|z\|_1$)	90.08 ± 0.8	53.33 ± 3.6	93.61	28.65	85.13 ± 0.1	59.76 ± 2.0	91.69 ± 0.5	37.97 ± 1.5	92.89 ± 0.3	36.88 ± 1.0	93.73 ± 0.7	33.81 ± 1.7
		(DR, KNN, Res., $\|z\|_1$)	89.99 ± 0.8	53.93 ± 3.0	93.84	25.91	85.19 ± 0.1	58.49 ± 2.1	91.94 ± 0.4	34.44 ± 1.1	93.16 ± 0.3	**33.55 ± 0.8**	93.98 ± 0.6	30.16 ± 1.2
		($-\mathcal{H}$, KNN, Res., $\|z\|_1$)	89.74 ± 0.8	55.38 ± 2.6	**94.15**	**24.90**	85.40 ± 0.1	58.73 ± 2.1	92.43 ± 0.4	32.88 ± 1.0	**93.42 ± 0.3**	33.59 ± 1.1	**94.28 ± 0.6**	**29.65 ± 1.2**
		MSP	90.31 ± 0.8	51.83 ± 2.9	91.91	38.79	85.34 ± 0.1	59.20 ± 1.0	91.23 ± 0.5	41.38 ± 2.2	91.78 ± 0.4	44.00 ± 1.2	93.32 ± 0.7	37.10 ± 2.9
		DOCTOR	**90.32 ± 0.8**	51.76 ± 3.3	92.19	35.23	85.48 ± 0.1	57.16 ± 1.8	91.52 ± 0.5	37.60 ± 1.2	92.03 ± 0.4	40.42 ± 1.3	93.60 ± 0.7	33.15 ± 1.9
		$-\mathcal{H}$	90.04 ± 0.9	53.35 ± 2.5	92.72	32.66	85.70 ± 0.1	56.98 ± 1.7	92.11 ± 0.4	34.77 ± 1.7	92.43 ± 0.5	39.13 ± 2.5	94.00 ± 0.6	31.35 ± 3.1
		$\|z\|_1$	36.87 ± 2.3	98.70 ± 0.4	63.78	79.71	42.10 ± 1.9	98.50 ± 0.6	55.63 ± 4.5	91.89 ± 2.1	53.48 ± 3.7	91.83 ± 2.4	64.95 ± 5.9	87.01 ± 6.2
		Residual	46.07 ± 1.1	95.47 ± 0.8	70.49	71.40	44.75 ± 0.9	96.14 ± 0.1	55.02 ± 1.6	91.28 ± 0.9	70.77 ± 2.2	78.27 ± 1.5	64.90 ± 6.1	90.94 ± 2.5
		KNN	71.83 ± 0.8	86.41 ± 2.0	90.10	43.62	71.86 ± 0.8	86.86 ± 0.8	87.04 ± 0.6	61.20 ± 2.8	86.89 ± 1.3	58.41 ± 2.7	83.18 ± 2.4	72.57 ± 5.4
		Max Logit	83.29 ± 1.3	62.31 ± 2.7	91.93	34.33	82.50 ± 0.3	63.05 ± 1.8	91.57 ± 0.6	37.88 ± 2.1	89.71 ± 1.1	48.53 ± 2.5	91.28 ± 1.2	42.80 ± 4.6
		Energy	82.12 ± 1.3	66.57 ± 3.4	91.54	36.74	81.82 ± 0.4	66.34 ± 1.6	91.14 ± 0.7	40.73 ± 2.4	88.88 ± 1.2	52.81 ± 3.0	90.37 ± 1.4	47.68 ± 5.2
		Gradnorm	50.18 ± 2.5	95.19 ± 1.6	76.84	61.85	54.43 ± 2.0	93.93 ± 1.9	72.05 ± 4.0	75.00 ± 4.6	67.42 ± 3.6	79.78 ± 4.3	77.17 ± 4.8	67.53 ± 8.2
		ViM	76.63 ± 1.3	84.78 ± 1.9	91.14	42.06	75.68 ± 0.7	86.14 ± 1.7	87.10 ± 1.1	64.29 ± 4.3	89.79 ± 0.9	51.78 ± 4.0	89.00 ± 1.5	59.41 ± 5.9
		Mahal	56.96 ± 1.1	94.58 ± 1.1	73.64	72.31	53.97 ± 1.1	95.39 ± 0.3	60.81 ± 1.5	90.96 ± 1.0	76.57 ± 1.6	76.51 ± 2.5	69.67 ± 4.7	89.80 ± 3.9

Model		Method	Textures		SpaceNet		Colonoscopy		Colorectal		Noise		ImageNet-O	
			%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓
DenseNet-121 ID %Error: 17.20	SIRC	(MSP, $\|z\|_1$)	93.68 ± 0.2	32.44 ± 0.8	96.37 ± 1.0	22.65 ± 5.3	95.39 ± 0.9	27.08 ± 5.5	96.44 ± 2.3	23.52 ± 14.1	94.31 ± 8.3	28.78 ± 25.0	83.32 ± 0.7	58.18 ± 1.3
		(MSP, Res.)	96.55 ± 0.1	16.62 ± 0.9	96.69 ± 0.8	19.56 ± 4.4	95.14 ± 0.8	27.73 ± 4.6	96.89 ± 1.7	19.74 ± 10.6	99.70 ± 1.1	2.02 ± 8.6	84.59 ± 0.6	52.70 ± 2.1

Table 4 continued

Model	Method	Textures		SpaceNet		Colonoscopy		Colorectal		Noise		ImageNet-O	
		%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓	%AUROC↑	%FPR@95↓
	(MSP, KNN)	95.43 ± 0.1	22.31 ± 0.7	96.93 ± 0.8	18.75 ± 4.4	95.83 ± 0.8	24.00 ± 4.2	97.10 ± 2.0	18.88 ± 13.1	97.31 ± 8.2	11.47 ± 23.3	84.22 ± 0.6	53.67 ± 1.3
	(DR, $\|z\|_1$)	94.02 ± 0.2	28.79 ± 0.4	96.69 ± 1.1	19.13 ± 5.3	95.73 ± 0.9	22.94 ± 5.4	96.84 ± 2.2	19.38 ± 13.2	94.72 ± 8.3	24.53 ± 24.1	83.52 ± 0.7	55.50 ± 1.7
	(DR, Res.)	96.77 ± 0.1	14.28 ± 1.2	97.02 ± 0.8	15.70 ± 4.3	95.42 ± 0.8	23.65 ± 4.6	97.24 ± 1.6	15.56 ± 9.2	99.78 ± 0.7	1.23 ± 5.3	84.84 ± 0.6	49.92 ± 3.2
	(DR, KNN)	95.76 ± 0.1	19.11 ± 1.3	97.29 ± 0.9	15.16 ± 4.7	96.21 ± 0.8	19.81 ± 4.1	97.51 ± 1.9	14.88 ± 11.5	97.55 ± 8.2	9.42 ± 22.4	84.52 ± 0.6	50.48 ± 2.4
	($-\mathcal{H}$, $\|z\|_1$)	94.48 ± 0.2	26.16 ± 1.3	97.15 ± 1.1	15.71 ± 6.5	96.43 ± 0.9	17.86 ± 5.1	97.38 ± 1.9	15.65 ± 12.1	95.17 ± 8.0	22.30 ± 19.5	83.89 ± 0.7	54.47 ± 2.0
	($-\mathcal{H}$, Res.)	97.07 ± 0.1	13.01 ± 1.0	97.44 ± 0.9	13.46 ± 5.4	96.21 ± 0.8	18.55 ± 4.3	97.78 ± 1.4	12.57 ± 8.5	99.91 ± 0.3	0.32 ± 1.4	85.22 ± 0.6	49.43 ± 1.5
	($-\mathcal{H}$, KNN)	96.04 ± 0.1	17.93 ± 0.8	97.62 ± 0.8	13.17 ± 5.4	96.76 ± 0.8	16.15 ± 4.0	97.89 ± 1.7	12.49 ± 10.7	97.81 ± 7.8	8.06 ± 22.1	84.80 ± 0.6	50.36 ± 1.7
SIRC+	(MSP, KNN, Res., $\|z\|_1$)	97.06 ± 0.1	13.41 ± 0.7	97.30 ± 0.8	15.58 ± 4.6	95.50 ± 0.8	25.09 ± 4.4	97.82 ± 1.6	13.46 ± 10.0	99.90 ± 0.4	0.72 ± 3.2	85.14 ± 0.6	49.87 ± 2.4
	(DR, KNN, Res., $\|z\|_1$)	97.26 ± 0.1	11.51 ± 0.8	97.58 ± 0.8	12.54 ± 4.5	95.78 ± 0.8	21.08 ± 3.7	98.15 ± 1.4	10.08 ± 8.2	99.94 ± 0.2	0.26 ± 1.1	85.41 ± 0.6	46.95 ± 2.0
	($-\mathcal{H}$, KNN, Res., $\|z\|_1$)	97.48 ± 0.1	10.83 ± 0.9	**97.87 ± 0.8**	**10.99 ± 4.7**	96.42 ± 0.8	17.25 ± 3.7	**98.46 ± 1.2**	**8.21 ± 7.0**	**99.97 ± 0.1**	**0.02 ± 0.1**	85.74 ± 0.7	**46.90 ± 1.7**
	MSP	93.25 ± 0.2	35.02 ± 0.8	96.12 ± 0.9	23.96 ± 4.4	95.42 ± 0.8	26.81 ± 5.1	95.92 ± 2.1	26.92 ± 13.1	93.44 ± 8.4	35.16 ± 28.5	83.31 ± 0.6	58.38 ± 1.3
	DOCTOR	93.54 ± 0.2	31.47 ± 0.5	96.44 ± 1.0	19.94 ± 5.0	95.78 ± 0.9	22.54 ± 4.9	96.25 ± 2.1	23.07 ± 12.5	93.72 ± 8.3	31.27 ± 28.9	83.50 ± 0.6	55.70 ± 1.9
	$-\mathcal{H}$	94.14 ± 0.2	28.09 ± 1.8	97.00 ± 1.0	16.39 ± 6.6	96.47 ± 0.8	17.68 ± 5.3	96.99 ± 1.9	18.26 ± 12.1	94.43 ± 7.8	29.17 ± 22.9	83.89 ± 0.7	54.77 ± 2.3
	$\|z\|_1$	75.96 ± 4.4	69.62 ± 9.0	65.98 ± 11.8	73.95 ± 4.6	47.58 ± 11.0	97.80 ± 0.4	87.28 ± 9.2	60.31 ± 32.1	95.31 ± 5.3	29.94 ± 34.9	49.50 ± 4.0	96.31 ± 1.0
	Residual	92.00 ± 0.4	29.65 ± 0.9	80.47 ± 2.0	71.98 ± 2.6	49.87 ± 4.5	99.46 ± 1.0	76.95 ± 1.7	72.89 ± 7.1	98.07 ± 2.1	9.12 ± 6.3	72.05 ± 0.9	74.29 ± 1.4
	KNN	97.36 ± 0.2	13.51 ± 1.4	95.52 ± 1.1	27.71 ± 6.3	93.21 ± 2.2	51.66 ± 17.0	98.04 ± 0.5	10.43 ± 3.3	99.70 ± 0.7	1.92 ± 4.7	**88.24 ± 0.7**	51.95 ± 4.0
	Max Logit	94.08 ± 0.7	27.48 ± 2.1	97.31 ± 1.5	15.60 ± 11.0	**96.96 ± 0.8**	**15.90 ± 4.8**	98.20 ± 1.4	10.34 ± 10.1	96.09 ± 6.8	20.58 ± 18.2	81.65 ± 1.9	61.14 ± 1.8
	Energy	93.82 ± 0.8	29.15 ± 2.8	97.10 ± 1.8	17.47 ± 13.5	96.80 ± 0.9	16.88 ± 5.1	98.26 ± 1.5	9.99 ± 10.6	96.13 ± 6.7	21.79 ± 20.4	81.07 ± 2.0	64.54 ± 2.3
	Gradnorm	86.17 ± 2.9	46.07 ± 7.8	82.81 ± 10.6	55.21 ± 18.2	74.52 ± 11.0	73.99 ± 12.1	95.79 ± 4.2	24.50 ± 22.9	97.25 ± 4.5	14.87 ± 21.6	60.78 ± 3.7	87.58 ± 3.6
	ViM	**98.07 ± 0.2**	**9.84 ± 1.5**	96.88 ± 0.5	18.02 ± 5.5	92.00 ± 0.8	63.10 ± 7.3	97.70 ± 0.5	11.15 ± 3.8	99.85 ± 0.1	0.04 ± 0.1	85.30 ± 1.4	56.84 ± 3.4
	Mahal	91.55 ± 0.4	31.71 ± 0.8	78.10 ± 5.7	75.91 ± 10.1	60.34 ± 9.6	98.88 ± 1.7	70.77 ± 4.0	79.77 ± 7.1	96.66 ± 2.9	12.67 ± 5.6	77.93 ± 1.2	71.53 ± 2.4

We show the mean ±2 SD. over 5 independent training runs. Bold indicates best performance, underline 2nd or 3rd best

Table 5 %AUROC and %FPR@95 results for single pre-trained ImageNet-1k models

Model		Method	ID✗		OOD mean		Openimage-O		iNaturalist		Textures	
			AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓
ResNetV2-101 ID %Error: 22.63	SIRC	(MSP, $\|z\|_1$)	86.17	63.37	90.95	34.42	90.25	47.10	94.37	29.14	88.74	43.86
		(MSP, Res.)	86.31	62.36	93.63	22.76	92.69	34.26	94.60	27.14	96.80	10.04
		(MSP, KNN)	86.30	62.30	93.03	25.19	91.69	39.33	94.33	29.62	94.40	18.41
		(DR, $\|z\|_1$)	85.36	66.05	91.25	32.57	90.35	47.21	94.75	26.92	89.19	41.71
		(DR, Res.)	85.55	64.65	94.06	20.37	**93.25**	**31.27**	94.99	24.27	97.15	8.06
		(DR, KNN)	85.52	65.23	93.53	22.77	92.14	37.89	94.60	29.05	95.06	15.12
		($-\mathcal{H}$, $\|z\|_1$)	83.43	69.61	91.75	33.89	90.46	51.55	94.53	33.10	89.06	47.03
		($-\mathcal{H}$, Res.)	83.50	68.94	94.33	23.04	92.78	40.25	94.63	33.52	97.05	10.16
		($-\mathcal{H}$, KNN)	83.44	68.86	93.59	26.36	91.61	46.97	94.24	37.04	94.61	20.37
	SIRC+	(MSP, KNN, Res., $\|z\|_1$)	85.88	64.52	94.23	20.15	92.92	33.19	94.99	25.00	97.39	7.79
		(DR, KNN, Res., $\|z\|_1$)	84.87	71.02	94.51	**18.16**	93.20	32.35	**95.17**	**23.23**	97.64	5.93
		($-\mathcal{H}$, KNN, Res., $\|z\|_1$)	83.18	70.42	94.83	20.74	93.05	38.49	95.06	30.84	97.62	7.95
		MSP	**86.35**	**61.92**	90.09	38.78	90.13	47.49	93.70	32.99	87.04	51.88
		DOCTOR	85.67	64.52	90.48	37.46	90.33	47.66	93.95	32.04	87.27	52.66
		$-\mathcal{H}$	83.49	69.09	91.07	38.24	90.23	54.07	93.80	38.95	87.47	54.92
		$\|z\|_1$	47.75	95.46	69.79	67.10	53.48	87.82	73.95	78.05	73.89	66.14
		Residual	50.18	94.87	87.03	45.97	80.17	68.36	76.76	80.57	97.67	11.01
		KNN	54.05	94.53	90.29	38.84	78.72	71.08	76.21	79.99	97.78	11.12
		Max Logit	77.26	71.05	91.11	38.57	88.12	59.64	91.87	48.87	87.08	55.70
		Energy	74.68	77.14	90.27	43.03	85.86	68.87	89.27	59.78	85.85	61.59
		Gradnorm	64.65	88.00	85.19	47.56	73.53	76.05	87.99	53.66	85.04	50.85
		ViM	70.30	86.88	**95.51**	22.49	92.08	41.80	91.67	47.46	**99.17**	**3.39**
		Mahal	56.82	93.95	90.63	42.52	86.43	61.38	85.09	73.13	98.19	9.19
DenseNet-121 ID %Error: 25.58	SIRC	(MSP, $\|z\|_1$)	85.99	63.14	90.49	30.65	90.93	39.50	95.36	21.61	89.65	37.34
		(MSP, Res.)	85.97	63.33	90.92	29.55	91.17	38.58	94.08	27.50	93.38	22.93
		(MSP, KNN)	**86.13**	62.82	91.27	27.70	91.17	38.45	94.39	26.08	92.33	25.78
		(DR, $\|z\|_1$)	85.77	64.51	90.98	27.62	91.55	36.00	95.98	17.81	90.32	33.12
		(DR, Res.)	85.72	65.09	91.31	26.65	91.72	35.28	94.50	24.33	93.85	19.42
		(DR, KNN)	85.90	64.09	91.80	24.57	91.78	34.70	94.82	22.78	93.04	21.72
		($-\mathcal{H}$, $\|z\|_1$)	84.90	67.31	91.78	25.80	92.41	34.47	**96.52**	**16.28**	91.05	32.85
		($-\mathcal{H}$, Res.)	84.85	67.87	92.31	24.20	92.64	34.09	95.67	20.33	94.42	19.07
		($-\mathcal{H}$, KNN)	84.97	66.96	92.48	23.65	92.60	34.20	95.88	19.28	93.42	22.91
	SIRC+	(MSP, KNN, Res., $\|z\|_1$)	85.81	64.69	92.14	23.89	92.16	33.64	95.21	21.24	95.03	15.70

Table 5 continued

Model	Method	ID✗		OOD mean		Openimage-O		iNaturalist		Textures	
		AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓
	(DR, KNN, Res., $\|z\|_1$)	85.47	66.75	92.53	21.15	92.70	**30.14**	95.66	17.46	95.44	**13.29**
	($-\mathcal{H}$, KNN, Res., $\|z\|_1$)	84.68	68.91	**93.19**	**20.75**	**93.39**	30.43	96.34	16.29	95.82	14.11
	MSP	86.11	**62.67**	89.79	34.54	90.26	43.08	94.26	27.56	88.31	43.72
	DOCTOR	85.93	63.43	90.27	31.73	90.82	39.93	94.83	23.95	88.85	41.01
	$-\mathcal{H}$	84.97	66.76	91.36	28.00	91.91	37.18	95.83	19.56	90.08	37.56
	$\|z\|_1$	47.53	94.93	79.08	54.68	69.94	70.15	89.06	39.14	84.61	49.73
	Residual	51.52	94.26	73.41	66.12	69.78	78.27	61.14	93.61	90.21	33.64
	KNN	52.73	95.71	86.28	52.37	74.30	79.50	62.50	94.18	94.40	26.05
	Max Logit	77.97	71.35	92.51	25.98	92.19	38.48	96.07	20.57	91.59	34.32
	Energy	76.13	75.77	92.38	27.01	91.54	42.66	95.60	23.50	91.39	35.43
	Gradnorm	55.44	92.10	86.34	40.70	78.97	58.55	93.87	25.24	89.62	37.81
	ViM	70.16	88.53	90.43	44.51	88.40	56.49	88.74	66.34	**96.64**	17.69
	Mahal	57.26	94.15	68.29	83.16	69.08	86.75	50.15	97.75	82.82	55.47

Model		Method	SpaceNet		Colonoscopy		Colorectal		Noise		ImageNet-O	
			AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓
ResNetV2-101 ID %Error: 22.63	SIRC	(MSP, $\|z\|_1$)	96.98	15.71	97.10	16.69	93.95	30.28	99.26	4.32	66.92	88.25
		(MSP, Res.)	98.81	5.58	97.18	14.70	96.88	14.04	99.99	0.00	72.07	76.35
		(MSP, KNN)	98.41	8.66	98.13	11.03	97.41	12.28	99.99	0.05	69.92	82.15
		(DR, $\|z\|_1$)	96.93	14.56	97.13	15.25	94.85	23.82	99.48	2.93	67.33	88.15
		(DR, Res.)	99.08	3.72	97.36	12.44	97.52	10.08	99.99	0.00	73.12	73.10
		(DR, KNN)	98.79	6.05	98.53	7.84	98.20	7.76	100.00	0.00	70.94	78.50
		($-\mathcal{H}$, $\|z\|_1$)	97.16	14.06	98.37	9.66	95.72	23.16	99.48	3.80	69.25	88.75
		($-\mathcal{H}$, Res.)	98.96	5.08	98.33	9.02	98.13	8.62	100.00	0.00	74.76	77.70
		($-\mathcal{H}$, KNN)	98.46	8.23	98.85	6.89	98.44	7.88	100.00	0.03	72.49	83.50
	SIRC+	(MSP, KNN, Res., $\|z\|_1$)	99.11	3.43	97.37	12.91	98.37	6.40	100.00	0.00	73.71	72.50
		(DR, KNN, Res., $\|z\|_1$)	99.21	1.93	97.38	9.99	98.69	3.84	100.00	0.00	74.83	68.00
		($-\mathcal{H}$, KNN, Res., $\|z\|_1$)	99.23	3.01	98.33	8.50	99.00	3.94	**100.00**	**0.00**	76.34	73.20
		MSP	96.62	17.64	97.47	14.92	91.55	44.24	97.37	12.53	66.86	88.55
		DOCTOR	96.85	16.28	97.89	12.63	92.34	39.78	97.94	9.74	67.25	88.90
		$-\mathcal{H}$	96.80	16.78	98.52	8.70	94.02	34.70	98.46	7.71	69.24	90.10
		$\|z\|_1$	62.63	69.91	58.42	89.18	86.16	51.20	99.61	1.36	50.15	93.10
		Residual	97.09	17.78	67.58	98.55	95.43	25.32	99.95	0.00	81.57	66.20

Table 5 continued

Model		Method	SpaceNet		Colonoscopy		Colorectal		Noise		ImageNet-O	
			AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓
		KNN	97.04	18.60	92.44	57.34	98.59	6.70	100.00	0.00	81.56	65.90
		Max Logit	97.18	16.64	99.04	4.64	96.25	25.94	98.79	6.47	70.57	90.70
		Energy	96.26	27.62	**99.19**	**3.17**	96.56	23.82	98.83	6.83	70.33	92.55
		Gradnorm	87.56	57.50	94.56	29.39	95.95	22.54	99.82	0.67	57.06	89.85
		ViM	**99.43**	**0.54**	95.59	29.94	**99.30**	**1.26**	100.00	0.00	**86.80**	**55.55**
		Mahal	97.72	12.45	77.36	98.76	96.09	22.40	99.88	0.00	84.28	62.85
DenseNet-121 ID %Error: 25.58	SIRC	(MSP, $\|z\|_1$)	97.34	14.13	96.79	17.15	96.06	20.94	99.74	1.10	58.10	93.45
		(MSP, Res.)	97.05	15.03	96.18	19.71	95.51	23.60	99.67	0.60	60.35	88.45
		(MSP, KNN)	97.57	13.31	97.56	12.63	97.07	15.54	99.90	0.49	60.20	89.35
		(DR, $\|z\|_1$)	97.86	10.34	97.10	14.22	96.88	15.68	99.79	0.83	58.36	93.00
		(DR, Res.)	97.46	11.56	96.30	16.79	96.21	17.90	99.62	0.54	60.83	87.35
		(DR, KNN)	98.11	9.77	98.02	8.96	97.86	10.28	99.92	0.34	60.83	88.00
		($-\mathcal{H}$, $\|z\|_1$)	98.48	7.33	97.89	9.86	97.79	12.36	99.83	0.68	60.31	92.60
		($-\mathcal{H}$, Res.)	98.26	8.34	97.45	11.37	97.56	12.30	99.79	0.39	62.71	87.70
		($-\mathcal{H}$, KNN)	98.62	7.12	**98.31**	**7.41**	98.40	8.86	99.95	0.24	62.67	89.20
	SIRC+	(MSP, KNN, Res., $\|z\|_1$)	98.07	9.02	96.75	15.04	97.76	10.52	99.91	0.20	62.23	85.75
		(DR, KNN, Res., $\|z\|_1$)	98.38	6.44	96.93	11.71	98.23	6.78	99.91	0.12	62.95	83.25
		($-\mathcal{H}$, KNN, Res., $\|z\|_1$)	98.77	**5.30**	97.72	9.15	98.71	5.68	99.94	0.09	64.82	84.95
		MSP	96.60	18.93	96.90	17.10	94.44	30.72	99.55	1.69	57.97	93.55
		DOCTOR	97.20	14.70	97.33	13.52	95.24	25.94	99.64	1.37	58.23	93.45
		$-\mathcal{H}$	98.15	9.23	97.95	9.42	97.00	17.20	99.76	1.15	60.17	92.70
		$\|z\|_1$	83.12	60.68	58.85	89.84	92.89	34.90	99.88	0.49	54.29	92.50
		Residual	83.61	63.65	37.94	99.37	75.49	70.74	97.32	14.40	71.83	75.25
		KNN	90.86	51.13	88.01	82.20	96.91	19.86	99.80	0.11	**83.48**	**65.95**
		Max Logit	**98.78**	5.76	98.20	8.77	98.62	6.48	99.89	0.49	64.77	92.95
		Energy	98.76	5.94	97.87	11.18	**98.86**	**5.02**	99.91	0.39	65.12	91.95
		Gradnorm	93.58	31.31	81.08	68.36	97.63	13.10	**99.96**	0.09	56.04	91.15
		ViM	96.38	21.43	82.83	89.17	95.19	31.76	99.57	**0.01**	75.66	73.20
		Mahal	63.65	93.88	66.15	97.28	58.79	96.32	75.23	68.31	80.48	69.50

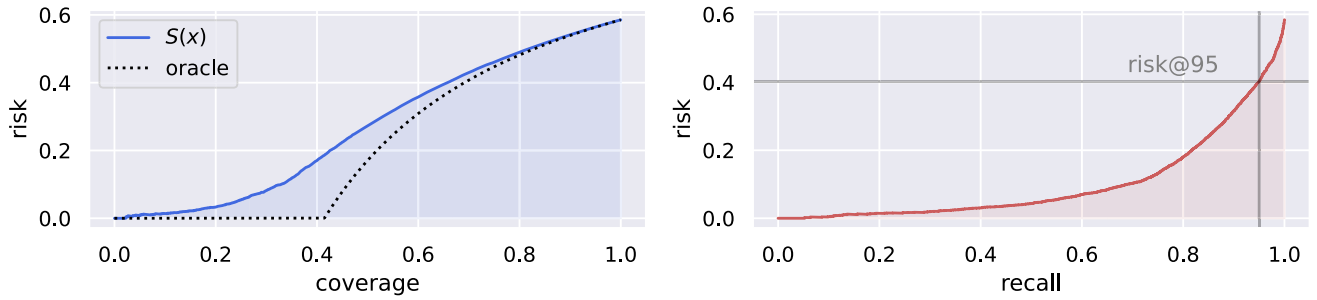

Fig. 17 Visualisations of different evaluation metrics for SCOD. We aim to minimise risk over different selection thresholds t. Left: Risk-Coverage curve (coverage is the proportion of all data accepted). We aggregate performance over t by taking the area under the curve. The oracle represents perfect separation of OOD, ID✗ |ID✓. Right: Risk-Recall curve. We consider both the area under the curve as well as risk@recall=0.95

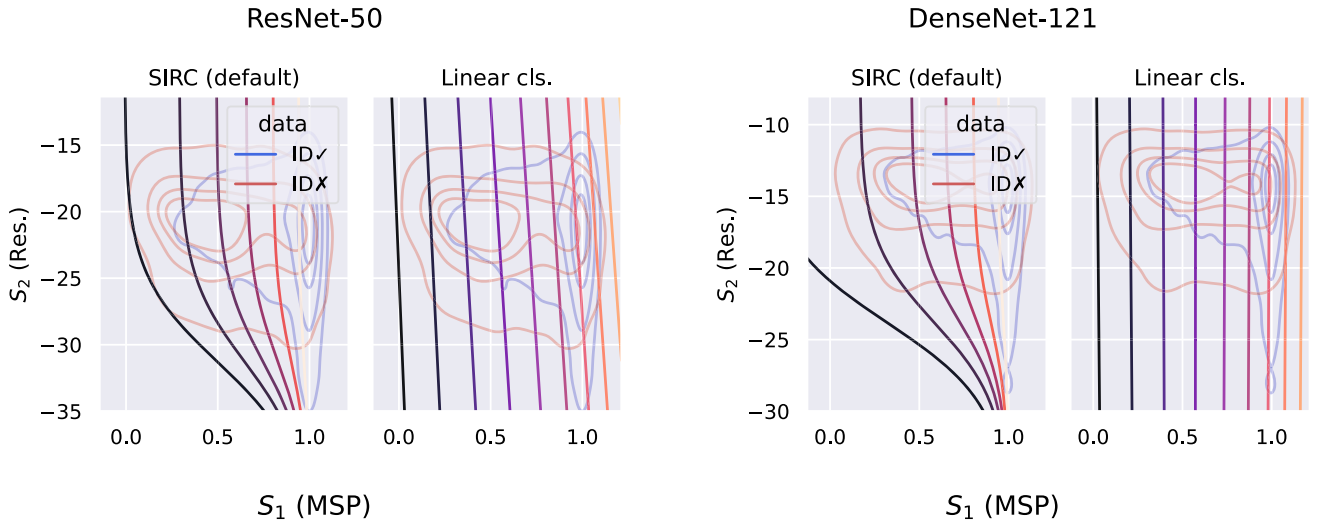

Fig. 18 2D KDE plots of ID✗ and ID✓ samples on ImageNet-200 with decision boundaries from SIRC as well as linear logistic classifiers (decision boundary at 0.5) trained with different class weightings

configurations (Figs. 27, 28, 29, 30). If there are multiple training runs, we plot the distributions corresponding to the outputs of the 1st run. Decision contours corresponding to the default parameter setting for SIRC are also overlayed. We note that the inconsistency of Residual can be observed here, where in some cases the OOD distribution is much lower than ID, whilst in others, there is almost complete overlap. In the case of MobileNetV2 on iNaturalist it is in fact higher for OOD than ID, although the nature of SIRC means that it is robust to such S_2 failure (as discussed in Sect. 5.2).

B.5 SIRC+ on Other Architectures

As in Fig. 14, we plot the change in SCOD performance relative to only using S_1 $(-\mathcal{H})$, for CNN architectures DenseNet-121 and MobileNet-V2 (Figs. 31, 32). Results tell a similar story to Sect. 7, with SIRC+ providing more consistent and overall better improvements. We note that for DenseNet-121 there is a slightly larger drop in ID✗ |ID✓ performance compared to the other two architectures for SIRC+. From the perspective of a practitioner, this cost should be visible on a validation set, and so the trade-off between ID✗ and OOD should be considered when choosing which version of SIRC to deploy.

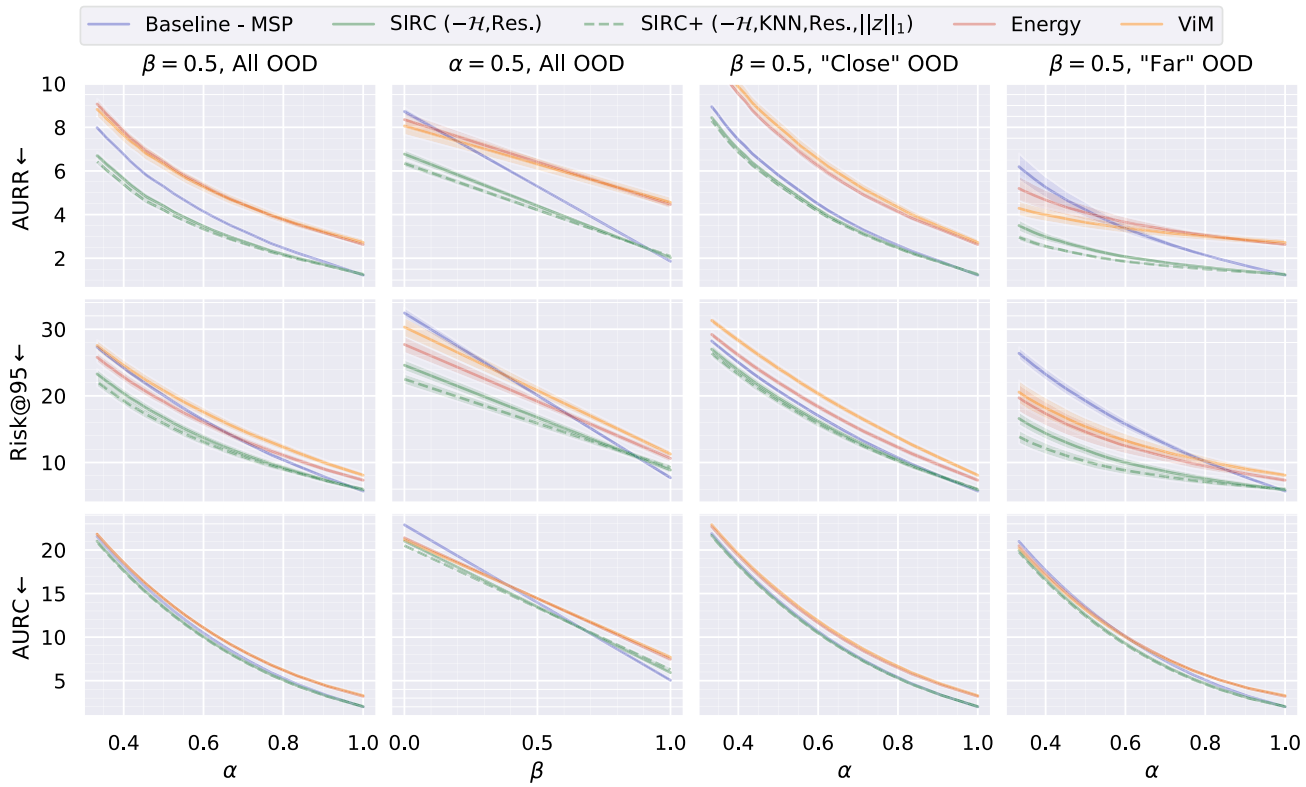

Fig. 19 Varying α and β for ResNet-50 (ImageNet-200) (values $\times 10^2$)

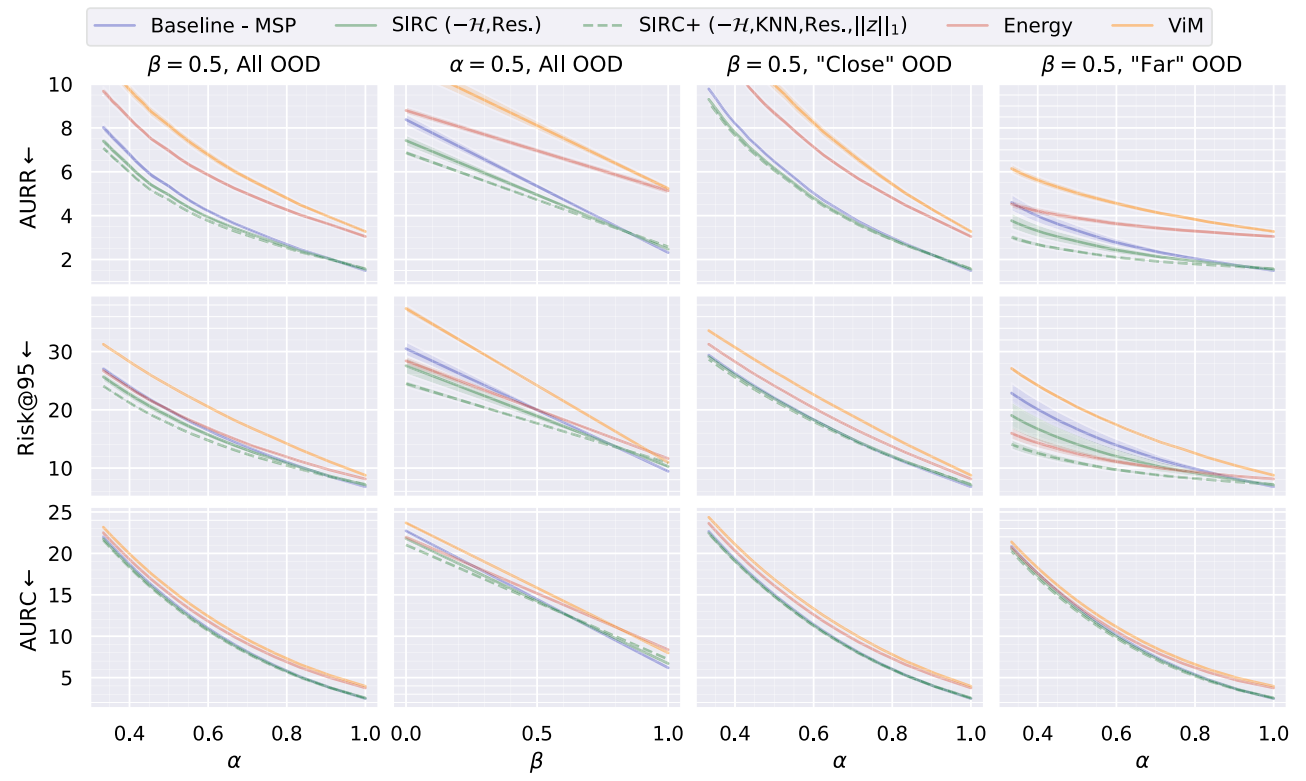

Fig. 20 Varying α and β for MobileNetV2 (ImageNet-200) (values $\times 10^2$)

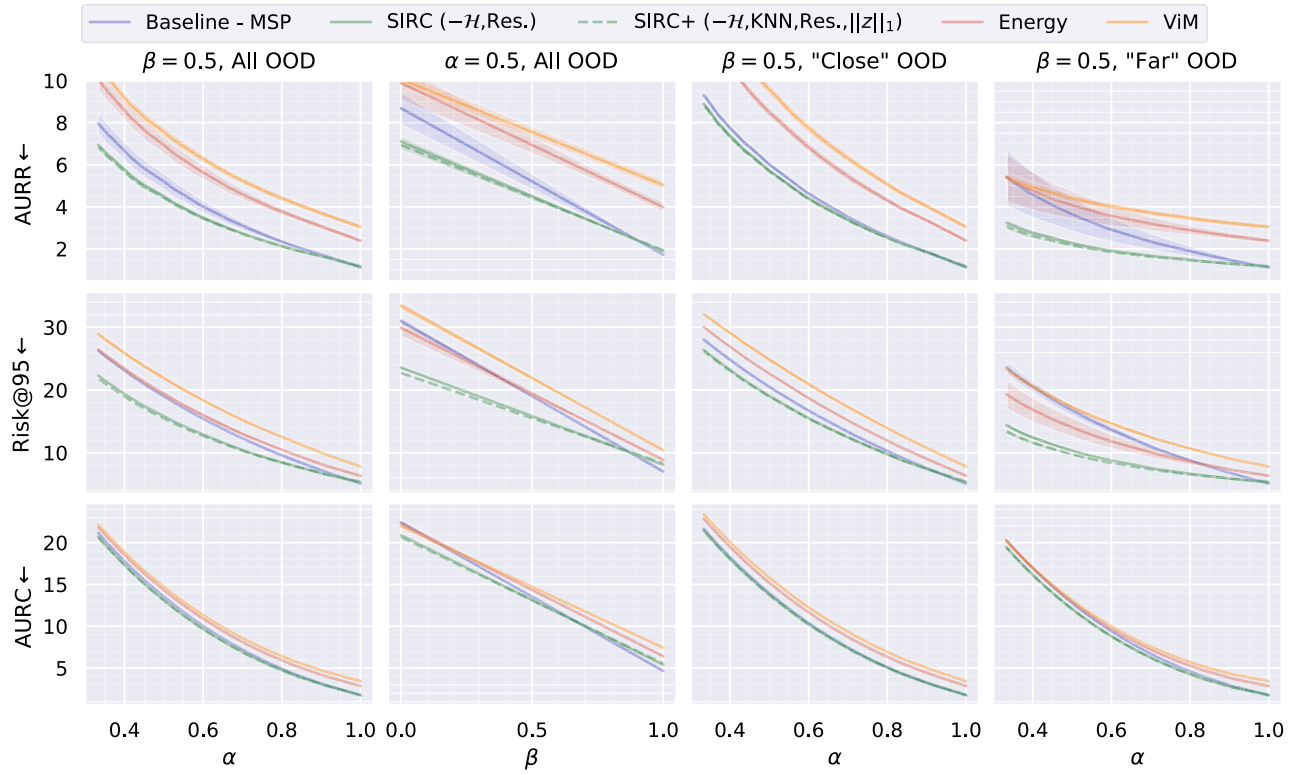

Fig. 21 Varying α and β for DenseNet-121 (ImageNet-200) (values $\times 10^2$)

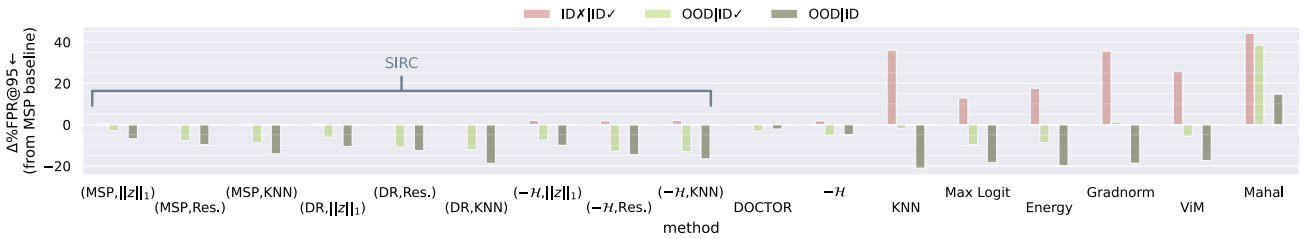

Fig. 22 ResNet-50 (ImageNet-200), comparing the change in %FPR@95 relative to the MSP baseline for different detection methods and data groups

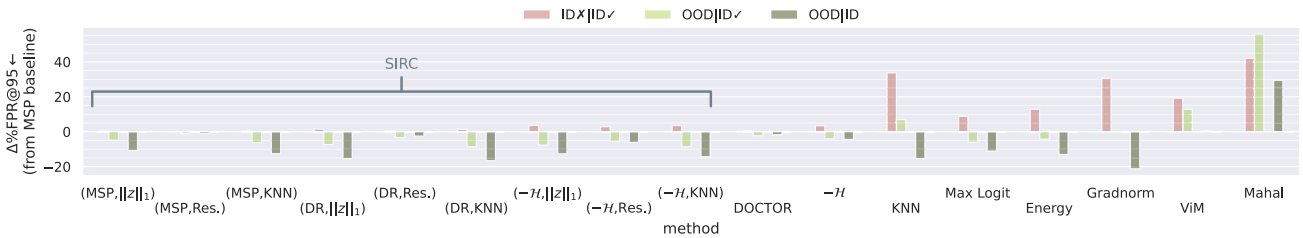

Fig. 23 MobileNetV2 (ImageNet-200), comparing the change in %FPR@95 relative to the MSP baseline for different detection methods and data groups

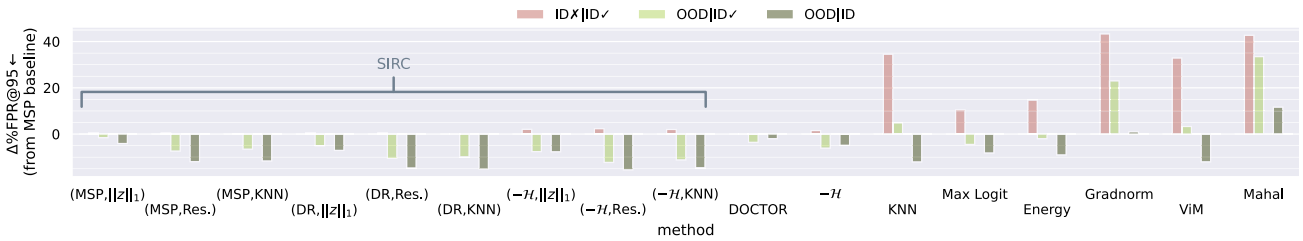

Fig. 24 DenseNet-121 (ImageNet-200), comparing the change in %FPR@95 relative to the MSP baseline for different detection methods and data groups

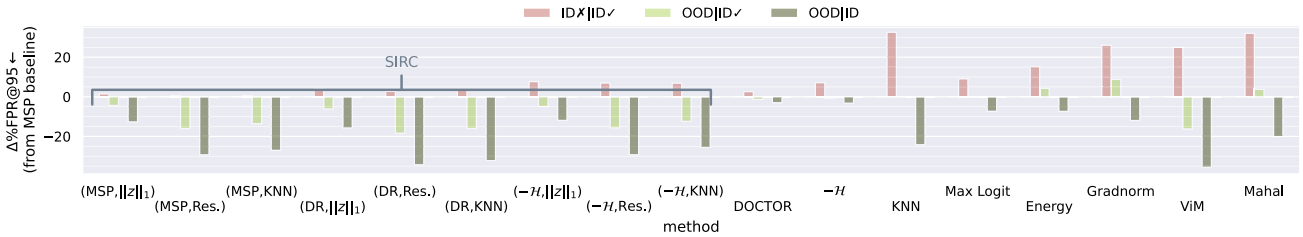

Fig. 25 ResNetV2-101 (ImageNet-1k), comparing the change in %FPR@95 relative to the MSP baseline for different detection methods and data groups

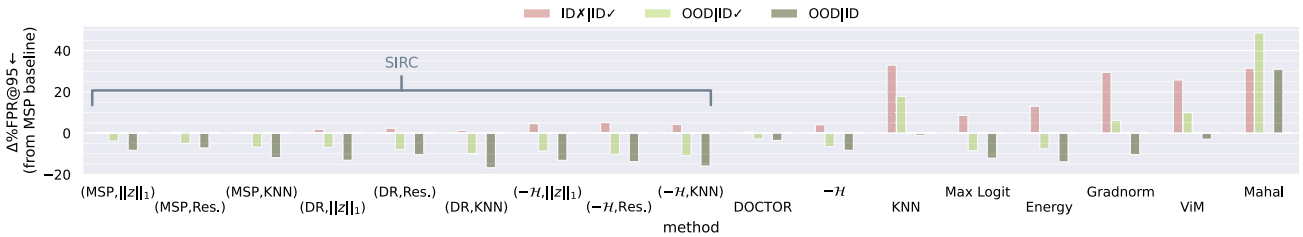

Fig. 26 DenseNet-121 (ImageNet-1k), comparing the change in %FPR@95 relative to the MSP baseline for different detection methods and data groups

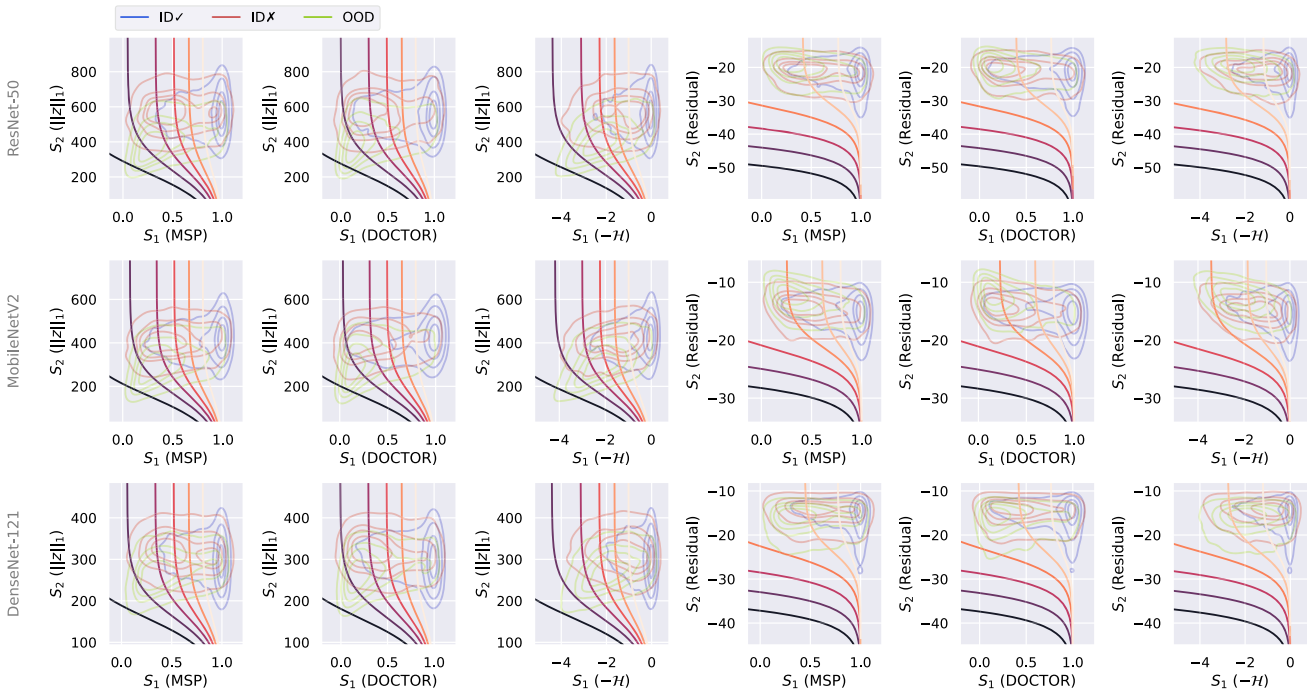

Fig. 27 SIRC combinations on the S_1, S_2-plane, ID: ImageNet-200, OOD: iNaturalist

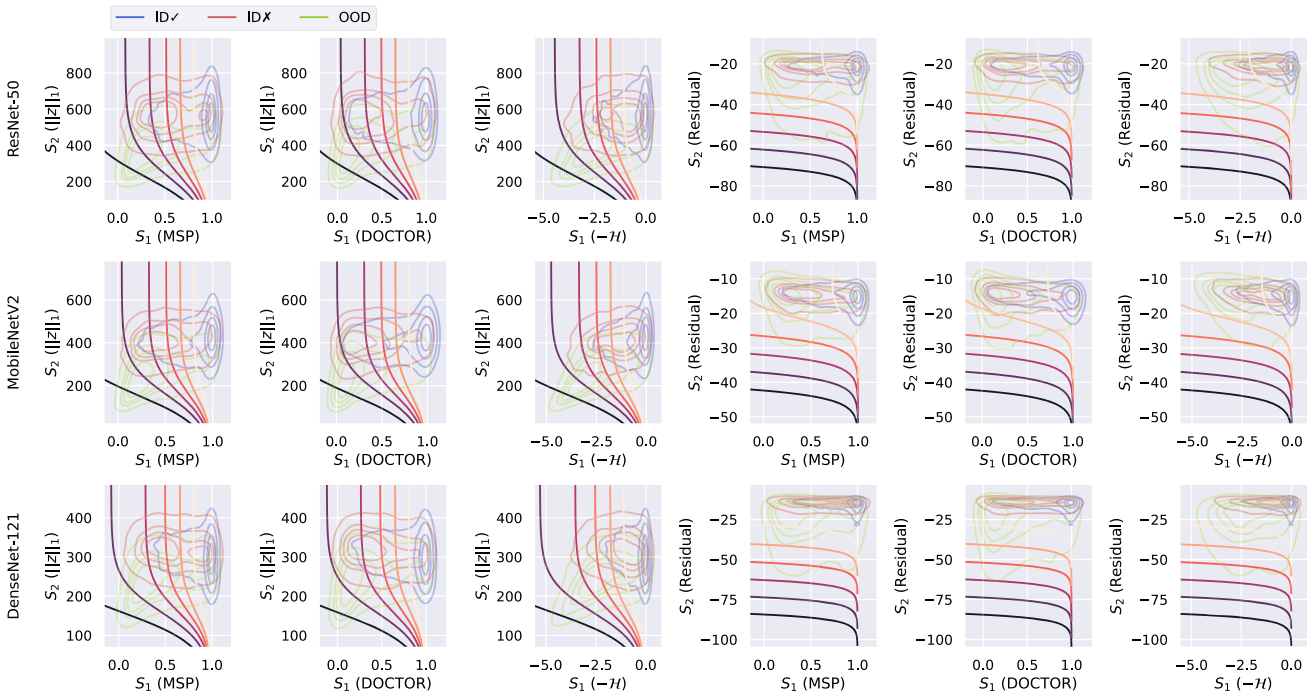

Fig. 28 SIRC combinations on the S_1, S_2-plane, ID: ImageNet-200, OOD: Textures

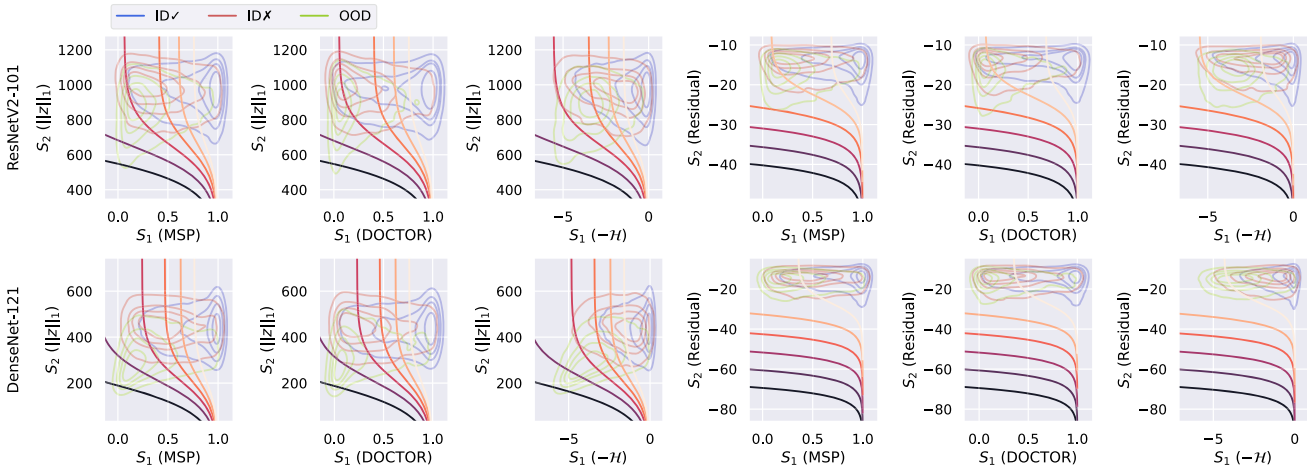

Fig. 29 SIRC combinations on the S_1, S_2-plane, ID ImageNet-1k, OOD: iNaturalist

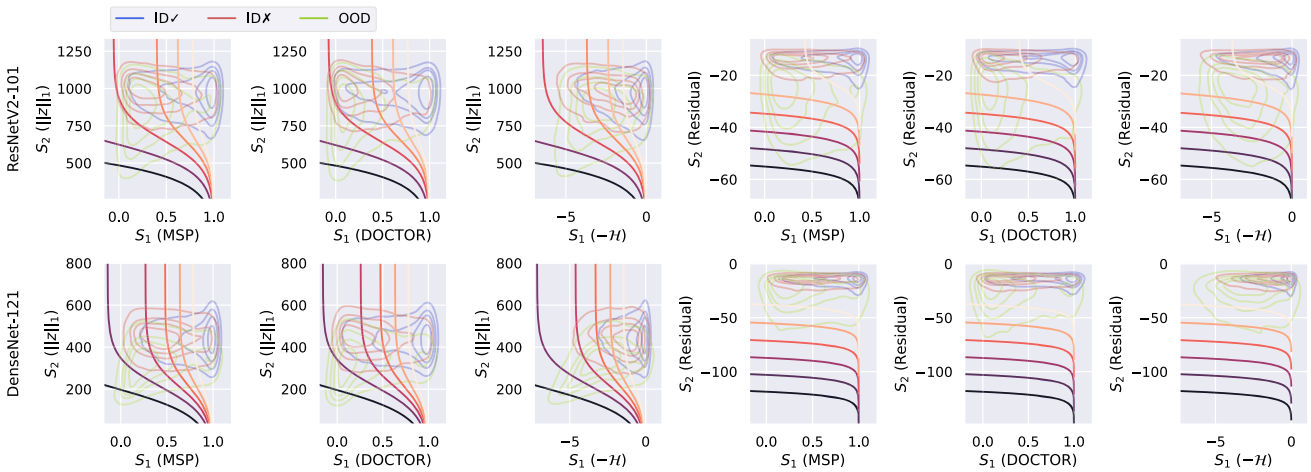

Fig. 30 SIRC combinations on the S_1, S_2-plane, ID ImageNet-1k, OOD: Textures

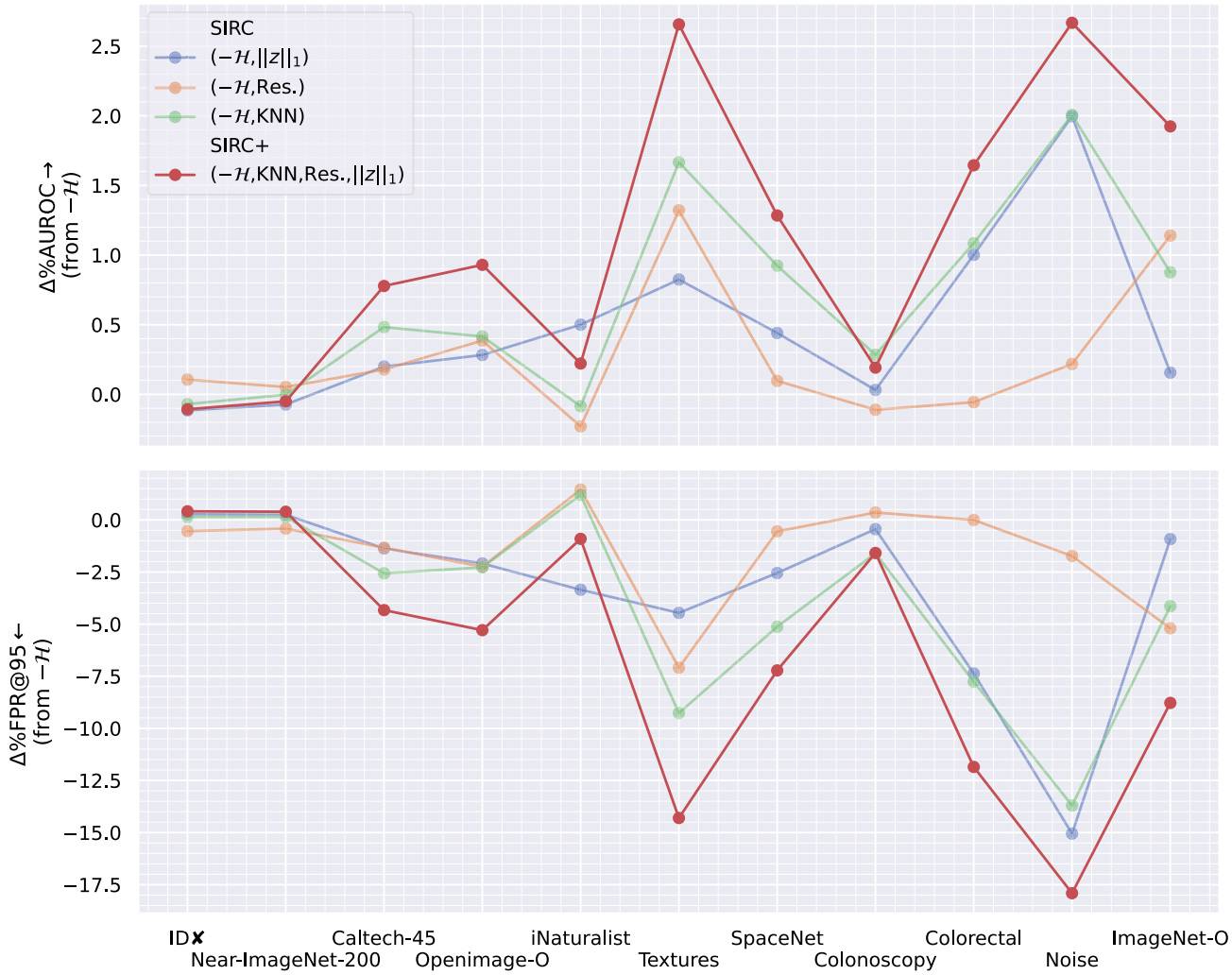

Fig. 31 Δ%AUROC and Δ%FPR@95 with respect to $-\mathcal{H}$ (S_1 only). Results are for MobileNet-V2 trained on ImageNet-200

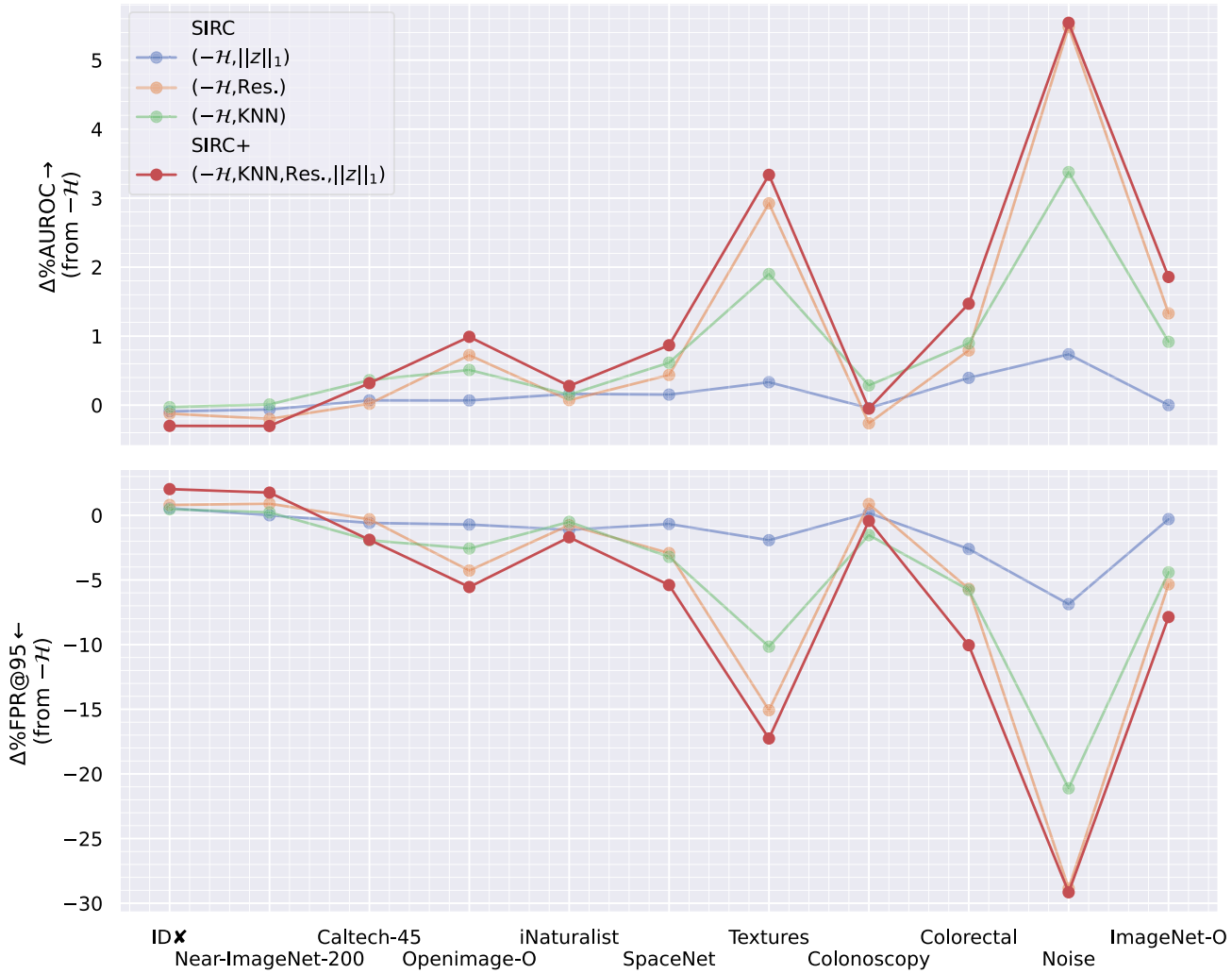

Fig. 32 Δ%AUROC and Δ%FPR@95 with respect to $-\mathcal{H}$ (S_1 only). Results are for DenseNet-121 trained on ImageNet-200

References

Caterini, A. L., & Loaiza-Ganem, G. (2021). Entropic issues in likelihood-based ood detection. arXiv:2109.10794.

Cen, J., Luan, D., Zhang, S., Pei, Y., Zhang, Y., Zhao, D., Shen, S., & Chen, Q. (2023). The devil is in the wrongly-classified samples: Towards unified open-set recognition. In *The 11th international conference on learning representations*.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., & Vedaldi, A. (2014). Describing textures in the wild. In *2014 IEEE conference on computer vision and pattern recognition* (pp. 3606–3613).

Corbière, C., Thome, N., Bar-Hen, A., Cord, M., & Pérez, P. (2019). Addressing failure prediction by learning model confidence. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 2902–2913). Curran Associates Inc.

Dhamija, A., Gunther, M., Ventura, J., & Boult, T. (2020). The overlooked elephant of object detection: Open set. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*.

Du, X., Wang, Z., Cai, M., & Li, Y. (2022). VOS: Learning what you don't know by virtual outlier synthesis. In *The 10th international conference on learning representations, ICLR 2022, virtual event, April 25–29, 2022*. OpenReview.net.

El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research, 11*, 1605–1641.

Etten, A. V., Lindenbaum, D., & Bacastow, T. M. (2018). Spacenet: A remote sensing dataset and challenge series. arXiv:1807.01232.

Fort, S., Ren, J., & Lakshminarayanan, B. (2021). Exploring the limits of out-of-distribution detection. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems 34: Annual conference on neural information processing systems 2021, NeurIPS 2021, December 6–14, 2021, virtual* (pp. 7068–7081). Curran Associates, Inc.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In M. F. Balcan, & K. Q. Weinberger (Eds.), *Proceedings of the 33rd international conference on machine learning, volume 48 of proceedings of machine learning research* (pp. 1050–1059). PMLR.

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R., (Eds.), *Advances in neural information processing systems*, (Vol. 30). Curran Associates, Inc.

Geifman, Y., & El-Yaniv, R. (2019). Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning* (pp. 2151–2159). PMLR.

Geifman, Y., Uziel, G., & El-Yaniv, R. (2019). Bias-reduced uncertainty estimation for deep neural classifiers. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, conference track proceedings*.

Granese, F., Romanelli, M., Gorla, D., Palamidessi, C., & Piantanida, P. (2021). DOCTOR: A simple method for detecting misclassification errors. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems 34: Annual conference on neural information processing systems 2021, NeurIPS 2021, December 6–14, 2021, virtual* (pp. 5669–5681). Curran Associates, Inc.

Griffin, G., Holub, A., & Perona, P. (2007). Caltech-256 object category dataset.

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

Hendrycks, D., & Dietterich, T. G. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *7th international conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net.

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, conference track proceedings*. OpenReview.net.

Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., & Song, D. (2022). Scaling out-of-distribution detection for real-world settings. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th international conference on machine learning, volume 162 of proceedings of machine learning research* (pp. 8759–8773). PMLR.

Hendrycks, D., Mazeika, M., & Dietterich, T. G. (2019). Deep anomaly detection with outlier exposure. In *7th international conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. X. (2021). Natural adversarial examples. *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 15257–15266).

Hsu, Y.-C., Shen, Y., Jin, H., & Kira, Z. (2020). Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 10948–10957).

Huang, R., & Li, Y. (2021). Mos: Towards scaling out-of-distribution detection for large semantic space. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 8706–8715).

Huang, R., Geng, A., & Li, Y. (2021). On the importance of gradients for detecting distributional shifts in the wild. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems 34: Annual conference on neural information processing systems 2021, NeurIPS 2021, December 6–14, 2021, virtual* (pp. 677–689). Curran Associates, Inc.

Huang, G., Liu, Z., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2261–2269).

Jaeger, P. F., Lüth, C. T., Klein, L., & Bungert, T. J. (2023). A call to reflect on evaluation practices for failure detection in image classification. In *The 11th international conference on learning representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023*. OpenReview.net.

Jospin, L. V., Laga, H., Boussaid, F., Buntine, W., & Bennamoun, M. (2022). Hands-on Bayesian neural networksG-a tutorial for deep learning users. *IEEE Computational Intelligence Magazine, 17*(2), 29–48.

Kamath, A., Jia, R., & Liang, P. (2020). Selective question answering under domain shift. In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics, ACL 2020, Online, July 5–10, 2020* (pp. 5684–5696). Association for Computational Linguistics.

Kather, J. N., Weis, C.-A., Bianconi, F., Melchers, S. M., Schad, L. R., Gaiser, T., Marx, A., & Zöllner, F. G. (2016). Multi-class texture analysis in colorectal cancer histology. *Scientific Reports*, 6.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st international conference on neural information processing systems, NIPS'17* (pp. 5580–5590). Curran Associates Inc.

Kim, J., Koo, J., & Hwang, S. (2021). A unified benchmark for the unknown detection capability of deep neural networks. arXiv:2112.00337.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (2020). Big transfer (bit): General visual representation learning. lecture notes in computer scienceIn A. Vedaldi, H. Bischof, T. Brox, & J. Frahm (Eds.), *Computer vision–ECCV 2020–16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, Part V* (Vol. 12350, pp. 491–507). Springer.

Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., & Murphy, K. (2017). Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github.com/openimages.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, CA, USA* (pp. 6402–6413). Curran Associates, Inc.

Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada* (pp. 7167–7177). Curran Associates, Inc.

Liang, S., Li, Y., & Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, conference track proceedings*. OpenReview.net.

Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., & Lakshminarayanan, B. (2020a). Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin, (Eds.), *Advances in neural information processing systems*, (Vol. 33, pp. 7498–7512). Curran Associates, Inc.

Liu, W., Wang, X., Owens, J., & Li, Y. (2020b). Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information pro-*

cessing systems (Vol. 33, pp. 21464–21475). Curran Associates, Inc.

MacKay, D. J. C. (1995). Probable networks and plausible predictions—a review of practical bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469.

Malinin, A., & Gales, M. J. F. (2018). Predictive uncertainty estimation via prior networks. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett, (Eds.), *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada* (pp. 7047–7058). Curran Associates, Inc.

Malinin, A., & Gales, M. J. F. (2021). Uncertainty estimation in autoregressive structured prediction. In *9th international conference on learning representations, ICLR 2021, virtual event, Austria, May 3–7, 2021*. OpenReview.net.

Malinin, A., Band, N., Gal, Y., Gales, M., Ganshin, A., Chesnokov, G., Noskov, A., Ploskonosov, A., Prokhorenkova, L., Provilkov, I., Raina, V., Raina, V., Roginskiy, D., Shmatova, M., Tigas, P., & Yangel, B. (2021). Shifts: A dataset of real distributional shift across multiple large-scale tasks. In *35th conference on neural information processing systems datasets and benchmarks track (round 2)*.

Malinin, A., Mlodozeniec, B., & Gales, M. J. F. (2020). Ensemble distribution distillation. In *8th international conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*. OpenReview.net.

Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O., Beorchia, S., Poincloux, L., & Bartoli, A. (2016). Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE Transactions on Medical Imaging, 35*(9), 2051–2063.

Moon, J., Kim, J., Shin, Y., & Hwang, S. (2020). Confidence-aware learning for deep neural networks. In *Proceedings of the 37th international conference on machine learning, ICML 2020, 13–18 July 2020, virtual event, volume 119 of Proceedings of machine learning research* (pp. 7034–7044). PMLR.

Mukhoti, J., Kirsch, A., van Amersfoort, J. R., Torr, P. H. S., & Gal, Y. (2021). Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. arXiv:2102.11582.

Nalisnick, E. T., Matsukawa, A., Teh, Y. W., Görür, D., & Lakshminarayanan, B. (2019). Do deep generative models know what they don't know? In *7th international conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc.

Pearce, T., Brintrup, A., & Zhu, J. (2021). Understanding softmax confidence and uncertainty. arXiv:2106.04972.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., & Lakshminarayanan, B. (2019). Likelihood ratios for out-of-distribution detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*, 211–252.

Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 4510–4520).

Sun, Y., Guo, C., & Li, Y. (2021). React: Out-of-distribution detection with rectified activations. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems 34: Annual conference on neural information processing systems 2021, NeurIPS 2021, December 6–14, 2021, virtual* (pp. 144–157). Curran Associates, Inc.

Sun, Y., Ming, Y., Zhu, X., & Li, Y. (2022). Out-of-distribution detection with deep nearest neighbors. *162*, 20827–20840.

Techapanurak, E., Suganuma, M., & Okatani, T. (2020). Hyperparameter-free out-of-distribution detection using cosine similarity. In *Proceedings of the Asian conference on computer vision (ACCV)*.

Ulmer, D. T., Hardmeier, C., & Frellsen, J. (2023). Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. In *Transactions on machine learning research: OpenReview.net*.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., & Belongie, S. (2017). The inaturalist species classification and detection dataset.

Wang, H., Li, Z., Feng, L., & Zhang, W. (2022). Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 4921–4930).

Wilson, A. G. (2020). The case for bayesian deep learning. arXiv:2001.10995.

Xia, G., & Bouganis, C.-S. (2022a). Augmenting softmax information for selective classification with out-of-distribution data. In *Proceedings of the Asian conference on computer vision (ACCV)* (pp. 1995–2012).

Xia, G., & Bouganis, C.-S. (2022b). On the usefulness of deep ensemble diversity for out-of-distribution detection. arXiv:2207.07517.

Xia, G., & Bouganis, C.-S. (2023). Window-based early-exit cascades for uncertainty estimation: When deep ensembles are more efficient than single models. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*.

Yang, J., Zhou, K., Li, Y., & Liu, Z. (2021). Generalized out-of-distribution detection: A survey. arXiv:2110.11334.

Zhang, M., Zhang, A., & McDonagh, S. (2021). On the out-of-distribution generalization of probabilistic image modelling. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems 34: Annual conference on neural information processing systems 2021, NeurIPS 2021, December 6–14, 2021, virtual* (pp. 3811–3823). Curran Associates, Inc.