
CHALMERS
UNIVERSITY OF TECHNOLOGY

# GEN: Pushing the Limits of Softmax-Based Out-of-Distribution Detection



Citation for the original published paper (version of record):
Liu, X., Lochman, Y., Zach, C. (2023). GEN: Pushing the Limits of Softmax-Based Out-of-Distribution Detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2023-June: 23946-23955.
http://dx.doi.org/10.1109/CVPR52729.2023.02293



(article starts on next page)

# GEN: Pushing the Limits of Softmax-Based Out-of-Distribution Detection

Xixi Liu
xixil@chalmers.se

Yaroslava Lochman
lochman@chalmers.se

Christopher Zach
zach@chalmers.se

Chalmers University of Technology

## Abstract

*Out-of-distribution (OOD) detection has been extensively studied in order to successfully deploy neural networks, in particular, for safety-critical applications. Moreover, performing OOD detection on large-scale datasets is closer to reality, but is also more challenging. Several approaches need to either access the training data for score design or expose models to outliers during training. Some post-hoc methods are able to avoid the aforementioned constraints, but are less competitive. In this work, we propose Generalized ENtropy score (GEN), a simple but effective entropy-based score function, which can be applied to any pre-trained softmax-based classifier. Its performance is demonstrated on the large-scale ImageNet-1k OOD detection benchmark. It consistently improves the average AUROC across six commonly-used CNN-based and visual transformer classifiers over a number of state-of-the-art post-hoc methods. The average AUROC improvement is at least 3.5%. Furthermore, we used GEN on top of feature-based enhancing methods as well as methods using training statistics to further improve the OOD detection performance. The code is available at:* https://github.com/XixiLiu95/GEN.

## 1. Introduction

In order to make the usage of deep learning methods in real-word applications safer, it is crucial to distinguish whether an input at test time is a valid in-distribution (ID) sample or a previously unseen out-of-distribution (OOD) sample. Thus, a trained deep neural network (DNN) should ideally know what it does not know [30]. This ability is particularly important for high-stake applications in autonomous driving [8] and medical image analysis [35]. However, it is common for neural networks to make overconfident predictions even for OOD samples. A recent survey on OOD detection [45] identifies several scenarios requiring the detection of OOD samples, with covariate shift (change in the input distribution) and semantic shift (change in the label distribution) being two important settings.

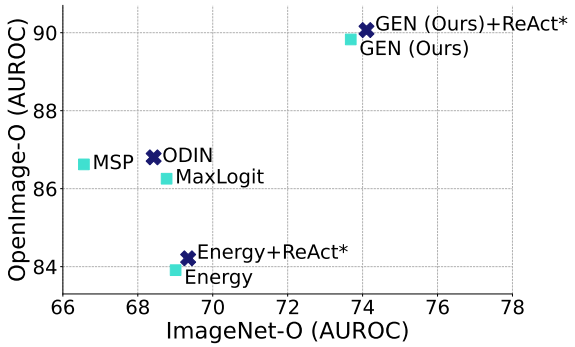


Figure 1. *Performance of Post-hoc OOD Detection Methods Applied to 6 Classifiers Trained on ImageNet-1K.* Reported are AUROC values (%) averaged over the models. Methods marked with light squares use information from logits / probabilities. Methods marked with dark crosses also use information from features. ReAct* corresponds to performing extra feature clipping before computing the score.

In this work, we focus on the semantic shift scenario, meaning that we aim to detect inputs with semantic labels not present in the training set. When solving the OOD detection problem, the idea is to design a scalar score function of a data sample as an argument that assigns higher values to true ID samples. The semantic shift scenario also allows us to mainly focus on the predictive distribution as provided by a DNN classifier to design such score function.

A number of existing works for OOD detection rely on the predictive distribution [14, 28], but often a better OOD detection performance can be achieved when also incorporating feature statistics for ID data [13, 24, 37, 38, 43]. These high-performing methods have practical constraints that can be challenging to eliminate: some methods require access to at least a portion of training data [13, 24, 37, 43] while others need access to internal feature activations [38]. However, commercially deployed DNNs are often black-

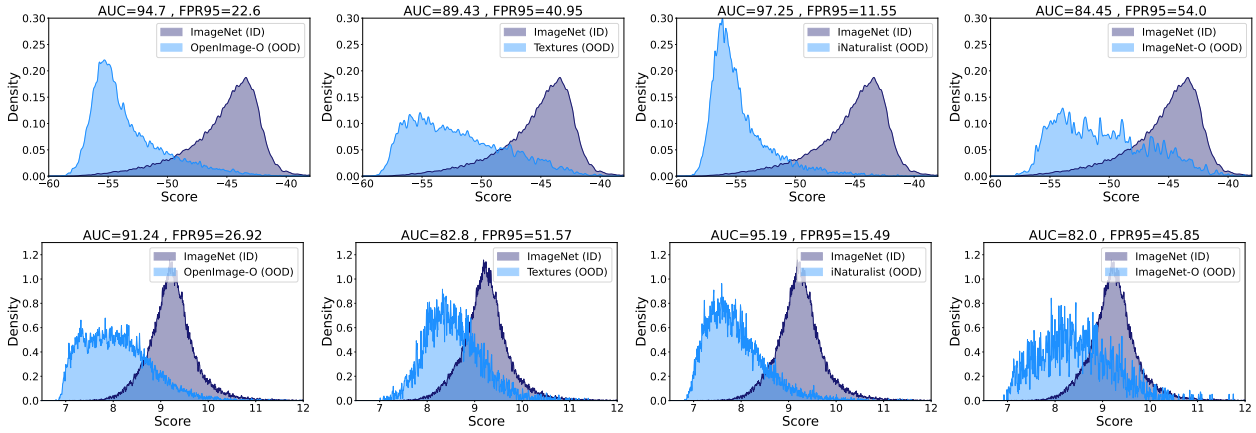


Figure 2. *Score Distributions*. The top row is GEN, and the bottom one is Energy [28]. The distributions are shown for the ID ImageNet-1K dataset (dark blue) and four OOD datasets (light blue). The classification model used here is Swin [29].

box classifiers, and the training data is likely to be confidential. Hence, the goal of this work is to explore and push the limits of OOD detection when the output of a softmax layer is the only available source of information. Our method therefore falls under the *post-hoc* category of OOD detection frameworks, where only a trained DNN is used without the need for training data. Fig. 1 highlights its performance compared to other methods in this category.

**Contribution** We propose GEN, a simple but effective entropy-based method for OOD detection. (i) GEN uses predictive distribution only. It does not require re-training and/or outlier exposure, it does not use any training data statistics. (ii) Yet it performs very well (see Figs. 1 and 2), meaning that it can potentially be used in more constrained model deployment scenarios. Compared to other post-hoc methods, score distributions produced by GEN lead to a better ID/OOD separation. We show that our method consistently achieves better results in terms of AUROC (and usually in terms of FPR95) compared to other post-hoc methods. In particular, GEN on average outperforms other post-hoc methods on the largest and carefully constructed OOD dataset OpenImage-O as well as on the very challenging ImageNet-O dataset based on natural adversarial examples.

## 2. Related Work

**Score design** Given a pre-trained softmax neural classifier, designing a proper score function that aims to separate ID from OOD data is essential to successfully perform OOD detection. [14] proposes the maximum predicted softmax probability (MSP) and thereby establishes an initial baseline for such scores. Subsequently, [24] defines the score as the minimum Mahalanobis distance between features and the empirical class-wise centroids, which are computed from training samples. The energy score is suggested in [28] and is computed via $\mathrm{LogSumExp}$, which is the soft maximum of the logits. This energy score can also be understood as the unnormalized log data density [10]. Unlike [28], [13] proposes the (hard) maximum of the logits as OOD score. [13] also gives a statistics-based alternative called KL Matching, where the posterior distribution template $\bar{p}_k$ for each class $k$ is computed from training data. At test time, the KL divergence between the predictive distribution and each posterior distribution template is calculated for a given sample. The negative minimum KL divergence is taken as the score.

Previous methods use information from one of the spaces, feature, logit, or predictive distribution. GradNorm [18] incorporates the information from both feature and predictive distribution. Specifically, this score is a product of the feature norm and the distance from the predictive to the uniform distribution. Predictive Normalized Maximum Likelihood (pNML) [1] derives a score function based on the generalization error (the regret), which needs to access the empirical correlation matrix of training features and the predictive distribution. ViM [43] uses information from all spaces via introducing a virtual logit with corresponding rescaling factor $\alpha$. First, the residual of the feature $\mathbf{z}$ is calculated as $\|\mathbf{z}^{P^\perp}\|$, where $P$ is the so-called principal space (i.e. the principal component of the features). A mixing coefficient $\alpha$ is computed in order to match the scale of the virtual logits to the real maximum logits over the training set. The final score is calculated as the softmax probability of the virtual logit and can be also interpreted as a combination of the energy score $\mathrm{LogSumExp}\, f(\mathbf{z})$ and rescaled residual $-\alpha\|\mathbf{z}^{P^\perp}\|$. Our GEN score can actually replace the energy in this formulation to further improve the OOD detection performance.

| Method | Equation | Free of | | Space | | |
|---|---|---|---|---|---|---|
| | | ID train data | ID labels | features | logits | probs |
| MSP [14] | $\max_c p_c$ | ✓ | ✓ | | | ✓ |
| MaxLogit [13] / Energy [28] | $\max_c f_c(\mathbf{z})$ / LogSumExp $f(\mathbf{z})$ | ✓ | ✓ | | ✓ | |
| GradNorm [18] | $\|\mathbf{p} - \mathbf{1}/C\|_1 \cdot \|\mathbf{z}\|_1$ | ✓ | ✓ | ✓ | | ✓ |
| ODIN [25] | $\tilde{\mathbf{x}} = \mathbf{x} + \varepsilon \text{sign}\left(\nabla_{\mathbf{x}} \log \max_c p_c(\mathbf{x})\right)$ | ✓ | ✓ | | | ✓ |
| ReAct [38] | $\tilde{\mathbf{z}} = \min(\mathbf{z}, b)$ | ✗, $b$ | ✓ | ✓ | | |
| RankFeat [37] | $\tilde{\mathbf{o}} = \mathbf{o} - s_1 \mathbf{u}_1 \mathbf{v}_1^\top$ | ✓ | ✓ | ✓ | | |
| Mahalanobis [24] | $\max_c -(\mathbf{z} - \hat{\mu}_c)^\top \hat{\Sigma}^{-1} (\mathbf{z} - \hat{\mu}_c)$ | ✗, $\hat{\Sigma}, \hat{\mu}_c$ | ✗ | ✓ | | |
| pNML [1] | $\log \sum_{c=1}^{C} \frac{p_c}{p_c + p_c^\kappa (1-p_c)}, \quad \kappa = \frac{\mathbf{z}^\top \Sigma_{\text{corr}} \mathbf{z}}{1 + \mathbf{z}^\top \Sigma_{\text{corr}} \mathbf{z}}$ | ✗, $\Sigma_{\text{corr}}$ | ✓ | | | |
| KL Matching [13] | $-\min_c D_{\text{KL}}(\mathbf{p} \parallel \mathbf{d}_c)$ | ✗, $\mathbf{d}_c$ | ✗ | | | ✓ |
| Residual [43] | $-\|\mathbf{z}^{P^\perp}\|_2$ | ✗, $P$ | ✓ | ✓ | | |
| ViM [43] | $-\alpha \|\mathbf{z}^{P^\perp}\|_2 + \text{LogSumExp}\, f(\mathbf{z})$ | ✗, $\alpha$, $P$ | ✓ | ✓ | ✓ | |
| Shannon Entropy | $-\sum_{m=1}^{M} p_{i_m} \log p_{i_m}, \quad p_{i_1} \geq \cdots \geq p_{i_C}, \ \gamma \in (0,1)$ | ✓ | ✓ | | | ✓ |
| **GEN (Ours)** | $G_\gamma(\mathbf{p}) = -\sum_{m=1}^{M} p_{i_m}^\gamma (1 - p_{i_m})^\gamma, \quad p_{i_1} \geq \cdots \geq p_{i_C}, \ \gamma \in (0,1)$ | ✓ | ✓ | | | ✓ |
| **GEN (Ours)** + ReAct [38] | $G_\gamma\left(\text{Softmax}(f(\tilde{\mathbf{z}})\right), \quad \tilde{\mathbf{z}} = \min(\mathbf{z}, b)$ | ✗, $b$ | ✓ | ✓ | | ✓ |
| **GEN (Ours)** + Residual [43] | $G_\gamma(\mathbf{p}) \cdot \|\mathbf{z}^{P^\perp}\|_2$ | ✗, $P$ | ✓ | ✓ | | ✓ |

Table 1. *Technical Comparison of OOD Detection Methods.* $\mathbf{x}$ is an input, $\mathbf{z}$ is an output of the penultimate layer (also called features), $f(\mathbf{z})$ denotes logits, $\mathbf{p} = \text{Softmax}(f(\mathbf{z}))$ is predictive distribution, and $C$ is the number of classes. Enhancing methods work in the input / feature space, *i.e.* they perturb original inputs $\mathbf{x}$, features $\mathbf{z}$, or intermediate convolutional features $\mathbf{o}$ (where the perturbed result of e.g. $\mathbf{x}$ is $\tilde{\mathbf{x}}$). Several methods require pre-computation of training data statistics. In particular, Mahalanobis [24] needs the empirical per-class mean $\hat{\mu}_c$ and tied covariance $\hat{\Sigma}$ of the training features. pNML [1] needs the empirical correlation matrix $\Sigma_{\text{corr}}$. KL Matching [13] requires the knowledge of per-class predictive distributions $\mathbf{d}_c$. Residual and ViM [43] require the principal space $P$ of the training features. Our method GEN uses information from the probability space only, does not perturb the inputs nor does it need ID data.

**Score enhancing methods** There is also a line of research that aims to enhance the OOD detection performance for given score functions [17, 25, 37, 38]. ODIN [25] uses a temperature scaling $T$ for logits and adds perturbation to the input sample to enhance the reliability of OOD detection when MSP score is used. Specifically, each logit is divided by a temperature $T$, and the perturbed input can be calculated as $\tilde{\mathbf{x}} = \mathbf{x} + \varepsilon \text{sign}(\nabla_{\mathbf{x}} \log S_{\hat{y}}(\mathbf{x}; T))$, where $S_{\hat{y}}(\mathbf{x}; T)$ is the maximum softmax probability. However, $T$ needs to be tuned with OOD samples. Generalized ODIN [17] aims to free ODIN [25] from the need of OOD samples without decreasing the OOD performance. ReAct [38] applies feature clipping on the penultimate layer of neural networks. Specifically, an operation $\min(f(\mathbf{z}), c)$ is applied element-wise to the feature vector $f(\mathbf{z})$. This enhancing method is compatible with MSP score [14] and energy score [28]. RankFeat [37] looked into the distribution of the singular values for ID and OOD samples and found that OOD samples appear to have larger dominant singular values than ID samples. Instead of using the largest singular value as the score, they remove the rank-1 matrix $s_1 \mathbf{u}_1 \mathbf{v}_1^\top$ composed of the largest singular value $s_1$ and the associated singular vectors $\mathbf{u}_1$, $\mathbf{v}_1$ from the intermediate (flattened) feature maps $\mathbf{o}$. The modified features $\tilde{\mathbf{o}}$ are processed by the remaining part of the neural network, and the energy score is computed. A summary of the aforementioned score design and enhancing methods is given in Table 1.

**Modifying the training loss** An alternative to the OOD detection score design for fixed networks is to incorporate the OOD samples into the training procedure. Specifically, adding a separate network head (and a suitable loss) for confidence prediction [5], reinterpreting logits as joint log-probabilities (over inputs and labels) and training using a log-evidence term in addition to the standard cross-entropy loss [10], or incorporating a subspace prior on features [47] are approaches to obtain DNNs better suited for OOD detection (besides solving a classification task). [19] addresses the fine-grained classification setting in particular and leverages semantic groups (and a dedicated out-of-group label), which simplifies decision boundaries and therefore helps to identify OOD samples. It is further possible to explicitly include OOD data into the training phase of a DNN. Joint minimization of a classification loss (over ID data) and a regularization term favoring highly uncertain predictive distributions for OOD data is suggested in [15, 31].

**Classifier calibration** Supervised training usually leads to uncalibrated classifiers, which tend to be either over-confident (usually) or under-confident (rarely) in their prediction confidence. In short, "a predicted probability (vector) should match empirical (observed) accuracy" [36]. The calibration of a pre-trained classifier can be improved by post-processing the logits [11, 33, 46] or by using classifier ensembles [23]. Since a number of OOD detection ap-

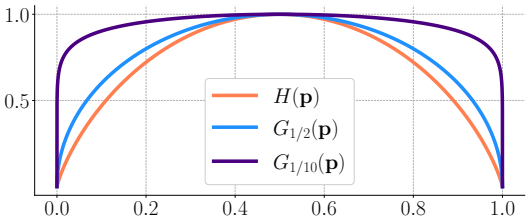


Figure 3. Generalized entropies: $G_{1/2}$, $G_{1/10}$ and the Shannon entropy $H(p)$ for a Bernoulli random variable (all scaled to the same range).

proaches uses solely the logits or resulting predictive distribution as input, the OOD detection performance may vary between the trained vanilla and the calibrated classifiers. At least for monotone transformations of logits [33,46] the performance of MSP [14], MaxLogit [13], and Energy [28] scores should be unaffected (in terms of AUROC). Other OOD detection scores (e.g. GradNorm) will be affected.

**Notation** The penultimate layer output is denoted as $\mathbf{z}$, which is the feature vector occurring immediately before the logit layer. The vector of logits is $f(\mathbf{z})$ and is typically computed via a linear layer, $f(\mathbf{z}) = \mathbf{W}\mathbf{z} + \mathbf{b}$ for a weight matrix $\mathbf{W}$ and bias vector $\mathbf{b}$. The output of a classifier network is the predictive distribution $\mathbf{p} = \mathrm{Softmax}(f(\mathbf{z}))$. Categorical distributions over $C$ classes are elements of the $C$-dimensional unit simplex $\Delta^C$. Equality up to an irrelevant constant is denoted by $\doteq$.

## 3. Generalized Entropy Score

Our aim is to rely solely on the logits and in particular on the predictive distribution as much as possible for OOD detection, because relatively simple scores using only this information are performing surprisingly well [13, 14, 18, 28]. Further, such an approach is agnostic to any information on the classifier training, the training set, or explicit OOD samples. The backbone of a classifier can be even a black box computation. Finally, the neural collapse hypothesis [32] states that the features from the penultimate layer have very limited additional information compared with the logits.

Our main assumption is, that the training loss for a classifier is dominated by a term that is minimal for a "pure" one-hot predictive distribution, which is a valid assumption for a wide range of losses (such as cross-entropy, squared Euclidean loss, label smoothing loss [39], focal loss [27] and more). Hence, ID test samples close to the training data are expected to result in a confident prediction. The prediction confidence can be measured in a variety of ways, and a statistical distance to either the uniform distribution or to a one-hot distribution. Common statistical distances are in the $f$-divergence family (e.g. [26]), Wasserstein metric [20, 42] and the total variation distance.

Here we borrow the concept of generalized entropy from the literature on proper scoring rules [4, 9]: a *generalized entropy* $G$ is a differentiable and concave function on the space of categorical distributions $\Delta^C$. The Bregman divergence $D_G(\mathbf{p}\|\mathbf{q})$ between 2 elements $\mathbf{p}, \mathbf{q} \in \Delta^C$ is the linearization error

$$D_G(\mathbf{p}\|\mathbf{q}) := G(\mathbf{q}) - G(\mathbf{p}) + (\mathbf{p} - \mathbf{q})^\top \nabla G(\mathbf{q}), \qquad (1)$$

which is non-negative for concave $G$. We assume that $G$ is invariant under permutations of the elements in $\mathbf{p}$ (all class labels are treated equally). Now the Bregman divergence between $\mathbf{p}$ and the uniform categorical distribution $\mathbf{u} = \mathbf{1}/C$ reduces to the negated generalized entropy (up to additive constants),

$$\begin{aligned} D_G(\mathbf{p}\|\mathbf{u}) &= G(\mathbf{u}) - G(\mathbf{p}) + (\mathbf{p} - \mathbf{u})^\top \nabla G(\mathbf{u}) \\ &\doteq -G(\mathbf{p}) + \underbrace{(\mathbf{p} - \mathbf{u})^\top \nabla G(\mathbf{u})}_{=0}. \end{aligned} \qquad (2)$$

The last term vanishes since $\nabla G(\mathbf{u}) = \nabla G(\mathbf{1}/C) = \kappa \mathbf{1}$ (for some $\kappa \in \mathbb{R}$, using our assumption of permutation invariance for $G$) and therefore $(\mathbf{p} - \mathbf{u})^\top \nabla G(\mathbf{u}) \propto \mathbb{E}_{\mathbf{p}}[\kappa] - \mathbb{E}_{\mathbf{u}}[\kappa] = 0$. Overall, using a negated entropy as score can be interpreted as a statistical distance between the predictive distribution $\mathbf{p}$ and the uniform distribution $\mathbf{u}$.

Our particular attention is on the following family of generalized entropies,

$$G_\gamma(\mathbf{p}) = \sum\nolimits_j p_j^\gamma (1 - p_j)^\gamma \qquad (3)$$

qfor a $\gamma \in (0, 1)$. It is straightforward to verify that the mapping $p \mapsto p^\gamma (1-p)^\gamma$ is concave in the domain $[0, 1]$ for all $\gamma \in [0, 1]$. The choice $\gamma = 1/2$, i.e.

$$G_{1/2}(\mathbf{p}) = \sum\nolimits_j \sqrt{p_j(1 - p_j)}, \qquad (4)$$

is connected to the (non-robust) exponential loss occurring in the boosting method (as detailed in [2]), and therefore considered to be more sensitive than e.g. the Shannon entropy $H(\mathbf{p}) = -\sum_j p_j \log p_j$[1]. Lower values of $\gamma$ amplify this behavior: Fig. 3 depicts the graphs of $H$, $G_{1/2}$ and $G_{1/10}$ for a Bernoulli random variable with parameter $p$. In particular the entropy $G_{1/10}$ increases rapidly near $p = 0$ and $p = 1$. Hence, $G_{1/10}$ can be seen as very sensitive detector for uncertainties in the predictive distribution.

To sum up, the motivation behind GEN is simple and straightforward. The aim of using a generalized entropy is to amplify minor deviations of a predictive distribution from the ideal one-hot encoding. In practice, this high sensitivity turns out to require some degree of robustness (and

[1]The regular Shannon entropy in analogy leads to the soft-plus loss in logistic regression.

numerical stability) in the fine-grained classification setting, which we achieve by "trimming" the predictive distribution described next.

**Truncation** If we consider sorted predictive probabilities, $p_{j_1} \geq p_{j_2} \geq \cdots \geq p_{i_C}$, then the generalized entropy $G_\gamma$ as a sum over all classes can be dominated by the tail, i.e. the large fraction of very small probabilities. Random but small variations in those probabilities have a significant impact on the score. With growing $C$, extremely small but random tails can change the sort order of discrete probabilities w.r.t. the generalized entropy. Hence, the ability of generalized entropies to discriminate finely between probability vectors near the boundary (compared to the regular Shannon entropy) comes at a cost in the many-class setting. Using a truncated sum over the top-$M$ classes made $G_\gamma$ robust in synthetic setups. Overall, our score is designed to capture small entropy variations in the top-$M$ classes.

## 4. Experiments

OOD detection benchmarks have matured over the years—there has been a transition from small scale datasets such as CIFAR-10, CIFAR-100 to more realistic large-scale dataset such as ImageNet-1K [34] and OpenImage-O [22], and the evaluation metrics have converged to AUROC and FPR95 values. We follow the recent development in evaluation strategy which we describe in Sec. 4. In our experiments, we closely follow the large-scale evaluation protocol conducted in ViM [43]. In particular, the choice of discriminative models with officially released pre-trained weights as well as the large-scale ID / OOD datasets. Note that all the methods studied in this work are deterministic.

**Models** We used several commonly-used convolutional and transformer-based architectures for large-scale image classification. These include Big Transfer [21], Vision Transformer [7], RepVGG [6], ResNet-50-D [12], DeiT [40], and Swin [29]. Big Transfer (*BiT*) [21] refers to the set of large neural network architectures and techniques (such as large batches, group normalization and weight standardization) for an efficient transfer learning and improved generalization. We utilized a variant with ResNet-v2-101 (BiT-S-R101x1 checkpoint). Vision Transformer (*ViT*) [7] is a pure transformer-based model for image classification. Its input image is cut into a sequence of patches with corresponding position embeddings. We use ViT-B/16 version in our experiments. *RepVGG* [6] model combines VGG and ResNet architectures in a way that allows for structural re-parameterizations. In particular, RepVGG is turned from a multi-branch ResNet-like network topology (used for training) into a plain VGG-like architecture with only $3 \times 3$ convolutions (used for inference). *ResNet-50-D* is one of the refined versions of ResNet architecture proposed by [12] to improve its performance. Shifted WINdows (*Swin*) transformer [29] injects priors coming from vision, such as hierarchy, locality and translational invariance, into a vision transformer network. Data-efficient image Transformers (*DeiT*) [40] is a token-based strategy for transformer distillation that enables efficient training and produces competitive results on downstream tasks. Specifications of the aforementioned architectures are summarized in Table 2.

| Classifier | Feat. | Top-1 (%) | Params |
|---|---|---|---|
| BiT-S-R101x1 [21] | 2048 | 81.30 | 44.54M |
| BiT-S-R101x1 [21] (ckpt [18]) | 2048 | 75.19 | 44.54M |
| ViT-B/16 [7] | 768 | 85.43 | 86.86M |
| RepVGG-B3 [6] | 2560 | 80.52 | 120.52M |
| ResNet-50-D [12] | 2048 | 80.52 | 23.53M |
| DeiT-B/16 [40] | 768 | 81.98 | 85.80M |
| SWIN-B/4 [29] | 1024 | 85.27 | 86.74M |

Table 2. Specifications of different architectures: dimensionality of the feature (penultimate layer output) space, top-1 accuracy on ImageNet-1K validation dataset, and the number of parameters.

**Datasets** We perform OOD detection on a large-scale OOD detection benchmark with ImageNet-1K [34] as ID dataset. We evaluate our methods using four commonly-used OOD datasets, which include OpenImage-O [43], Texture [3], iNaturalist [41], and ImageNet-O [16]. These datasets cover different domains including fine-grained images, scene images, textures images, *etc*. In particular, ImageNet-O consists of natural adversarial examples that are unforeseen classes in ImageNet-1K and cause model's performance to significantly degrade. OpenImage-O is the largest OOD dataset for ImageNet-1K released by ViM [43]. The authors discover that previous datasets like SUN [44], Places [48], and Texture [3] have a subset of images that is indistinguishable from ID data and thus manually select images from OpenImage-v3 dataset [22] that are OOD w.r.t. ImageNet-1K. Specifications of the used datasets are summarized in the supplementary material.

**Post-hoc methods** First and foremost, we compare GEN to the scores within the same family of post-hoc methods, *i.e*. not requiring prior access to the training dataset with or without labels. The first group of methods includes MSP [14], MaxLogit [13], Energy [28], and GradNorm [18] that operate on the output space. In addition, the score function that uses negative Shannon entropy is also considered. The second group comprises input / feature enhancing methods like ODIN [25] and ReAct*. ReAct* is a local version of ReAct [38] that clips penultimate activations of the current sample based on the values alone. We furthermore combine GEN with ReAct* to achieve better performance.

| | *Classifier* + OOD Method | OpenImage-O | | Textures | | iNaturalist | | ImageNet-O | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ |
| | *BiT-S-R101x1* | | | | | | | | | | |
| Post-hoc | MSP [14] | 83.05 | 76.21 | 79.76 | 77.13 | 87.90 | 64.53 | 57.16 | 96.90 | 76.97 | 78.69 |
| Post-hoc | MaxLogit [13] | 82.33 | 79.75 | 81.65 | 73.59 | 86.78 | 70.52 | 62.99 | 96.90 | 78.44 | 80.19 |
| Post-hoc | Energy [28] | 80.59 | 82.00 | 81.10 | 73.91 | 84.52 | 74.93 | 63.56 | 96.35 | 77.44 | 81.80 |
| Post-hoc | GradNorm [18] | 70.68 | 79.34 | **83.12** | **55.72** | 86.13 | **58.34** | 53.73 | **91.90** | 73.42 | **71.33** |
| Post-hoc | ODIN [25] | **85.64** | **72.83** | 81.60 | 74.07 | 86.73 | 70.75 | 63.00 | 96.85 | 79.24 | 78.63 |
| Post-hoc | ReAct* | 80.83 | 81.85 | 81.44 | 73.74 | 84.77 | 74.80 | 63.63 | 96.30 | 77.67 | 81.67 |
| Post-hoc | Shannon Entropy | 83.98 | 80.48 | 81.30 | 76.32 | 88.73 | 69.66 | 60.42 | 97.30 | 78.61 | 80.94 |
| Post-hoc | **GEN (Ours)** | 83.77 | 80.43 | 81.48 | 77.93 | 88.67 | 68.32 | 66.09 | 97.30 | 80.00 | 81.00 |
| Post-hoc | **GEN (Ours)** + ReAct* | 83.99 | 80.35 | 81.80 | 77.87 | **88.90** | 68.03 | **66.18** | 97.25 | **80.22** | 80.88 |
| Require ID | KL Matching [13] | 87.94 | 54.92 | 86.91 | 50.89 | 92.95 | **33.19** | 65.76 | 86.80 | 83.39 | 56.45 |
| Require ID | Mahalanobis [24] | 82.62 | 66.24 | 97.33 | 13.95 | 85.79 | 64.71 | 80.37 | 70.20 | 86.53 | 53.77 |
| Require ID | ReAct [38] | 85.43 | 67.45 | 90.65 | 50.14 | 91.50 | 48.65 | 67.04 | 91.50 | 83.66 | 64.44 |
| Require ID | pNML [1] | 88.62 | 55.27 | 93.59 | 22.25 | **93.12** | 38.21 | 67.27 | 86.35 | 85.65 | 50.52 |
| Require ID | Residual [43] | 80.20 | 68.05 | 97.67 | 11.14 | 76.93 | 80.18 | 81.58 | 65.60 | 84.09 | 56.24 |
| Require ID | ViM [43] | 89.96 | 49.01 | **98.92** | **4.63** | 89.38 | 55.09 | 83.85 | **61.25** | 90.53 | 42.50 |
| Require ID | **GEN (Ours)** + ReAct [38] | 85.36 | 78.22 | 84.68 | 74.09 | 90.27 | 62.36 | 67.54 | 97.10 | 81.96 | 77.94 |
| Require ID | **GEN (Ours)** + Residual [43] | **91.75** | **43.83** | 98.54 | 5.78 | 92.25 | 47.13 | **83.88** | 63.70 | **91.61** | **40.11** |
| | *Swin* | | | | | | | | | | |
| Post-hoc | MSP [14] | 91.38 | 34.81 | 85.31 | 51.74 | 94.76 | 22.97 | 78.86 | 63.90 | 87.58 | 43.36 |
| Post-hoc | MaxLogit [13] | 92.09 | 26.70 | 84.81 | 47.23 | 95.71 | 15.34 | 81.07 | 52.10 | 88.42 | 35.34 |
| Post-hoc | Energy [28] | 91.24 | 26.92 | 82.80 | 51.57 | 95.19 | 15.49 | 82.00 | **45.85** | 87.81 | 34.96 |
| Post-hoc | GradNorm [18] | 45.52 | 77.94 | 37.12 | 93.02 | 33.79 | 88.81 | 50.27 | 78.05 | 41.68 | 84.45 |
| Post-hoc | ODIN [25] | 91.38 | 28.42 | 85.74 | 44.59 | 94.24 | 19.65 | 80.62 | 53.65 | 88.00 | 36.58 |
| Post-hoc | ReAct* | 91.23 | 26.98 | 82.79 | 51.69 | 95.18 | 15.50 | 82.00 | 45.90 | 87.80 | 35.02 |
| Post-hoc | Shannon Entropy | 93.16 | 25.61 | 87.15 | 43.84 | 95.95 | 16.21 | 82.13 | 51.95 | 89.60 | 34.40 |
| Post-hoc | **GEN (Ours)** | **94.70** | **22.60** | **89.43** | **40.95** | **97.25** | **11.55** | **84.45** | 54.00 | **91.46** | **32.28** |
| Post-hoc | **GEN (Ours)** + ReAct* | 94.69 | 22.62 | 89.42 | 41.01 | **97.25** | 11.56 | 84.44 | 54.00 | 91.45 | 32.30 |
| Require ID | KL Matching [13] | 91.86 | 39.93 | 86.82 | 53.24 | 94.75 | 27.76 | 81.78 | 67.30 | 88.80 | 47.06 |
| Require ID | Mahalanobis [24] | 94.35 | 34.85 | 89.95 | 49.09 | 98.69 | 5.38 | 85.43 | 73.65 | 92.11 | 40.74 |
| Require ID | ReAct [38] | 93.71 | 22.61 | 85.62 | 47.79 | 97.49 | 9.99 | 83.83 | **44.95** | 90.16 | 31.34 |
| Require ID | pNML [1] | 95.53 | **19.29** | 91.55 | **33.29** | 97.84 | 8.98 | 87.22 | 45.05 | 93.03 | **26.65** |
| Require ID | Residual [43] | 94.44 | 33.40 | 91.36 | 43.26 | 98.90 | 4.79 | 86.66 | 68.65 | 92.84 | 37.53 |
| Require ID | ViM [43] | **95.93** | 24.43 | **92.40** | 37.98 | **99.29** | **2.62** | **88.74** | 59.00 | **94.09** | 31.01 |
| Require ID | **GEN (Ours)** + ReAct [38] | 94.80 | 22.23 | 89.47 | 40.85 | 97.42 | 10.67 | 84.48 | 54.25 | 91.54 | 32.00 |
| Require ID | **GEN (Ours)** + Residual [43] | 95.73 | 25.06 | 92.23 | 37.66 | 99.13 | 3.10 | 88.07 | 61.50 | 93.79 | 31.83 |

Table 3. *Per-Dataset Performance of OOD Detection Methods.* The classifiers are BiT [21] and Swin [29]. The ID dataset is ImageNet-1K, the OOD datasets are OpenImage-O, Textures, iNaturalist and ImageNet-O. For GEN, the number of maximal logits is set to 10% and $\gamma = 0.1$. Clipping quantile for ReAct* is set to 0.9995, and for ReAct [38] — to 0.999. The best performing method is in bold, the second best is underlined.

**Methods requiring ID train data** One of the advantages of GEN is that it does not require ID training dataset. Nevertheless, when the training data is available, it is potentially beneficial to combine this information with GEN (see Table. 1). We compare it to existing methods that require pre-computation of the training data statistics, such as KL Matching [13], Mahalanobis [24], pNML [1], Residual, and ViM [43].

**Evaluation metrics** We use two standard evaluation metrics for OOD detection. The first one is the area under the receiver operating characteristic curve (AUROC), for which higher values indicate better performance. The second one is FPR95 — the false positive rate when the true positive rate is 95%. Lower FPR95 values are better. The reported units for both metrics in all tables are percentages.

| | OOD Method | RepVGG [6] | | ResNet-50-D [12] | | ViT [7] | | DeiT [40] | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ |
| Post-hoc | MSP [14] | 78.02 | 70.83 | 77.99 | 68.10 | 89.33 | 41.89 | 79.44 | 66.22 | 81.19 | 61.76 |
| | MaxLogit [13] | 77.47 | 73.55 | 75.47 | 69.28 | 94.56 | 24.34 | 76.77 | 64.37 | 81.07 | 57.89 |
| | Energy [28] | 76.29 | 79.11 | 71.25 | 78.01 | **94.89** | 22.54 | 72.81 | 69.88 | 78.81 | 62.39 |
| | GradNorm [18] | 52.98 | 94.98 | 44.04 | 96.08 | 90.32 | 28.66 | 32.05 | 97.47 | 54.85 | 79.30 |
| | ODIN [25] | 77.72 | 72.68 | 75.27 | 68.56 | 94.57 | 24.25 | 77.13 | 63.92 | 81.17 | 57.35 |
| | ReAct* | 77.60 | 78.57 | 71.55 | 77.70 | **94.89** | **22.83** | 72.82 | 69.87 | 79.21 | 62.24 |
| | Shannon Entropy | 79.01 | 71.81 | 78.82 | 66.41 | 91.91 | 30.41 | 80.61 | 61.78 | 82.59 | 57.60 |
| | **GEN (Ours)** | 81.33 | 66.00 | 82.75 | **62.08** | 94.31 | 26.14 | **84.61** | **59.68** | 85.75 | **53.47** |
| | **GEN (Ours)** + ReAct* | **82.88** | **65.64** | **82.80** | 62.29 | 94.31 | 26.23 | 84.60 | 59.77 | **86.15** | 53.48 |
| Require ID | KL Matching [13] | 81.29 | 61.65 | 82.66 | 64.83 | 90.81 | 36.04 | 83.46 | 64.66 | 85.40 | 54.85 |
| | Mahalanobis [24] | 85.91 | 59.80 | 88.11 | 56.38 | 95.96 | 19.68 | 85.08 | 72.75 | 89.43 | 49.87 |
| | ReAct [38] | 65.42 | 96.29 | 77.68 | 66.45 | 95.13 | 21.93 | 73.95 | 68.39 | 78.04 | 63.27 |
| | pNML [1] | 83.23 | 55.37 | 84.19 | **50.20** | 92.75 | 28.12 | 83.09 | 61.39 | 85.81 | 48.77 |
| | Residual [43] | 83.96 | 59.44 | 86.72 | 59.44 | 92.71 | 31.50 | 84.18 | 73.97 | 88.08 | 52.37 |
| | ViM [43] | **87.65** | **50.95** | **89.03** | 53.28 | **96.16** | **18.46** | 85.28 | 69.81 | **89.53** | **48.12** |
| | **GEN (Ours)** + ReAct [38] | 86.32 | 56.08 | 84.58 | 59.08 | 94.44 | 25.80 | 84.65 | **60.06** | 87.50 | 50.26 |
| | **GEN (Ours)** + Residual [43] | 87.49 | 51.67 | 89.07 | 53.44 | 95.73 | 20.69 | **85.59** | 67.51 | 89.47 | 48.33 |

Table 4. *Average Performance of OOD Detection Methods.* Results are shown for RepVGG [6], ResNet-50-D [12], ViT [7], and DeiT [40] architectures with ImageNet-1K as ID data. The reported are averaged results over four OOD datasets: OpenImage-O, Textures, iNaturalist and ImageNet-O. For GEN, the number of maximal logits is set to 10% and $\gamma = 0.1$. Clipping quantile for ReAct* is set to 0.9995, and for ReAct [38] — to 0.999. The best performing method is in bold, the second best is underlined.

### 4.1. OOD Detection Performance Results

In this section, the results of the OOD detection benchmark are presented. We reproduce the results for all methods (except for ODIN [25]) and obtain slightly different results than reported in [43]. In our experiments, we used NVIDIA GeForce RTX 3080, CUDA 11.5 + PyTorch 1.11.

**Results on BiT and Swin** We show detailed results on BiT [21] and Swin [29] architectures, since BiT is commonly used for large-scale OOD detection [18, 19, 37, 43] and Swin [29] is the recent transformer-based architecture.

The results of BiT [21] are presented in the top half of Table 3. First, one can see from the "Post hoc" block that our score achieves the highest average AUROC (across four datasets) compared to other post-hoc methods. In particular, we obtain the highest AUROC on ImageNet-O and iNaturalist. Furthermore, using feature clipping further improves the performance in terms of AUROC and FPR95. For this classifier, GradNorm [18] gives lower FPR95 values. We think this could be connected to the lower classification accuracy of the pre-trained models (see Tab. 2) and/or model specifics because GradNorm performs significantly worse for other classifiers (see Tab. 4 and Tab. 2 in Supplementary). Then we look into the methods using ID data statistics. Our score is combined with ReAct [38] and Residual [43] methods, which compute compressed information of feature space from all training data. Results from the "Require ID" block show that using information from feature space could further improve our score. Specifically, our method combined with Residual [43] achieves the state-of-the-art results on in terms of the averaged AUROC and FPR95 (over four datasets) when using BiT [21], in particular on the challenging OpenImage-O dataset.

The results of Swin [29] are shown in the bottom half of Table 3. Results from the "Post-hoc" block show that our score is consistently better in terms of AUROC values than all other post-hoc methods. Particularly, our method outperforms MaxLogit [13] by 3% on average in terms of AUROC. According to the "Require ID" block, our performance is comparable to ViM [43].

To visualize OOD performance, we present the score distributions using our score and Energy [28] score in Figure 2 and it shows that our method makes ID/OOD separation better. Interestingly, the score distribution drawn based on our score function is smoother.

**Averaged results for other architectures** To further investigate the effectiveness and robustness of our score, we perform OOD detection on four remaining architectures, RepVGG [6], ResNet-50-D [12], ViT [7], and DeiT [40]. The averaged results over four datasets are shown in Table 4. First, it is apparent that our score continually gains the best AUROC on different architectures compared to all post-hoc methods. Specifically, our method outperforms MSP [14] on average with a notable margin, almost 5%. Moreover, our score also obtains the lowest aver-

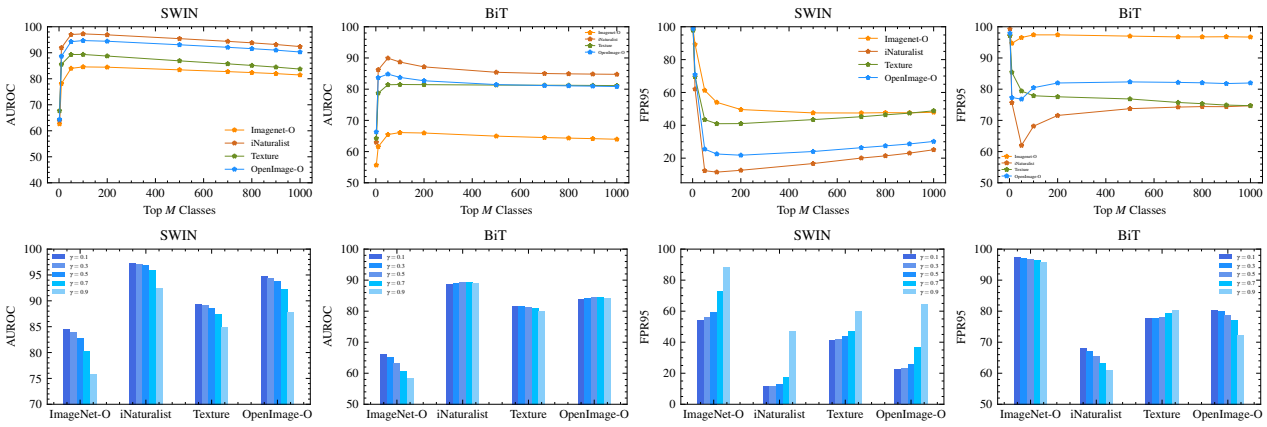


Figure 4. *Effective value of $M$ and $\gamma$.* GEN Performance for varying values of (top row) the number of largest probabilities referred as top $M$ classes, and (bottom row) the exponential scale $\gamma$ of the entropy. The left two columns correspond to Swin [29] architecture, and the right two columns correspond to BiT [21].

aged FPR95 over four datasets and four architectures. It is significantly better than all other post-hoc methods in terms of FPR95, with a nearly $4\%$ margin. The results of combining our score with information from ID dataset can be found in the bottom half of the Table 4. It shows that our method achieves competitive results compared with ViM [43].

### 4.2. Choice of $M$ and $\gamma$

We empirically show how the performance of our method varies with different $M$ and $\gamma$ in terms of AUROC and FPR95. First, we investigate the effective value of $M$ for $C = 1000$ semantic classes. The first row of Figure 4 shows the results (with $\gamma = 0.1$). The results of BiT [21] are illustrated in the two rightmost columns (with AUROC and FPR95, respectively) and the results of Swin [29] are presented in the two leftmost columns (with AUROC and FPR95, respectively). It shows that it is sufficient to use the top $M = 100$ classes for the score.

We also look into the effectiveness of using different $\gamma$. The second row of Figure 4 (with $M = 100$) shows that it is adequate to obtain better OOD performance via setting $\gamma = 0.1$ for different OOD datasets. On average, AUROC and FPR95 values are better when using lower $\gamma$. Setting $\gamma = 0.1$ also works well on other architectures. Results for the remaining architectures and the dependence of $(\gamma, M)$ on the architecture (which led to our choice of $(\gamma = 0.1, M = 100)$) can be found in the supplementary material.

The current evaluation protocol for OOD detection is performed on the test dataset directly, which is not suitable for real applications. We therefore evaluate the methods on completely unseen datasets, SUN [44] and Places [48][2]. GEN achieves the state-of-the-art performance with 1% and 3% margin in terms of both AUROC and FPR95 for post-hoc and ID requiring methods, respectively. The detailed results are in the supplementary material.

[2]We followed GradNorm [18] by taking the non-overlapping classes w.r.t. ImageNet-1k

## 5. Discussion and Conclusions

In this work, we challenged ourselves to narrow the gap between simple and fast post-hoc OOD detection methods—those working on top of (nearly) black-box classifiers—and the "white-box" methods—those benefiting from extra information such as large and representative ID dataset with or without corresponding labels. The proposed entropy-based method GEN is as easy to implement as previous methods, and the only requirement it has is that the classifier admits class probabilities. Combining GEN with more feature-based and enhancing methods is one of the potential future directions for improvement.

We found that GEN performs best when using $\approx 10\%$ of the logits with the maximal response. Interestingly, a similar observation also applies to some other post-hoc scores (with different fractions of logits), *i.e.* that it might generally be a good idea to use only partial information coming from the largest logits. The lowest logits seem to introduce noise that might be particularly damaging for OOD detection in large-scale and fine-grained classification tasks with thousands of semantic classes. More details on our experiments can be found in the supplementary material.

## Acknowledgements

This work was supported by the Wallenberg Artificial Intelligence, Autonomous Systems and Software Program (WASP), funded by Knut and Alice Wallenberg Foundation.

## References

[1] Koby Bibas, Meir Feder, and Tal Hassner. Single layer predictive normalized maximum likelihood for out-of-

distribution detection. In *NeurIPS*, 2021. 2, 3, 6, 7

[2] Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November*, 2005. 4

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 5

[4] Alexander Philip Dawid and Monica Musio. Theory and applications of proper scoring rules. *Metron*, 2014. 4

[5] Terrance DeVries and Graham W. Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 3

[6] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *CVPR*, 2021. 5, 7

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 5, 7

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1

[9] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 2007. 4

[10] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *ICLR*, 2020. 2, 3

[11] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 3

[12] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *CVPR*, 2019. 5, 7

[13] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, 2022. 1, 2, 3, 4, 5, 6, 7

[14] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 1, 2, 3, 4, 5, 6, 7

[15] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019. 3

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

[17] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data, 2020. 3

[18] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *NeurIPS*, 2021. 2, 3, 4, 5, 6, 7, 8

[19] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *CVPR*, 2021. 3, 7

[20] Leonid V Kantorovich. Mathematical methods of organizing and planning production. *Management science*, 1960. 4

[21] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *ECCV*, 2020. 5, 6, 7, 8

[22] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017. 5

[23] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017. 3

[24] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018. 1, 2, 3, 6, 7

[25] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 3, 5, 6, 7

[26] Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006. 4

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 4

[28] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020. 1, 2, 3, 4, 5, 6, 7

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2, 5, 6, 7, 8

[30] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *ICLR*, 2019. 1

[31] Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing*, 441:138–150, 2021. 3

[32] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. 4

[33] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 3, 4

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,

Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015. 5

[35] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *IPMI*, 2017. 1

[36] Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. Classifier calibration: How to assess and improve predicted class probabilities: a survey. *arXiv preprint arXiv:2112.10327*, 2021. 3

[37] Yue Song, Nicu Sebe, and Wei Wang. Rankfeat: Rank-1 feature removal for out-of-distribution detection. In *NeurIPS*, 2022. 1, 3, 7

[38] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021. 1, 3, 5, 6, 7

[39] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 4

[40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 5, 7

[41] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 5

[42] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009. 4

[43] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7, 8

[44] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 5, 8

[45] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 1

[46] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *ACM SIGKDD*, 2002. 3, 4

[47] Alireza Zaeemzadeh, Niccolò Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *CVPR*, 2021. 3

[48] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 5, 8