Operational Open-Set Recognition and PostMax Refinement

Steve Cruz[1], Ryan Rabinowitz[2], Manuel Günther[3], and Terrance E. Boult[2]

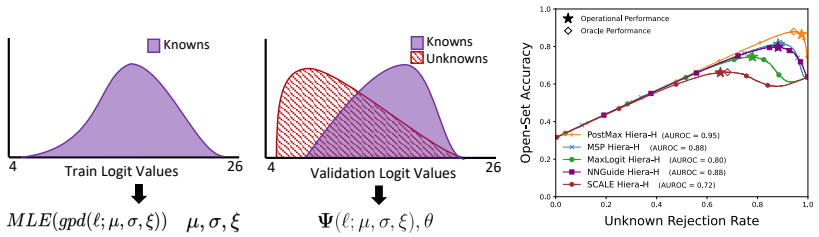
[1] University of Notre Dame
[2] University of Colorado Colorado Springs
[3] University of Zurich
stevecruz@nd.edu {rrabinow,tboult}@uccs.edu guenther@ifi.uzh.ch

Abstract. Open-Set Recognition (OSR) is a problem with mainly practical applications. However, recent evaluations have largely focused on small-scale data and tuning thresholds over the test set, which disregard the real-world operational needs of parameter selection. Thus, we revisit the original goals of OSR and propose a new evaluation metric, Operational Open-Set Accuracy (OOSA), which requires predicting an operationally relevant threshold from a validation set with known and a surrogate set with unknown samples, and then applying this threshold during testing. With this new measure in mind, we develop a large-scale evaluation protocol suited for operational scenarios. Additionally, we introduce the novel PostMax algorithm that performs post-processing refinement of the logit of the maximal class. This refinement involves normalizing logits by deep feature magnitudes and utilizing an extreme-value-based generalized Pareto distribution to map them into proper probabilities. We evaluate multiple pre-trained deep networks, including leading transformer and convolution-based architectures, on different selections of large-scale surrogate and test sets. Our experiments demonstrate that PostMax advances the state of the art in open-set recognition, showing statistically significant improvements in our novel OOSA metric as well as in previously used metrics such as AUROC, FPR95, and others.

1 Introduction

The original Open-Set Recognition (OSR) formulation [1, 38, 39] gained attention due to the flaws in existing recognition systems when handling inputs from outside the training classes. These systems lack control over inputs, and any unknown sample that is not rejected leads to a misclassification, thereby reducing accuracy. Therefore, early open-set work [1, 3, 35] introduced protocols tailored to emulate real-world systems, allowing for a proper evaluation of open-set performance.

For a decade, researchers explored techniques that enhanced OSR performance [1,5,7,10,11,20,25,27,28,35,39,55,57]. While contributing novel ideas, recent works have diverged from the original goals of OSR. Instead of leveraging large-scale data as in early research on open-set deep networks [1], recent OSR evaluations have



(a) Train **(b)** Val **(c)** Test (OOSA & AUROC)

Fig. 1: PostMax Pipeline. During train (a), post-processed maximum logits of correctly classified train samples are normalized by feature magnitude ℓ; the set is then used in a Maximum Likelihood Estimation (MLE) to obtain Generalized Pareto Distribution (GPD) [33] with parameters μ, σ, and ξ. During validation (b), we use a validation and a surrogate set where the cumulative distribution of the GPD (Ψ) gives the probability of the normalized max logit being from the train distribution and predict an operational threshold. During operation (c), we deploy each method's threshold and apply it to determine OSR and compute Operational Open Set Accuracy (OOSA). For evaluation we compare against Maximum Softmax Probability (MSP) [16,44], Maximum Logit (MaxLogit) [15,44], Nearest Neighbor Guidance (NNGuide) [32], and SCALE [50]. Operational Open-Set Accuracy (OOSA) performance is ($\star$) and ($\diamond$) is the optimal Open-Set Accuracy (OSA) for an oracle that knows the final test set (shown to highlight differences from OOSA). AUROC scores are shown for comparison.

focused on small-scale datasets with few classes, with only a few notable exceptions [30,35,44]. Moreover, the prevalent reliance on metrics such as the Area Under the Receiver Operating Characteristics (AUROC) curve for performance evaluation has led to inopportune assessments since such metrics do not offer a realistic perspective on OSR. Fig. 1 illustrates an example where AUROC scores exhibit one ranking, yet in an operational setting, there is a slight difference. Inherently, common measures overlook the real-world operational needs of threshold selection. Some may argue for their use with a dedicated validation set; however, it is not clear how one would extract a threshold from these metrics and deploy it at test time. This is problematic as methods have largely ignored a fundamental practice of machine learning systems – having separate training, validation, and test sets to ensure that models generalize, avoid overfitting, and provide unbiased evaluation and reliable performance prediction on new data.

Initial aspirations behind OSR aimed to enhance *real systems* [4]. Imagine an engineer tasked with implementing an OSR algorithm into their system. As they analyze results from various algorithms through tables and plots to inform decisions, a formidable challenge emerges. While AUROC and similar metrics assess a model's discriminative capability across all possible classification thresholds, they do not guide the engineer toward an optimal threshold for decision-making. There is a pressing need for a metric that enables confident and statistically rigorous evaluations with anticipated data, while also demonstrating robustness against varying numbers of unknowns. When deploying an open-set system, engineers must, at a minimum, consider two factors: **(1)** how an

approach scales with operational data, and **(2)** how to choose an operating point that balances their desired known classification accuracy and unknown rejection. These considerations underscore the importance of achieving overall accuracy with a predicted operational threshold. Therefore, we introduce a new metric called Operational Open-Set Accuracy (OOSA), illustrated in Fig. 1, which facilitates algorithm and operational threshold selection. To ensure real-world applicability, we implement a large-scale evaluation protocol with proper training/validation/test splits. We determine a threshold on the validation set and deploy it during testing on unseen unknowns. Additionally, for completeness, fairness, and robustness, we evaluate state-of-the-art algorithms on ImageNet-scale open-set recognition using common metrics (AUROC, FPR95, AUOSCR [6], and OpenAUC [46]), many of which are detailed in the supplemental.

Before incorporating an operational threshold, it is natural to ask about the distribution of scores above this threshold. This naturally leads to a Peak-over-Threshold (POT) formulation of the distribution, best modeled via Extreme Value Theory (EVT). The Pickands–Balkema–De Haan theorem of EVT [33] yields a Generalized Pareto Distribution (GPD) over normalized scores, as depicted in Fig. 1. This GPD-based score transformation provides theoretically grounded probabilities rather than raw, network-dependent ad-hoc confidence scores (see supplemental for details). While GPD can be applied directly to Softmax confidences or logit values, our approach goes further. We introduce a novel algorithm, **PostMax**, which applies GPD to post-processed normalized maximum logits. Here, normalization by the deep feature magnitude enhances the separation of known and unknown classes. Prior observations and theoretical explanations [10, 13, 31, 44] suggest that deep feature magnitudes extracted from inputs that the network was not trained on are generally smaller than those for known samples. However, our findings show the exact opposite, *i.e.*, that modern networks trained on the large-scale ILSVRC2012 dataset generally exhibit larger deep feature magnitudes for unknown samples (details in supplemental). Therefore, applying GPD to logit values *divided* by the deep feature magnitude can enhance OSR beyond the strong baselines of using raw Softmax confidence [16], maximum logit values [15], overconfidence reduction with features and logits [32], or scaled logit values [50]. The contributions of this paper are as follows:

- We design a novel evaluation metric, Operational Open-Set Accuracy (OOSA), that emphasizes real-world usage by predicting an operational threshold on a validation set (knowns and unknowns surrogate), which is then deployed during testing on unseen data.
- We introduce a novel algorithm, PostMax, that post-processes maximum logits by normalizing them by deep feature magnitude and then applies GPD to provide probabilities. Our code is publicly available.[4]
- We develop new large-scale evaluation protocols suited for assessing algorithms in operational scenarios.
- We showcase that PostMax advances the state of the art in large-scale open-set recognition with statistical significance on OOSA and prior metrics.

[4] `https://github.com/Vastlab/PostMax-OOSA`

2 Related Work

Improving OSR, distinct from OOD detection, anomaly detection, or novel category discovery, can be approached in two ways: providing better features through improved network training, or through post-processing where a pre-trained network is trained for closed-set tasks and then adapted for OSR. Our work is rooted in the post-processing approaches.

2.1 Related Problems

A common question regarding open-set is, *What about a two-stage system with OOD followed by classification, does that not solve OSR?* This hypothesis is rejected with an example from one of our operational systems where we have to recognize objects in novel contexts, *e.g.*, in snowy or foggy conditions. Such conditions are OOD with respect to training samples, but they should **not** be rejected from an open-set point of view. Out-Of-Distribution (OOD) detection, though distinct from OSR, has been more consistent in its use of large-scale experimentation as were used in early OSR work [15,16,32,45,47,50,52,54]. Even when OOD is restricted to out-of-class, the two-stage system needs to be evaluated in that context, including a process for selecting an OOD threshold and then evaluating the resulting classification performance. The OpenOOD Benchmark [52,54] has implemented various OSR and OOD techniques such as OpenGan [20], MOS [19], ReAct [41], ViM [45], GEN [23], NNGuide [32], and SCALE [50]. While we caution the use of OOD algorithms in OSR evaluations/settings, based on the OpenOOD Benchmark leaderboard, we compare with the state-of-the-art, NNGuide & SCALE.

2.2 Closed-Set Classifiers

Recently, Vaze *et al.* [44] argued that closed-set classifiers are sufficient. This aligns with Hendrycks *et al.* [15,16], who demonstrated thresholding on Softmax confidences or, especially, on logits provide unreasonably good baselines. Our work shows that such approaches/networks can be improved with normalization.

We evaluate several leading pre-trained architectures trainined on ILSVRC2012-1K with no additional data. Particularly, Meta's Vision Transformer Hiera-H [37], which strips non-essential components making it faster and more accurate during training/inference. Also, we use their Masked Autoencoders (MAE) model, ViT-H [14], which masks random patches of an image and reconstructs the missing pixels. To show generalization, we also utilize CNNs, including Meta's ConvNeXtV2-H [48] and ConvNeXt-L [24] which modernize a standard ResNet. Other networks are found in the supplemental material.

2.3 Post-processing Approaches

Besides thresholding softmax scores or logits [15, 16, 44], OSR methods take features and use Weibull-calibrated Support Vector Machines (W-SVM) [38]

or Extreme Value Machines (EVM) [35] to estimate a probability of unknown. OpenMax [1] was the first method to add an artificial probability of unknown to closed-set networks by computing a logit score for the unknown class based on deep feature similarities to features of known classes. Other approaches include a small adaptor network [13, 42] that adds open-set capabilities [10] to features extracted from closed-set networks.

2.4 Learning-based Approaches

Most methods use *negative* aka. *known unknown* samples in training, which represent samples of no interest to the classifier and should be classified as unknown, with a few exemptions such as replacing the final Softmax classifier with a set of binary classifiers [40, 43]. While some use real negative samples [10, 13, 30], the majority of research tries to artificially create them. For example, combinations of knowns [57], added noise to knowns [47], or Generative Adversarial Networks (GANs) [11, 20]. In our approach, we avoid the use of any negative samples.

2.5 Evaluation Techniques

Evaluating OSR correctly in a real-world setting is difficult and current measures do not satisfy operational requirements. For example, the widely used AUROC metric [5, 7, 27, 28, 44, 51] only looks at the binary decision of known *vs*. unknown, but ignores the task of assigning the correct class for known samples. Such metrics are often combined with closed-set classification accuracy in order to evaluate OSR. These evaluations are reasonable, but provide no information under realistic scenarios, *i.e.*, an operational threshold classifying samples as known or unknown.

Another metric often employed is the macro-averaged F1 score [1, 27, 35, 53], which treats unknown as a separate class. This measure has many counter-intuitive properties that make it difficult to interpret the results. For example, a different threshold is required for each class so a sample could be classified as several known classes and unknown (at the same time). Thus, incorrect classification of unknown samples can easily be overlooked.

Recently, Dhamija *et al.* [10] introduced the Open-Set Classification Rate (OSCR) curve. This curve handles knowns and unknowns separately, evaluating the Correct Classification Rate (CCR) at various thresholds corresponding to specific False Positive Rates (FPR). While valuable for high-risk applications like open-set face recognition, its practicality in general OSR tasks is questionable. In scenarios where a threshold is chosen based on a specific FPR, the curve utilizes unknowns from the test set, raising doubts about its applicability to unseen classes. Moreover, the curve faces the challenge of not providing a single value for comparing different methods. Additionally, the proposed metrics, Area Under the OSCR (AUOSCR) [6] and Open Area Under the ROC Curve (OpenAUC) [46], do not provide an intuition on how to select an operating threshold.

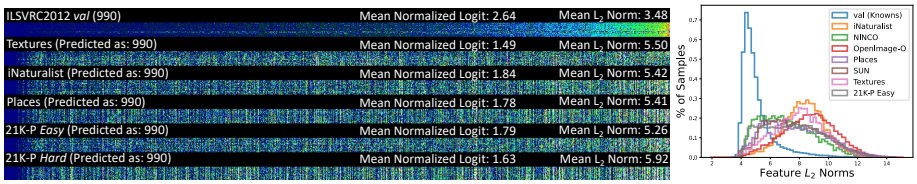


Fig. 2: Motivation for Normalization. We illustrate why PostMax divides by the feature magnitude norm before applying GPD. The comparison above shows squared deep feature values extracted from a pre-trained ILSVRC2012-1K Transformer (ViT-H [14]). We sorted the deep features of 20 randomly sampled images from class 990 based on increasing mean square magnitude. Next, we applied this sorted order to the features of 20 images misclassified as class 990 from each unknown dataset. Unknowns exhibit high feature values (bright colors) in regions where known sample activations are low. These large, uncorrelated responses accumulate to give it a high score for the class but also result in unknown dataset features having, on average, larger L_2 norms than those of the correct class, depicted in the magnitude histogram on the right. Therefore, dividing by magnitude enhances separation.

3 Approach

Several approaches advocate selecting the maximum class and then applying thresholding to certain scores, *e.g.*, Softmax [16] or logits [15], to address OSR. While such thresholding proves effective in some recognition tasks [44], its relevance often extends only to establishing a ranked order of classes for individual samples. We focus on leveraging information from the features themselves and the formal distribution of these maxima. The result is PostMax, an effective approach with theoretically grounded probabilities to address OSR through (1) normalization and (2) score distribution.

3.1 Normalization & GPD

From Fig. 2, unknown samples activate high dimensions in the feature space, which goes against our intuition that deep networks would only learn features necessary to classify known classes and disregard those belonging to other uninteresting ones. We attribute these findings to high logit values and Softmax scores, where positions with high feature magnitudes align with larger weights of certain classes, even when an input exhibits a large feature magnitude and the high-scoring class has a small or negative weight. The observed effect may depend on the network's loss function and the training dataset; in small-scale datasets such as MNIST or CIFAR, which only differentiate between 10 classes, the opposite has been observed [10, 18, 31, 44]. Further details regarding the normalizations in these works and how they differ from our findings are provided in the supplemental.

As we classify samples based on their "maximum score" threshold, a key question arises: *Does this score belong to the training distribution of maximum scores?* Statistical Extreme Value Theory (EVT) provides a grounded approach to

Algorithm 1 PostMax Fitting. This implements the probability distribution estimation of the proposed PostMax algorithm.

Require: D_{train}, $\text{FV}(x)$, $\text{L}(x)$
  $M \leftarrow \{\}$
  for each $(x, t) \in D_{\text{train}}$ **do**
    if $\arg\max \text{L}(x) = t$ **then**
      $\ell \leftarrow \frac{\max \text{L}(x)}{\|\text{FV}(x)\|}$
      $M \leftarrow M \cup \{\ell\}$
    end if
  end for
  $\mu, \sigma, \xi \leftarrow$ `scipy.stats.genpareto.fit`(M)

answering this question. Unlike ad-hoc Softmax confidences, transforming logits into real probabilities allows probabilities to be more interpretable compared to Softmax, probabilities do not need to sum up to one (images may contain multiple objects), and EVT models probabilities on are based on thousands/millions of samples rather than the current sample. Given that thresholdling maximum logits or Softmax scores serves as a reasonable baseline [15, 32, 44], applying EVT to this problem is both theoretically justified and intuitive. Note, prior OSR research [1, 26, 35, 38] utilizes the Fisher–Tippett–Gnedenko EVT [12], resulting in a Weibull distribution. However, practical effectiveness was limited due to challenges in parameter selection affecting EVT modeling. In contrast, we employ the Generalized Pareto Distribution (GPD) derived from the Peak-over-threshold (POT) approach [33], which models extreme values above a threshold.

3.2 PostMax

Using these intuitions and observations, we define our Postnormalization of Maxima (PostMax) algorithm. We aim to find a general distribution Ψ of logit values that is valid for all of our known classes $c \in \{1, \ldots, C\}$. As observed in Fig. 2, modeling raw logit values might not be fruitful since unknown samples have generally higher activation of deep features, which often leads to high logit values. For a sample, we utilize *normalized* logit values by dividing the original logit values by their deep feature magnitude via (1).

Based on POT EVT [33], we understand that if we consider all maximum values above a threshold (such as the smallest correct logit observed during training), the resulting distribution follows a Generalized Pareto Distribution (GPD). Let $D_{\text{train}} = \{(x_i, t_i)\}$ be the collection of all samples x_i with their respective ground-truth target label t_i in the training set. Let $\text{FV}(x)$ be the function returning the feature vector of a sample x. Let $\text{L}(x)$ be the function returning the logit vector of a sample. We collect the normalized logits from the target class for all correctly classified training samples, which are then used to model a Generalized Pareto Distribution $\Psi_{\mu,\sigma,\xi}$. Details can be found in Alg. 1.

At inference for test sample x we use $\text{FV}(x)$ and $\text{L}(x)$ to extract the deep features and logits of x and then for class c assign probability $p_c(x)$ using the

normalized logit value:

$$\ell_c = \frac{\mathrm{L}(x)[c]}{\|\mathrm{FV}(x)\|} \qquad\qquad p_c(x) = \Psi_{\mu,\sigma,\xi}(\ell_c) \tag{1}$$

where $\Psi_{\mu,\sigma,\xi}(\ell)$ is the cumulative distribution of the GPD with location μ, scale σ and shape ξ computed via Alg. 1. If we just want the top class (as all experiments), we compute the class with the max logit and apply the above to that logit.

Note that we do not utilize any unknowns or negatives in this process, nor do we explicitly model a probability of unknowns. Instead, we threshold based on the cumulative GPD probability of known classes, computing the maximum value of normalized logits.

3.3 Operational Open-Set Accuracy

By designing the OpenAUC metric, Wang *et al.* [46] proposed four different conditions that a good open-set evaluation metric should fulfill, which we rephrase:

P1 The metric needs to check that known samples are classified correctly with high probability.

P2 The metric needs to evaluate if unknown samples are assigned to known classes with low probability.

P3 The metric should be insensitive to a score threshold.

P4 The metric should be a single number.

We totally agree with **P1**, **P2** and **P4**. However, we reject **P3** because it means that the metric cannot be used to select a classifier that works well in a specific operational setting, which requires obtaining an operational threshold. In our view, a useful open-set algorithm must include that step. Therefore, we reformulate condition **P3'** to be operationally relevant:

P3' The metric should indicate an operational threshold that is optimal under specified circumstances and can be applied to unseen data and unseen unknown classes.

These four conditions (**P1**, **P2**, **P3'**, **P4**) exclude all existing open-set evaluation metrics; we need a new metric to satisfy them. As discussed in Sec. 2.5, metrics such as AUROC, AUOSCR, and OpenAUC simply average over all possible operational thresholds, while F1 and normalized accuracy fail to use a single threshold [46]. On the other hand, the Open-Set Classification Rate (OSCR) curve [10] does not provide a single value for comparison, which makes it hard to compare different algorithms. Therefore, we introduce Operational Open-Set Accuracy (OOSA), which is inspired by OSCR but fulfills all four conditions.

From an engineer's view, the real question is how well a specific method can perform under the presence of both known and unknown samples in their data. Most applications of open-set classification are not security-sensitive, so there is no need to have very low False Positive Rates (FPR) as typically evaluated in the OSCR [10, 30].

Usually, an engineer has a rough estimate of how many known and unknown samples the system will observe – so they are rather interested in the total accuracy

the system will achieve and which kind of operational threshold will lead to this performance. To provide both, we exploit the FPR and CCR calculations used in OSCR [10] to define the *Unknown Rejection Rate* (URR) and the *Open-Set Accuracy* (OSA), both of which rely on the operational threshold θ and only make use of the probabilities of known classes $P_c(x)$:

$$\begin{aligned} \mathrm{CCR}(\theta) &= \frac{\left|\{x \mid x \in D_c \wedge \arg\max_c P_c(x) = \hat{c} \wedge P_c(x) \geq \theta\}\right|}{|D_c|} \\ \mathrm{URR}(\theta) &= \frac{\left|\{x \mid x \in D_a \wedge \max_c P_c(x) < \theta\}\right|}{|D_a|} \\ \mathrm{OSA}(\theta) &= \alpha \cdot \mathrm{CCR}(\theta) + (1-\alpha) \cdot \mathrm{URR}(\theta) \end{aligned} \tag{2}$$

where D_c is the collection of all known test samples with class labels $\hat{c}$, while D_a collects all unknown test samples [10]. Please note that $\mathrm{URR}(\theta) = 1 - \mathrm{FPR}(\theta)$ provides a positive view on the unknown samples, *i.e.*, it calculates how many of them are correctly rejected under threshold θ. OSA includes a weighted average over CCR and URR to provide a holistic view of the expected total accuracy, where the application-dependent α can be used to provide a stronger focus on correctly classifying known or unknown samples. When both types of samples should be treated similarly, $\alpha = \frac{|D_c|}{|D_c|+|D_a|}$ can be selected, which we do in most of our experiments. In that case, OSA can be thought of as evaluating a C class classifier using $C+1$ class accuracy where the unknown class $C+1$ is classified when the maximal probability does not exceed our operational threshold θ.

When plotting OSA over URR, by varying the operational threshold, we arrive at the plots seen in Fig. 1. This plot is *not monotonic*, but rather provides a peak where the OSA is maximized. The maximum value can be used to select a threshold θ^* on a given set of scores S effectively:

$$\begin{aligned} \theta^* &= \arg\max_{\theta \in S} \mathrm{OSA}_S(\theta) \\ S &= \left\{P_{\hat{c}}(x) \mid x \in D_c\right\} \cup \left\{\max_c P_c(x) \mid x \in D_a\right\} \end{aligned} \tag{3}$$

Here, S is the collection of all maximum scores for known and unknown samples. When applying this calculation on the test set S_{test}, we will find the optimal threshold θ^*_{test}, indicated by $\diamond$ in Fig. 1. However, since test set samples are unknown in operational settings, we make use of the validation set D_{val} of known samples as D_c and an additional surrogate set D_{sur} of unknown samples to replace D_a in (3) for obtaining the operational threshold θ^*_{sur}, which we will then apply to the test set: $\mathrm{OSA}_{S_{test}}(\theta^*_{sur})$, which is indicated by $\star$ in Fig. 1. How to optimally select this surrogate set is discussed below.

4 Experiments

Although a few OSR methods have explored large-scale datasets [1, 30, 35, 44], many recent techniques [5–7, 10, 11, 20, 25, 27, 28, 55, 57] still lack comprehensive large-scale evaluations. The performance on small-scale data such as MNIST [22],

Table 1: Operational Open-Set Accuracy (OOSA). The mean OOSA (↑) of all methods. To compute, we validate methods on ImageNetV2 [34] ($10K$ images) as knowns and 20% of our surrogate 21K-P *Hard* [44] ($9.8K$ images) as unknowns and predict an operational threshold. Then, we deploy each method's threshold and test on five different ILSVRC2012 *val* [36] splits (each $10K$ images) and specified unknowns. Standard deviation is omitted as it is < 0.005 for all methods/unknowns. OSR is performed on extractions from two state-of-the-art pre-trained architectures on ILSVRC2012 - (1) Transformer Hiera-H [37] (2) CNN ConvNeXtV2-H [48]. The best scores are in **bold**.

Unknowns (# images)	**OOSA ↑**									
	Hiera-H [37]					**ConvNeXtV2-H** [48]				
	PostMax (Ours)	**MSP** [16, 44]	**MaxLogit** [15, 44]	**NNGuide** [32]	**SCALE** [50]	**PostMax** (Ours)	**MSP** [16, 44]	**MaxLogit** [15, 44]	**NNGuide** [32]	**SCALE** [50]
iNaturalist [17] ($10K$)	**0.825**	0.815	0.780	0.782	0.727	**0.812**	0.796	0.786	0.713	0.772
NINCO [2] ($\sim 5.8K$)	**0.740**	0.739	0.707	0.625	0.667	**0.732**	0.723	0.718	0.597	0.705
OpenImage-O [45] ($\sim 17.6K$)	**0.864**	0.814	0.745	0.796	0.663	**0.849**	0.820	0.792	0.745	0.767
Places [56] ($10K$)	**0.785**	0.754	0.691	0.745	0.625	**0.779**	0.749	0.710	0.693	0.676
SUN [49] ($10K$)	**0.795**	0.758	0.709	0.749	0.651	**0.783**	0.751	0.721	0.693	0.690
Textures [8] ($\sim 5.1K$)	**0.764**	0.756	0.736	0.660	0.715	**0.744**	0.736	0.740	0.601	0.740
21K-P *Easy* [44] ($50K$)	**0.818**	0.740	0.662	0.744	0.568	**0.813**	0.763	0.691	0.783	0.633

CIFAR [21], or SVHN [29] does not reflect operational scenarios and lacks generalizability due to small image crops and a limited number of classes [30].

An open-set protocol is essential and necessitates a large-scale setting for the proper evaluation of real-world scenarios [30, 44]. Additionally, it is paramount to employ proper machine-learning methodology, including a validation set (and a surrogate set) for operational threshold selection. Thus, we evaluate the robustness and generalization of PostMax on multiple large-scale datasets and splits with realistic and high-quality images. We also leverage the recently introduced large-scale ImageNet open-set splits [44]. Lastly, we conduct ablations to understand the performance impact of our approach.

4.1 Details

Architectures. We use state-of-the-art pre-trained models on ILSVRC2012 [36] with standard 224×224 image crops - Meta's Hiera-H [37] and ConvNeXtV2-H [48]. Meta fine-tuned these models with no additional/external data. Note, we do not perform additional training or fine-tuning.

Datasets. For knowns, we utilize the large-scale ImageNet [9] subset ILSVRC2012 [36]. This dataset provides an advantage due to the extensive variety of available pre-trained architectures, enabling us to compare with more diverse deep networks. Our PostMax model is trained on the ILSVRC2012 *train* split, which contains $\sim 1.28M$ images and $1K$ classes (PostMax training requires no additional negative/unknown data). For validation, we employ ImageNetV2 [34], which contains $10K$ images. For testing, we split ILSVRC2012 *val* into five equal splits of $10K$ images each.

For unknowns, we use iNaturalist [17], NINCO [2], OpenImage-O [45], Places [56], SUN [49], Textures [8], and ImageNet-21K-P Open-Set splits [44]. The vision

Table 2: Common Metrics. The mean AUROC (↑) and FPR95 (↓) of all methods. To compute, we test each method on five different ILSVRC2012 *val* [36] splits (each $10K$ images) as knowns and specified unknowns. Standard deviation is omitted as it is < 0.01 for all methods/unknowns. OSR is performed on extractions from two state-of-the-art pre-trained architectures on ILSVRC2012-1K - (1) Transformer Hiera-H [37] (2) CNN ConvNeXtV2-H [48]. The best scores are in **bold**.

Unknowns (# images)	**AUROC ↑ / FPR95 ↓**									
	Hiera-H [37]					**ConvNeXtV2-H** [48]				
	PostMax (Ours)	**MSP** [16, 44]	**MaxLogit** [15, 44]	**NNGuide** [32]	**SCALE** [50]	**PostMax** (Ours)	**MSP** [16, 44]	**MaxLogit** [15, 44]	**NNGuide** [32]	**SCALE** [50]
iNaturalist [17] ($10K$)	**.96** / **.21**	.92 / .37	.88 / .36	.90 / .50	.81 / .51	**.95** / **.26**	.91 / .42	.90 / .37	.86 / .58	.88 / .44
NINCO [2] ($\sim 5.8K$)	**.87** / **.58**	.83 / .62	.79 / .61	.69 / .78	.73 / .67	**.86** / **.56**	.83 / .64	.82 / .60	.71 / .81	.80 / .62
OpenImage-O [45] ($\sim 17.6K$)	**.95** / **.25**	.88 / .49	.80 / .49	.88 / .45	.72 / .61	**.94** / **.29**	.88 / .49	.87 / .44	.81 / .70	.85 / .47
Places [56] ($10K$)	**.89** / **.49**	.83 / .60	.75 / .61	.85 / .55	.67 / .71	**.89** / **.50**	.84 / .62	.79 / .60	.82 / .64	.76 / .64
SUN [49] ($10K$)	**.90** / **.46**	.84 / .57	.78 / .57	.85 / .56	.71 / .66	**.89** / **.47**	.84 / .59	.81 / .57	.83 / .60	.77 / .61
Textures [8] ($\sim 5.1K$)	**.93** / **.32**	.88 / .47	.86 / .42	.74 / .86	.83 / .46	**.91** / **.34**	.87 / .49	.88 / .40	.73 / .85	.88 / .38
21K-P *Easy* [44] ($50K$)	**.86** / **.50**	.79 / .64	.72 / .64	.80 / .68	.65 / .73	**.85** / **.51**	.79 / .65	.75 / .64	.78 / .72	.72 / .68

community [15, 16, 32, 44, 45, 50, 52, 54] has utilized these datasets as they present unique challenges not seen in small-scale settings. For validation, to predict an operational threshold, we utilize 20% of ImageNet-21K-P *Hard* ($9.8K$ images) as our surrogate unknowns. For testing, NINCO, OpenImage-O, Textures, and ImageNet-21K-P *Easy* are entirely used, while others sample $10K$ images from a manual selection of classes. Further descriptions of all datasets and splits are provided in the supplemental.

Evaluation Metrics. To align with existing literature, we utilized common metrics in Tab. 2, namely, Area Under the Receiver Operating Curve (AUROC) and False Positive Rate at a fixed True Positive Rate of 95% (FPR95). Additional metrics like Area Under Open-Set Classification Rate (AUOSCR) [6] and Open Area Under ROC Curve (OpenAUC) [46] are provided in the supplemental. Alongside these, we introduce Operational Open-Set Accuracy (OOSA) in Tab. 1, where we predict an operational threshold on a validation and a surrogate set and deploy it at test time. This measure, described in detail in Sec. 3.3, serves as a more realistic alternative to existing measures. When evaluating OSR methods in real-world scenarios, other measures are insufficient as they do not provide an operating point for deployment and are usually only evaluated with a test set.

4.2 Results

Our main results are presented in Tab. 1 and 2, where we report performance on each unknown dataset mentioned in Sec. 4.1. We compare our approach with two commonly seen methods, Maximum Softmax Probability (MSP) [16, 44] and Maximum Logit (MaxLogit) [15, 44], along with the recently introduced Nearest Neighbor Guidance (NNGuide) [32] and SCALE [50]. We omit comparisons with older techniques such as OpenMax [1] and EVM [35] as they are much worse. According to the OpenOOD [52, 54] Benchmark leaderboard[5] for ImageNet-1K,

[5] `https://zjysteven.github.io/OpenOOD`

Table 3: Different Validation - Knowns & Unknowns Surrogate. The Operational Open-Set Accuracy (OOSA) (↑) of all methods on a different validation set. To compute, we validate methods on ILSVRC2012 *val* [36] ($50K$ images) as knowns and 21K-P *Easy* [44] ($50K$ images) as unknowns surrogate and predict an operational threshold. Then, we deploy each method's threshold and test on ImageNetV2 [34] ($10K$ images) and specified unknowns. OSR is performed again on extractions from Hiera-H [37] and ConvNeXtV2-H [48]. The best scores are in **bold**.

Unknowns (# images)	**OOSA ↑**									
	Hiera-H [37]					**ConvNeXtV2-H** [48]				
	PostMax (Ours)	**MSP** [16, 44]	**MaxLogit** [15, 44]	**NNGuide** [32]	**SCALE** [50]	**PostMax** (Ours)	**MSP** [16, 44]	**MaxLogit** [15, 44]	**NNGuide** [32]	**SCALE** [50]
iNaturalist [17] ($10K$)	**0.863**	0.808	0.780	0.808	0.716	**0.842**	0.806	0.794	0.759	0.763
NINCO [2] ($\sim 5.8K$)	**0.733**	0.730	0.704	0.655	0.654	**0.759**	0.733	0.726	0.643	0.695
OpenImage-O [45] ($\sim 17.6K$)	**0.885**	0.810	0.736	0.809	0.649	**0.865**	0.821	0.786	0.746	0.757
Places [56] ($10K$)	**0.801**	0.748	0.684	0.766	0.611	**0.795**	0.753	0.708	0.731	0.664
SUN [49] ($10K$)	**0.812**	0.752	0.702	0.771	0.635	**0.802**	0.757	0.722	0.738	0.679
Textures [8] ($\sim 5.1K$)	**0.813**	0.746	0.739	0.691	0.706	**0.784**	0.750	0.757	0.657	0.732
21K-P *Hard* [44] ($49K$)	0.634	**0.673**	0.583	0.535	0.515	0.647	**0.672**	0.593	0.559	0.566

NNGuide stands out as the leader in the OOD space, closely followed by SCALE. We caution against relying solely on OOD methods and metrics, as they may not accurately reflect OSR performance. Note, some OSA and ROC curves are found in the supplemental.

When utilizing both architectures (Hiera-H & ConvNeXtV2-H) as feature extractors, PostMax outperforms all other methods in every measure (OOSA, AUROC, FPR95) on each unknown dataset. PostMax demonstrates superior performance, emphasizing robustness and generalizability. Effectively, PostMax achieves an excellent trade-off between classification accuracy and unknown rejection. Overall, the results suggest that a system operator can seamlessly integrate PostMax, enhance the performance beyond good closed-set classifiers, and confidently select the operational threshold that best suits its operation.

Statistical Testing. To underscore the significance of performance differences between algorithms in our experiments, we prioritized the assessment of statistical significance. As a result, we performed paired, two-tailed t-tests, with Bonferroni corrections, to compare PostMax with the four methods (MSP, MaxLogit, NNGuide, and SCALE) using both architectures (Hiera-H and ConvNeXtV2-H) across 5 fold runs for each metric (OOSA, AUROC, and FPR95). On Hiera-H, PostMax **(1)** OOSA scores improvements range from **2.5**% to **12.4**% with corresponding P-values all less than $\mathbf{1.0E^{-7}}$, **(2)** AUROC scores improvements range from **5.7**% to **17.9**% with corresponding P-values all less than $\mathbf{1.0E^{-8}}$, and **(3)** FPR95 scores improvements range from **12.7**% to **22.5**% with corresponding P-values all less than $\mathbf{1.0E^{-8}}$. Similarly, on ConvNeXtV2-H, PostMax **(1)** OOSA scores improvements range from **2.3**% to **10.4**%, **(2)** AUROC scores improvements range from **4.7**% to **10.6**%, and **(3)** FPR95 scores improvements range from **9.6**% to **28.1**%, where all corresponding P-values are less than $\mathbf{1.0E^{-8}}$ for all tests on all three metrics. In summary, the t-tests conducted for PostMax, considering the combined scores of all metrics across all datasets and architectures, revealed very statistically significant differences compared to every other method.

Table 4: Different Testing - Prior Metrics. The AUROC (↑) and FPR95 (↓) of all methods on different test sets. To compute, we test each method on ImageNetV2 [34] ($10K$ images) as knowns and specified unknowns. OSR is performed again on extractions from Hiera-H [37] and ConvNeXtV2-H [48]. The best scores are in **bold**.

Unknowns (# images)	AUROC ↑ / FPR95 ↓									
	Hiera-H [37]					ConvNeXtV2-H [48]				
	PostMax (Ours)	MSP [16, 44]	MaxLogit [15, 44]	NNGuide [32]	SCALE [50]	PostMax (Ours)	MSP [16, 44]	MaxLogit [15, 44]	NNGuide [32]	SCALE [50]
iNaturalist [17] ($10K$)	**.96** / **.16**	.92 / .37	.88 / .37	.91 / .40	.80 / .55	**.95** / **.23**	.91 / .41	.90 / .37	.87 / .53	.87 / .44
NINCO [2] ($\sim 5.8K$)	**.88** / **.52**	.83 / .61	.77 / .62	.71 / .72	.70 / .71	**.87** / **.54**	.83 / .63	.81 / .60	.73 / .78	.79 / .63
OpenImage-O [45] ($\sim 17.6K$)	**.96** / **.20**	.87 / .48	.79 / .49	.90 / .38	.70 / .64	**.94** / **.27**	.88 / .49	.86 / .43	.82 / .65	.84 / .48
Places [56] ($10K$)	**.90** / **.44**	.83 / .60	.74 / .62	.87 / .48	.65 / .74	**.89** / **.48**	.83 / .61	.78 / .60	.84 / .59	.74 / .64
SUN [49] ($10K$)	**.91** / **.41**	.84 / .57	.77 / .58	.87 / .49	.69 / .70	**.90** / **.44**	.84 / .59	.80 / .57	.84 / .55	.76 / .61
Textures [8] ($\sim 5.1K$)	**.94** / **.27**	.88 / .47	.85 / .42	.77 / .79	.81 / .49	**.92** / **.32**	.87 / .49	.88 / .39	.75 / .82	.87 / .39
21K-P *Hard* [44] ($49K$)	**.77** / **.68**	.73 / .73	.68 / .73	.68 / .76	.62 / .82	**.76** / **.72**	.74 / .75	.70 / .74	.67 / .82	.67 / .79

4.3 Ablation Study

In Tab. 3 and 4, we investigate the effects of varying the validation datasets (knowns and unknowns surrogate). We interchanged ImageNetV2 with ILSVRC2012 *val* for knowns and 21K-P *Hard* with 21K-P *Easy* for the unknowns surrogate. Unlike in Tab. 1, this involves validating on a larger number of knowns and unknowns, with an easier unknowns surrogate. Performance remained consistent regardless of the datasets.

Also, we investigated the impact of other architectures as extractors and interchanged pre-trained models (ILSVRC2012-1K only w/ 224 crops) with Meta's ViT-H [14] and ConvNeXt-L [24]. Results showcasing PostMax's robustness are in the supplemental.

Lastly, we explored the impact of varying the application-dependent α from Eq. (2) and the normalization (L_2) within PostMax. Details and results are in the supplemental.

5 Discussion

This paper introduces the new evaluation metric Open-Set Accuracy (OSA) (2); we framed the evaluation of an open-set problem as predicting a threshold and then evaluating open-set accuracy at that threshold. This supports an engineer selecting an algorithm and a threshold to be applied in some Operational setting.

In real-world scenarios, unknown samples can be drawn from an unlimited variety, unlike known samples, which are assumed to be related to the training data. Thus, the evaluation must consider the computation of the predictive threshold on validation data and a separate evaluation on test data. The quality of the predicted threshold will depend on many factors with the two most influential being (1) how representative are the surrogate samples of the real unknowns, and (2) the relative proportion of knowns and unknowns. In Sec. 4.3 and our supplemental, we explore some variations of those factors.

Maybe not surprisingly, if we train with "harder" unknowns, we generally see better performance. As shown in Tab. 1 and 2, the novel PostMax algorithm is better on unknowns across two networks (other networks in the supplemental).

However, as shown in Sec. 4.3, when 21K-P *Easy* is used as our unknowns surrogate, PostMax does not exhibit an advantage on 21K-P *Hard*. In general, 21K-P *Hard* poses the most significant challenge for all algorithms. Notably, the sets close visual and semantic similarity to known samples suggests it represents a fine-grained categorization challenge rather than a typical open-set scenario.

One can adjust the importance of knowns *vs*. unknowns either by explicitly weighting them or implicitly through their relative percentages in the data. Although performance degrades when there is a significant disparity in their importance during threshold prediction stages, PostMax consistently outperforms and demonstrates greater stability across all thresholds. Refer to the supplemental for further details.

Finally, to account for inherent random variations in predictions and tests, we developed training/validation/test splits, which are available on the GitHub repo. This protocol enables rigorous experimentation to assess the statistical significance of our evaluation. While the paper demonstrates statistically significant improvements across all results, some ablations in the supplemental do not.

6 Conclusion

This paper contributes to OSR research by shifting focus from common metrics on small-scale experiments to more realistic scenarios and datasets, thereby fostering more practically relevant research. We revisited and revised the criteria for effective open-set metrics as suggested by Wang *et al.* [46], making them more applicable to real-world scenarios. Our proposed Operational Open-Set Accuracy (OOSA), which incorporates known and unknown samples, offers a comprehensive metric and facilitates the selection of the most effective method and operating threshold for system operators.

Our investigation of deep feature magnitudes of unknown samples from various ILSVRC2012 pre-trained networks yielded surprising findings. Contrary to smaller-scale studies [10, 31], we observed that these magnitudes are often larger for unknown than for known samples. This observation was consistent across multiple networks and types of unknown samples. Thus, we introduced our novel PostMax algorithm, refining normalized logit values from closed-set architectures into real probabilities of class inclusion using a Generalized Pareto Distribution (GPD). This distribution method offers advantages over previous extreme-value-based techniques, including negating the need to select a tail size and minimizing the influence of outliers. Also, it eliminates the need to define negative training samples as representatives of unknown classes, a common approach in much of the OSR literature that has not yielded a sufficient solution.

While a good closed-set classifier is an important place to start [44], our novel **PostMax** algorithm shows that one can do better than "all you need" and that for operational usage, one needs to predict a threshold to use. With this paper, we empower engineers to operationalize their usage of OSR.

References

1. Bendale, A., Boult, T.E.: Towards open set deep networks. In: Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2016)
2. Bitterwolf, J., Mueller, M., Hein, M.: In or out? fixing ImageNet out-of-distribution detection evaluation. In: ICLR Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models (2023)
3. Bodesheim, P., Freytag, A., Rodner, E., Denzler, J.: Local novelty detection in multi-class recognition problems. In: Winter Conference on Applications of Computer Vision (WACV) (2015)
4. Boult, T.E., Cruz, S., Dhamija, A.R., Günther, M., Henrydoss, J., Scheirer, W.J.: Learning and the unknown: Surveying steps toward open world recognition. In: AAAI Conference on Artificial Intelligence. vol. 33, pp. 9801–9807 (2019)
5. Cevikalp, H., Uzun, B., Salk, Y., Saribas, H., Köpüklü, O.: From anomaly detection to open set recognition: Bridging the gap. Pattern Recognition **138**, 109385 (2023)
6. Chen, G., Peng, P., Wang, X., Tian, Y.: Adversarial reciprocal points learning for open set recognition. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **44**(11), 8065–8081 (2021)
7. Chen, G., Qiao, L., Shi, Y., Peng, P., Li, J., Huang, T., Pu, S., Tian, Y.: Learning open set network with discriminative reciprocal points. In: European Conference on Computer Vision (ECCV). Springer (2020)
8. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255. IEEE (2009)
10. Dhamija, A.R., Günther, M., Boult, T.: Reducing network agnostophobia. Advances in Neural Information Processing Systems (NeurIPS) **31** (2018)
11. Ge, Z., Demyanov, S., Garnavi, R.: Generative OpenMax for multi-class open set classification. In: British Machine Vision Conference (BMVC) (2017)
12. Gumbel, E.J.: Statistical Theory of Extreme Values and Some Practical Applications: A Series of Lectures, vol. 33. US Government Printing Office (1954)
13. Günther, M., Dhamija, A.R., Boult, T.E.: Watchlist adaptation: Protecting the innocent. In: International Conference of the Biometrics Special Interest Group (BIOSIG) (2020)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (2022)
15. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: International Conference on Machine Learning (ICML). PMLR (2022)
16. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. International Conference on Learning Representations (ICLR) (2017)
17. van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The INaturalist species classification and detection dataset. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
18. Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. Advances in Neural Information Processing Systems (NeurIPS) **34**, 677–689 (2021)

19. Huang, R., Li, Y.: Mos: Towards scaling out-of-distribution detection for large semantic space. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8710–8719 (2021)
20. Kong, S., Ramanan, D.: Opengan: Open-set recognition via open data generation. In: International Conference on Computer Vision (ICCV). pp. 813–822 (2021)
21. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)
22. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
23. Liu, X., Lochman, Y., Zach, C.: Gen: Pushing the limits of softmax-based out-of-distribution detection. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23946–23955 (2023)
24. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11976–11986 (2022)
25. Lu, J., Xu, Y., Li, H., Cheng, Z., Niu, Y.: Pmal: Open set recognition via robust prototype mining. In: AAAI Conference on Artificial Intelligence. vol. 36:2, pp. 1872–1880 (2022)
26. Lyu, Z., Gutierrez, N.B., Beksi, W.J.: Metamax: Improved open-set deep neural networks via weibull calibration. In: Winter Conference on Applications of Computer Vision Workshops (WACVW) (2023)
27. Moon, W., Park, J., Seong, H.S., Cho, C.H., Heo, J.P.: Difficulty-aware simulator for open set recognition. In: European Conference on Computer Vision (ECCV). pp. 365–381. Springer (2022)
28. Neal, L., Olson, M., Fern, X., Wong, W.K., Li, F.: Open set learning with counterfactual images. In: European Conference on Computer Vision (ECCV) (2018)
29. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011)
30. Palechor, A., Bhoumik, A., Günther, M.: Large-scale open-set classification protocols for imagenet. In: Winter Conference on Applications of Computer Vision (WACV). pp. 42–51. CVF/IEEE (2023)
31. Park, J., Chai, J.C.L., Yoon, J., Teoh, A.B.J.: Understanding the feature norm for out-of-distribution detection. In: International Conference on Computer Vision (ICCV). pp. 1557–1567 (2023)
32. Park, J., Jung, Y.G., Teoh, A.B.J.: Nearest neighbor guidance for out-of-distribution detection. In: International Conference on Computer Vision (ICCV). pp. 1686–1695 (2023)
33. Pickands III, J.: Statistical inference using extreme order statistics. The Annals of Statistics **3**(1), 119 – 131 (1975)
34. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do Imagenet classifiers generalize to Imagenet? In: International Conference on Machine Learning (ICML). pp. 5389–5400. PMLR (2019)
35. Rudd, E.M., Jain, L.P., Scheirer, W.J., Boult, T.E.: The extreme value machine. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2017)
36. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015)

37. Ryali, C., Hu, Y.T., Bolya, D., Wei, C., Fan, H., Huang, P.Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., Malik, J., Li, Y., Feichtenhofer, C.: Hiera: A hierarchical vision transformer without the bells-and-whistles. International Conference on Machine Learning (ICML) (2023)
38. Scheirer, W.J., Jain, L.P., Boult, T.E.: Probability models for open set recognition. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **36**(11), 2317–2324 (2014)
39. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boult, T.E.: Towards open set recognition. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **35**(7) (2013)
40. Shu, L., Xu, H., Liu, B.: DOC: Deep open classification of text documents. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics (2017)
41. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. Advances in Neural Information Processing Systems (NeurIPS) **34**, 144–157 (2021)
42. Vareto, R.H., Linghu, Y., Boult, T.E., Schwartz, W.R., Günther, M.: Open-set face recognition with maximal entropy and Objectosphere loss. Image and Vision Computing (IMAVIS) **141** (2024)
43. Vareto, R.H., Schwartz, W.R., Günther, M.: Toward open-set face recognition with neural ensemble, maximal entropy loss and feature augmentation. In: Conference on Graphics, Patterns and Images (SIBGRAPI) (2023)
44. Vaze, S., Han, K., Vedaldi, A., Zissermann, A.: Open-set recognition: A good closed-set classifier is all you need? In: International Conference on Learning Representations (ICLR) (2022)
45. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4921–4930 (2022)
46. Wang, Z., Xu, Q., Yang, Z., He, Y., Cao, X., Huang, Q.: OpenAUC: Towards auc-oriented open-set recognition. Advances in Neural Information Processing Systems (NeurIPS) **35**, 25033–25045 (2022)
47. Wilson, S., Fischer, T., Dayoub, F., Miller, D., Sünderhauf, N.: Safe: Sensitivity-aware features for out-of-distribution object detection. In: International Conference on Computer Vision (ICCV). pp. 23565–23576 (2023)
48. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16133–16142 (2023)
49. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3485–3492. IEEE (2010)
50. Xu, K., Chen, R., Franchi, G., Yao, A.: Scaling for training time and post-hoc out-of-distribution detection enhancement. In: International Conference on Learning Representations (ICLR) (2024)
51. Yang, H.M., Zhang, X.Y., Yin, F., Yang, Q., Liu, C.L.: Convolutional prototype network for open set recognition. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2020)
52. Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., et al.: OpenOOD: Benchmarking generalized out-of-distribution detection. Advances in Neural Information Processing Systems (NeurIPS) **35**, 32598–32611 (2022)

53. Yoshihashi, R., Shao, W., Kawakami, R., You, S., Iida, M., Naemura, T.: Classification-reconstruction learning for open-set recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
54. Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Zhou, K., Zhang, W., Li, Y., Liu, Z., Chen, Y., Li, H.: OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. In: NeurIPS Workshop on Distribution Shifts: New Frontiers with Foundation Models (2023)
55. Zhang, X., Cheng, X., Zhang, D., Bonnington, P., Ge, Z.: Learning network architecture for open-set recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36(3), pp. 3362–3370 (2022)
56. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **40**(6), 1452–1464 (2017)
57. Zhou, D.W., Ye, H.J., Zhan, D.C.: Learning placeholders for open-set recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)