IN OR OUT? FIXING IMAGENET OUT-OF-DISTRIBUTION DETECTION EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) detection is the problem of identifying inputs which are unrelated to the in-distribution task. The OOD detection performance when the in-distribution (ID) is ImageNet-1K is commonly being tested on a small range of test OOD datasets. We find that most of the currently used test OOD datasets have severe issues, in some cases more than 50% of the dataset contains objects belonging to one of the ID classes. These erroneous samples heavily distort the evaluation of OOD detectors. As a solution, we introduce with NINCO a novel test OOD dataset, each sample checked to be ID free, which with its fine-grained range of OOD classes allows for a detailed analysis of an OOD detector's strengths and failure modes, particularly when paired with a number of synthetic "OOD unit-tests". We provide detailed evaluations across a large set of architectures and OOD detection methods on NINCO and the unit-tests, revealing new insights about model weaknesses and the effects of pretraining on OOD detection performance. We provide code and data here.

1 INTRODUCTION

While deep learning based models have shown impressive performance on many real world tasks, they often exhibit unforeseen behaviour when confronted with unknown situations like receiving an input that is not related to the task it has been trained on. Such samples are regarded as out-of-distribution (OOD) and deep neural network classifiers are known to make very confident predictions that those belong to one of the **in-distribution (ID)** classes (Hendrycks & Gimpel, 2017; Hein et al., 2019). This unwanted behaviour is a serious obstacle when applying classifiers in real world applications. The purpose of OOD detectors is to reject OOD inputs, which depending on the application can mean requesting human intervention, steering towards a safe state, or simply abstaining from making a prediction, while at the same time letting ID inputs pass through.

Current OOD detection evaluations in image classification rely on the assumption that there is no ID class present in an OOD test image, not even in the background. We follow this definition and consider an input to be **out-of-distribution (OOD)** if it does not contain any of the in-distribution classes. However, we show that this assumption is not fulfilled for most of the current test OOD datasets for ImageNet-1K (IN-1K) of Russakovsky et al. (2015). We demonstrate that occurrences of objects from ID classes in these test OOD datasets are often correctly recognized by state-of-the-are OOD detectors, but as an unwarranted consequence held against them as mistakes in OOD detection evaluations (false "false positive"). Even in cases where current models struggle to identify ID content, e.g. if ID objects are partially occluded or in the background, OOD datasets containing ID objects are not future proof: when evaluating on them, one would not realize if a future model correctly predicts the class of a visible ID object.

The erroneous occurrences of ID objects in existing OOD datasets can be characterized into two failure modes, which we illustrate in Figure 1 and define as follows. **Categorical ID contaminations** show objects from ID classes which already are classes in a base dataset from which the test OOD dataset has been built. Their label coincides with an ID class or semantically designates a subset of an ID class, e.g. the class *hayfield* from the PLACES datset and the IN-1k class *hay*. **Incidental ID contaminations** on the other hand occur in images which are supposed to belong to an OOD category but which contain an ID object. The object can be in the background or a feature of the specific instance of the shown main object, e.g. the IN-1k class *plane* in an image of the OOD category *sky*.

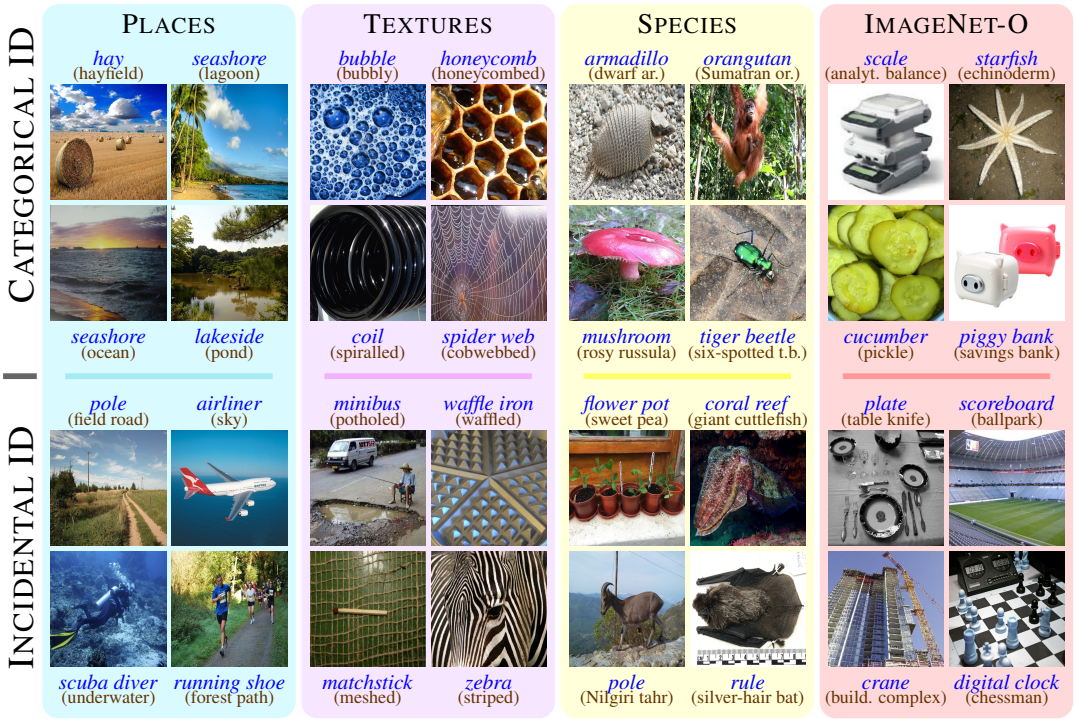

Figure 1: **Contamination of OOD test sets with ID samples (ImageNet).** *Blue:* ImageNet-1K class found in the image. (Brown): Label of the image in the original source dataset. **Top:** Samples from classes of the OOD dataset that by class meaning categorically overlap with ImageNet-1K classes. **Bottom:** Labels alone do not reveal that the images are ID, but incidental ID objects can be found.

We show that ID contaminations strongly impact the conclusions which can be drawn from evaluating OOD detection methods by (1) systematically underestimating the true OOD detection performance and (2) unrightfully punishing stronger OOD detectors.

Probing the true performance of OOD detectors for IN-1K requires a range of OOD classes that are challenging, diverse, and most importantly actually OOD. Compiling a test OOD dataset is indeed a challenging task, as the 1000 classes of IN-1K cover a fair portion of the images found in general image datasets. In this paper we introduce the **NINCO (No ImageNet Class Objects) dataset** which contains 4 293 images that we individually checked not to contain any ID object from the classes in IN-1K. These images are ordered into 46 OOD classes, which facilitates a specific analysis of the failure modes of an OOD detector. Additionally, we provide a dataset of **"OOD unit-tests"**, synthetic images which do not resemble real world photos, but are designed to test specific weaknesses that might have impact in real-world applications (e.g. due to a camera failure). We find that surprisingly many OOD detectors struggle to detect these supposedly easy unit-tests, in particular methods that work well on natural test data.

We provide a detailed OOD detection evaluation on NINCO for a range of eleven OOD detection methods across a large number of architectures and training schemes. Surprisingly, it turns out to be difficult for many OOD detectors to improve consistently over the baseline of Maximum Softmax Probability (MSP). While we confirm the observation that pretraining on larger datasets generally helps OOD detectors and particularly methods relying on feature-information, we find that the type of pretraining has a strong impact.

2 PREVALENCE OF ID SAMPLES IN POPULAR OOD DATASETS

In Appendix B we give an overview of the datasets that have been used to evaluate OOD detection performance for IN-1K as ID. In the following we use *blue* for the name of an ImageNet class and *brown* for the category name in the source dataset used for the generation of the test OOD dataset. Concerningly, several test OOD datasets for IN-1K that are in use by the community contain a substantial fraction of samples that show ID objects. Figure 1 shows some typical appearances of ID

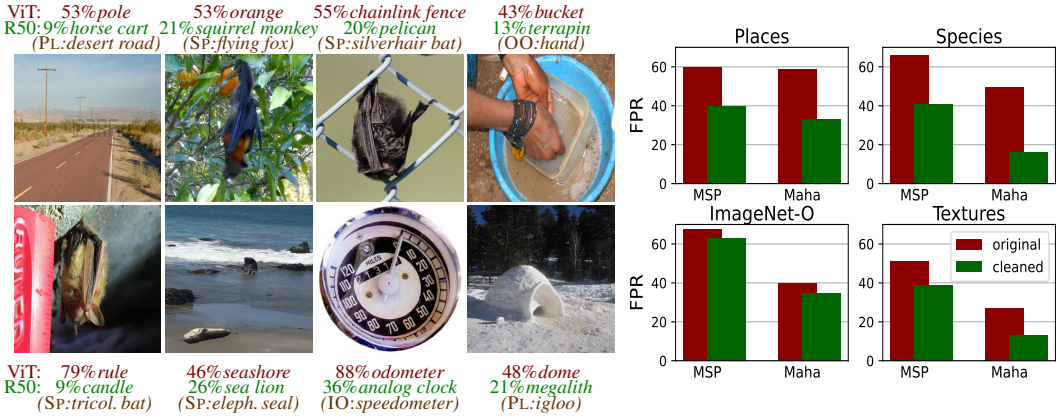

Figure 2: **Left: A Vision Transformer confidently classifies ID objects in samples from popular OOD datasets** *(source label in parentheses)* as the correct IN-1K class, but is marked down with false positives in OOD detection evaluation when using MSP (Max Softmax Prob.) as criterion. The weaker **ResNet-50**, in contrast, doesn't recognize the ID objects and hence the MSP is low enough to reject all images wrongly as OOD. This illustrates how a better model (ViT in our case) can be unjustly punished when the test OOD dataset contains ID objects. For both models, the 95%TPR threshold is at a MSP of 38%. **Right: OOD-detection before and after removing samples with ID-objects:** We show FPR (lower is better) of two OOD detectors (MSP and Mahalanobis distance) for a ViT, evaluated on cleaned and full subsets of four popular OOD datasets.

data in supposedly OOD datasets. The categorical ID failure mode illustrated in the top part is the inclusion of samples from explicitly ID classes of the source dataset from which the OOD dataset has been built. For instance, the class *hayfield* from the PLACES-dataset overlaps with the IN-1K class *hay*. However, also in principally innocuous classes (bottom part), many incidental ID samples can still be found. Here, the occurring failure modes are numerous: some ID objects happen to be in the background, some are a prominent part of the depicted scene, and some happen to realize both the original class and the ID class. For instance, the class *table knife* contains samples which also show a *plate*, and the class *striped* from the TEXTURES-dataset often shows the stripes of a *zebra*.

In order to quantify the severity of ID objects in test OOD datasets, we manually check for ID objects in 400 random samples from each of the most commonly used datasets. For fair treatment, unclear and ambiguous samples, which we would exclude from NINCO introduced below, are ignored in this survey. The results in Table 11 show that for many of these common OOD detection benchmarks, a substantial fraction of samples is actually ID: For both the PLACES and SPECIES datasets, it is more than 50%. Only iNATURALIST OOD PLANTS (2.5% of samples ID) and OPENIMAGE- (4.9% ID) contain comparably few ID images.

2.1 EFFECT OF ID CONTAMINATION ON OOD EVALUATION

In Figure 2 (left), we show how OOD detection evaluation with incidental ID samples can unrightfully punish strong OOD-detectors: A better model can correctly recognize ID objects with high confidence even if they are in the background of the image, leading to a false "false positive" in the evaluation, while a weaker model not recognizing the ID object and providing a low-confidence prediction is "rewarded" with a false "true negative". For example, the strong VisionTransformer (ViT) (Dosovitskiy et al., 2021) identifies the *pole* besides an otherwise empty desert road, and thus has high confidence on the image where the weaker ResNet-50 does not recognize any ID class with high confidence. Similarly, in the second example, the ViT is punished with a false "false positive" for recognizing (above the detection threshold) the *oranges* in the background while ignoring the unknown flying fox (truly OOD), whereas the ResNet-50 even does predict a wrong ID class, namely *squirrel monkey*, but does so with low confidence (below the detection threshold), and is thus rewarded with a false "true negative".

We quantify the effect of ID contaminations on evaluation results in customary OOD datasets in Figure 2 (right) for the MSP baseline and the Mahalanobis OOD detection method (Lee et al., 2018). For the test OOD datasets which showed a large portion of ID samples in Table 11, we report

Figure 3: **Difficult OOD classes in NINCO**: Examples of images from some of NINCO's most difficult (see Table 4) *OOD classes* (first row) and from the *ImageNet-1K class* (second row) which the *OOD class* is most frequently confused for.

the FPR at 95% TPR obtained with a ViT when evaluating on the original 400 samples and our cleaned subsample of it not containing any more ID objects (detailed results for a range of models and methods can be found in Appendix K). We find that ID contaminations strongly impact the conclusions which can be drawn from evaluating OOD detection methods on those datasets. Most clearly, both methods perform substantially *better* after removing the images with ID objects from the OOD datasets, in some cases reducing the FPR by more than 50%. This is unsurprising: If a significant fraction of the dataset is actually ID, this fraction should not be detected as OOD by a well-performing method. Hence, evaluating OOD detection performance with partially ID data leads to a systematic *overestimation* of the true FPR of the OOD detection method and disadvantages better models as they are more likely to detect ID objects as discussed above. Additionally, we observe that the differences between OOD detectors become more pronounced. In Figure 2 (right) it can be seen that for each dataset, the FPR for the Mahalanobis OOD detector decreases more than for the MSP-baseline. The effect is particularly strong for SPECIES (25.5% gain of MSP vs. 33.2% gain of Mahalanobis) and PLACES (20.1% gain vs. 25.7% gain), which are the two datasets we found to contain most ID samples.

3 A NEW OOD TEST SET FOR IMAGENET-1K

As discussed in Section 1, an **OOD input for IN-1K** is an image that does not contain an object from one (or several) of the 1 000 IN-1K classes. These ImageNet classes are based on individual WordNet (Fellbaum, 1998) synsets, each consisting of one or more keywords that are synonymous in some context. During the ImageNet creation process (Deng et al., 2009), images were first collected from the web by using variations of each keyword of a respective class and then verified by humans to fit its synset's definition. In reality, many ImageNet samples fit one but not necessarily *all* keywords of their class label. This means that to make sure that OOD detectors are treated fairly[1], OOD test samples cannot fall into the definition of any keyword of any IN-1K class. For example, photos of the Sumatran orangutan cannot be considered OOD, since they could be included in the IN-1K class *(orangutan, orang, orangutang, Pongo pygmaeus)*, even though *Pongo pygmaeus* only refers to the Bornean orangutan. To determine what counts as an ID object, we follow the WordNet glosses (one can look up synsets with glosses here). as well as dictionary definitions of keywords and source dataset class labels. For difficult cases, we consult additional sources like Wikipedia. For example, the species *northern elephant seal* does not fall into the ID class *sea lion*, among other biological criteria distinguished by the fact that the former do not have ears while the latter do. An image of an OOD dataset can furthermore not incidentally contain ID objects, to avoid cases as in Figure 1 (bottom) and Figure 2. We give an overview of the problem of OOD detection including definitions like that of FPR, as well as state-of-the-art detection methods in Appendix D.

3.1 NINCO DATASET CONSTRUCTION

For each OOD class of our new NINCO dataset, we start by **choosing a base class** which consists of all samples from a named class of an existing or newly scraped dataset. The majority of the NINCO base classes are sourced from SPECIES (Hendrycks et al., 2022), which provides images scraped from

[1]For fair treatment of previous OOD *datasets*, such unclear samples that don't fit all keywords were ignored in Table 11.

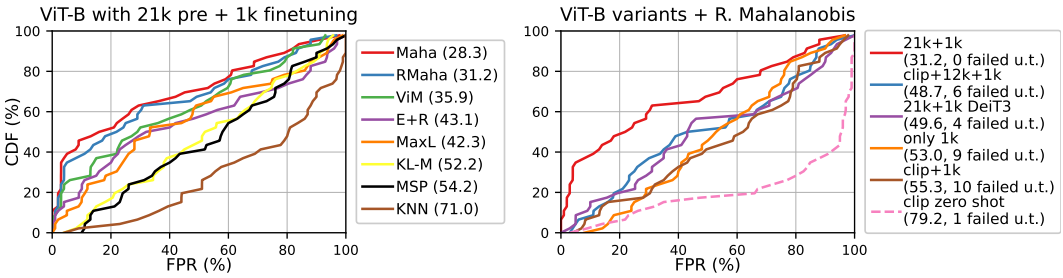

Figure 4: Cumulative distribution of the % of NINCO-classes for which an FPR at least as low as a given x-value is achieved. The area over this curve corresponds to the mean FPR. The further in the top left corner, the better. **The best methods leverage features (Left):** Different OOD detection methods with a ViT-B pretrained on IN-21k (mean FPR in parentheses). **Not all pretraining helps (Right):** RMaha applied to ViT-B with different training variants (MCM for CLIP zero-shot). Only the top model does not fail OOD unit-tests.

iNaturalist. For each base class, we carefully decide, based on WordNet glosses, iNaturalist taxonomy details and Wikipedia, whether it can be included according to the non-permissive interpretation described at the beginning of Section 3. The choice of base classes is not random, since there is no way to randomly sample from the set of concepts that might occur at test time. Rather, we aim for a variety of classes that are challenging, diverse and, most importantly, not actually being categorically ID to begin with. Then for each base class, we **individually inspect each image** for ID objects. To help remembering the 1000 ID classes, we display the 5 top ID classes of a ViT's prediction on each image. If an ID object is at least partially visible, the corresponding sample is removed. In cases where it is ambiguous whether we see an ID object in the image, the sample is not included in the cleaned dataset. As the iNaturalist data (including the SPECIES dataset) has been curated by experts and can be considered very reliable, we generally trust in the main object belonging to the species it is labelled as. For data from PLACES and FOOD-101, we consider ourselves competent to verify whether the label is correct. In addition to samples showing ID objects, we also remove images where no object from the OOD class is visible, e.g. we exclude pictures of animal traces or remains which frequently appear in iNaturalist. While for most existing datasets, the cleaning has been outsourced to external services like Amazon Mechanical Turk or student labellers. By researching all OOD classes and visually inspecting all their samples ourselves, we as authors of NINCO were able to do more in-depth research for each ambiguous case and obtain more coherent decisions, which we are positive leads to a higher quality dataset.

The NINCO (**N**o **I**mageNet **C**lass **O**bjects) dataset consists of 46 OOD classes with a total of 4 293 samples. The base classes which we cleaned to obtain NINCO were sourced from SPECIES (35 classes) (Hendrycks et al., 2022), PLACES (3 classes) (Zhou et al., 2017), which both are discussed in Section 2 and Appendix B, as well as from the FOOD-101 dataset (6 classes) (Bossard et al., 2014) and newly scraped from iNaturalist (2 classes) (Van Horn et al., 2018). Details for all NINCO OOD classes are given in Appendix G. We show samples from all NINCO classes in Figures 8 and 9 in Appendix I. In addition to NINCO, we also provide the 1673 OOD images obtained from cleaning 400 samples of six test OOD datasets as discussed in Section 2.1. In order to notice ID contaminations potentially biasing the drawn conclusions, we recommend to also evaluate on these cleaned versions when evaluating on those original benchmarks.

3.2 OOD UNIT-TESTS

Following common practice (e.g. Hendrycks et al. (2022)), we argue that evaluating an OOD detector on a range of simple, synthetic classes *besides* the variably challenging natural image classes of an OOD dataset can give additional insights about its OOD detection weaknesses. Example images and reproducibility details for all 17 pre-existing and newly proposed OOD unit-tests are included in Appendices H and I. Since these **OOD unit-tests** do not represent a diverse distribution of photos, but different modes of simple, synthetically generated image inputs which any good OOD detector should be expected to detect, we don't include them in summary metrics or distribution plots. Instead, we suggest to count an OOD unit test as **failed** if a method has an FPR above a user-defined threshold, which we suggest setting at 10%, and to report the number of *failed* OOD unit-tests (which should be 0 for a strong OOD detector) alongside the aggregate results on a test OOD dataset like NINCO.

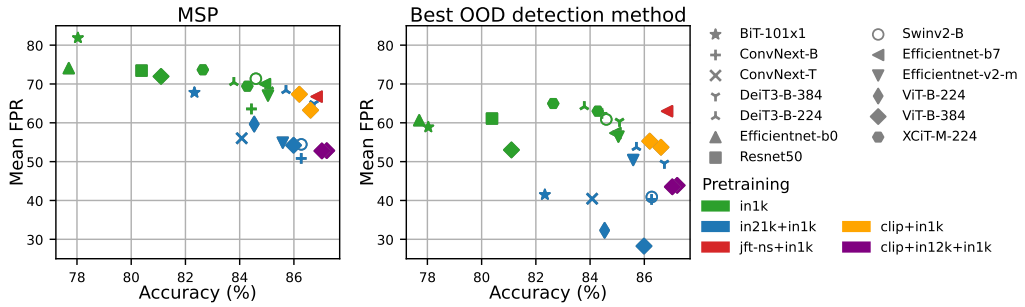

Figure 5: **IN-21K pretraining boosts feature-dependent OOD detectors on NINCO:** Mean FPR vs. accuracy for MSP and each model's best detector (which is always feature-dependent). OOD detection strongly improves when using models pretrained on IN-21K. Additional CLIP-pretraining or on JFT can yield higher accuracy, but OOD detection need not be better than with IN-21K pretraining.

For each OOD unit-test, we provide a set of 400 samples in typical ImageNet format, by mirroring the sizes and file formats of random ImageNet samples. While some OOD unit-tests may appear redundant at first sight, we find that they provide important information as some detectors e.g. mostly pass the *monochrome* test but completely fail on *black*, which reveals a specific weakness that is very realistic to be encountered in practice.

4 Overview of Evaluation Results for OOD Detectors

In Appendix A, we conduct an in-depth analysis of the evaluation results for eleven state-of-the-art OOD-detection methods (described in Appendix D) combined with a wide range of models (described in Appendix C). We summarize the main take-aways here.

- For most models, like for the ViT in Figure 4 (left), methods that leverage the features of the model's penultimate layer out-perform methods purely based on quantities derived from the logits.
- Many otherwise strong OOD detectors struggle to distinguish supposedly easy OOD unit-tests from ID-data. Concerningly, many detectors are even fooled by completely *black*, *grey* or *white* images, which are for many systems likely to be recorded in practice.
- Mahalanobis distance (Lee et al., 2018) on ViT features is the single best detector in terms of mean FPR and also passes all unit-tests, while Relative Mahalanobis distance (Ren et al., 2021) shows the most consistently good results across models.
- Results on the cleaned subsets of six previous benchmark vary, but overall lead to similar conclusions as NINCO.
- Whether and how a model was pretrained (before fine-tuning on IN-1K) has a strong effect on OOD detection performance: models using 'traditional' pretraining on IN-21K have a large advantage (even compared to higher accuracy models pre-trained on CLIP), as can be observed in Figures 4 (right) and 5.
- We notice no substantially different benefit of IN-21K pre-training for detecting OOD classes that overlap with with IN-21K classes.
- Inspecting the behaviour on individual OOD classes, which the NINCO dataset facilitates, helps understanding the specific weaknesses and strengths of a considered OOD detector.

5 Conclusions

We introduce with NINCO a novel, ID-contamination-free and challenging OOD test-dataset for IN-1K with fine-grained class-resolution. We find that many OOD detectors work better than previously thought, when their recorded number of undetected OOD inputs is not inflated by ID contaminations. However, most detection methods cannot reliably be applied with arbitrary classifier models, as even OOD unit-tests are failed by many combinations. We are hopeful for NINCO and the cleaned test OOD subsets to facilitate the more precise development of reliable OOD detectors which do not try to avoid presumed failures which are actually correct decisions.

REFERENCES

Anonymous. A framework for benchmarking class-out-of-distribution detection and its application to imagenet. In *Submitted to The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Iuubb9W6Jtk. under review.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – mining discriminative components with random forests. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision – ECCV 2014*, pp. 446–461, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10599-4.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *CVPR*, 2014.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Fellbaum, C. Wordnet: An electronic lexical database. *Cambridge, MA: MIT Press*, 1998.

Fort, S., Ren, J., and Lakshminarayanan, B. Exploring the limits of out-of-distribution detection. In *NeurIPS*, 2021. URL https://openreview.net/forum?id=j5NrN8ffXC.

Hein, M., Andriushchenko, M., and Bitterwolf, J. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*, 2019.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. In *ICLR*, 2019. https://github.com/hendrycks/outlier-exposure.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *CVPR*, 2021.

Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., and Song, D. Scaling out-of-distribution detection for real-world settings. In *ICML*, 2022.

Huang, R. and Li, Y. Mos: Towards scaling out-of-distribution detection for large semantic space. In *CVPR*, 2021.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Big transfer (bit): General visual representation learning. In *ECCV*, 2020. URL https://doi.org/10.1007/978-3-030-58558-7_29.

Koner, R., Sinhamahapatra, P., Roscher, K., Günnemann, S., and Tresp, V. Oodformer: Out-of-distribution detection transformer. *arXiv preprint arXiv:2107.08976*, 2021.

Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Kamali, S., Malloci, M., Pont-Tuset, J., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., and Murphy, K. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from storage.googleapis.com/openimages/web/index.html*, 2017.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020.

Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., and Li, Y. Delving into out-of-distribution detection with vision-language representations. In *NeurIPS*, 2022. URL https://openreview.net/forum?id=KnCS9390Va.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Ren, J., Fort, S., Liu, J., Roy, A. G., Padhy, S., and Lakshminarayanan, B. A simple fix to mahalanobis distance for improving near-ood detection, 2021. URL https://arxiv.org/abs/2106.09022.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015.

Salehi, M., Mirzaei, H., Hendrycks, D., Li, Y., Rohban, M. H., and Sabokrou, M. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*, 2021.

Song, Y., Sebe, N., and Wang, W. Rankfeat: Rank-1 feature removal for out-of-distribution detection. *arXiv preprint arXiv:2209.08590*, 2022.

Steiner, A. P., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. How to train your vit? data, augmentation, and regularization in vision transformers. *TMLR*, 2022. URL https://openreview.net/forum?id=4nPswr1KcP.

Sun, Y., Guo, C., and Li, Y. React: Out-of-distribution detection with rectified activations. *NeurIPS*, 2021.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. *ICML*, 2022.

Tack, J., Mo, S., Jeong, J., and Shin, J. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *NeurIPS*, 2020.

Techapanurak, E., Suganuma, M., and Okatani, T. Hyperparameter-free out-of-distribution detection using cosine similarity. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Touvron, H., Cord, M., and Jegou, H. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *CVPR*, 2018.

Wang, H., Liu, W., Bocchieri, A., and Li, Y. Can multi-label classification networks know what they don't know? *NeurIPS*, 2021.

Wang, H., Li, Z., Feng, L., and Zhang, W. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, 2022.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Xia, G. and Bouganis, C.-S. On the usefulness of deep ensemble diversity for out-of-distribution detection. *arXiv preprint arXiv:2207.07517*, 2022.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with noisy student improves imagenet classification. *arXiv preprint arXiv:1911.04252*, 2019.

Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., et al. Openood: Benchmarking generalized out-of-distribution detection. *arXiv preprint arXiv:2210.07242*, 2022.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464, 2017.

A EVALUATION RESULTS FOR OOD DETECTORS

We evaluate a range of IN-1K models obtained from the public timm-library (Wightman, 2019) and state-of-the-art OOD-detection methods on NINCO. We focus on transformer architectures and convolutional networks, both with and without pretraining. While most pretrained models were initially trained on IN-21K, we also include an EfficientNet trained via noisy student (Xie et al., 2019) on the JFT-300M dataset, and four ViTs with CLIP-pretraining (Radford et al., 2021) and subsequent fine-tuning, as well as a zero-shot CLIP model. A detailed description of all models can be found in Appendix C. We investigate the following commonly used OOD detection methods, which can be grouped into two categories: Max-Softmax (MSP) (Hendrycks & Gimpel, 2017), Max-Logit (Hendrycks et al., 2022), Energy (Liu et al., 2020) and KL-Matching (Hendrycks et al., 2022) derive an OOD-score exclusively from logit outputs, whereas Mahalanobis distance (Maha) (Lee et al., 2018), Virtual Logit Matching (ViM) (Wang et al., 2022), ReAct (Sun et al., 2021), Relative Mahalanobis distance (RMaha) (Ren et al., 2021), and K-Nearest-Neighbours (KNN) (Sun et al., 2022) also leverage information from the features of the DNN's penultimate (pre-logit) layer. For the zero-shot evaluation of CLIP, we use Maximum-Concept-Matching (MCM) (Ming et al., 2022) and Cosine-similarity (Cos) (Anonymous, 2023) to class-specific text-embeddings. Noting that OOD detection based on softmax of a cosine similarity to a specific feature vector has been proposed in different variants (Tack et al. (2020), Techapanurak et al. (2020) and MCM), we find that using it with classifier class means produces reasonable OOD detection results, marked below as relative cosine class similarity (RCos). We provide an overview over all methods in Appendix D.1.

A.1 RESULTS ON NINCO

Comparison of OOD detection Methods. In Figure 4 (left), we illustrate the performance of a single ViT when combined with a range of OOD-methods. Overall, most feature-based methods, like Maha, RMaha and ViM, outperform the MSP-baseline by a clear margin. Notably, MaxLogit, which utilizes no feature information, is also able to strongly improve over MSP, while KL-Matching performs roughly on par, and KNN much worse. We observe that while Maha, RMaha and ViM improve over MSP in all FPR ranges, this is different for e.g. MaxLogit: For large FPR, it is similar to MSP, indicating that the method brings no advantage over MSP for hard test classes, and its improved mean performance is mainly due to lower FPR for the easier OOD classes. When regarding the mean FPR values of all method-model-combinations shown in Table 2 in Appendix A.3, we observe that while Maha in combination with a (pretrained) ViT is the single best OOD-detector, this method often performs worse when combined with other models. RMaha, however, yields good results with *all* models, and is the only method which can consistently improve over the MSP baseline in terms of mean FPR. For most models, it is either the best-performing method, or close to the best-performing method, which is somewhat surprising, given its relatively poor performance on the unit-tests. We further note that for all models, the best-performing method always uses feature-information, and that in contrast to e.g. KNN, Energy and ReAct, even the adapted methods based on feature space cosine similarity Cos and MCM/RCos fairly consistently improve over the MSP-baseline. Each OOD dataset representing a different out-distribution that can be relevant for certain applications, we find that results vary on the cleaned subsets of six previous benchmarks which we evaluate in Appendix K, while the overall conclusions on the methods and models resemble those on NINCO.

Pretraining matters. In Figure 5, we plot the mean FPR on NINCO over the accuracy for all investigated models for both the MSP-baseline (left) and the best-performing OOD detector per model (right). For MSP, the mean FPR decreases roughly linearly with accuracy. Since most pretrained models (blue) have higher accuracy, they typically also show better OOD-detection performance, but also between models of similar accuracy, the pretrained ones achieve better mean FPR. For the best-performing OOD detector, improvements can be observed for models both with and without pretraining. Notably, the linear relation between FPR and accuracy disappears, and all purely 1K models (green) perform roughly on one level. In comparison, the gains for the majority of models pretrained on IN-21K (blue) are significantly larger. In particular ViT, ConvNext and BiT benefit strongly from leveraging their respective best method, which as discussed above always uses feature-information. In other words, pretraining helps in two ways: First, it leads to higher ID-performance (accuracy), which benefits methods like the MSP-baseline. Second, it creates better feature-embeddings for this task, which lead to improvements beyond the accuracy-MSP correlation. This is most clearly visible for the pretrained BiT-m, which has comparably low accuracy (82%)

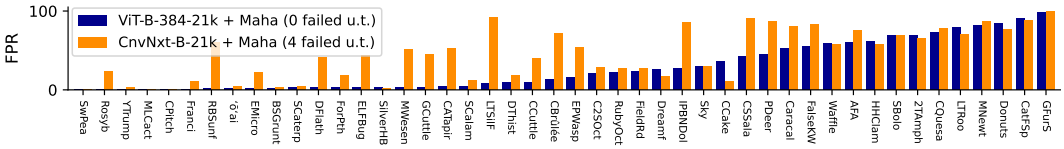

Figure 6: FPR of a pretrained ViT-B and pretrained ConvNext-B for all classes of NINCO.

and hence no outstanding MSP-performance, but outperforms all 1k-models by a significant margin with features leveraging ViM. However, as we observe in Figure 4 (right), the benefit of pre-training depends strongly on the specific data and training method: With RMaha, the ViT with 'traditional' IN-21K pretraining from (Steiner et al., 2022) clearly outperforms models with the distillation-based training of DeiT3 (Touvron et al., 2022), CLIP-pretraining or even CLIP with interjected IN-12K training, which barely improve over ViT without pretraining. The zero-shot methods for CLIP, despite having shown promising results in (Anonymous, 2023) and (Ming et al., 2022) and performing well on the unit tests, are not competitive to IN-1k classifiers on NINCO. Regarding all methods, the five models trained with different pre-training strategies (EfficentNet-b7 with noisy student and four ViTs with CLIP-pretraining (Radford et al., 2021) and subsequent fine-tuning) show some of the highest accuracies in our survey, yet, their OOD-detection performance is surprisingly poor. Overall, we see strong indication hat the precise type of pretraining has a large impact on whether it produces a feature space that is beneficial for feature based methods. In Appendix L we investigate whether IN-21K-pretraining particularly benefits detection of OOD classes that overlap with IN-21K classes, but we notice no substantially different changes between the model with and without pretraining.

Analysis of failure cases. In Figure 6 we plot the individual FPR for each OOD class of NINCO for the combination ViT+Maha, the overall best OOD detector in terms of mean FPR, and contrast it with ConvNext+Maha, which also shows good mean FPR. Performance varies widely between OOD classes, with both models severely struggling for some classes. Where the ViT shows large FPR, the ConvNext rarely performs better, while it also fails to detect certain classes like the *long-tailed silverfish* where the ViT does well. We illustrate samples from hard classes in Figure 3. Both models struggle to detect the *Galápagos fur seal* (98% FPR for the ViT), often confused with the IN-1K class *sea lion*, and *cat-faced spider* (confused with *barn spider*, 91% FPR). From a human perspective, those classes are arguably hard to detect. We note, however, that it is possible to tell them apart, as a ViT IN-21K-classifier e.g. identifies the *Galápagos fur seal* as a *fur seal* (IN-21K class) in 92% of samples and misclassifies only 6% of them as a *sea lion*. The networks however also fail for classes more obvious to humans: *donut* (84% FPR ViT, confused with *bagel*), *spaghetti bolognese* (69% FPR, *carbonara*) and *chicken quesadilla* (73% FPR, *burrito*) also confuse both models.

Table 1: **Some detectors fail OOD unit-tests:** FPR for a ViT and a ConvNext (with and without pretraining) on selected unit-tests. FPR larger than 10% count as failed an are thus marked red. Especially for methods relying on feature representations (like ViM and Maha) the OOD unit-tests reveal difficulties.

	method	*black*	*white*	*grey*	*hor*	*SmN*	*Rad*	*mon*
ViT21k	MSP	0.0	0.0	0.0	0.2	0.5	0.0	0.0
	ViM	0.0	100.0	46.0	0.0	0.0	0.0	0.5
	Maha	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Cos	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Cnv1k	MSP	0.0	0.0	0.0	65.8	0.8	0.0	0.0
	ViM	100.0	100.0	100.0	98.5	23.8	100.0	100.0
	Maha	100.0	100.0	100.0	88.2	25.0	100.0	100.0
	Cos	0.0	0.0	0.0	35.0	0.0	0.0	0.0
Cnv21k	MSP	0.0	0.0	0.0	6.0	1.8	0.0	0.0
	ViM	100.0	100.0	100.0	0.0	0.0	41.2	0.5
	Maha	100.0	100.0	100.0	0.0	0.0	42.5	2.8
	Cos	0.0	0.0	0.0	0.0	0.0	0.0	0.0

A.2 Results on the OOD unit-tests

Auditing OOD detectors on the OOD unit-tests, we find that surprisingly many combinations of models and OOD detection methods struggle to distinguish supposedly easy inputs from ID-data. While results for all models and methods can be found in Appendix J, we provide some illustrative

unit-test results in Table 1 for a ViT pretrained on IN-21k and a ConvNext both with and without IN-21K pretraining. In general, most methods fail fewer unit tests when applied to pretrained models, however there are still many severely flawed combinations, often involving methods that would otherwise shine based on their detection of natural OOD data discussed above: especially the feature-based methods ViM, Maha and RMaha reveal weaknesses, each failing multiple unit-tests on at least 21 of 26 models. Many tested OOD detectors are vulnerable to *black*, *white* and *grey*, which is concerning as encountering inputs of this kind could occur in many real-world applications due to camera malfunction or occlusion. Here those feature methods only provide trustworthy results in combination with ViTs pretrained on IN-21k, the BiT-models and a pretrained EfficientNet-V2. Methods like Cos (7/26 models fail multiple tests) and MCM/RCos (7/26), originally designed for cosine-trained features as in CLIP, achieve remarkably strong OOD-detection performance on the unit-tests across a broad range of models, both with and without CLIP-pretraining. While taking note of these general trends, each OOD detector's robustness to the OOD unit-tests should be examined individually.

A.3 Detailed results on NINCO

A detailed overview over the results on the NINCO benchmark is presented in Table 2, where we show the mean FPR for all models and methods across the dataset's OOD classes. The best method per model is marked bold, and the difference to the MSP-baseline is shown in green where a model outperforms the MSP-baseline and in red if it performs worse than MSP. It is clearly visible that there is no one-fits-all method. Instead, different models synergize with different methods. Overall, the two ViT models pretrained only on IN-21K in combination with Mahalanobis distance outperform other models and methods by a clear margin. This is in line with the observations of previous works (Koner et al., 2021; Fort et al., 2021; Anonymous, 2023), which also found the ViTs to perform exceptionally well. In terms of MSP, the ViTs are not better than e.g. the ConvNext, indicating that their improved OOD detection capabilities stem from a favourably structured feature-space. It is further interesting to see that for models without pretraining, out of all methods only Relative Mahalanobis consistently improves over the MSP-baseline. Apart from KL-Matching and KNN, most methods improve fairly consistently over the MSP-baseline for pretrained models and the CLIP-methods Cosine and RCos perform comparably well, yielding their best results with models pretrained *both* on CLIP and IN-12k. Since CLIP models are trained with cosine-similarity, it is likely that the structure of the feature space after finetuning remains favorable to cosine-based methods, while it might harm the performance of other feature-based methods like Mahanobis compared to models pretrained *only* on IN-21k.

It has been remarked (Hendrycks et al., 2022) that the advantage of models pretrained with IN-21K in the OOD detection task CIFAR-10 vs. CIFAR-100 (Krizhevsky & Hinton, 2009) might partially be explained by the CIFAR-100 classes not truly being unseen at train time, as they have a large overlap with IN-21K classes. We checked each NINCO class for overlap with the 21 843 classes of IN-21K with the help of a ViT classifier for IN-21K, see Table 6. This allows us to test whether the pretrained models have a larger advantage over purely IN-1K-trained models when trying to detect those classes with overlap compared to the classes without overlap. In Appendix L notice no substantially different changes between the models with and without pretraining. We remark, however, that even for several models without pretraining, the subselections of classes show quite different results.

Table 2: **Mean FPR on our NINCO dataset.** Lower is better. The difference to MSP is shown in red if a method performs worse, and in green if it improves. Bold values mark the best-performing method per model.

pre	acc.	model	MSP	MaxL	Ener	KL-M	Maha	RMaha	ViM	E+R	KNN	Cos	MCM/RCos
21k	86.0	ViT-B-384	54.2	42.3 −12	41.8 −12	52.2 −2	**28.3** −26	31.2 −23	35.9 −18	43.1 −11	71.0 +17	50.2 −4	49.6 −5
	84.5	ViT-B-224	59.7	49.8 −10	49.7 −10	57.7 −2	**32.3** −27	35.5 −24	42.2 −17	53.0 −7	76.3 +17	58.2 −2	59.6 −0
	86.3	Swinv2-B-256	54.4	44.5 −10	43.3 −11	59.2 +5	59.5 +5	50.9 −4	54.9 +0	**40.9** −14	67.5 +13	55.1 +1	51.1 −3
	86.7	Deit3-B-384	64.9	60.6 −4	61.8 −3	63.0 −2	53.9 −11	**49.6** −15	58.7 −6	57.1 −8	58.4 −6	52.0 −13	51.5 −13
	85.7	Deit3-B-224	68.3	64.4 −4	64.6 −4	68.9 +1	59.4 −9	**53.7** −15	64.5 −4	62.0 −6	64.4 −4	58.0 −10	57.0 −11
	86.3	CnvNxt-B	50.8	45.2 −6	47.5 −3	57.3 +7	45.0 −6	**40.3** −11	42.2 −9	43.5 −7	58.2 +7	46.9 −4	45.6 −5
	84.1	CnvNxt-T	56.0	50.7 −5	48.6 −7	61.7 +6	45.9 −10	45.4 −11	**40.4** −16	46.8 −9	66.4 +10	54.2 −2	52.4 −4
	82.3	BiT-m	67.8	59.0 −9	59.2 −9	64.8 −3	50.3 −17	**41.5** −26	41.6 −26	49.9 −18	59.4 −8	53.4 −14	54.3 −13
	85.6	EffNetv2-M	54.9	52.9 −2	59.1 +4	57.2 +2	63.2 +8	52.0 −3	58.6 +4	90.6 +36	74.0 +19	**50.4** −4	53.2 −2
none	81.1	ViT-B-384	71.9	70.1 −2	69.9 −2	67.5 −4	57.9 −14	**53.0** −19	70.4 −2	66.2 −6	78.7 +7	72.5 +1	73.8 +2
	84.6	Swinv2-B-256	71.4	68.8 −3	73.9 +3	69.4 −2	65.8 −6	**60.9** −10	70.3 −1	72.2 +1	74.5 +3	68.2 −3	66.0 −5
	85.1	Deit3-B-384	68.4	73.6 +5	88.0 +20	67.0 −1	67.5 −1	61.3 −7	61.1 −7	89.6 +21	81.5 +13	72.5 +4	**60.4** −8
	83.8	Deit3-B-224	70.4	71.9 +1	82.2 +12	69.7 −1	71.7 +1	65.4 −5	**64.2** −6	84.1 +14	87.1 +17	78.7 +8	64.2 −6
	82.6	XCiT-M-224	73.7	73.4 −0	79.1 +5	74.4 +1	68.5 −5	**65.0** −9	67.1 −7	77.8 +4	77.8 +4	72.1 −2	70.2 −3
	84.3	XCiT-M-224-d	69.4	66.4 −3	71.9 +2	68.4 −1	68.8 −1	**63.0** −6	64.9 −5	75.1 +6	77.2 +8	69.6 +0	66.9 −2
	84.4	CnvNxt-B	63.6	67.8 +4	86.3 +23	72.6 +9	69.8 +6	63.4 −0	69.4 +6	83.0 +19	77.9 +14	66.5 +3	**62.4** −1
	78.0	BiT-s	81.9	83.6 +2	84.7 +3	67.2 −15	84.9 +3	**58.9** −23	77.2 −5	77.3 −5	85.0 +3	74.0 −8	85.3 +3
	85.0	EffNetv2-M	67.0	67.1 +0	76.7 +10	64.4 −3	65.0 −2	**56.4** −11	74.8 +8	73.3 +6	70.3 +3	63.8 −3	58.8 −8
	84.9	EffNetb7	70.0	72.3 +2	84.5 +15	65.5 −5	72.0 +2	**57.4** −13	77.9 +8	80.8 +11	72.5 +3	64.9 −5	58.7 −11
	77.7	EffNet-B0	74.1	73.3 −1	79.1 +5	72.8 −1	85.5 +11	73.0 −1	79.6 +6	75.8 +2	87.0 +13	**60.7** −13	67.2 −7
	80.4	ResNet50	73.4	76.1 +3	80.6 +7	69.6 −4	86.8 +13	66.3 −7	83.3 +10	98.2 +25	79.1 +6	61.5 −12	**61.1** −12
JFT	86.8	EffNetb7-ns	66.7	63.5 −3	71.7 +5	66.0 −1	82.6 +16	**63.0** −4	85.4 +19	70.8 +4	75.3 +9	65.5 −1	63.2 −4
clip +12k	87.2	ViT-B-384-l2b	52.8	50.5 −2	53.7 +1	54.6 +2	52.0 −1	47.0 −6	46.7 −6	48.9 −4	50.7 −2	43.9 −9	**43.9** −9
	87.0	ViT-B-384-oai	52.8	48.3 −5	48.3 −4	54.4 +2	55.8 +3	48.7 −4	52.0 −1	47.1 −6	50.8 −2	43.7 −9	**43.5** −9
clip	86.6	ViT-B-384-l2b	63.3	63.1 −0	67.8 +5	60.3 −3	57.6 −6	54.7 −9	57.7 −6	65.7 +2	62.9 −0	55.8 −7	**53.7** −10
	86.2	ViT-B-384-oai	67.4	67.3 −0	70.8 +3	64.8 −3	57.6 −10	**55.3** −12	59.4 −8	69.2 +2	65.5 −2	60.3 −7	57.8 −10
clip	74.2	clip-ViT-L-336	—	—	—	—	—	—	—	—	—	**66.5**	67.5
z. shot	66.6	clip-ViT-B-224	—	—	—	—	—	—	—	—	—	**78.5**	79.2

Table 3: **Mean AUROC on our NINCO dataset.** Higher is better. The difference to MSP is shown in red if a method performs worse, and in green if it improves. Bold values mark the best-performing method per model.

pre	acc.	model	MSP	MaxL	Ener	KL-M	Maha	RMaha	ViM	E+R	KNN	Cos	MCM/RCos
21k	86.0	ViT-B-384	86.2	91.4 +5	91.5 +5	85.4 −1	**94.5** +8	93.7 +7	93.2 +7	91.3 +5	83.1 −3	90.3 +4	90.4 +4
	84.5	ViT-B-224	84.8	89.5 +5	89.6 +5	84.1 −1	**93.5** +9	92.6 +8	91.5 +7	88.8 +4	79.5 −5	88.1 +3	87.8 +3
	86.3	Swinv2-B-256	85.2	86.1 +1	85.0 −0	84.6 −1	87.2 +2	88.2 +3	**88.8** +4	88.0 +3	81.5 −4	87.6 +2	88.5 +3
	86.7	Deit3-B-384	79.8	75.2 −5	71.8 −8	82.4 +3	88.8 +9	89.2 +9	87.6 +8	78.0 −2	84.9 +5	88.9 +9	**89.2** +9
	85.7	Deit3-B-224	78.7	74.8 −4	71.6 −7	80.8 +2	87.3 +9	**87.9** +9	86.0 +7	76.5 −2	82.6 +4	87.3 +9	87.6 +9
	86.3	CnvNxt-B	86.7	86.3 −0	84.4 −2	86.7 −0	90.9 +4	91.5 +5	**91.7** +5	86.8 +0	85.8 −1	90.6 +4	90.9 +4
	84.1	CnvNxt-T	84.0	84.4 +0	83.4 −1	84.1 +0	88.9 +5	88.4 +4	**91.2** +7	84.4 +0	83.3 −1	87.5 +4	88.2 +4
	82.3	BiT-m	82.2	84.7 +3	84.3 +2	81.9 −0	90.0 +8	89.8 +8	**91.9** +10	89.0 +7	84.6 +2	88.5 +6	87.4 +5
	85.6	EffNetv2-M	85.0	83.5 −1	81.0 −4	86.0 +1	86.4 +1	87.7 +3	86.0 +1	72.5 −13	80.9 −4	**88.3** +3	87.2 +2
none	81.1	ViT-B-384	80.1	82.9 +3	82.9 +3	78.9 −1	85.6 +6	**87.0** +7	81.0 +1	83.7 +4	76.2 −4	82.5 +2	82.0 +2
	84.6	Swinv2-B-256	80.2	78.0 −2	75.4 −5	80.8 +1	85.2 +5	**85.8** +6	78.6 −2	79.6 −1	79.3 −1	84.3 +4	84.9 +5
	85.1	Deit3-B-384	81.4	76.4 −5	67.4 −14	82.1 +1	85.6 +4	**87.1** +6	84.4 +3	64.4 −17	76.9 −4	83.6 +2	85.6 +4
	83.8	Deit3-B-224	81.3	79.5 −2	76.2 −5	81.2 −0	84.3 +3	**86.2** +5	84.1 +3	74.5 −7	74.0 −7	81.9 +1	84.6 +3
	82.6	XCiT-M-224	77.9	73.6 −4	67.0 −11	80.0 +2	83.8 +6	84.8 +7	**85.0** +7	73.5 −4	77.9 −0	82.9 +5	83.4 +5
	84.3	XCiT-M-224-d	83.0	82.1 −1	77.2 −6	82.0 −1	84.5 +2	**86.1** +3	85.0 +2	79.0 −4	78.6 −4	84.3 +1	84.7 +2
	84.4	CnvNxt-B	81.5	78.6 −3	69.1 −12	81.9 +0	83.4 +2	**85.3** +4	80.6 −1	75.4 −6	77.8 −4	83.8 +2	85.2 +4
	78.0	BiT-s	78.7	74.6 −4	72.7 −6	81.1 +2	69.6 −9	**85.7** +7	73.6 −5	75.0 −4	63.9 −15	75.0 −4	67.0 −12
	85.0	EffNetv2-M	80.6	77.3 −3	71.1 −10	82.7 +2	85.1 +5	**87.9** +7	77.6 −3	76.9 −4	80.7 +0	85.8 +5	86.6 +6
	84.9	EffNetb7	77.8	70.7 −7	62.5 −15	82.6 +5	82.4 +5	**87.5** +10	78.5 +1	68.5 −9	79.2 +1	85.1 +7	86.0 +8
	77.7	EffNet-B0	80.2	78.2 −2	74.9 −5	81.1 +1	74.5 −6	82.4 +2	79.3 −1	78.2 −2	74.9 −5	**83.9** +4	81.4 +1
	80.4	ResNet50	80.6	80.5 −0	80.1 −0	77.7 −3	74.0 −7	**84.8** +4	79.3 −1	62.7 −18	73.5 −7	83.6 +3	84.4 +4
JFT	86.8	EffNetb7-ns	79.6	76.5 −3	71.1 −8	80.9 +1	79.7 +0	**86.6** +7	76.1 −4	74.7 −5	77.0 −3	84.2 +5	84.2 +5
clip +12k	87.2	ViT-B-384-l2b	84.5	80.7 −4	76.0 −8	87.5 +3	89.8 +5	90.1 +6	**91.2** +7	81.1 −3	87.5 +3	90.9 +6	90.9 +6
	87.0	ViT-B-384-oai	85.3	83.6 −2	82.1 −3	86.7 +1	88.9 +4	90.0 +5	90.0 +5	84.3 −1	87.9 +3	91.2 +6	**91.2** +6
clip	86.6	ViT-B-384-l2b	79.8	72.4 −7	67.4 −12	84.3 +5	87.3 +7	87.5 +8	86.5 +7	70.2 −10	83.4 +4	88.1 +8	**88.2** +8
	86.2	ViT-B-384-oai	77.5	69.1 −8	63.8 −14	82.5 +5	87.2 +10	87.2 +10	86.9 +9	66.6 −11	82.0 +5	87.0 +10	**87.7** +10
clip	74.2	clip-ViT-L-336	—	—	—	—	—	—	—	—	—	**82.3**	79.4
z. shot	66.6	clip-ViT-B-224	—	—	—	—	—	—	—	—	—	**74.4**	73.9

Table 4: FPR on all classes of NINCO (lower is better) for a pretrained ViT-B and a pretrained ConvNext-B.

	ViT-B-384-21k											CnvNxt-B-21k										
Dataset	MSP	MaxL	Ener	KL-M	Maha	RMaha	ViM	E+R	KNN	Cos	RCos	MSP	MaxL	Ener	KL-M	Maha	RMaha	ViM	E+R	KNN	Cos	RCos
Caracal	79.0	75.0	73.0	68.0	53.0	47.0	82.0	88.0	87.0	83.0	81.0	78.0	72.0	67.0	78.0	80.0	75.0	88.0	67.0	88.0	77.0	78.0
2TAmph	75.6	90.9	91.5	70.5	68.8	54.0	92.0	93.8	95.5	86.9	84.7	80.7	86.4	91.5	68.8	64.8	56.8	72.2	89.2	84.7	73.3	71.6
AFA	95.7	87.0	82.6	93.5	60.9	71.7	65.2	80.4	80.4	84.8	82.6	89.1	80.4	69.6	89.1	76.1	80.4	47.8	60.9	82.6	80.4	80.4
CatFSp	90.0	96.0	96.0	94.0	91.0	88.0	91.0	97.0	100.0	97.0	97.0	81.0	79.0	88.0	92.0	88.0	79.0	89.0	84.0	98.0	96.0	96.0
GFurS	100.0	98.9	96.7	95.6	97.8	96.7	79.1	100.0	100.0	100.0	100.0	98.9	98.9	81.3	96.7	100.0	100.0	100.0	86.8	100.0	100.0	100.0
CSSala	76.0	86.0	89.0	76.0	42.0	60.0	57.0	93.0	100.0	94.0	94.0	80.0	78.0	73.0	80.0	91.0	73.0	93.0	72.0	97.0	89.0	86.0
CQuesa	86.0	95.0	97.0	81.0	73.0	77.0	82.0	88.0	99.0	98.0	98.0	89.0	92.0	96.0	84.0	78.0	72.0	83.0	94.0	95.0	85.0	83.0
DThist	52.0	22.0	19.0	48.0	9.0	13.0	6.0	20.0	51.0	34.0	32.0	44.0	41.0	47.0	52.0	19.0	30.0	10.0	43.0	42.0	38.0	38.0
CBrûlée	52.5	33.3	30.3	65.7	13.1	24.2	22.2	27.3	77.8	60.6	53.5	36.4	24.2	25.3	64.6	71.7	44.4	42.4	22.2	53.5	41.4	40.4
LTSilF	59.0	28.0	26.0	51.0	8.0	4.0	36.0	40.0	79.0	44.0	40.0	35.0	18.0	12.0	66.0	92.0	62.0	93.0	11.0	98.0	77.0	66.0
CCake	81.2	65.0	55.0	75.0	36.2	83.8	23.8	56.2	88.8	51.2	45.0	77.5	83.8	90.0	66.2	11.2	47.5	13.8	87.5	7.5	15.0	15.0
CPitch	23.0	5.0	6.0	14.0	0.0	0.0	0.0	1.0	51.0	14.0	15.0	19.0	18.0	31.0	32.0	1.0	7.0	3.0	20.0	18.0	9.0	10.0
LTRoo	68.0	88.0	95.0	72.0	79.0	68.0	93.0	97.0	100.0	96.0	98.0	65.0	74.0	84.0	64.0	71.0	43.0	85.0	84.0	95.0	80.0	80.0
Donuts	81.0	79.0	80.0	86.0	84.0	88.0	74.0	77.0	97.0	86.0	86.0	71.0	66.0	70.0	83.0	77.0	81.0	68.0	67.0	80.0	78.0	79.0
EMicro	35.0	24.0	23.0	26.0	2.0	1.0	22.0	23.0	44.0	25.0	19.0	48.0	42.0	41.0	46.0	22.0	15.0	21.0	37.0	45.0	32.0	29.0
Franci	26.0	9.0	3.0	18.0	0.0	0.0	0.0	4.0	81.0	15.0	17.0	25.0	9.0	11.0	32.0	11.0	16.0	2.0	5.0	41.0	23.0	23.0
FieldRd	26.0	17.7	18.8	34.4	24.0	21.9	22.9	20.8	51.0	24.0	25.0	25.0	28.1	46.9	35.4	27.1	29.2	25.0	33.3	46.9	24.0	24.0
ForPth	24.0	5.0	3.0	38.0	3.0	11.0	3.0	4.0	29.0	19.0	18.0	16.0	14.0	18.0	23.0	18.0	23.0	7.0	16.0	26.0	18.0	17.0
MLCact	13.0	0.0	0.0	15.0	0.0	0.0	0.0	0.0	9.0	1.0	0.0	31.0	21.0	14.0	42.0	1.0	6.0	1.0	13.0	3.0	8.0	8.0
BSGrunt	43.4	18.2	12.1	43.4	2.0	4.0	4.0	10.1	55.6	17.2	19.2	39.4	30.3	33.3	50.5	3.0	8.1	2.0	24.2	35.4	16.2	13.1
HHClam	93.5	93.5	90.3	93.5	61.3	67.7	77.4	93.5	96.8	96.8	93.5	71.0	64.5	48.4	71.0	58.1	67.7	41.9	35.5	71.0	71.0	67.7
SilverHB	11.1	2.0	2.0	8.1	3.0	2.0	3.0	2.0	4.0	3.0	3.0	11.1	11.1	15.2	14.1	2.0	3.0	2.0	11.1	5.1	3.0	3.0
SwPea	10.0	1.0	1.0	12.0	0.0	1.0	1.0	0.0	33.0	2.0	2.0	9.0	10.0	10.0	22.0	0.0	1.0	0.0	6.0	2.0	1.0	1.0
RBSunf	76.0	12.0	5.0	69.0	2.0	2.0	4.0	1.0	65.0	14.0	9.0	33.0	8.0	7.0	51.0	60.0	44.0	15.0	6.0	38.0	24.0	13.0
ELFBug	37.0	26.0	49.0	38.0	3.0	3.0	44.0	71.0	98.0	67.0	83.0	44.0	21.0	18.0	67.0	44.0	30.0	24.0	18.0	63.0	40.0	36.0
MWesen	97.0	33.3	18.2	90.9	3.0	30.3	3.0	9.1	57.6	57.6	39.4	78.8	48.5	33.3	87.9	51.5	78.8	6.1	30.3	81.8	84.8	84.8
C2SOct	40.0	49.0	58.0	36.0	21.0	17.0	61.0	62.0	87.0	54.0	52.0	44.0	47.0	57.0	43.0	29.0	13.0	38.0	51.0	59.0	30.0	28.0
RubyOct	42.0	48.0	48.0	39.0	22.0	14.0	55.0	48.0	88.0	54.0	54.0	24.0	30.0	41.0	25.0	27.0	12.0	34.0	35.0	49.0	27.0	29.0
PDeer	81.7	61.0	59.8	89.0	45.1	58.5	52.4	85.4	91.5	80.5	80.5	68.3	37.8	34.1	93.9	86.6	76.8	78.0	32.9	96.3	84.1	82.9
DFlath	58.0	33.0	32.0	52.0	3.0	18.0	13.0	32.0	62.0	31.0	33.0	65.0	53.0	49.0	67.0	41.0	42.0	27.0	45.0	67.0	53.0	50.0
EPWasp	73.0	55.0	56.0	64.0	16.0	29.0	29.0	63.0	90.0	75.0	72.0	36.0	28.0	22.0	69.0	54.0	42.0	40.0	22.0	73.0	51.0	50.0
FalseKW	80.6	74.6	74.6	74.6	55.2	55.2	59.7	86.6	98.5	91.0	89.6	71.6	56.7	56.7	73.1	83.6	76.1	91.0	56.7	97.0	85.1	83.6
Sky	23.5	25.0	25.0	20.6	29.4	25.0	39.7	29.4	44.1	14.7	16.2	25.0	25.0	66.2	22.1	29.4	27.9	32.4	54.4	32.4	14.7	16.2
Dreamf	60.0	44.0	45.0	63.0	26.0	31.0	30.0	42.0	69.0	48.0	50.0	63.0	62.0	73.0	64.0	17.0	29.0	13.0	59.0	68.0	56.0	54.0
YTrump	14.0	1.0	0.0	4.0	0.0	0.0	0.0	0.0	54.0	10.0	14.0	16.0	8.0	5.0	26.0	3.0	2.0	0.0	4.0	5.0	0.0	0.0
GCuttle	31.3	10.1	8.1	34.3	3.0	4.0	14.1	9.1	36.4	14.1	13.1	33.3	26.3	37.4	43.4	45.5	10.1	65.7	36.4	47.5	23.2	21.2
CCuttle	34.0	23.0	24.0	24.0	9.0	8.0	27.0	25.0	44.0	22.0	20.0	26.0	26.0	32.0	36.0	40.0	8.0	57.0	26.0	40.0	16.0	15.0
SCalam	21.2	11.1	13.1	21.2	5.1	4.0	13.1	10.1	39.4	16.2	15.2	33.3	23.2	27.3	34.3	12.1	6.1	15.2	22.2	34.3	18.2	17.2
SCaterp	11.0	12.0	19.0	10.0	3.0	3.0	13.0	16.0	80.0	33.0	31.0	27.0	23.0	27.0	26.0	5.0	8.0	3.0	23.0	21.0	11.0	10.0
SBolo	67.2	47.8	49.3	83.6	68.7	79.1	50.7	22.4	100.0	100.0	100.0	74.6	76.1	83.6	77.6	68.7	61.2	76.1	82.1	100.0	91.0	88.1
Rosyb	65.0	28.0	17.0	45.0	0.0	2.0	2.0	11.0	86.0	36.0	37.0	73.0	68.0	70.0	72.0	23.0	38.0	22.0	60.0	72.0	52.0	52.0
CATapir	13.0	10.0	11.0	12.0	4.0	3.0	15.0	15.0	23.0	13.0	13.0	26.0	30.0	37.0	48.0	53.0	10.0	71.0	36.0	53.0	22.0	22.0
MNewt	90.0	94.0	95.0	91.0	82.0	83.0	80.0	96.0	99.0	98.0	98.0	93.0	95.0	98.0	93.0	87.0	82.0	92.0	98.0	99.0	93.0	93.0
IPBNDol	57.0	47.0	45.0	56.0	27.0	29.0	47.0	64.0	90.0	64.0	69.0	48.0	33.0	36.0	63.0	86.0	59.0	93.0	39.0	94.0	71.0	68.0
'ō'ai	63.0	28.0	23.0	51.0	2.0	5.0	3.0	18.0	71.0	33.0	32.0	53.0	45.0	34.0	52.0	4.0	14.0	1.0	27.0	27.0	18.0	17.0
Waffle	57.4	54.1	59.0	55.7	59.0	49.2	60.7	52.5	83.6	52.5	57.4	62.3	65.6	75.4	50.8	57.4	62.3	55.7	68.9	45.9	57.4	55.7
mean	54.2	42.3	41.8	52.2	28.3	31.2	35.9	43.1	71.0	50.2	49.6	50.8	45.2	47.5	57.3	45.0	40.3	42.2	43.5	58.2	46.9	45.6

B POPULAR TEST OOD DATASETS FOR IMAGENET-1K

We use *blue* for the name of an ImageNet class and *brown* for the category name in the source dataset used for the generation of the test OOD dataset. **iNATURALIST OOD PLANTS** is a subset of 10 000 images curated by Huang & Li (2021) from 110 OOD plant species of iNat2017 (Van Horn et al., 2018) which is sourced from the iNaturalist project. It is frequently used as test OOD dataset (Xia & Bouganis, 2022; Ming et al., 2022).

PLACES is a subset of Places365 (Zhou et al., 2017) curated by Huang & Li (2021) as "50 categories [...] that are not present in IN-1K". It is used as test OOD dataset in (Huang & Li, 2021; Sun et al., 2021; Ming et al., 2022). The dataset contains 9 822 images from 50 environment classes. We find that several of these classes are either subsets of ID classes, e.g. *hayfield* (*hay*), *cornfield* (*corn*), *lagoon* (*seashore* and *lakeshore*), or contain mostly ID objects, e.g. *underwater* (*coral reef* and *scuba diver*), *ocean* (*seashore*).

TEXTURES (Cimpoi et al., 2014) contains 5640 images of various objects that show one of 47 patterns. It is used as test OOD dataset in (Huang & Li, 2021; Sun et al., 2021; Wang et al., 2021; Xia & Bouganis, 2022; Ming et al., 2022) and others. Wang et al. (2022) address the issue of overlap with IN-1K and remove four categorically ID textures (*bubbly* (*bubble*), *honeycombed* (*honeycomb*), *cobwebbed* (*spider web*), *spiralled* (*spiral*)). We find that even their version (denoted as TEXTURES43) contains about 20% ID images.

SPECIES was proposed in (Hendrycks et al., 2022) as OOD dataset for IN-21K (Deng et al., 2009) and should thus also be OOD for the IN-1K subset. Sourced from iNaturalist, it consists of 700 000 images from 1 316 species which were selected for not being in IN-21K. They sort the species into 10 superclasses. The largest superclass *Fungi* largely coincides with the IN-1K class *mushroom*, and also many of the remaining species are ID. Papers evaluating on SPECIES for IN-1K OOD detection include (Salehi et al., 2021; Yang et al., 2022; Song et al., 2022).

IMAGENET-O (Hendrycks et al., 2021) contains 2 000 images from IN-21K, excluding its subset IN-1K. To make the dataset challenging it was composed from images where a ResNet-50 classifier for a subset of 200 IN-1K classes attains high confidence. The samples being OOD relies on the assumption that IN-21K without IN-1K is OOD for IN-1K. However, this assumption does not hold, due to a significant overlap between ImageNet classes from IN-1K and IN-21K, e.g. *analytical balance*/*scale* and *pickle*/*cucumber*, and a lack of filtering for incidental ID objects.

OPENIMAGE-O (Wang et al., 2022) consists of 17 632 images from the OpenImage-v3 (Krasin et al., 2017) test set which their human labellers categorize as OOD. It is also used in Yang et al. (2022).

C MODELS

In Table 5 we give an overview over the evaluated models. All model implementation and model weights were taken from the publicly available timm-repository (Wightman, 2019), except for the BiT-s weights, which can be obtained via the github repository of (Kolesnikov et al., 2020), and the zero-shot CLIP models, which are also available via github. For the ViTs finetuned from CLIP and the ViT without pretraining we used the timm-version *0.8.0dev0*, for all other models version *0.6.12*. IN-12k (description and defining synsets) is a subset of IN-21k, for which the classes with few samples are excluded, leading to an overlap of roughly 85%.

Table 5: Overview over the evaluated models.

model	pretraining	top-1 acc.	params	timm name
ViT-B-384-l2b-12k	laion2b + IN-12k	87.2	87M	vit_base_patch16_clip_384.laion2b_ft_in12k_in1k
ViT-B-384-oai-12k	openai + IN-12k	87.0	87M	vit_base_patch16_clip_384.openai_ft_in12k_in1k
ViT-B-384-l2b	laion2b	86.6	87M	vit_base_patch16_clip_384.laion2b_ft_in1k
ViT-B-384-oai	openai	86.2	87M	vit_base_patch16_clip_384.openai_ft_in1k
ViT-B-384-21k	IN-21k	86.0	87M	vit_base_patch16_384
ViT-B-224-21k	IN-21k	84.5	87M	vit_base_patch16_224
Swinv2-B-256-21k	IN-21k	86.3	88M	swinv2_base_window12to16_192to256_22kft1k
Deit3-B-384-21k	IN-21k	86.7	87M	deit3_base_patch16_384_in21ft1k
Deit3-B-224-21k	IN-21k	85.7	87M	deit3_base_patch16_224_in21ft1k
CnvNxt-B-21k	IN-21k	86.3	89M	convnext_base_in22ft1k
CnvNxt-T-21k	IN-21k	84.1	29M	convnext_tiny_384_in22ft1k
BiT-m	IN-21k	82.3	45M	resnetv2_101x1_bitm
EffNetv2-M-21k	IN-21k	85.6	54M	tf_efficientnetv2_m_in21ft1k
EffNetb7-ns	JFT - noisy student	86.8	66M	tf_efficientnet_b7_ns
ViT-B-384	—	81.1	87M	vit_base_patch16_384.augreg_in1k
Swinv2-B-256	—	84.6	88M	swinv2_base_window16_256
Deit3-B-384	—	85.1	87M	deit3_base_patch16_384
Deit3-B-224	—	83.8	87M	deit3_base_patch16_224
XCiT-M-224	—	82.6	84M	xcit_medium_24_p16_224
XCiT-M-224-d	—	84.3	84M	xcit_medium_24_p16_224_dist
CnvNxt-B	—	84.4	89M	convnext_base
BiT-s	—	78.0	45M	resnetv2_101x1_bitm
EffNetv2-M	—	85.0	54M	tf_efficientnetv2_m
EffNetb7	—	84.9	66M	tf_efficientnet_b7
EffNet-B0	—	77.7	5M	efficientnet_b0
ResNet50	—	80.4	26M	resnet50
CLIP-ViT-B16	openai	66.6	150M	—
CLIP-ViT-B16	openai	74.2	428M	—

D OOD DETECTORS AND HOW TO EVALUATE THEM

An **OOD detector** for inputs from the domain X of possible input images is represented by a score function $S : X \to \mathbb{R} \cup \{\pm\infty\}$ which is generally supposed to be larger on ID inputs than on OOD inputs. One example is the Maximum Softmax Probability (MSP) or confidence $S_{\text{MSP}}(x) = \max_{k=1,\dots,K} p_k(x)$ of a classifier with output probabilities p for K ID classes. The MSP is the standard baseline OOD detection method (Hendrycks & Gimpel, 2017), since it is intuitively expected to be low on OOD compared to ID inputs. Observing that standard classifiers are frequently overconfident on OOD inputs, OOD detection research aims at finding detectors that improve on this baseline. In Appendix D.1, we give an overview of a range of OOD detection methods which have been proposed for IN-1K as ID. An OOD detector is usually obtained by combining such an OOD detection method with a concrete classifier model. We analyze OOD detectors in terms of the fraction of falsely accepted OOD inputs at a true positive rate of 95%, short FPR. Detailed definitions can be found in Appendix E.

Different OOD classes (and similarly also different OOD test datasets) represent different probabilistic distributions of inputs that a detector is tested against. An important arising question is how the collective of individual performance measurements can be interpreted and whether they can be aggregated into one number that can be used to make an informed decision on which OOD detector works best. Certainly, the notion of 'best' may notably vary depending on the application and situation and we often cannot hope to model a 'true' out-distribution, or even be sure that it meaningfully exists. An aggregate number which gives a good overview of an OOD detector's performance on the class based NINCO dataset is the **mean FPR** of the individual FPR values for each of the 46 OOD classes of NINCO.

However, for many applications it is not possible to model the potential OOD inputs that might be encountered at test time with a fixed probability distribution. Thus a single aggregate number cannot tell the full story, and may hide outliers in the FPR values. For one, some errors might be *less acceptable* than others, e.g. a FPR of 20.0% might be very bad for monochrome inputs, but would lose much significance when subsumed into a mean. For OOD unit tests, where OOD detectors can be expected to be very robust, we therefore propose regarding pass-fail statistics instead of mean FPR. Also, an evaluator might want to be informed about the *concrete failure modes* of the model, e.g. all OOD classes with a particular high FPR. An OOD detector showing consistent improvements on most of the OOD classes (instead of only in terms of the mean) can be seen as strong evidence for the method yielding actual improvement, as opposed to the detector overfitting to a limited scope of test OOD data, which Wang et al. (2022) describe as a form of hackability. Due to these considerations, and with the OOD data being organized into *OOD classes* as in NINCO, we suggest evaluations of OOD detectors to always provide the **distribution of results over OOD classes** and additionally to **make the individual results available**, such that the reader can make an informed comparison based on which types of OOD inputs are most relevant to them.

D.1 OOD DETECTION METHODS

Here we give an overview over the evaluated OOD detection methods. For clarity, we denote vectors in bold and lowercase letters and matrices in bold an uppercase letters. We write neural networks as functions n, which are parametrized by weights θ, take an input sample $\mathbf{x}$ and produce an output vector $\mathbf{o}$ of size C, where C is typically the number of classes in a classification task (1000 in the case of IN-1K). We refer to $\mathbf{o}$ as the logits of $\mathbf{x}$, which can be transformed to a probability vector $\mathbf{p}$ (also of size C) via the softmax function: $p_i = \exp(o_i) / \sum_c \exp(o_c)$. The network n can be decomposed into a feature extractor h and the networks last layer g:

$$\mathbf{o} = n(\mathbf{x}) = g(h(\mathbf{x})),$$

where g is a fully connected, linear layer, i.e. $g(\mathbf{h}) = \mathbf{W}^T\mathbf{h} + \mathbf{b}$ with weight $\mathbf{W}$ and bias $\mathbf{b}$. We refer to $\mathbf{h} = h(\mathbf{x})$ as the *features* or the *embeddings* of $\mathbf{x}$ w.r.t. the network n. As presented in Section A, for each sample $\mathbf{x}$, a method returns an OOD-score $s = f(\mathbf{x})$, a scalar value which is supposed to be larger for ID data and smaller for OOD data. In the following, we will describe how each methods computes the score s for a test input $\mathbf{x}$.

MSP (Hendrycks & Gimpel, 2017): The most popular OOD-detection baseline uses the confidence, i.e. the max softmax probability of a models probability output vector:

$$s = \max_c (p_c)$$

Max-Logit (Hendrycks et al., 2022): Similar to MSP, Max-Logit returns the largest entry of the logit-vector $\mathbf{o}$, i.e.

$$s = \max_c (o_c)$$

Energy (Liu et al., 2020): The Energy based OOD detection method uses the denominator of the softmax-function as OOD-score:

$$s = \log \sum_c^C \exp\left(o_c\right)$$

KL-Matching (Hendrycks et al., 2022): KL-Matching computes a mean probability vector $\mathbf{d}_c$ for each of the C classes. For a test input, the KL-distances of all $\mathbf{d}_c$ vectors to its probability vector $\mathbf{p}$ are computed, and the OOD-score is the negative of the smallest of those distances:

$$s = -\min_c \text{KL}[\mathbf{p}||\mathbf{d}_c]$$

In the original paper by (Hendrycks et al., 2022), the average for $\mathbf{d}_c$ is computed over an additional validation set. Since none of the other methods leverages extra data and we are interested in fair comparison, we deploy KL-Matching like in (Wang et al., 2022; Yang et al., 2022), where the average is computed over the train set.

KNN (Sun et al., 2022): KNN is a non-parametric method that computes distances in the feature-space. Specifically, the feature vector of a test input is normalized to $\mathbf{z} = \mathbf{h}/||\mathbf{h}||_2$ and the pairwise distances $r_i(\mathbf{z}) = ||\mathbf{z} - \mathbf{z}_i||_2$ to the normalized features $\mathcal{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_N\}$ of all samples of the training set are computed. The distances $r_i(\mathbf{z})$ are then sorted according to their magnitude and the Kth smallest distance, denoted $r^K(\mathbf{z})$ is used as negative OOD-score:

$$s = -r^K(\mathbf{z})$$

Like suggested in (Sun et al., 2022), we use $K = 1000$.

Mahalanobis distance (Lee et al., 2018): This popular method fits a class-conditional Gaussian with shared covariance matrix to the train set, i.e. computes

$$\hat{\mu}_c = \frac{1}{N_c} \sum_{i:y_i=c} \mathbf{h}_i, \qquad \hat{\mathbf{\Sigma}} = \frac{1}{N} \sum_c \sum_{i:y_i=c} (\mathbf{h}_i - \hat{\mu}_c)(\mathbf{h}_i - \hat{\mu}_c)^T$$

where N_c is the number of train samples in class c and N is the total number of train samples. The OOD-score of a test sample is then the Mahalanobis distance induced by $\hat{\mathbf{\Sigma}}$ between its feature $\mathbf{h}$ and the closest class mean:

$$s = -\min_c (\mathbf{h} - \hat{\mu}_c)\hat{\mathbf{\Sigma}}^{-1}(\mathbf{h} - \hat{\mu}_c)^T$$

Relative Mahalanobis distance (Ren et al., 2021): A modification of the Mahalanobis distance method, thought to improve near-OOD detection, is to additionally fit a global Gaussian distribution to the train set without taking class-information into account:

$$\hat{\mu}_{\text{global}} = \frac{1}{N} \sum_i \mathbf{h}_i, \qquad \hat{\mathbf{\Sigma}}_{\text{global}} = \frac{1}{N} \sum_i (\mathbf{h}_i - \hat{\mu}_{\text{global}})(\mathbf{h}_i - \hat{\mu}_{\text{global}})^T$$

The OOD-score is then defined as the difference between the original Mahalanobis distance and the Mahalanobis distance w.r.t. the global Gaussian distribution:

$$s = -\min_c \left((\mathbf{h} - \hat{\mu}_c)\hat{\mathbf{\Sigma}}^{-1}(\mathbf{h} - \hat{\mu}_c)^T - (\mathbf{h} - \hat{\mu}_{\text{global}})\hat{\mathbf{\Sigma}}_{\text{global}}^{-1}(\mathbf{h} - \hat{\mu}_{\text{global}})^T \right)$$

ReAct (Sun et al., 2021): The authors propose to perform a truncation of the feature vector, $\bar{\mathbf{h}} = \min(\mathbf{h}, r)$, where the $\min$ operation is to be understood element-wise and r is the truncation threshold. The truncated features can then be converted to so-called rectified logits via $\bar{\mathbf{o}} = g(\bar{\mathbf{h}}) = \mathbf{W}^T\bar{\mathbf{h}} + \mathbf{b}$.

While the rectified logits can now be used with a variety of existing detection methods, we follow (Sun et al., 2021) and use the rectified Energy as OOD-score:

$$s = \log \sum_{c}^{C} \exp\left(\bar{o}_c\right)$$

As suggested in (Wang et al., 2022), we set the threshold r such that 1% of the activations from the train set would be truncated.

Virtual Logit Matching (Wang et al., 2022): The idea behind ViM is that meaningful features are thought to lie in a low-dimensional manifold, called the principal space P, whereas features from OOD-samples should also lie in $P^{\perp}$, the space orthogonal to P. P is the D-dimensional subspace spanned by the eigenvectors with the largest D eigenvalues of the matrix $\mathbf{F}^T\mathbf{F}$, where $\mathbf{F}$ is the matrix of all train features offsetted by $\mathbf{u} = -(\mathbf{W^T})^{+}\mathbf{b}$ ($+$ denotes the Moore-Penrose inverse). A sample with feature vector $\mathbf{h}$ is then also offset to $\tilde{\mathbf{h}} = \mathbf{h} - \mathbf{u}$ and can be decomposed into $\tilde{\mathbf{h}} = \tilde{\mathbf{h}}^{P} + \tilde{\mathbf{h}}^{P^{\perp}}$, and $\tilde{\mathbf{h}}^{P^{\perp}}$ is referred to as the *Residual* of $\mathbf{h}$. ViM leverages the Residual and converts it to a virtual logit $o_0 = \alpha||\tilde{\mathbf{h}}^{P^{\perp}}||_2$, where

$$\alpha = \frac{\sum_{i=1}^{N} \max_c o_i^c}{\sum_{i=1}^{N} ||\mathbf{h}_i^{P^{\perp}}||_2}$$

is designed to match the scale of the virtual logit to the scale of the real train logits. The virtual logit is then appended to the original logits of the test sample, i.e. to $\mathbf{o}$, and a new probability vector is computed via the softmax function. The probability corresponding to the virtual logit is then the final OOD-score:

$$s = -\frac{\exp\left(o_0\right)}{\sum_{c=1}^{C} \exp\left(o_c\right) + \exp\left(o_0\right)}$$

Like suggested in (Wang et al., 2022), we use $D = 1000$ if the dimensionality of the feature space d is $d \geq 2048$, $D = 512$ if $2048 \geq d \geq 768$, and $D = d/2$ rounded to integers otherwise.

Cosine (Tack et al., 2020; Anonymous, 2023): This method computes the maximum cosine-similarity between the features of a test-sample and embedding vectors $\tilde{\mathbf{u}}_c$ (sometimes also called concept-vector):

$$s = \max_c \tilde{\mathbf{u}}_c^T \mathbf{h} / ||\tilde{\mathbf{u}}_c^T||_2 \tag{1}$$

For zero-shot CLIP, $\tilde{\mathbf{u}}_c$ can be obtained by creating text-embeddings from the ImageNet class names. Encoding *'A photo of a ...'* yields an embedding from the corresponding class. For classifiers, we use the class-wise train means $\hat{\mu}_c$, that are also used for Mahalanobis distance.

MCM/RCos (Ming et al., 2022; Techapanurak et al., 2020): Maximum-Concept-Matching was recently introduced as a zero-shot OOD detection method for CLIP and applies additional softmax-scaling to the cosine-similarities of the *Cosine* method, potentially with a temperature scaling (which we omit, following (Ming et al., 2022)). Again, we extend this method to work with conventional classifiers by using the class-means $\hat{\mu}_c$ like they are used for Mahalanobis distance as embedding/concept vectors. We then refer to it as relative cosine (short: MCM/RCos or just RCos) in order to distinguish it from CLIPs zero-shot method.

E Definitions of OOD detection metrics

The performance of OOD detectors is commonly reported in terms of the *false positive rate at a fixed true positive rate* Q, denoted as **FPR@TPRQ**, short **FPR**.This means that the detector is interpreted as making the decision to *accept* an unknown input x *if* $S(x) \geq \tau$, for a threshold τ that is chosen such that Q% of ID inputs are accepted, and *rejecting* the input as OOD *if* $S(x) < \tau$. The FPR@TPRQ counts the fraction of falsely accepted OOD inputs under this decision scheme. This means the *lower* the FPR@TPRQ, the *better* the OOD detection performance. In the OOD detection literature, the most commonly used value for Q is 95%, which we too use throughout this paper. We also report results in terms of the mean *area under the receiver-operator characteristic curve*, short **AUROC** in Table 3. It represents the probability that an ID input receives a higher score (equal scores counted half) than an OOD input when both are drawn randomly from their respective evaluation

datasets. Like for the FPR, the mean AUROC corresponds to first uniformly drawing an OOD class and then drawing a sample from that class.

F ILLUSTRATIVE EXAMPLES FROM THE CLEANING PROCESS

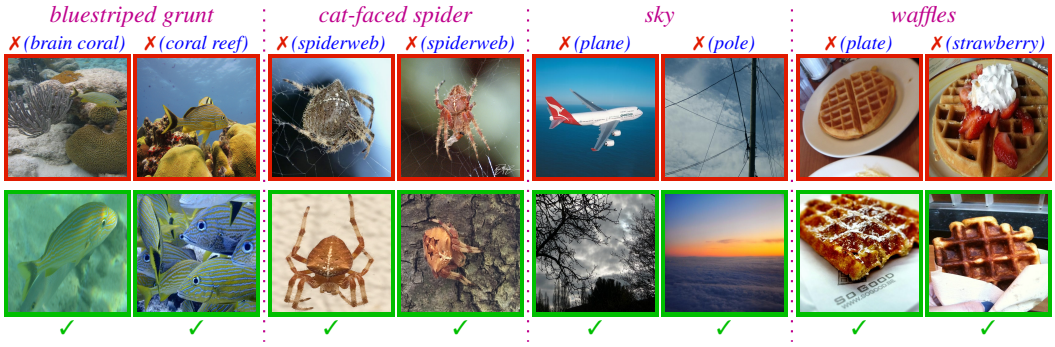

Figure 7: **Cleaning the OOD classes. Top:** Samples that were excluded due to overlap with ID classes. **Bottom:** Samples from the same OOD class that were included in the cleaned datasets.

G DETAILS OF THE NINCO DATASET.

Table 6: Detailed information for each OOD class. For determining overlap with classes of IN-21K, we checked the 8 most common predictions of a ViT classifier for IN-21K on the NINCO OOD class.

OOD class name	shortname	# samples	source dataset	ImageNet-21K overlap
AFA (cyanobacterium)	*AFA*	46	SPECIES	*microorganism*
bluestriped grunt	*BSGrunt*	99	SPECIES	*grunt*
California pitcher plant	*CPitch*	100	SPECIES	*pitcher plant*
California slender salamander	*CSSala*	100	SPECIES	*slender salamander*
California two-spot octopus	*C2SOct*	100	SPECIES	*octopus*
caracal	*Caracal*	100	iNat. Download	*caracal*
cat-faced spider	*CatFSp*	100	SPECIES	*unclear/very broad class*
Central American tapir	*CATapir*	100	SPECIES	*tapir*
chicken quesadilla	*CQuesa*	100	FOOD-101	*quesadilla*
common cuttlefish	*CCuttle*	100	SPECIES	*cuttlefish*
crème brûlée	*CBrûlée*	99	FOOD-101	*creme brulee*
cupcakes	*CCake*	80	FOOD-101	-
donuts	*Donuts*	100	FOOD-101	*doughnut*
dreamfish	*Dreamf*	100	SPECIES	*sea bream*
dune thistle	*DThist*	100	SPECIES	-
dusky flathead (fish)	*DFlath*	100	SPECIES	*flathead*
E. micromeris (cactus)	*EMicro*	100	SPECIES	-
Eastern leaf-footed bug	*ELFBug*	100	SPECIES	*leaf-footed bug*
European paper wasp	*EPWasp*	100	SPECIES	*paper wasp*
false killer whale	*FalseKW*	67	SPECIES	*unclear/very broad class*
field road	*FieldRd*	96	PLACES	*byway*
forest path	*ForPth*	100	PLACES	*unclear/very broad class*
Franciscan wallflower	*Franci*	100	SPECIES	*wallflower*
Galápagos fur seal	*GFurS*	91	SPECIES	*arcella*
giant cuttlefish	*GCuttle*	99	SPECIES	*cuttlefish*
Horse Hoof clam	*HHClam*	31	SPECIES	*seashell*
Indo-Pacific bottlenose dolphin	*IPBNDol*	100	SPECIES	*dolphin*
long-tailed silverfish	*LTSilF*	100	SPECIES	*silverfish*
Lumholtz's tree-kangaroo	*LTRoo*	100	SPECIES	*tree wallaby*
M. wesenbergii (cyanobacterium)	*MWesen*	33	SPECIES	*microorganism*
marbled newt	*MNewt*	100	SPECIES	*newt*
Mexican lime cactus	*MLCact*	100	SPECIES	*barrel cactus*
Pampas deer	*PDeer*	82	SPECIES	*buck*
redbreast sunfish	*RBSunf*	100	SPECIES	*sunfish*
rosybells (flowering plant)	*Rosyb*	100	SPECIES	-
ruby octopus	*RubyOct*	100	SPECIES	*octopus*
silver-haired bat	*SilverHB*	99	SPECIES	*bat*
skipper caterpillar	*SCaterp*	100	iNat. Download	*caterpillar*
sky	*Sky*	68	PLACES	*sky*
southern calamari	*SCalam*	99	SPECIES	*squid*
spaghetti bolognese	*SBolo*	67	FOOD-101	*spaghetti*
sweet pea	*SwPea*	100	SPECIES	*unclear/very broad class*
two-toed amphiuma (salamander)	*2TAmph*	176	SPECIES	*amphiuma*
waffles	*Waffle*	61	FOOD-101	-
yellow trumpets	*YTrump*	100	SPECIES	*yellow trumpet*
'ōhelo 'ai (flowering plant)	*'ō'ai*	100	SPECIES	-

H DETAILS AND RECIPES FOR OOD UNIT-TESTS

We provide 400 samples for each of 17 OOD unit-tests, mirroring the sizes and file formats of random ImageNet samples. Their reproducible definitions are given as follows:

- **uniform noise (Hendrycks & Gimpel, 2017):** Each RBG colour channel of each pixel is independently sampled uniformly between 0.0 or 1.0.
- **Gaussian noise (Hendrycks & Gimpel, 2017):** For each image, first σ is chosen randomly between $(0.05, 0.075, 0.1, 0.15, 0.2, 0.3, 0.5)$.
  Then each RBG colour channel of each pixel is independently sampled from $\mathcal{N}(0.5, \sigma)$.
- **Rademacher noise (Hendrycks et al., 2019):** Then each RBG colour channel of each pixel is independently set to 0.0 or 1.0 with 50% probability.
- **IN pixel permutations (Hein et al., 2019):** We choose a random IN-1K validation image and randomly shuffle its pixels (no remixing of colours).
- **black:** All colour channels are set to 0.0.
- **white:** All colour channels are set to 1.0.
- **shades of grey:** All colour channels are set to the same value, sampled uniformly between 0.0 or 1.0.
- **monochrome:** All pixels are set to a uniformly random RGB-colour (sampled uniformly from $[0.0, 1.0]^3$).
- **tricolour:** The image is split into three stripes of equal size, vertically or horizontally with probability 50%.
  Each stripe is set to an independent uniformly random RGB-colour.
- **primary tricolour:** The image is split into three stripes of equal size, vertically or horizontally with probability 50%.
  Each stripe is set to a colour where each RGB-channel value is chosen randomly as either 0.0 or 1.0.
- **horizontal stripes:** The image is split into a random number chosen between $(4, 5, 7, 10, 15, 20)$ of horizontal stripes of equal size.
  Each stripe is set to an independent uniformly random RGB-colour.
- **vertical stripes:** The image is split into a random number chosen between $(4, 5, 7, 10, 15, 20)$ of vertical stripes of equal size.
  Each stripe is set to an independent uniformly random RGB-colour.
- **smooth noise (Hein et al., 2019):** For each image, first σ is chosen randomly between $(10, 15, 25, 40, 60, 85)$.
  A uniform noise image is sampled.
  Then we apply a Gaussian filter with a kernel size of σ pixels.
  Finally, the pixel values are scaled linearly such that the minimum brightness over all channels and pixels is 0.0 and the maximum is 1.0.
- **smooth noise+:** For each image, first σ is chosen randomly between $(10, 15, 25, 40, 60, 85)$.
  A uniform noise image is sampled.
  Then we apply a Gaussian filter with a kernel size of σ pixels.
  Finally, each RGB channel is scaled linearly such that its minimum brightness over all pixels is 0.0 and the maximum is 1.0.
- **smooth color:** For each image, first σ is chosen randomly between $(10, 15, 25, 40, 60, 85)$, δ uniformly between 0.1 and 0.3, and a uniformly random RGB-colour c.
  A uniform noise image is sampled.
  Then we apply a Gaussian filter with a kernel size of σ pixels.
  Finally, each RGB channel is scaled linearly such that $c - \delta$ is the 2.5th quantile of its values and $c + \delta$ the 97.5th.
- **smooth IN pixel permutations (Hein et al., 2019):** For each image, first σ is chosen randomly between $(1, 1.5, 2, 3, 4, 6, 8)$.
  An IN pixel permutations image is sampled.
  Then we apply a Gaussian filter with a kernel size of σ pixels.

- **blobs (Hendrycks et al., 2019):** For each image, first σ is chosen randomly between $(1.5, 2, 2.5, 3, 3.5, 4)$.
  Each RBG colour channel of each pixel is independently set to 1.0 with 70% probability or 0.0 with 30%.
  Then we apply a Gaussian filter with a kernel size of σ pixels.
  Finally, all channel values below 0.75 are set to 0.0.

Where necessary, the resulting channel values are clipped to $[0, 1]$. We show samples of each unit-test in the following Appendix I in Figure 10.

I EXAMPLES IMAGES OF EACH OOD CLASS IN NINCO AND OOD UNIT-TESTS

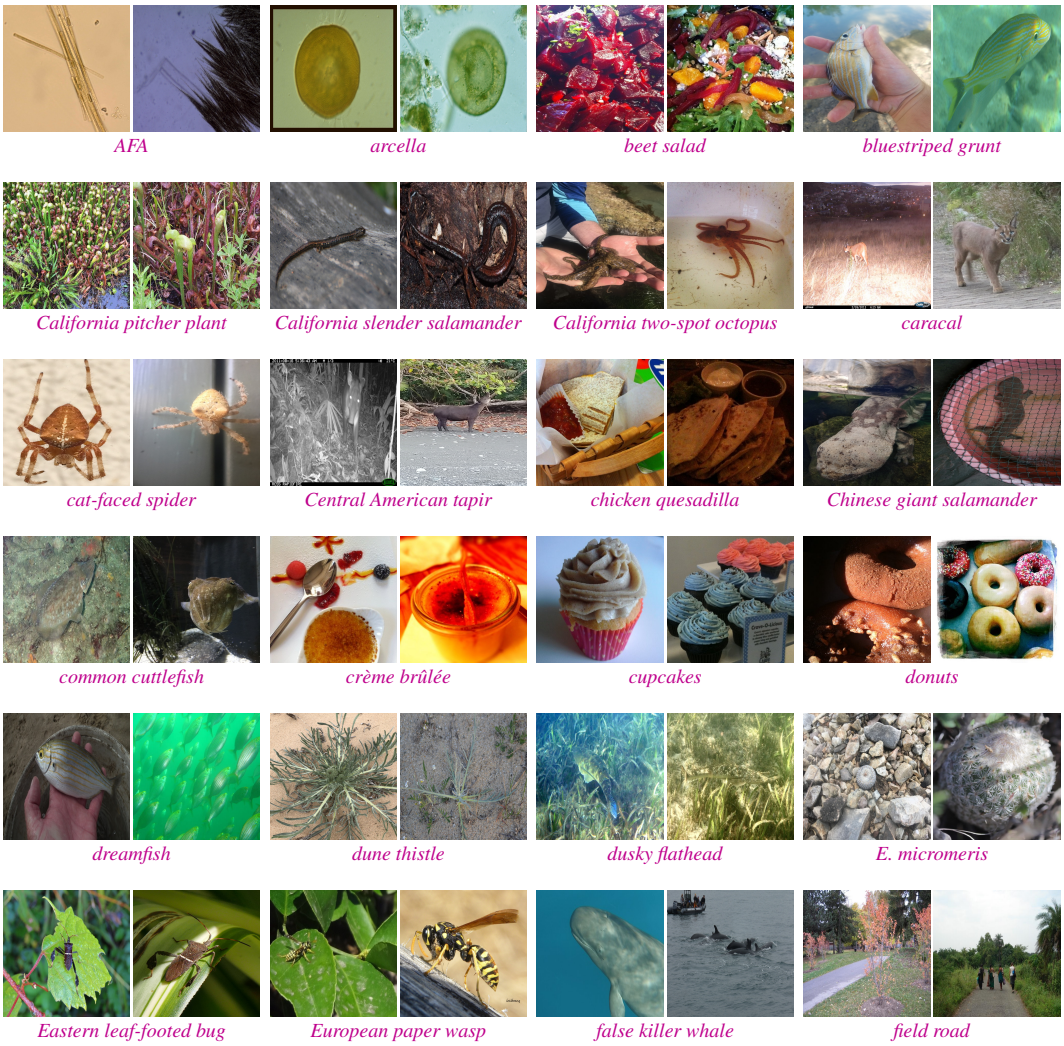

Figure 8: Samples of each class of the NINCO dataset (1/2).

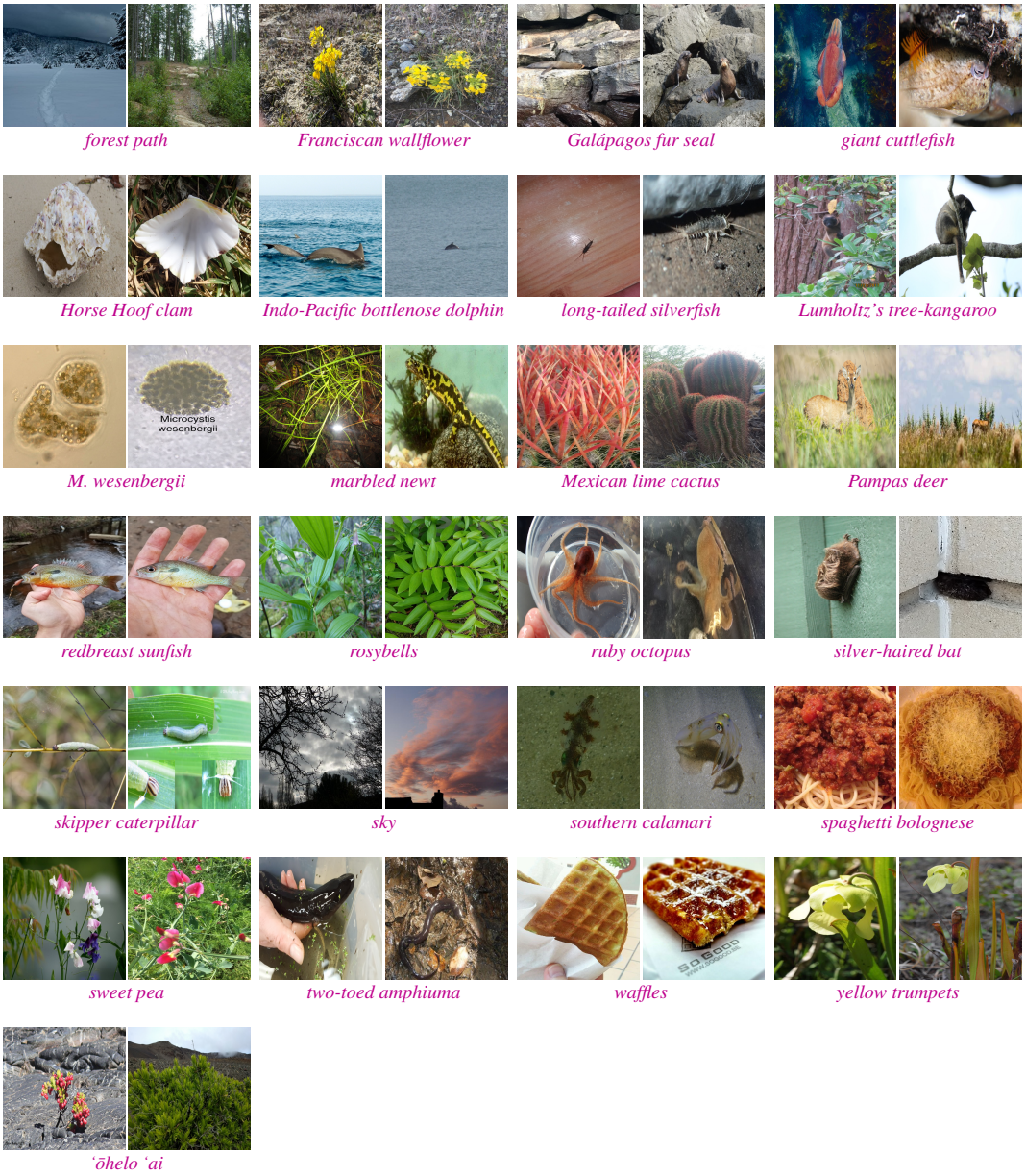

Figure 9: Samples of each class of the NINCO dataset (2/2).

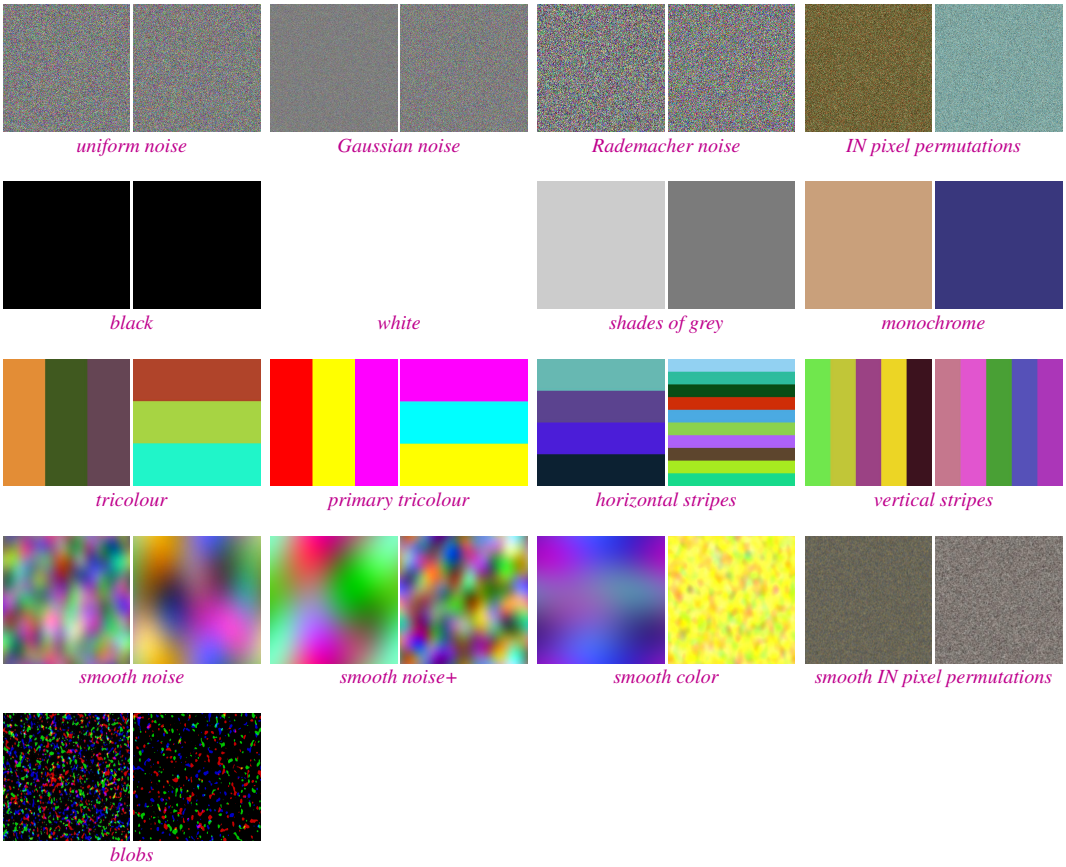

Figure 10: Samples of each OOD unit-test.

J TABLES - UNIT TESTS

Table 7: FPR of pretrained transformers for unit-tests. The ViTs pretrained *only* on ImageNet-21k fail only few unit tests, the other models struggle often with feature-dependent methods.

model	acc.	method	# fails	max	*Gauss*	*Rade*	*Black*	*Blob*	*Grey*	*Hor*	*SmN*	*SmN+*	*SmCol*	*SmPxPerm*	*Mono*	*PxPerm*	*Tri*	*PrTri*	*Uni*	*Ver*	*White*
ViT-B-384-l2b-12k	87.2	MSP	0	9.5	0.0	0.0	0.0	0.5	0.0	0.0	0.5	0.5	0.5	9.5	0.0	4.5	0.0	0.0	0.0	0.0	0.0
		MaxLogit	0	5.8	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.2	0.2	5.8	0.0	2.8	0.0	0.0	0.0	0.0	0.0
		Energy	0	6.5	0.0	0.0	0.0	0.2	0.0	0.0	0.5	0.8	0.5	6.5	0.0	2.8	0.0	0.0	0.0	0.0	0.0
		KL-Matching	0	9.5	0.0	0.0	0.0	3.5	0.0	0.0	0.5	1.0	2.5	9.5	0.0	6.0	0.0	0.0	0.0	0.0	0.0
		Mahalanobis	15	100.0	54.5	99.8	100.0	41.2	100.0	37.0	15.0	9.8	18.2	35.2	100.0	3.2	99.8	99.2	97.5	79.5	100.0
		Relative Mahalanobis	10	100.0	7.0	79.0	100.0	24.2	100.0	2.2	6.0	6.5	9.2	10.2	54.5	5.2	35.0	53.2	3.5	16.8	100.0
		ViM	5	100.0	0.0	18.8	100.0	0.0	96.2	0.0	0.2	0.0	1.0	1.5	13.5	0.2	0.8	1.2	0.2	0.2	100.0
		Energy+React	0	4.0	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.5	4.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0
		knn	0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	2.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
		cosine	0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0
		MCM/RCos	0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	1.2	0.0	0.5	0.0	0.0	0.0	0.0	0.0
ViT-B-384-l2b	86.6	MSP	1	18.5	0.0	0.0	0.0	0.0	0.0	0.8	0.8	0.5	0.5	8.2	0.0	18.5	0.0	0.0	0.0	0.0	0.0
		MaxLogit	1	11.2	0.0	0.0	0.0	0.0	0.0	0.2	0.5	0.2	0.2	2.8	0.0	11.2	0.0	0.0	0.0	0.0	0.0
		Energy	1	12.5	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.8	2.8	0.0	12.5	0.0	0.0	0.0	0.0	0.0
		KL-Matching	1	14.0	0.0	0.0	0.0	0.0	0.0	0.8	0.8	0.2	0.2	8.5	0.0	14.0	0.0	0.0	0.0	0.0	0.0
		Mahalanobis	11	100.0	64.5	100.0	100.0	24.5	100.0	1.0	0.0	0.2	0.2	14.2	100.0	2.5	16.8	40.8	99.8	0.2	100.0
		Relative Mahalanobis	8	100.0	47.0	98.5	100.0	3.2	100.0	0.0	0.0	0.0	1.5	11.8	42.5	9.0	0.8	9.5	92.2	0.0	100.0
		ViM	1	26.2	0.0	0.0	0.0	0.0	26.2	0.0	0.0	0.0	0.0	0.2	0.5	0.2	0.0	0.5	0.0	0.0	0.0
		Energy+React	0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	2.5	0.0	10.0	0.0	0.0	0.0	0.0	0.0
		knn	0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.8	0.0	0.0	0.0
		cosine	0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.8	0.0	0.0	0.0	0.0	0.0
		MCM/RCos	0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
ViT-B-384-21k	86.0	MSP	0	4.2	0.0	0.0	0.0	0.8	0.0	0.2	0.5	0.0	1.0	3.2	0.0	4.2	0.0	1.5	0.5	0.0	0.0
		MaxLogit	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Energy	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		KL-Matching	0	4.8	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	4.8	0.0	3.8	0.0	0.5	0.0	0.0	0.0
		Mahalanobis	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Relative Mahalanobis	0	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		ViM	2	100.0	0.0	0.0	0.0	0.0	46.0	0.0	0.0	0.0	0.5	0.0	0.5	0.0	0.0	5.0	0.0	0.0	100.0
		Energy+React	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		knn	0	4.8	0.0	0.0	0.0	0.0	0.0	0.0	3.2	1.8	4.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		cosine	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		MCM/RCos	0	3.5	3.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ViT-B-224-21k	84.5	MSP	0	7.5	1.0	0.0	0.0	1.8	0.0	1.5	0.5	0.2	2.5	2.2	0.0	6.0	0.2	0.5	0.2	7.5	0.0
		MaxLogit	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Energy	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		KL-Matching	2	16.5	1.8	0.0	0.0	0.0	0.0	12.0	0.5	0.0	1.8	4.0	0.0	5.8	0.2	0.5	0.2	16.5	0.0
		Mahalanobis	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Relative Mahalanobis	1	10.2	0.0	0.0	0.0	0.0	0.0	5.2	0.0	0.0	0.0	0.0	0.0	0.0	0.2	2.0	0.0	10.2	0.0
		ViM	3	100.0	0.0	0.0	100.0	0.0	55.5	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	1.5	0.0	0.0	100.0
		Energy+React	0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		knn	0	6.0	0.8	0.0	0.0	0.0	0.0	0.0	0.8	0.8	6.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		cosine	0	0.8	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0
		MCM/RCos	1	15.0	15.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
Swinv2-B-256-21k	86.3	MSP	4	27.3	0.0	0.0	0.0	0.0	0.0	1.5	22.2	27.3	5.2	11.0	0.0	7.2	2.5	3.5	0.2	14.5	0.0
		MaxLogit	2	16.0	0.0	0.0	0.0	0.0	0.0	0.0	13.2	16.0	2.8	7.8	0.0	4.5	0.0	2.5	0.0	8.0	0.0
		Energy	2	12.5	0.0	0.0	0.0	0.0	0.0	0.0	11.0	12.5	1.8	8.5	0.0	7.0	0.0	2.5	0.0	10.0	0.0
		KL-Matching	5	23.5	0.2	0.0	0.0	0.8	0.0	5.2	19.5	23.5	3.8	15.0	0.0	17.5	6.8	6.5	0.2	14.2	0.0
		Mahalanobis	13	100.0	56.8	38.2	100.0	13.2	100.0	46.2	0.8	0.8	3.0	19.8	100.0	4.5	91.0	91.8	26.0	26.2	100.0
		Relative Mahalanobis	10	100.0	17.5	8.8	100.0	0.2	100.0	45.0	0.0	0.0	0.5	24.2	100.0	6.2	77.0	79.2	0.5	15.5	100.0
		ViM	8	100.0	26.2	12.0	100.0	0.0	100.0	7.2	0.2	0.0	0.8	2.2	100.0	0.5	33.5	23.2	0.0	1.5	100.0
		Energy+React	0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	7.0	0.8	6.8	0.0	2.0	0.0	3.0	0.0	5.8	0.0
		knn	0	3.2	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.2	0.0	0.2	3.2	1.5	0.0	0.0	0.0
		cosine	0	3.8	0.0	0.0	0.0	0.0	0.0	3.8	0.0	0.0	0.0	0.2	0.0	0.0	1.5	1.5	0.0	0.0	0.0
		MCM/RCos	0	4.5	0.0	0.0	0.0	0.0	0.0	4.5	0.0	0.0	0.0	0.2	0.0	0.5	2.2	2.2	0.0	0.0	0.0
Deit3-B-384-21k	86.7	MSP	8	88.5	17.0	88.5	0.0	3.5	0.0	34.8	1.2	0.8	9.5	50.5	0.0	20.2	30.0	15.2	54.0	2.2	0.0
		MaxLogit	7	90.8	16.8	90.8	0.0	2.2	0.0	13.5	0.5	0.5	8.0	39.0	0.0	12.8	18.0	7.0	55.2	0.2	0.0
		Energy	6	93.5	22.2	93.5	0.0	2.2	0.0	7.5	0.2	0.0	8.0	31.5	0.0	11.5	11.0	4.0	66.8	0.0	0.0
		KL-Matching	8	70.8	12.8	70.8	0.0	3.0	0.0	42.2	1.5	1.2	10.0	49.0	0.0	17.2	32.8	17.0	46.5	4.0	0.0
		Mahalanobis	6	100.0	0.0	0.0	100.0	0.0	15.5	42.2	0.0	0.0	1.0	29.2	5.5	0.8	38.2	47.8	0.0	2.5	0.0
		Relative Mahalanobis	4	39.8	0.0	0.0	0.0	0.0	0.0	39.8	0.0	0.0	1.2	35.5	1.5	0.8	32.0	21.0	0.0	2.0	0.0
		ViM	7	100.0	0.2	0.0	100.0	0.0	27.0	30.5	0.8	0.0	2.8	27.5	36.2	0.5	55.2	71.2	0.0	5.8	0.0
		Energy+React	4	72.0	8.5	72.0	0.0	0.8	0.0	7.0	0.0	0.0	6.5	25.0	0.0	4.0	11.5	3.8	32.5	0.0	0.0
		knn	2	22.5	0.0	0.0	0.0	0.0	0.0	22.5	0.0	0.0	0.5	1.5	0.0	0.2	16.8	3.5	0.0	0.2	0.0
		cosine	3	22.2	0.0	0.0	0.0	0.0	0.0	22.2	0.0	0.0	0.8	13.8	0.0	0.8	16.5	3.0	0.0	0.2	0.0
		MCM/RCos	3	21.0	0.0	0.0	0.0	0.0	0.0	21.0	0.0	0.0	0.8	14.0	0.0	0.8	16.2	3.0	0.0	0.2	0.0
Deit3-B-224-21k	85.7	MSP	5	48.5	1.5	4.8	0.0	3.2	0.0	30.5	2.0	0.5	9.0	48.5	0.0	13.5	30.8	14.0	5.2	1.8	0.0
		MaxLogit	2	39.0	1.5	4.0	0.0	1.2	0.0	10.0	0.8	0.2	6.2	39.0	0.0	10.0	13.2	5.8	4.8	0.2	0.0
		Energy	1	35.8	1.5	5.2	0.0	1.0	0.0	6.5	0.5	0.0	5.2	35.8	0.0	8.8	6.2	1.5	6.2	0.2	0.0
		KL-Matching	5	47.8	1.8	4.2	0.0	3.2	0.0	38.8	3.8	1.8	9.0	47.8	0.2	14.0	38.0	19.5	4.5	6.5	0.0
		Mahalanobis	6	100.0	0.0	0.0	100.0	0.0	61.3	42.8	0.0	0.0	1.5	16.5	8.2	0.2	63.2	73.0	0.5	1.8	0.0
		Relative Mahalanobis	4	50.7	0.0	0.0	0.0	0.0	0.0	39.8	0.0	0.0	1.5	18.8	2.2	0.2	50.7	39.8	0.5	1.2	0.0
		ViM	6	100.0	0.2	0.0	100.0	0.0	53.8	38.0	0.5	0.2	3.2	15.0	9.0	0.2	61.5	80.2	0.2	1.8	0.0
		Energy+React	1	26.0	0.5	1.2	0.0	0.2	0.0	5.8	0.0	0.0	2.5	26.0	0.0	5.5	5.2	1.5	1.5	0.2	0.0
		knn	2	19.8	0.0	0.0	0.0	0.0	0.0	19.8	0.0	0.0	1.0	3.0	0.0	1.0	12.2	3.5	0.0	0.0	0.0
		cosine	3	18.2	0.0	0.0	0.0	0.0	0.0	18.2	0.0	0.0	0.8	10.8	0.0	0.2	10.2	4.2	0.2	0.0	0.0
		MCM/RCos	3	18.8	0.0	0.0	0.0	0.0	0.0	18.8	0.0	0.0	0.5	10.2	0.0	0.2	11.0	4.2	0.2	0.0	0.0
clip-ViT-B-224	66.6	MCM	1	19.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	4.5	4.2	0.0	19.8	0.0
		cosine	15	100.0	99.2	100.0	100.0	38.5	1.5	25.8	85.0	89.0	65.0	73.2	53.5	96.2	12.8	31.8	99.5	71.5	0.0
clip-ViT-L-336	74.2	MCM	0	4.2	0.0	0.0	0.0	4.2	0.0	0.2	0.0	0.0	0.0	1.5	1.2	0.2	4.2	0.5	0.0	0.2	0.0
		cosine	14	100.0	96.0	100.0	100.0	58.8	48.2	4.5	97.2	96.0	84.8	81.2	88.8	94.2	6.5	34.2	100.0	32.0	0.0

Table 8: FPR of pretrained convolutional networks for OOD unit-tests.

model	acc.	method	# fails	max	*Gauss*	*Rade*	*Black*	*Blob*	*Grey*	*Hor*	*SmN*	*SmN+*	*SmCol*	*SmPxPerm*	*Mono*	*PxPerm*	*Tri*	*PrTri*	*Uni*	*Ver*	*White*
CnvNxt-B-21k	86.3	MSP	0	8.2	0.0	0.0	0.0	0.2	0.0	6.0	1.8	5.0	2.0	5.8	0.0	0.5	1.2	8.2	0.2	0.2	0.0
		MaxLogit	0	5.2	0.0	0.0	0.0	0.0	0.0	1.5	2.0	5.2	1.0	2.8	0.0	0.0	0.2	4.0	0.0	0.2	0.0
		Energy	0	9.2	0.0	0.0	0.0	0.0	0.0	1.0	5.0	9.2	2.0	2.5	0.0	0.0	0.0	3.0	0.0	0.2	0.0
		KL-Matching	2	21.0	0.0	0.0	0.0	5.2	0.0	21.0	3.2	2.0	2.5	6.5	0.0	1.5	2.5	10.8	0.0	0.2	0.0
		Mahalanobis	4	100.0	0.8	42.5	100.0	0.0	100.0	0.0	0.0	0.0	0.0	2.5	2.8	0.0	0.0	4.8	0.2	0.0	100.0
		Relative Mahalanobis	6	100.0	0.5	41.0	100.0	0.0	100.0	1.0	0.0	0.0	0.0	16.5	1.5	0.0	0.0	11.5	0.0	0.0	100.0
		ViM	4	100.0	0.5	41.2	100.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	2.2	0.0	0.0	100.0
		Energy+React	0	3.8	0.0	0.0	0.0	0.0	0.0	0.5	1.8	3.8	0.0	1.0	0.0	0.0	0.0	1.8	0.0	0.2	0.0
		knn	0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0
		cosine	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		MCM/RCos	0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CnvNxt-T-21k	84.1	MSP	2	16.2	4.0	3.8	0.0	2.2	0.0	7.0	8.8	9.2	5.8	12.8	0.5	16.2	6.5	7.8	0.8	0.5	0.0
		MaxLogit	4	25.5	18.0	22.2	0.0	0.8	0.0	1.5	3.0	5.0	4.0	9.0	0.0	10.5	1.5	1.2	25.5	0.2	0.0
		Energy	3	65.0	35.0	65.0	0.0	0.5	0.0	0.5	1.0	2.5	2.5	7.2	0.0	8.8	0.2	0.0	44.8	0.0	0.0
		KL-Matching	7	26.5	9.2	2.5	0.0	9.5	0.0	23.0	12.0	13.2	6.8	15.2	1.8	26.5	19.8	17.2	1.0	4.8	0.0
		Mahalanobis	7	100.0	4.0	11.5	100.0	0.0	100.0	1.0	0.0	0.0	0.5	0.0	42.5	0.0	12.2	12.8	7.0	0.0	100.0
		Relative Mahalanobis	4	100.0	0.5	7.5	100.0	0.0	100.0	1.5	0.0	0.0	0.8	1.2	17.0	1.2	4.5	8.5	5.0	0.0	100.0
		ViM	4	100.0	1.0	7.2	100.0	0.0	100.0	0.2	0.0	0.0	0.0	0.0	14.0	0.0	1.0	5.5	5.8	0.0	100.0
		Energy+React	3	42.2	23.2	42.2	0.0	0.2	0.0	0.5	0.2	1.2	2.0	5.0	0.0	5.5	0.0	0.0	29.8	0.0	0.0
		knn	2	18.5	0.0	0.0	0.0	0.0	0.0	18.2	0.0	0.0	1.0	0.8	0.0	0.0	9.5	18.5	0.0	0.0	0.0
		cosine	0	6.8	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.8	0.2	0.0	1.2	3.2	6.8	0.0	0.0	0.0
		MCM/RCos	0	9.8	0.0	0.0	0.0	0.0	0.0	8.8	0.0	0.0	0.8	0.2	0.0	1.2	5.0	9.8	0.0	0.0	0.0
BiT-m	82.3	MSP	12	92.8	22.8	22.0	0.0	6.5	0.0	11.0	43.5	41.0	50.2	92.8	13.5	80.8	38.5	24.2	12.2	5.2	0.0
		MaxLogit	7	88.0	8.0	6.8	0.0	1.8	0.0	2.5	42.0	41.8	47.0	88.0	6.2	70.8	16.2	12.0	3.0	2.2	0.0
		Energy	8	86.8	8.2	10.2	0.0	1.5	0.0	1.2	66.2	68.5	60.5	86.8	6.0	71.5	13.2	13.5	3.5	7.0	0.0
		KL-Matching	11	90.2	41.2	48.0	0.0	5.0	0.0	8.5	36.0	35.5	38.5	90.2	12.8	73.8	31.5	19.5	36.8	4.2	0.0
		Mahalanobis	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Relative Mahalanobis	0	2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.2	1.2	1.0	2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		ViM	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Energy+React	8	63.7	2.5	3.0	0.0	0.5	0.0	1.0	61.8	62.7	57.2	63.7	4.0	31.8	15.2	16.2	3.8	32.8	0.0
		knn	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		cosine	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		MCM/RCos	0	1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	1.2	1.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0
EffNetv2-M-21k	85.6	MSP	4	52.8	0.8	0.0	0.0	1.2	0.0	0.5	10.2	9.0	19.8	52.8	0.5	13.8	0.5	0.0	0.5	0.0	0.0
		MaxLogit	3	48.5	0.5	0.0	0.0	0.8	0.0	0.5	6.0	5.5	13.8	48.5	1.2	14.8	0.0	0.0	0.0	0.0	0.0
		Energy	4	57.0	10.2	0.0	0.0	6.2	0.0	0.5	9.5	8.5	17.0	57.0	3.2	26.8	0.0	0.0	0.0	0.0	0.0
		KL-Matching	5	50.2	0.8	0.8	0.0	3.8	0.0	0.8	17.5	16.5	23.2	50.2	0.2	13.8	1.8	0.0	0.5	1.2	0.0
		Mahalanobis	0	3.8	0.0	0.0	0.0	0.0	0.0	0.2	1.8	1.5	3.8	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0
		Relative Mahalanobis	2	100.0	0.0	0.0	0.0	0.0	3.8	0.2	5.0	2.2	15.5	5.5	0.0	0.2	0.5	5.8	0.0	0.0	100.0
		ViM	0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Energy+React	14	100.0	91.5	99.2	0.0	100.0	1.0	59.8	99.0	99.5	92.8	98.0	28.5	83.8	30.2	27.0	98.5	79.5	0.0
		knn	0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		cosine	0	4.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		MCM/RCos	1	12.0	0.0	0.0	0.0	0.0	0.0	0.2	1.2	0.8	12.0	0.5	0.0	0.2	0.0	4.2	0.0	0.0	0.0
EffNetb7-ns	86.8	MSP	0	9.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.5	0.2	0.0	4.0	0.0	0.0	0.0	9.5	0.0
		MaxLogit	0	8.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.2	0.0	0.0	0.0	8.0	0.0
		Energy	3	29.5	15.8	29.5	0.0	4.0	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.5	0.0	0.0	12.8	9.2	0.0
		KL-Matching	1	24.2	0.0	0.0	0.0	0.5	0.0	2.8	0.0	1.2	0.8	3.8	0.0	24.2	0.2	0.2	0.8	9.8	0.0
		Mahalanobis	14	100.0	8.5	0.8	100.0	11.8	100.0	100.0	65.0	63.2	88.5	67.2	100.0	55.0	100.0	98.5	1.5	99.2	100.0
		Relative Mahalanobis	13	100.0	0.0	0.0	100.0	0.0	100.0	93.8	20.8	19.2	42.8	35.8	99.8	10.2	99.0	81.0	0.0	98.0	100.0
		ViM	15	100.0	10.5	1.2	100.0	21.8	100.0	100.0	70.8	67.0	90.8	70.0	100.0	55.5	100.0	99.0	3.0	99.2	100.0
		Energy+React	0	6.8	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	6.8	0.0
		knn	0	7.2	0.0	0.0	0.0	0.5	0.0	0.2	0.0	0.0	0.0	0.0	0.0	7.2	0.0	0.0	0.0	0.0	0.0
		cosine	0	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
		MCM/RCos	0	0.8	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0

Table 9: FPR of transformers without pretraining for OOD unit-tests.

model	acc.	method	# fails	max	*Gauss*	*Rade*	*Black*	*Blob*	*Grey*	*Hor*	*SmN*	*SmN+*	*SmCol*	*SmPxPerm*	*Mono*	*PxPerm*	*Tri*	*PrTri*	*Uni*	*Ver*	*White*
ViT-B-384	81.1	MSP	7	42.8	0.0	0.0	0.0	22.0	0.0	22.2	11.5	10.5	12.8	5.5	0.0	42.8	0.5	10.8	1.0	5.0	0.0
		MaxLogit	2	18.2	0.0	0.0	0.0	7.0	0.0	15.2	4.8	5.8	6.2	1.0	0.0	18.2	0.5	7.0	0.0	1.8	0.0
		Energy	2	15.2	3.0	0.0	0.0	1.5	0.0	15.2	3.0	3.0	5.8	0.5	0.0	12.0	0.5	8.2	0.0	2.0	0.0
		KL-Matching	6	42.8	0.0	0.0	0.0	18.5	3.8	23.8	7.8	8.2	8.8	4.2	2.8	42.8	11.5	12.2	0.2	27.5	0.0
		Mahalanobis	4	81.0	19.8	0.5	0.0	0.0	81.0	20.5	0.8	0.5	0.2	1.5	8.8	21.2	9.2	3.2	7.5	0.8	0.0
		Relative Mahalanobis	9	100.0	7.0	0.0	100.0	1.0	100.0	24.5	1.0	1.0	1.5	15.2	49.0	23.0	49.8	16.8	0.2	9.8	100.0
		ViM	4	90.0	71.2	4.5	0.0	0.0	0.0	28.2	4.2	3.8	0.5	2.2	0.0	28.2	0.5	1.8	90.0	0.5	0.0
		Energy+React	2	18.0	1.2	0.0	0.0	2.2	0.0	18.0	2.8	3.8	6.0	0.5	0.2	12.5	0.5	7.8	0.0	4.0	0.0
		knn	4	76.2	76.2	48.5	0.0	3.2	0.0	18.8	3.8	3.8	6.8	1.0	1.0	8.0	0.2	10.0	61.3	0.0	0.0
		cosine	3	20.2	20.2	0.0	0.0	3.5	0.0	11.0	2.8	3.5	3.2	0.8	0.8	13.2	0.2	6.8	1.0	0.0	0.0
		MCM/RCos	3	27.5	27.5	0.0	0.0	4.8	0.0	12.2	3.5	3.5	4.5	0.8	1.0	15.2	0.5	8.2	1.2	0.0	0.0
Swinv2-B-256	84.6	MSP	3	30.0	0.0	0.0	0.0	14.5	0.0	30.0	9.8	9.8	3.8	9.0	0.0	23.0	3.2	2.2	0.5	4.0	0.0
		MaxLogit	1	21.2	0.0	0.0	0.0	9.2	0.0	9.0	5.2	4.5	2.0	4.8	0.0	21.2	0.5	1.2	0.0	2.0	0.0
		Energy	1	29.5	0.0	0.0	0.0	10.0	0.0	4.0	2.8	2.8	1.0	3.8	0.0	29.5	0.2	0.2	0.0	2.2	0.0
		KL-Matching	6	51.0	0.2	0.0	0.0	16.0	0.0	51.0	15.2	12.5	6.5	10.5	0.0	23.5	9.0	4.5	0.5	4.2	0.0
		Mahalanobis	12	100.0	55.0	57.8	100.0	5.8	100.0	61.0	7.0	5.0	5.5	18.2	100.0	2.5	97.8	93.2	54.5	34.2	100.0
		Relative Mahalanobis	12	100.0	23.0	47.5	100.0	1.5	100.0	62.7	5.8	3.8	3.0	13.5	100.0	2.2	97.5	88.8	49.2	34.0	100.0
		ViM	12	100.0	21.0	46.0	100.0	1.0	100.0	53.5	9.0	5.5	7.8	13.5	100.0	4.0	97.5	98.0	49.0	34.0	100.0
		Energy+React	0	7.8	0.0	0.2	0.0	4.0	0.0	0.8	2.0	1.2	1.0	2.5	0.0	7.8	0.2	0.2	0.0	0.2	0.0
		knn	2	37.5	0.0	0.0	0.0	0.0	0.0	37.5	4.5	3.0	0.8	0.0	6.2	1.5	12.0	0.5	0.0	0.0	0.0
		cosine	1	28.5	0.0	0.0	0.0	0.0	0.0	28.5	3.0	1.0	1.0	0.0	0.2	0.8	4.5	0.0	0.0	0.0	0.0
		MCM/RCos	1	38.2	0.0	0.0	0.0	0.0	0.0	38.2	3.0	1.5	0.8	0.5	0.0	1.0	6.0	0.5	0.0	0.0	0.0
Deit3-B-384	85.1	MSP	2	37.2	0.0	0.0	0.0	0.0	0.0	37.2	9.5	16.8	3.8	4.2	0.0	9.0	0.0	1.2	0.0	0.5	0.0
		MaxLogit	1	16.2	0.0	0.0	0.0	0.0	0.0	16.2	3.8	8.8	2.5	1.2	0.0	7.5	0.0	0.8	0.0	0.0	0.0
		Energy	0	6.5	0.0	0.0	0.0	0.0	0.0	4.0	3.2	6.2	2.8	1.5	0.0	6.5	0.0	0.0	0.0	0.2	0.0
		KL-Matching	4	63.0	0.0	0.0	0.0	0.0	0.0	63.0	13.2	19.2	6.5	6.5	0.0	11.5	0.8	5.2	0.0	3.5	0.0
		Mahalanobis	14	100.0	99.2	99.8	100.0	74.2	100.0	98.8	6.5	1.5	32.8	47.2	100.0	7.2	100.0	100.0	99.8	97.0	100.0
		Relative Mahalanobis	14	100.0	77.8	98.5	100.0	31.2	100.0	97.0	2.5	1.0	15.0	40.2	100.0	4.5	100.0	100.0	85.8	85.8	100.0
		ViM	8	100.0	0.0	0.0	100.0	0.0	100.0	82.5	0.2	0.5	1.0	0.0	100.0	0.0	99.2	88.0	0.0	47.2	100.0
		Energy+React	6	58.0	0.0	0.0	0.0	0.0	0.0	35.8	48.5	58.0	26.0	23.0	0.0	53.2	0.0	1.5	0.0	0.5	0.0
		knn	14	100.0	91.8	99.0	100.0	58.0	100.0	96.8	8.5	2.8	32.2	12.8	100.0	2.8	99.8	95.8	98.2	56.2	100.0
		cosine	7	100.0	19.2	0.5	100.0	0.0	66.5	89.0	0.5	0.5	2.8	4.8	96.8	1.5	51.7	22.0	0.2	0.5	0.0
		MCM/RCos	1	11.8	0.0	0.0	0.0	0.0	0.0	11.8	0.0	0.2	0.5	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0
Deit3-B-224	83.8	MSP	3	57.5	0.0	0.2	0.0	0.2	0.0	57.5	20.5	25.0	5.8	5.2	0.0	6.8	10.0	7.8	0.0	0.8	0.0
		MaxLogit	3	51.2	0.0	0.0	0.0	0.2	0.0	51.2	15.2	23.2	4.8	3.8	0.0	5.2	4.0	3.8	0.0	0.5	0.0
		Energy	3	47.5	0.0	0.0	0.0	0.0	0.0	47.5	12.5	22.0	2.2	3.0	0.0	8.2	0.5	1.0	0.0	5.8	0.0
		KL-Matching	7	71.0	0.5	2.0	0.0	2.8	0.0	71.0	25.2	30.0	11.8	6.2	0.0	13.0	34.5	25.0	1.5	5.5	0.0
		Mahalanobis	16	100.0	100.0	100.0	100.0	63.0	100.0	88.2	31.0	25.8	61.3	55.0	100.0	2.8	100.0	100.0	100.0	86.2	100.0
		Relative Mahalanobis	16	100.0	92.8	87.2	100.0	38.5	100.0	85.5	16.0	15.5	29.8	40.2	100.0	1.2	100.0	100.0	84.8	69.5	100.0
		ViM	8	100.0	0.2	0.0	100.0	0.0	100.0	69.2	2.0	5.2	0.5	0.8	97.8	0.2	49.0	63.0	0.0	27.5	100.0
		Energy+React	5	45.8	0.8	0.0	0.0	0.5	0.0	45.8	21.2	27.5	5.2	10.8	0.0	28.0	0.2	1.2	0.0	0.0	0.0
		knn	15	100.0	38.5	33.0	100.0	32.0	100.0	74.8	35.5	31.2	53.8	8.5	100.0	4.0	95.8	91.0	14.5	31.2	100.0
		cosine	7	100.0	8.5	2.5	100.0	0.2	100.0	70.0	4.2	7.5	2.0	1.2	100.0	1.2	49.2	35.8	1.2	0.8	100.0
		MCM/RCos	1	58.0	0.0	0.0	0.0	0.0	0.0	58.0	0.5	2.0	0.5	0.2	0.0	0.2	8.2	4.2	0.0	0.0	0.0

Table 10: FPR of convolutional networks without pretraining for OOD unit-tests.

model	acc.	method	# fails	max	*Gauss*	*Rade*	*Black*	*Blob*	*Grey*	*Hor*	*SmN*	*SmN+*	*SmCol*	*SmPxPerm*	*Mono*	*PxPerm*	*Tri*	*PrTri*	*Uni*	*Ver*	*White*
XCiT-M-224	82.6	MSP	2	47.2	0.0	0.0	0.0	2.5	0.0	37.2	1.8	3.2	3.2	2.0	0.0	10.0	13.0	4.5	0.0	0.0	0.0
		MaxLogit	1	39.0	0.0	0.0	0.0	0.8	0.0	39.0	2.0	4.2	1.5	1.0	0.0	4.8	7.0	3.5	0.0	0.0	0.0
		Energy	3	57.0	0.0	0.0	0.0	1.0	0.0	57.0	19.0	31.2	4.0	1.8	0.0	9.8	5.5	6.8	0.0	3.5	0.0
		KL-Matching	4	50.0	0.5	0.0	0.0	4.8	0.0	50.0	5.8	7.5	6.8	9.8	0.0	15.8	38.0	16.8	0.0	1.5	0.0
		Mahalanobis	16	100.0	100.0	100.0	100.0	96.0	100.0	100.0	13.5	10.2	43.2	56.5	100.0	9.0	99.8	98.8	99.8	53.0	100.0
		Relative Mahalanobis	15	100.0	99.8	99.8	100.0	90.0	100.0	100.0	13.5	9.0	41.8	54.8	100.0	9.0	99.8	98.5	99.8	52.5	100.0
		ViM	12	100.0	63.5	98.8	100.0	10.0	100.0	99.8	2.8	1.5	9.0	16.0	100.0	6.0	96.2	86.2	96.0	27.5	100.0
		Energy+React	3	67.0	1.8	0.2	0.0	1.0	0.0	67.0	10.2	16.8	3.8	1.0	0.0	8.2	6.2	7.5	0.0	1.0	0.0
		knn	6	77.5	12.2	77.5	0.0	3.8	0.0	49.0	0.5	0.0	1.2	0.2	68.5	3.0	23.0	7.5	27.8	0.0	0.0
		cosine	1	27.0	0.0	0.0	0.0	3.0	0.0	27.0	0.5	0.0	1.0	0.5	0.0	2.2	9.2	1.5	0.0	0.0	0.0
		MCM/RCos	2	54.2	0.0	0.0	0.0	3.2	0.0	54.2	0.5	0.0	1.8	1.0	0.0	2.8	18.2	2.0	0.0	0.0	0.0
XCiT-M-224-d	84.3	MSP	6	100.0	8.0	2.2	0.0	0.8	54.2	64.5	6.5	8.8	6.2	17.8	3.2	34.5	10.2	7.8	2.5	0.8	100.0
		MaxLogit	1	12.8	0.0	0.0	0.0	0.0	0.0	12.8	1.2	3.8	3.5	7.5	0.0	9.8	0.2	0.0	0.0	0.0	0.0
		Energy	1	11.8	0.0	0.0	0.0	0.0	0.0	1.8	1.8	5.2	3.8	4.8	0.0	11.8	0.0	0.0	0.0	0.0	0.0
		KL-Matching	12	100.0	24.0	16.8	100.0	3.5	82.5	84.8	5.0	5.2	6.2	22.0	14.5	38.5	36.2	24.8	12.5	1.8	100.0
		Mahalanobis	12	100.0	17.2	92.2	100.0	8.5	100.0	86.2	1.5	1.2	5.0	12.2	100.0	4.0	100.0	96.8	23.0	48.0	100.0
		Relative Mahalanobis	10	100.0	5.5	56.0	100.0	1.8	100.0	85.8	2.0	1.5	3.2	13.5	85.8	3.5	98.8	94.8	7.2	47.8	100.0
		ViM	9	100.0	4.0	57.2	100.0	0.0	100.0	79.2	0.0	0.0	0.2	2.8	96.8	1.0	97.0	80.8	5.8	38.2	100.0
		Energy+React	0	7.2	0.2	0.0	0.0	0.0	0.0	3.8	0.0	0.2	3.0	3.8	0.0	7.2	0.0	0.2	0.0	0.0	0.0
		knn	2	69.2	0.0	0.0	0.0	0.0	0.0	69.2	0.0	0.0	0.2	1.8	0.0	1.0	26.5	6.2	0.0	0.0	0.0
		cosine	4	100.0	0.0	0.0	100.0	0.0	21.0	64.8	0.0	0.0	0.0	1.5	0.2	0.5	22.0	4.0	0.0	0.0	0.0
		MCM/RCos	6	100.0	0.0	0.0	100.0	0.0	52.8	74.8	0.0	0.0	0.2	2.2	1.5	0.8	28.2	11.5	0.0	0.0	100.0
CnvNxt-B	84.4	MSP	3	65.8	0.0	0.0	0.0	12.2	0.0	65.8	0.8	1.5	3.0	10.0	0.0	30.8	0.2	3.5	0.0	0.0	0.0
		MaxLogit	4	73.0	2.2	0.2	0.0	15.5	0.0	73.0	2.5	4.2	6.8	12.2	0.0	51.2	0.2	4.2	0.0	0.5	0.0
		Energy	12	100.0	61.5	49.2	0.0	79.8	0.0	90.8	97.8	100.0	86.0	50.7	3.0	98.0	9.5	36.5	40.5	62.5	0.0
		KL-Matching	12	100.0	25.8	46.0	100.0	25.5	100.0	58.8	1.2	1.5	7.8	15.2	68.5	22.2	9.5	16.8	48.8	1.2	100.0
		Mahalanobis	16	100.0	100.0	100.0	100.0	71.0	100.0	88.2	25.0	15.0	49.0	72.0	100.0	1.5	100.0	100.0	100.0	78.2	100.0
		Relative Mahalanobis	15	100.0	100.0	99.0	100.0	57.0	100.0	91.8	15.8	8.0	36.0	69.8	100.0	1.0	100.0	99.8	99.8	70.0	100.0
		ViM	16	100.0	100.0	100.0	100.0	69.2	100.0	98.5	23.8	14.0	47.5	78.2	100.0	1.0	100.0	100.0	100.0	88.5	100.0
		Energy+React	10	86.8	51.0	46.8	0.0	41.8	0.0	83.2	49.0	60.8	38.5	28.2	1.8	86.8	1.8	9.5	36.5	7.0	0.0
		knn	11	100.0	84.0	58.2	0.0	5.0	56.2	84.5	6.5	2.8	19.5	0.8	88.5	0.5	100.0	96.0	83.8	59.2	100.0
		cosine	1	35.0	0.0	0.0	0.0	0.0	0.0	35.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.8	0.0	0.0	0.0
		MCM/RCos	1	54.5	0.0	0.0	0.0	0.0	0.0	54.5	0.0	0.0	0.0	2.2	0.0	0.2	1.2	1.2	0.0	0.0	0.0
BiT-s	78.0	MSP	8	86.5	4.2	6.8	0.0	86.5	0.0	77.0	37.0	36.5	9.8	9.0	6.5	46.0	11.0	10.2	4.5	70.0	0.0
		MaxLogit	6	81.8	3.2	6.0	0.0	81.8	0.0	55.0	22.0	21.5	3.8	2.5	3.5	26.2	0.8	1.0	3.2	61.0	0.0
		Energy	6	82.2	3.0	6.0	0.0	82.2	0.0	46.5	16.0	16.0	1.5	0.5	1.8	17.8	0.2	0.8	3.0	58.5	0.0
		KL-Matching	14	100.0	4.0	4.8	100.0	36.2	100.0	70.0	19.0	15.2	21.2	11.8	91.0	22.5	47.8	52.2	8.0	71.0	100.0
		Mahalanobis	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Relative Mahalanobis	2	13.8	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	12.0	0.0	13.8	0.2	0.0	0.0	0.0	0.0
		ViM	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		Energy+React	3	69.8	1.0	1.2	0.0	69.8	0.0	21.2	0.2	0.0	0.0	0.2	0.0	1.0	0.0	0.0	0.5	40.5	0.0
		knn	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		cosine	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		MCM/RCos	9	94.5	3.8	4.2	0.0	6.0	0.0	94.5	18.0	19.0	12.2	5.0	34.0	23.5	93.5	84.0	3.8	53.8	0.0
EffNetv2-M	85.0	MSP	1	11.5	0.0	0.0	0.0	4.5	0.0	11.5	2.0	3.5	0.5	0.0	0.0	3.8	0.0	0.0	0.0	0.0	0.0
		MaxLogit	0	1.8	0.0	0.0	0.0	1.0	0.0	0.5	1.0	1.8	0.2	0.0	0.0	1.5	0.0	0.0	0.0	0.0	0.0
		Energy	0	1.8	0.0	0.0	0.0	0.0	0.0	0.2	0.5	0.8	1.8	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
		KL-Matching	2	42.8	0.2	0.0	0.0	14.8	0.0	42.8	3.5	4.0	2.5	0.5	0.0	7.0	1.2	0.5	0.0	0.0	0.0
		Mahalanobis	12	100.0	38.0	4.8	100.0	27.3	100.0	91.2	2.0	2.2	17.8	67.0	100.0	1.2	99.2	99.8	7.2	70.0	100.0
		Relative Mahalanobis	9	100.0	0.2	0.2	100.0	2.2	100.0	86.8	1.0	0.5	6.5	25.2	100.0	0.5	96.5	89.0	0.0	65.0	100.0
		ViM	14	100.0	79.0	13.0	100.0	41.5	100.0	90.8	4.5	3.7	32.8	65.2	100.0	1.5	91.8	97.5	40.2	64.2	100.0
		Energy+React	0	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		knn	7	100.0	0.0	0.0	100.0	0.2	100.0	65.5	0.5	1.0	1.8	1.0	39.2	0.2	28.2	19.8	0.0	3.2	100.0
		cosine	1	39.5	0.0	0.0	0.0	0.0	0.0	39.5	0.2	0.5	0.2	0.0	0.0	0.2	9.2	4.5	0.0	0.0	0.0
		MCM/RCos	1	14.5	0.0	0.0	0.0	0.0	0.0	14.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0
EffNetb7	84.9	MSP	3	17.0	0.0	0.0	0.0	1.2	0.0	17.0	12.8	14.0	4.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0
		MaxLogit	1	10.5	0.0	0.0	0.0	0.8	0.0	5.0	8.2	10.5	3.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
		Energy	0	7.0	0.0	0.2	0.0	0.2	0.0	0.8	4.8	6.0	7.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
		KL-Matching	3	29.0	0.0	0.0	0.0	4.5	0.0	29.0	16.8	15.5	4.8	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.0
		Mahalanobis	16	100.0	30.5	15.0	100.0	94.5	100.0	99.2	75.5	66.2	75.2	88.8	100.0	16.8	100.0	100.0	7.5	99.5	100.0
		Relative Mahalanobis	14	100.0	1.8	7.8	100.0	31.0	100.0	97.2	35.5	31.0	34.8	88.2	100.0	12.8	100.0	99.5	4.2	99.2	100.0
		ViM	15	100.0	17.5	12.5	100.0	82.5	100.0	99.0	73.2	62.7	75.8	76.5	100.0	8.5	100.0	100.0	4.2	99.5	100.0
		Energy+React	0	6.5	0.0	0.0	0.0	0.2	0.0	0.5	5.5	6.5	3.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		knn	6	100.0	0.0	0.0	0.0	0.0	36.8	54.2	15.8	17.0	4.8	0.0	45.8	0.5	1.0	3.2	0.0	0.0	100.0
		cosine	4	46.8	0.0	0.0	0.0	0.2	19.2	46.8	14.5	13.5	2.8	0.0	3.5	0.0	0.8	0.8	0.0	0.0	0.0
		MCM/RCos	1	34.5	0.0	0.0	0.0	0.0	0.0	34.5	9.8	8.8	1.5	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0
EffNet-B0	77.7	MSP	5	28.2	0.0	0.0	0.0	4.0	0.0	20.5	23.5	28.2	14.5	1.8	0.0	21.0	0.0	0.8	0.0	3.0	0.0
		MaxLogit	5	23.2	0.0	0.0	0.0	2.0	0.0	12.0	22.5	25.2	12.2	0.5	0.0	15.8	0.0	0.5	0.0	5.8	0.0
		Energy	5	31.8	0.0	0.0	0.0	1.2	0.0	7.5	30.5	31.8	18.2	0.5	0.0	26.5	0.2	3.5	0.0	26.8	0.0
		KL-Matching	17	100.0	39.5	93.0	100.0	17.8	100.0	77.5	43.2	43.5	30.0	25.0	89.2	22.5	95.2	67.0	73.8	68.2	100.0
		Mahalanobis	17	100.0	100.0	100.0	100.0	99.0	100.0	91.2	58.2	51.0	70.5	99.2	100.0	21.0	100.0	99.2	100.0	73.5	100.0
		Relative Mahalanobis	17	100.0	98.2	99.2	100.0	86.0	100.0	84.5	59.0	54.0	64.0	99.0	100.0	51.5	99.8	95.8	100.0	78.2	100.0
		ViM	15	100.0	62.0	99.5	100.0	44.2	100.0	73.8	14.5	9.5	41.0	76.8	100.0	0.2	98.2	78.5	99.2	57.5	100.0
		Energy+React	0	7.5	0.0	0.0	0.0	0.2	0.0	0.0	3.0	5.5	5.0	0.0	0.0	0.0	0.0	1.2	0.0	7.5	0.0
		knn	2	62.5	6.5	62.5	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	1.5	0.2	0.0	0.0	18.2	0.0	0.0
		cosine	0	2.5	0.0	0.0	0.0	0.0	0.0	0.0	2.5	2.5	0.2	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
		MCM/RCos	9	97.0	42.0	97.0	0.0	4.2	34.0	17.5	26.5	30.0	10.0	0.8	15.2	6.5	8.5	20.8	95.0	4.0	0.0
ResNet50	80.4	MSP	3	73.0	0.0	0.0	0.0	11.8	0.0	73.0	5.5	9.2	7.5	6.8	0.0	12.2	4.8	4.2	0.0	2.5	0.0
		MaxLogit	9	86.5	0.0	0.0	0.0	13.0	0.0	86.5	11.0	14.8	11.2	11.5	0.0	19.0	11.8	13.0	0.0	8.8	0.0
		Energy	17	100.0	98.2	92.8	100.0	87.8	100.0	99.8	63.0	61.5	84.0	96.0	100.0	49.0	99.2	97.0	95.5	84.8	100.0
		KL-Matching	4	81.8	0.0	0.0	0.0	13.0	0.0	81.8	3.0	5.2	8.0	5.5	0.0	9.8	11.0	12.2	0.0	7.5	0.0
		Mahalanobis	17	100.0	100.0	100.0	100.0	78.8	100.0	85.0	46.2	49.8	71.5	89.0	95.8	15.0	96.0	100.0	100.0	96.2	100.0
		Relative Mahalanobis	17	100.0	100.0	100.0	100.0	99.5	100.0	98.0	83.2	81.5	99.0	99.8	100.0	99.0	100.0	100.0	100.0	100.0	100.0
		ViM	16	100.0	38.5	32.2	100.0	44.8	100.0	81.0	19.8	17.0	36.2	60.0	87.2	1.2	74.2	89.0	32.2	37.5	100.0
		Energy+React	17	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		knn	0	5.8	2.5	3.0	0.0	0.0	0.0	1.2	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.8	5.8	0.0	0.0
		cosine	0	9.8	0.0	0.0	0.0	0.0	0.0	9.8	0.0	0.0	0.5	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0
		MCM/RCos	1	37.0	0.0	0.0	0.0	0.0	0.0	37.0	0.5	0.0	1.2	0.0	0.0	0.0	0.8	0.2	0.0	0.0	0.0

K EFFECT OF ID CONTAMINATION ON ALL MODELS

Table 11: Percentage of ID samples, $p = \frac{\text{ID}}{\text{ID}+\text{OOD}}$, in commonly used test OOD datasets found by visual inspection of 400 random samples per dataset. Unclear samples are ignored (which are at most 6.7% (for PLACES) of the 400 samples).

Dataset	ID samples	Dataset	ID samples
PLACES	59.5%	SPECIES	57.0%
IMAGENET-O	20.2%	TEXTURES	25.6%
iNAT. PLANTS	2.5%	TEXTURES43	20.0%
OPENIM.-O	4.9%		

In Table 12 we show the FPR values averaged across the cleaned subsampled datasets presented in Table 11, and in Tables 13-16 the results on the individual datasets. Since TEXTURES and iNATURALIST are fairly easy OOD datasets, the FPR values of most models in Table 12 are lower than on NINCO. In general, the results are however similar: Feature-dependent methods outperform methods not leveraging feature-information, yet still fail for some models, and pretraining only on IN-21k yields the best OOD-detectors. Again, Cosine and MCM/RCos improve fairly consistently over MSP, and are in some cases even the best-performing method.

Table 12: Mean FPR on subsampled datasets (averaged).

pre	acc.	model	MSP	MaxL	Ener	KL-M	Maha	RMaha	ViM	E+R	KNN	Cos	MCM/RCos
21k	86.0	ViT-B-384	36.1	22.9 -13	21.4 -15	33.8 -2	**17.8** -18	20.4 -16	18.2 -18	21.9 -14	43.9 $+8$	24.0 -12	24.4 -12
	84.5	ViT-B-224	38.7	26.1 -13	24.0 -15	37.4 -1	**18.0** -21	22.8 -16	19.7 -19	26.3 -12	47.0 $+8$	30.7 -8	30.4 -8
	86.3	Swinv2-B-256	39.4	31.5 -8	32.3 -7	41.9 $+3$	34.1 -5	31.1 -8	**25.5** -14	26.6 -13	32.4 -7	26.4 -13	26.5 -13
	86.7	Deit3-B-384	52.4	44.9 -7	48.6 -4	49.3 -3	32.9 -20	31.0 -21	32.4 -20	38.6 -14	31.9 -21	29.9 -23	**29.8** -23
	85.7	Deit3-B-224	52.3	47.3 -5	50.5 -2	52.4 $+0$	37.3 -15	35.1 -17	37.5 -15	42.4 -10	36.9 -15	34.3 -18	**33.7** -19
	86.3	CnvNxt-B	36.3	32.7 -4	37.2 $+1$	38.7 $+2$	25.7 -11	24.5 -12	**20.9** -15	30.7 -6	29.7 -7	24.2 -12	24.1 -12
	84.1	CnvNxt-T	40.6	35.2 -5	33.2 -7	45.9 $+5$	27.1 -14	30.7 -10	**20.5** -20	30.4 -10	41.0 $+0$	33.8 -7	32.4 -8
	82.3	BiT-m	56.7	47.5 -9	48.6 -8	51.9 -5	20.2 -36	26.0 -31	**16.9** -40	44.2 -12	24.3 -32	23.7 -33	30.8 -26
	85.6	EffNetv2-M	40.6	40.7 $+0$	49.1 $+9$	43.0 $+2$	31.4 -9	33.3 -7	24.1 -16	90.0 $+49$	38.2 -2	**24.0** -17	31.4 -9
none	81.1	ViT-B-384	59.8	52.6 -7	50.3 -10	54.5 -5	**42.8** -17	42.9 -17	54.9 -5	48.2 -12	58.2 -2	52.7 -7	55.1 -5
	84.6	Swinv2-B-256	60.1	57.8 -2	64.6 $+4$	58.2 -2	43.5 -17	**41.0** -19	44.6 -16	54.2 -6	50.8 -9	45.8 -14	43.9 -16
	85.1	Deit3-B-384	56.6	60.6 $+4$	80.0 $+23$	54.4 -2	46.1 -11	43.8 -13	**38.2** -18	89.6 $+33$	61.5 $+5$	51.5 -5	38.7 -18
	83.8	Deit3-B-224	56.4	56.6 $+0$	70.8 $+14$	54.9 -1	52.7 -4	48.5 -8	**40.9** -16	76.1 $+20$	71.5 $+15$	59.7 $+3$	44.2 -12
	82.6	XCiT-M-224	62.8	62.0 -1	67.9 $+5$	63.6 $+1$	53.0 -10	50.8 -12	**50.0** -13	62.5 -0	57.3 -6	51.4 -11	50.5 -12
	84.3	XCiT-M-224-d	61.2	56.6 -5	65.3 $+4$	58.9 -2	50.5 -11	48.2 -13	**44.0** -17	61.1 -0	54.1 -7	48.0 -13	46.3 -15
	84.4	CnvNxt-B	60.2	69.6 $+9$	92.7 $+33$	59.8 -0	51.0 -9	48.1 -12	48.5 -12	88.1 $+28$	55.5 -5	47.6 -13	**45.1** -15
	78.0	BiT-s	74.0	74.4 $+0$	75.5 $+1$	54.8 -19	57.5 -17	**43.1** -31	47.9 -26	56.5 -17	60.2 -14	44.6 -29	81.5 $+7$
	85.0	EffNetv2-M	55.5	56.1 $+1$	67.8 $+12$	53.8 -2	39.9 -16	37.0 -19	48.6 -7	52.2 -3	48.0 -7	43.5 -12	**36.1** -19
	84.9	EffNetb7	56.3	60.4 $+4$	73.8 $+18$	51.0 -5	51.8 -4	43.8 -12	57.8 $+2$	61.5 $+5$	53.0 -3	46.5 -10	**37.2** -19
	77.7	EffNet-B0	68.9	70.9 $+2$	78.7 $+10$	63.9 -5	68.1 -1	61.9 -7	56.8 -12	66.0 -3	68.6 -0	**40.5** -28	56.5 -12
	80.4	ResNet50	65.8	68.1 $+2$	76.6 $+11$	62.3 -3	77.4 $+12$	77.4 $+12$	69.7 $+4$	97.7 $+32$	56.5 -9	**44.9** -21	49.8 -16
JFT	86.8	EffNetb7-ns	47.5	44.1 -3	56.3 $+9$	45.0 -2	77.9 $+30$	47.7 $+0$	82.6 $+35$	51.2 $+4$	45.8 -2	**37.4** -10	37.7 -10
clip	87.2	ViT-B-384-l2b	34.5	33.0 -2	36.3 $+2$	35.3 $+1$	31.4 -3	27.0 -7	26.1 -8	30.6 -4	26.4 -8	22.9 -12	**22.6** -12
+12k	87.0	ViT-B-384-oai	36.4	33.0 -3	33.5 -3	36.6 $+0$	34.5 -2	29.7 -7	28.1 -8	31.1 -5	25.6 -11	23.0 -13	**23.0** -13
clip	86.6	ViT-B-384-l2b	51.0	49.1 -2	54.1 $+3$	45.0 -6	30.4 -21	31.6 -19	**28.6** -22	49.4 -2	33.6 -17	31.0 -20	30.7 -20
	86.2	ViT-B-384-oai	52.8	51.5 -1	55.5 $+3$	49.4 -3	**32.4** -20	34.8 -18	32.6 -20	52.0 -1	36.0 -17	33.0 -20	32.4 -20
clip	74.2	clip-ViT-L-336	—	—	—	—	—	—	—	—	—	53.4	**39.6**
z. shot	66.6	clip-ViT-B-224	—	—	—	—	—	—	—	—	—	67.7	**48.8**

Table 13: Comparing the cleaned and original datasets in terms of FPR. The best method per model and dataset is marked bold.

									fpr						
model	acc.	method	Pl-f	Pl-c	Spc-f	Spc-c	IN-f	IN-c	txt-f	txt-43	txt-c	OpO-f	OpO-c	iNat-f	iNat-c
		MSP	72.6	61.4	78.0	59.3	84.0	83.2	62.7	61.1	54.5	54.5	55.2	46.5	45.4
		MaxL	68.9	**52.3**	75.8	54.1	81.8	81.9	55.2	53.9	45.8	43.8	44.0	38.2	37.3
		Ener	69.1	54.2	73.8	51.2	79.8	78.3	50.5	49.6	41.3	42.8	42.7	35.0	33.9
		KL-M	72.6	60.8	72.8	52.3	**78.5**	**77.3**	58.5	57.1	51.0	51.5	51.6	35.0	33.7
		Maha	70.1	54.9	61.8	34.3	79.2	77.7	**47.0**	**45.8**	**39.9**	**36.8**	**35.9**	15.8	14.1
ViT-B-384	81.1	RMaha	69.1	54.2	**59.5**	**32.6**	80.0	79.0	48.8	47.5	41.7	38.8	38.0	**13.5**	**11.7**
		ViM	73.8	64.1	72.0	50.6	82.2	81.6	54.5	54.2	47.2	47.8	48.1	38.2	37.6
		E+R	**67.7**	53.6	71.2	46.5	81.8	80.6	50.5	49.6	42.0	40.0	39.7	28.0	26.9
		KNN	75.1	64.7	80.0	61.6	79.2	78.0	49.0	47.7	42.0	48.8	49.5	54.5	53.5
		Cos	70.4	55.6	75.8	55.2	80.5	80.3	50.0	49.1	41.3	43.5	43.8	41.2	39.9
		RCos	73.1	60.8	77.2	58.7	80.5	80.3	50.7	49.9	42.0	45.5	45.9	44.5	43.1
		MSP	69.1	52.9	78.0	64.0	83.8	84.1	66.0	63.8	58.0	55.2	54.3	48.0	47.3
		MaxL	69.4	53.6	76.5	62.2	79.2	79.3	56.2	55.8	48.6	58.0	56.8	46.8	46.2
		Ener	76.5	66.7	82.5	73.8	76.2	76.4	56.2	55.5	50.7	64.5	63.6	56.8	56.4
		KL-M	73.1	57.5	75.0	58.7	79.5	80.3	63.0	62.2	55.6	52.5	51.9	45.8	45.2
		Maha	60.0	42.5	73.2	50.0	77.5	77.0	51.7	50.1	41.0	31.2	31.2	21.0	19.3
Swinv2-B-256	84.6	RMaha	**59.3**	**40.5**	**69.5**	**44.2**	77.0	76.4	50.7	48.0	38.9	**30.8**	**30.4**	**17.5**	**15.7**
		ViM	63.2	45.1	76.5	54.7	**75.8**	**75.1**	48.5	46.9	38.2	32.0	32.6	23.2	21.9
		E+R	67.2	49.7	81.5	66.3	76.0	75.7	47.8	45.8	38.9	50.2	50.3	45.0	44.4
		KNN	65.2	48.4	79.5	60.5	79.2	78.3	49.5	48.0	39.2	37.0	38.3	41.5	40.2
		Cos	64.4	44.4	76.2	55.2	77.2	76.7	47.0	44.8	36.1	33.8	34.2	29.5	28.2
		RCos	63.5	43.1	75.5	53.5	76.8	75.7	**44.8**	**43.7**	**33.7**	32.2	32.6	26.0	24.5
		MSP	64.2	51.6	76.5	60.5	80.8	79.9	58.0	55.2	50.0	56.5	57.1	41.5	40.5
		MaxL	69.9	56.9	80.0	62.8	82.2	82.5	54.8	52.0	47.9	64.2	63.9	49.5	49.3
		Ener	87.7	80.4	91.5	86.0	86.0	87.1	66.5	66.0	61.5	83.5	82.1	83.5	83.0
		KL-M	70.6	54.9	77.2	58.1	74.5	73.1	56.5	53.6	49.3	51.5	52.4	39.5	38.6
		Maha	63.7	47.1	71.0	45.9	76.5	74.1	59.0	57.6	50.7	37.8	39.7	20.5	19.1
Deit3-B-384	85.1	RMaha	61.2	43.1	70.2	44.2	76.0	73.8	57.0	55.5	48.3	35.5	37.2	17.5	15.9
		ViM	**57.5**	**34.0**	70.5	44.2	71.2	69.3	42.0	39.9	31.9	33.8	34.8	**16.8**	**15.1**
		E+R	94.1	90.8	94.5	93.6	90.8	92.6	81.5	80.7	78.5	91.8	91.0	90.8	91.1
		KNN	66.4	51.0	84.8	69.8	82.5	81.6	64.2	62.7	55.2	50.5	53.3	58.5	58.2
		Cos	64.2	47.1	79.0	58.7	79.5	77.3	57.0	55.2	46.2	40.8	42.7	38.0	37.1
		RCos	70.9	52.3	**69.0**	**41.9**	**66.2**	**63.1**	**39.5**	**37.5**	**29.2**	**29.2**	**29.9**	17.0	15.7
		MSP	65.2	48.4	73.2	58.1	88.8	88.0	58.0	55.2	51.0	52.8	51.9	42.2	41.0
		MaxL	64.7	47.1	74.0	58.1	86.8	85.4	55.0	52.3	47.2	55.8	55.4	47.5	46.2
		Ener	75.6	63.4	82.5	74.4	85.8	85.8	63.7	62.5	58.3	74.5	74.7	68.8	68.4
		KL-M	72.6	52.3	72.8	55.2	83.2	81.6	58.2	56.0	51.7	50.5	50.5	39.5	38.1
		Maha	67.4	49.7	75.2	52.3	85.8	84.1	65.2	65.1	60.8	42.0	42.9	28.0	26.6
Deit3-B-224	83.8	RMaha	64.7	46.4	68.0	43.6	85.5	84.1	62.3	61.4	58.3	37.8	38.3	22.0	20.4
		ViM	**57.0**	**36.6**	**67.2**	**40.7**	80.8	79.3	**44.0**	**42.9**	**35.8**	**35.8**	**35.3**	**19.0**	**17.8**
		E+R	83.0	75.2	84.5	72.7	84.8	86.1	69.5	68.1	63.5	79.8	80.4	78.8	78.6
		KNN	78.5	66.0	89.2	79.1	87.8	86.7	63.2	61.4	57.3	63.5	65.5	74.2	74.2
		Cos	73.1	54.2	82.2	66.3	84.2	82.8	57.0	54.7	49.7	52.2	53.5	52.0	51.4
		RCos	68.4	49.0	70.0	45.9	**74.0**	**71.2**	45.0	43.7	37.2	38.0	38.6	24.8	23.5

Table 14: Comparing the cleaned and original datasets in terms of FPR. The best method per model and dataset is marked bold.

			fpr												
model	acc.	method	Pl-f	Pl-c	Spc-f	Spc-c	IN-f	IN-c	txt-f	txt-43	txt-c	OpO-f	OpO-c	iNat-f	iNat-c
		MSP	56.5	37.3	64.0	42.4	54.8	49.8	48.5	45.8	37.2	26.2	26.1	16.0	14.1
		MaxL	56.5	37.9	65.8	40.1	52.2	46.6	47.0	44.5	34.7	26.5	26.1	13.8	12.3
		Ener	60.5	41.2	69.0	43.6	50.2	47.2	51.2	49.3	40.6	30.8	29.6	16.8	15.4
		KL-M	57.8	37.3	66.0	43.0	55.5	51.5	45.0	42.4	35.1	27.3	27.2	20.0	17.8
		Maha	46.7	20.3	67.2	41.9	65.5	61.8	48.2	45.6	35.4	22.0	21.7	8.8	7.3
ViT-B-384-l2b-12k	87.2	RMaha	47.9	21.6	63.0	33.7	59.5	55.7	42.5	39.9	29.2	17.8	16.8	6.5	5.0
		ViM	**43.0**	**15.7**	63.7	36.6	61.8	57.6	41.8	39.1	27.4	15.8	15.5	5.2	3.7
		E+R	52.8	32.0	67.0	40.1	**46.0**	**42.4**	44.5	42.1	32.6	25.0	24.5	13.0	11.7
		KNN	45.9	24.8	68.2	40.7	54.8	49.8	36.0	33.0	20.8	17.0	16.3	7.5	6.0
		Cos	47.7	24.2	61.3	31.4	50.2	45.6	**35.5**	**32.4**	**19.4**	**14.0**	**13.6**	**4.8**	**3.1**
		RCos	46.9	22.2	**60.5**	**30.8**	50.2	45.6	36.0	33.0	19.8	**14.0**	13.9	**4.8**	**3.1**
		MSP	57.3	38.6	64.2	42.4	55.5	53.4	49.0	46.9	38.2	30.8	30.7	16.8	14.9
		MaxL	55.3	36.6	65.0	39.5	49.0	46.3	47.2	45.0	35.4	27.8	27.7	14.2	12.3
		Ener	57.0	39.2	67.8	40.7	45.8	42.7	49.8	47.7	38.2	26.0	26.6	15.0	13.3
		KL-M	63.2	43.8	65.8	40.1	55.2	53.4	46.2	44.8	36.1	31.0	31.5	16.8	14.9
		Maha	60.0	38.6	68.0	40.7	69.5	68.3	40.5	38.6	27.1	21.5	21.7	12.0	10.7
ViT-B-384-oai-12k	87.0	RMaha	58.3	35.9	63.2	33.7	61.0	58.6	40.5	38.3	26.7	16.8	17.4	7.5	6.0
		ViM	**52.6**	**26.1**	65.0	36.0	62.7	61.5	38.2	36.2	25.0	15.8	15.8	**6.0**	**4.4**
		E+R	55.6	37.3	66.5	39.0	**44.2**	**41.4**	47.0	45.0	34.7	22.8	23.1	13.0	11.2
		KNN	57.3	33.3	67.5	39.0	47.8	44.3	**30.2**	**27.9**	**15.3**	15.0	15.5	8.0	6.3
		Cos	55.3	30.1	**61.5**	**29.7**	46.8	42.4	33.0	30.6	18.4	12.2	12.5	7.0	5.2
		RCos	54.8	29.4	62.0	**29.7**	46.8	42.4	33.5	31.1	19.1	**12.0**	**12.2**	6.8	5.2
		MSP	65.4	51.0	70.2	52.3	78.8	77.3	54.5	51.7	44.4	43.5	44.3	38.0	36.6
		MaxL	62.7	47.1	73.5	54.7	73.8	72.2	54.8	52.0	44.1	39.0	39.9	37.5	36.6
		Ener	63.0	45.1	78.8	61.6	72.8	70.9	61.0	58.2	50.3	42.8	44.3	52.5	52.2
		KL-M	64.7	47.7	69.0	47.1	69.2	68.9	48.0	45.3	36.1	38.0	38.6	32.8	31.3
		Maha	55.6	30.1	64.5	**36.6**	62.5	60.8	40.2	37.5	27.1	18.5	19.3	**10.5**	**8.6**
ViT-B-384-l2b	86.6	RMaha	56.5	30.7	64.5	39.5	64.0	61.8	40.2	37.5	26.4	19.8	20.4	12.5	10.7
		ViM	**50.1**	**22.9**	66.8	38.4	**60.0**	**57.9**	39.8	36.7	25.0	**17.8**	**18.5**	10.8	9.1
		E+R	60.0	39.9	77.5	58.7	71.2	69.3	57.5	54.7	45.8	38.2	39.1	44.5	43.9
		KNN	55.8	32.7	69.2	43.6	63.5	63.4	37.8	35.1	24.7	22.5	22.8	16.0	14.4
		Cos	54.3	30.7	66.2	39.5	63.2	61.2	36.8	33.8	22.2	20.5	21.2	12.8	11.0
		RCos	55.1	32.0	**63.2**	37.2	63.5	61.2	**35.8**	**32.7**	**21.2**	20.2	21.2	13.2	11.5
		MSP	64.9	50.3	71.2	54.1	78.0	76.7	54.5	52.5	44.1	51.0	50.8	42.2	41.0
		MaxL	64.2	51.0	73.5	54.1	73.8	71.5	56.5	54.7	46.5	47.5	47.6	39.2	38.4
		Ener	66.4	53.6	75.2	56.4	73.5	72.5	63.2	61.7	55.6	50.7	51.4	43.8	43.3
		KL-M	65.2	49.0	70.2	50.0	71.0	69.6	52.2	50.7	42.4	48.2	48.6	38.2	36.6
		Maha	53.3	29.4	64.5	37.2	65.0	61.8	41.5	39.1	29.2	**24.2**	25.3	**13.2**	**11.5**
ViT-B-384-oai	86.2	RMaha	54.8	31.4	**63.0**	37.8	68.5	66.7	43.5	41.3	31.9	26.0	26.6	16.0	14.4
		ViM	**50.9**	**27.5**	66.5	**36.6**	67.0	65.0	42.5	39.9	30.6	**24.2**	**23.9**	**13.2**	11.7
		E+R	64.2	49.0	75.8	55.8	71.5	70.2	61.3	59.5	52.8	46.0	46.5	38.0	37.9
		KNN	53.8	31.4	72.5	45.9	64.2	62.1	40.8	38.3	28.1	28.7	29.3	20.5	19.1
		Cos	55.1	30.7	66.8	39.5	64.5	61.5	39.8	37.3	26.0	25.2	25.8	16.2	14.6
		RCos	54.6	30.1	64.2	38.4	**63.7**	**60.8**	**38.5**	**35.9**	**25.0**	**24.2**	25.0	16.8	15.1
		MSP	60.5	37.9	64.8	41.3	62.7	58.3	54.0	52.5	43.4	28.2	27.2	10.5	8.9
		MaxL	50.6	27.5	65.0	33.7	45.8	40.5	36.2	33.5	21.2	12.2	12.2	3.5	2.1
		Ener	**49.1**	**25.5**	64.2	31.4	43.0	38.5	34.8	32.4	20.5	11.2	10.9	3.2	1.8
		KL-M	64.4	43.1	66.8	36.0	57.5	53.4	50.0	48.8	37.5	26.8	25.8	8.5	6.8
		Maha	57.5	34.6	**48.0**	**16.9**	**34.8**	**28.8**	29.0	25.7	16.3	**9.5**	**9.5**	2.0	0.8
ViT-B-384-21k	86.0	RMaha	55.8	30.1	51.2	19.8	42.5	36.2	37.5	35.1	23.3	12.2	12.2	2.0	0.8
		ViM	50.1	26.8	55.0	23.3	37.8	34.6	**24.0**	**21.2**	**12.5**	12.5	11.7	**1.5**	**0.5**
		E+R	53.3	30.7	60.5	30.8	44.0	38.5	34.0	31.6	19.4	10.5	10.1	2.8	1.6
		KNN	69.4	50.3	82.5	62.2	51.5	47.6	37.5	36.2	24.7	36.5	37.5	41.0	41.3
		Cos	57.0	32.7	67.2	34.3	42.2	35.9	36.5	33.2	21.2	15.5	15.2	6.2	5.0
		RCos	56.8	33.3	68.2	36.6	40.2	34.6	34.0	31.4	18.8	17.0	16.8	7.2	6.0
		MSP	60.0	39.9	66.2	40.7	67.8	62.8	51.2	48.5	38.5	35.8	36.7	15.0	13.8
		MaxL	51.1	28.8	65.5	35.5	50.7	45.3	38.0	35.7	24.0	17.8	17.9	6.8	5.2
		Ener	48.6	**26.1**	65.2	34.9	46.8	41.7	35.8	33.0	22.2	14.2	14.1	6.5	5.0
		KL-M	64.7	43.8	69.0	38.4	62.3	58.3	50.5	48.5	37.8	33.8	34.8	12.8	11.2
		Maha	59.0	33.3	**49.5**	**16.3**	**39.8**	**34.6**	27.0	24.7	13.2	**10.5**	**10.3**	**2.0**	**0.5**
ViT-B-224-21k	84.5	RMaha	59.8	37.3	51.0	17.4	47.5	41.7	39.0	37.0	24.0	16.0	15.8	2.2	**0.5**
		ViM	**48.1**	26.8	58.0	26.2	41.0	36.9	**23.0**	**20.6**	**12.8**	15.5	14.1	2.8	1.6
		E+R	53.1	27.5	64.8	36.6	50.7	45.3	38.8	35.7	25.3	17.0	17.1	7.5	6.0
		KNN	70.4	53.6	83.5	65.7	56.5	52.1	37.2	35.9	26.0	44.5	45.7	39.0	38.6
		Cos	60.7	39.9	70.5	41.9	51.5	46.9	35.8	33.0	22.6	22.8	23.1	11.2	9.9
		RCos	62.0	40.5	71.2	42.4	45.2	41.1	36.2	33.8	22.6	23.5	24.2	13.2	11.7
		MSP	58.5	37.9	65.0	44.2	62.0	59.9	53.5	52.0	44.1	33.2	33.2	19.2	17.0
		MaxL	53.3	31.4	61.3	36.6	48.5	45.6	50.7	47.7	41.3	23.2	22.8	13.0	11.0
		Ener	54.8	31.4	62.0	35.5	43.0	40.1	54.8	50.9	45.5	31.0	30.4	12.8	11.0
		KL-M	64.2	49.0	70.0	44.8	66.0	64.1	52.5	52.8	42.0	32.5	33.7	20.0	17.8
		Maha	60.0	38.6	67.8	38.4	71.0	68.0	47.0	45.8	36.5	20.2	20.4	3.5	2.6
Swinv2-B-256-21k	86.3	RMaha	58.0	35.3	62.7	32.0	67.8	64.7	44.5	42.4	34.0	17.2	17.7	4.2	3.1
		ViM	51.1	27.5	65.0	32.6	59.0	53.4	37.8	36.5	25.7	12.8	12.5	**2.0**	**1.3**
		E+R	**48.9**	**24.8**	**59.2**	**30.8**	**42.0**	**38.8**	49.0	47.2	37.8	21.0	20.4	8.5	6.8
		KNN	61.0	38.6	73.5	44.8	63.2	58.6	38.8	36.7	27.4	18.5	18.5	7.8	6.5
		Cos	54.3	30.1	67.0	37.8	57.5	53.4	**34.5**	**32.4**	**22.2**	**12.2**	**12.2**	4.0	2.6
		RCos	57.0	34.0	65.8	36.6	55.2	50.5	36.0	**32.4**	22.9	12.8	**12.2**	4.2	2.9
		MSP	67.7	53.6	72.8	54.7	73.5	72.2	64.8	62.7	55.9	50.7	50.5	29.8	27.7
		MaxL	61.5	41.8	72.8	46.5	59.5	57.0	63.5	61.1	53.1	44.2	44.3	28.7	26.9
		Ener	63.5	44.4	76.5	53.5	53.2	51.1	67.0	65.1	56.6	48.0	47.8	39.2	38.1
		KL-M	67.9	51.0	72.0	52.3	69.5	68.9	59.0	57.6	49.7	47.8	47.3	28.5	26.4
		Maha	53.3	30.7	66.0	40.1	66.2	63.8	48.2	46.1	37.5	22.0	21.5	6.0	3.9
Deit3-B-384-21k	86.7	RMaha	51.6	28.1	**64.2**	**39.0**	62.0	59.5	46.2	44.2	34.0	21.2	20.9	6.8	4.7
		ViM	**50.1**	28.8	68.8	43.6	63.7	61.5	50.2	48.0	38.9	**20.8**	**20.1**	**3.8**	**1.8**
		E+R	56.3	32.7	73.0	45.9	**51.7**	**49.5**	58.2	55.8	45.5	36.0	35.3	24.0	22.7
		KNN	54.1	28.8	72.0	45.9	59.2	56.0	44.8	42.6	32.6	21.2	20.7	9.2	7.3
		Cos	52.6	**26.8**	68.0	41.3	58.5	55.7	42.8	**40.2**	29.9	21.5	20.9	6.5	4.7
		RCos	53.1	28.8	67.8	41.3	57.0	54.7	**42.2**	**40.2**	**29.5**	**20.8**	**20.1**	6.5	4.7
		MSP	66.2	48.4	73.2	53.5	73.5	72.5	66.8	64.9	59.4	46.8	45.4	36.2	35.0
		MaxL	62.2	45.1	72.0	48.3	61.0	58.6	64.5	62.5	55.6	41.8	40.5	36.8	35.8
		Ener	64.2	48.4	74.2	51.2	55.0	52.8	63.5	61.7	54.2	44.5	44.3	53.0	52.5
		KL-M	68.4	52.3	74.5	55.8	71.8	70.9	63.0	61.4	55.6	48.5	46.7	34.2	32.9
		Maha	53.8	33.3	71.0	45.9	70.2	68.9	52.5	51.2	42.7	26.2	26.4	8.5	6.3
Deit3-B-224-21k	85.7	RMaha	**52.3**	**31.4**	**67.8**	**41.9**	66.8	65.7	52.0	50.1	42.7	**23.0**	**22.8**	8.2	6.3
		ViM	54.8	34.0	74.8	50.0	68.2	66.7	53.2	51.5	43.1	26.2	26.6	**7.0**	**4.7**
		E+R	57.0	35.3	73.0	47.7	**52.2**	**50.5**	59.0	56.3	48.6	34.5	34.2	39.0	37.9
		KNN	56.3	35.9	75.8	51.2	62.7	60.2	47.8	45.3	36.8	24.2	23.9	15.2	13.3
		Cos	53.3	32.0	71.8	45.9	60.5	59.2	47.8	45.3	37.2	24.5	24.2	9.2	7.3
		RCos	53.6	32.7	71.0	44.8	60.2	58.6	**47.2**	**44.8**	**36.1**	23.5	23.1	9.0	7.0

Table 15: Comparing the cleaned and original datasets in terms of FPR. The best method per model and dataset is marked bold.

			fpr												
model	acc.	method	Pl-f	Pl-c	Spc-f	Spc-c	IN-f	IN-c	txt-f	txt-43	txt-c	OpO-f	OpO-c	iNat-f	iNat-c
		MSP	70.4	58.8	79.8	69.8	86.2	84.8	62.3	60.6	54.5	60.0	60.1	50.0	49.1
		MaxL	70.1	60.8	81.5	70.3	83.5	82.2	58.8	57.1	50.7	60.0	60.6	48.2	47.3
		Ener	77.5	69.9	82.8	72.7	85.2	84.8	62.5	61.7	56.2	65.2	65.5	59.2	58.0
		KL-M	72.8	60.8	79.8	69.8	85.2	83.8	62.3	60.9	56.6	59.2	59.2	52.2	51.4
		Maha	64.4	48.4	74.0	52.9	86.2	84.1	65.2	64.1	57.3	46.5	47.3	29.0	27.9
XCiT-M-224	82.6	RMaha	64.4	47.1	**72.0**	**50.6**	85.5	83.2	64.0	62.7	55.6	43.2	44.0	**25.8**	**24.5**
		ViM	65.2	48.4	73.8	53.5	83.8	**80.9**	58.0	56.6	49.0	**42.5**	**42.4**	27.0	25.8
		E+R	70.9	59.5	82.5	69.2	**83.0**	82.2	61.8	60.6	54.2	60.2	60.9	50.0	49.3
		KNN	65.4	50.3	82.8	65.7	84.2	82.8	55.0	53.6	45.8	50.5	51.6	48.0	47.5
		Cos	**62.0**	**44.4**	78.2	58.7	83.8	81.9	**53.8**	**52.3**	**44.4**	44.2	45.1	34.5	33.7
		RCos	63.0	45.1	76.0	55.8	83.8	81.6	54.8	53.4	45.1	43.2	44.0	32.5	31.6
		MSP	73.3	59.5	75.0	63.4	88.0	88.0	59.5	57.9	52.1	56.5	56.0	48.8	48.0
		MaxL	71.1	56.9	72.5	59.3	86.0	85.4	51.5	49.9	43.8	55.0	54.6	40.5	39.4
		Ener	77.0	63.4	80.8	73.8	87.2	86.7	58.5	56.8	50.7	67.5	67.1	50.7	49.9
		KL-M	75.3	62.7	77.2	61.6	**83.8**	83.8	55.5	53.6	45.8	55.5	55.2	44.8	44.1
		Maha	72.1	52.9	72.5	50.6	88.2	87.1	56.2	54.7	46.5	41.8	41.6	25.5	24.0
XCiT-M-224-d	84.3	RMaha	71.9	51.6	**69.2**	45.3	88.0	86.7	55.2	53.6	45.8	39.0	38.6	22.5	20.9
		ViM	69.4	49.0	70.2	**44.2**	84.2	**82.8**	48.2	46.6	37.5	**34.0**	**33.4**	**19.0**	**17.2**
		E+R	74.1	59.5	82.2	70.3	85.8	85.1	54.5	53.1	45.1	58.8	59.0	48.8	47.5
		KNN	73.6	56.2	82.0	65.1	86.0	85.1	**47.0**	**44.8**	**35.1**	41.5	42.4	41.8	40.7
		Cos	68.9	48.4	75.0	54.1	85.0	83.5	48.5	46.4	36.5	38.5	38.6	28.7	27.2
		RCos	**67.7**	**45.8**	73.0	51.2	85.0	84.1	47.8	45.8	36.5	37.0	37.0	25.0	23.5
		MSP	70.4	56.9	**66.8**	50.6	91.5	91.3	69.5	68.1	61.8	63.0	63.6	38.2	36.8
		MaxL	79.5	68.6	71.8	59.3	94.0	93.9	75.5	76.1	69.4	78.5	78.8	48.2	47.5
		Ener	95.1	95.4	85.8	90.7	96.2	96.1	93.0	93.8	92.0	96.0	96.2	85.8	85.6
		KL-M	74.3	57.5	79.0	62.8	85.2	85.8	61.0	59.5	52.1	54.0	54.9	47.8	46.0
		Maha	65.9	43.1	74.8	55.2	87.8	87.1	54.8	52.3	44.4	37.5	38.9	38.8	37.1
CnvNxt-B	84.4	RMaha	**64.7**	43.1	70.8	50.6	87.8	87.1	54.2	51.7	43.8	35.0	36.4	30.0	27.9
		ViM	66.9	44.4	73.8	51.7	84.8	83.8	51.7	49.1	40.6	38.2	38.9	32.8	31.3
		E+R	92.6	90.8	84.2	86.0	95.8	95.5	89.8	89.8	87.8	92.8	92.9	76.0	75.7
		KNN	75.1	56.9	82.2	65.7	84.5	**83.2**	**48.8**	**46.4**	**36.1**	44.0	45.1	46.5	46.0
		Cos	67.7	43.8	72.8	53.5	85.5	84.8	49.2	47.5	37.2	34.8	**35.3**	32.0	30.8
		RCos	66.4	**41.2**	70.0	**48.8**	**84.2**	**83.2**	49.0	47.7	36.5	**34.2**	**35.3**	**27.0**	**25.6**
		MSP	79.3	66.7	84.2	74.4	97.0	96.8	76.0	74.8	71.9	72.8	73.9	61.3	60.3
		MaxL	73.8	57.5	85.0	74.4	98.0	98.1	72.2	70.8	68.4	71.0	72.3	75.2	75.5
		Ener	74.3	59.5	85.2	75.6	97.8	97.7	73.0	71.3	68.1	70.5	71.5	79.2	80.4
		KL-M	76.8	62.7	70.0	52.9	87.8	87.7	52.2	51.7	46.2	50.7	50.5	30.5	28.7
		Maha	80.2	71.9	83.2	70.3	70.8	67.6	17.0	16.1	13.9	62.7	60.6	61.3	60.6
BiT-s	78.0	RMaha	79.5	67.3	**56.0**	**34.3**	77.8	77.7	24.2	22.8	15.3	**43.2**	**41.3**	**24.8**	**22.7**
		ViM	69.4	57.5	80.8	66.3	**64.8**	**62.8**	**6.2**	**5.9**	**3.5**	45.5	43.2	54.2	54.3
		E+R	**56.3**	**39.2**	76.8	61.6	90.8	90.6	51.0	49.3	45.1	54.0	53.5	48.8	49.1
		KNN	84.0	74.5	85.2	73.3	65.5	63.1	13.0	13.4	9.4	62.7	60.6	79.5	80.2
		Cos	70.9	56.9	74.5	57.6	67.2	64.4	11.8	11.0	7.6	45.0	42.1	39.2	38.9
		RCos	90.1	85.6	87.8	83.7	93.5	93.5	66.2	64.6	60.8	75.2	76.9	87.8	88.3
		MSP	65.9	48.4	72.0	52.9	87.5	87.4	54.5	52.0	46.9	53.8	54.9	43.5	42.6
		MaxL	68.9	50.3	73.0	54.1	86.5	87.1	53.8	50.7	45.5	56.5	56.5	44.0	43.1
		Ener	76.5	64.1	80.8	69.8	86.5	87.7	61.5	59.5	54.9	67.2	67.7	63.0	62.9
		KL-M	70.4	55.6	70.8	50.6	81.5	81.6	54.2	52.5	46.5	50.2	50.3	39.2	38.4
		Maha	66.2	47.1	68.2	41.9	75.5	73.8	40.2	36.7	29.5	30.2	30.4	18.0	16.7
EffNetv2-M	85.0	RMaha	64.0	43.1	**62.7**	**34.9**	75.5	74.1	38.8	35.9	27.8	**28.0**	**28.5**	**14.8**	**13.3**
		ViM	68.1	51.6	74.5	52.3	80.2	79.0	47.8	44.5	37.5	38.2	38.9	33.0	32.4
		E+R	70.1	50.3	78.5	59.3	79.5	78.6	45.5	43.4	35.1	47.2	47.8	42.5	42.0
		KNN	69.9	51.0	74.8	51.7	80.5	79.3	43.5	41.3	32.6	38.2	39.7	35.0	33.9
		Cos	65.4	42.5	71.2	47.7	80.2	79.3	43.0	40.5	31.6	33.5	34.5	26.8	25.6
		RCos	**58.5**	**35.3**	65.8	39.5	**74.8**	**72.8**	**34.5**	**31.9**	**22.9**	28.7	28.8	18.5	17.0
		MSP	66.2	58.8	68.5	50.6	86.5	86.4	55.5	52.8	46.5	51.5	51.9	44.5	43.3
		MaxL	70.1	61.4	76.8	60.5	85.8	85.8	57.5	55.5	47.6	57.0	57.6	50.7	49.6
		Ener	83.7	77.1	87.0	76.2	87.2	88.0	65.8	63.5	56.6	74.0	74.2	71.2	70.8
		KL-M	68.9	56.2	70.2	48.8	79.0	78.3	50.0	47.5	42.4	43.5	44.0	38.2	36.6
		Maha	74.3	61.4	73.5	52.3	73.5	73.5	59.5	58.7	52.8	42.0	42.7	29.2	28.2
EffNetb7	84.9	RMaha	71.6	57.5	**63.2**	**37.2**	74.5	73.1	51.2	49.3	42.7	32.0	32.9	20.5	19.3
		ViM	74.1	62.7	79.8	64.5	73.5	72.5	61.3	60.3	55.2	50.2	50.8	41.5	41.0
		E+R	76.5	65.4	82.0	65.7	80.2	80.6	54.2	52.5	42.4	57.5	58.7	57.0	56.4
		KNN	75.1	59.5	77.8	58.1	77.0	76.7	47.5	44.8	35.4	43.8	45.7	44.0	42.8
		Cos	70.1	50.3	71.8	50.0	76.5	74.4	50.7	48.5	38.9	34.2	35.9	31.0	29.8
		RCos	**62.7**	**39.2**	65.5	39.5	**72.0**	**70.2**	**39.0**	**37.0**	**27.8**	**28.7**	**29.9**	**17.8**	**16.4**
		MSP	76.3	68.0	79.5	69.8	95.0	94.2	67.0	65.7	62.5	65.2	65.5	54.0	53.3
		MaxL	81.5	71.9	82.2	72.7	93.8	93.2	66.5	65.1	60.4	66.5	66.3	61.5	61.1
		Ener	87.2	79.1	86.5	79.1	91.5	91.9	72.5	72.1	69.4	77.0	76.9	76.0	76.0
		KL-M	80.5	68.0	76.5	64.0	90.5	90.6	61.3	60.6	57.3	60.0	59.5	45.2	44.1
		Maha	90.1	85.0	83.5	73.8	78.0	76.4	44.5	44.5	38.2	71.0	71.2	65.5	64.2
EffNet-B0	77.7	RMaha	83.2	69.3	71.5	57.6	88.5	87.4	54.8	53.9	50.0	62.0	62.0	46.5	45.2
		ViM	84.2	74.5	81.0	67.4	74.2	71.2	25.2	24.1	17.0	58.0	56.5	55.5	54.3
		E+R	81.7	70.6	83.2	72.1	83.8	82.5	49.5	49.1	43.8	64.0	63.9	63.7	63.2
		KNN	94.1	90.8	87.8	83.1	**70.5**	**67.3**	30.2	29.0	22.2	74.5	73.1	75.5	75.2
		Cos	**71.1**	**53.6**	**59.2**	**43.0**	73.0	71.8	**22.8**	**21.4**	**13.5**	**40.8**	**37.8**	**24.8**	**23.0**
		RCos	76.8	61.4	65.2	49.4	89.8	89.6	52.8	51.5	45.1	56.0	55.7	39.0	37.9
		MSP	**73.8**	**58.8**	78.5	68.0	96.0	95.5	64.5	63.0	59.0	62.5	62.8	51.7	50.7
		MaxL	76.0	64.1	79.0	68.0	96.8	96.4	67.2	65.7	61.8	64.8	64.9	54.5	53.5
		Ener	84.4	79.1	81.5	72.1	95.8	95.1	75.8	75.1	73.3	73.8	73.6	67.5	66.6
		KL-M	76.0	64.1	75.8	64.5	93.0	92.2	59.0	57.4	52.1	57.0	57.3	44.8	43.6
		Maha	94.1	90.2	88.8	82.6	79.8	78.3	53.5	53.4	49.7	77.5	78.5	85.8	85.4
ResNet50	80.4	RMaha	93.1	89.5	70.5	62.2	79.8	79.6	75.8	75.6	73.3	75.0	76.4	83.5	83.3
		ViM	87.4	80.4	84.5	75.6	89.0	88.3	40.8	39.9	33.7	65.0	66.8	73.5	73.1
		E+R	99.8	100.0	99.0	99.4	91.5	91.6	98.2	98.1	98.3	96.5	97.0	99.8	99.7
		KNN	84.4	74.5	82.0	68.0	**68.0**	**64.4**	**24.2**	**23.3**	**16.3**	57.5	56.5	59.2	59.3
		Cos	76.3	60.8	66.8	**45.9**	78.0	76.1	32.0	29.8	20.8	**42.5**	**42.4**	**25.0**	**23.5**
		RCos	79.8	64.7	**65.2**	47.1	82.2	79.6	39.0	36.7	27.8	48.0	49.2	31.8	30.3

Table 16: Comparing the cleaned and original datasets in terms of FPR. The best method per model and dataset is marked bold.

			fpr												
model	acc.	method	Pl-f	Pl-c	Spc-f	Spc-c	IN-f	IN-c	txt-f	txt-43	txt-c	OpO-f	OpO-c	iNat-f	iNat-c
		MSP	57.5	35.3	63.2	40.1	61.0	57.9	51.2	47.5	38.5	30.5	29.3	18.5	16.4
		MaxL	57.3	35.9	60.8	36.6	53.5	49.5	47.2	43.4	31.9	26.0	24.7	18.8	17.2
		Ener	62.7	41.8	63.2	39.0	49.5	45.0	55.0	52.5	43.4	31.0	29.6	25.2	24.3
		KL-M	65.7	43.1	69.5	44.2	59.8	57.6	49.0	46.6	36.5	34.5	33.7	19.5	17.0
		Maha	59.0	32.7	58.0	29.7	59.2	56.3	32.8	30.0	19.1	15.0	13.9	3.5	2.3
CnvNxt-B-21k	86.3	RMaha	60.7	34.0	**56.5**	**26.7**	55.8	51.1	33.8	30.3	18.8	15.2	14.7	3.2	1.8
		ViM	**53.3**	**29.4**	**56.5**	30.8	**44.5**	**40.8**	**29.0**	**24.7**	**14.9**	**10.2**	**9.0**	**1.8**	**0.8**
		E+R	56.0	32.7	60.2	33.1	48.8	43.7	48.2	46.4	34.4	24.2	22.8	18.5	17.2
		KNN	68.9	50.3	70.0	42.4	48.8	44.0	31.8	27.9	17.0	17.2	17.1	9.2	7.3
		Cos	59.8	33.3	63.5	33.1	50.2	45.3	32.0	29.0	16.3	14.2	14.1	4.5	3.1
		RCos	58.8	32.7	63.5	32.6	50.0	45.6	32.5	29.2	16.7	14.5	14.4	4.5	2.9
		MSP	65.4	43.1	68.8	45.9	67.8	65.4	52.2	53.9	42.0	33.8	33.7	15.8	13.6
		MaxL	61.5	38.6	67.8	41.3	58.8	56.0	47.5	47.2	36.5	29.5	29.3	12.0	9.7
		Ener	61.5	37.9	67.8	39.5	51.5	47.6	47.5	46.6	36.8	29.2	28.5	11.2	8.9
		KL-M	71.9	53.6	72.8	49.4	71.5	71.8	51.2	52.0	41.0	41.0	41.0	20.5	18.3
		Maha	56.8	34.0	57.0	29.7	64.5	62.1	31.2	29.5	17.7	17.2	16.8	3.8	2.3
CnvNxt-T-21k	84.1	RMaha	60.5	39.9	59.0	32.0	63.7	61.5	38.0	36.2	25.3	21.5	21.7	5.5	3.7
		ViM	**53.3**	**29.4**	**51.5**	**23.8**	**46.5**	**42.4**	**28.0**	**27.3**	**15.6**	**11.2**	**10.6**	**2.5**	**1.0**
		E+R	58.3	33.3	67.5	38.4	51.2	47.2	45.0	45.3	33.7	24.0	23.6	8.2	6.0
		KNN	76.3	58.2	73.2	47.1	60.2	56.0	36.8	34.6	22.6	34.5	35.1	28.7	27.2
		Cos	65.7	42.5	67.2	40.1	62.3	58.6	38.2	36.2	25.3	24.8	25.5	12.5	10.4
		RCos	63.5	39.9	65.8	37.2	62.3	58.6	38.8	36.2	26.0	22.8	23.4	11.0	9.4
		MSP	75.3	55.6	72.5	54.7	83.2	82.2	72.8	72.4	68.1	52.8	52.4	29.8	27.2
		MaxL	69.9	47.7	65.8	41.3	80.2	79.3	66.5	67.8	59.7	43.0	42.9	16.8	13.8
		Ener	70.6	50.3	66.5	42.4	79.8	79.0	69.0	68.9	62.8	43.8	43.8	15.5	13.1
		KL-M	73.8	54.9	74.8	54.1	74.0	73.5	64.0	65.1	58.7	46.5	45.9	27.3	24.5
		Maha	64.2	40.5	52.5	25.0	**36.8**	**33.3**	6.5	5.9	**2.1**	20.0	17.1	4.8	3.1
BiT-m	82.3	RMaha	63.7	43.1	49.5	**20.3**	56.2	54.4	23.8	22.8	13.9	23.0	22.0	3.8	2.3
		ViM	**55.1**	**28.1**	47.0	22.7	38.2	34.6	**5.5**	**5.1**	**2.1**	**14.8**	**12.8**	**2.2**	**1.0**
		E+R	64.9	44.4	**46.5**	29.1	75.8	74.8	67.8	67.6	64.6	44.5	44.0	10.2	8.4
		KNN	70.4	49.0	58.8	32.6	42.2	39.5	11.2	10.7	4.9	20.2	16.8	5.0	3.1
		Cos	64.7	37.9	58.8	30.2	50.5	46.9	16.5	14.7	6.6	19.2	17.9	4.5	2.9
		RCos	66.7	39.2	62.5	36.0	64.5	61.8	31.2	29.0	18.4	25.8	24.7	6.2	4.7
		MSP	61.0	39.2	67.0	45.3	65.2	62.5	52.2	51.2	42.4	36.5	36.7	19.2	17.2
		MaxL	62.2	44.4	68.8	45.9	61.5	58.6	52.8	51.2	41.7	34.8	34.5	20.8	19.1
		Ener	69.4	54.2	75.0	54.7	64.5	63.1	65.5	63.3	56.9	40.5	39.9	27.3	25.6
		KL-M	65.2	45.8	70.5	47.7	64.8	63.4	51.7	49.6	42.4	36.8	37.5	22.8	21.1
		Maha	69.4	47.1	68.2	39.0	58.0	53.1	27.3	25.2	15.6	28.2	27.4	7.8	6.0
EffNetv2-M-21k	85.6	RMaha	69.9	52.9	61.3	33.7	59.8	56.0	34.5	34.3	24.0	27.0	26.6	8.5	6.5
		ViM	62.0	**36.6**	65.2	33.7	47.2	42.1	**21.8**	**18.8**	**10.8**	**20.5**	**19.0**	**3.8**	**2.3**
		E+R	94.6	93.5	93.5	91.3	89.2	89.3	87.2	88.7	89.2	87.5	87.0	89.5	89.6
		KNN	83.0	71.2	75.2	52.3	45.5	41.7	26.0	24.7	14.6	40.5	38.0	13.2	11.5
		Cos	**60.7**	37.3	**59.8**	**29.7**	**41.8**	**37.5**	23.8	22.3	12.8	23.0	22.6	5.2	3.9
		RCos	63.5	39.9	62.7	37.8	57.2	53.7	37.8	35.4	25.7	25.5	25.5	6.8	5.5
		MSP	56.8	37.3	69.2	52.9	68.2	66.0	55.5	55.5	47.9	44.2	43.8	38.8	37.1
		MaxL	**50.4**	34.6	**66.5**	48.3	65.0	64.1	51.5	51.7	43.4	39.8	39.4	36.8	34.7
		Ener	54.3	42.5	74.8	61.6	72.8	72.5	60.0	61.1	54.2	52.5	52.7	55.8	54.6
		KL-M	63.7	39.2	73.5	54.1	62.3	59.9	53.2	52.5	45.5	43.2	42.4	31.0	29.0
		Maha	94.6	88.9	89.0	80.2	76.8	75.7	82.2	79.1	78.5	72.5	72.6	71.8	71.3
EffNetb7-ns	86.8	RMaha	78.3	59.5	69.2	**45.9**	70.2	68.9	63.2	60.9	55.2	41.0	39.9	**18.8**	**16.4**
		ViM	95.8	91.5	92.2	85.5	77.8	77.3	87.2	82.8	85.1	77.2	77.2	79.2	78.9
		E+R	58.8	39.9	74.0	58.1	65.5	66.0	58.2	57.1	50.7	44.0	44.6	49.2	47.8
		KNN	62.5	39.9	83.0	66.3	65.2	63.8	42.2	40.5	28.5	42.5	43.2	35.0	33.4
		Cos	56.3	**31.4**	74.8	54.7	**61.0**	**58.6**	**38.8**	**36.2**	**24.7**	**33.8**	**33.7**	23.5	21.7
		RCos	57.8	32.0	73.8	54.7	61.5	59.9	40.0	38.1	26.4	34.0	**33.7**	21.8	19.8

L RESULTS ON NINCO CLASSES WITH AND WITHOUT OVERLAP WITH IN-21K

Since the classes of NINCO can be distinguished by whether they belong to an IN-21k class or not, we present results on both of these groups here. We note that they should be taken with care, since the groups differ both in size (7 vs. 39 classes) and difficulty of the individual classes. Most models and methods perform better on the classes *without* IN-21k overlap, and ViT+Maha is the best OOD-detector in both cases. While RMaha yields the most consistent improvements over MSP in both cases, ViM performs comparably better on the classes without overlap. Pretraining *only* on IN-21k yields the best OOD-detectors in both cases.

Table 17: Mean FPR for classes without 21k overlap.

pre	acc.	model	MSP	MaxL	Ener	KL-M	Maha	RMaha	ViM	E+R	KNN	Cos	MCM/RCos
21k	86.0	ViT-B-384	63.4	45.3 -18	41.7 -22	56.4 -7	**28.5** -35	34.6 -29	29.8 -34	39.7 -24	74.9 $+12$	47.0 -16	45.6 -18
	84.5	ViT-B-224	66.1	53.2 -13	50.0 -16	61.3 -5	**26.7** -39	37.2 -29	32.3 -34	49.8 -16	76.8 $+11$	56.9 -9	58.5 -8
	86.3	Swinv2-B-256	69.4	59.7 -10	57.8 -12	66.1 -3	38.7 -31	44.1 -25	**33.0** -36	48.9 -21	58.3 -11	49.3 -20	48.1 -21
	86.7	Deit3-B-384	74.7	75.0 $+0$	79.7 $+5$	71.7 -3	**44.1** -31	45.0 -30	45.4 -29	70.3 -4	51.4 -23	50.2 -25	49.7 -25
	85.7	Deit3-B-224	75.9	75.8 -0	79.4 $+4$	73.5 -2	47.0 -29	49.1 -27	**45.4** -31	73.9 -2	57.9 -18	54.1 -22	54.7 -21
	86.3	CnvNxt-B	62.7	60.6 -2	63.6 $+1$	61.6 -1	32.1 -31	40.8 -22	**30.4** -32	58.2 -4	48.2 -14	44.1 -19	43.2 -19
	84.1	CnvNxt-T	57.8	54.6 -3	53.3 -5	56.9 -1	31.5 -26	40.7 -17	**28.9** -29	50.3 -8	58.3 $+0$	45.9 -12	44.4 -13
	82.3	BiT-m	69.9	59.9 -10	59.3 -11	59.7 -10	29.0 -41	31.8 -38	**26.7** -43	51.7 -18	34.7 -35	30.3 -40	36.7 -33
	85.6	EffNetv2-M	59.2	53.2 -6	56.2 -3	59.2 $+0$	47.5 -12	49.1 -10	36.2 -23	96.0 $+37$	49.9 -9	**32.5** -27	42.6 -17
none	81.1	ViT-B-384	71.7	68.3 -3	66.9 -5	65.5 -6	55.9 -16	**53.6** -18	66.5 -5	62.8 -9	75.5 $+4$	68.5 -3	70.6 -1
	84.6	Swinv2-B-256	73.6	71.2 -2	73.4 -0	70.6 -3	62.9 -11	**59.4** -14	62.0 -12	71.6 -2	73.5 -0	67.4 -6	64.7 -9
	85.1	Deit3-B-384	73.0	77.8 $+5$	89.2 $+16$	69.8 -3	66.4 -7	62.4 -11	**58.2** -15	93.4 $+20$	79.0 $+6$	71.0 -2	58.4 -15
	83.8	Deit3-B-224	74.6	75.6 $+1$	82.8 $+8$	70.4 -4	65.9 -9	62.7 -12	**58.3** -16	79.5 $+5$	84.5 $+10$	74.7 $+0$	60.7 -14
	82.6	XCiT-M-224	77.4	77.0 -0	79.9 $+3$	76.7 -1	72.2 -5	**69.5** -8	71.2 -6	78.4 $+1$	82.8 $+5$	77.3 -0	76.0 -1
	84.3	XCiT-M-224-d	72.6	71.2 -1	75.7 $+3$	68.1 -5	66.9 -6	**62.6** -10	63.8 -9	78.2 $+6$	77.2 $+5$	68.1 -4	65.4 -7
	84.4	CnvNxt-B	68.3	70.6 $+2$	85.9 $+18$	72.4 $+4$	68.2 -0	65.5 -3	64.6 -4	80.9 $+13$	76.5 $+8$	68.0 -0	**64.5** -4
	78.0	BiT-s	78.6	78.2 -0	78.8 $+0$	56.6 -22	85.6 $+7$	**42.2** -36	67.7 -11	69.0 -10	82.6 $+4$	63.2 -15	76.1 -3
	85.0	EffNetv2-M	72.3	71.6 -1	78.6 $+6$	67.0 -5	**54.1** -18	55.2 -17	55.4 -17	63.7 -9	67.2 -5	62.2 -10	58.4 -14
	84.9	EffNetb7	73.4	75.9 $+3$	87.1 $+14$	65.4 -8	54.5 -19	**54.0** -19	58.9 -14	74.7 $+1$	69.0 -4	65.6 -8	56.6 -17
	77.7	EffNet-B0	75.8	76.9 $+1$	84.8 $+9$	68.6 -7	71.0 -5	62.8 -13	67.0 -9	77.9 $+2$	86.2 $+10$	**49.1** -27	54.9 -21
	80.4	ResNet50	77.2	77.4 $+0$	81.6 $+4$	68.1 -9	78.8 $+2$	69.2 -8	80.3 $+3$	96.2 $+19$	64.5 -13	54.3 -23	**53.4** -24
JFT	86.8	EffNetb7-ns	80.1	80.3 $+0$	87.1 $+7$	73.9 -6	73.8 -6	64.2 -16	76.3 -4	80.5 $+0$	73.0 -7	66.6 -13	**63.5** -17
clip +12k	87.2	ViT-B-384-l2b	58.8	56.5 -2	61.8 $+3$	58.6 -0	40.9 -18	41.9 -17	**35.1** -24	54.2 -5	45.7 -13	42.1 -17	42.4 -16
	87.0	ViT-B-384-oai	62.2	58.0 -4	58.0 -4	61.2 -1	39.9 -22	42.6 -20	**34.7** -27	56.2 -6	37.0 -25	38.8 -23	38.7 -24
clip	86.6	ViT-B-384-l2b	69.7	68.1 -2	70.4 $+1$	64.3 -5	53.9 -16	54.2 -16	**53.8** -16	67.2 -3	58.7 -11	55.0 -15	54.2 -15
	86.2	ViT-B-384-oai	69.7	69.1 -1	70.7 $+1$	65.7 -4	56.3 -13	58.5 -11	**52.4** -17	68.5 -1	60.6 -9	58.4 -11	58.5 -11
clip	74.2	clip-ViT-L-336	—	—	—	—	—	—	—	—	—	61.4	**60.6**
z. shot	66.6	clip-ViT-B-224	—	—	—	—	—	—	—	—	—	72.5	**68.5**

Table 18: Mean FPR for classes with 21k overlap.

pre	acc.	model	MSP	MaxL	Ener	KL-M	Maha	RMaha	ViM	E+R	KNN	Cos	MCM/RCos
21k	86.0	ViT-B-384	52.6	41.8 -11	41.9 -11	51.5 -1	**28.2** -24	30.6 -22	37.0 -16	43.8 -9	70.3 $+18$	50.8 -2	50.3 -2
	84.5	ViT-B-224	58.5	49.2 -9	49.6 -9	57.0 -2	**33.3** -25	35.2 -23	44.0 -15	53.6 -5	76.3 $+18$	58.4 -0	59.8 $+1$
	86.3	Swinv2-B-256	51.8	41.8 -10	40.7 -11	58.0 $+6$	63.3 $+12$	52.1 $+0$	58.8 $+7$	**39.5** -12	69.2 $+17$	56.1 $+4$	51.6 -0
	86.7	Deit3-B-384	63.2	58.0 -5	58.6 -5	61.4 -2	55.7 -7	**50.4** -13	61.1 -2	54.8 -8	59.7 -3	52.4 -11	51.8 -11
	85.7	Deit3-B-224	67.0	62.4 -5	62.0 -5	68.1 $+1$	61.6 -5	**54.5** -12	67.9 $+1$	59.8 -7	65.6 -1	58.6 -8	57.5 -10
	86.3	CnvNxt-B	48.7	42.4 -6	44.6 -4	56.6 $+8$	47.3 -1	**40.2** -9	44.3 -4	40.9 -8	60.0 $+11$	47.4 -1	46.0 -3
	84.1	CnvNxt-T	55.7	50.0 -6	47.7 -8	62.6 $+7$	48.5 -7	46.2 -9	**42.5** -13	46.2 -10	67.8 $+12$	55.7 $+0$	53.8 -2
	82.3	BiT-m	67.4	58.9 -9	59.2 -8	65.7 -2	54.2 -13	**43.2** -24	44.3 -23	49.6 -18	63.8 -4	57.5 -10	57.5 -10
	85.6	EffNetv2-M	54.1	52.9 -1	59.6 $+6$	56.8 $+3$	66.0 $+12$	**52.5** -2	62.7 $+9$	89.6 $+36$	78.3 $+24$	53.6 -0	55.0 $+1$
none	81.1	ViT-B-384	72.0	70.5 -2	70.4 -2	67.8 -4	58.2 -14	**52.9** -19	71.0 -1	66.8 -5	79.3 $+7$	73.2 $+1$	74.4 $+2$
	84.6	Swinv2-B-256	71.0	68.4 -3	74.0 $+3$	69.1 -2	66.3 -5	**61.1** -10	71.8 $+1$	72.3 $+1$	74.7 $+4$	68.3 -3	66.2 -5
	85.1	Deit3-B-384	67.6	72.9 $+5$	87.8 $+20$	66.5 -1	67.7 $+0$	61.1 -6	61.7 -6	89.0 $+21$	81.9 $+14$	72.7 $+5$	**60.7** -7
	83.8	Deit3-B-224	69.6	71.2 $+2$	82.1 $+12$	69.6 -0	72.7 $+3$	65.9 -4	65.2 -4	85.0 $+15$	87.6 $+18$	79.4 $+10$	**64.8** -5
	82.6	XCiT-M-224	73.0	72.7 -0	79.0 $+6$	74.0 $+1$	67.9 -5	**64.2** -9	66.4 -7	77.7 $+5$	76.9 $+4$	71.2 -2	69.1 -4
	84.3	XCiT-M-224-d	68.9	65.6 -3	71.2 $+2$	68.5 -0	69.2 $+0$	**63.1** -6	65.1 -4	74.6 $+6$	77.2 $+8$	69.9 $+1$	67.2 -2
	84.4	CnvNxt-B	62.7	67.3 $+5$	86.4 $+24$	72.7 $+10$	70.0 $+7$	63.0 $+0$	70.3 $+8$	83.4 $+21$	78.1 $+15$	66.3 $+4$	**62.1** -1
	78.0	BiT-s	82.5	84.6 $+2$	85.8 $+3$	69.1 -13	84.8 $+2$	**61.9** -21	78.9 -4	78.8 -4	85.5 $+3$	76.0 -6	87.0 $+4$
	85.0	EffNetv2-M	66.0	66.3 $+0$	76.4 $+10$	64.0 -2	66.9 $+1$	**56.6** -9	78.3 $+12$	75.0 $+9$	70.9 $+5$	64.1 -2	58.9 -7
	84.9	EffNetb7	69.4	71.7 $+2$	84.0 $+15$	65.5 -4	75.2 $+6$	**57.9** -11	81.3 $+12$	81.9 $+13$	73.1 $+4$	64.7 -5	59.1 -10
	77.7	EffNet-B0	73.8	72.7 -1	78.1 $+4$	73.6 -0	88.2 $+14$	74.8 $+1$	81.9 $+8$	75.4 $+2$	87.1 $+13$	**62.7** -11	69.4 -4
	80.4	ResNet50	72.8	75.9 $+3$	80.4 $+8$	69.9 -3	88.3 $+15$	65.8 -7	83.9 $+11$	98.6 $+26$	81.7 $+9$	62.8 -10	**62.5** -10
JFT	86.8	EffNetb7-ns	64.3	**60.5** -4	68.9 $+5$	64.6 $+0$	84.2 $+20$	62.8 -2	87.1 $+23$	69.1 $+5$	75.7 $+11$	65.3 $+1$	63.1 -1
clip +12k	87.2	ViT-B-384-l2b	51.8	49.4 -2	52.3 $+1$	53.9 $+2$	53.9 $+2$	47.9 -4	48.8 -3	48.0 -4	51.6 -0	44.2 -8	**44.2** -8
	87.0	ViT-B-384-oai	51.1	46.5 -5	46.5 -5	53.1 $+2$	58.7 $+8$	49.8 -1	55.0 $+4$	45.4 -6	53.2 $+2$	44.6 -6	**44.4** -7
clip	86.6	ViT-B-384-l2b	62.1	62.3 $+0$	67.4 $+5$	59.6 -3	58.3 -4	54.8 -7	58.4 -4	65.4 $+3$	63.6 $+2$	56.0 -6	**53.6** -8
	86.2	ViT-B-384-oai	66.9	67.0 $+0$	70.8 $+4$	64.7 -2	57.8 -9	**54.7** -12	60.6 -6	69.3 $+2$	66.4 -1	60.7 -6	57.7 -9
clip	74.2	clip-ViT-L-336	—	—	—	—	—	—	—	—	—	**67.4**	68.8
z. shot	66.6	clip-ViT-B-224	—	—	—	—	—	—	—	—	—	**79.6**	81.1