Augmenting Softmax Information for Selective Classification with Out-of-Distribution Data Supplemental Material

Guoxuan Xia and Christos-Savvas Bouganis

Imperial College London
{g.xia21,christos-savvas.bouganis}@imperial.ac.uk

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 [4] using the default hyperparameters found in PyTorch's examples.[1] 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 [18] 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 [8]. 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[2] [13], specifically `BiT-S-R101x1`, and DenseNet-121 provided by PyTorch.[3]. 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.

A.2 ImageNet-Scale Datasets

Figure 1 shows a number of random examples from each dataset introduced in the main paper, 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 [7, 10–12, 19]. We note that there is a slight discrepancy between the number of samples reported in [12]

[1] `https://github.com/pytorch/examples/tree/main/imagenet`
[2] `https://github.com/google-research/big_transfer`
[3] `https://pytorch.org/vision/stable/models.html`

for ImageNet-200 and in the authors' provided datasets,[4] but we do not believe this affects the validity of our results.

Noise We randomly generate 10000 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.[5] The resulting data thus varies in both scale and contrast (see Fig. 1).

Colonoscopy We separate out frames as individual images from videos provided in [16].[6] 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 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 $\{\text{MSP,DOCTOR,}\mathcal{H}\}\times\{||\boldsymbol{z}||_1\text{,Residual}\}$.
- Maximum Softmax Probability (MSP)[6]: a baseline score that takes the max value from the softmax $\pi_{\max} = \max_k \pi_k$.
- DOCTOR [3]: 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 [9], $||\boldsymbol{z}||_1$.
- Residual: used in ViM [9], this score measures the component of the feature vector that is outside of a principal subspace defined using ID data, $||\boldsymbol{z}^{P^\perp}||_2$. We follow [19] in setting the dimensionality of the subspace to 1000 if the

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

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

[6] `http://www.depeca.uah.es/colonoscopy_dataset/`

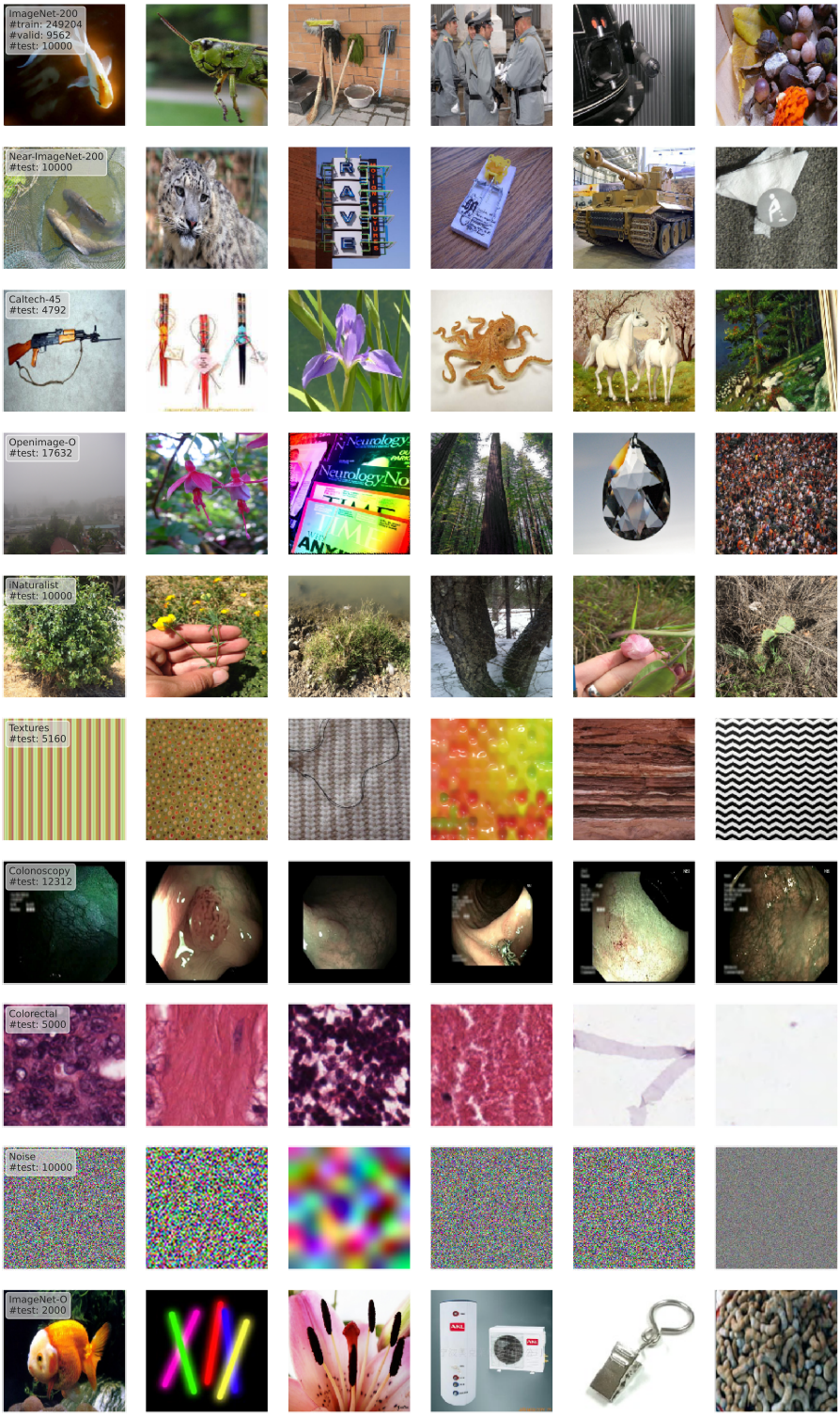

Fig. 1. Random examples from each ImageNet-scale dataset, with the #samples in each.

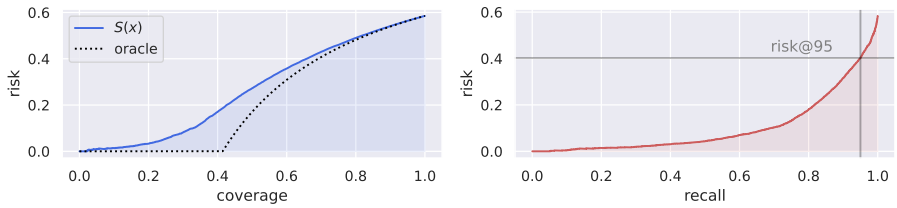

Fig. 2. Visualisations of different evaluation metrics for SCOD. We aim to minimise risk over different selection thresholds t. Left: Risk-Coverage curve (coverage is the proportion of all data accepted). We aggregate performance over t by taking the area under the curve. The oracle represents perfect separation of OOD,ID✗ |ID✓. Right: Risk-Recall curve. We consider both the area under the curve as well as risk@recall=0.95.

dimensionality of $\boldsymbol{z}$, $L > 1500$ and 512 otherwise. Like Entropy, we use the negative of the score $S = -||\boldsymbol{z}^{P^\perp}||_2$ as this score is meant to be higher for OOD data. Please refer to Wang et al. [19]'s paper for full details.

- Max Logit [5]: Max Logit is similar to MSP, but the score is taken from the logits before the softmax $v_{\max} = \max_k v_k$.
- Energy [15]: this score aggregates over all logit values as $\log \sum_k \exp v_k$.
- Gradnorm [9]: although this score was originally motivated by gradients, we can view it simply as the combination of two scores, $C = ||\boldsymbol{\pi} - \mathbf{1}/K||_1 ||\boldsymbol{z}||_1$.
- ViM [19]: this linearly combines Energy and Residual, $C = \log \sum_k \exp v_k - c||\boldsymbol{z}^{P^\perp}||_2$. 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 [14]: this score involves building a classwise gaussian mixture model over the features with tied covariance matrix. The confidence is then calculated as $-\min_k (\boldsymbol{z} - \boldsymbol{\mu}_k)^T \tilde{\boldsymbol{\Sigma}} (\boldsymbol{z} - \boldsymbol{\mu}_k)$. We use the approach in [1, 19] where only the final layer features are considered.

A.4 Evaluation Metrics

Other than the metrics specified in the main paper, we additionally use Area Under the Risk-Coverage Curve (AURC)↓, from [2, 12]. 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. Fig. 2 illustrates graphically some of the metrics we use to evaluate SCOD.

B Additional Results

We provide more complete versions of the results presented in the main paper 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. We also include mean±2std. for experiments with multiple training runs. SIRC performs as expected in all cases – a negligible reduction in ID✗ |ID✓ in exchange for a meaningful uplift in OOD|ID✓ compared to only using S_1. 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 [19]). 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. [17] 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, SVHN), have especially high variation.

B.2 Varying α and β

We plot performance under varying α and β for all 3 ImageNet-200 architectures (Figs. 3 to 5). We also present the mean±std. 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

We plot the change in %FPR@95↓ relative to the MSP baseline for all architectures and confidence scores (Figs. 6 to 10). The behaviour is as discussed in the main paper, 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

We plot different SIRC combinations on the S_1, S_2-plane for different experimental configurations (Figs. 11 to 14). If there are multiple training runs, we plot

Table 1. Full %AUROC and %FPR@95 results for all models trained on ImageNet-200. We show the mean±2std. over 5 independent training runs. **Bold** indicates best performance, underline 2nd or 3rd best.

Model		Method	ID✗ %AUROC↑	ID✗ %FPR@95↓	OOD mean %AUROC↑	OOD mean %FPR@95↓	Near-IN-200 %AUROC↑	Near-IN-200 %FPR@95↓	Caltech-45 %AUROC↑	Caltech-45 %FPR@95↓	Openimage-O %AUROC↑	Openimage-O %FPR@95↓	iNaturalist %AUROC↑	iNaturalist %FPR@95↓
ResNet-50 ID %Error: 19.01	SIRC	(MSP,$\|\|z\|\|_1$)	90.34 ±0.2	52.70 ±3.2	91.51	40.27	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
	SIRC	(MSP,Res.)	**90.43 ±0.3**	52.10 ±3.0	92.56	34.98	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
	SIRC	(DR,$\|\|z\|\|_1$)	90.29 ±0.3	52.54 ±3.4	91.83	37.08	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
	SIRC	(DR,Res.)	90.40 ±0.4	**51.81 ±2.9**	92.83	31.76	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
	SIRC	($-\mathcal{H}$,$\|\|z\|\|_1$)	90.00 ±0.4	54.26 ±2.7	92.24	35.85	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**
	SIRC	($-\mathcal{H}$,Res.)	90.13 ±0.4	54.01 ±3.2	**93.36**	**30.05**	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
		MSP	90.41 ±0.3	52.13 ±2.0	91.00	43.25	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	91.26	40.22	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	91.81	38.24	**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	78.22	58.70	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	58.45	78.97	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
		Max Logit	83.21 ±0.6	65.16 ±3.4	92.33	34.15	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	92.06	35.32	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	85.22	44.41	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	92.34	38.14	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	61.66	78.92	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	ID✗ %AUROC↑	ID✗ %FPR@95↓	Textures %AUROC↑	Textures %FPR@95↓	Colonoscopy %AUROC↑	Colonoscopy %FPR@95↓	Colorectal %AUROC↑	Colorectal %FPR@95↓	Noise %AUROC↑	Noise %FPR@95↓	ImageNet-O %AUROC↑	ImageNet-O %FPR@95↓
ResNet-50 ID %Error: 19.01	SIRC	(MSP,$\|\|z\|\|_1$)	90.34 ±0.2	52.70 ±3.2	93.64 ±0.7	32.02 ±3.3	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
	SIRC	(MSP,Res.)	**90.43 ±0.3**	52.10 ±3.0	96.00 ±0.5	19.81 ±2.1	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
	SIRC	(DR,$\|\|z\|\|_1$)	90.29 ±0.3	52.54 ±3.4	94.01 ±0.7	28.62 ±2.6	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
	SIRC	(DR,Res.)	90.40 ±0.4	**51.81 ±2.9**	96.28 ±0.5	17.29 ±2.0	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
	SIRC	($-\mathcal{H}$,$\|\|z\|\|_1$)	90.00 ±0.4	54.26 ±2.7	94.38 ±0.7	27.38 ±2.7	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
	SIRC	($-\mathcal{H}$,Res.)	90.13 ±0.4	54.01 ±3.2	96.68 ±0.5	15.70 ±2.1	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
		MSP	90.41 ±0.3	52.13 ±2.0	92.88 ±0.8	36.61 ±3.1	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	90.39 ±0.3	51.87 ±1.6	93.16 ±0.8	33.46 ±3.6	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
		$-\mathcal{H}$	90.07 ±0.4	54.05 ±2.9	93.77 ±0.8	30.79 ±3.7	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
		$\|\|z\|\|_1$	48.06 ±1.1	94.70 ±1.4	88.90 ±1.5	39.67 ±2.6	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	47.59 ±1.8	96.45 ±1.1	82.84 ±2.4	46.63 ±3.8	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
		Max Logit	83.21 ±0.6	65.16 ±3.4	95.44 ±0.8	22.04 ±2.8	**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	82.05 ±0.6	69.79 ±3.9	95.37 ±0.8	22.50 ±2.8	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	60.17 ±1.5	87.88 ±2.5	93.00 ±1.1	26.57 ±3.0	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	80.62 ±0.7	78.13 ±2.3	**98.46 ±0.4**	**7.62 ±2.1**	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	49.96 ±2.0	96.36 ±0.9	84.64 ±2.1	46.98 ±3.2	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

Model		Method	ID✗ %AUROC↑	ID✗ %FPR@95↓	OOD mean %AUROC↑	OOD mean %FPR@95↓	Near-IN-200 %AUROC↑	Near-IN-200 %FPR@95↓	Caltech-45 %AUROC↑	Caltech-45 %FPR@95↓	Openimage-O %AUROC↑	Openimage-O %FPR@95↓	iNaturalist %AUROC↑	iNaturalist %FPR@95↓
MobileNetV2 ID %Error: 21.35	SIRC	(MSP,$\|\|z\|\|_1$)	89.53 ±0.3	55.51 ±1.0	92.27	34.82	84.78 ±0.3	61.33 ±1.1	90.46 ±0.3	43.03 ±0.9	91.27 ±0.4	44.05 ±1.5	94.20 ±0.8	31.88 ±3.7
	SIRC	(MSP,Res.)	**89.67 ±0.3**	55.10 ±1.4	91.78	38.56	84.84 ±0.4	61.18 ±0.3	90.25 ±0.4	44.42 ±2.3	91.20 ±0.5	44.83 ±1.8	93.22 ±0.9	37.97 ±3.8
	SIRC	(DR,$\|\|z\|\|_1$)	89.40 ±0.2	56.49 ±2.2	92.66	32.30	84.90 ±0.3	61.26 ±0.8	90.82 ±0.3	40.52 ±2.2	91.61 ±0.4	**42.36 ±1.0**	94.63 ±0.7	**29.15 ±2.8**
	SIRC	(DR,Res.)	89.60 ±0.3	55.69 ±2.0	92.08	36.21	84.98 ±0.3	60.92 ±1.1	90.58 ±0.4	41.98 ±2.3	91.51 ±0.5	43.22 ±1.9	93.40 ±0.9	36.31 ±4.2
	SIRC	($-\mathcal{H}$,$\|\|z\|\|_1$)	88.90 ±0.2	58.64 ±2.1	**92.92**	**32.16**	84.96 ±0.2	62.72 ±0.9	91.35 ±0.3	**39.89 ±2.1**	91.82 ±0.4	43.99 ±1.4	**94.74 ±0.7**	30.47 ±3.9
	SIRC	($-\mathcal{H}$,Res.)	89.12 ±0.3	57.85 ±3.1	92.69	34.20	**85.08 ±0.2**	62.06 ±0.8	91.33 ±0.3	39.93 ±2.2	**91.93 ±0.4**	43.80 ±1.2	94.01 ±0.8	35.24 ±3.7
		MSP	89.64 ±0.3	**55.03 ±1.5**	91.54	39.73	84.84 ±0.3	61.03 ±0.9	90.17 ±0.3	44.77 ±1.7	90.91 ±0.5	46.34 ±1.8	93.57 ±0.9	35.85 ±3.8
		DOCTOR	89.57 ±0.2	55.48 ±2.1	91.86	37.43	84.99 ±0.3	**60.61 ±0.3**	90.52 ±0.3	42.46 ±2.1	91.20 ±0.5	44.96 ±1.0	93.91 ±0.8	33.40 ±3.4
		$-\mathcal{H}$	89.02 ±0.2	58.43 ±2.2	92.37	36.04	85.03 ±0.2	62.53 ±0.4	91.16 ±0.3	41.27 ±2.0	91.54 ±0.4	46.11 ±1.4	94.24 ±0.7	33.85 ±3.8
		$\|\|z\|\|_1$	53.56 ±0.7	93.40 ±0.4	81.06	53.50	56.05 ±0.7	92.65 ±0.5	75.15 ±1.4	73.17 ±2.2	74.05 ±1.4	68.93 ±1.7	86.03 ±1.7	48.35 ±5.0
		Residual	41.99 ±0.8	97.30 ±0.3	41.42	94.11	42.46 ±0.7	97.37 ±0.3	40.90 ±1.2	96.70 ±0.9	44.63 ±1.1	94.39 ±0.6	22.88 ±4.7	99.18 ±0.4
		Max Logit	83.14 ±0.6	63.85 ±1.8	92.08	34.64	81.75 ±0.4	67.36 ±1.3	**91.40 ±0.2**	42.44 ±2.1	89.70 ±0.8	50.66 ±3.0	92.63 ±1.0	39.76 ±4.9
		Energy	81.87 ±0.7	67.98 ±2.0	91.68	36.68	80.87 ±0.4	70.81 ±1.2	90.93 ±0.3	45.77 ±1.9	88.86 ±0.8	54.53 ±3.1	91.76 ±1.0	44.23 ±5.7
		Gradnorm	65.27 ±1.1	85.73 ±1.1	87.25	40.67	66.07 ±0.7	85.13 ±1.0	83.94 ±1.0	56.57 ±2.9	81.94 ±1.2	58.20 ±1.7	90.73 ±1.3	36.80 ±3.4
		ViM	80.21 ±0.4	74.36 ±2.1	89.46	51.97	79.15 ±0.3	75.78 ±1.4	89.17 ±0.4	58.45 ±0.4	87.66 ±1.0	59.54 ±2.1	81.93 ±3.0	81.71 ±6.4
		Mahal	44.44 ±1.0	97.14 ±0.6	43.65	94.20	44.57 ±0.7	97.23 ±0.4	42.82 ±1.1	96.64 ±0.8	48.03 ±1.2	94.11 ±0.8	27.31 ±4.8	99.07 ±0.3

Model		Method	ID✗ %AUROC↑	ID✗ %FPR@95↓	Textures %AUROC↑	Textures %FPR@95↓	Colonoscopy %AUROC↑	Colonoscopy %FPR@95↓	Colorectal %AUROC↑	Colorectal %FPR@95↓	Noise %AUROC↑	Noise %FPR@95↓	ImageNet-O %AUROC↑	ImageNet-O %FPR@95↓
MobileNetV2 ID %Error: 21.35	SIRC	(MSP,$\|\|z\|\|_1$)	89.53 ±0.3	55.51 ±1.0	94.05 ±0.4	28.69 ±0.5	96.64 ±0.8	19.38 ±4.8	96.98 ±1.4	19.39 ±8.8	98.77 ±1.1	7.61 ±7.7	83.30 ±0.3	57.98 ±1.2
	SIRC	(MSP,Res.)	**89.67 ±0.3**	55.10 ±1.4	94.26 ±0.2	28.37 ±1.2	96.25 ±0.9	21.57 ±4.3	95.45 ±1.7	29.21 ±9.2	96.45 ±2.5	24.49 ±24.8	84.09 ±0.4	55.02 ±1.7
	SIRC	(DR,$\|\|z\|\|_1$)	89.40 ±0.2	56.49 ±2.2	94.54 ±0.4	25.70 ±1.1	97.07 ±0.7	15.64 ±3.1	97.61 ±1.3	14.88 ±7.2	99.25 ±0.9	4.38 ±5.6	83.56 ±0.2	56.83 ±1.6
	SIRC	(DR,Res.)	89.60 ±0.3	55.69 ±2.0	94.68 ±0.2	25.37 ±0.7	96.61 ±0.8	17.89 ±3.8	95.77 ±1.6	26.30 ±9.1	96.68 ±2.7	21.41 ±30.2	84.49 ±0.4	52.49 ±2.2
	SIRC	($-\mathcal{H}$,$\|\|z\|\|_1$)	88.90 ±0.2	58.64 ±2.1	94.88 ±0.4	25.26 ±1.1	97.71 ±0.5	11.67 ±3.3	97.82 ±1.1	13.78 ±7.6	99.07 ±1.2	4.28 ±6.6	83.93 ±0.3	57.35 ±1.6
	SIRC	($-\mathcal{H}$,Res.)	89.12 ±0.3	57.85 ±3.1	95.37 ±0.2	22.62 ±0.8	97.57 ±0.6	12.44 ±3.8	96.76 ±1.4	21.12 ±10.8	97.29 ±2.8	17.55 ±30.2	84.91 ±0.3	53.07 ±1.2
		MSP	89.64 ±0.3	**55.03 ±1.5**	92.93 ±0.5	35.28 ±0.8	96.51 ±0.8	20.36 ±4.6	95.63 ±1.6	28.52 ±9.4	96.22 ±2.7	26.33 ±25.6	83.11 ±0.3	59.05 ±1.3
		DOCTOR	89.57 ±0.2	55.48 ±2.1	93.32 ±0.4	32.59 ±1.6	96.99 ±0.8	16.46 ±3.8	96.04 ±1.5	25.23 ±8.9	96.46 ±2.9	23.31 ±31.4	83.34 ±0.3	57.87 ±1.3
		$-\mathcal{H}$	89.02 ±0.2	58.43 ±2.2	94.05 ±0.4	29.72 ±0.9	97.68 ±0.6	12.13 ±3.5	96.82 ±1.4	21.15 ±10.2	97.07 ±2.8	19.33 ±30.7	83.77 ±0.3	58.30 ±1.2
		$\|\|z\|\|_1$	53.56 ±0.7	93.40 ±0.4	92.88 ±0.3	27.55 ±1.5	79.90 ±5.7	80.70 ±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	41.99 ±0.8	97.30 ±0.3	56.86 ±1.4	78.82 ±2.1	27.32 ±6.6	99.68 ±1.1	28.41 ±8.1	99.38 ±1.2	49.61 ±18.3	93.66 ±5.7	59.77 ±1.2	87.83 ±0.7
		Max Logit	83.14 ±0.6	63.85 ±1.8	95.13 ±0.4	25.45 ±1.1	**98.12 ±0.5**	**9.64 ±3.7**	98.27 ±1.0	9.72 ±7.5	98.74 ±0.9	3.42 ±5.1	83.02 ±0.6	63.26 ±1.5
		Energy	81.87 ±0.7	67.98 ±2.0	95.02 ±0.4	26.47 ±2.4	97.99 ±0.5	10.64 ±4.2	98.43 ±1.1	8.65 ±8.3	98.93 ±0.8	2.55 ±4.4	82.36 ±0.7	66.51 ±1.2
		Gradnorm	65.27 ±1.1	85.73 ±1.1	**95.47 ±0.2**	**19.24 ±1.3**	93.35 ±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	70.89 ±2.3
		ViM	80.21 ±0.4	74.36 ±2.1	95.46 ±0.2	26.00 ±1.6	93.62 ±1.5	50.41 ±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	44.44 ±1.0	97.14 ±0.6	58.12 ±1.6	79.53 ±2.3	29.12 ±7.4	99.67 ±1.0	30.35 ±8.1	99.15 ±2.2	51.22 ±15.5	94.62 ±5.4	61.28 ±1.1	87.74 ±1.1

Model		Method	ID✗ %AUROC↑	ID✗ %FPR@95↓	OOD mean %AUROC↑	OOD mean %FPR@95↓	Near-IN-200 %AUROC↑	Near-IN-200 %FPR@95↓	Caltech-45 %AUROC↑	Caltech-45 %FPR@95↓	Openimage-O %AUROC↑	Openimage-O %FPR@95↓	iNaturalist %AUROC↑	iNaturalist %FPR@95↓
DenseNet-121 ID %Error: 17.20	SIRC	(MSP,$\|\|z\|\|_1$)	90.22 ±0.8	52.41 ±2.7	91.68	38.83	85.28 ±0.1	59.35 ±1.0	91.33 ±0.5	40.80 ±2.0	91.88 ±0.4	43.34 ±1.4	93.52 ±0.8	35.88 ±2.8
	SIRC	(MSP,Res.)	90.20 ±0.8	52.42 ±4.2	92.81	32.68	85.18 ±0.1	59.65 ±1.7	91.29 ±0.5	40.51 ±1.6	92.55 ±0.3	39.24 ±1.0	93.42 ±0.7	35.98 ±1.9
	SIRC	(DR,$\|\|z\|\|_1$)	90.21 ±0.8	52.36 ±2.9	91.99	35.30	85.42 ±0.1	57.49 ±1.9	91.63 ±0.5	37.15 ±1.1	92.14 ±0.4	39.89 ±1.2	93.85 ±0.7	31.95 ±2.0
	SIRC	(DR,Res.)	90.18 ±0.8	52.46 ±4.3	93.04	29.68	85.28 ±0.1	57.98 ±2.6	91.53 ±0.5	36.84 ±1.4	92.83 ±0.3	35.57 ±1.1	93.69 ±0.7	32.08 ±1.0
	SIRC	($-\mathcal{H}$,$\|\|z\|\|_1$)	89.95 ±0.9	53.96 ±2.5	92.42	32.92	85.64 ±0.1	**56.98 ±1.5**	**92.18 ±0.4**	**34.18 ±1.7**	92.50 ±0.5	38.43 ±2.0	**94.16 ±0.7**	**30.23 ±2.7**
	SIRC	($-\mathcal{H}$,Res.)	89.92 ±0.8	54.17 ±3.2	**93.45**	**27.97**	85.50 ±0.1	57.90 ±1.8	92.13 ±0.4	34.47 ±1.1	**93.16 ±0.4**	**34.87 ±1.5**	94.07 ±0.6	30.60 ±1.9
		MSP	90.30 ±0.8	51.85 ±2.9	91.44	40.44	85.34 ±0.1	59.20 ±1.0	91.23 ±0.5	41.37 ±2.1	91.78 ±0.4	43.99 ±1.3	93.32 ±0.7	37.11 ±2.8
		DOCTOR	**90.32 ±0.8**	**51.80 ±3.3**	91.71	36.95	85.48 ±0.1	57.16 ±1.8	91.51 ±0.5	37.63 ±1.2	92.03 ±0.4	40.45 ±1.3	93.60 ±0.7	33.19 ±1.9
		$-\mathcal{H}$	90.04 ±0.9	53.41 ±2.3	92.24	34.49	**85.70 ±0.1**	57.01 ±1.5	92.11 ±0.4	34.78 ±1.7	92.43 ±0.5	39.17 ±2.5	94.00 ±0.7	31.37 ±3.1
		$\|\|z\|\|_1$	36.87 ±2.3	98.70 ±0.4	63.53	80.35	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	46.08 ±1.1	95.44 ±0.8	69.38	71.33	44.76 ±0.9	96.14 ±0.1	55.03 ±1.6	91.27 ±0.9	70.77 ±2.2	78.26 ±1.6	64.91 ±6.1	90.93 ±2.5
		Max Logit	83.28 ±1.3	62.32 ±2.7	91.34	36.42	82.50 ±0.3	63.06 ±1.8	91.57 ±0.6	37.89 ±2.2	89.71 ±1.1	48.54 ±2.6	91.28 ±1.2	42.80 ±4.7
		Energy	82.12 ±1.3	66.54 ±3.4	90.92	38.87	81.82 ±0.4	66.32 ±1.6	91.14 ±0.7	40.74 ±2.4	88.88 ±1.2	52.80 ±3.0	90.37 ±1.4	47.67 ±5.2
		Gradnorm	50.18 ±2.6	95.19 ±1.6	76.18	62.58	54.43 ±2.0	93.93 ±1.9	72.05 ±4.0	74.98 ±4.7	67.42 ±3.6	79.78 ±4.3	77.17 ±4.8	67.53 ±8.2
		ViM	76.63 ±1.3	84.73 ±1.9	90.50	44.71	75.68 ±0.8	86.14 ±1.7	87.10 ±1.1	64.27 ±4.3	89.79 ±0.9	51.76 ±4.0	89.00 ±1.5	59.37 ±5.9
		Mahal	56.97 ±1.1	94.57 ±1.1	73.14	71.91	53.97 ±1.1	95.39 ±0.3	60.81 ±1.5	90.95 ±1.0	76.58 ±1.6	76.51 ±2.6	69.67 ±4.7	89.79 ±4.0

Model		Method	ID✗ %AUROC↑	ID✗ %FPR@95↓	Textures %AUROC↑	Textures %FPR@95↓	Colonoscopy %AUROC↑	Colonoscopy %FPR@95↓	Colorectal %AUROC↑	Colorectal %FPR@95↓	Noise %AUROC↑	Noise %FPR@95↓	ImageNet-O %AUROC↑	ImageNet-O %FPR@95↓
DenseNet-121 ID %Error: 17.20	SIRC	(MSP,$\|\|z\|\|_1$)	90.22 ±0.8	52.41 ±2.7	93.68 ±0.2	32.48 ±0.9	95.38 ±0.9	27.10 ±5.5	96.44 ±2.3	23.54 ±14.1	94.31 ±8.3	28.79 ±25.0	83.32 ±0.7	58.19 ±1.4
	SIRC	(MSP,Res.)	90.20 ±0.8	52.42 ±4.2	96.55 ±0.1	16.60 ±0.9	95.14 ±0.8	27.71 ±4.6	96.89 ±1.7	19.72 ±10.6	99.70 ±1.1	2.01 ±8.6	84.59 ±0.6	52.69 ±2.1
	SIRC	(DR,$\|\|z\|\|_1$)	90.21 ±0.8	52.36 ±2.9	94.02 ±0.2	28.81 ±0.3	95.73 ±0.9	22.94 ±5.4	96.84 ±2.2	19.40 ±13.2	94.72 ±8.3	24.54 ±24.1	83.52 ±0.7	55.50 ±1.7
	SIRC	(DR,Res.)	90.18 ±0.8	52.46 ±4.3	96.77 ±0.1	14.26 ±1.2	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	49.91 ±3.2
	SIRC	($-\mathcal{H}$,$\|\|z\|\|_1$)	89.95 ±0.9	53.96 ±2.5	94.48 ±0.2	26.16 ±1.3	96.43 ±0.9	17.89 ±5.2	97.38 ±1.9	15.64 ±12.0	95.17 ±8.0	22.30 ±19.5	83.89 ±0.7	54.48 ±1.9
	SIRC	($-\mathcal{H}$,Res.)	89.92 ±0.8	54.17 ±3.2	97.07 ±0.1	13.02 ±1.0	96.21 ±0.8	18.57 ±4.3	97.78 ±1.4	12.58 ±8.4	**99.91 ±0.3**	0.32 ±1.4	85.22 ±0.6	**49.44 ±1.4**
		MSP	90.30 ±0.8	51.85 ±2.9	93.25 ±0.2	35.01 ±0.8	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	58.37 ±1.3
		DOCTOR	**90.32 ±0.8**	**51.80 ±3.3**	93.54 ±0.2	31.47 ±0.5	95.78 ±0.9	22.56 ±4.8	96.25 ±2.1	23.09 ±12.5	93.72 ±8.3	31.28 ±29.0	83.49 ±0.6	55.73 ±1.9
		$-\mathcal{H}$	90.04 ±0.9	53.41 ±2.3	94.14 ±0.2	28.14 ±1.7	96.47 ±0.8	17.71 ±5.2	96.99 ±1.9	18.30 ±12.3	94.43 ±7.8	29.18 ±22.9	83.89 ±0.7	54.76 ±2.3
		$\|\|z\|\|_1$	36.87 ±2.3	98.70 ±0.4	75.96 ±4.4	69.61 ±9.0	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	46.08 ±1.1	95.44 ±0.8	92.00 ±0.4	29.64 ±0.9	49.87 ±4.5	99.46 ±1.0	76.95 ±1.7	72.88 ±7.1	98.07 ±2.1	9.12 ±6.3	72.05 ±0.9	74.29 ±1.4
		Max Logit	83.28 ±1.3	62.32 ±2.7	94.08 ±0.7	27.50 ±2.1	**96.96 ±0.8**	**15.92 ±4.8**	98.20 ±1.4	10.35 ±10.1	96.09 ±6.8	20.58 ±18.2	81.65 ±1.9	61.15 ±1.9
		Energy	82.12 ±1.3	66.54 ±3.4	93.82 ±0.8	29.16 ±2.8	96.80 ±0.9	16.88 ±5.1	**98.26 ±1.5**	**9.99 ±10.5**	96.13 ±6.7	21.78 ±20.4	81.07 ±2.0	64.51 ±2.2
		Gradnorm	50.18 ±2.6	95.19 ±1.6	86.17 ±2.9	46.07 ±7.8	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	76.63 ±1.3	84.73 ±1.9	**98.07 ±0.2**	**9.82 ±1.4**	92.00 ±0.8	63.03 ±7.1	97.70 ±0.5	11.13 ±3.7	99.85 ±0.1	**0.04 ±0.1**	**85.30 ±1.4**	56.82 ±3.3
		Mahal	56.97 ±1.1	94.57 ±1.1	91.55 ±0.4	31.71 ±0.8	60.34 ±9.6	98.88 ±1.7	70.77 ±4.0	79.75 ±7.1	96.66 ±2.9	12.67 ±5.6	77.93 ±1.2	71.51 ±2.4

Table 2. %AUROC and %FPR@95 results for single pre-trained ImageNet-1k models.

Model	Method	ID✗		OOD mean		Openimage-O		iNaturalist		Textures		Colonoscopy		Colorectal		Noise		ImageNet-O	
		AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	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	90.08	37.09	90.25	47.09	94.37	29.13	88.74	43.84	97.10	16.69	93.94	30.28	99.26	4.32	66.93	88.30
	SIRC (MSP,Res.)	86.31	62.35	92.89	25.23	92.69	34.29	94.61	27.17	96.80	10.04	97.18	14.69	96.88	14.06	99.99	**0.00**	72.07	76.35
	SIRC (DR,$\|\|z\|\|_1$)	85.36	66.04	90.44	35.13	90.35	47.22	94.75	26.89	89.19	41.72	97.13	15.22	94.85	23.82	99.48	2.93	67.33	88.10
	SIRC (DR,Res.)	85.55	64.66	93.34	**22.76**	**93.25**	**31.27**	**94.99**	**24.30**	97.15	8.06	97.36	12.46	97.52	10.10	99.99	0.00	73.12	73.10
	SIRC ($-\mathcal{H}$,$\|\|z\|\|_1$)	83.43	69.62	90.98	36.74	90.46	51.57	94.53	33.11	89.06	47.03	98.37	9.66	95.71	23.18	99.48	3.80	69.25	88.80
	SIRC ($-\mathcal{H}$,Res.)	83.50	68.91	93.67	25.61	92.78	40.20	94.63	33.51	97.05	10.19	98.33	9.02	98.12	8.62	100.00	0.00	74.76	77.70
	MSP	**86.35**	**61.93**	89.16	41.81	90.13	47.48	93.70	32.96	87.04	51.90	97.47	14.92	91.55	44.28	97.37	12.53	66.87	88.60
	DOCTOR	85.67	64.49	89.57	40.48	90.33	47.64	93.95	32.04	87.27	52.64	97.89	12.61	92.34	39.78	97.94	9.74	67.25	88.90
	$-\mathcal{H}$	83.49	69.09	90.25	41.32	90.23	54.09	93.80	38.97	87.47	54.92	98.52	8.69	94.02	34.70	98.46	7.73	69.24	90.10
	$\|\|z\|\|_1$	47.74	95.45	70.81	66.69	53.48	87.82	73.95	78.05	73.89	66.14	58.42	89.18	86.16	51.18	99.61	1.36	50.14	93.10
	Residual	50.18	94.86	85.59	50.02	80.17	68.38	76.76	80.65	97.67	10.99	67.60	98.55	95.43	25.34	99.95	0.00	81.57	66.20
	Max Logit	77.25	71.07	90.24	41.72	88.11	59.64	91.87	48.90	87.08	55.70	99.04	4.64	96.25	25.98	98.79	6.47	70.57	90.70
	Energy	74.68	77.15	89.41	45.23	85.86	68.88	89.27	59.78	85.85	61.61	**99.19**	**3.17**	96.56	23.82	98.83	6.83	70.33	92.55
	Gradnorm	64.64	88.00	84.85	46.15	73.53	76.05	87.99	53.65	85.04	50.85	94.56	29.39	95.94	22.56	99.82	0.67	57.05	89.85
	ViM	70.30	86.87	**94.95**	25.61	92.08	41.79	91.68	47.40	**99.17**	**3.39**	95.59	29.88	**99.30**	**1.26**	**100.00**	0.00	**86.80**	**55.55**
	Mahal	56.82	93.95	89.62	46.82	86.43	61.39	85.09	73.14	98.19	9.19	77.36	98.76	96.09	22.40	99.88	0.00	84.28	62.85

Model	Method	ID✗		OOD mean		Openimage-O		iNaturalist		Textures		Colonoscopy		Colorectal		Noise		ImageNet-O	
		AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓
DenseNet-121 ID %Error: 25.38	SIRC (MSP,$\|\|z\|\|_1$)	85.99	63.14	89.52	33.01	90.93	39.50	95.36	21.61	89.65	37.34	96.79	17.15	96.06	20.94	99.74	1.10	58.10	93.45
	SIRC (MSP,Res.)	85.97	63.33	90.05	31.62	91.17	38.58	94.08	27.50	93.38	22.93	96.18	19.71	95.51	23.60	99.67	0.60	60.35	88.45
	SIRC (DR,$\|\|z\|\|_1$)	85.77	64.51	90.00	30.09	91.55	36.00	95.98	17.81	90.32	33.12	97.10	14.22	96.88	15.68	99.79	0.83	58.36	93.00
	SIRC (DR,Res.)	85.72	65.09	90.43	28.80	91.72	35.28	94.50	24.33	93.85	19.42	96.30	16.79	96.21	17.90	99.62	0.54	60.83	87.35
	SIRC ($-\mathcal{H}$,$\|\|z\|\|_1$)	84.90	67.31	90.83	28.44	92.41	34.47	**96.52**	**16.28**	91.05	32.85	97.89	9.86	97.79	12.36	99.83	0.68	60.31	92.60
	SIRC ($-\mathcal{H}$,Res.)	84.85	67.87	91.46	**26.46**	**92.64**	**34.09**	95.67	20.33	94.42	19.07	97.45	11.37	97.56	12.30	99.79	0.39	62.71	87.70
	MSP	**86.11**	**62.67**	88.81	36.77	90.26	43.08	94.26	27.56	88.31	43.72	96.90	17.10	94.44	30.72	99.55	1.69	57.97	93.55
	DOCTOR	85.93	63.43	89.28	34.17	90.82	39.93	94.83	23.95	88.85	41.01	97.33	13.52	95.24	25.94	99.64	1.37	58.23	93.45
	$-\mathcal{H}$	84.97	66.76	90.39	30.68	91.91	37.18	95.83	19.56	90.08	37.56	97.95	9.42	97.00	17.20	99.76	1.15	60.17	92.70
	$\|\|z\|\|_1$	47.53	94.93	78.50	53.82	69.94	70.15	89.06	39.14	84.61	49.73	58.85	89.84	92.89	34.90	99.88	0.49	54.29	92.50
	Residual	51.52	94.26	71.96	66.47	69.78	78.27	61.14	93.61	90.21	33.64	37.94	99.37	75.49	70.74	97.32	14.40	71.83	75.25
	Max Logit	77.97	71.35	**91.62**	28.87	92.19	38.48	96.07	20.57	91.59	34.32	**98.20**	**8.77**	98.62	6.48	99.89	0.49	64.77	92.95
	Energy	76.13	75.77	91.47	30.02	91.54	42.66	95.60	23.50	91.39	35.43	97.87	11.18	**98.86**	**5.02**	99.91	0.39	65.12	91.95
	Gradnorm	55.44	92.10	85.31	42.04	78.97	58.55	93.87	25.24	89.62	37.81	81.08	68.36	97.63	13.10	**99.96**	0.09	56.04	91.15
	ViM	70.16	88.53	89.58	47.81	88.40	56.49	88.74	66.34	**96.64**	**17.69**	82.83	89.17	95.19	31.76	99.57	**0.01**	75.66	73.20
	Mahal	57.28	94.10	68.90	81.55	69.02	86.67	49.94	97.79	82.79	55.43	66.51	96.97	58.34	96.34	75.13	68.35	**80.53**	**69.30**

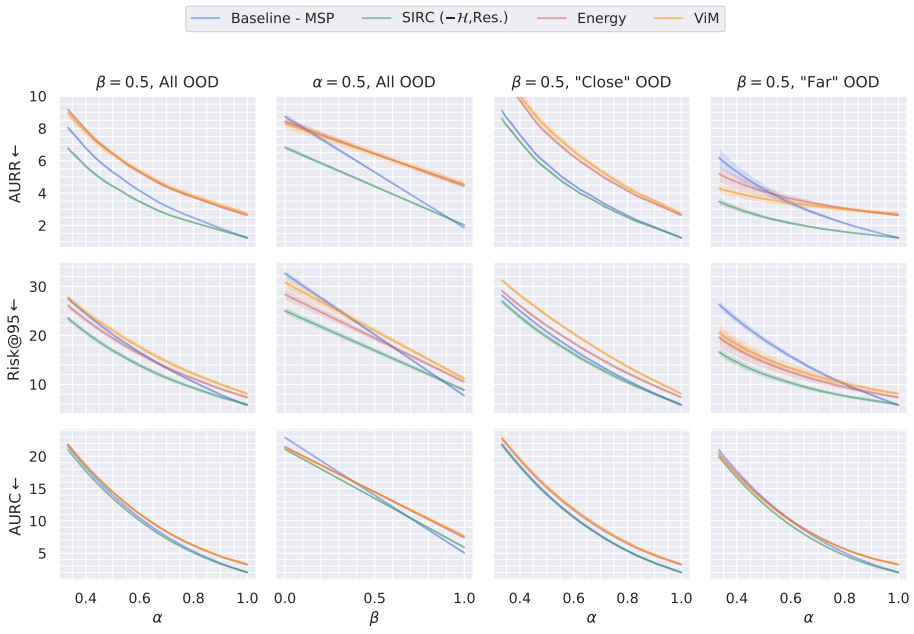

Fig. 3. Varying α and β for ResNet-50 (ImageNet-200) (values $\times 10^2$).

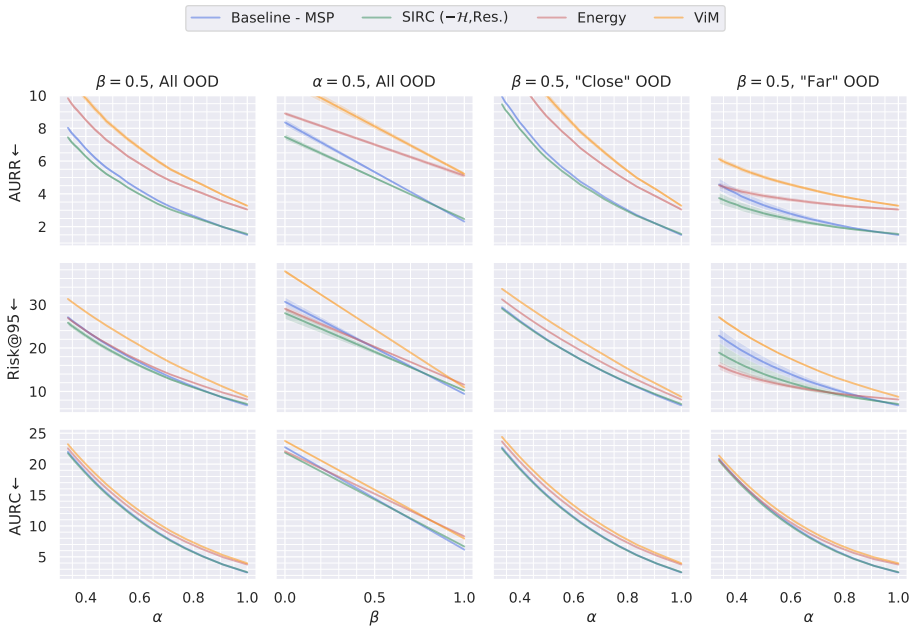

Fig. 4. Varying α and β for MobileNetV2 (ImageNet-200) (values $\times 10^2$).

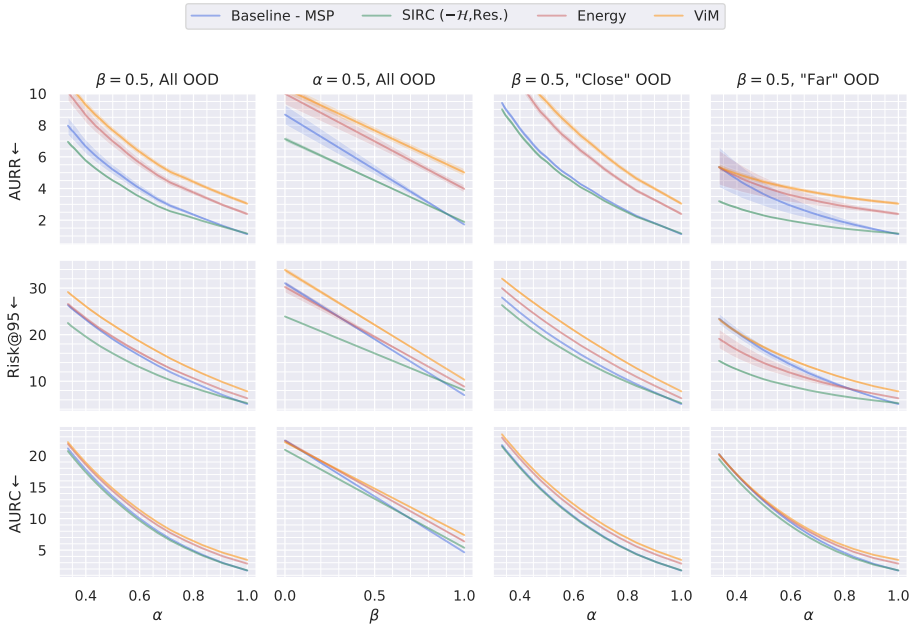

Fig. 5. Varying α and β for DenseNet-121 (ImageNet-200) (values $\times 10^2$).

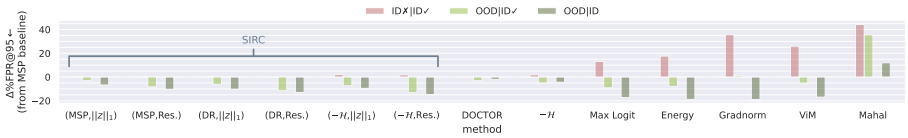

Fig. 6. ResNet-50 (ImageNet-200), comparing the change in %FPR@95 relative to the MSP baseline for different detection methods and data groups.

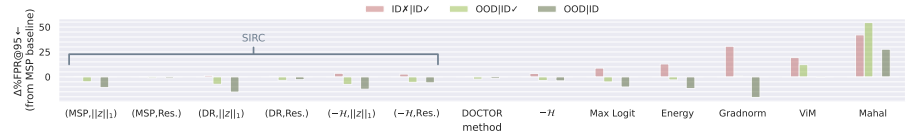

Fig. 7. MobileNetV2 (ImageNet-200), comparing the change in %FPR@95 relative to the MSP baseline for different detection methods and data groups.

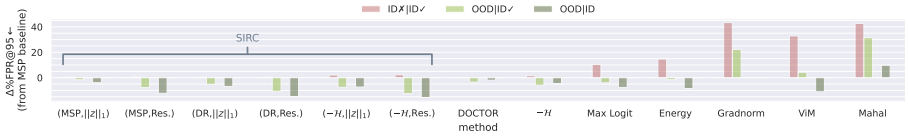

Fig. 8. DenseNet-121 (ImageNet-200), comparing the change in %FPR@95 relative to the MSP baseline for different detection methods and data groups.

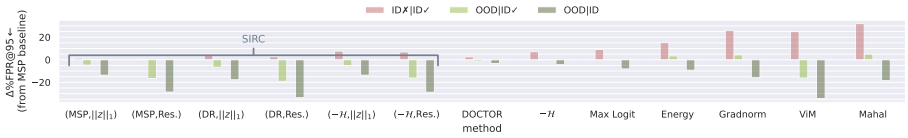

Fig. 9. ResNetV2-101 (ImageNet-1k), comparing the change in %FPR@95 relative to the MSP baseline for different detection methods and data groups.

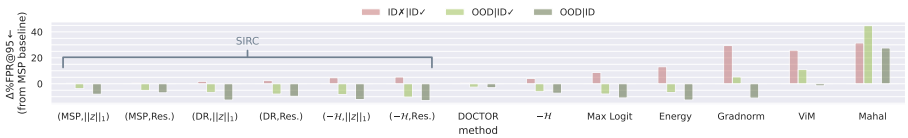

Fig. 10. DenseNet-121 (ImageNet-1k), comparing the change in %FPR@95 relative to the MSP baseline for different detection methods and data groups.

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 the main paper).

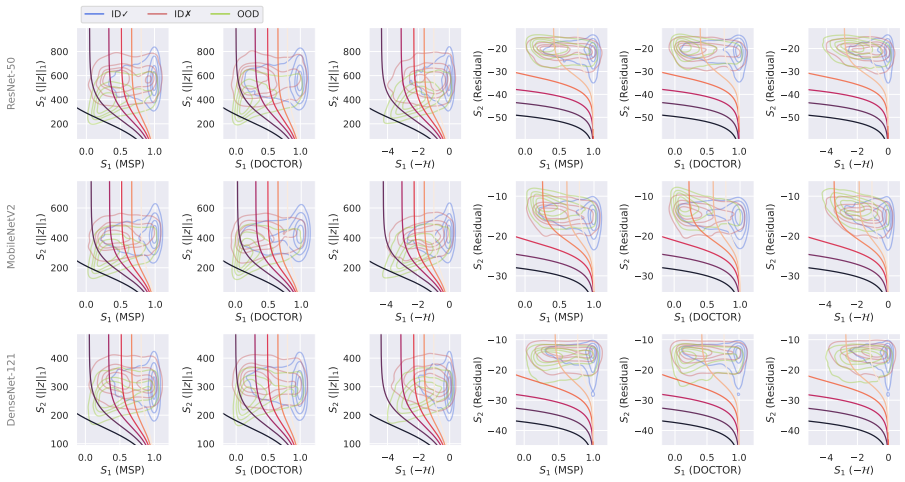

Fig. 11. SIRC combinations on the S_1, S_2-plane, ID: ImageNet-200, OOD: iNaturalist.

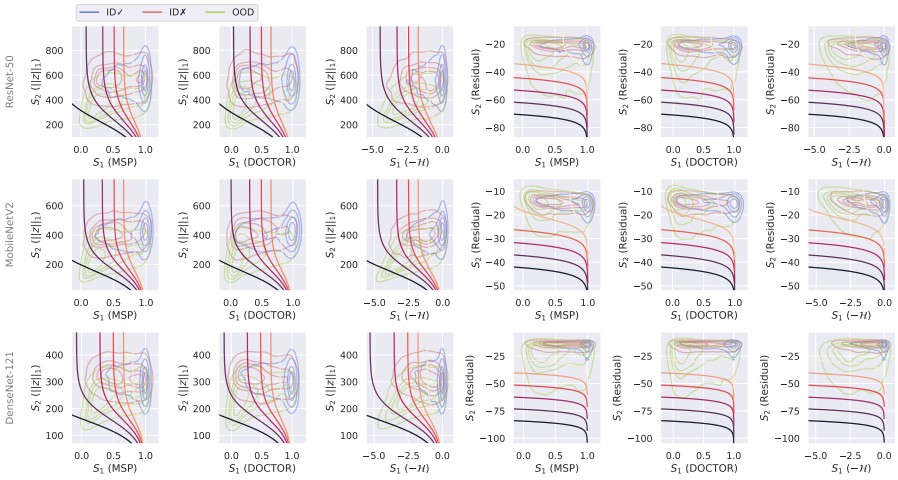

Fig. 12. SIRC combinations on the S_1, S_2-plane, ID: ImageNet-200, OOD: Textures.

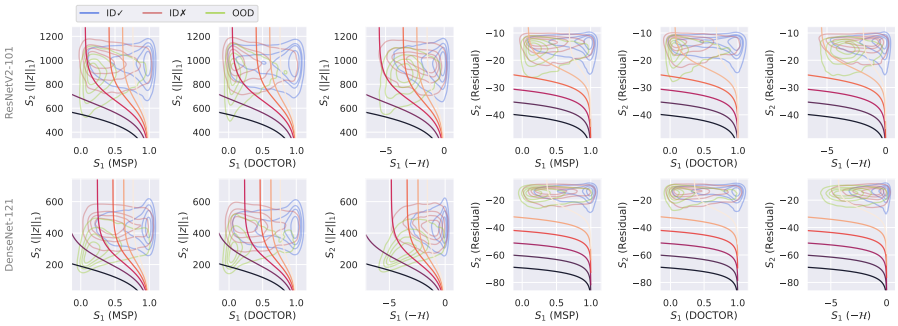

Fig. 13. SIRC combinations on the S_1, S_2-plane, ID ImageNet-1k, OOD: iNaturalist.

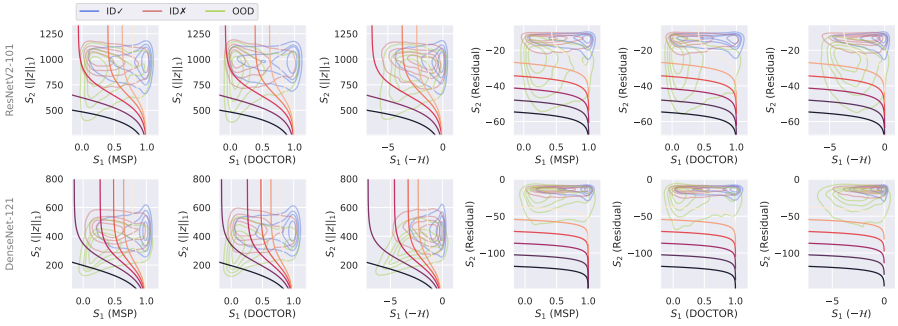

Fig. 14. SIRC combinations on the S_1, S_2-plane, ID ImageNet-1k, OOD: Textures.

References

[1] Fort, S., Ren, J., Lakshminarayanan, B.: Exploring the limits of out-of-distribution detection. In: NeurIPS (2021)

[2] Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: NIPS (2017)

[3] Granese, F., Romanelli, M., Gorla, D., Palamidessi, C., Piantanida, P.: Doctor: A simple method for detecting misclassification errors. In: NeurIPS (2021)

[4] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2016)

[5] Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., Song, D.X.: Scaling out-of-distribution detection for real-world settings. arXiv: Computer Vision and Pattern Recognition (2020)

[6] Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. ArXiv abs/1610.02136 (2017)

[7] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.X.: Natural adversarial examples. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 15257–15266 (2021)

[8] Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2261–2269 (2017)

[9] Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. In: NeurIPS (2021)

[10] Huang, R., Li, Y.: Mos: Towards scaling out-of-distribution detection for large semantic space. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 8706–8715 (2021)

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

[12] Kim, J., Koo, J., Hwang, S.: A unified benchmark for the unknown detection capability of deep neural networks. ArXiv abs/2112.00337 (2021)

[13] Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (bit): General visual representation learning. In: ECCV (2020)

[14] Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: NeurIPS (2018)

[15] Liu, W., Wang, X., Owens, J.D., Li, Y.: Energy-based out-of-distribution detection. ArXiv abs/2010.03759 (2020)

[16] Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O.Y., Béorchia, S., Poincloux, L., Bartoli, A.: Computer-aided classification of gastrointestinal lesions in regular colonoscopy. IEEE transactions on medical imaging (2016)

[17] Mukhoti, J., Kirsch, A., van Amersfoort, J.R., Torr, P.H.S., Gal, Y.: Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. ArXiv abs/2102.11582 (2021)
[18] Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 4510–4520 (2018)
[19] Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. ArXiv abs/2203.10807 (2022)