COSTARR: Consolidated Open Set Technique with Attenuation for Robust Recognition

Supplementary Material

6. Extended discussion

While we stated that probabilist interpretation of COSTARR depends on an accurate classifier, it is interesting to note that in 2 the gap over the prior state of the art (PostMax) is largest for ResNet-50, which is the weakest network in terms of closed set accuracy, suggesting there is more to the improvements than just the probabilistic interpretation.

In the ablation, 3, we see that for ResNet50, the Hadamard only version, with only post-attentive features, did much worse than using only the Features. We hypothesize that maybe the gains for ResNet50 in other tables are from the concatenation with Features improving class separation beyond that the logit/Hadamard feature can provide. For Resnet50 with Textures as the unknown, adding the logit did not really improve things, again, maybe suggesting the post-attenuation Hadamard features are weaker in ResNet.

The weakest element of the ablation was the difference between NoLogit and COSTARR, which for two networks there was one dataset where the results without the logit were nearly identical to the overall COSTARR, though the differences were not statistically different. We note that the full COSTARR with logit version is actually cheap at test time since we access only the mean for the max logit class and already have the logit value from that computation.

7. Metrics Cont.

For completeness, we include Open-Set Accuracy (OSA) curves in Fig. 4-6. Note, the operational performance (OOSA) reported in Tab. 1, is indicated by a $\star$ on each curve. Consistent with the tabular results, COSTARR achieves maximum OSA across all datasets.

In addition to OOSA and AUOSCR metrics included in the main paper, we also evaluated methods using Open-Set Classification Rate (OSCR) curves, shown in Fig. 7-9. Similar to OOSA (Tab. 1) and AUROC in (Tab. 5), COSTARR outperforms all methods across all datasets.

8. Different OOSA Validation

For consistency of comparison with prior work, we also report performance on the same validation from Cruz et al. [7] in Tab. 4: ImageNetV2 (10K images) as knowns and 21K-P Hard (9.8K images) as unknowns with their "contaminated" validation process. Similar to the results in Tab. 1, COSTARR outperforms all methods across all architectures except ViT-H.

Arch	Method	iNat	NINCO	Open-O	Text
ResNet-50	SCALE	0.532	0.402	0.634	0.386
	NNGuide	0.513	0.385	0.647	0.357
	MaxLogit	0.669	0.584	0.707	0.565
	MSP	0.702	0.617	0.747	0.606
	PostMax	0.571	0.452	0.689	0.431
	COSTARR	**0.718**	**0.627**	**0.789**	**0.627**
ConvNeXt-L	SCALE	0.683	0.638	0.664	0.659
	NNGuide	0.583	0.447	0.662	0.450
	MaxLogit	0.740	0.682	0.728	0.693
	MSP	0.773	0.701	0.791	0.706
	PostMax	0.794	0.698	0.830	0.722
	COSTARR	**0.801**	**0.721**	**0.846**	**0.730**
ConvNeXtV2-H	SCALE	0.772	0.705	0.767	0.740
	NNGuide	0.713	0.597	0.745	0.601
	MaxLogit	0.786	0.718	0.792	0.740
	MSP	0.796	0.723	0.820	0.736
	PostMax	0.812	0.732	0.849	0.744
	COSTARR	**0.819**	**0.739**	**0.857**	**0.752**
ViT-H	SCALE	0.748	0.671	0.706	0.718
	NNGuide	0.636	0.552	0.661	0.544
	MaxLogit	0.788	0.711	0.755	0.742
	MSP	0.794	0.717	0.809	0.727
	PostMax	0.826	0.732	0.861	0.761
	COSTARR	**0.834**	**0.759**	**0.872**	**0.773**
Hiera-H	SCALE	0.727	0.667	0.663	0.715
	NNGuide	0.782	0.625	0.796	0.660
	MaxLogit	0.780	0.707	0.745	0.736
	MSP	0.815	0.739	0.814	0.756
	PostMax	0.825	0.740	0.864	0.764
	COSTARR	**0.852**	**0.782**	**0.883**	**0.797**

Table 4. OPERATIONAL OPEN-SET ACCURACY. The mean OOSA (↑) of all methods on the validation from Cruz et al. [7]. To predict an operational threshold, we validate the methods using ImageNetV2 [31] ($10K$ images) as knowns and 21K-P Hard [40] ($9.8K$ images) as unknowns. Then, each method's threshold is deployed and tested on five different ILSVRC2012 *val* [33] splits (each containing $10K$ images) and specified unknowns. OSR is performed on extractions from various pre-trained architectures. COSTARR (ours) is the final method in each network series and the best scores for each respective architecture and unknowns dataset are in **bold**.

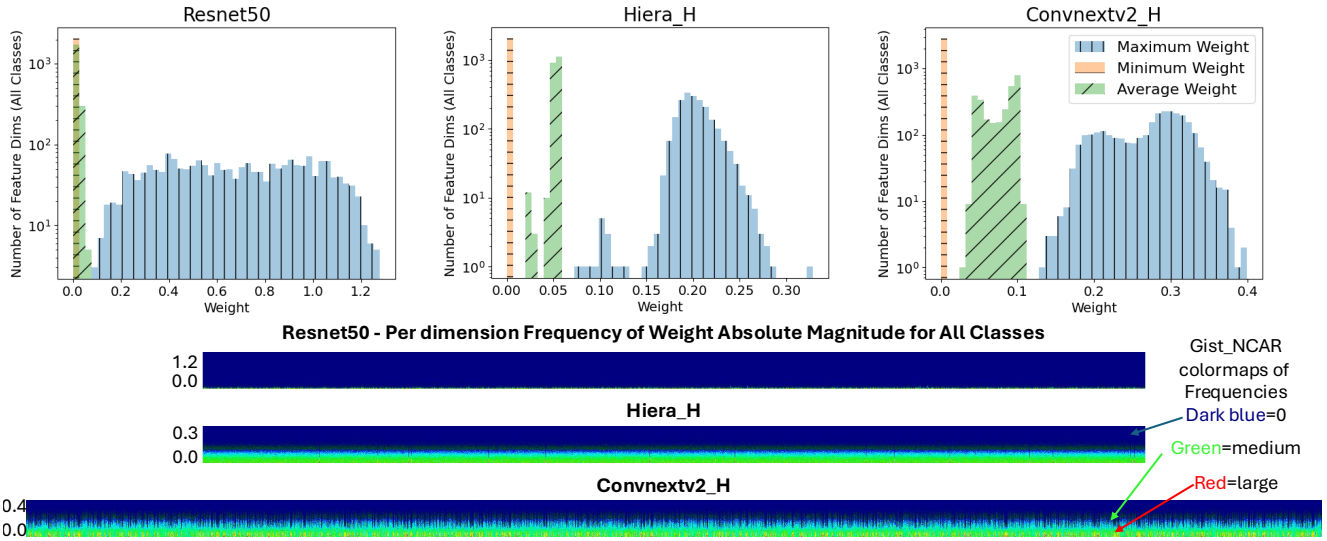
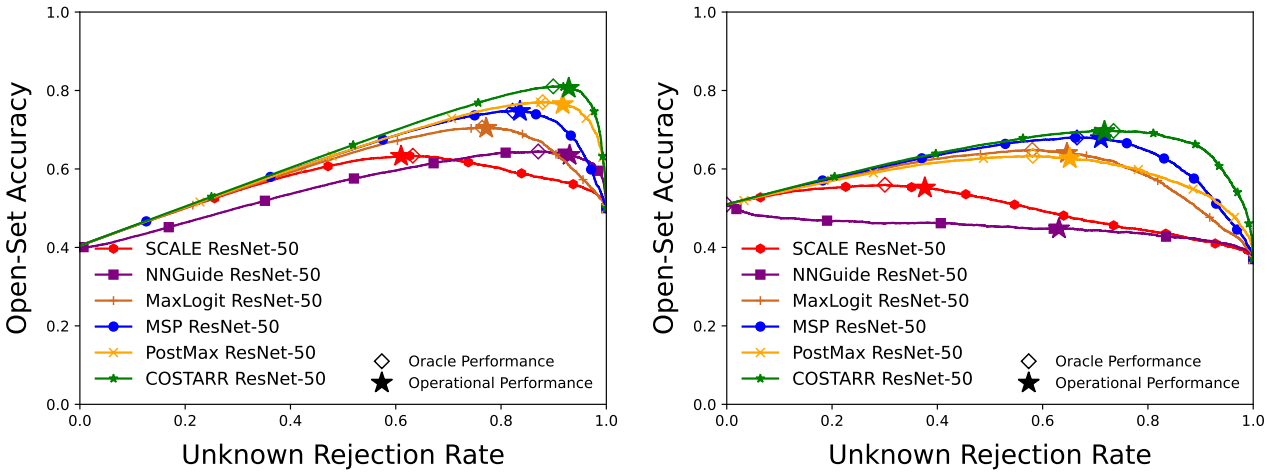
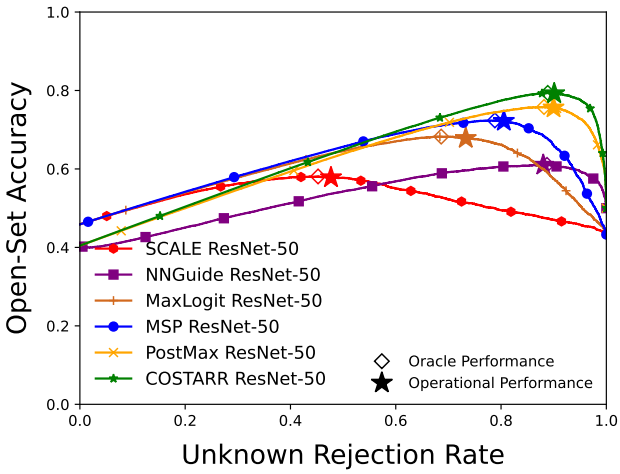


Figure 3. HISTOGRAM OF CLASS-WEIGHTS. The spread of Min (horizontal stripes), Mean (diagonal stripes), and Max (vertical stripes) weights per dimension across all classes illustrate how any given feature can be discarded by the classification layer of a DNN (Hiera-H pictured). The Maximum weights are all reasonably above 0, indicating that each feature dimension is useful to some class. The minimum weights all approach 0, indicating that there is a class that ignores each feature. Since the Mean weights tend closer to 0 than to the mean of the max weights, feature dimensions are more often ignored (by having their contribution attenuated) rather than being strong contributors to the final logit. The bottom three color bars visualize how often this occurs across every class in ImageNet, with the absolute value per-dimension frequency of each weight visualized. The bars are different widths due to the different feature dimensions from each network. Each pixel in the 80-pixel-tall columns represents a bin from the range between 0 and the absolute maximum weight of all dimensions and classes; brightness indicates frequency (using the "gist_ncar" color scheme). The concentration near of bright green near the bottom (approaching 0 weight) shows that every feature dimension most often has its contributions to maximum logit confidence attenuated, depending on which class' weights are used. While these dimensions have been weighted to discriminate between known classes, they ignore information that could help differentiate known from unknown.

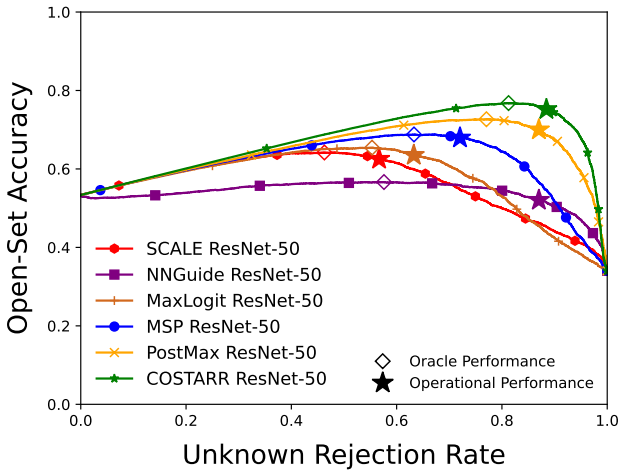


(a) iNaturalist

(b) NINCO

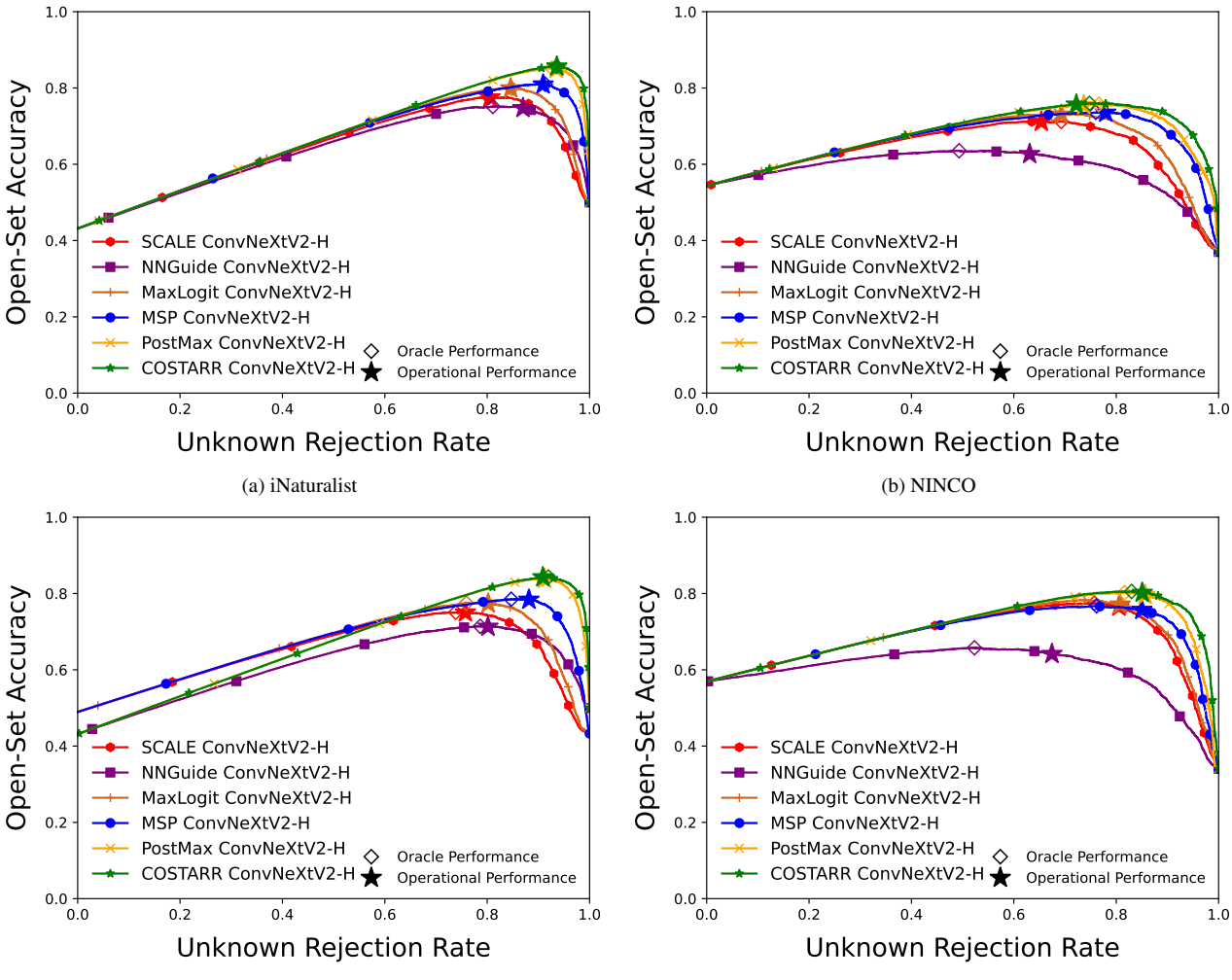


(c) OpenImage-O



(d) Textures

Figure 4. OPEN-SET ACCURACY CURVES. The OSA curves of all methods for ResNet-50 (same experimental setup as Tab. 1). A $\star$ signifies the peak performance (OOSA) of each method.



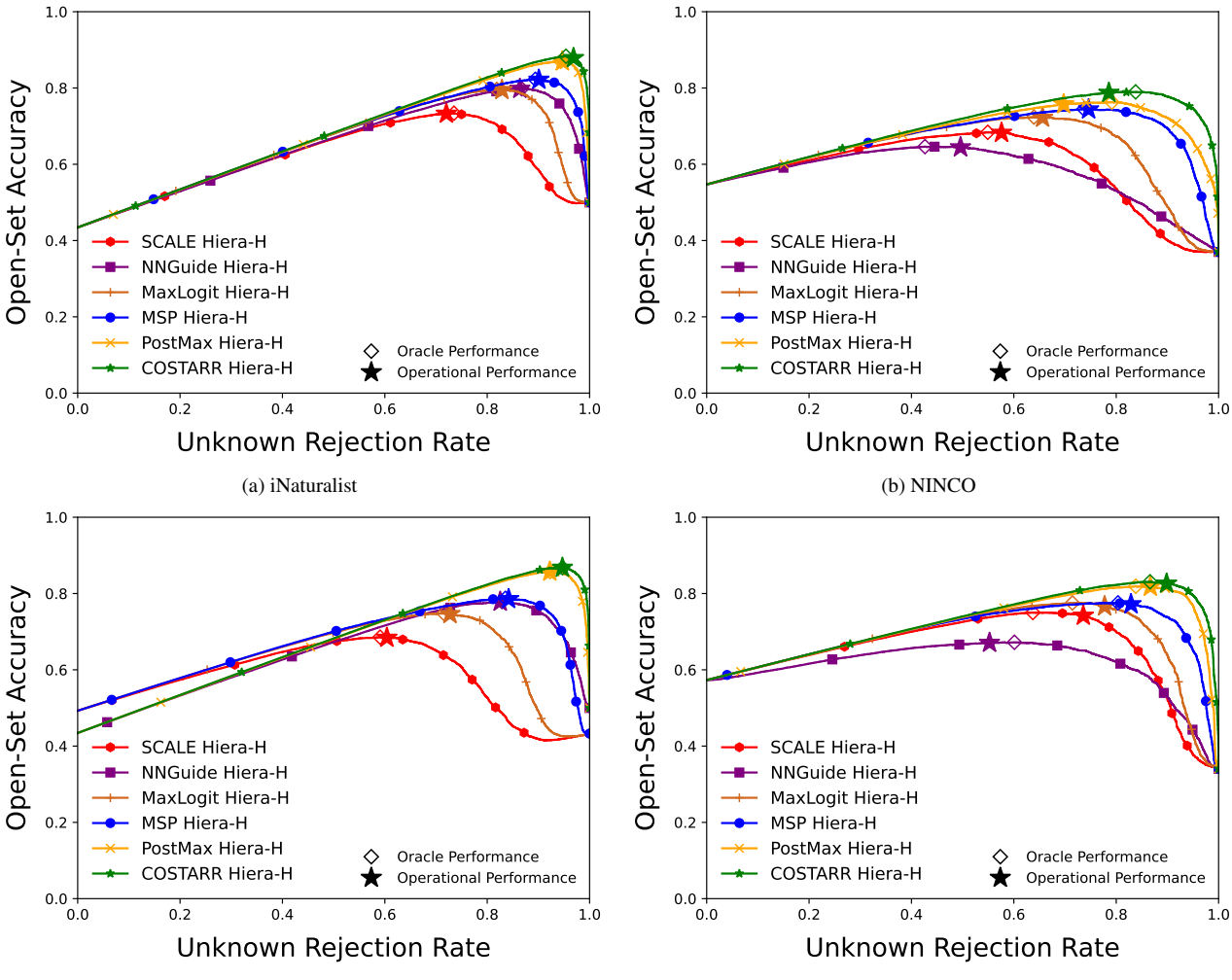
(a) iNaturalist

(b) NINCO

(c) OpenImage-O

(d) Textures

Figure 5. OPEN-SET ACCURACY CURVES. The OSA curves of all methods for ConvNeXtV2-H (same experimental setup as Tab. 1). A $\star$ signifies the peak performance (OOSA) of each method.



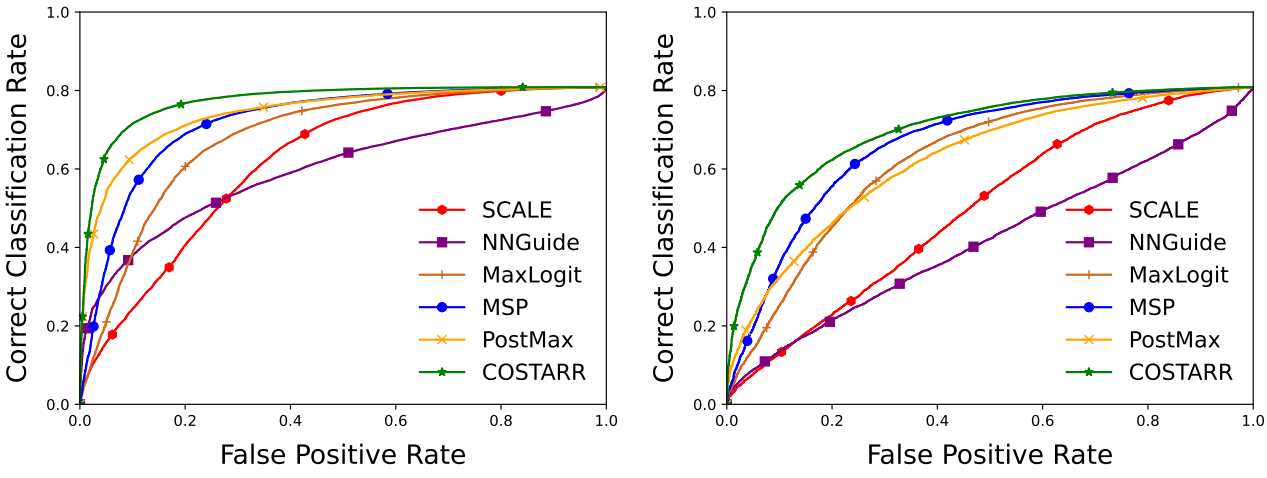
(a) iNaturalist

(b) NINCO

(c) OpenImage-O

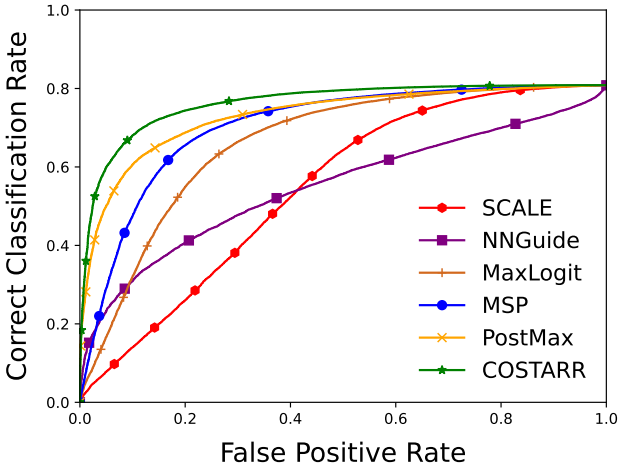
(d) Textures

Figure 6. OPEN-SET ACCURACY CURVES. The OSA curves of all methods for Hiera-H (same experimental setup as Tab. 1). A $\star$ signifies the peak performance (OOSA) of each method.

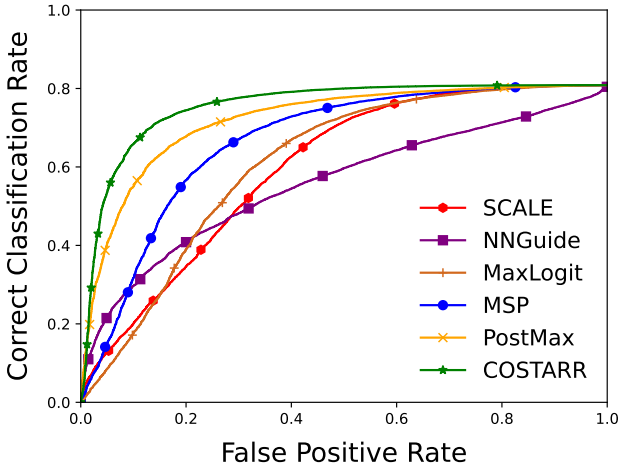


(a) iNaturalist

(b) NINCO

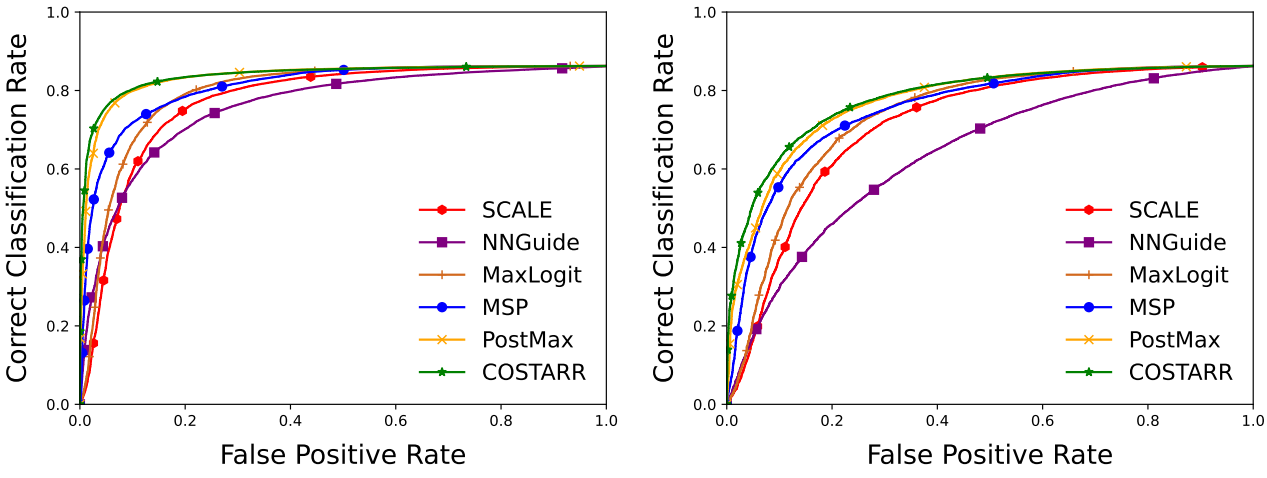


(c) OpenImage-O



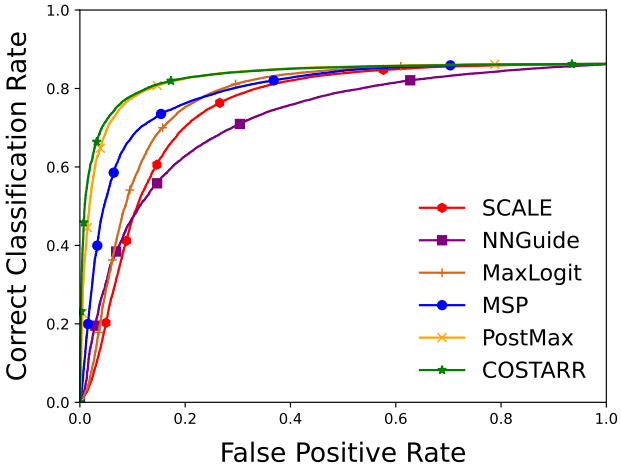
(d) Textures

Figure 7. OPEN-SET CLASSIFICATION RATE CURVES. The OSCR curves of all methods for ResNet-50 (same experimental setup as Tab. 5).

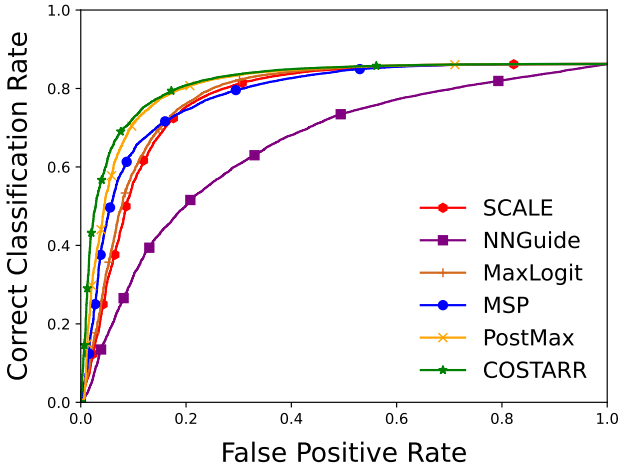


(a) iNaturalist

(b) NINCO

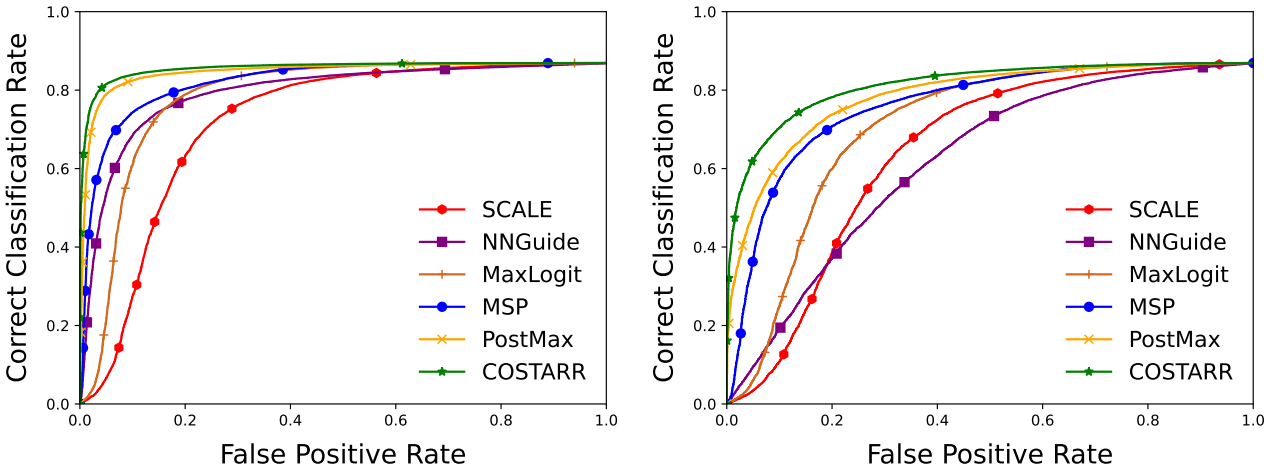


(c) OpenImage-O



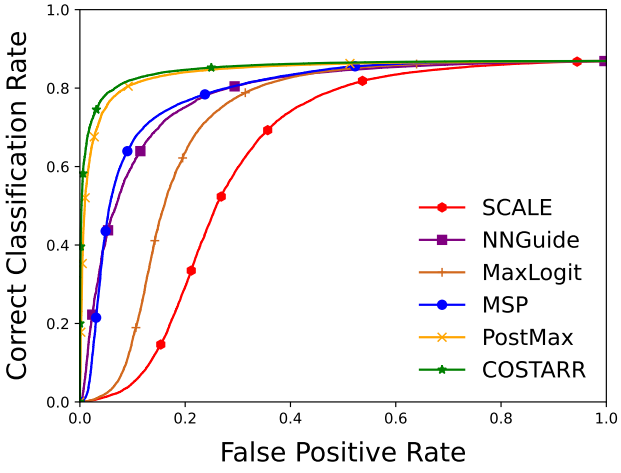
(d) Textures

Figure 8. OPEN-SET CLASSIFICATION RATE CURVES. The OSCR curves of all methods for ConvNeXtV2-H (same experimental setup as Tab. 5).

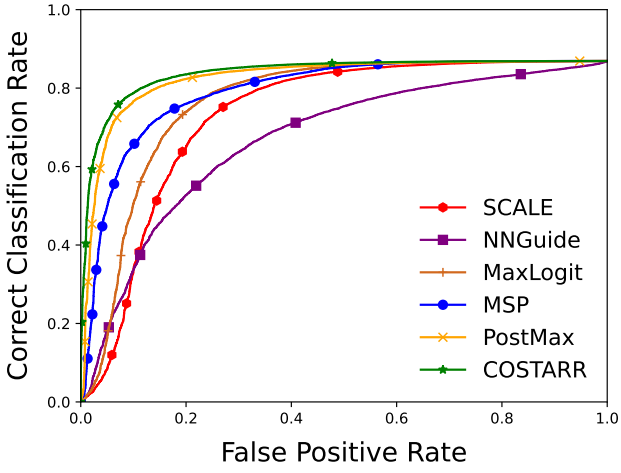


(a) iNaturalist

(b) NINCO



(c) OpenImage-O



(d) Textures

Figure 9. OPEN-SET CLASSIFICATION RATE CURVES. The OSCR curves of all methods for Hiera-H (same experimental setup as Tab. 5).

Arch	Method	iNat	NINCO	Open-O	Text
ResNet-50	SCALE	0.756	0.604	0.663	0.729
	NNGuide	0.715	0.492	0.656	0.665
	MaxLogit	0.804	0.729	0.777	0.714
	COMBOOD†	0.871	N/A	0.866	**0.970**
	MSP	0.848	0.768	0.823	0.774
	PostMax	0.882	0.722	0.862	0.852
	COSTARR	**0.931**	**0.816**	**0.910**	0.905
ConvNeXt-L	SCALE	0.773	0.721	0.734	0.76
	NNGuide	0.676	0.52	0.704	0.507
	MaxLogit	0.836	0.765	0.794	0.792
	MSP	0.876	0.799	0.848	0.823
	PostMax	0.920	0.817	0.911	0.880
	COSTARR	**0.944**	**0.863**	**0.934**	**0.912**
ConvNeXtV2-H	SCALE	0.878	0.803	0.849	0.879
	NNGuide	0.860	0.715	0.811	0.731
	MaxLogit	0.899	0.820	0.869	0.881
	MSP	0.908	0.831	0.883	0.873
	PostMax	0.949	0.864	0.940	0.912
	COSTARR	**0.954**	**0.875**	**0.946**	**0.925**
ViT-H	SCALE	0.847	0.742	0.767	0.842
	NNGuide	0.728	0.525	0.758	0.493
	MaxLogit	0.889	0.786	0.817	0.858
	MSP	0.910	0.828	0.869	0.867
	PostMax	0.957	0.857	0.947	0.925
	COSTARR	**0.970**	**0.896**	**0.959**	**0.942**
Hiera-H	SCALE	0.814	0.725	0.716	0.829
	NNGuide	0.896	0.690	0.881	0.744
	MaxLogit	0.883	0.786	0.802	0.860
	MSP	0.920	0.835	0.875	0.881
	PostMax	0.959	0.872	0.953	0.930
	COSTARR	**0.973**	**0.907**	**0.963**	**0.947**

Table 5. AREA UNDER RECEIVER OPERATING CHARACTERISTIC CURVE. The AUROC (↑) of all methods. To compute, we tested methods using ILSVRC2012 *val* [33] ($50K$ images) as knowns and specified unknowns. COSTARR (ours) is the final method in each network series and the best scores for each respective architecture and unknowns dataset are in **bold**. † indicates a result has been transposed from [30].

9. Statistical Testing

In Tables 6, 7, 8, and 9 we present P-values from statistical testing of COSTARR vs different algorithms on different architectures and datasets for validation and testing. COSTARR is statistically significantly better in almost all cases and is never statistically worse.

While [7] introduce splits and a testing procedure, we modified that process somewhat for this paper. For the splits for the knowns, there are non-overlapping samples from each class, and a set of unknowns varies by dataset, so they assumed independence and used two-sided paired t-tests. But when combining data over different architectures or different unknown sets, the independence breaks down. Thus we use Wilcoxon signed rank test, as implemented in scipy, with Bonferroni correction to account for the different number of tests being considered.

We obtained the splits from them and recomputed statistical significance using Wilcoxon signed rank test, as implemented in scipy, with Bonferroni correction. The results do not significantly change the conclusions from their paper, but we could only answer that by doing the proper non-parametric test. This also sets up the work so that future tests can properly compare with COSTARR. The splits will be included in with released code.

Net / Alg	**PostMax**	**MSP**	**MaxLogit**	**NNGuide**	**SCALE**
Reset-50 P-Value (N=15)	1.82E-08	1.16E-09	7.93E-12	3.87E-12	2.88E-15
ConvNext-L P-Value (N=15)	1.31E-07	1.24E-11	1.82E-10	5.70E-17	2.11E-11
ConvNextV2-H P-Value (N=15)	5.23E-10	1.79E-09	5.03E-09	1.73E-15	2.08E-10
ViT-H P-Value (N=15)	8.81E-06	1.09E-10	9.13E-11	2.15E-14	5.99E-12
Hiera-H P-Value (N=15)	9.24E-05	2.24E-11	8.01E-11	1.45E-09	2.26E-11
Overall P-Value (N=75)	9.35E-12	1.78E-37	3.02E-39	1.48E-30	5.28E-30

Table 6. STATISTICS COMPARING OOSA PERFORMANCE OF COSTARR VS. DIFFERENT ALGORITHMS ON "CLEAN DATA" WHERE OPEN-IMAGES IS A SURROGATE SET. P-values of the null hypothesis: that the mean performance is the same as COSTARR. This uses a subset of Open-Images as the surrogate set to select the threshold for the given network. The statistics corresponding to each architecture use five different splits of the validation data and unknowns drawn from the full set of iNaturalist, NINCO, and OpenImage_O. There are N=15 runs per architecture and 75 runs overall. Since OOSA computes thresholds directly on the surrogate set, there are *NO free/tuned parameters* in these experiments for any architecture. This is computed by Wilcoxon signed rank test with Bonferroni correction. All tests are statistically very significant.

Net / Alg	**PostMax**	**MSP**	**MaxLogit**	**NNGuide**	**SCALE**
Reset-50 P-value (N=40)	4.163E-08	3.067E-11	2.341E-21	3.955E-11	3.953E-21
ConvNext-L P-value (N=40)	6.710E-05	6.181E-08	4.354E-20	8.862E-15	7.392E-23
ConvNextV2-H P-value (N=40)	1.433E-03	2.152E-01	3.098E-15	3.385E-09	1.910E-18
ViT-H P-value (N=40)	1.006E-04	2.967E-07	3.704E-19	2.368E-18	6.521E-24
Hiera-H P-value (N=40)	1.372E-04	4.188E-11	5.033E-23	1.849E-13	1.549E-24
Overall P-value (N=200)	9.354-12	1.784-37	3.025E-39	1.484E-30	5.282E-30

Table 7. STATISTICS COMPARING OOSA PERFORMANCE OF COSTARR VS. DIFFERENT ALGORITHMS ON "CONTAMINATED DATA" WHERE OPEN-IMAGES IS SURROGATE SET. P-values of null hypothesis: that the mean performance is the same as COSTARR. This uses a subset of Open-Images as the surrogate set to select the threshold for the given network. The statistics corresponding to each architecture use five different splits of the validation data and unknowns drawn from the full set of iNaturalist, NINCO, OpenImage_O, Places, SUN, Textures, easy_i21k, and hard_i21k, so it includes results where many corrupted/overlapping classes/images, with N=40 runs per architecture and 200 overall. Since OOSA computes thresholds directly on the surrogate set, there are *NO free/tuned parameters* in these experiments for any architecture. This is computed by Wilcoxon signed rank test with Bonferroni correction. All tests except ConvNextV2-H for MSP (shown in purple) are statistically very significant.

Net/Alg	PostMax	MSP	MaxLogit	NNGuide	SCALE
Reset-50 P-value (N=15)	1.25E-11	7.72E-09	1.28E-09	3.83E-13	6.07E-13
ConvNext-L P-value (N=15)	3.95E-08	1.58E-08	2.36E-09	4.11E-13	1.19E-10
ConvNextV2-H P-value (N=15)	9.43E-08	4.15E-08	2.15E-08	4.23E-10	9.27E-09
ViT-H P-value (N=15)	4.34E-08	8.18E-09	1.41E-09	1.66E-11	1.08E-10
Hiera-H P-value (N=15)	3.60E-08	1.43E-08	1.62E-09	1.19E-09	7.53E-11
Overall P-value (N=75)	3.01E-06	6.41E-30	7.56E-33	7.95E-30	3.32E-32

Table 8. STATISTICS COMPARING OOSA PERFORMANCE OF COSTARR VS. DIFFERENT ALGORITHMS WHERE HARD_I21K IS A SURROGATE SET WITH NON-CONTAMINATED UNKNOWN SETS. P-values of the null hypothesis: that the mean performance is the same as COSTARR. This uses a subset of hard_i21k as the surrogate set to select the threshold for the given network. The statistics corresponding to each architecture use five different splits of the validation data and unknowns drawn from the non-contaminated set of iNaturalist, NINCO, and OpenImage_O. It has N=15 runs per architecture and 75 overall. This is computed by Wilcoxon signed rank test with Bonferroni correction. All tests are statistically very significant.

Net / Alg	**PostMax**	**MSP**	**MaxLogit**	**NNGuide**	**SCALE**
Reset-50 P-Value (N=36)	2.89E-12	4.27E-24	1.31E-30	3.11E-13	1.40E-17
ConvNext-L P-Value (N=36)	4.18E-07	9.78E-18	3.43E-14	1.58E-13	2.36E-17
ConvNextV2-H P-Value (N=36)	1.41E-02	1.59E-10	4.42E-09	1.43E-15	1.66E-09
ViT-H P-value (N=36)	7.77E-01	1.29E-15	1.33E-09	3.98E-16	1.56E-13
Hiera-H P-Value (N=36)	4.28E-06	2.07E-19	6.17E-19	4.64E-14	3.91E-18
Overall P-Value (N=176)	3.83E-10	4.85E-71	5.31E-60	1.50E-53	1.27E-60

Table 9. STATISTICS COMPARING OOSA PERFORMANCE OF COSTARR VS. DIFFERENT ALGORITHMS WHERE HARD_I21K IS A SURROGATE SET, SIMILAR TO THAT USED IN POSTMAX PAPER. P-values of the null hypothesis that the mean performance is the same as COSTARR. This uses a subset of hard_i21k as the surrogate set to select the threshold for the given network. The statistics corresponding to each architecture use five different splits of the validation data and unknowns drawn from the full set of iNaturalist, NINCO, OpenImage_O, Places, SUN, Textures, easy_i21k, so it includes results where many corrupted/overlapping classes/images, with N=36 runs per architecture and 176 overall, which is slightly less than when Open-Images is used for the surrogate set. Since OOSA computes thresholds directly on the surrogate set, there are *NO free/tuned parameters* in these experiments for any architecture. This is computed by paired two-sided t-tests with Bonferroni correction. All tests except ViT-H for PostMax (shown in purple) are statistically very significant.

10. Contamination

Given the recent in-distribution contamination analysis performed by Bitterwolf et al. [3], we excluded any datasets with >20% contamination from the main paper. However, to report performance on datasets used in prior SOTA evaluations [7], we show results in Tab. 10. Note, COSTARR models per-class confidence, so any significant overlap will hinder performance. Nonetheless, overall performance is still statistically significant as shown in Sec. 9.

To analyze how in-distribution contaminated unknown datasets degrade evaluations, we examined the images responsible for the performance difference between PostMax and COSTARR for the Places and SUN datasets. At the validation-predicted OOSA threshold, we looked at samples which PostMax said were unknown, but COSTARR labeled as known. We compared these images with ImageNet training data from the closed-set predicted class. From Places, we found every image within this performance gap has visually significant overlap with ImageNet training data, from SUN we found the same holds for nearly every image. We include *all* images from these examinations in Figures 10 & 11 for Places and Figures 12 and 13. Labeling images as unknown while they are visually represented in the training set hinders OSR and OOD evaluations. We can see from these figures, the images responsible for COSTARR's seemingly inferior performance to PostMax are actually knowns which have been mislabeled as unknowns. Of course, all algorithms are affected by the presence of these mislabeled images. When contaminated dataset is used as unknowns, even a perfect OSR system will have a reduced OOSA and AUROC score.

Arch	Method	Places	SUN	21K E	21K H
ResNet-50	SCALE	0.503	0.539	0.382	0.355
	NNGuide	0.607	0.626	**0.776**	0.487
	MaxLogit	0.631	0.652	0.632	0.535
	MSP	0.679	0.696	0.687	**0.573**
	PostMax	0.690	0.719	0.739	0.485
	COSTARR	**0.691**	**0.726**	0.727	0.535
ConvNeXt-L	SCALE	0.588	0.617	0.514	0.492
	NNGuide	0.683	0.699	0.656	0.464
	MaxLogit	0.652	0.675	0.586	0.535
	MSP	0.722	0.733	0.704	**0.620**
	PostMax	**0.752**	**0.773**	0.733	0.533
	COSTARR	0.739	0.767	**0.743**	0.582
ConvNeXtV2-H	SCALE	0.667	0.681	0.580	0.515
	NNGuide	0.720	0.727	0.720	0.552
	MaxLogit	0.707	0.721	0.637	0.560
	MSP	0.750	0.755	0.728	**0.648**
	PostMax	**0.779**	**0.790**	**0.740**	0.572
	COSTARR	0.730	0.748	0.712	0.565
ViT-H	SCALE	0.637	0.672	0.569	0.510
	NNGuide	0.700	0.696	0.554	0.393
	MaxLogit	0.692	0.717	0.632	0.566
	MSP	0.739	0.751	0.704	**0.627**
	PostMax	**0.765**	**0.784**	0.732	0.541
	COSTARR	0.754	0.781	**0.742**	0.588
Hiera-H	SCALE	0.624	0.649	0.520	0.480
	NNGuide	0.749	0.756	0.677	0.485
	MaxLogit	0.693	0.710	0.608	0.548
	MSP	0.746	0.752	0.698	**0.622**
	PostMax	**0.773**	0.785	0.727	0.548
	COSTARR	0.765	**0.786**	**0.743**	0.612

Table 10. OPERATIONAL OPEN-SET ACCURACY. The mean OOSA (↑) of all methods on the contaminated datasets (same experimental setup as Tab. 1), which have significant overlap with ImageNet data [3]. We additionally validated this overlap for Places [49] and SUN [43] in Figures 10, 11, 12 and 11. COSTARR (ours) is, and the best scores for each respective architecture and unknowns dataset are in **bold**.

Figure 10. DATA CONTAMINATION IN PLACES. We analyzed images from the OOD dataset Places [49] (red bordered images) compared with Imagenet training images (green bordered images). To select OOD images from Places[49], we examined the validation threshold for PostMax and COSTARR on ViT-H, then selected those images from Places which PostMax would label unknown, but COSTARR would label known. Between this figure and Figure 11, we present *every image satisfying this constraint*, hence, these images are directly responsible for the performance difference between the algorithms (in terms of OOSA). To select ImageNet training images, we examined all images from the closed-set prediction class (the known class the network predicted the unknown image was) and selected the closest one. As all the images from Places (red bordered) are treated as unknowns and the networks has seen all of the ImageNet training data (green bordered), these falsely labeled unknowns are actually hindering the evaluation, given the clear overlap between known training data and supposed unknowns. The presence of these mislabeled unknowns is consistent with NINCO's [3] observations, partially validating their claim of in-distribution dataset contamination.

Figure 11. DATA CONTAMINATION IN PLACES. Continuation of Figure 10, showing overlap between Places [49] (red bordered images) and ImageNet training data (green bordered images). The overlap between known training data and test data mislabeled as unknowns hinders evaluations, as correctly identifying a known image (which is mislabeled as unknown) will incorrectly penalize an algorithm's score.

Figure 12. DATA CONTAMINATION IN SUN. We analyzed images from the OOD dataset SUN [43] (red bordered images) compared with ImageNet training images (green bordered images). To select OOD images from SUN [43], we examined the validation threshold for PostMax and COSTARR on ViT-H, then selected those images from Places which PostMax would label unknown, but COSTARR would label known. Between this figure and Figure 13, we present *every image satisfying this constraint*, hence, these images are directly responsible for the performance difference between the algorithms (in terms of OOSA). To select ImageNet training images, we examined all images from the closed-set prediction class (the known class the network predicted the unknown image was) and selected the closest one. As all the images from SUN (red bordered) are treated as unknowns and the network has seen all of the ImageNet training data (green bordered), these falsely labeled unknowns are actually hindering the evaluation, given the clear overlap between known training data and supposed unknowns.

Figure 13. DATA CONTAMINATION IN SUN. Continuation of Figure 12, showing overlap between SUN [43] (red bordered images) and ImageNet training data (green bordered images). The overlap between known training data and test data mislabeled as unknowns hinders evaluations, as correctly identifying a known image (which is mislabeled as unknown) will incorrectly penalize an algorithm's score.