Supplementary Material: 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

1 Evaluation Metrics Cont.

Discussed in main Section 4.1 (Details).

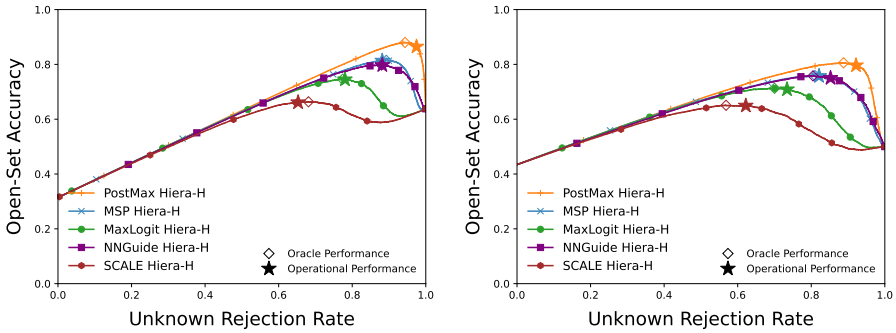
For completeness, fairness, and robustness, we also assessed methods using recent metrics designed for open-set scenarios: Area Under the Open-Set Classification Rate (OSCR) curve (AUOSCR) [2] in Tab. 1 and Open Area Under the ROC Curve (OpenAUC) [23] in Tab. 2. As explained in the main paper, neither metric offers a threshold to evaluate operational performance. Nevertheless, PreMax still outperforms all methods on both metrics/architectures.

Table 1: AUOSCR Results. The mean AUOSCR (↑) of all methods. To compute, we test each method on five different ILSVRC2012 *val* [19] 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 [20] (2) CNN ConvNeXtV2-H [24]. The best scores are in **bold**.

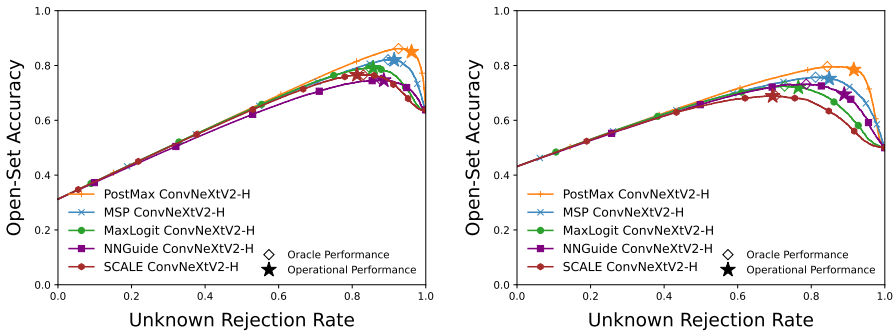
Unknowns (# imgs)	**AUOSCR ↑**									
	Hiera-H [20]					**ConvNeXtV2-H** [24]				
	PreMax (Ours)	**MSC** [9, 21]	**MaxLogit** [8, 21]	**NNGuide** [16]	**SCALE** [26]	**PreMax** (Ours)	**MSC** [9, 21]	**MaxLogit** [8, 21]	**NNGuide** [16]	**SCALE** [26]
iNaturalist [10] ($10K$)	**0.85**	0.82	0.78	0.79	0.72	**0.83**	0.81	0.79	0.76	0.77
NINCO [1] ($\sim 5.8K$)	**0.78**	0.76	0.71	0.62	0.64	**0.77**	0.75	0.73	0.64	0.71
OpenImage-O [22] ($\sim 17.6K$)	**0.84**	0.79	0.72	0.78	0.63	**0.83**	0.79	0.77	0.72	0.75
Places [27] ($10K$)	**0.80**	0.75	0.68	0.76	0.59	**0.79**	0.76	0.71	0.73	0.67
SUN [25] ($10K$)	**0.80**	0.76	0.70	0.76	0.63	**0.79**	0.76	0.72	0.73	0.69
Textures [4] ($\sim 5.1K$)	**0.83**	0.79	0.76	0.67	0.73	**0.80**	0.78	0.78	0.65	0.77
21K-P *Easy* [21] ($50K$)	**0.77**	0.72	0.65	0.72	0.58	**0.76**	0.72	0.68	0.69	0.64

Table 2: OpenAUC Results. The mean OpenAUC (↑) of all methods. To compute, we test each method on five different ILSVRC2012 *val* [19] 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 [20] (2) CNN ConvNeXtV2-H [24]. The best scores are in **bold**.

Unknowns (# imgs)	**OpenAUC ↑**									
	Hiera-H [20]					**ConvNeXtV2-H** [24]				
	PreMax (Ours)	**MSC** [9, 21]	**MaxLogit** [8, 21]	**NNGuide** [16]	**SCALE** [26]	**PreMax** (Ours)	**MSC** [9, 21]	**MaxLogit** [8, 21]	**NNGuide** [16]	**SCALE** [26]
iNaturalist [10] ($10K$)	**0.98**	0.95	0.91	0.92	0.85	**0.97**	0.95	0.93	0.90	0.91
NINCO [1] ($\sim 5.8K$)	**0.91**	0.89	0.84	0.75	0.77	**0.91**	0.89	0.87	0.78	0.85
OpenImage-O [22] ($\sim 17.6K$)	**0.97**	0.92	0.85	0.91	0.76	**0.96**	0.93	0.91	0.86	0.89
Places [27] ($10K$)	**0.93**	0.89	0.81	0.89	0.72	**0.93**	0.89	0.85	0.87	0.81
SUN [25] ($10K$)	**0.93**	0.89	0.83	0.89	0.76	**0.93**	0.90	0.86	0.87	0.82
Textures [4] ($\sim 5.1K$)	**0.96**	0.92	0.90	0.80	0.86	**0.94**	0.92	0.92	0.79	0.91
21K-P *Easy* [21] ($50K$)	**0.91**	0.85	0.78	0.85	0.71	**0.90**	0.86	0.81	0.83	0.78

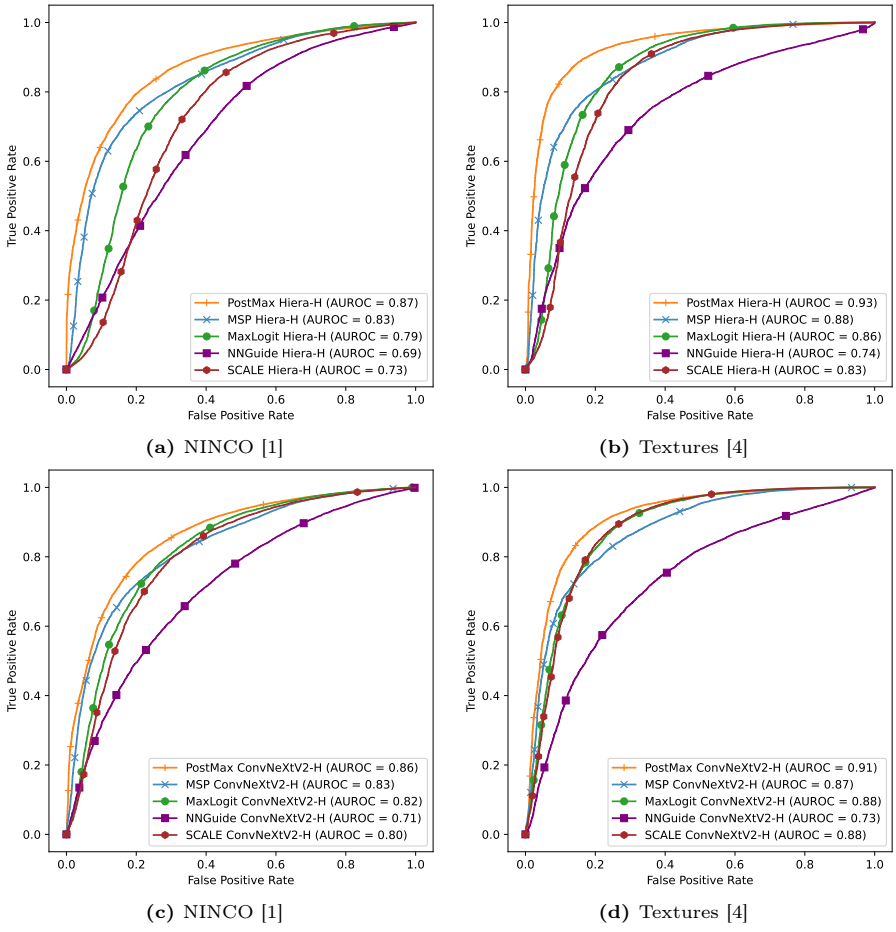


(a) OpenImage-O [22] **(b)** SUN [25]



(c) OpenImage-O [22] **(d)** SUN [25]

Fig. 1: Open-Set Accuracy Curves. The Open-Set Accuracy curves of all methods on two unknown datasets from Table 1 in the main paper. OSR is performed on extractions from the same pre-trained architectures – state-of-the-art Hiera-H and ConvNeXtV2-H. ⋆ signifies the peak performance (OOSA) of each method.



(a) NINCO [1] **(b)** Textures [4]

(c) NINCO [1] **(d)** Textures [4]

Fig. 2: ROC Curves. The Receiver Operating Characteristic curves of all methods on two unknown datasets from Table 2 in the main paper. OSR is performed on extractions from the same pre-trained architectures – state-of-the-art Hiera-H and ConvNeXtV2-H.

2 Ablation Study Cont.

Discussed in main Section 4.3 (Ablation Study).

2.1 Varying α

We also explore the impact of varying the application-dependent α from Eq. 2 in the main paper. In high-risk applications, operators may prioritize a high rejection rate over correct classification to ensure the exclusion of harmful samples ($\alpha = 0.17$). In low-risk applications, accuracy maximization becomes a higher priority ($\alpha = 0.50$). Therefore, in Tab. 3, we varied α to analyze different operational

Table 3: VARYING α. The mean and standard deviation of Operational Open-Set Accuracy (OOSA) (↑) across all datasets (5 ILSVRC2012 *val* [19] splits and various unknowns). OSR performed is on extractions from a pre-trained architecture on ILSVRC2012-1K - (1) Transformer Hiera-H [20]. The best performance is in **bold**.

α	**PostMax** (Ours)	**MSP** [9, 21]	**MaxLogit** [8, 21]	**NNGuide** [16]	**SCALE** [26]
0.67	**0.77** ± .10	0.74 ± .07	0.68 ± .10	0.69 ± .11	0.61 ± .12
0.50	**0.79** ± .07	0.76 ± .05	0.71 ± .07	0.70 ± .09	0.65 ± .08
0.29	**0.76** ± .04	0.73 ± .04	0.53 ± .16	0.66 ± .07	0.53 ± .16
0.20	**0.69** ± .07	0.53 ± .16	0.53 ± .16	0.62 ± .09	0.53 ± .16
0.17	**0.66** ± .09	0.53 ± .16	0.53 ± .16	0.60 ± .10	0.53 ± .16

requirements. Note, varying α effectively evaluates methods on different thresholds. Compared to other methods, PostMax demonstrates a higher average OOSA, lower variance, and greater stability across different thresholds.

2.2 Additional Architectures

To further showcase robustness, we present two other architectures: Meta's ViT-H [7] and ConvNeXt-L [13]. The experimental details are the same as the experiments in the main paper. From Tab. 4, it is clear PostMax maintains performance against MSP [9,21], MaxLogit [8,21], NNGuide [16], and SCALE [26] in Operational Open-Set Accuracy (OOSA) across other architectures.

Table 4: ADDITIONAL ARCHITECTURE RESULTS. The mean OOSA (↑) of all methods. To compute, we validate methods on ImageNetV2 [17] ($10K$ images) as knowns and 20% of our surrogate 21K-P *Hard* [21] ($9.8K$ images) as unknowns and predict an operational threshold. Then, we deploy each method's threshold and test on five different ILSVRC2012 *val* [19] 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 pre-trained architectures on ILSVRC2012-1K - (1) Transformer ViT-H [7] (2) CNN ConvNeXt-L [13]. The best scores are in **bold**.

Unknowns (# **images**)	**OOSA** ↑									
	ViT-H [7]					**ConvNeXt-L** [13]				
	PostMax (Ours)	**MSC** [9, 21]	**MaxLogit** [8, 21]	**NNGuide** [16]	**SCALE** [26]	**PostMax** (Ours)	**MSC** [9, 21]	**MaxLogit** [8, 21]	**NNGuide** [16]	**SCALE** [26]
iNaturalist [10] ($10K$)	**0.826**	0.794	0.788	0.636	0.748	**0.794**	0.773	0.740	0.583	0.683
NINCO [1] ($\sim 5.8K$)	**0.732**	0.717	0.711	0.552	0.671	0.698	**0.701**	0.682	0.447	0.638
OpenImage-O [22] ($\sim 17.6K$)	**0.861**	0.809	0.755	0.661	0.706	**0.830**	0.791	0.728	0.662	0.664
Places [27] ($10K$)	**0.780**	0.741	0.695	0.699	0.640	**0.756**	0.725	0.654	0.621	0.588
SUN [25] ($10K$)	**0.792**	0.749	0.717	0.696	0.672	**0.771**	0.734	0.674	0.623	0.617
Textures [4] ($\sim 5.1K$)	**0.761**	0.727	0.742	0.544	0.718	**0.722**	0.706	0.693	0.450	0.659
21K-P *Easy* [21] ($50K$)	**0.810**	0.770	0.675	0.543	0.626	**0.794**	0.739	0.624	**0.794**	0.552

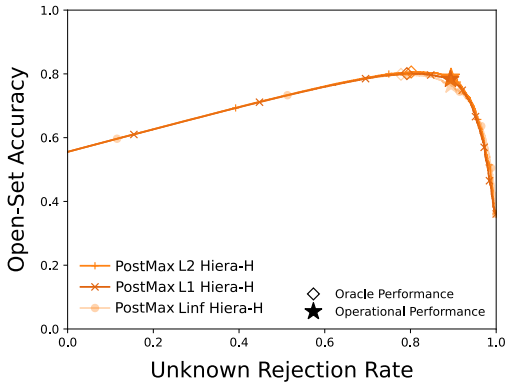


Fig. 3: Normalization Ablation. The Open-Set Accuracy curve of PostMax with L_1, L_2, and L_∞ on Hiera-H [20] with ILSVRC2012 [19] *val* as knowns and Textures [4] as unknowns.

2.3 PostMax Norm

In Fig. 3 we show PostMax performance is invariant regardless of different normalizations. We showcase this on the state-of-the-art Hiera-H [20].

3 Normalizations

Discussed in main Section 3.2 (Feature Norms).

In [6], it is argued that unknowns have smaller norms, and similar observations are made in [11, 15, 21]—in other words, deep feature magnitudes of unknown samples tend to be *smaller* than those of knowns. However, these studies were confined to a few classes, while our findings indicate the opposite for modern ImageNet-scale networks. If, for a particular network, a preliminary test reveals that unknowns have smaller norms, our normalization method's division could be substituted with multiplication or another operation that increases with the magnitude.

We hypothesize that the smaller magnitudes observed in datasets with fewer classes may be due to the lower likelihood of random features combining to produce a significant response in the final network. In contrast, larger networks with more classes often experience confusion with unknowns due to many small features adding up to sufficient magnitude to stimulate class responses.

While the study in [3] observes an increase in norm values for unknowns, it pertains to the norms of backpropagated gradients. In contrast, our observations concern the norms of forward pass features. Additionally, while [15] presents intriguing theorems on feature norms, these are based on assumptions that may not hold true for larger networks or a higher number of classes.

4 Datasets Cont.

Discussed in main Section 4.1 (Details).

Works in Open-Set Recognition (OSR) and Out-Of-Distribution (OOD) detection commonly use the same dataset for validation and testing. Consequently, thresholds and parameters in algorithms are often tuned based on the test set. However, such an approach deviates from real-world operational usage. During testing, unknowns should not have been encountered during validation. It is paramount for engineers to utilize an unknown set that is representative of what their system expects to encounter, which we term *unknowns surrogate*. The surrogate set should act as a substitute for expected unknowns, providing engineers with the capability to predict operational performance.

ILSVRC2012 [19]. The dataset consists of $1K$ object categories encountered in the real world and is commonly used as a large-scale benchmark for computer vision research. It is based on the original ImageNet database [5], which was organized according to the WordNet hierarchy [14] with over $14M$ annotated images. The *train* split, comprising $\sim 1.2M$ images, is exclusively used for knowns during training. The *val* split, containing $50K$ images, is exclusively used for knowns during testing. Note, while ILSVRC2012 has a *test* split, it lacks the labels necessary for performance evaluation. We aimed to compute the mean and standard deviation during testing, so we divided the *val* set into five equal subsets, each consisting of $10K$ images. Importantly, each subset still encompasses $1K$ classes.

ImageNetV2 [17]. The dataset comprises three distinct $10K$ images splits, sampling the same categories as ILSVRC2012 [19]. These splits were meticulously curated with the aim of closely replicating the distribution of the original *val* set. Each split employs a unique sampling strategy implemented by Amazon Mechanical Turk workers. Notably, during validation, we selected *TopImages* ($10K$ images) as our knowns, given its highest selection frequency among the workers.

ImageNet-21K-P Open-Set splits [21]. Recently, the large-scale ImageNet-21K-P - Winter21 [18] dataset released, a subset of the ImageNet database [5] that removed small classes, resulting in approximately $11K$ categories. Thus, Vaze *et al.* [21] leveraged the subset and constructed two sets of 1000 classes. The curation process was based on the sorted total semantic distance between ImageNet-21K-P - Winter21 and ILSVRC2012, providing a measure of open-set difficulty. The easier (larger semantic novelty) split resulted in *Easy* ($50K$ images), while the more

difficult (smaller semantic novelty) resulted in *Hard* ($49K$ images). Importantly, neither split contains any overlap with ILSVRC2012. During validation, we use 20% of ImageNet-21K-P *Hard* ($9.8K$ images) as surrogate unknowns, while the *Easy* split is entirely reserved for unknowns during testing. Our decision to use *Hard* as surrogate unknowns was driven by its difficulty.

iNaturalist [10]. The dataset comprises over $859K$ images encompassing more than $5K$ different species of plants and animals. It distinguishes itself from other vision datasets by being unbiased and more representative of real-world scenarios. We exclusively consider 110 classes that do not overlap with ILSVRC2012 [19], sampling $10K$ images from these classes for use as unknowns during testing.

NINCO [1]. Bitterwolf *et al.* [1] discovered that recent large-scale datasets exhibit categorical contamination, where labels coincide with a class or act as a subset of a class, as well as incidental contamination, where objects are found in the background or are an aspect of another class. In response, they curated the No ImageNet Class Objects (NINCO) dataset, which comprises $5,879$ images from 64 object categories. Each image was individually checked to ensure it does not contain any objects found in [19]. We utilize the entire dataset as unknowns during testing.

OpenImage-O [22]. Inspired by the shortcomings of existing datasets, particularly the unreliability of creating datasets through tag queries or the absence of human inspection to confirm validity, Wang *et al.* [22] curated the OpenImage-O dataset. This dataset comprises $17,632$ images and is a subset of the Open Images Dataset V3 [12] *test* set. The manually annotated subset boasts a large-scale and naturally diverse distribution. We employ the entire dataset as unknowns during testing.

Places [27]. The database comprises over $10M$ images captured from 434 scenes, each carefully selected to represent 98% of the various types of places humans might encounter worldwide. Throughout experimentation, the curators focus exclusively on the Places365-Standard subset. From this subset, only 50 classes, distinct from those in ILSVRC2012 [19], are considered. We then sample $10K$ images from these classes to serve as unknowns during testing.

SUN [25]. The scene understanding database comprises over $130K$ images collected from 899 environments, each offering a snapshot of diverse and rich

daily encounters. In the course of experimentation, the curators narrowed their focus to only 397 classes. Among these, we specifically selected 50 classes that do not overlap with ILSVRC2012 [19]. From this subset, we sampled $10K$ images to serve as unknowns during testing.

Textures [4]. The dataset comprises $5,640$ images, each representing one of 47 texture attributes. Captured *in the wild*, these images aim to provide the best representation for recognizing describable texture attributes. Recently, Wang *et al.* [22] removed four categories (*bubbly*, *honeycombed*, *cobwebbed*, *spiralled*) that overlapped with ILSVRC2012. The resulting $5,160$ images are exclusively employed as unknowns during testing.

References

1. 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)
2. 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)
3. Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A.: Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: International conference on machine learning. pp. 794–803. PMLR (2018)
4. 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)
5. 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)
6. Dhamija, A.R., Günther, M., Boult, T.: Reducing network agnostophobia. Advances in Neural Information Processing Systems (NeurIPS) **31** (2018)
7. 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)
8. 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)
9. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. International Conference on Learning Representations (ICLR) (2017)
10. 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)
11. 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)

12. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., et al.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github.com/openimages (2017)
13. 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)
14. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
15. 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)
16. 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)
17. 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)
18. Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik, L.: Imagenet-21k pretraining for the masses. In: Vanschoren, J., Yeung, S. (eds.) Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1 (2021)
19. 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)
20. 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)
21. 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)
22. 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)
23. 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)
24. 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)
25. 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)
26. 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)
27. 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)