COSTARR: Consolidated Open Set Technique with Attenuation for Robust Recognition

Ryan Rabinowitz[1*] Steve Cruz[2*] Walter Scheirer[2] Terrance E. Boult[1]

[1]University of Colorado Colorado Springs, USA
[2]University of Notre Dame, USA
{rrabinow,tboult}@uccs.edu {stevecruz,walter.scheirer}@nd.edu

Abstract

Handling novelty remains a key challenge in visual recognition systems. Existing open-set recognition (OSR) methods rely on the familiarity hypothesis, detecting novelty by the absence of familiar features. We propose a novel attenuation hypothesis: small weights learned during training attenuate features and serve a dual role—differentiating known classes while discarding information useful for distinguishing known from unknown classes. To leverage this overlooked information, we present COSTARR, a novel approach that combines both the requirement of familiar features and the lack of unfamiliar ones. We provide a probabilistic interpretation of the COSTARR score, linking it to the likelihood of correct classification and belonging in a known class. To determine the individual contributions of the pre- and post-attenuated features to COSTARR's performance, we conduct ablation studies that show both pre-attenuated deep features and the underutilized post-attenuated Hadamard product features are essential for improving OSR. Also, we evaluate COSTARR in a large-scale setting using ImageNet2012-1K as known data and NINCO, iNaturalist, OpenImage-O, and other datasets as unknowns, across multiple modern pretrained architectures (ViTs, ConvNeXts, and ResNet). The experiments demonstrate that COSTARR generalizes effectively across various architectures and significantly outperforms prior state-of-the-art methods by incorporating previously discarded attenuation information, advancing open-set recognition capabilities.

1. Introduction

Dealing with unknown inputs in a recognition system is an important and widely recognized problem, which can be formalized as Open-Set Recognition (OSR) [35]. The recently introduced familiarity hypothesis [9] frames existing DNN open-set systems as "detecting the absence of familiar [deep] learned features rather than the presence of novelty".

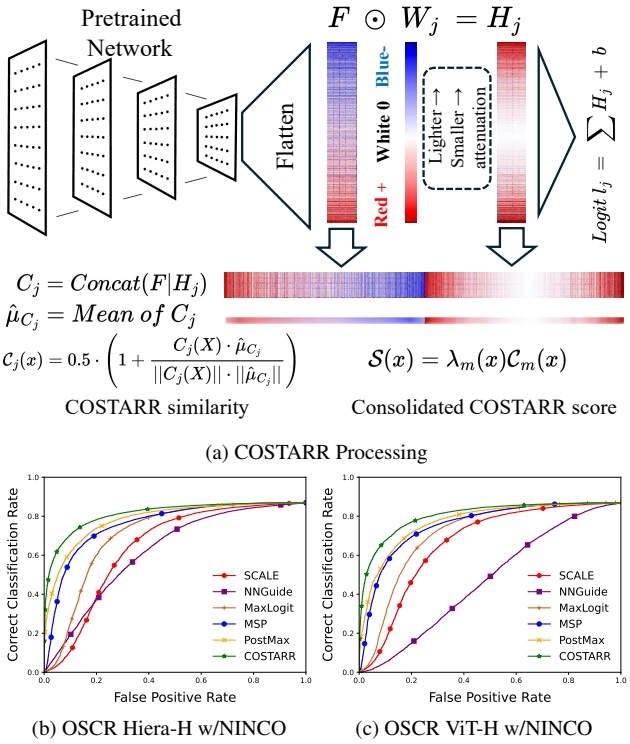

Figure 1. OVERVIEW. (a) The Hadamard product H of the feature vector F and the weight vector W for class j is combined with the per-class mean comparison $\hat{\mu}_{C_j}$ to compute the COSTARR similarity score $\mathcal{C}_j(x)$. This similarity is then scaled by the normalized logit λ_m to produce the final COSTARR score $\mathcal{S}(x)$. See Sec. 3 for details. Our attenuation hypothesis motivates the need for both components. (b) and (c) demonstrate that COSTARR outperforms current state-of-the-art algorithms, including PostMax, in terms of Open Set Classification Rate (OSCR), highlighting how our novel approach effectively leverages consolidated information to advance the state of the art.

*Equal contribution.

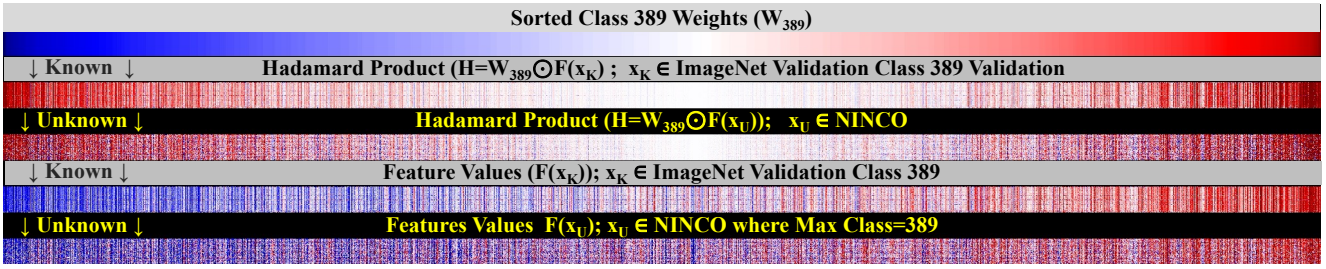

Figure 2. INTUITION BEHIND THE ATTENUATION HYPOTHESIS. The logit for class j is computed as the dot product between its weight vector W_j and the input features F, but features associated with low weights are attenuated and thus ignored by this class's projection. The top color bar shows the weights (W) for ImageNet2012 class 389 from Hiera-H's classification layer, sorted from low to high (left to right); the white region indicates where weights approach zero. This sorting index is used to order all subsequent color bars. The second bar displays the Hadamard product (H) for a known input x_k, and the third shows the same for an unknown input x_u, for which class 389 produced the max logit. Each column represents a single feature dimension, for 40 different images. The fourth and fifth bars represent the feature vectors from the same 40 images and the most confident unknown samples from NINCO [3], respectively. The consistent low-saturation regions in the weight and Hadamard bars demonstrate that many feature dimensions are attenuated before contributing to classification, regardless of which features are present. Notably, while the feature vectors of NINCO samples are visually dissimilar to those from the validation set, attenuation in the Hadamard and logit computations diminishes their effect. The average logit for the known samples is 8.959, while for the unknowns it is 8.143-making them difficult to distinguish by logit alone. In contrast, the average COSTARR score for samples from known class 389 is 0.647, compared to 0.369 for the unknowns, providing a clearer separation. (Figure best viewed in color.)

Our novel COSTARR approach (Fig. 1) stems from us asking *why do systems implicitly focus on only familiar features* and presents our novel *attenuation hypothesis* to explain it.

Let $F(x)$ represent the pre-attenuation deep features extracted from input x. Let W_j be the weight vector for class j. The Hadamard product $H_j = F(x) \odot W_j$ represents the pointwise multiplication that occurs in the classification layer and is a post-attenuated feature. These two sides of the attenuation, pre- and post- attenuation vectors, form the foundation of our hypothesis detailed in Sec. 3.

In OSR, there exists an inherent trade-off between accurately classifying knowns and effectively rejecting unknowns, with each objective benefiting from either pre– or post–attenuation representations. We empirically demonstrate that by not ignoring either features, i.e., by incorporating both pre–attenuation and post–attenuation features into our overall confidence measure, we improve a DNN's ability to recognize known classes among unknown samples. Ablations show that we need to consolidate information from both sides of the attenuation to provide robust recognition.

We formally introduce the attenuation hypothesis in Sec. 3.1, but we use Fig. 2 here to build intuition and help readers visualize the underlying concept. The figure illustrates how weights and features contribute to the computation of the maximum classification logit. In many dimensions, the corresponding weights attenuate the associated features, resulting in a marginal contribution to the final logit. This attenuation can allow an unknown input to have high confidence for a known class. Due to the design of fully connected layers, such dimensions are uninformative for the current class but likely play important roles in distinguishing other classes. However, unknown inputs often exhibit unexpected activation patterns in these attenuated dimensions. By incorporating both F and H, our novel COSTARR score effectively captures these discrepancies. COSTARR is applicable to any pre-trained network containing a classification layer and incurs minimal computational and storage overhead.

The contributions of this paper are as follows:

- A novel *attenuation hypothesis* explaining the **necessary** roles of pre- and post-attenuated features in OSR systems.
- COSTARR: an efficient, state-of-the-art OSR post-processing algorithm that combines pre-attenuation features, Hadamard product-based post-attenuation features, and logits for robust recognition. Code is publicly available.[1]
- Ablation studies confirm the necessity of both similarity components proposed in the attenuation hypothesis.
- Sec 3.3 formalizes COSTARR and provides an interpretation of it as a probability estimate that an input is from a known class and the selected class is correct.
- Experiments on ImageNet-1K across five leading architectures and multiple unknown datasets show statistically significant improvements on several OSR metrics.

2. Related Work

Numerous works [1, 4, 5, 8, 11, 18, 19, 25–28, 36, 46, 48, 50] have trained Deep Neural Networks (DNNs) to mitigate the problems posed by unknowns, however, they are categorically different from ours. We focus exclusively on improving open-set recognition (OSR) after the training pro-

[1] https://github.com/Vastlab/COSTARR

cess. Accordingly, we consider other work that focuses on improving OSR for pre-trained networks.

2.1. Open-Set Post-processors

OpenMax [2], the foundational work for adapting DNNs to open-set, sought to overcome the overconfidence of DNNs trained with softmax activations. Using mean activation vectors from the penultimate layer, OpenMax fit Weibull distributions for each class using Extreme Value Theory (EVT). At test time, OpenMax converted sample distance from the top predicted classes to probabilities and, using a softmax-like operation, revised these probabilities into a vector summing to 1 with an explicit probability of unknown. Another open-set DNN adaptation, the Extreme Value Machine (EVM) [32], fits Weibull models between crucial in-class samples (extreme vectors) to negative samples from other classes. Using the distance of a sample to extreme vectors, the EVM generated probabilities of which class the sample belongs to. By selecting extreme vectors, the work aimed to minimize open space risk. Both were interesting but experimentally they failed to produce algorithms much better than just using a good close-set classifier [40].

Multiple works have proposed using either Maximum Logit (MaxLogit) [16, 40] or Maximum Softmax Probability (MSP) [15, 40] as a baseline for OSR. These methods are intuitive and widely applicable, as they simply threshold on existing network outputs. However, PostMax [7], the recent state-of-the-art for large-scale OSR, demonstrated that there is still much room for improvement.

The current state of the art is PostMax [7] which uses EVT and feature magnitudes to normalize logit "confidence". Its predictions are transformed into a probability space, similar to SoftMax, except that there is only one probability per sample rather than a probability for each class. While their normalizing with feature magnitudes was effective, it relies on the observation that unknowns have a higher feature magnitude than knowns, which contradicts observations of others [8, 40] and need not be true. In contrast, we make no such reliance and compare the similarity between known class centers and test samples to capture the information that is generally ignored. We note that PostMax did not offer a compelling reason for its improvement, just an observation that feature magnitudes were different between known and unknowns. Based on our analysis we hypothesize that PostMax's improvement is because that, consistent with the *attenuation hypothesis*, they exploit raw feature magnitudes thus their normalization uses some of the ignored information compared, albeit in a weak way.

2.2. Out-of-Distribution Post-processors

Distinct from OSR, Out-of-Distribution (OOD) only focuses on the detection of samples as in-distribution or OOD. Various approaches have been proposed [10, 21–23, 29, 30, 37, 38, 41, 44] and adopted similar large-scale evaluations, departing from small-scale experimentation in early OOD works [17, 20] which relied on expensive strategies like reprocessing the inputs. Given the abundance of methods, OpenOOD [45, 47] curated a large-scale benchmark to provide an accurate, standardized, and unified evaluation of OOD detection. We compare our method to others using the large-scale ImageNet-1K benchmark. Specifically, following prior work [7], we select the best-performing post-processors: SCALE [44] and NNGuide [29]. SCALE rescales intermediate features, which affects the final output confidence of a network. This is different from ReAct [37], which pruned features outside of a given threshold. NNGuide [29], drawing off the K-Nearest Neighbor (KNN) ideology, retains a "bank set" of features from the training set. Using the bank set and test-time samples, they form a "guided" score, which reduces over-confidence on far-OOD samples. Also, we include results provided by the very recent COMBOOD [30], which combines nearest-neighbor and Mahalanobis distances to form an OOD score. However, due to difficulties integrating this recent algorithm with our evaluations, we only compare with results from their publication, which did not utilize open-set metrics. Accordingly, we present this comparison in the supplemental.

2.3. Metrics

In addition to introducing PostMax, Cruz et al. [7] proposed the Operational Open-Set Accuracy (OOSA), as a new metric for evaluating open-set recognition systems. The work was motivated by giving an example of an engineer preparing to deploy an OSR system and the need to determine the overall accuracy of methods before deployment. OOSA highlights how proper training / validation / testing splits can be used to model real-world settings. We chose to utilize OOSA as our main evaluation measure as it allows for the prediction of an operational threshold but also recognize other common metrics in the literature and include them for completeness: the Open-Set Classification Rate [8] curve (in our secondary evaluation) and Area Under the Receiver Operating Characteristic curve (in the supplemental material).

3. Approach

Improving algorithm performance is important, but addressing causality and explaining **why** is even more important for science. Our approach had two main hypotheses: 1) the novel *attenuation hypothesis* which leads us to combine per-class models using the class mean of both pre- and post-attenuation features, and 2) to maximize usage of training information and provide a consistent probabilistic interpretation, we scaled it using normalized logits to provide the full consolidated model. We discuss each of these elements.

3.1. Attenuation Hypothesis

For class j on input x, let Hadamard product $H_j = F(x) \odot W_j$ be the point-wise product of features $F(x)$ with weight vector W_j. The hypothesis states weights that attenuate select features in W_j are important for two competing reasons:

1. To maintain known class accuracy, features used in one class often need to be ignored by others. For class j, training yields many small weights that reduce the impact on H_j of some features used in classes $k \neq j$. Similarity to the class H_j mean can measure known class similarity.
2. For unknowns, large feature magnitudes can overwhelm small weight attenuation, causing large logit or softmax values leading to misclassification of unknowns as known. Unknowns can be detected using measures that consider feature magnitude without attenuating weights, e.g., using feature vector $F(x)$ similarity to class j mean so that the added information is not attenuated.

The Attenuation Hypothesis: *A robust open-set recognition system should integrate both the pre-attenuation features (F) and the post-attenuation features (H) to optimize the dual objectives of maximizing known class accuracy while minimizing false acceptance of unknown inputs.*

We argue that this hypothesis applies to any network that uses a linear classifier, regardless of the linearity of the input or loss. We explicitly evaluate the hypothesis in Sec. 4 across five major networks, including convolutional-based and transformer-based models. In addition, the hypothesis also helps explain prior work. The "familiarity hypothesis," [9] suggests that traditional algorithms relying on logits or softmax-based analysis tend to reject unknowns only when they lack familiar features, but fail when encountering inputs with unexpected ones. We contend this is precisely because such methods rely solely on information in H. Other systems such as OpenMax [2], Extreme Value Machine (EVM) [32], and many prototype-based few-shot open-set learning methods, use only the deep features F, and therefore tend to suffer in accuracy on known classes. To our knowledge only the most recent approach–PostMax [7], which was the state of the art for large-scale OSR– attempts to combine information from both F and H, albeit in a weak manner and without a clear explanation.

The Attenuation Hypothesis is a key contribution of this paper. While there may be better ways to leverage the information than our COSTARR approach, we are the first to explain why robust OSR solutions must consider both pre– and post–attenuation features.

Although not explicitly part of the hypothesis, we also note that the deep network training learns the bias b as part of the final classifier. This term likely contains useful information that should be incorporated–e.g., through the use of softmax or logits. Below, we show how to use logits and provide ablations to analyze their importance.

To see this in action, consider the operation performed by the linear layer used to produce classification outputs in a DNN. As illustrated in Fig. 2, when computing the logit for a single class, the layer's weights attenuate corresponding features–weights approaching zero effectively suppress those features when the Hadamard product is applied. The resulting values are then summed to produce the logit, which is subsequently scaled to generate the softmax score. This works well for known inputs.

The supplement includes a frequency chart showing that many feature dimensions are commonly attenuated to have only a marginal effect. Analyzing class weights shows that for every feature, there exists at least one class where it is weighted highly and others where it is largely ignored.

When the Hadamard product values are summed, the attenuated features contribute little to the resulting logit or softmax score. Yet, the response of these features still encodes information about the sample, even if it is not directly useful for the closed-set prediction of the given class–it may be informative for other classes. Our hypothesis suggests that this discarded information is in fact valuable for distinguishing known from unknown samples, as unknowns tend to exhibit feature responses and Hadamard values that are less consistent with class mean.

3.2. Per-class Models

Many models have used per-class information, but none have incorporated both forms of per-class information that we consider. While OpenMax and EVM utilize per-class information–such as Mean Activation Vectors and Extreme Vectors–the approaches are weakened by ignoring the additional per-class information in the classification weights.

Recall that PostMax [7] operates by computing the Euclidean norm of feature vectors and dividing the maximum logit by this norm. PostMax does not incorporate class-centric information into its normalization factor; regardless of the predicted class, the same Euclidean norm is applied. While PostMax argues that the method works because unknowns tend to have larger magnitudes, we contend that this observation is just an accidental side effect of the attenuation hypothesis–and for some networks, this assumption does not hold. For example, several papers have observed unknowns with smaller magnitudes [8, 40], in which case PostMax's normalization would worsen performance. Moreover, examining figures such as Fig. 7f in [40] reveals that some classes exhibit very different feature magnitude ranges–some smaller than those of unknowns, others larger. There appears to be no consistent pattern of unknowns having either larger or smaller magnitudes, suggesting the need for a per-class model instead.

3.3. Formalizing COSTARR

Given the attenuation hypothesis, we aim to define a similarity measure based on the combination of the attenuation-

related quantities, H and F. Let N be some pretrained network that produced deep features $F(x)$ on input x and $l_j(x)$ be the logit for class j. Let K be the set of known classes and $\Omega(x)$ be the oracle operator that returns the correct class or -1 for an unknown class index, and $\Lambda(x)$ be the class with maximum logit from the classifier. Let T be training data for model building, with T_j the subset correctly classified as class j – ideally we would use separate data from the network training but for the pretrained networks herein we use the ImageNet training data and use the val set for knowns for testing. We define a globalized normalized logit (GNL) function to normalize our logits using a global min-max normalization based on training to ensure all normalized logits are in the [0,1] range:

$$GNL(l) = \max(0, \min(1, \frac{l - l_{tmin}}{l_{tmax} - l_{tmin}})) \quad (1)$$

where l_{tmin}, l_{tmax} are the minimum and maximum logits overall classes for correctly classified inputs $x \in T, \Lambda(x) = \Omega(x)$. Note we avoid per-instance normalization or the use of softmax since there is important information in the relative scale of logits, e.g. smaller versions of object will often have smaller logits. At inference time we let

$$\lambda_m(x) = \max_j \ GNL(l_j(x)) \quad (2)$$

be the maximum overall all classes GNL of the logit on input x, and note since this is a global linear transform, it does not change which class is maximum, so $\Lambda(x) = m$ is the associated class index.

We formalize COSTARR scoring through the following steps. First, our concatenated attenuation-related vector is:

$$C_j(x) = Concat(F(x), H_j) \quad (3)$$

where $H_j = F(x) \odot W_j$ is the Hadamard product, i.e., a pointwise product of weight vector W_j and feature vector $F(x)$ for class j on input x. Let $\hat{\mu}_{C_j}$ be the pre-computed mean of $C_j(x) \forall x \in T_j$. Given the co-staring role of both F and H in this we define our COSTARR similarity $\mathcal{C}_j(x)$ for class j and consolidated COSTARR score $\mathcal{S}(x)$ as:

$$\mathcal{C}_j(x) = 0.5 \cdot \left(1 + \frac{C_j(X) \cdot \hat{\mu}_{C_j}}{||C_j(X)|| \cdot ||\hat{\mu}_{C_j}||}\right) \quad (4)$$

$$\mathcal{S}(x) = \lambda_m(x)\mathcal{C}_m(x) \quad (5)$$

where the 0.5 in Eq. 4 rescales the similarity into [0,1]. In processing, we select the class $m = \Lambda(x)$ with maximum GNL logit λ_m and then use the precomputed mean for the consolidated attenuation-related vector of class m to compute the similarity. Thus we interpret the COSTARR similarity as an approximation of $P(x \in K | \Lambda(x) = m)$, i.e. the conditional probability the input is from a known class given we observed the input is from class m. If we assume that our GNL normalized max logit estimates the probability for that class, which is a good approximation of the correct class for known inputs, i.e. $\lambda_m(x) \approx P(\Lambda(x) = m)$, then we can interpret the COSTARR score as

$$\mathcal{S}(x) \approx P(\Lambda(x) = m) \cdot P((x \in K) | (\Lambda(x) = m)) \quad (6)$$

$$= P(\Lambda(x) = m) \cdot \frac{P((x \in K) \cap (\Lambda(x) = m))}{P(\Lambda(x) = m)} \quad (7)$$

$$= P((x \in K) \cap (\Lambda(x) = m)) \quad (8)$$

$$\approx P((x \in K) \cap (\Omega(x) = m)) \quad (9)$$

where Eq. 7 follows from 6 from the definition of conditional probability and 9 uses the assumption that on known inputs our classifier is sufficiently accurate. Thus the novel COSTARR score is an approximation of the probability that an input is both known and of the correct class. Hence, it can be thresholded to separate correctly classified knowns from unknown or misclassified ones.

Since we condition our algorithm–and the associated probabilities–on the max logit class, a limitation of this approach is that the probability approximation can fail on inputs from known classes when the pretrained network misclassifies them; that is, when $x \in K$ and $\Lambda(x) \neq \Omega(x)$. In such cases, the mean of the incorrect class is used, resulting in a lower COSTARR score. Consequently, the model is more likely to declare the input as unknown. From an operational perspective, this "limitation"–rejecting incorrectly classified known inputs–may actually be beneficial, as such misclassifications are likely to cause downstream errors.

This analysis also explains why we multiply by the normalized logit: in Eq. 7, it cancels out the denominator that arises in the expansion of the conditional probability. Without this term, variations in per-class accuracy (i.e., errors in $P(\Lambda(x) = m)$)) or poor calibration would directly affect the similarity score. By canceling this out, we ensure that the sorted order of scores remains reliable for thresholding, even if the scores themselves are not well calibrated.

While we define a particular similarity above, the attenuation hypothesis does not specify exactly how to compute similarity between the related terms. Although a global monotonic transformation of $\lambda(x)$ does not affect sorted order, per-instance normalization or changes to the similarity function can change it, which impacts open-set performance. In our ablations, we demonstrate the importance of considering both the attenuated H and the original features F, validating the attenuation hypothesis. We do not claim that 9 is the optimal combination. We expect "optimal" will depend on the network and the data used, as both impact F and H and the accuracy of logits on which we condition. Practitioners applying this approach may want to explore alternate combinations, normalizations, or optimizations, particularly after fine-tuning models for their specific tasks. Nonetheless, a key contribution of this paper is the attenuation hypothesis

itself. Those aiming to optimize performance should ensure both F and H are incorporated into their methods.

4. Experiments

The theory behind COSTARR was presented above, but experimental validation is critical to show the underlying assumptions hold in the real world. To effectively evaluate our OSR performance in a large-scale setting, we adapt the protocol recently established by Cruz et al. [7], with their provided code. Their protocol consists of large-scale datasets and modern architectures and allows the prediction of an operational threshold closely resembling a real-world system. Additionally, we conduct ablations to understand the performance impact of our approach and its components.

4.1. Details

4.1.1. Datasets

We utilize the large-scale ImageNet2012 [33] as our known class training and test sets. For OOSA threshold prediction, we employ ImageNetV2 [31] as our known validation set.

Bitterwolf et al. [3] recently showed that the 21K-P Open-Set splits [40] commonly used for OOD/unknowns evaluation contain over 40% overlap with ImageNet-1K, raising concerns about in-distribution contamination. Using such splits to estimate operational thresholds or evaluate performance can be misleading—we elaborate on this in the supplemental with supporting examples. To avoid this issue, we follow the protocol by Cruz et al. [7], but replace their unknowns surrogate with 10K images from OpenImage-O, which has no significant overlap. For completeness, we also report results on 21K-P splits in the supplement to match PostMax's evaluation.

At test time, we use 10K images from the plant classes in iNaturalist (iNat) [39] and the remaining 7.6K images from OpenImage-O (Open-O) [41] as unknowns. We also include the NINCO [3] (5.8K images) and Textures (Text) [6] (5.1K images) datasets as unknowns. Bitterwolf et al. [3] reported significant contamination in several commonly used datasets, including Places [49] (59.5%) and Textures [6] (20.0%). Using contaminated datasets for OSR is problematic, as high confidence on mislabeled unknowns can compromise evaluation; properly labeled unknowns typically yield better performance. In the supplemental, we provide examples where contaminated samples closely resemble ImageNet-1K training classes. While Textures and OpenImage-O each have 20% overlap, we include them to increase dataset diversity, with further analysis provided in the supplemental. However, we caution against over-interpreting results on these contaminated datasets.

4.1.2. Architectures

By utilizing ImageNet2012 [33], we can select well-studied pre-trained architectures for our evaluation. We evaluate performance on the traditional ResNet-50 [13] and two recent state-of-the-art architectures pre-trained on ImageNet2012 only: ConvNeXtV2-H [42] and Hiera-H [34]. We also use a Vision Transformer trained with masked-auto-encoder ViT/MAE[14] and the original ConvNeXt [24]. For each network, we extract features from the penultimate layer and the added special code to extract the Hadamard product.

4.1.3. Metrics

Our main results utilize the recent Operational Open-Set Accuracy (OOSA) [7], using code obtained from the authors. OOSA evaluates algorithm performance using the predicted threshold from validation. Additionally, we also utilize the traditional OSR metric, the Open-Set Classification Rate (OSCR) [8] as area under the curve in tables here and plots in the supplemental. In the supplemental, we also include Area Under the Receiver Operator Curve (AUROC), commonly used in the OOD literature but stress that AUROC should be interpreted with caution as it is generally incoherent for comparisons of algorithms [12].

4.2. Results

We compared COSTARR with recent approaches: PostMax [7] - current state-of-the-art for OSR, Maximum Logit (MaxLogit) [16, 40], and Maximum Softmax Probability (MSP) [15, 40]. Additionally, we include recent state-of-the-art OOD methods (as determined by the OpenOOD Benchmark leaderboard [45, 47]): SCALE [44] and NNGuide [29]. Since COMBOOD [30] is not yet integrated within OpenOOD, we cannot reproduce its results; however, we include results from their paper in the supplemental.

As shown in Tab. 1, COSTARR outperforms other methods on OOSA, a powerful metric used to measure the deployment characteristics of Open-Set algorithms. Notably, NINCO [3], a purpose-built OOD dataset designed to avoid contamination with ImageNet2012, provides a powerful point of comparison at which COSTARR excels. To ensure the results are not just random effects, dataset-dependent, architecture-specific, or validation tuning results, we computed Wilcoxon signed rank test, as implemented in scipy, with Bonferroni correction across Post-Max's five different splits of the validation data with each architecture and algorithm – see supplemental for more details. Note this is different than the t-tests used in [7], because the data is reused in ways that likely violate the independence needed for t-tests. All statistical claims in the paper use this test process. Since OOSA computes thresholds directly, there are *NO free/tuned parameters* in these experiments. The result of the statistical testing is that COSTARR is statistically significantly better ($p < 10^{-5}$) on each architecture separately, as well as very significantly ($p < 10^{-6}$) across all architectures combined. Additionally, while Places [49], 21K-P Easy/Hard [40], and SUN [43] have significant ImageNet2012 contamination, COSTARR still performs significantly ($p < 10^{-}3$)) better

Arch	Method	iNat	NINCO	Open-O	Text
ResNet-50	SCALE	0.635	0.555	0.581	0.628
	NNGuide	0.633	0.445	0.571	0.518
	MaxLogit	0.705	0.641	0.679	0.636
	MSP	0.745	0.675	0.719	0.676
	PostMax	0.762	0.623	0.730	0.696
	COSTARR	**0.804**	**0.693**	**0.773**	**0.749**
ConvNeXt-L	SCALE	0.681	0.650	0.668	0.679
	NNGuide	0.601	0.513	0.615	0.505
	MaxLogit	0.744	0.692	0.718	0.711
	MSP	0.778	0.707	0.752	0.722
	PostMax	0.820	0.714	0.798	0.761
	COSTARR	**0.837**	**0.744**	**0.814**	**0.778**
ConvNeXtV2-H	SCALE	0.777	0.713	0.752	0.767
	NNGuide	0.748	0.628	0.700	0.643
	MaxLogit	0.800	0.732	0.775	0.773
	MSP	0.808	0.734	0.782	0.756
	PostMax	0.850	0.755	0.830	0.796
	COSTARR	**0.856**	**0.756**	**0.835**	**0.802**
ViT-H/MAE-H	SCALE	0.758	0.689	0.710	0.750
	NNGuide	0.639	0.550	0.672	0.541
	MaxLogit	0.797	0.721	0.750	0.765
	MSP	0.812	0.737	0.777	0.761
	PostMax	0.865	0.749	0.841	0.808
	COSTARR	**0.876**	**0.780**	**0.851**	**0.820**
Hiera-H	SCALE	0.732	0.682	0.684	0.742
	NNGuide	0.798	0.644	0.770	0.670
	MaxLogit	0.796	0.723	0.747	0.768
	MSP	0.822	0.744	0.785	0.772
	PostMax	0.870	0.758	0.850	0.818
	COSTARR	**0.879**	**0.788**	**0.857**	**0.826**

Table 1. OPERATIONAL OPEN-SET ACCURACY. The mean OOSA (↑) of all methods. To predict an operational threshold, we validate the methods using ImageNetV2 [31] ($10K$ images) as knowns and OpenImage-O [41] ($10K$ images) as unknowns. Then, each method's threshold is deployed and tested on five ILSVRC2012 *val* [33] splits (each containing $10K$ images) and specified unknowns. OSR is performed on extractions from various pre-trained architectures. COSTARR, our novel algorithm, has the best scores (**bold**) for each respective architecture and unknowns dataset.

Arch	Method	iNat	NINCO	Open-O	Text
ResNet-50	SCALE	0.621	0.499	0.545	0.598
	NNGuide	0.596	0.419	0.551	0.557
	MaxLogit	0.682	0.625	0.661	0.609
	MSP	0.720	0.664	0.703	0.664
	PostMax	0.743	0.624	0.730	0.720
	COSTARR	**0.773**	**0.699**	**0.760**	**0.755**
ConvNeXt-L	SCALE	0.669	0.628	0.636	0.657
	NNGuide	0.594	0.462	0.616	0.453
	MaxLogit	0.729	0.677	0.695	0.693
	MSP	0.771	0.717	0.751	0.731
	PostMax	0.800	0.725	0.792	0.770
	COSTARR	**0.815**	**0.761**	**0.809**	**0.793**
ConvNeXtV2-H	SCALE	0.773	0.712	0.748	0.771
	NNGuide	0.759	0.639	0.719	0.653
	MaxLogit	0.794	0.735	0.771	0.779
	MSP	0.810	0.754	0.791	0.782
	PostMax	0.834	0.773	0.827	0.805
	COSTARR	**0.838**	**0.782**	**0.831**	**0.814**
ViT-H/MAE-H	SCALE	0.749	0.661	0.679	0.744
	NNGuide	0.646	0.469	0.670	0.442
	MaxLogit	0.790	0.709	0.730	0.764
	MSP	0.816	0.755	0.785	0.782
	PostMax	0.846	0.772	0.839	0.822
	COSTARR	**0.854**	**0.803**	**0.848**	**0.835**
Hiera-H	SCALE	0.718	0.643	0.633	0.730
	NNGuide	0.793	0.620	0.781	0.669
	MaxLogit	0.783	0.706	0.715	0.765
	MSP	0.822	0.760	0.788	0.792
	PostMax	0.846	0.784	0.842	0.825
	COSTARR	**0.856**	**0.812**	**0.850**	**0.838**

Table 2. AREA UNDER OPEN SET CLASSIFICATION RATE CURVE. The AUOSCR (↑) of all methods. To compute, we tested methods using ILSVRC2012 *val* [33] ($50K$ images) as knowns and specified unknowns. The best scores for each respective architecture and unknowns dataset are in **bold**.

than all methods except PostMax, but was never statistically worse. We include those datasets in the supplemental, with a discussion about data contamination and statistical testing.

Using AUOSCR as a secondary metric, Tab. 2 shows COSTARR outperforms all methods; again, the differences are statistically significant. In the supplemental we include AUROC tables where again COSTARR is statistically significantly better overall ($p < 10^{-3}$).

4.3. Ablations

We ran an ablation study to examine the key elements of our attenuation hypothesis and to analyze which components of COSTARR contribute to the observed performance gains. All ablations use AUOSCR as the evaluation metric, as it is not sensitive to threshold selection.

The hypothesis states that both pre– and post–attenuation features should be used. Also, it suggests that the final logits, including the learned bias, likely contain valuable information necessary for the probabilistic interpretation to function properly.

We build an ablation version using only COSTARR Similarity (Eq. 4), employing either the Hadamard product H or deep features F alone, and excluding the Logit. Treat-

Arch	Method	iNat	NINCO	Open-O	Text
ResNet-50	PostMax	0.743	0.624	0.730	0.720
	Hadamard	0.426	0.365	0.393	0.384
	Features	0.746	0.681	0.736	0.740
	NoLogit	0.771	0.690	0.757	**0.756**
	CO-SM	0.757	0.693	0.743	0.725
	COSTARR	**0.773**	**0.699**	**0.760**	0.755
ConvNV2	PostMax	0.834	0.773	0.827	0.805
	Hadamard	0.834	0.772	0.829	0.807
	Features	0.835	0.773	0.830	0.810
	NoLogit	0.837	0.775	0.833	0.813
	CO-SM	0.832	0.778	0.823	0.807
	COSTARR	**0.838**	**0.782**	**0.831**	**0.814**
Hiera-H	PostMax	0.846	0.784	0.842	0.825
	Hadamard	0.853	0.807	0.848	0.833
	Features	0.854	0.809	0.849	0.835
	NoLogit	0.855	0.809	**0.850**	0.836
	CO-SM	0.851	0.807	0.843	0.831
	COSTARR	**0.856**	**0.812**	**0.850**	**0.838**

Table 3. ABLATION STUDY. AUOSCR scores from our ablation studies; the best scores are in **bold**. Hadamard refers to a version of COSTARR that uses only the Hadamard product features (H) without concatenation, i.e. post-attenuation for similarity and selection. Features uses only pre-attenuation features (F) for similarity and selection. NoLogit shows performance when using the concatenated F and H similarity, with the class selected based on the maximum similarity. For comparison, we also include PostMax, the prior state of the art. We evaluate across three networks (as used in supplemental Fig. 3: ResNet-50 [13], ConvNeXtV2-H [42], and Hiera-H [34]. In all cases, the ablations perform worse than COSTARR, providing empirical evidence that incorporating features typically ignored (discarded or marginalized by the Hadamard product) improves known/unknown differentiation. The comparison of softmax (CO-SM) *vs*. logits (COSTARR) further confirms that normalization using logits yields superior performance.

ing the resulting vector as a type of prototype, we select the class with the maximum similarity. Although there are some architecture and dataset-specific variations, both of these variants perform statistically significantly worse than COSTARR overall ($p < .05$).

We also include an ablation with CO-SM, which uses softmax instead of logits for scaling – this does not cancel out in the probabilistic interpretation. Another ablation demonstrates that even when combining both pre– and post–attenuation features (F and H), the logits still provide additional value. To test this, we introduce a variant called NoLogit, which uses the full COSTARR similarity from Eq. 4 to both select the winning class and reject unknowns. This variant performs slightly better than using either feature alone, but remains weaker than the overall COSTARR.

We were initially surprised by the much lower performance and greater variability of the ablation version using softmax instead of logit (CO-SM). For some networks and datasets, it outperformed the NoLogit ablation, while for others, it performed worse. Although both max logit and max softmax select the same class, the difference likely arises from softmax's per-instance normalization, which affects the score and the probabilistic interpretation in Eq.9. Since CO-SM never outperformed COSTARR with logits, we did not explore it further.

From Tab. 3, COSTARR consistently outperforms the ablation variants. These provide strong evidence in support of our *attenuation hypothesis* and indicate that each component of COSTARR contributes meaningfully to its performance. Additional discussion is provided in the supplemental, particularly regarding COSTARR's performance.

5. Discussion and Conclusion

Exploring our novel *attenuation hypothesis*, we analyzed deep features and classification layer weights, finding that multiple networks rely on classification weights that attenuate general features and why that attenuation is problematic for OSR. From our analysis of networks (see also Fig. 3 in supplement), we showed that each feature has low weight for some classes and high for others, supporting our hypothesis.

We exploited our *attenuation hypothesis* in our proposed OSR algorithm, COSTARR, which introduces only a constant to test-time complexity. Our main results on operational open-set accuracy (Tab. 1) demonstrate that COSTARR outperforms prior approaches, including the recent PostMax [7]. These results are practically better, statistically significant, and achieved at almost no added computational cost. Additionally, results with a secondary OSR evaluation metric, AUOSCR (Tab. 2), show statistically significant performance gains over prior approaches. The ablations (Tab. 3) validated that COSTARR's performance derives from three components: the use of both pre– and post-attenuated features as well as a small increase from using logits. Through this analysis, we found direct evidence supporting the *attenuation hypothesis* as well as the benefit of using logits and per-class models for OSR. While the ablations show the effectiveness of each component varies across networks and datasets, COSTARR achieves unprecedented performance across all networks by consolidating all three sources of information. We leave the exploration of these insights on non-pretrained networks for future work.

Through this exploration, we have advanced the understanding of the weak performance of different open-set classifiers because they discard information useful to either known or unknown samples. COSTARR takes the first steps toward consolidating pre-attenuation and post-attenuation information to improve robust recognition.

References

[1] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13349–13358, 2021. 2

[2] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. 3, 4

[3] Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICLR Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023. 2, 6, 13, 14

[4] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 507–522. Springer, 2020. 2

[5] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8065–8081, 2021. 2

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 6

[7] Steve Cruz, Ryan Rabinowitz, Manuel Günther, and Terrance E Boult. Operational open-set recognition and postmax refinement. In *European Conference on Computer Vision*, pages 475–492. Springer, 2025. 3, 4, 6, 8, 1, 10, 13

[8] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2, 3, 4, 6

[9] Thomas G Dietterich and Alex Guyer. The familiarity hypothesis: Explaining the behavior of deep open set methods. *Pattern Recognition*, 132:108931, 2022. 1, 4

[10] Ke Fan, Tong Liu, Xingyu Qiu, Yikai Wang, Lian Huai, Zeyu Shangguan, Shuang Gou, Fengjian Liu, Yuqian Fu, Yanwei Fu, et al. Test-time linear out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23752–23761, 2024. 3

[11] Yunrui Guo, Guglielmo Camporese, Wenjing Yang, Alessandro Sperduti, and Lamberto Ballan. Conditional variational capsule network for open set recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 103–111, 2021. 2

[12] David J. Hand and Christoforos Anagnostopoulos. When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letters*, 34(5):492–495, 2013. Publisher: Elsevier. 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 8

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6

[15] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations (ICLR)*, 2017. 3, 6

[16] 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 *International Conference on Machine Learning (ICML)*, 2022. 3, 6

[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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10951–10960, 2020. 3

[18] Hongzhi Huang, Yu Wang, Qinghua Hu, and Ming-Ming Cheng. Class-specific semantic reconstruction for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4214–4228, 2022. 2

[19] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 813–822, 2021. 2

[20] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 3

[21] Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. *ICML*, 2024. 3

[22] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[23] Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23946–23955, 2023. 3

[24] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 6

[25] Jing Lu, Yunlu Xu, Hao Li, Zhanzhan Cheng, and Yi Niu. Pmal: Open set recognition via robust prototype mining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1872–1880, 2022. 2

[26] Dimity Miller, Niko Sunderhauf, Michael Milford, and Feras Dayoub. Class anchor clustering: A loss for distance-based open set recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3570–3578, 2021.

[27] WonJun Moon, Junho Park, Hyun Seok Seong, Cheol-Ho Cho, and Jae-Pil Heo. Difficulty-aware simulator for open set recognition. In *European Conference on Computer Vision*, pages 365–381. Springer, 2022.

[28] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual

images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[29] Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *International Conference on Computer Vision (ICCV)*, pages 1686–1695, 2023. 3, 6

[30] Magesh Rajasekaran, Md Saiful Islam Sajol, Frej Berglind, Supratik @articleliang2017enhancing, title=Enhancing the reliability of out-of-distribution image detection in neural networks, author=Liang, Shiyu and Li, Yixuan and Srikant, Rayadurgam, journal=arXiv preprint arXiv:1706.02690, year=2017 Mukhopadhyay, and Kamalika Das. Combood: A semiparametric approach for detecting out-of-distribution data for image classification. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pages 643–651. SIAM, 2024. 3, 6, 9

[31] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do Imagenet classifiers generalize to Imagenet? In *International Conference on Machine Learning (ICML)*, pages 5389–5400. PMLR, 2019. 6, 7, 1

[32] Ethan M. Rudd, Lalit P. Jain, Walter J. Scheirer, and Terrance E. Boult. The extreme value machine. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. 3, 4

[33] 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. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 6, 7, 1, 9

[34] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. *International Conference on Machine Learning*, 2023. 6, 8

[35] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(7), 2012. 1

[36] Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, and Guohao Peng. Conditional gaussian distribution learning for open set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13480–13489, 2020. 2

[37] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021. 3

[38] Keke Tang, Chao Hou, Weilong Peng, Runnan Chen, Peican Zhu, Wenping Wang, and Zhihong Tian. Cores: Convolutional response-based score for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10916–10925, 2024. 3

[39] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[40] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *International Conference on Learning Representations (ICLR)*, 2022. 3, 4, 6, 1

[41] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022. 3, 6, 7

[42] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. 6, 8

[43] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 6, 13, 16, 17

[44] Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In *International Conference on Learning Representations (ICLR)*, 2024. 3, 6

[45] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. OpenOOD: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:32598–32611, 2022. 3, 6

[46] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[47] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. In *NeurIPS Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2023. 3, 6

[48] Xuelin Zhang, Xuelian Cheng, Donghao Zhang, Paul Bonnington, and Zongyuan Ge. Learning network architecture for open-set recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3362–3370, 2022. 2

[49] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(6), 2017. 6, 13, 14, 15

[50] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2021. 2