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

Julian Bitterwolf*
University of Tübingen
Tübingen AI Center

Maximilian Müller*
University of Tübingen
Tübingen AI Center

Matthias Hein
University of Tübingen
Tübingen AI Center

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-art 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**

*Equal contribution. Correspondence to: julian.bitterwolf@uni-tuebingen.de.

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.

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 an aspect of the specific instance of the shown main object, e.g. the IN-1k class *plane* in an image of the OOD category *sky*. 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 5 879 images that we individually checked not to contain any ID object from the classes in IN-1K. These images are ordered into 64 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 using pre-logit feature-information directly, 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

Table 1: Percentage of ID samples, $p = \frac{\text{ID}}{\text{ID+OOD}}$, 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). An overview over the datasets is given in Appendix B.

Dataset	% ID	Dataset	% ID	Dataset	% ID	Dataset	% ID
PLACES	59.5	SPECIES	57.0	IMAGENET-O	20.2	SSB-EASY	53.4
INAT. PLANTS	2.5	TEXTURES	25.6	OPENIM.-O	4.9	SSB-HARD	41.6
COOD	38.2	TEXTURES43	20.0	360OPENSET	26.9	IN-1K-OOD	32.1

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 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 1 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

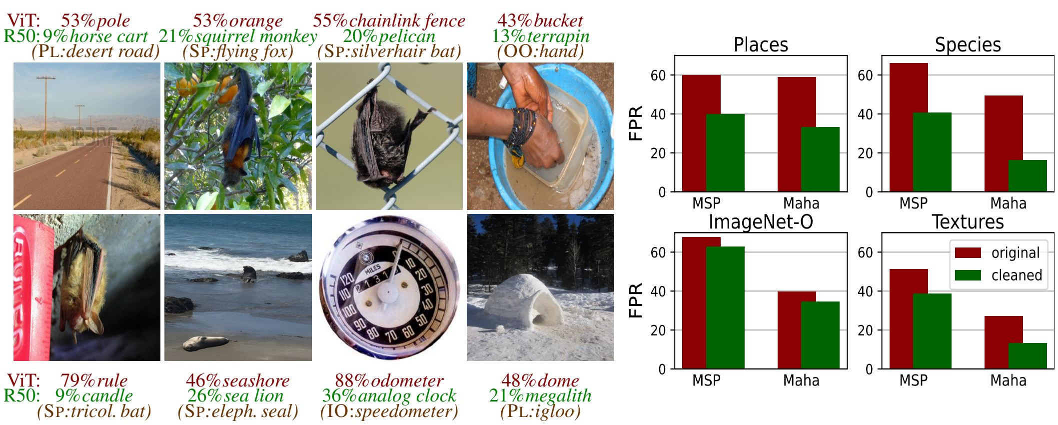

Figure 2: **Left: A Vision Transformer confidently classifies ID objects in samples from popular OOD datasets** *(source label in parentheses)* as the correct IN-1K class, but is marked down with false positives in OOD detection evaluation when using MSP (Max Softmax Prob.) as criterion. The weaker **ResNet-50**, 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.

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

Figure 3: **Difficult OOD classes in NINCO**: Examples of images from some of NINCO's most difficult (see Table 7) *OOD classes* (first row) and from the *ImageNet-1K class* (second row) which the *OOD class* is most frequently confused for.

"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 1, we report 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 (25.6% gain of MSP vs. 33.2% gain of Mahalanobis) and PLACES (19.6% gain vs. 26.3% gain), which are two of the datasets we found to contain most ID samples. We further emphasize that due to the presence of large fractions of ID samples in most common benchmarks, even a perfect detector's measured performance would saturate significantly above 0% FPR. For example with Species, we find that for a strong current detector already more than 85% of the 'false positives' contain ID objects.

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.

Sourcing OOD test samples for ImageNet-1K from ImageNet-21K (or its subsets) based on class-labels has been leading to highly contaminated datasets (5 of the datasets in Table 1 are sourced from ImageNet-21K and all contain between 20% and 53% ID samples and significant categorical contamination). This is partly due to the class-structure of those datasets: Both ImageNet-1K and ImageNet-21K contain leaf and internal nodes of the WordNet-tree as classes. While the internal

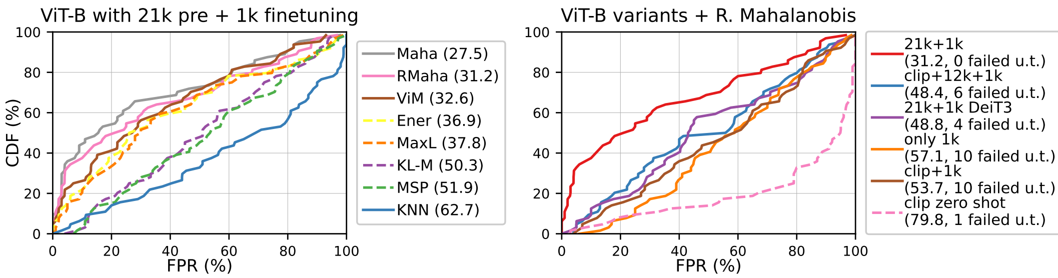

Figure 4: Cumulative distribution of the % of NINCO-classes for which an FPR at least as low as a given x-value is achieved. The area over this curve corresponds to the mean FPR. The further in the top left corner, the better. **The best methods access pre-logit features (Left):** Different OOD detection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

nodes of ImageNet-1K are not ancestors to other Imagenet-1K classes, ImageNet-21K internal nodes can be ancestors to ImageNet-1K nodes, and vice versa. Moreover, there are ambigous class-definitions in WordNet, like e.g. *police dog*, which is not parent or child of another dog class, but mostly shows a *german shepherd*, or an *alley cat* showing one of the many cat classes without being parent or child to other cat classes. Besides, there is significant incidental contamination even for nominally disjoint classes. Since the automation of filtering for challenging OOD data would require a strong detector that already solves the problems that the dataset is meant to pose, we conclude that it is impossible to construct a clean and challenging OOD dataset without manually checking the OOD samples for ID contamination.

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[2] 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.

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 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

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

[2]One can look up synsets with glosses here.

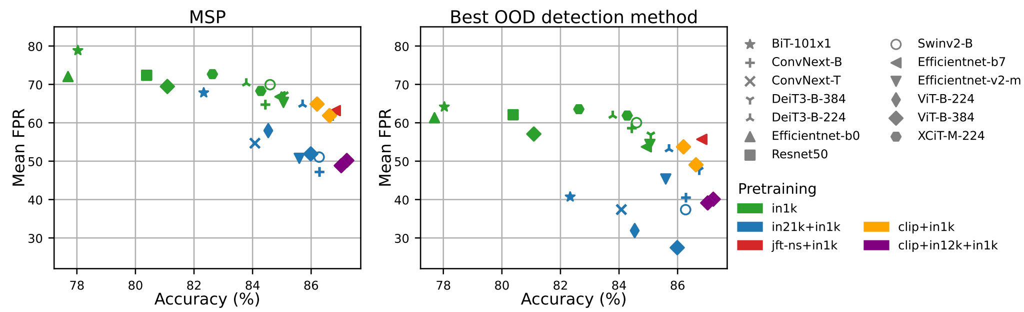

Figure 5: **IN-21K pretraining boosts OOD detectors accessing pre-logit features on NINCO:** Mean FPR vs. accuracy for MSP and each model's best detector (which except for the noisy-student model always accesses the pre-logit features). OOD detection strongly improves when using models pretrained on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

is labelled as. For data from other base classes, 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. Such high data quality is crucial for in-depth evaluations (Vasudevan et al., 2022; Shankar et al., 2021), as only being completely in-distribution free allows understanding a detector's individual mistakes.

The NINCO (**N**o **I**mageNet **C**lass **O**bjects) dataset consists of 64 OOD classes with a total of 5 879 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 Appendix B, as well as from the FOOD-101 dataset (7 classes) (Bossard et al., 2014), CALTECH-101 (4 classes) (Li et al., 2022), MYNURSINGHOME (4 classes) (Ismail et al., 2020), ImageNet-21k (1 class) and newly scraped from `iNaturalist.org` (2 classes) or other websites like Flickr (8 classes). Details for all NINCO OOD classes are given in Appendix G. We show samples from all NINCO classes in Figures 9, 10 and 11 in Appendix I. In addition to NINCO, we also provide the 2715 OOD images obtained from cleaning 400 samples of eleven 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. 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 directly 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) and Cosine-based methods show the most consistently good results across models.
- Results on the cleaned subsets of eleven 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.

Acknowledgements

We thank Vaclav Voracek for helpful discussions and suggesting the cdf plots. We acknowledge support from the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039A) and from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC number 2064/1, Project number 390727645), as well as from the DFG TRR 248 (Project number 389792660) and of the Carl Zeiss Foundation in the project "Certification and Foundations of Safe Machine Learning Systems in Healthcare". We also thank the European Laboratory for Learning and Intelligent Systems (ELLIS) for supporting Maximilian Müller.

References

Bendale, A. and Boult, T. E. Towards open set deep networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. URL `https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.173`.

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.

Galil, I., Dabbah, M., and El-Yaniv, R. A framework for benchmarking class-out-of-distribution detection and its application to imagenet. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Iuubb9W6Jtk.

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.

Ismail, A., Ahmad, S. A., Che Soh, A., Hassan, M. K., and Harith, H. H. Mynursinghome: A fully-labelled image dataset for indoor object classification. *Data in Brief*, 2020. doi: https://doi.org/10.1016/j.dib.2020.106268. URL https://www.sciencedirect.com/science/article/pii/S2352340920311628.

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.

Li, F.-F., Andreeto, M., Ranzato, M., and Perona, P. Caltech 101, 2022.

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.

Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., and Schmidt, L. Evaluating machine accuracy on imagenet. In *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*, 2021. URL `https://openreview.net/forum?id=Q3R088EFtng`.

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.

Vasudevan, V., Caine, B., Gontijo-Lopes, R., Fridovich-Keil, S., and Roelofs, R. When does dough become a bagel? analyzing the remaining mistakes on imagenet. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=mowt1WNhTC7`.

Vaze, S., Han, K., Vedaldi, A., and Zisserman, A. Open-set recognition: a good closed-set classifier is all you need? In *International Conference on Learning Representations*, 2022.

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*, 2022a.

Wang, H., Zhang, A., Zhu, Y., Zheng, S., Li, M., Smola, A., and Wang, Z. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In *ICML 2022*, 2022b.

Wen, X., Zhao, B., and Qi, X. A simple parametric classification baseline for generalized category discovery. *ArXiv*, abs/2211.11727, 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., 2022a), 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 directly. For the zero-shot evaluation of CLIP, we use Maximum-Concept-Matching (MCM) (Ming et al., 2022) and Cosine-similarity (Cos) (Galil et al., 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 call methods which directly access the pre-logit feature layer *feature-based* and 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 and Energy, which do not access the pre-logit features directly, are 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 3 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 together with (Relative) Cosine the only method which can fairly 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 except the noisy-student model, the best-performing method always is always feature-based, 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 eleven 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 is feature-based. 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

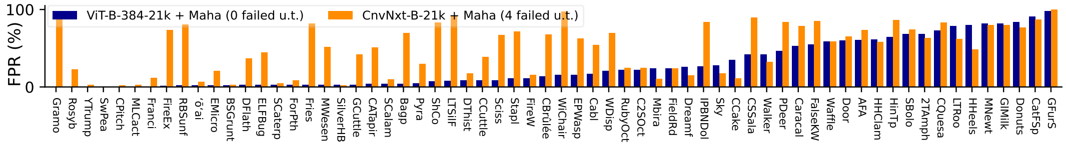

Figure 6: FPR of a pretrained ViT-B and pretrained ConvNext-B for all classes of NINCO.

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%) 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 (Galil et al., 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 2: **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	*bla*	*whi*	*gre*	*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	60.5	0.8	0.0	0.0
	ViM	100.0	100.0	100.0	98.0	24.5	100.0	100.0
	Maha	100.0	100.0	100.0	87.5	27.5	100.0	100.0
	Cos	0.0	0.0	0.0	27.5	0.0	0.0	0.0
Cnv21k	MSP	0.0	0.0	0.0	13.5	2.2	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 2 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 3, where we show the mean FPR for all models and methods across the dataset's OOD classes. Tables 4-6 show AUROC, AUPR-S and AUPR-E with the same conclusions. 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; Galil et al., 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 9. 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.

In Figure 7 we contrast the results on NINCO with the results from previously used datasets. We show all methods for a pretrained ViT-B-384 and all models for the MSP-baseline. In both cases we observe several ranking changes: For the ViT, the best-performing method changes from ViM to Mahalanobis, and Relative Mahalanobis improves from sixth to second place. For the MSP-baseline, the clip-pretrained ViTs were the strongest OOD detectors on the previously used datasets, but are outperformend by the ConvNext-B on NINCO.

Table 3: **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	51.9	37.8 −14	36.9 −15	50.3 −2	**27.5** −24	31.2 −21	32.6 −19	38.5 −13	62.7 +11	46.0 −6	45.0 −7
	84.5	ViT-B-224	58.0	46.5 −12	46.1 −12	57.2 −1	**31.9** −26	36.8 −21	38.4 −20	49.4 −9	68.8 +11	54.7 −3	54.3 −4
	86.3	Swinv2-B-256	51.1	41.1 −10	40.0 −11	56.0 +5	62.8 +12	53.8 +3	54.8 +4	**37.4** −14	61.9 +11	51.4 +0	48.2 −3
	86.7	Deit3-B-384	61.8	56.0 −6	56.3 −5	60.3 −1	53.9 −8	48.8 −13	56.9 −5	51.6 −10	53.4 −8	48.4 −13	**47.7** −14
	85.7	Deit3-B-224	64.8	59.2 −6	58.1 −7	65.2 +0	60.0 −5	53.8 −11	62.5 −2	55.2 −10	58.7 −6	54.2 −11	**53.2** −12
	86.3	CnvNxt-B	47.2	41.1 −6	43.3 −4	54.9 +8	49.6 +2	42.4 −5	41.5 −6	**40.5** −7	51.8 +5	44.2 −3	42.6 −5
	84.1	CnvNxt-T	54.7	47.9 −7	45.4 −9	60.7 +6	46.9 −8	45.8 −9	**37.4** −17	44.1 −11	56.6 +2	51.2 −4	49.2 −5
	82.3	BiT-m	67.8	62.0 −6	63.2 −5	64.9 −3	50.0 −18	45.1 −23	**40.7** −27	57.1 −11	58.0 −10	51.6 −16	54.4 −13
	85.6	EffNetv2-M	50.7	48.3 −2	54.1 +3	54.6 +4	62.9 +12	51.6 +1	53.5 +3	89.8 +39	67.5 +17	**45.4** −5	50.6 −0
none	81.1	ViT-B-384	69.5	67.7 −2	68.1 −1	66.7 −3	60.0 −9	**57.1** −12	69.4 −0	65.8 −4	73.6 +4	68.7 −1	69.8 +0
	84.6	Swinv2-B-256	69.9	67.6 −2	72.2 +2	67.5 −2	63.9 −6	**60.0** −10	66.5 −3	68.8 −1	69.2 −1	63.5 −6	62.0 −8
	85.1	Deit3-B-384	67.3	72.8 +5	87.6 +20	64.6 −3	64.0 −3	59.4 −8	60.0 −7	90.2 +23	74.4 +7	67.1 −0	**56.9** −10
	83.8	Deit3-B-224	70.3	71.9 +2	82.3 +12	68.4 −2	69.0 −1	64.3 −6	63.5 −7	83.1 +13	80.4 +10	73.0 +3	**61.9** −8
	82.6	XCiT-M-224	72.7	73.3 +1	79.2 +6	71.8 −1	66.2 −6	**63.5** −9	64.9 −8	76.4 +4	71.8 −1	67.1 −6	66.0 −7
	84.3	XCiT-M-224-d	68.3	66.2 −2	73.1 +5	66.9 −1	66.4 −2	**61.9** −6	62.3 −6	72.4 +4	70.4 +2	64.6 −4	62.6 −6
	84.4	CnvNxt-B	64.7	71.5 +7	89.1 +24	68.0 +3	65.8 +1	60.6 −4	65.4 +1	85.9 +21	70.5 +6	61.3 −3	**58.6** −6
	78.0	BiT-s	78.8	81.2 +2	82.9 +4	68.4 −10	83.5 +5	**64.1** −15	73.5 −5	77.8 −1	83.2 +4	72.1 −7	84.1 +5
	85.1	EffNetv2-M	65.3	65.3 +0	74.5 +9	62.8 −2	62.5 −3	54.9 −10	72.5 +7	69.6 +4	64.4 −1	59.6 −6	**54.4** −11
	84.9	EffNetb7	66.8	69.0 +2	81.5 +15	62.7 −4	68.1 +1	54.6 −12	72.7 +6	76.3 +10	66.8 +0	60.5 −6	**53.7** −13
	77.7	EffNet-B0	72.0	72.4 +0	79.6 +8	72.3 +0	83.3 +11	74.0 +2	75.2 +3	75.1 +3	86.9 +15	**61.3** −11	69.8 −2
	80.4	ResNet50	72.4	74.3 +2	77.9 +6	69.0 −3	85.9 +13	69.5 −3	78.6 +6	97.4 +25	77.9 +6	63.0 −9	**62.1** −10
JFT	86.8	EffNetb7-ns	63.2	**55.7** −7	61.5 −2	64.5 +1	87.4 +24	68.7 +6	89.2 +26	61.7 −1	73.8 +11	65.2 +2	63.7 +1
clip +12k	87.2	ViT-B-384-l2b	50.2	47.4 −3	50.3 +0	52.2 +2	52.6 +2	47.3 −3	45.8 −4	44.9 −5	45.4 −5	**40.1** −10	40.2 −10
	87.0	ViT-B-384-oai	48.8	43.7 −5	44.1 −5	49.6 +1	57.7 +9	48.4 −0	52.5 +4	42.2 −7	45.0 −4	39.3 −10	**39.1** −10
clip	86.6	ViT-B-384-l2b	61.9	61.6 −0	65.8 +4	57.5 −4	52.7 −9	50.5 −11	51.7 −10	63.2 +1	57.0 −5	50.8 −11	**49.1** −13
	86.2	ViT-B-384-oai	64.9	64.9 +0	69.7 +5	61.8 −3	55.7 −9	**53.7** −11	56.9 −8	67.3 +2	61.4 −4	56.6 −8	54.3 −11
clip z. shot	74.3	clip-ViT-L-336	—	—	—	—	—	—	—	—	—	72.5	**67.1**
	66.6	clip-ViT-B-224	—	—	—	—	—	—	—	—	—	**79.1**	79.8

Table 4: **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	87.2	92.5 +5	92.7 +5	86.9 −0	**95.0** +8	94.0 +7	94.0 +7	92.5 +5	85.9 −1	91.5 +4	91.7 +4
	84.5	ViT-B-224	85.5	90.6 +5	90.8 +5	85.2 −0	**94.0** +8	92.8 +7	92.5 +7	90.1 +5	82.6 −3	89.5 +4	89.4 +4
	86.3	Swinv2-B-256	86.3	87.0 +1	85.8 −0	86.1 −0	88.0 +2	89.0 +3	**89.9** +4	88.8 +3	84.3 −2	89.1 +3	89.8 +4
	86.7	Deit3-B-384	81.1	77.7 −3	74.9 −6	83.6 +2	89.7 +9	90.0 +9	89.0 +8	80.7 −0	87.0 +6	90.1 +9	**90.4** +9
	85.7	Deit3-B-224	80.3	77.3 −3	74.8 −6	82.4 +2	88.3 +8	88.8 +9	87.6 +7	79.5 −1	85.2 +5	88.8 +8	**89.0** +9
	86.3	CnvNxt-B	87.9	87.6 −0	85.8 −2	88.0 +0	90.9 +3	91.8 +4	**92.5** +5	87.9 −0	87.5 −0	91.5 +4	91.8 +4
	84.1	CnvNxt-T	85.1	86.0 +1	85.3 +0	85.2 +0	89.8 +5	89.5 +4	**92.5** +7	86.0 +1	86.0 +1	89.1 +4	89.6 +5
	82.3	BiT-m	82.2	83.3 +1	82.5 +0	82.8 +1	90.6 +8	90.0 +8	**92.2** +10	85.8 +4	86.4 +4	89.6 +7	88.5 +6
	85.6	EffNetv2-M	86.3	85.1 −1	82.7 −4	87.3 +1	87.6 +1	88.9 +3	87.9 +2	73.0 −13	83.8 −3	**90.1** +4	88.8 +3
none	81.1	ViT-B-384	81.4	84.2 +3	84.2 +3	80.7 −1	86.5 +5	**87.3** +6	82.6 +1	84.6 +3	79.7 −2	84.4 +3	84.1 +3
	84.6	Swinv2-B-256	80.4	77.8 −3	75.0 −5	81.9 +1	86.2 +6	**86.7** +6	81.1 +1	80.2 −0	81.9 +1	85.8 +5	86.3 +6
	85.1	Deit3-B-384	81.7	76.4 −5	66.6 −15	83.5 +2	86.9 +5	**88.0** +6	84.8 +3	61.8 −20	80.4 −1	85.5 +4	87.0 +5
	83.8	Deit3-B-224	81.0	78.8 −2	74.8 −6	82.3 +1	85.5 +5	**86.9** +6	84.0 +3	74.6 −6	77.6 −3	83.8 +3	85.6 +5
	82.6	XCiT-M-224	77.9	72.2 −6	64.5 −13	81.2 +3	85.1 +7	85.7 +8	**86.0** +8	73.3 −5	80.9 +3	84.7 +7	85.0 +7
	84.3	XCiT-M-224-d	82.7	80.2 −3	74.1 −9	82.9 +0	85.5 +3	**86.8** +4	85.3 +3	78.6 −4	81.4 −1	85.8 +3	86.1 +3
	84.4	CnvNxt-B	81.1	76.2 −5	64.6 −16	83.4 +2	85.2 +4	86.6 +6	82.5 +1	72.9 −8	81.3 +0	85.8 +5	**86.9** +6
	78.0	BiT-s	80.1	77.3 −3	75.6 −5	82.3 +2	71.2 −9	**84.9** +5	77.9 −2	76.2 −4	68.8 −11	78.3 −2	69.8 −10
	85.1	EffNetv2-M	81.8	78.3 −3	71.8 −10	84.0 +2	86.5 +5	**88.9** +7	80.1 −2	79.0 −3	83.4 +2	87.3 +6	88.1 +6
	84.9	EffNetb7	79.6	72.8 −7	64.6 −15	84.2 +5	84.5 +5	**88.6** +9	81.6 +2	71.8 −8	82.5 +3	86.9 +7	87.9 +8
	77.7	EffNet-B0	80.8	78.5 −2	74.9 −6	81.9 +1	76.7 −4	82.7 +2	81.6 +1	79.1 −2	76.2 −5	**85.0** +4	82.3 +1
	80.4	ResNet50	81.5	81.5 +0	81.2 −0	79.3 −2	75.8 −6	85.0 +4	81.3 −0	64.6 −17	76.3 −5	84.9 +3	**85.5** +4
JFT	86.8	EffNetb7-ns	83.6	82.5 −1	78.6 −5	83.1 −0	78.1 −5	**86.6** +3	74.6 −9	81.1 −2	79.2 −4	85.2 +2	85.0 +1
clip +12k	87.2	ViT-B-384-l2b	86.1	82.6 −4	78.1 −8	88.8 +3	90.5 +4	91.1 +5	91.9 +6	83.4 −3	89.5 +3	**92.2** +6	92.1 +6
	87.0	ViT-B-384-oai	87.2	85.8 −1	84.2 −3	88.4 +1	89.6 +2	91.1 +4	90.8 +4	86.5 −1	89.9 +3	92.5 +5	**92.5** +5
clip	86.6	ViT-B-384-l2b	81.1	73.5 −8	68.5 −13	85.9 +5	89.1 +8	89.1 +8	88.8 +8	71.7 −9	86.1 +5	89.8 +9	**90.0** +9
	86.2	ViT-B-384-oai	78.8	70.4 −8	65.0 −14	84.4 +6	88.6 +10	88.5 +10	88.3 +10	68.1 −11	84.6 +6	88.6 +10	**89.2** +10
clip z. shot	74.3	clip-ViT-L-336	—	—	—	—	—	—	—	—	—	79.7	**81.1**
	66.6	clip-ViT-B-224	—	—	—	—	—	—	—	—	—	74.0	**74.9**

Table 5: **Mean AUPR-S 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
	86.0	ViT-B-384	97.2	98.4 +1	98.4 +1	96.6 −1	**99.0** +2	98.8 +2	98.7 +2	98.4 +1	97.0 −0	98.2 +1	98.3 +1
	84.5	ViT-B-224	96.7	97.9 +1	98.0 +1	96.1 −1	**98.7** +2	98.5 +2	98.4 +2	97.8 +1	96.1 −1	97.8 +1	97.8 +1
	86.3	Swinv2-B-256	96.2	95.8 −0	95.2 −1	96.6 +0	97.4 +1	97.5 +1	97.9 +2	96.6 +0	96.5 +0	97.7 +2	**97.9** +2
	86.7	Deit3-B-384	94.1	91.7 −2	90.4 −4	95.9 +2	97.9 +4	97.9 +4	97.6 +4	93.1 −1	97.1 +3	97.9 +4	**98.0** +4
21k	85.7	Deit3-B-224	94.1	91.7 −2	90.6 −3	95.5 +1	97.5 +3	97.6 +3	97.3 +3	92.7 −1	96.6 +3	97.6 +4	**97.6** +4
	86.3	CnvNxt-B	96.8	96.3 −1	95.7 −1	97.2 +0	98.1 +1	98.2 +1	**98.4** +2	96.4 −0	97.2 +0	98.2 +1	98.3 +1
	84.1	CnvNxt-T	96.0	95.9 −0	95.6 −0	96.4 +0	97.7 +2	97.6 +2	**98.4** +2	95.8 −0	96.9 +1	97.6 +2	97.8 +2
	82.3	BiT-m	95.7	95.7 −0	95.5 −0	95.3 −0	98.0 +2	97.7 +2	**98.3** +3	96.6 +1	97.0 +1	97.7 +2	97.5 +2
	85.6	EffNetv2-M	96.3	95.7 −1	95.0 −1	97.0 +1	97.3 +1	97.5 +1	97.2 +1	93.6 −3	96.5 +0	**97.8** +1	97.5 +1
	81.1	ViT-B-384	95.5	96.1 +1	96.2 +1	94.6 −1	96.9 +1	**97.1** +2	95.7 +0	96.1 +1	95.2 −0	96.4 +1	96.3 +1
	84.6	Swinv2-B-256	94.5	92.6 −2	91.4 −3	95.2 +1	96.9 +2	**97.0** +2	94.9 +0	94.0 −1	95.8 +1	96.9 +2	96.9 +2
	85.1	Deit3-B-384	95.2	92.9 −2	89.6 −6	95.7 +1	97.1 +2	**97.4** +2	96.1 +1	88.2 −7	95.4 +0	96.8 +2	96.8 +2
	83.8	Deit3-B-224	95.1	94.1 −1	93.0 −2	95.5 +0	96.8 +2	**97.1** +2	95.9 +1	93.2 −2	94.8 −0	96.3 +1	96.6 +2
	82.6	XCiT-M-224	93.7	90.7 −3	87.6 −6	95.3 +2	96.5 +3	**96.8** +3	96.7 +3	91.6 −2	95.4 +2	96.5 +3	96.5 +3
	84.3	XCiT-M-224-d	95.8	94.1 −2	92.1 −4	95.6 −0	96.7 +1	**97.0** +1	96.5 +1	93.9 −2	95.6 −0	96.8 +1	96.9 +1
none	84.4	CnvNxt-B	94.9	93.0 −2	89.8 −5	95.8 +1	96.7 +2	96.9 +2	95.6 +1	92.8 −2	95.5 +1	96.8 +2	**97.0** +2
	78.0	BiT-s	95.3	94.7 −1	94.3 −1	95.3 +0	92.7 −3	**96.5** +1	94.8 −0	94.4 −1	92.0 −3	94.7 −1	92.1 −3
	85.1	EffNetv2-M	95.1	92.9 −2	90.4 −5	96.0 +1	97.0 +2	**97.6** +3	94.9 −0	93.8 −1	96.2 +1	97.2 +2	97.1 +2
	84.9	EffNetb7	94.0	90.7 −3	87.9 −6	96.2 +2	96.5 +3	**97.5** +3	95.6 +2	90.9 −3	95.9 +2	97.0 +3	97.3 +3
	77.7	EffNet-B0	95.1	94.1 −1	93.1 −2	95.4 +0	94.6 −0	96.1 +1	95.9 +1	94.6 −1	94.6 −1	**96.5** +1	95.8 +1
	80.4	ResNet50	95.5	95.6 +0	95.5 −0	94.2 −1	94.3 −1	96.7 +1	95.6 +0	91.7 −4	94.1 −1	96.5 +1	**96.7** +1
JFT	86.8	EffNetb7-ns	95.8	94.7 −1	92.9 −3	95.8 +0	95.0 −1	**97.2** +1	94.1 −2	94.4 −1	95.3 −1	96.8 +1	96.6 +1
clip	87.2	ViT-B-384-l2b	95.9	94.0 −2	92.6 −3	97.5 +2	98.0 +2	98.1 +2	98.3 +2	94.5 −1	97.7 +2	98.4 +2	**98.4** +3
+12k	87.0	ViT-B-384-oai	96.6	95.5 −1	94.9 −2	97.3 +1	97.8 +1	98.1 +2	98.1 +2	95.9 −1	97.8 +1	**98.5** +2	98.4 +2
clip	86.6	ViT-B-384-l2b	94.2	90.8 −3	89.2 −5	96.6 +2	97.6 +3	97.5 +3	97.4 +3	90.2 −4	96.8 +3	97.8 +4	**97.8** +4
	86.2	ViT-B-384-oai	93.1	89.6 −3	88.0 −5	96.1 +3	97.5 +4	97.5 +4	97.4 +4	89.2 −4	96.4 +3	97.5 +4	**97.7** +5
clip	74.3	clip-ViT-L-336	—	—	—	—	—	—	—	—	—	95.2	**95.5**
z. shot	66.6	clip-ViT-B-224	—	—	—	—	—	—	—	—	—	93.6	**93.9**

Table 6: **Mean AUPR-E 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
	86.0	ViT-B-384	60.2	69.8 +10	71.3 +11	60.4 +0	**78.9** +19	74.9 +15	74.2 +14	69.4 +9	53.8 −6	65.3 +5	65.3 +5
	84.5	ViT-B-224	56.1	64.9 +9	65.4 +9	56.4 +0	**75.3** +19	72.4 +16	71.3 +15	62.0 +6	48.2 −8	59.5 +3	59.6 +3
	86.3	Swinv2-B-256	61.9	68.2 +6	69.8 +8	54.7 −7	54.7 −7	59.9 −2	59.8 −2	**71.6** +10	52.6 −9	60.1 −2	63.0 +1
	86.7	Deit3-B-384	54.1	55.8 +2	54.7 +1	53.0 −1	59.2 +5	62.0 +8	57.2 +3	60.3 +6	59.0 +5	62.4 +8	**63.4** +9
21k	85.7	Deit3-B-224	50.4	53.5 +3	53.3 +3	49.6 −1	54.9 +4	58.2 +8	53.8 +3	56.7 +6	55.8 +5	59.2 +9	**60.0** +10
	86.3	CnvNxt-B	64.0	68.3 +4	67.0 +3	57.1 −7	65.3 +1	68.2 +4	69.6 +6	**70.2** +6	59.5 −5	67.2 +3	68.1 +4
	84.1	CnvNxt-T	58.3	63.4 +5	65.1 +7	52.4 −6	64.8 +7	65.3 +7	**71.4** +13	65.0 +7	55.3 −3	60.3 +2	62.6 +4
	82.3	BiT-m	47.7	52.5 +5	51.8 +4	51.7 +4	63.5 +16	66.2 +18	**68.8** +21	55.7 +8	56.6 +9	62.1 +14	60.2 +13
	85.6	EffNetv2-M	60.9	62.3 +1	58.7 −2	56.5 −4	54.5 −6	59.6 −1	58.6 −2	31.0 −30	49.7 −11	**65.6** +5	61.5 +1
	81.1	ViT-B-384	47.1	49.6 +2	50.3 +3	49.6 +2	56.6 +9	**58.5** +11	48.3 +1	51.8 +5	43.9 −3	49.7 +3	49.2 +2
	84.6	Swinv2-B-256	46.8	46.9 +0	42.7 −4	48.7 +2	52.1 +5	**54.8** +8	48.3 +1	46.0 −1	46.3 −1	51.7 +5	53.6 +7
	85.1	Deit3-B-384	49.8	43.1 −7	29.0 −21	51.4 +2	51.9 +2	55.5 +6	53.5 +4	25.4 −24	43.1 −7	49.3 −0	**57.0** +7
	83.8	Deit3-B-224	47.0	45.2 −2	35.7 −11	48.5 +2	49.4 +2	51.9 +5	50.9 +4	35.5 −11	38.5 −9	45.8 −1	**52.7** +6
	82.6	XCiT-M-224	42.9	40.1 −3	33.5 −9	46.8 +4	52.6 +10	**54.2** +11	52.6 +10	38.6 −4	44.6 +2	50.3 +7	50.3 +7
	84.3	XCiT-M-224-d	49.0	48.4 −1	41.7 −7	50.2 +1	51.3 +2	**53.7** +5	53.1 +4	43.6 −5	46.0 −3	51.8 +3	52.6 +4
none	84.4	CnvNxt-B	49.6	43.7 −6	27.7 −22	48.8 −1	51.4 +2	**55.5** +6	50.3 +1	33.3 −16	45.6 −4	53.1 +3	55.5 +6
	78.0	BiT-s	40.5	37.5 −3	36.5 −4	49.1 +9	33.6 −7	**51.5** +11	42.1 +2	39.4 −1	32.8 −8	43.1 +3	33.1 −7
	85.1	EffNetv2-M	50.1	48.6 −1	39.0 −11	52.3 +2	53.8 +4	58.8 +9	45.3 −5	45.0 −5	52.2 +2	55.1 +5	**59.1** +9
	84.9	EffNetb7	48.3	43.9 −4	31.5 −17	53.8 +5	49.5 +1	59.2 +11	45.5 −3	39.4 −9	49.8 +2	54.0 +6	**60.3** +12
	77.7	EffNet-B0	45.3	44.1 −1	37.9 −7	45.9 +1	36.3 −9	45.1 −0	43.0 −2	43.0 −2	33.7 −12	**54.2** +9	48.1 +3
	80.4	ResNet50	44.8	44.7 −0	42.2 −3	48.0 +3	34.2 −11	48.4 +4	41.5 −3	22.0 −23	39.4 −5	53.1 +8	**54.3** +9
JFT	86.8	EffNetb7-ns	52.7	**56.3** +4	50.6 −2	49.3 −3	36.1 −17	50.2 −3	33.1 −20	50.6 −2	43.1 −10	50.4 −2	51.5 −1
clip	87.2	ViT-B-384-l2b	61.0	61.4 +0	58.3 −3	58.2 −3	62.4 +1	65.2 +4	66.9 +6	63.4 +2	64.2 +3	**68.2** +7	68.0 +7
+12k	87.0	ViT-B-384-oai	62.2	64.9 +3	64.5 +2	60.5 −2	59.5 −3	64.2 +2	62.3 +0	66.8 +5	63.8 +2	69.4 +7	**69.7** +7
clip	86.6	ViT-B-384-l2b	51.9	48.7 −3	43.8 −8	56.1 +4	61.0 +9	61.3 +9	60.2 +8	46.3 −6	55.9 +4	60.7 +9	**63.0** +11
	86.2	ViT-B-384-oai	50.2	44.6 −6	39.5 −11	53.9 +4	59.8 +10	59.8 +10	58.9 +9	42.1 −8	53.5 +3	58.3 +8	**60.2** +10
clip	74.3	clip-ViT-L-336	—	—	—	—	—	—	—	—	—	44.2	**48.5**
z. shot	66.6	clip-ViT-B-224	—	—	—	—	—	—	—	—	—	37.5	**38.0**

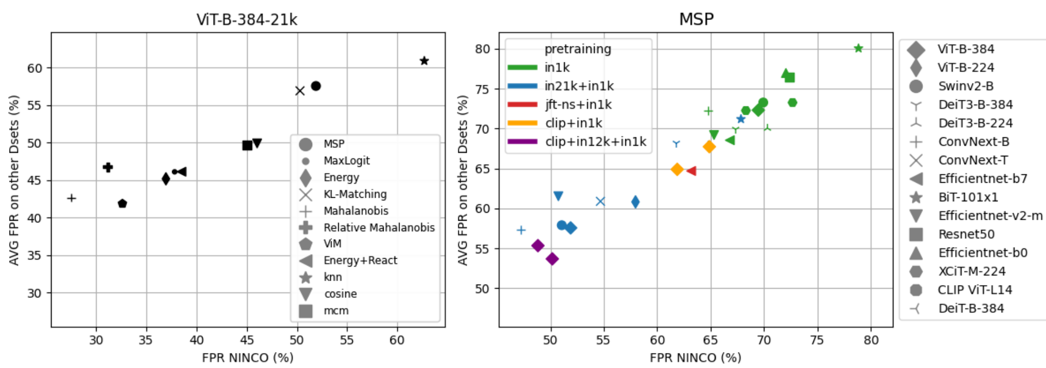

Figure 7: **Mean FPR on NINCO vs. mean-FPR on previously used datasetes** with fixed model (left) and fixed method (right). We observe several ranking changes, including the best-performing method and model.

Table 7: 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	77.0	74.0	67.0	79.0	79.0	76.0	86.0	67.0	87.0	77.0	77.0
2TAmph	75.6	90.9	91.5	70.5	68.8	54.0	92.0	93.8	95.5	86.9	84.7	79.5	85.8	91.5	69.3	63.6	54.5	72.7	86.9	83.5	72.2	71.0
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	82.6	69.6	89.1	73.9	80.4	45.7	65.2	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	78.0	77.0	86.0	92.0	87.0	81.0	88.0	86.0	98.0	96.0	95.0
GFurS	100.0	98.9	96.7	95.6	97.8	96.7	75.8	100.0	98.9	100.0	100.0	98.9	98.9	81.3	96.7	100.0	100.0	100.0	87.9	100.0	100.0	100.0
Bagp	8.2	2.1	2.1	12.4	4.1	4.1	4.1	1.0	11.3	6.2	4.1	16.5	5.2	2.1	37.1	70.1	20.6	18.6	3.1	12.4	8.2	7.2
CSSala	76.0	86.0	89.0	76.0	42.0	60.0	57.0	93.0	100.0	94.0	94.0	78.0	71.0	66.0	79.0	90.0	70.0	92.0	66.0	97.0	86.0	82.0
Cabl	46.6	15.9	14.8	56.8	17.0	42.0	11.4	21.6	45.5	48.9	38.6	35.2	36.4	48.9	52.3	54.5	60.2	29.5	48.9	23.9	34.1	31.8
CQuesa	86.0	95.0	97.0	81.0	73.0	77.0	82.0	88.0	99.0	98.0	98.0	85.0	87.0	93.0	85.0	83.0	78.0	81.0	92.0	96.0	89.0	87.0
DThist	52.0	22.0	19.0	48.0	9.0	13.0	6.0	20.0	51.0	34.0	32.0	41.0	42.0	49.0	46.0	18.0	27.0	10.0	46.0	41.0	34.0	34.0
CBrûlée	53.5	33.3	32.3	66.7	14.1	24.2	23.2	28.3	79.8	62.6	54.5	30.3	19.2	24.2	58.6	67.7	45.5	36.4	19.2	45.5	41.4	39.4
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	17.0	11.0	68.0	92.0	63.0	94.0	11.0	98.0	75.0	66.0
CCake	77.5	62.5	55.0	71.2	35.0	83.8	18.8	56.2	88.8	51.2	46.2	81.2	85.0	93.8	66.2	11.2	40.0	10.0	88.8	12.5	17.5	20.0
CPitch	23.0	5.0	6.0	14.0	0.0	0.0	0.0	1.0	51.0	14.0	15.0	22.0	18.0	26.0	30.0	2.0	6.0	3.0	22.0	18.0	10.0	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	62.0	70.0	84.0	60.0	62.0	39.0	84.0	84.0	93.0	80.0	78.0
Donuts	81.0	79.0	80.0	86.0	84.0	88.0	74.0	77.0	97.0	86.0	86.0	76.0	70.0	71.0	82.0	77.0	82.0	68.0	71.0	82.0	82.0	82.0
Door	54.0	26.0	26.0	56.0	60.0	77.0	29.0	25.0	67.0	49.0	52.0	30.0	29.0	39.0	43.0	65.0	60.0	33.0	32.0	40.0	34.0	32.0
WDisp	52.5	36.4	35.4	44.4	21.2	35.4	24.2	38.4	61.6	28.3	30.3	58.6	51.5	54.5	59.6	69.7	60.6	62.6	55.6	49.5	37.4	36.4
EMicro	35.0	24.0	23.0	26.0	2.0	1.0	22.0	23.0	44.0	25.0	19.0	47.0	39.0	45.0	48.0	21.0	17.0	24.0	38.0	45.0	34.0	28.0
Franci	26.0	9.0	3.0	18.0	0.0	0.0	0.0	4.0	81.0	15.0	16.0	28.0	10.0	13.0	30.0	12.0	18.0	5.0	11.0	44.0	24.0	24.0
FieldRd	26.0	17.7	15.6	33.3	24.0	25.0	21.9	18.8	50.0	22.9	25.0	29.2	26.0	41.7	33.3	24.0	29.2	20.8	33.3	37.5	19.8	20.8
ForPth	24.0	5.0	3.0	38.0	3.0	11.0	3.0	4.0	29.0	19.0	18.0	13.0	11.0	18.0	21.0	9.0	18.0	4.0	12.0	20.0	15.0	15.0
MLCact	13.0	0.0	0.0	15.0	0.0	0.0	0.0	0.0	9.0	1.0	0.0	32.0	20.0	17.0	44.0	3.0	5.0	2.0	14.0	2.0	9.0	9.0
FireEx	16.0	2.8	0.9	13.2	1.9	2.8	0.9	0.0	5.7	1.9	1.9	5.7	3.8	4.7	10.4	73.6	26.4	41.5	3.8	0.9	1.9	1.9
FireW	50.0	30.0	29.0	45.0	11.0	10.0	28.0	28.0	32.0	26.0	24.0	58.0	62.0	69.0	47.0	16.0	22.0	19.0	68.0	45.0	35.0	35.0
Fries	38.0	30.0	39.0	37.0	3.0	1.0	22.0	6.0	99.0	70.0	76.0	67.0	53.0	56.0	57.0	82.0	55.0	73.0	53.0	96.0	82.0	73.0
GlMilk	83.1	64.0	55.1	86.5	82.0	89.9	46.1	68.5	77.5	86.5	86.5	61.8	56.2	46.1	73.0	79.8	82.0	49.4	47.2	65.2	77.5	75.3
Gramo	7.1	1.8	0.0	16.1	0.0	0.0	1.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.1	87.5	67.9	60.7	0.0	0.0	0.0	0.0
BSGrunt	49.0	20.8	12.5	51.0	2.1	4.2	2.1	10.4	56.2	17.7	20.8	31.2	26.0	29.2	42.7	3.1	7.3	1.0	20.8	24.0	18.8	15.6
HHeels	46.5	35.4	32.3	64.6	79.8	87.9	40.4	70.7	83.8	78.8	48.5	41.4	35.4	45.5	64.6	48.5	54.5	17.2	41.4	63.6	45.5	44.4
HinTp	88.2	84.3	82.4	82.4	64.7	56.9	76.5	80.4	98.0	88.2	90.2	68.6	54.9	41.2	80.4	86.3	86.3	68.6	39.2	92.2	86.3	88.2
HHClam	93.5	93.5	90.3	93.5	61.3	67.7	77.4	93.5	96.8	96.8	93.5	77.4	54.8	51.6	74.2	58.1	67.7	41.9	45.2	71.0	71.0	71.0
SilverHB	12.1	2.0	2.0	8.1	3.0	2.0	4.0	2.0	6.1	3.0	3.0	11.1	10.1	16.2	15.2	2.0	3.0	2.0	11.1	6.1	4.0	4.0
SwPea	10.0	1.0	1.0	12.0	0.0	1.0	1.0	0.0	33.0	2.0	2.0	19.0	14.0	19.0	29.0	0.0	1.0	0.0	9.0	1.0	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	36.0	13.0	7.0	53.0	81.0	57.0	30.0	7.0	44.0	35.0	18.0
ELFBug	37.0	26.0	49.0	38.0	3.0	3.0	44.0	71.0	98.0	67.0	83.0	35.0	20.0	17.0	66.0	45.0	29.0	22.0	16.0	67.0	41.0	36.0
Mbira	28.4	19.4	20.9	25.4	23.9	16.4	32.8	20.9	17.9	19.4	17.9	40.3	40.3	71.6	29.9	10.4	9.0	11.9	61.2	6.0	10.4	9.0
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	51.5	30.3	87.9	51.5	78.8	6.1	33.3	81.8	84.8	84.8
C2SOct	40.0	49.0	58.0	36.0	22.0	17.0	61.0	62.0	87.0	54.0	52.0	40.0	48.0	58.0	45.0	25.0	13.0	34.0	51.0	59.0	34.0	32.0
RubyOct	42.0	48.0	48.0	39.0	22.0	14.0	55.0	48.0	88.0	54.0	54.0	30.0	34.0	44.0	32.0	25.0	13.0	32.0	40.0	50.0	28.0	28.0
PDeer	81.7	61.0	59.8	89.0	46.3	58.5	52.4	85.4	91.5	80.5	80.5	64.6	34.1	31.7	93.9	84.1	73.2	75.6	29.3	95.1	81.7	80.5
DFlath	58.0	33.0	32.0	52.0	3.0	18.0	13.0	32.0	62.0	31.0	33.0	62.0	58.0	53.0	68.0	37.0	39.0	26.0	47.0	64.0	52.0	49.0
EPWasp	73.0	55.0	56.0	64.0	16.0	29.0	29.0	63.0	90.0	75.0	72.0	39.0	27.0	24.0	73.0	63.0	43.0	45.0	24.0	76.0	65.0	59.0
FalseKW	80.6	74.6	74.6	74.6	55.2	55.2	59.7	86.6	98.5	91.0	89.6	73.1	53.7	53.7	74.6	85.1	77.6	86.6	56.7	97.0	85.1	83.6
Pyra	11.0	5.0	6.0	12.0	5.0	6.0	7.0	6.0	11.0	7.0	5.0	21.0	14.0	19.0	26.0	30.0	9.0	12.0	15.0	13.0	11.0	10.0
Sky	22.1	23.5	25.0	22.1	27.9	25.0	38.2	29.4	44.1	14.7	16.2	20.6	25.0	64.7	17.6	17.6	23.5	25.0	54.4	25.0	13.2	13.2
Dreamf	60.0	44.0	45.0	63.0	26.0	31.0	30.0	42.0	69.0	48.0	50.0	64.0	63.0	67.0	65.0	15.0	29.0	11.0	59.0	65.0	56.0	53.0
YTrump	14.0	1.0	0.0	4.0	0.0	0.0	0.0	0.0	54.0	10.0	14.0	14.0	7.0	5.0	23.0	3.0	1.0	0.0	1.0	3.0	0.0	0.0
Sciss	29.0	9.0	10.0	42.0	9.0	12.0	11.0	11.0	19.0	22.0	26.0	27.0	24.0	24.0	44.0	67.0	64.0	31.0	22.0	16.0	31.0	26.0
GCuttle	30.3	10.1	8.1	33.3	3.0	4.0	15.2	10.1	37.4	14.1	14.1	35.4	30.3	35.4	47.5	42.4	11.1	62.6	35.4	42.4	18.2	15.2
CCuttle	34.0	23.0	24.0	24.0	9.0	8.0	27.0	25.0	44.0	22.0	20.0	22.0	22.0	30.0	34.0	39.0	8.0	57.0	27.0	36.0	16.0	15.0
SCalam	21.2	11.1	12.1	21.2	4.0	4.0	13.1	8.1	40.4	15.2	15.2	29.3	25.3	27.3	35.4	10.1	5.1	11.1	22.2	34.3	18.2	18.2
ShCo	58.2	4.5	4.5	43.3	7.5	1.5	37.3	7.5	3.0	1.5	1.5	13.4	7.5	6.0	64.2	83.6	17.9	76.1	6.0	10.4	10.4	10.4
SCaterp	11.0	13.0	19.0	11.0	3.0	3.0	14.0	15.0	81.0	28.0	29.0	31.0	21.0	27.0	29.0	5.0	8.0	4.0	21.0	22.0	13.0	12.0
SBolo	67.2	47.8	49.3	83.6	68.7	79.1	50.7	22.4	100.0	100.0	100.0	71.6	76.1	82.1	82.1	74.6	61.2	73.1	82.1	100.0	94.0	91.0
Stapl	34.0	14.0	12.0	31.0	11.0	21.0	13.0	18.0	20.0	17.0	19.0	27.0	23.0	24.0	30.0	72.0	59.0	38.0	22.0	22.0	27.0	26.0
Rosyb	65.0	28.0	17.0	45.0	0.0	2.0	2.0	11.0	86.0	36.0	37.0	75.0	72.0	70.0	72.0	23.0	40.0	23.0	65.0	71.0	53.0	53.0
CATapir	13.0	10.0	11.0	12.0	4.0	3.0	15.0	15.0	23.0	13.0	13.0	29.0	32.0	39.0	48.0	51.0	10.0	74.0	36.0	58.0	24.0	24.0
MNewt	90.0	94.0	95.0	91.0	82.0	83.0	80.0	96.0	99.0	98.0	98.0	90.0	93.0	95.0	93.0	80.0	74.0	86.0	95.0	99.0	92.0	91.0
IPBNDol	57.0	47.0	45.0	56.0	27.0	29.0	47.0	64.0	90.0	64.0	69.0	47.0	36.0	39.0	64.0	84.0	59.0	92.0	41.0	94.0	72.0	69.0
'ō'ai	63.0	28.0	23.0	51.0	2.0	5.0	3.0	18.0	71.0	33.0	32.0	54.0	47.0	35.0	50.0	7.0	14.0	2.0	25.0	30.0	19.0	19.0
Waffle	57.4	54.1	59.0	55.7	59.0	49.2	60.7	52.5	83.6	52.5	57.4	70.5	72.1	80.3	55.7	59.0	62.3	59.0	75.4	54.1	62.3	62.3
Walker	77.8	52.5	46.5	53.5	42.4	57.6	44.4	35.4	56.6	56.6	52.5	34.3	17.2	15.2	35.4	32.3	22.2	18.2	14.1	12.1	12.1	11.1
WiChair	97.2	40.8	25.4	90.1	15.5	32.4	12.7	43.7	31.0	32.4	25.4	81.7	49.3	25.4	94.4	97.2	95.8	84.5	28.2	95.8	91.5	88.7
mean	51.9	37.8	36.9	50.3	27.5	31.2	32.6	38.5	62.7	46.0	45.0	47.2	41.1	43.3	54.9	49.6	42.4	41.5	40.5	51.8	44.2	42.6

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. (2022a) 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., 2022a) 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).

360OpenSetClasses (Bendale & Boult, 2016) uses those 360 classes (15.000 samples) from ILSVRC2010 which are not part of ILSVRC2012. Like for ImageNet-O, this leads to large semantic overlap, e.g. the class *organ pipe* coinciding with the ID class *organ/pipe organ*.

Semantic Shift Benchmark (SSB) (Vaze et al., 2022) contains a *hard* and *easy* benchmark, each consisting of 1000 classes, that were created by regarding the distances between nodes in the WordNet tree. Similar to 360OpenSetClasses, we find both categorical and incidental ID contamination, e.g. *rainbow lorikeet*/*lorikeet*. Papers evaluating on SSB include (Wen et al., 2022).

ImageNet-1K-OOD (Wang et al., 2022b) contains 50.000 images from 1.000 classes randomly sampled from ImageNet-21K, such that those classes don't overlap with ImageNet-1K and ImageNet-LT, another dataset introduced by the authors. Categorical examples include *bobwhite quail*/*quail* and *king vulture*/*vulture*.

COOD-benchmark (Galil et al., 2023) is a general framework for benchmarking ImageNet-1K OOD detection. Their test set consists of ImageNet-21K samples which were filtered by class. It includes severe contamination, including categorical cases like *orange, orange tree*/*orange* and *cup*/*cup* (with different ids).

C MODELS

In Table 8 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 8: 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 $S_{\text{MSP}}(x) = \max_{k=1,...,K} p_k(x)$ 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 64 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. (2022a) 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. Methods accessing $h(\boldsymbol{x})$ directly in order to compute the OOD-score are referred to as feature-based methods, in contrast to methods that derive their OOD-score from the logits $\boldsymbol{o}$ (even though the logits clearly implicitly also depend on the features). 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:

$$s = \log \sum_c^C \exp{(o_c)}$$

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., 2022a; 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 K^{th} 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., 2022a), we set the threshold r such that 1% of the activations from the train set would be truncated.

Virtual Logit Matching (Wang et al., 2022a): 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., 2022a), 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; Galil et al., 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 4. 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

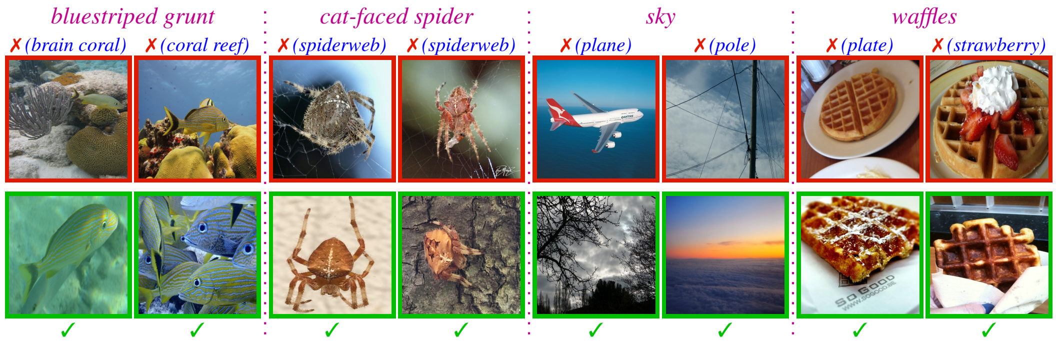

Figure 8: **Cleaning the OOD classes. Top:** Samples that were excluded due to overlap with ID classes. **Bottom:** Samples from the same OOD class that were included in the cleaned datasets.

G DETAILS OF THE NINCO DATASET.

Table 9: 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*
bagpipe	*Bagp*	97	Imagenet-21k	*bagpipe*
bluestriped grunt	*BSGrunt*	96	SPECIES	*grunt*
cable	*Cabl*	88	scraped	*cable television*
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	-
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*
door	*Door*	100	MyNursingHome	*interior door*
dreamfish	*Dreamf*	100	SPECIES	*sea bream*
dune thistle	*DThist*	100	SPECIES	*creme brulee*
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*
fire extinguisher	*FireEx*	106	MyNursingHome	*fire extinguisher*
fireworks	*FireW*	100	scraped	-
forest path	*ForPth*	100	PLACES	*unclear/very broad class*
Franciscan wallflower	*Franci*	100	SPECIES	*wallflower*
French fries	*Fries*	100	FOOD-101	*french fries*
Galápagos fur seal	*GFurS*	91	SPECIES	*arcella*
giant cuttlefish	*GCuttle*	99	SPECIES	*cuttlefish*
glass of milk	*GlMilk*	89	scraped	*milk*
gramophone	*Gramo*	56	scraped	*gramophone*
high heels	*HHeels*	99	scraped	-
Hindu temple	*HinTp*	51	scraped	*unclear/very broad class*
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*
mbira	*Mbira*	67	scraped	-
Mexican lime cactus	*MLCact*	100	SPECIES	*barrel cactus*
Pampas deer	*PDeer*	82	SPECIES	*buck*
pyramid	*Pyra*	100	caltech-101	*Cheops*
redbreast sunfish	*RBSunf*	100	SPECIES	*sunfish*
rosybells (flowering plant)	*Rosyb*	100	SPECIES	-
ruby octopus	*RubyOct*	100	SPECIES	*octopus*
scissors	*Sciss*	100	caltech-101	*scissors*
shuttlecock	*ShCo*	67	scraped	*shuttlecock*
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*
stapler	*Stapl*	100	caltech-101	*stapler*
sweet pea	*SwPea*	100	SPECIES	*unclear/very broad class*
two-toed amphiuma (salamander)	*2TAmph*	176	SPECIES	*amphiuma*
waffles	*Waffle*	61	FOOD-101	-
walker	*Walker*	99	MyNursingHome	*walker*
water dispenser (jugless)	*WDisp*	100	MyNursingHome	*water cooler*
Windsor chair	*WiChair*	71	caltech-101	*Windsor chair*
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 12.

I Examples images of each OOD class in NINCO and OOD unit-tests

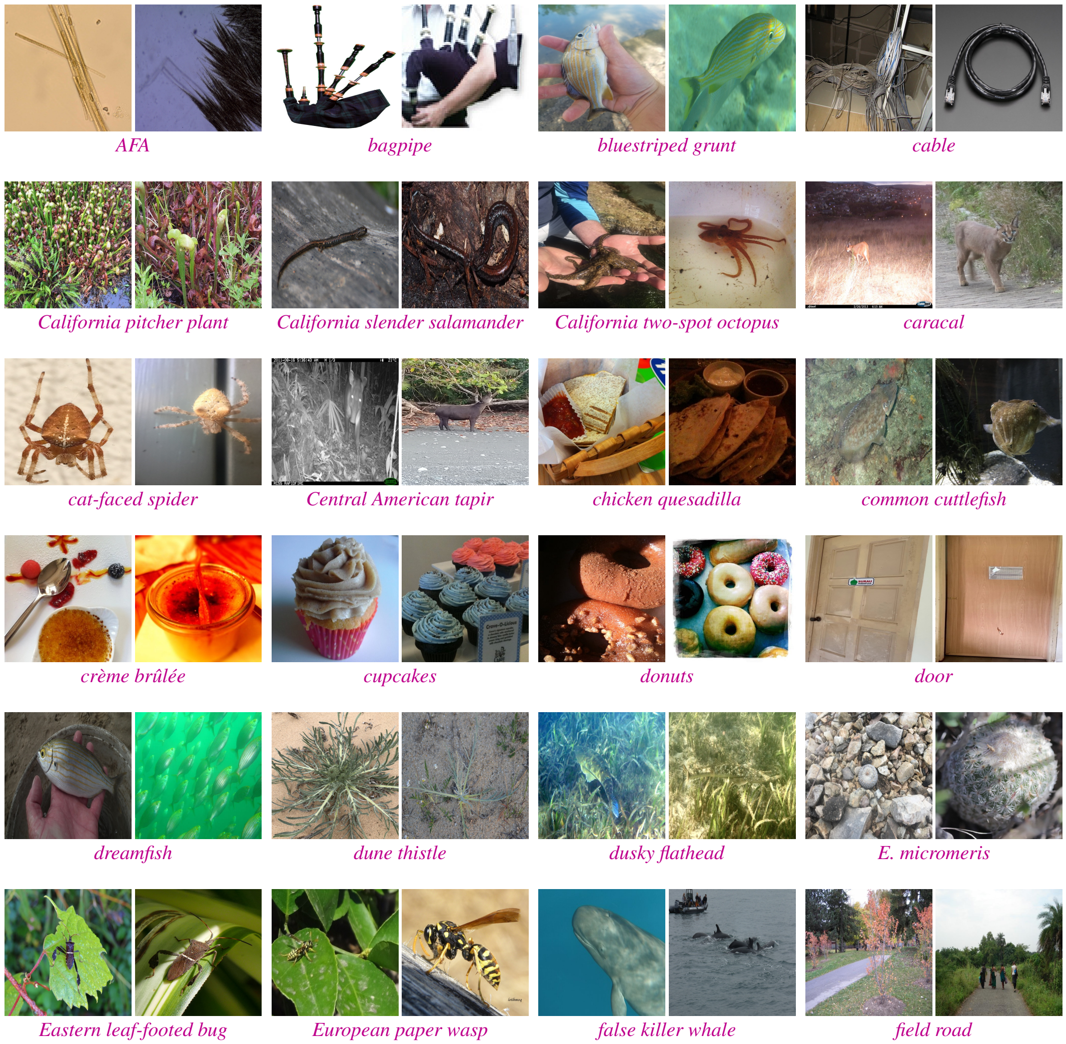

Figure 9: Samples of each class of the NINCO dataset (1/3).

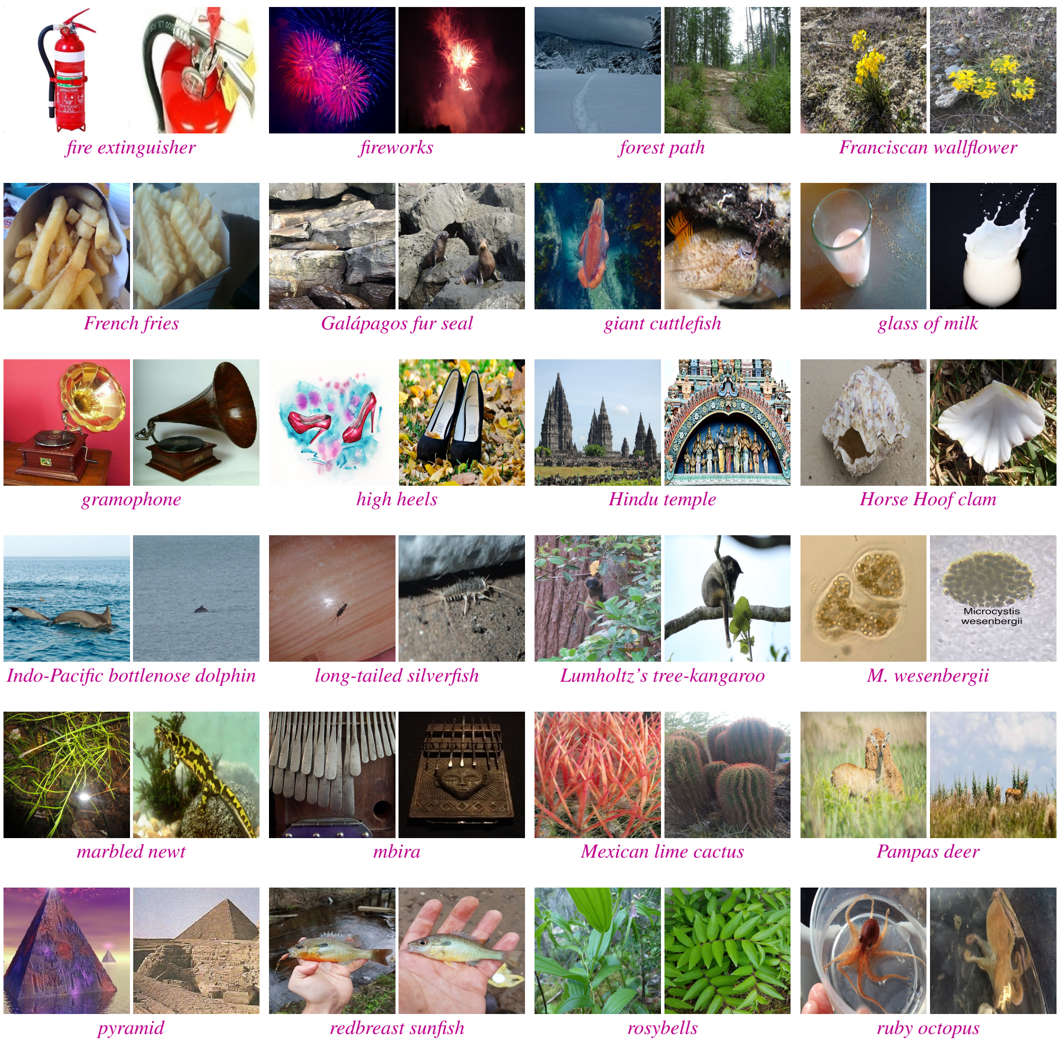

Figure 10: Samples of each class of the NINCO dataset (2/3).

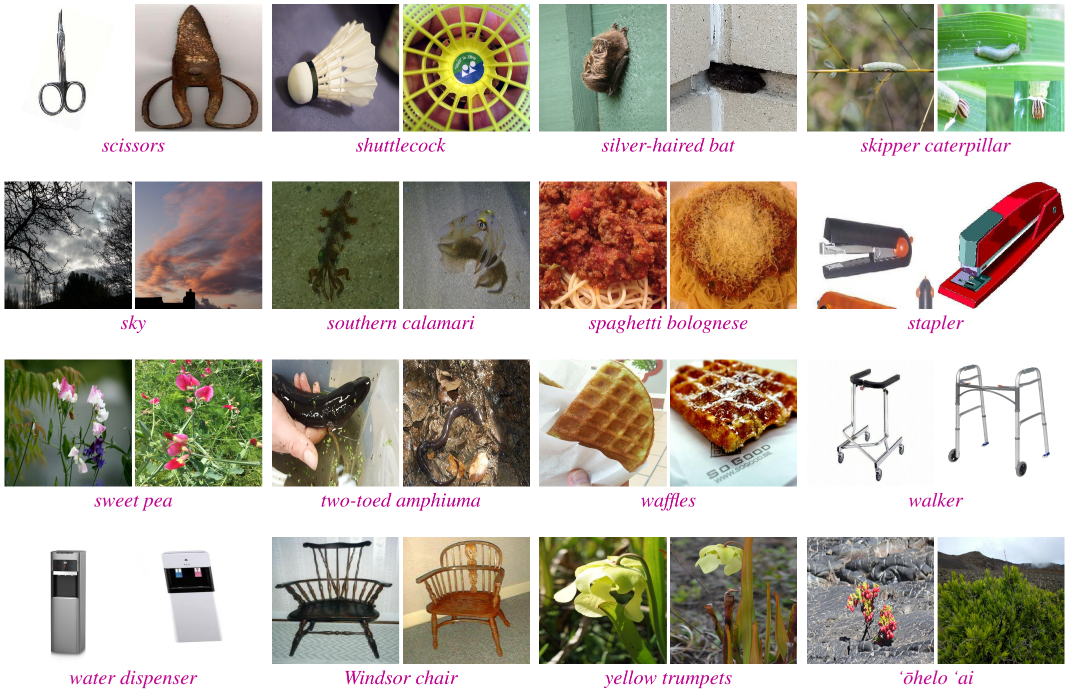

Figure 11: Samples of each class of the NINCO dataset (3/3).

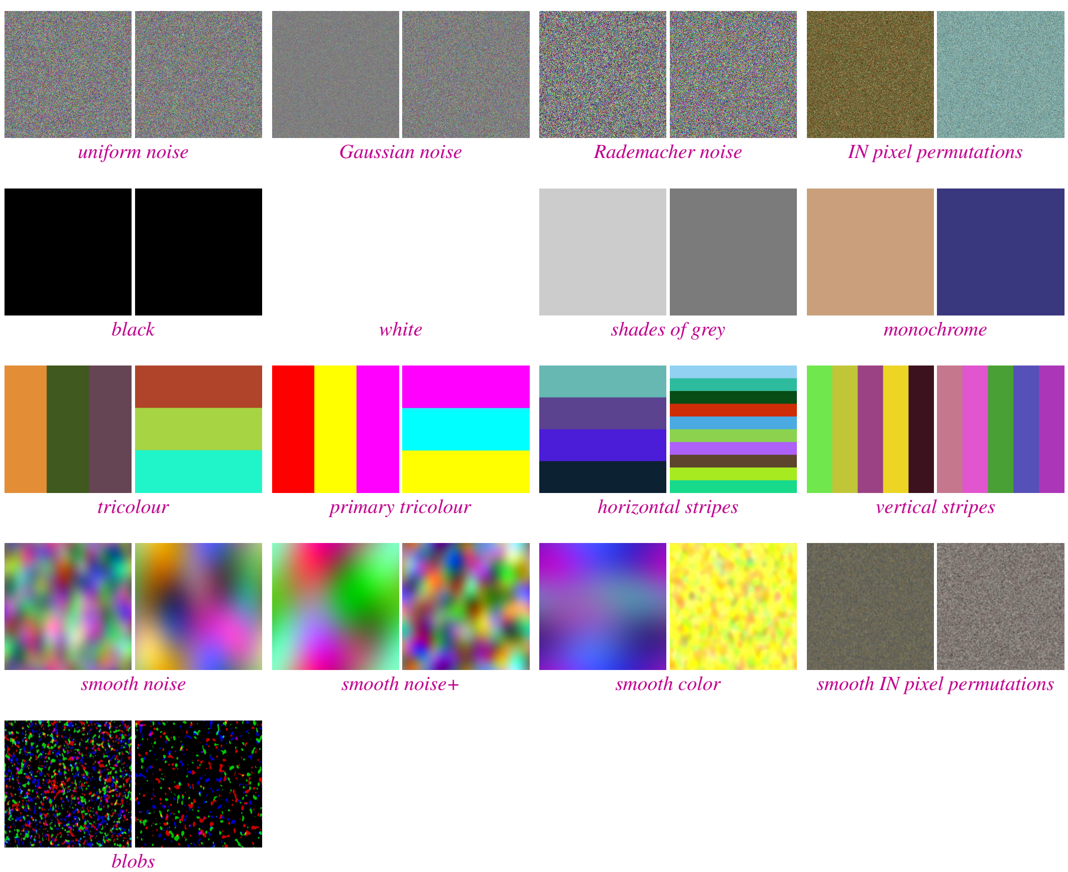

Figure 12: Samples of each OOD unit-test.

J Tables - Unit tests

Table 10: 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-based 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-384-21k	86.0	MSP	0	4.2	0.5	0.0	0.0	4.2	0.0	0.0	0.0	0.0	0.0	1.5	0.2	0.0	0.5	0.0	1.0	3.0	0.8
		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
		ViM	2	100.0	0.0	0.0	0.0	0.0	100.0	0.0	46.0	0.5	0.0	5.0	0.0	0.0	0.0	0.0	0.5	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
		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
		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	3.8	0.0	0.0	0.0	0.0	0.0	0.5	1.2	0.0	0.0	0.0	0.0	4.8	0.0
		knn	0	4.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.2	1.8	4.8	0.0	0.0
		Relative Mahalanobis	0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0
		MCM/RCos	0	3.5	0.0	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
		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
ViT-B-224-21k	84.5	MSP	0	7.0	0.5	4.5	0.0	6.5	0.0	0.0	0.0	0.0	0.2	0.8	1.5	7.0	0.5	0.2	2.5	2.2	1.0
		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
		ViM	3	100.0	0.0	0.0	0.0	0.0	100.0	100.0	55.5	0.0	0.0	1.5	0.0	0.0	0.0	0.0	0.8	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
		Energy+React	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	0.0	0.0	0.2	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	16.2	0.5	3.2	0.0	7.0	0.0	0.0	0.0	0.0	0.2	0.2	12.0	16.2	0.5	0.2	1.8	3.5	0.0
		knn	0	6.0	0.0	1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.8	6.0	0.2	0.0
		Relative Mahalanobis	0	9.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	2.2	5.2	9.8	0.0	0.0	0.0	0.0	0.0
		MCM/RCos	1	16.2	0.5	16.2	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	0.0	0.0
		cosine	0	1.8	0.0	1.8	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
Swinv2-B-256-21k	86.3	MSP	4	28.0	0.0	0.0	0.0	7.2	0.0	0.0	0.0	0.0	2.5	3.0	1.2	12.0	22.5	28.0	5.2	11.0	0.0
		MaxLogit	2	17.2	0.0	0.0	0.0	4.5	0.0	0.0	0.0	0.0	1.0	2.2	0.2	6.8	13.5	17.2	2.8	7.8	0.0
		ViM	8	100.0	0.2	20.2	12.0	0.5	100.0	100.0	100.0	100.0	31.0	22.8	6.8	2.0	0.0	0.0	0.8	2.2	0.0
		Mahalanobis	13	100.0	19.5	48.8	36.5	4.5	100.0	100.0	100.0	100.0	93.0	91.8	45.8	27.5	1.0	0.8	2.8	19.0	13.2
		Energy+React	0	8.2	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.8	2.8	0.0	5.0	5.2	8.2	0.8	7.0	0.0
		Energy	2	14.2	0.0	0.0	0.0	7.0	0.0	0.0	0.0	0.0	0.2	2.5	0.0	7.5	11.8	14.2	2.2	9.0	0.0
		KL-Matching	5	23.8	0.2	0.0	0.0	18.8	0.0	0.0	0.0	0.0	7.2	6.5	4.8	12.8	19.8	23.8	4.0	15.5	0.8
		knn	0	3.5	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	3.5	1.8	1.8	0.0	0.0	0.0	0.0	0.2	0.0
		Relative Mahalanobis	9	100.0	0.5	5.5	8.8	7.0	100.0	100.0	100.0	100.0	77.5	77.2	44.0	16.2	0.0	0.0	0.5	23.8	0.2
		MCM/RCos	0	4.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	1.5	1.2	4.0	0.0	0.0	0.0	0.0	0.2	0.0
		cosine	0	3.8	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	1.5	1.0	3.8	0.0	0.0	0.0	0.0	0.2	0.0
Deit3-B-384-21k	86.7	MSP	8	72.8	58.0	16.5	72.8	20.8	0.0	0.0	0.0	0.0	29.8	13.5	33.5	2.5	1.0	0.5	9.5	50.2	3.2
		MaxLogit	7	75.5	62.5	17.8	75.5	13.5	0.0	0.0	0.0	0.0	17.2	6.0	12.0	0.0	0.5	0.5	8.0	39.2	2.0
		ViM	7	100.0	0.0	0.0	0.0	0.5	0.0	100.0	27.0	36.2	56.0	73.8	30.0	5.8	0.2	0.0	2.2	27.5	0.0
		Mahalanobis	6	100.0	0.0	0.0	0.0	0.8	0.0	100.0	15.5	5.5	39.5	50.2	42.0	2.5	0.0	0.0	0.8	29.8	0.0
		Energy+React	3	60.5	35.8	8.2	60.5	3.8	0.0	0.0	0.0	0.0	8.2	3.5	6.8	0.0	0.0	0.0	6.8	24.8	0.8
		Energy	5	84.2	73.5	24.2	84.2	11.8	0.0	0.0	0.0	0.0	8.5	4.0	6.8	0.0	0.2	0.0	8.2	31.2	2.0
		KL-Matching	8	62.0	47.2	12.0	62.0	17.8	0.0	0.0	0.0	0.0	32.5	15.0	39.5	3.8	1.2	0.8	10.0	48.5	2.8
		knn	2	20.2	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	14.2	3.0	20.2	0.2	0.0	0.0	0.5	1.5	0.0
		Relative Mahalanobis	4	39.5	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.5	32.2	21.8	39.5	2.0	0.0	0.0	0.8	35.2	0.0
		MCM/RCos	3	19.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	14.0	2.2	19.0	0.2	0.0	0.0	0.8	14.2	0.0
		cosine	3	20.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	14.2	3.0	20.0	0.2	0.0	0.0	0.8	13.5	0.0
Deit3-B-224-21k	85.7	MSP	5	48.5	7.0	2.5	5.5	13.2	0.0	0.0	0.0	0.0	29.5	13.0	31.0	2.0	2.0	0.5	9.2	48.5	3.5
		MaxLogit	2	39.0	6.5	2.5	4.5	9.8	0.0	0.0	0.0	0.0	11.5	5.8	9.8	0.2	0.8	0.0	6.0	39.0	1.2
		ViM	6	100.0	0.5	0.2	0.0	0.2	0.0	100.0	53.8	9.0	61.5	82.2	39.0	1.5	0.5	0.2	3.5	15.2	0.0
		Mahalanobis	6	100.0	0.5	0.0	0.0	0.2	0.0	100.0	61.3	8.2	64.0	73.8	43.5	2.0	0.0	0.0	1.5	16.8	0.0
		Energy+React	1	27.0	2.2	0.8	2.0	5.5	0.0	0.0	0.0	0.0	5.8	1.5	5.0	0.2	0.0	0.0	2.5	27.0	0.0
		Energy	1	36.2	8.8	3.0	6.5	8.5	0.0	0.0	0.0	0.0	6.2	1.8	5.5	0.2	0.5	0.0	5.2	36.2	0.8
		KL-Matching	5	47.0	6.8	2.5	4.5	13.8	0.0	0.0	0.0	0.2	38.5	17.8	38.0	6.8	3.5	1.8	9.5	47.0	3.8
		knn	2	18.8	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	13.8	3.5	18.8	0.0	0.0	0.0	1.0	3.8	0.0
		Relative Mahalanobis	4	51.5	0.5	0.0	0.0	0.2	0.0	0.0	0.0	2.2	51.5	40.2	40.0	1.2	0.0	0.0	1.8	18.5	0.0
		MCM/RCos	2	19.2	0.2	0.0	0.0	0.2	0.0	0.0	0.0	0.0	12.5	4.0	19.2	0.0	0.0	0.0	0.8	10.0	0.0
		cosine	3	17.2	0.2	0.0	0.0	0.2	0.0	0.0	0.0	0.0	11.0	4.2	17.2	0.0	0.0	0.0	0.8	11.0	0.0
ViT-B-384-l2b-12k	87.2	MSP	0	7.5	0.0	0.0	0.0	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.5	0.5	7.5	0.2
		MaxLogit	0	4.5	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.2	0.0	4.5	0.0
		ViM	5	100.0	0.2	0.0	18.0	0.5	100.0	100.0	97.0	14.8	0.8	2.2	0.0	1.2	0.5	0.0	1.2	1.5	0.0
		Mahalanobis	16	100.0	83.0	43.8	99.8	4.2	100.0	100.0	100.0	100.0	99.8	99.5	37.5	80.2	18.0	14.8	19.2	38.8	38.8
		Energy+React	0	3.8	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.2	3.8	0.0
		Energy	0	5.2	0.0	0.0	0.0	2.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	0.2	5.2	0.0
		KL-Matching	0	8.5	0.0	0.0	0.0	5.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.8	1.8	2.8	8.5	3.2
		knn	0	2.2	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	0.5	2.2	0.0
		Relative Mahalanobis	10	100.0	3.2	4.5	44.2	5.8	100.0	100.0	100.0	53.5	30.5	56.2	2.5	16.5	10.0	8.5	10.0	12.2	24.5
		MCM/RCos	0	1.2	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.2	0.0	0.8	1.2	0.0
		cosine	0	1.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	0.0
ViT-B-384-oai-12k	87.0	MSP	1	12.8	0.0	0.0	0.0	8.8	0.0	0.0	0.0	0.0	0.2	1.0	0.2	0.0	2.0	3.8	1.5	12.8	0.5
		MaxLogit	1	11.0	0.0	0.0	0.0	5.5	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	4.8	5.0	1.5	11.0	0.0
		ViM	2	100.0	0.0	0.0	0.0	0.0	0.0	100.0	38.2	0.2	0.0	0.2	0.0	0.0	0.5	0.0	0.5	0.0	0.0
		Mahalanobis	7	100.0	5.0	6.5	46.0	0.2	100.0	100.0	100.0	62.3	32.2	72.2	7.0	0.0	9.2	5.2	7.2	1.8	0.8
		Energy+React	3	15.0	0.0	2.0	0.0	4.5	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	15.0	10.8	3.8	11.0	0.2
		Energy	3	14.5	0.0	1.8	0.0	6.5	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	14.5	10.2	3.8	11.8	0.2
		KL-Matching	0	9.8	0.0	0.0	0.0	9.8	0.0	0.0	0.0	0.0	0.5	2.8	1.0	0.2	2.2	4.5	5.5	7.5	2.2
		knn	0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.8	0.0	0.0	0.8	0.2	0.5	0.0	0.0
		Relative Mahalanobis	6	100.0	0.0	0.0	2.8	1.0	0.0	100.0	87.2	16.5	15.5	28.0	3.0	0.2	10.8	6.0	8.0	1.2	0.2
		MCM/RCos	0	5.5	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.2	1.5	0.0	0.0	5.5	3.5	3.0	0.0	0.0
		cosine	0	4.5	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.2	1.0	0.0	0.0	4.5	1.5	2.5	0.0	0.0
ViT-B-384-l2b	86.6	MSP	1	18.0	0.0	0.0	0.0	18.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.8	0.5	0.2	7.2	0.0
		MaxLogit	1	11.5	0.0	0.0	0.0	11.5	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.5	0.2	0.2	3.2	0.0
		ViM	1	25.5	0.0	0.0	0.0	0.2	0.0	0.0	25.5	0.5	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.2	0.0
		Mahalanobis	11	100.0	99.5	62.7	100.0	3.0	100.0	100.0	100.0	100.0	15.8	41.2	1.0	0.2	0.0	0.2	0.2	15.8	18.2
		Energy+React	0	9.8	0.0	0.0	0.0	9.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	2.8	0.0
		Energy	1	12.0	0.0	0.0	0.0	12.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.5	3.0	0.0
		KL-Matching	1	12.8	0.0	0.0	0.0	12.8	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.8	0.2	0.2	8.5	0.0
		knn	0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.5	0.0
		Relative Mahalanobis	8	100.0	75.2	33.8	95.2	8.8	100.0	100.0	100.0	42.2	0.8	10.0	0.0	0.0	0.0	0.0	1.8	11.5	2.2
		MCM/RCos	0	0.2	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	0.0	0.0
		cosine	0	1.0	0.0	0.0	0.0	1.0	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
ViT-B-384-oai	86.2	MSP	1	17.5	0.0	0.0	0.0	17.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5	9.0	0.0
		MaxLogit	1	13.5	0.0	0.0	0.0	13.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	5.5	0.0
		ViM	2	100.0	0.0	0.0	0.0	2.0	0.0	100.0	30.0	0.5	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.0
		Mahalanobis	11	100.0	15.5	6.0	47.5	1.5	100.0	100.0	100.0	100.0	95.0	90.8	26.5	26.8	0.0	0.2	1.8	8.5	21.2
		Energy+React	1	12.2	0.0	0.2	0.0	12.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.5	0.0
		Energy	1	13.5	0.0	0.0	0.0	13.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0
		KL-Matching	1	16.8	0.0	0.0	0.0	16.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.8	9.5	0.0
		knn	0	4.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	1.2	0.8	1.2	0.0	0.0	0.0	0.0	3.8	0.0
		Relative Mahalanobis	10	100.0	0.2	1.0	34.2	2.0	100.0	100.0	100.0	89.0	69.0	64.2	14.5	29.5	0.0	0.2	4.2	7.8	25.5
		MCM/RCos	0	0.8	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
		cosine	0	1.5	0.0	0.0	0.0	1.5	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
clip-ViT-B-224	66.6	mcm-clip	1	17.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	4.0	4.2	0.0	17.0	0.0	0.0	0.0	0.0	0.0
		cosine-clip	15	100.0	100.0	99.5	100.0	98.0	0.0	100.0	1.5	53.2	14.2	31.5	23.5	72.5	83.8	87.8	60.2	72.2	38.2
clip-ViT-L-336	74.3	mcm-clip	0	5.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0	1.2	5.0	0.5	0.0	1.0	0.0	0.0	0.0	1.2	3.8
		cosine-clip	14	100.0	100.0	99.8	100.0	98.2	0.0	100.0	48.2	88.8	5.0	33.2	3.2	31.8	94.2	88.5	83.2	79.0	73.2

Table 11: 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*
		MSP	1	13.5	0.0	0.0	0.0	0.2	0.0	13.5	2.2	5.5	2.0	5.8	0.0	0.5	2.5	8.2	0.2	0.2	0.0
		MaxLogit	0	6.0	0.0	0.0	0.0	0.0	0.0	4.5	2.5	6.0	1.0	2.8	0.0	0.0	0.2	4.0	0.2	0.2	0.0
		Energy	1	10.8	0.0	0.0	0.0	0.0	0.0	2.5	5.2	10.8	2.2	2.8	0.0	0.0	0.0	3.0	0.0	0.2	0.0
		KL-Matching	2	29.2	0.0	0.0	0.0	2.0	0.0	29.2	2.5	2.0	2.5	6.2	0.0	1.2	4.2	10.8	0.2	0.0	0.0
		Mahalanobis	4	100.0	0.8	42.5	100.0	1.0	100.0	0.0	0.0	0.0	0.0	2.0	2.8	0.0	0.0	5.0	0.2	0.0	100.0
CnvNxt-B-21k	86.3	Relative Mahalanobis	5	100.0	0.5	41.0	100.0	0.0	100.0	1.0	0.0	0.0	0.0	13.0	1.5	0.0	0.2	9.8	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	2.0	0.0	0.0	100.0
		Energy+React	0	4.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0	4.0	0.0	1.0	0.0	0.0	0.0	2.0	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
		MSP	3	16.0	3.2	1.2	0.0	2.2	0.0	7.2	8.5	11.2	5.8	13.2	0.5	16.0	6.0	7.8	0.5	0.8	0.0
		MaxLogit	4	24.8	17.0	16.5	0.0	0.2	0.0	1.0	3.5	5.2	3.8	9.5	0.0	11.0	1.8	1.5	24.8	0.2	0.0
		Energy	3	60.2	29.8	60.2	0.0	0.2	0.0	0.5	0.8	2.5	2.8	7.2	0.0	9.5	0.8	0.2	38.2	0.0	0.0
		KL-Matching	8	27.0	8.5	2.2	0.0	13.2	0.0	22.5	11.0	12.0	6.5	16.5	1.8	27.0	17.8	16.8	0.8	7.5	0.0
		Mahalanobis	6	100.0	4.0	11.5	100.0	0.0	100.0	1.2	0.0	0.0	0.0	0.0	40.5	0.0	9.0	14.0	7.0	0.0	100.0
CnvNxt-T-21k	84.1	Relative Mahalanobis	4	100.0	0.5	7.5	100.0	0.0	100.0	1.5	0.0	0.0	0.2	1.2	17.0	1.0	3.0	9.5	5.0	0.0	100.0
		ViM	4	100.0	1.0	7.0	100.0	0.0	100.0	0.2	0.0	0.0	0.0	0.0	13.0	0.0	1.2	5.2	5.5	0.0	100.0
		Energy+React	3	34.8	22.2	34.8	0.0	0.0	0.0	0.5	0.0	1.2	1.8	5.0	0.0	5.2	0.2	0.0	28.7	0.0	0.0
		knn	2	18.5	0.0	0.0	0.0	0.0	0.0	12.8	0.0	0.0	0.8	0.8	0.0	0.0	4.5	18.5	0.0	0.0	0.0
		cosine	0	8.0	0.0	0.0	0.0	0.0	0.0	3.5	0.0	0.0	0.5	0.2	0.0	0.8	0.2	8.0	0.0	0.0	0.0
		MCM/RCos	1	10.5	0.0	0.0	0.0	0.0	0.0	6.0	0.0	0.0	0.5	0.2	0.0	0.8	1.5	10.5	0.0	0.0	0.0
		MSP	11	93.0	48.2	51.7	0.0	5.0	0.0	9.8	41.5	41.5	51.5	93.0	13.5	81.0	34.8	23.5	48.8	4.2	0.0
		MaxLogit	10	87.8	36.0	40.0	0.0	1.2	0.0	1.8	40.0	42.2	49.5	87.8	6.2	71.5	13.2	11.2	35.2	1.8	0.0
		Energy	10	86.8	33.8	32.0	0.0	1.0	0.0	0.8	65.2	68.0	62.5	86.8	6.0	72.0	11.2	12.8	27.3	5.2	0.0
		KL-Matching	11	90.5	51.5	51.7	0.0	5.0	0.0	8.2	33.5	32.5	39.8	90.5	12.8	72.5	28.7	18.2	50.2	3.8	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
BiT-m	82.3	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	63.2	6.2	4.2	0.0	0.0	0.0	0.8	61.5	63.2	57.8	62.7	4.0	31.8	15.8	14.5	4.8	29.5	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	0	1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.2	1.0	1.5	1.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0
		MSP	4	52.2	0.2	0.0	0.0	1.0	0.0	0.2	10.2	9.0	20.0	52.2	0.5	14.2	0.5	0.0	0.5	0.0	0.0
		MaxLogit	3	48.8	0.5	0.0	0.0	0.8	0.0	0.2	6.2	6.0	14.0	48.8	1.2	14.5	0.0	0.0	0.5	0.0	0.0
		Energy	4	57.0	11.0	0.0	0.0	10.0	0.0	0.2	9.2	8.2	16.5	57.0	3.2	27.0	0.0	0.0	1.5	0.0	0.0
		KL-Matching	5	50.2	0.8	0.0	0.0	2.8	0.0	0.8	18.0	16.2	23.5	50.2	0.2	14.5	1.0	0.0	0.2	1.0	0.0
		Mahalanobis	0	3.2	0.0	0.0	0.0	0.0	0.0	0.5	2.0	1.5	3.2	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.0
EffNetv2-M-21k	85.6	Relative Mahalanobis	2	100.0	0.0	0.0	0.0	0.0	3.8	0.2	5.5	2.8	16.5	5.5	0.2	0.5	0.5	5.8	0.0	0.2	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	99.5	0.0	100.0	1.0	59.2	98.8	99.5	92.5	98.2	27.8	83.2	30.8	25.5	98.5	80.0	0.0
		knn	0	5.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		cosine	0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		MCM/RCos	1	12.5	0.0	0.0	0.0	0.0	0.0	0.2	1.2	0.8	12.5	0.2	0.0	0.2	0.0	3.5	0.0	0.0	0.0
		MSP	0	9.5	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.8	0.2	0.2	0.0	5.2	0.0	0.0	0.0	9.5	0.0
		MaxLogit	0	8.8	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.2	0.0	0.0	0.0	8.8	0.0
		Energy	3	43.0	18.0	43.0	0.0	11.8	0.0	0.0	0.0	0.2	0.5	0.2	0.0	0.5	0.0	0.0	9.5	9.2	0.0
		KL-Matching	2	26.5	0.2	0.0	0.0	1.2	0.0	3.8	0.0	1.2	0.5	5.2	0.0	26.5	0.5	0.2	0.8	11.0	0.0
		Mahalanobis	15	100.0	36.5	1.0	100.0	14.5	100.0	100.0	68.0	64.5	88.2	70.8	100.0	60.8	100.0	98.5	6.8	99.8	100.0
EffNetb7-ns	86.8	Relative Mahalanobis	13	100.0	0.0	0.0	100.0	0.0	100.0	94.2	25.5	24.8	46.2	39.5	100.0	12.5	98.5	80.0	0.0	99.0	100.0
		ViM	16	100.0	37.8	1.5	100.0	25.0	100.0	100.0	73.2	68.5	89.5	72.8	100.0	61.3	100.0	99.0	10.8	99.8	100.0
		Energy+React	0	8.2	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	8.2	0.0
		knn	0	9.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	9.2	0.0	0.0	0.0	0.0	0.0
		cosine	0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	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	0.2	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

Table 12: 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*
		MSP	7	43.8	0.0	0.0	0.0	15.8	0.0	26.2	11.8	11.2	13.2	5.5	0.0	43.8	0.5	10.5	1.0	4.5	0.0
		MaxLogit	2	19.2	0.0	0.0	0.0	4.0	0.0	16.8	5.2	6.2	6.5	1.0	0.0	19.2	0.2	7.8	0.0	2.2	0.0
		Energy	2	17.0	3.0	0.0	0.0	1.2	0.0	17.0	2.8	2.8	5.8	0.5	0.0	12.2	0.8	8.8	0.0	2.8	0.0
		KL-Matching	6	44.8	0.0	0.0	0.0	13.8	6.2	28.0	8.2	8.2	9.0	4.8	3.0	44.8	10.2	12.2	0.2	27.5	0.0
		Mahalanobis	4	80.8	19.0	0.5	0.0	0.0	80.8	22.0	0.8	0.2	0.2	1.2	8.8	22.8	9.8	3.2	8.0	1.0	0.0
ViT-B-384	81.1	Relative Mahalanobis	10	100.0	6.8	0.0	100.0	0.8	100.0	26.8	1.0	1.0	1.8	15.2	49.8	24.0	49.8	17.0	0.2	12.2	100.0
		ViM	5	94.8	74.5	12.2	0.0	0.0	0.0	28.0	4.8	4.0	0.5	2.2	0.0	27.5	0.5	1.8	94.8	0.0	0.0
		Energy+React	2	21.0	1.0	0.0	0.0	1.5	0.0	21.0	2.8	3.2	6.2	0.5	0.2	12.8	0.8	8.8	0.0	3.5	0.0
		knn	5	85.2	82.2	85.2	0.0	3.0	0.0	18.8	4.0	4.0	6.8	0.8	1.0	7.2	0.5	10.5	84.8	0.0	0.0
		cosine	3	23.2	23.2	0.0	0.0	3.0	0.0	14.0	2.8	3.5	3.2	0.8	0.8	14.0	0.2	7.2	1.0	0.0	0.0
		MCM/RCos	4	38.2	38.2	3.2	0.0	3.8	0.0	14.8	3.5	3.5	4.2	0.8	1.0	15.8	0.2	8.5	16.8	0.0	0.0
		MSP	4	27.8	0.0	0.0	0.0	20.5	0.0	27.8	10.0	10.8	3.8	9.0	0.0	25.0	3.2	2.8	0.2	3.5	0.0
		MaxLogit	2	22.8	0.0	0.0	0.0	18.0	0.0	8.2	6.0	5.0	2.0	5.0	0.0	22.8	0.8	1.5	0.0	2.2	0.0
		Energy	2	31.5	0.0	0.0	0.0	20.8	0.0	4.0	2.8	2.8	1.0	3.8	0.0	31.5	0.0	0.5	0.0	2.8	0.0
		KL-Matching	6	49.8	0.0	0.0	0.0	21.0	0.0	49.8	15.0	13.8	6.5	10.5	0.0	24.5	8.5	6.0	0.2	4.2	0.0
		Mahalanobis	12	100.0	51.2	49.5	100.0	5.0	100.0	60.0	7.0	5.0	5.5	18.2	100.0	3.0	97.8	90.8	52.5	35.2	100.0
Swinv2-B-256	84.6	Relative Mahalanobis	12	100.0	22.8	44.2	100.0	1.5	100.0	61.5	5.5	3.8	3.2	13.5	100.0	2.8	97.2	86.5	48.8	34.5	100.0
		ViM	12	100.0	20.8	45.2	100.0	1.0	100.0	53.8	9.0	5.8	8.0	13.0	100.0	4.5	97.5	96.5	48.8	33.8	100.0
		Energy+React	1	10.2	0.0	0.0	0.0	10.2	0.0	0.8	2.0	1.5	1.0	2.5	0.0	8.5	0.0	0.2	0.0	0.2	0.0
		knn	1	32.0	0.0	0.0	0.0	0.0	0.0	32.0	5.0	3.2	1.0	0.0	6.2	1.8	8.0	0.8	0.0	0.0	0.0
		cosine	1	25.5	0.0	0.0	0.0	0.0	0.0	25.5	3.0	1.0	1.0	0.0	0.2	1.2	3.0	0.2	0.0	0.0	0.0
		MCM/RCos	1	33.5	0.0	0.0	0.0	0.0	0.0	33.5	3.0	1.2	0.8	0.5	0.0	1.5	4.2	0.5	0.0	0.0	0.0
		MSP	3	34.8	0.0	0.0	0.0	0.0	0.0	34.8	9.5	16.2	3.8	4.2	0.0	10.8	0.0	0.8	0.0	0.5	0.0
		MaxLogit	1	15.8	0.0	0.0	0.0	0.0	0.0	15.8	3.8	8.5	2.5	1.2	0.0	8.0	0.0	0.5	0.0	0.0	0.0
		Energy	0	7.0	0.0	0.0	0.0	0.0	0.0	4.0	2.8	6.0	3.0	1.5	0.0	7.0	0.0	0.0	0.0	0.0	0.0
		KL-Matching	4	61.3	0.2	0.0	0.0	0.0	0.0	61.3	13.2	19.2	6.5	6.0	0.0	13.2	0.5	4.2	0.5	3.5	0.0
		Mahalanobis	14	100.0	99.2	99.8	100.0	66.5	100.0	98.5	6.2	1.5	32.8	46.8	100.0	6.8	100.0	100.0	99.5	96.2	100.0
Deit3-B-384	85.1	Relative Mahalanobis	14	100.0	78.0	97.0	100.0	23.8	100.0	95.8	2.5	1.0	14.8	38.8	100.0	4.0	100.0	100.0	84.2	85.2	100.0
		ViM	8	100.0	0.0	0.0	100.0	0.0	100.0	82.8	0.2	0.5	1.0	0.0	100.0	0.0	99.2	89.5	0.0	47.8	100.0
		Energy+React	6	57.5	0.0	0.0	0.0	0.0	0.0	35.0	48.5	57.5	25.2	22.2	0.0	54.0	0.0	1.0	0.0	0.5	0.0
		knn	14	100.0	91.8	99.8	100.0	53.2	100.0	96.5	8.5	2.5	31.5	12.8	100.0	2.5	99.8	96.0	98.5	52.5	100.0
		cosine	7	100.0	20.0	0.5	100.0	0.0	66.5	88.2	0.5	0.5	2.8	4.5	96.8	1.5	51.2	21.5	0.2	0.8	0.0
		MCM/RCos	1	10.8	0.0	0.0	0.0	0.0	0.0	10.8	0.0	0.2	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		MSP	4	57.5	0.0	0.2	0.0	0.2	0.0	57.5	20.5	25.0	5.8	5.0	0.0	7.5	10.5	8.0	0.0	0.5	0.0
		MaxLogit	3	51.5	0.0	0.0	0.0	0.0	0.0	51.5	15.2	23.2	4.8	3.8	0.0	4.5	4.0	4.2	0.0	0.2	0.0
		Energy	3	49.0	0.0	0.0	0.0	0.0	0.0	49.0	12.5	22.0	2.2	3.0	0.0	8.8	0.5	0.8	0.0	5.5	0.0
		KL-Matching	7	70.5	0.2	2.2	0.0	4.0	0.0	70.5	25.2	30.0	11.8	6.0	0.0	12.8	34.2	26.0	2.0	7.2	0.0
		Mahalanobis	16	100.0	100.0	100.0	100.0	61.8	100.0	87.8	29.8	26.0	60.8	55.2	100.0	2.8	100.0	100.0	99.8	85.2	100.0
Deit3-B-224	83.8	Relative Mahalanobis	16	100.0	93.2	90.2	100.0	36.5	100.0	84.5	16.0	15.0	29.5	40.0	100.0	1.2	100.0	100.0	85.5	69.5	100.0
		ViM	8	100.0	0.2	0.0	100.0	0.0	100.0	68.5	2.0	5.2	0.5	0.8	97.8	0.2	49.0	63.0	0.0	26.2	100.0
		Energy+React	5	46.0	1.2	0.2	0.0	0.5	0.0	46.0	21.5	27.8	5.2	10.8	0.0	28.2	0.2	1.5	0.2	0.0	0.0
		knn	15	100.0	40.8	35.5	100.0	27.3	100.0	74.8	35.2	31.0	53.2	8.5	100.0	4.0	96.2	91.0	15.5	31.5	100.0
		cosine	7	100.0	8.8	2.0	100.0	0.2	100.0	70.8	4.2	7.5	2.0	1.2	100.0	1.8	49.0	36.8	1.0	0.8	100.0
		MCM/RCos	1	57.2	0.0	0.0	0.0	0.0	0.0	57.2	0.5	2.0	0.5	0.2	0.0	0.2	9.0	4.0	0.0	0.0	0.0

Table 13: 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*
		MSP	3	48.0	0.0	0.0	0.0	5.2	0.0	48.0	1.8	4.0	3.2	2.8	0.0	11.2	12.0	5.2	0.0	0.2	0.0
		MaxLogit	1	40.2	0.0	0.0	0.0	2.2	0.0	40.2	2.0	4.2	1.5	1.0	0.0	6.0	6.5	3.2	0.0	0.0	0.0
		Energy	4	56.0	0.0	0.0	0.0	1.0	0.0	56.0	18.2	32.2	4.0	1.8	0.0	10.5	5.5	6.5	0.0	3.0	0.0
		KL-Matching	4	52.8	0.5	0.0	0.0	8.0	0.0	52.8	6.5	8.2	6.8	9.8	0.0	14.8	37.0	18.2	0.0	1.2	0.0
		Mahalanobis	16	100.0	100.0	99.8	100.0	91.5	100.0	100.0	13.5	11.0	43.0	57.2	100.0	8.5	99.8	98.5	99.8	51.2	100.0
XCiT-M-224	82.6	Relative Mahalanobis	15	100.0	99.5	99.2	100.0	82.5	100.0	100.0	12.5	10.0	40.8	54.0	100.0	8.5	99.8	98.2	99.8	52.0	100.0
		ViM	12	100.0	60.5	98.2	100.0	9.5	100.0	100.0	2.8	1.8	9.0	16.0	100.0	5.5	96.2	87.2	96.5	25.5	100.0
		Energy+React	3	66.8	1.0	0.5	0.0	1.0	0.0	66.8	10.5	17.5	3.8	0.8	0.0	8.5	5.8	7.2	0.0	1.2	0.0
		knn	4	68.8	2.5	44.5	0.0	7.5	0.0	49.5	0.5	0.2	1.0	0.2	68.8	3.2	21.0	7.5	2.5	0.0	0.0
		cosine	1	24.8	0.0	0.0	0.0	5.0	0.0	24.8	0.5	0.0	1.5	0.2	0.0	2.2	8.2	1.2	0.0	0.0	0.0
		MCM/RCos	2	54.2	0.0	0.0	0.0	6.0	0.0	54.2	0.5	0.0	1.8	1.0	0.0	2.5	17.8	7.2	0.0	0.5	0.0
		MSP	9	100.0	21.0	45.0	0.0	1.2	54.2	60.5	6.5	9.0	6.2	18.5	3.2	36.8	10.2	8.5	36.8	0.2	100.0
		MaxLogit	1	10.5	0.0	0.0	0.0	0.0	0.0	10.0	1.2	3.8	3.5	7.5	0.0	10.5	0.2	0.0	0.0	0.0	0.0
		Energy	1	12.5	0.0	0.0	0.0	0.0	0.0	1.8	1.8	5.2	3.8	4.8	0.0	12.5	0.0	0.0	0.0	0.0	0.0
		KL-Matching	12	100.0	41.0	48.0	100.0	4.2	82.5	82.8	5.0	5.2	6.5	22.2	14.0	39.5	36.0	25.2	42.8	1.8	100.0
		Mahalanobis	12	100.0	24.0	95.0	100.0	5.0	100.0	85.2	1.5	1.2	4.8	12.2	100.0	4.5	100.0	97.2	50.2	46.5	100.0
XCiT-M-224-d	84.3	Relative Mahalanobis	12	100.0	11.5	75.8	100.0	1.2	100.0	84.5	2.0	1.5	3.8	13.5	85.2	4.5	98.8	95.0	24.8	45.5	100.0
		ViM	10	100.0	10.0	65.2	100.0	0.0	100.0	77.5	0.0	0.0	0.2	2.8	96.0	1.0	97.0	79.5	27.3	37.2	100.0
		Energy+React	0	9.2	0.2	2.0	0.0	0.0	0.0	2.5	0.0	0.2	2.8	3.8	0.0	9.2	0.0	0.0	0.0	0.0	0.0
		knn	2	67.8	0.0	0.0	0.0	0.0	0.0	67.8	0.0	0.0	0.2	1.8	0.0	1.0	25.2	5.8	0.0	0.0	0.0
		cosine	4	100.0	0.0	0.0	100.0	0.0	19.5	62.7	0.0	0.0	0.0	1.5	0.5	0.5	21.0	4.0	0.0	0.0	0.0
		MCM/RCos	6	100.0	0.0	0.0	100.0	0.0	53.8	72.8	0.0	0.0	0.2	2.2	1.5	0.8	25.8	11.2	0.0	0.0	100.0
		MSP	4	60.5	0.0	0.0	0.0	21.5	0.0	60.5	0.8	1.5	3.0	10.8	0.0	31.2	0.2	4.2	0.0	0.0	0.0
		MaxLogit	4	69.2	2.5	0.0	0.0	25.0	0.0	69.2	2.5	4.5	7.2	12.5	0.0	52.5	0.2	4.8	0.0	0.0	0.0
		Energy	12	100.0	66.2	52.5	0.0	86.2	0.0	90.0	97.8	100.0	86.8	50.7	3.0	98.0	7.8	38.0	49.8	58.2	0.0
		KL-Matching	13	100.0	25.2	44.0	100.0	26.0	100.0	56.0	1.2	1.8	7.8	15.5	68.5	22.8	11.8	17.5	46.8	1.5	100.0
		Mahalanobis	16	100.0	100.0	100.0	100.0	63.0	100.0	87.5	27.5	17.0	52.0	74.0	100.0	1.0	100.0	100.0	100.0	79.2	100.0
CnvNxt-B	84.4	Relative Mahalanobis	15	100.0	99.5	99.2	100.0	51.5	100.0	89.2	16.5	9.0	38.5	70.2	100.0	1.2	100.0	100.0	99.8	72.0	100.0
		ViM	16	100.0	100.0	100.0	100.0	63.0	100.0	98.0	24.5	14.8	48.5	78.8	100.0	0.8	100.0	100.0	100.0	89.2	100.0
		Energy+React	10	87.0	58.5	50.5	0.0	53.8	0.0	79.0	51.2	64.5	39.8	29.0	1.8	87.0	1.8	9.5	47.2	5.2	0.0
		knn	12	100.0	84.5	58.8	0.0	7.2	56.2	81.2	13.8	8.5	29.0	2.2	88.5	0.2	100.0	97.2	84.2	65.0	100.0
		cosine	1	27.5	0.0	0.0	0.0	0.0	0.0	27.5	0.0	0.0	0.0	0.5	0.0	0.0	0.2	0.0	0.0	0.0	0.0
		MCM/RCos	1	49.0	0.0	0.0	0.0	0.2	0.0	49.0	0.0	0.0	0.0	1.2	0.0	0.2	0.2	1.8	0.0	0.0	0.0
		MSP	8	89.0	4.8	13.8	0.0	89.0	0.0	76.0	37.2	36.8	10.0	9.0	6.5	47.0	10.5	8.2	7.5	68.5	0.0
		MaxLogit	6	87.5	3.2	7.2	0.0	87.5	0.0	54.0	22.0	21.8	3.8	2.5	3.5	26.5	0.5	1.0	3.2	59.0	0.0
		Energy	6	89.8	3.2	7.2	0.0	89.8	0.0	44.5	16.5	15.2	1.5	0.5	1.8	16.5	0.0	0.0	3.2	56.8	0.0
		KL-Matching	16	100.0	5.0	27.0	100.0	50.0	100.0	73.0	18.8	15.5	21.8	12.2	91.0	24.0	48.5	51.7	12.0	71.5	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
BiT-s	78.0	Relative Mahalanobis	2	15.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	12.0	0.0	15.5	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	75.8	0.8	2.5	0.0	75.8	0.0	20.0	0.2	0.0	0.0	0.2	0.0	0.5	0.0	0.0	1.5	39.8	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	10	95.0	4.2	15.5	0.0	6.2	0.0	95.0	17.5	18.5	12.0	5.2	34.0	24.5	93.0	85.8	5.5	54.2	0.0
		MSP	1	18.0	0.0	0.0	0.0	18.0	0.0	6.0	2.2	4.0	0.5	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0
		MaxLogit	1	12.0	0.0	0.0	0.0	12.0	0.0	0.5	1.0	1.8	0.2	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0
		Energy	0	5.5	0.0	0.0	0.0	5.5	0.0	0.5	0.5	0.8	1.8	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0
		KL-Matching	2	35.0	0.0	0.0	0.0	23.0	0.0	35.0	3.8	4.0	2.5	0.5	0.0	7.0	0.0	0.5	0.0	0.0	0.0
		Mahalanobis	12	100.0	40.5	3.8	100.0	14.0	100.0	90.5	2.0	2.2	18.5	67.0	100.0	1.5	99.5	100.0	6.2	71.8	100.0
EffNetv2-M	85.1	Relative Mahalanobis	9	100.0	0.2	0.0	100.0	0.0	100.0	84.5	1.0	0.8	6.5	25.2	100.0	0.5	96.8	87.0	0.0	65.8	100.0
		ViM	14	100.0	80.0	10.8	100.0	27.0	100.0	91.8	4.5	3.2	33.2	65.5	100.0	1.5	94.2	98.8	40.8	67.2	100.0
		Energy+React	0	2.8	0.0	0.0	0.0	2.8	0.0	0.0	0.2	0.2	2.0	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	0.8	100.0	60.2	0.5	1.2	1.8	0.8	39.0	0.2	22.2	16.2	0.0	2.0	100.0
		cosine	1	35.2	0.0	0.0	0.0	0.5	0.0	35.2	0.2	0.8	0.0	0.0	0.0	0.2	4.0	3.0	0.0	0.0	0.0
		MCM/RCos	1	10.2	0.0	0.0	0.0	0.0	0.0	10.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0
		MSP	3	17.8	0.0	0.0	0.0	1.5	0.0	17.8	12.5	14.0	4.5	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0
		MaxLogit	0	10.0	0.0	0.0	0.0	1.0	0.0	5.2	8.0	10.0	3.2	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	0.5	0.0	1.2	4.5	5.8	7.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0	0.0
		KL-Matching	3	27.0	0.0	0.0	0.0	5.0	0.0	27.0	16.5	15.8	5.0	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.0
		Mahalanobis	16	100.0	32.8	11.8	100.0	94.2	100.0	99.2	76.2	67.5	75.8	88.5	100.0	14.5	100.0	100.0	6.5	99.5	100.0
EffNetb7	84.9	Relative Mahalanobis	14	100.0	1.8	5.2	100.0	30.8	100.0	97.5	35.5	30.5	32.8	88.2	100.0	12.5	100.0	99.2	3.8	98.5	100.0
		ViM	14	100.0	17.0	10.0	100.0	82.2	100.0	99.2	74.8	63.5	76.0	76.5	100.0	8.0	100.0	100.0	3.8	99.5	100.0
		Energy+React	0	6.0	0.0	0.0	0.0	0.5	0.0	0.2	5.0	6.0	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	52.8	15.5	17.0	5.0	0.0	45.8	0.5	1.0	2.0	0.0	0.0	100.0
		cosine	4	47.0	0.0	0.0	0.0	0.0	19.2	47.0	14.2	14.0	3.0	0.0	3.2	0.0	0.2	0.5	0.0	0.0	0.0
		MCM/RCos	1	34.2	0.0	0.0	0.0	0.0	0.0	34.2	9.8	9.0	1.5	0.0	0.0	0.0	0.2	0.2	0.0	0.0	0.0
		MSP	5	28.5	0.0	0.0	0.0	4.0	0.0	18.8	23.2	28.5	14.5	1.8	0.0	22.5	0.0	0.8	0.0	2.2	0.0
		MaxLogit	5	25.0	0.0	0.0	0.0	1.8	0.0	11.0	22.5	25.0	12.2	0.5	0.0	17.2	0.0	0.2	0.0	4.8	0.0
		Energy	5	31.5	0.0	0.5	0.0	0.8	0.0	6.5	30.8	31.5	18.2	0.5	0.0	27.3	0.2	2.5	0.0	26.5	0.0
		KL-Matching	17	100.0	27.5	56.5	100.0	21.8	100.0	74.2	43.2	44.2	29.8	25.5	89.2	24.8	96.0	65.5	37.5	69.0	100.0
		Mahalanobis	17	100.0	100.0	100.0	100.0	99.8	100.0	93.0	57.5	50.2	70.2	99.2	100.0	19.2	100.0	99.8	100.0	75.0	100.0
EffNet-B0	77.7	Relative Mahalanobis	17	100.0	97.5	99.2	100.0	88.2	100.0	86.0	59.2	54.0	64.2	99.0	100.0	51.5	99.8	97.0	99.8	80.0	100.0
		ViM	15	100.0	56.2	99.0	100.0	44.2	100.0	77.8	15.2	6.5	40.8	76.5	100.0	0.0	99.2	81.0	98.2	59.8	100.0
		Energy+React	0	6.5	0.0	0.0	0.0	0.2	0.0	0.0	3.0	5.2	5.0	0.0	0.0	0.0	0.0	0.8	0.0	6.5	0.0
		knn	1	27.0	1.5	27.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	1.5	0.2	0.0	0.0	2.0	0.0	0.0
		cosine	0	2.8	0.0	0.0	0.0	0.0	0.0	0.0	2.5	2.8	0.2	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
		MCM/RCos	9	96.8	39.5	96.8	0.0	6.2	33.0	13.8	25.0	29.5	9.0	0.8	15.2	7.2	8.5	20.0	94.5	4.2	0.0
		MSP	3	73.2	0.0	0.0	0.0	23.5	0.0	73.2	5.5	9.0	7.2	6.8	0.0	12.5	4.2	3.5	0.0	2.0	0.0
		MaxLogit	8	86.0	0.0	0.0	0.0	25.8	0.0	86.0	11.0	14.8	11.0	12.0	0.0	19.8	9.8	12.2	0.0	7.8	0.0
		Energy	17	100.0	97.8	94.5	100.0	87.8	100.0	99.8	64.8	60.0	83.2	96.0	100.0	48.8	99.2	97.0	96.8	83.0	100.0
		KL-Matching	4	81.5	0.0	0.0	0.0	23.8	0.0	81.5	3.2	5.2	8.0	5.8	0.0	10.0	11.5	10.5	0.0	6.2	0.0
		Mahalanobis	17	100.0	100.0	100.0	100.0	66.8	100.0	83.2	46.2	49.2	71.2	88.5	95.8	14.5	95.5	100.0	100.0	96.2	100.0
ResNet50	80.4	Relative Mahalanobis	17	100.0	100.0	100.0	100.0	89.5	100.0	98.2	82.5	81.2	99.0	99.8	100.0	99.0	100.0	99.8	100.0	100.0	100.0
		ViM	16	100.0	44.5	43.2	100.0	34.5	100.0	77.8	20.2	17.0	36.0	60.0	87.2	1.2	73.2	89.2	52.5	36.0	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	3	28.0	11.0	12.5	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	28.0	0.0	0.0
		cosine	0	8.8	0.0	0.0	0.0	0.0	0.0	8.8	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0
		MCM/RCos	1	33.5	0.0	0.0	0.0	0.0	0.0	33.5	0.5	0.0	1.2	0.0	0.0	0.0	0.5	0.5	0.0	0.0	0.0

K EFFECT OF ID CONTAMINATION ON ALL MODELS

In Table 14 we show the FPR values averaged across the cleaned subsampled datasets presented in Table 1, and in Tables 17-16 the results on the individual datasets. Since TEXTURES and INATURALIST are fairly easy OOD datasets, the FPR values of most models in Table 14 are lower than on NINCO. In general, the results are however similar: Feature-based methods outperform methods not leveraging feature-information directly, 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 14: 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	39.7	27.0 −13	25.7 −14	38.4 −1	22.4 −17	25.5 −14	**22.4** −17	27.5 −12	48.2 +8	30.6 −9	30.4 −9
	84.5	ViT-B-224	43.3	30.8 −13	29.3 −14	42.7 −1	**23.8** −19	28.2 −15	24.7 −19	32.6 −11	53.3 +10	37.0 −6	36.1 −7
	86.3	Swinv2-B-256	41.9	32.3 −10	31.5 −10	46.4 +4	47.4 +5	40.4 −2	37.5 −4	**27.8** −14	43.1 +1	35.5 −6	34.2 −8
	86.7	Deit3-B-384	53.4	45.4 −8	46.4 −7	52.5 −1	40.8 −13	37.8 −16	41.2 −12	39.9 −13	40.1 −13	36.3 −17	**36.0** −17
	85.7	Deit3-B-224	55.1	46.9 −8	47.2 −8	56.1 +1	46.6 −9	42.6 −12	47.5 −8	42.0 −13	45.1 −10	41.4 −14	**40.4** −15
	86.3	CnvNxt-B	38.6	32.9 −6	35.3 −3	43.6 +5	36.3 −2	30.5 −8	29.9 −9	31.1 −8	37.0 −2	30.0 −9	**29.5** −9
	84.1	CnvNxt-T	44.1	37.6 −7	35.7 −8	50.7 +7	36.2 −8	37.0 −7	**27.7** −16	34.0 −10	44.1 −0	40.2 −4	38.9 −5
	82.3	BiT-m	59.9	52.0 −8	52.6 −7	55.3 −5	30.9 −29	32.7 −27	**26.9** −33	46.3 −14	37.2 −23	32.9 −27	38.2 −22
	85.6	EffNetv2-M	43.4	42.5 −1	49.7 +6	46.3 +3	43.7 +0	41.1 −2	37.0 −6	89.0 +46	50.2 +7	**32.4** −11	38.5 −5
none	81.1	ViT-B-384	63.5	59.4 −4	58.8 −5	59.6 −4	49.1 −14	**48.2** −15	61.4 −2	55.4 −8	64.0 +0	59.1 −4	60.9 −3
	84.6	Swinv2-B-256	63.5	63.0 −1	68.6 +5	60.9 −3	49.4 −14	**46.0** −17	52.0 −11	60.5 −3	57.0 −6	52.1 −11	50.4 −13
	85.1	Deit3-B-384	60.0	64.8 +5	83.2 +23	57.8 −2	51.2 −9	48.5 −11	44.9 −15	89.2 +29	65.6 +6	57.2 −3	**43.8** −16
	83.8	Deit3-B-224	60.4	62.2 +2	76.1 +16	58.9 −1	57.6 −3	52.8 −8	**48.9** −11	80.4 +20	73.7 +13	64.4 +4	49.5 −11
	82.6	XCiT-M-224	65.8	65.2 −1	71.4 +6	65.4 −0	58.3 −7	55.7 −10	**55.4** −10	66.9 +1	63.1 −3	57.3 −8	56.4 −9
	84.3	XCiT-M-224-d	63.9	61.6 −2	69.9 +6	61.0 −3	55.4 −8	52.8 −11	**50.4** −13	66.4 +3	59.5 −4	53.6 −10	52.3 −12
	84.4	CnvNxt-B	63.1	72.3 +9	92.1 +29	62.8 −0	55.5 −8	52.1 −11	53.7 −9	88.7 +26	60.8 −2	53.6 −9	**50.6** −12
	78.0	BiT-s	75.3	77.7 +2	79.8 +5	59.8 −15	68.9 −6	**51.2** −24	60.1 −15	65.8 −10	71.2 −4	56.0 −19	84.0 +9
	85.1	EffNetv2-M	59.0	59.4 +0	70.5 +12	56.8 −2	48.2 −11	**42.9** −16	57.4 −2	59.9 +1	54.7 −4	50.2 −9	43.5 −16
	84.9	EffNetb7	60.1	63.4 +3	75.7 +16	56.0 −4	57.9 −2	47.6 −13	63.4 +3	66.6 +6	58.4 −2	52.7 −7	**44.4** −16
	77.7	EffNet-B0	69.3	69.9 +1	77.3 +8	68.2 −1	75.9 +7	68.6 −1	65.7 −4	67.8 −1	77.0 +8	**51.7** −18	63.8 −6
	80.4	ResNet50	68.3	70.0 +2	76.6 +8	64.5 −4	81.0 +13	75.9 +8	73.0 +5	97.6 +29	65.9 −2	**51.7** −17	55.0 −13
JFT	86.8	EffNetb7-ns	53.8	49.9 −4	62.5 +9	52.7 −1	79.5 +26	53.2 −1	82.4 +29	57.0 +3	55.0 +1	47.0 −7	**46.5** −7
clip +12k	87.2	ViT-B-384-l2b	37.3	33.7 −4	35.6 −2	40.5 +3	43.6 +6	36.9 −0	36.7 −1	31.6 −6	35.0 −2	29.5 −8	**29.3** −8
	87.0	ViT-B-384-oai	38.7	33.1 −6	32.9 −6	40.7 +2	45.9 +7	37.4 −1	38.4 −0	31.2 −8	33.7 −5	29.2 −9	**29.1** −10
clip	86.6	ViT-B-384-l2b	54.2	52.5 −2	57.2 +3	51.0 −3	40.2 −14	40.4 −14	**38.4** −16	54.0 −0	44.0 −10	40.0 −14	39.5 −15
	86.2	ViT-B-384-oai	56.7	55.0 −2	59.0 +2	53.9 −3	40.6 −16	40.8 −16	41.4 −15	56.0 −1	45.6 −11	41.3 −15	**40.3** −16
clip z. shot	74.3	clip-ViT-L-336	—-	—-	—-	—-	—-	—-	—-	—-	—-	64.4	**51.8**
	66.6	clip-ViT-B-224	—-	—-	—-	—-	—-	—-	—-	—-	—-	71.4	**60.0**

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	IN1K-f	IN1K-c	OS-f	OS-c	SBe-f	SBe-c	SBh-f	SBh-c	CO-f	CO-c
ViT-B-384-21k		MSP	60.5	37.9	65.8	41.9	63.0	58.3	54.2	52.3	43.4	28.2	27.2	10.5	8.9	83.0	61.3	71.0	32.6	70.7	33.8	77.6	53.4	48.5	38.2
		MaxL	50.6	27.5	65.2	33.7	46.0	40.8	36.5	33.5	21.2	12.2	12.2	3.5	2.1	75.2	54.9	58.9	21.2	**59.0**	16.6	66.7	43.8	33.2	22.6
		ViM	49.9	26.1	53.8	22.1	38.2	35.3	**24.0**	**21.2**	**12.2**	12.5	11.7	**1.5**	**0.5**	78.8	57.4	**54.8**	14.4	60.2	17.9	**59.9**	**32.2**	**26.8**	**16.0**
		Maha	57.5	34.6	**47.5**	**15.7**	**35.2**	**29.1**	28.7	25.7	15.6	**9.5**	**9.5**	2.0	0.8	74.5	**52.8**	**54.8**	**12.3**	65.9	17.9	64.6	41.8	29.0	16.5
		E+R	53.3	30.7	60.2	32.0	43.8	38.5	34.5	31.6	20.1	10.5	10.1	2.8	1.6	83.0	63.8	**54.8**	17.8	63.1	**15.9**	68.8	50.5	32.5	21.7
	86.0	Ener	**49.1**	**25.5**	64.2	30.8	43.2	38.5	35.0	32.4	20.8	11.0	10.6	3.2	1.8	75.8	57.4	58.9	17.4	60.2	16.6	64.6	42.3	31.5	21.2
		KL-M	64.4	43.1	68.5	39.5	57.5	53.7	50.7	48.8	38.5	26.8	25.8	8.5	6.8	82.4	62.1	69.4	29.2	71.9	33.8	77.1	51.4	48.8	37.7
		KNN	69.4	50.3	81.5	59.9	50.2	46.0	38.2	36.2	25.3	36.5	37.5	40.8	41.0	91.5	86.0	64.5	42.4	67.9	24.5	75.0	71.2	54.2	46.2
		RMaha	55.8	30.1	52.0	20.3	42.8	36.6	37.8	35.1	23.3	12.2	12.2	2.0	0.8	**72.7**	53.2	61.3	16.9	68.7	19.9	71.4	44.2	37.8	23.1
		RCos	56.8	33.3	67.8	35.5	40.8	35.3	34.5	31.4	19.1	17.0	16.8	7.2	6.0	83.0	67.2	64.5	24.2	62.7	17.2	72.4	51.0	39.5	28.8
		Cos	57.0	32.7	67.0	33.7	42.5	36.6	36.2	33.2	20.8	15.5	15.2	6.2	5.0	83.0	66.0	66.1	25.4	62.7	19.9	72.4	51.4	40.0	30.2
ViT-B-224-21k		MSP	59.5	39.9	66.2	40.7	67.8	62.8	51.2	49.3	38.5	35.2	36.1	15.2	14.1	89.7	65.5	69.4	39.4	79.5	35.8	81.8	58.7	53.2	44.3
		MaxL	50.1	28.1	65.5	35.5	50.7	45.3	38.0	35.7	24.0	18.0	18.2	6.8	5.2	81.2	59.6	62.1	23.7	65.9	21.9	74.0	48.1	39.2	28.8
		ViM	**48.4**	27.5	58.0	26.2	41.0	36.9	**23.0**	**20.1**	**12.8**	14.5	13.6	3.0	1.8	79.4	57.4	**58.1**	19.1	**61.0**	**17.2**	**64.1**	**38.5**	**29.0**	**20.3**
		Maha	58.3	32.0	**49.5**	**16.3**	**39.8**	**34.6**	27.0	24.7	13.2	**10.2**	**10.1**	**2.0**	**0.5**	78.2	**54.0**	**58.1**	**14.4**	69.9	21.9	68.2	43.8	32.2	20.8
		E+R	53.1	27.5	64.8	36.6	50.7	45.3	38.8	35.9	25.3	17.0	17.1	7.2	5.7	87.3	69.4	60.5	24.6	65.5	24.5	75.5	54.8	38.0	27.8
	84.5	Ener	48.9	**26.8**	65.2	34.9	46.8	41.7	35.8	33.2	22.2	14.0	13.9	6.8	5.2	78.8	60.4	63.7	23.3	63.1	20.5	68.8	47.6	37.0	25.9
		KL-M	64.4	43.8	69.0	38.4	62.3	58.3	50.5	48.5	37.8	34.2	35.3	12.5	10.7	87.9	63.4	66.9	35.2	77.5	41.1	83.9	60.1	53.8	45.3
		KNN	70.4	53.6	83.5	65.7	56.5	52.1	37.2	35.7	26.0	44.5	45.7	39.5	39.2	90.9	86.8	72.6	49.6	75.9	29.8	81.8	75.0	65.0	62.7
		RMaha	60.2	37.9	51.0	17.4	47.5	41.7	39.0	36.2	24.0	16.0	15.8	2.2	**0.5**	**77.0**	54.9	62.1	20.3	77.5	25.8	76.0	46.2	40.8	25.5
		RCos	61.7	39.9	71.2	42.4	45.2	41.1	36.2	33.8	22.6	23.0	23.6	12.8	11.2	83.6	70.2	71.0	30.9	71.1	22.5	78.6	57.7	45.0	35.4
		Cos	60.0	39.2	70.5	41.9	51.5	46.9	35.8	33.2	22.6	22.0	22.6	11.5	10.2	85.5	71.1	70.2	31.8	71.5	22.5	77.6	59.6	48.0	38.2
Swinv2-B-256-21k		MSP	58.8	37.9	65.0	44.2	62.0	59.9	53.5	51.5	44.1	34.5	34.5	19.8	17.5	80.6	61.7	70.2	33.1	67.9	35.8	76.6	54.3	48.0	38.2
		MaxL	53.8	30.7	61.3	36.6	48.5	45.6	50.7	48.8	41.3	25.2	25.0	13.2	11.2	75.2	47.7	64.5	22.9	69.5	27.2	64.6	42.8	35.2	24.5
		ViM	49.9	25.5	65.0	32.6	59.0	53.4	37.8	36.2	25.7	**14.2**	**14.1**	**1.8**	**1.0**	92.1	86.8	66.9	29.2	67.9	32.5	82.3	74.0	44.5	37.7
		Maha	58.8	36.6	67.8	38.4	71.0	68.0	47.0	45.6	36.5	22.2	22.3	3.5	2.6	93.3	87.7	70.2	44.1	71.5	49.0	89.6	84.6	56.5	51.9
		E+R	**46.9**	**21.6**	**59.2**	**30.8**	**42.0**	**38.8**	49.0	46.6	37.8	21.8	21.5	9.0	7.3	75.2	48.1	**55.6**	**17.8**	**65.5**	23.8	**62.5**	37.5	**27.8**	**20.3**
	86.3	Ener	54.8	30.7	62.0	35.5	43.0	40.1	54.8	52.8	45.5	30.8	29.9	13.8	12.0	**73.9**	**44.3**	62.1	21.2	70.7	30.5	64.6	**34.6**	31.8	22.2
		KL-M	64.2	47.7	70.0	44.8	66.0	64.1	52.5	50.9	42.0	33.2	33.7	20.0	17.8	86.7	70.6	67.7	39.4	69.9	36.4	82.3	66.3	56.2	47.6
		KNN	60.2	35.3	73.5	44.8	63.2	58.6	38.8	37.3	27.4	20.2	20.1	7.8	6.5	95.2	90.2	66.1	34.7	70.3	29.8	84.4	79.8	53.0	46.7
		RMaha	57.5	35.3	62.7	32.0	67.8	64.7	44.5	42.9	34.0	17.5	17.4	4.2	3.1	87.9	71.1	66.9	33.1	71.9	40.4	87.5	70.7	51.0	42.5
		RCos	56.3	32.7	65.8	36.6	55.2	50.5	36.0	33.8	22.9	14.8	14.4	4.2	2.9	87.3	69.8	61.3	26.7	69.1	**21.9**	82.8	62.5	45.0	34.9
		Cos	52.1	26.1	67.0	37.8	57.5	53.4	**34.5**	**32.7**	**22.2**	**14.2**	**14.1**	4.0	2.6	89.7	74.5	62.1	28.4	69.5	26.5	84.9	67.3	47.0	37.7
Deit3-B-384-21k		MSP	67.2	52.9	72.8	54.7	73.5	72.2	64.8	63.5	55.9	49.5	48.9	30.0	27.9	90.9	66.8	78.2	41.5	76.7	47.0	83.9	65.9	62.3	53.3
		MaxL	61.5	41.8	72.8	46.5	59.5	57.0	63.5	61.7	53.1	43.8	43.5	29.0	27.2	83.0	60.0	70.2	37.3	76.3	33.1	78.1	**53.4**	54.5	46.2
		ViM	**50.1**	28.1	68.8	43.6	63.7	61.5	50.2	48.3	38.9	**20.5**	**19.8**	**3.8**	**1.8**	92.7	86.4	66.1	32.6	**62.7**	29.1	79.7	72.6	45.5	38.7
		Maha	52.6	29.4	66.0	40.1	66.2	63.8	48.2	46.1	37.5	22.2	22.0	5.8	3.7	90.3	79.1	66.1	33.5	65.1	32.5	82.3	68.8	46.5	38.7
		E+R	56.0	32.7	73.0	45.9	**51.7**	**49.5**	58.2	55.8	45.5	34.8	33.7	23.2	21.9	**81.2**	62.6	**63.7**	34.7	69.1	**22.5**	**73.4**	**53.4**	44.0	**36.3**
	86.7	Ener	62.7	43.1	76.5	53.5	53.2	51.1	67.0	64.9	56.6	47.2	47.0	39.8	38.6	81.8	**56.6**	66.9	36.0	75.1	31.8	75.5	**53.4**	51.7	42.5
		KL-M	67.9	50.3	72.0	52.3	69.5	68.9	59.0	57.4	49.7	47.8	47.3	28.2	26.1	91.5	71.9	75.8	41.9	72.7	45.0	85.9	69.7	62.5	54.7
		KNN	53.3	28.1	72.0	45.9	59.2	56.0	44.8	42.1	32.6	**20.5**	20.1	9.0	7.0	92.7	81.3	65.3	33.5	63.1	27.2	82.3	70.2	45.0	38.7
		RMaha	51.6	28.1	**64.2**	**39.0**	62.0	59.5	46.2	43.4	34.0	21.5	21.2	6.8	4.7	90.3	71.9	64.5	30.1	64.3	29.1	80.7	62.0	**43.5**	**36.3**
		RCos	52.8	27.5	67.8	41.3	57.0	54.7	**42.2**	**39.7**	**29.5**	21.2	20.7	6.5	4.7	90.3	68.9	65.3	**28.0**	65.1	24.5	81.8	59.1	**43.5**	37.3
		Cos	52.8	**26.8**	68.0	41.3	58.5	55.7	42.8	40.2	29.9	21.5	20.9	6.5	4.7	90.3	69.8	65.3	**28.0**	64.7	25.8	82.8	59.6	43.8	37.3
Deit3-B-224-21k		MSP	65.4	48.4	73.2	53.5	73.5	72.5	66.8	64.6	59.4	47.2	45.9	36.2	35.0	89.7	70.2	75.8	51.7	75.1	49.0	82.8	67.3	61.5	53.3
		MaxL	61.0	43.8	72.0	48.3	61.0	58.6	64.5	62.5	55.6	41.2	39.9	36.5	35.5	84.8	58.7	63.7	41.1	69.9	36.4	76.0	56.2	54.5	42.0
		ViM	54.8	34.0	74.8	50.0	68.2	66.7	53.2	51.5	43.1	25.5	25.5	**7.5**	**5.2**	92.7	91.5	68.5	39.0	67.9	41.1	84.9	79.3	52.5	47.6
		Maha	53.3	33.3	71.0	45.9	70.2	68.9	52.5	50.9	42.7	26.0	26.1	8.0	5.7	92.7	85.5	70.2	39.0	69.9	43.0	84.9	74.0	53.2	48.1
		E+R	57.0	34.6	73.0	47.7	**52.2**	**50.5**	59.0	56.8	48.6	33.8	33.4	38.8	37.6	83.6	56.6	**60.5**	36.9	**65.1**	**26.5**	72.9	52.9	49.0	**37.3**
	85.7	Ener	63.5	47.1	74.2	51.2	55.0	52.8	63.5	61.4	54.2	44.2	43.8	53.0	52.5	**81.2**	**53.2**	63.7	38.6	67.9	33.1	**70.8**	**51.0**	53.2	42.5
		KL-M	68.1	51.6	74.5	55.8	71.8	70.9	63.0	61.1	55.6	48.2	46.5	34.2	32.9	91.5	75.7	74.2	55.5	75.1	47.7	84.9	69.2	63.7	55.7
		KNN	56.3	35.9	75.8	51.2	62.7	60.2	47.8	45.3	36.8	24.2	23.9	14.8	12.8	93.9	83.8	70.2	42.8	65.9	29.1	85.4	73.6	51.0	45.8
		RMaha	**52.1**	**31.4**	**67.8**	**41.9**	66.8	65.7	52.0	50.4	42.7	**23.2**	**23.1**	7.8	5.7	90.9	76.6	69.4	**34.3**	67.5	37.1	84.4	67.8	49.8	42.5
		RCos	53.3	32.0	71.0	44.8	60.2	58.6	**47.2**	**44.8**	**36.1**	23.8	**23.1**	9.0	7.0	89.7	73.2	65.3	35.6	66.7	29.8	84.9	65.9	**47.5**	38.2
		Cos	53.6	32.7	71.8	45.9	60.5	59.2	47.8	45.3	37.2	24.2	23.6	9.2	7.3	89.1	74.9	67.7	36.9	65.9	31.1	84.9	66.8	48.0	39.6
ViT-B-384-12b-12k		MSP	55.8	38.6	64.0	42.4	54.5	49.5	47.5	45.0	36.1	26.5	26.4	16.0	14.1	80.6	60.0	64.5	28.8	69.9	29.1	69.8	54.3	41.2	30.7
		MaxL	55.1	35.9	65.0	39.5	52.2	46.6	46.0	43.7	34.0	24.8	24.5	13.0	11.2	78.8	54.9	60.5	24.2	65.9	25.2	63.0	48.1	**36.5**	**26.4**
		ViM	**43.5**	**17.0**	64.0	34.3	60.2	55.7	41.2	38.6	27.1	14.0	13.6	**4.5**	3.1	83.0	78.3	67.7	37.7	63.5	35.1	78.6	66.8	42.2	35.4
		Maha	48.9	26.1	67.2	40.1	65.2	61.5	46.0	43.2	32.6	20.0	19.3	8.2	6.8	87.9	84.7	69.4	45.3	68.3	45.7	83.3	75.0	49.8	42.0
		E+R	52.8	34.0	69.2	43.0	**45.0**	**41.4**	43.0	40.5	30.6	24.2	24.2	12.8	11.5	**77.0**	**51.9**	62.1	**21.2**	**63.1**	21.2	60.9	**40.9**	**36.5**	27.4
	87.2	Ener	60.0	41.2	71.0	45.9	49.5	46.0	50.0	48.0	39.6	29.2	28.3	16.2	14.9	**77.0**	52.8	62.1	22.5	67.9	27.2	**60.4**	42.3	40.2	31.1
		KL-M	55.8	37.9	64.2	40.7	56.5	52.8	43.8	41.0	33.0	27.5	27.4	19.0	17.0	84.2	70.6	63.7	33.1	71.9	33.0	76.6	62.0	46.5	37.7
		KNN	46.7	25.5	67.8	38.4	54.8	49.8	**34.8**	**31.6**	19.8	15.8	14.9	7.0	5.7	89.7	80.4	59.7	29.2	65.1	21.2	78.1	67.8	41.8	32.5
		RMaha	49.4	22.9	63.5	34.3	60.5	56.6	42.2	39.7	28.5	16.5	15.8	6.0	4.4	86.7	77.9	64.5	33.9	65.1	32.5	83.3	65.4	44.5	34.0
		RCos	46.4	22.9	**60.5**	**29.7**	50.0	45.3	35.2	32.2	19.4	**12.2**	**11.7**	**4.5**	**2.9**	84.2	66.0	58.9	22.9	65.1	**19.2**	76.0	55.8	38.0	26.9
		Cos	46.9	22.9	**60.5**	**29.7**	50.0	45.3	35.0	31.9	**19.1**	**12.2**	**11.7**	**4.5**	**2.9**	84.2	66.8	**58.1**	23.3	65.1	19.9	75.5	56.2	38.0	26.9
ViT-B-384-oai-12k		MSP	57.0	39.9	65.0	41.3	55.0	52.8	50.7	48.8	40.3	30.2	30.4	17.0	14.9	80.0	61.7	65.3	27.5	70.3	27.8	72.9	51.9	45.0	37.7
		MaxL	55.6	36.6	65.0	39.5	48.0	45.3	47.0	44.8	35.1	26.5	26.4	14.5	12.3	75.8	**51.9**	61.3	19.5	66.3	23.8	65.6	45.2	36.8	28.8
		ViM	**53.1**	**27.5**	64.0	33.1	62.7	61.2	37.8	35.7	24.7	16.0	16.3	**5.5**	**3.9**	86.7	83.8	62.9	34.3	65.5	29.1	78.6	69.2	44.0	39.2
		Maha	60.0	37.9	66.2	37.8	68.5	67.0	41.0	39.1	27.4	22.5	22.8	11.2	9.9	87.9	84.7	68.5	47.0	74.7	44.4	82.3	76.4	54.8	50.0
		E+R	55.6	37.3	67.0	39.0	**43.2**	**40.1**	46.8	44.8	34.4	23.8	23.4	12.5	10.7	75.8	**51.9**	**58.1**	**17.4**	62.2	22.5	60.4	**40.9**	**34.0**	25.5
	87.0	Ener	56.3	38.6	68.8	40.7	45.0	41.4	50.5	48.5	39.2	26.8	26.6	14.5	12.8	**73.9**	**51.9**	60.5	18.6	64.7	25.2	**58.9**	41.8	35.0	**24.5**
		KL-M	61.7	43.1	67.0	41.9	55.2	53.7	48.0	46.4	37.8	29.0	29.1	16.2	14.1	84.2	67.7	66.9	30.1	70.3	30.5	77.6	58.7	48.2	40.6
		KNN	57.5	34.6	67.8	39.0	47.2	44.0	**30.2**	**27.9**	**15.3**	15.0	15.2	8.0	6.3	89.1	77.0	62.1	28.0	**61.0**	**17.2**	75.5	60.6	40.2	33.0
		RMaha	58.0	35.3	62.3	32.0	60.8	58.3	40.8	38.6	27.1	16.8	17.1	8.0	6.5	86.1	73.6	61.3	31.8	71.5	29.1	80.2	61.5	48.5	39.2
		RCos	54.8	29.4	**61.5**	**29.1**	46.2	41.7	33.0	30.6	18.4	**12.0**	**12.5**	6.5	5.0	84.8	64.7	58.9	22.9	63.9	**17.2**	70.8	52.4	37.5	27.4
		Cos	55.3	30.7	**61.5**	**29.1**	46.0	41.4	32.5	30.0	17.7	**12.0**	**12.5**	6.5	5.0	84.8	66.0	58.9	22.0	63.9	18.5	70.8	51.4	37.0	27.4
ViT-B-384-l2b		MSP	63.7	50.3	67.8	48.8	78.2	76.1	55.2	52.5	44.1	43.5	44.3	37.2	35.8	83.6	70.2	72.6	53.4	73.9	49.0	77.6	69.7	60.2	54.2
		MaxL	62.0	44.4	71.2	51.7	73.2	71.2	55.0	52.0	43.4	38.2	39.4	37.2	36.3	83.0	**67.2**	71.0	51.3	75.5	49.7	77.6	69.7	60.2	52.8
		ViM	**50.9**	**23.5**	66.2	37.2	**59.5**	**57.6**	40.2	37.3	25.7	**18.0**	**18.8**	11.0	9.4	84.8	77.9	65.3	41.1	**61.8**	**21.9**	**74.5**	68.8	48.2	40.1
		Maha	56.3	32.7	65.0	**36.6**	63.0	61.2	39.0	36.2	26.0	19.0	19.8	**10.8**	**8.9**	84.2	77.4	**62.1**	40.3	63.1	28.5	76.0	69.7	49.5	40.6
		E+R	59.5	38.6	78.0	58.7	70.8	68.6	57.5	54.7	45.8	38.8	39.9	45.2	44.6	84.8	71.1	74.2	50.8	74.3	47.7	78.1	70.7	60.2	57.5
	86.6	Ener	62.5	44.4	79.0	61.6	72.2	70.2	60.5	57.6	49.3	42.8	44.3	52.0	51.7	86.1	71.5	77.4	53.4	76.3	51.7	79.2	70.2	64.8	61.3
		KL-M	64.7	48.4	67.2	44.8	69.0	68.9	49.2	46.1	37.5	39.0	39.7	33.2	31.9	**81.8**	71.5	66.1	45.8	72.7	48.3	81.2	70.2	59.0	53.8
		KNN	56.3	33.3	69.5	43.6	63.2	63.4	37.8	35.1	24.7	21.2	21.5	15.8	14.1	87.3	82.1	66.1	43.6	64.7	33.8	83.3	78.4	52.8	45.3
		RMaha	58.3	32.0	64.5	39.5	64.0	61.5	40.0	37.3	26.4	19.0	19.8	12.2	10.4	83.0	71.9	64.5	39.4	62.7	33.1	78.6	68.3	49.8	41.5
		RCos	54.8	32.0	**64.2**	37.8	64.0	61.8	**35.2**	**32.2**	**20.5**	20.5	21.5	13.5	11.7	**81.8**	73.6	65.3	**39.0**	64.3	32.5	79.2	**64.4**	**48.0**	**39.6**
		Cos	55.1	33.3	65.2	37.8	63.7	61.8	37.0	34.0	22.9	19.5	20.4	13.0	11.2	84.2	75.3	65.3	39.4	64.7	30.5	80.2	67.3	48.8	40.1
ViT-B-384-oai		MSP	64.7	49.7	69.2	54.1	78.2	76.1	54.2	52.5	44.4	52.2	52.2	40.8	39.7	84.8	74.9	76.6	52.5	74.3	51.0	86.5	72.1	63.2	56.6
		MaxL	63.5	49.7	71.8	54.7	73.2	70.9	55.2	53.6	45.1	49.2	49.5	37.8	36.8	85.5	72.8	75.8	48.7	75.1	49.7	85.9	73.1	59.5	54.2
		ViM	**50.9**	**28.1**	66.5	37.8	66.5	64.4	41.2	38.6	29.5	24.0	**23.6**	**12.8**	**11.2**	87.3	77.9	68.5	38.1	**63.1**	34.4	**79.2**	68.3	50.0	41.5
		Maha	53.1	29.4	65.2	**37.2**	65.0	62.1	41.2	38.9	29.2	**23.2**	24.2	14.0	12.3	86.1	73.2	**66.1**	37.7	67.1	33.8	81.8	65.9	51.2	41.5
		E+R	63.2	48.4	75.2	56.4	71.0	69.3	60.0	58.4	51.4	45.8	46.7	38.5	38.4	86.1	76.6	73.4	48.7	72.3	50.3	84.9	72.6	61.8	57.1
	86.2	Ener	65.7	52.3	76.2	58.1	73.8	72.5	62.7	61.1	54.5	51.7	52.4	44.5	44.1	86.7	75.7	78.2	52.5	75.1	53.0	84.9	74.5	64.5	59.0
		KL-M	66.2	51.0	72.0	54.7	71.5	69.6	49.2	47.5	39.2	48.8	48.6	38.8	37.3	87.3	75.3	71.0	47.5	72.7	47.7	85.4	69.2	61.0	52.8
		KNN	53.3	30.7	73.5	47.1	64.8	62.5	40.5	38.1	27.8	28.7	29.3	21.5	20.1	88.5	83.4	69.4	44.1	67.1	**32.5**	85.4	74.5	55.8	49.5
		RMaha	55.1	32.0	**64.5**	38.4	67.8	66.0	43.2	41.0	31.6	26.0	26.9	15.8	14.1	**81.8**	**69.4**	**66.1**	**36.0**	68.3	33.8	81.8	**63.0**	**48.5**	**37.7**
		RCos	55.3	30.1	65.8	39.5	**63.2**	**60.2**	**37.2**	**34.6**	**23.6**	24.5	25.3	16.8	15.1	83.6	71.1	67.7	**36.0**	68.3	33.1	83.9	66.8	51.0	42.0
		Cos	55.3	30.1	67.8	40.1	64.5	61.5	39.2	36.7	25.7	25.5	26.4	16.8	15.1	85.5	74.5	67.7	37.3	67.5	**32.5**	85.4	68.3	51.2	42.5

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	IN1K-f	IN1K-c	OS-f	OS-c	SBe-f	SBe-c	SBh-f	SBh-c	CO-f	CO-c
		MSP	58.3	35.9	63.5	40.1	61.3	58.3	51.2	48.8	38.5	32.0	30.2	18.0	16.2	80.6	56.6	70.2	30.5	71.5	30.5	77.1	51.0	46.5	36.8
		MaxL	59.5	37.9	61.0	37.2	53.8	49.8	47.2	44.8	31.9	28.2	25.8	18.5	17.0	76.4	49.8	66.9	22.5	64.7	25.2	67.2	35.6	40.5	28.8
		ViM	**49.6**	**24.2**	**56.2**	30.2	**44.2**	**40.5**	**29.0**	**26.0**	**14.6**	**12.0**	**10.9**	**1.8**	**0.8**	86.1	75.7	**57.3**	24.6	65.1	25.2	73.4	56.2	**35.5**	**26.4**
		Maha	55.3	28.8	57.8	29.1	59.0	56.3	32.0	29.5	18.1	15.5	14.4	3.0	1.8	84.2	75.3	65.3	35.2	68.7	41.7	83.3	63.5	45.2	34.9
		E+R	57.5	33.3	60.2	33.1	49.0	44.0	48.2	46.1	34.4	26.8	24.7	18.2	17.0	**71.5**	50.2	63.7	**17.8**	65.9	23.2	**60.4**	34.1	39.5	29.7
CnvNxt-B-21k	86.3	Ener	63.0	41.2	63.5	39.5	49.5	45.0	55.2	53.9	43.8	36.0	34.0	26.0	25.1	72.1	**48.5**	66.9	21.6	67.9	27.8	60.9	**31.2**	41.0	30.7
		KL-M	65.7	42.5	69.5	44.2	59.8	57.6	49.0	46.4	36.5	33.2	32.9	19.0	16.4	84.2	66.0	68.5	37.3	73.5	37.7	82.8	63.0	55.5	45.8
		KNN	66.9	47.7	69.5	41.9	48.2	43.4	31.5	28.7	17.0	19.5	19.0	7.8	6.3	90.3	79.1	63.7	28.4	**60.6**	19.2	81.8	67.3	45.5	37.3
		RMaha	57.8	32.0	56.5	**27.3**	54.5	50.5	33.5	30.6	18.4	16.2	15.5	2.8	1.3	80.0	56.6	60.5	24.2	68.7	27.8	76.6	54.3	40.5	27.8
		RCos	59.0	30.7	62.7	32.0	50.0	45.6	32.0	29.0	16.3	15.2	14.9	4.5	2.9	84.8	61.3	63.7	23.7	62.2	**17.2**	75.0	51.0	40.2	28.3
		Cos	59.3	32.7	63.5	32.6	49.2	44.3	32.0	29.2	16.3	14.8	14.7	4.0	2.6	85.5	64.3	64.5	23.3	62.2	**17.2**	76.6	51.9	41.5	29.7
		MSP	63.7	40.5	69.0	46.5	68.0	65.7	52.2	50.4	42.0	38.0	38.0	15.8	13.6	89.7	68.9	70.2	33.9	73.1	39.1	80.7	55.8	49.2	41.5
		MaxL	61.0	37.9	67.8	41.3	58.8	56.0	47.5	45.8	36.5	32.2	32.1	12.2	9.9	84.2	61.7	65.3	25.8	68.7	30.5	73.4	48.6	41.8	33.0
		ViM	**48.6**	**24.2**	**53.0**	**25.6**	**47.5**	**43.7**	**27.8**	**24.9**	**15.3**	**11.5**	**10.3**	**2.5**	**1.0**	84.8	68.5	**53.2**	**16.1**	**64.3**	**21.2**	**69.8**	53.8	**34.2**	**25.0**
		Maha	55.6	31.4	56.8	29.7	64.2	61.8	31.0	28.7	17.4	15.8	15.2	3.8	2.3	84.8	75.7	64.5	28.4	71.1	37.7	79.2	66.8	42.5	31.6
		E+R	57.8	34.6	67.5	38.4	51.2	47.2	45.0	43.4	33.7	28.5	27.7	8.2	5.7	**82.4**	59.1	64.5	23.3	68.7	29.8	71.9	45.2	37.0	29.2
CnvNxt-T-21k	84.1	Ener	61.0	40.5	67.8	39.5	51.5	47.6	47.5	46.1	36.8	31.8	31.2	12.0	9.4	**82.4**	**58.7**	64.5	24.6	69.5	29.8	71.9	**43.3**	38.8	31.1
		KL-M	70.4	52.3	72.8	49.4	71.2	71.5	51.0	49.1	41.0	42.5	42.4	19.5	17.2	89.7	73.2	70.2	43.2	73.5	53.0	83.9	67.3	56.0	47.2
		KNN	72.3	52.3	73.0	47.1	59.8	55.7	36.0	33.2	21.5	28.7	28.8	20.0	18.3	90.3	84.3	71.8	36.0	69.5	26.5	82.3	74.0	49.0	41.0
		RMaha	59.3	41.8	59.0	32.0	63.7	61.2	38.0	35.7	25.3	21.2	21.5	5.2	3.4	84.8	67.2	65.3	26.3	74.3	31.8	79.2	62.0	43.5	34.0
		RCos	63.0	39.9	65.2	36.0	62.0	58.3	38.2	36.2	25.3	24.2	24.5	10.8	8.9	86.7	71.5	67.7	30.5	69.1	32.5	81.8	65.4	44.0	35.4
		Cos	64.4	41.2	67.0	39.5	62.0	58.3	37.8	35.7	24.7	26.0	26.4	11.8	10.2	86.7	73.2	66.9	32.2	70.3	31.8	80.7	67.8	44.5	36.8
		MSP	74.3	56.9	72.5	54.7	83.2	82.2	72.8	71.8	68.1	52.5	52.2	28.7	26.1	86.7	76.2	79.0	50.0	78.3	59.6	90.6	76.0	64.2	57.1
		MaxL	69.6	47.7	65.8	41.3	80.2	79.3	66.5	65.4	59.7	43.5	43.5	17.0	14.1	83.6	74.0	77.4	36.9	78.3	57.0	87.0	68.3	58.2	50.0
		ViM	**54.1**	**28.1**	47.0	22.7	38.2	34.6	**5.5**	**5.1**	**2.1**	**14.2**	**12.2**	**2.2**	**1.0**	85.5	67.2	**64.5**	**13.6**	**73.9**	33.1	79.2	57.7	**39.2**	**24.1**
		Maha	64.4	41.8	52.5	25.0	**36.8**	**33.3**	6.5	5.9	**2.1**	19.0	16.0	4.5	2.9	84.2	68.1	69.4	17.8	80.3	39.1	83.9	62.5	48.0	31.6
		E+R	64.2	44.4	**46.5**	29.1	75.8	74.8	67.8	67.3	64.6	44.2	43.8	10.2	8.4	77.0	57.9	74.2	29.7	82.7	62.9	**70.3**	**48.6**	55.5	45.3
BiT-m	82.3	Ener	70.9	51.0	66.5	42.4	79.8	79.0	69.0	68.4	62.8	43.5	42.9	15.2	12.8	84.8	76.6	77.4	36.4	80.7	57.6	85.9	66.8	59.0	50.0
		KL-M	72.6	54.2	74.8	54.1	74.0	73.5	64.0	63.3	58.7	45.0	45.1	28.2	25.6	83.0	74.9	79.0	43.6	76.3	55.0	89.6	71.6	61.5	51.9
		KNN	69.4	47.7	58.8	32.6	42.2	39.5	11.2	10.5	4.9	19.0	16.0	4.5	2.6	93.3	88.9	76.6	21.2	83.9	38.4	87.0	80.3	54.5	37.3
		RMaha	65.2	45.8	49.5	**20.3**	56.2	54.4	23.8	22.3	13.9	23.0	22.0	4.0	2.6	**73.3**	**55.3**	72.6	19.5	80.7	37.1	78.6	56.7	49.2	31.6
		RCos	66.2	39.9	62.5	36.0	64.5	61.8	31.2	29.5	18.4	24.2	23.4	6.0	4.4	83.6	72.8	73.4	28.4	74.3	33.8	82.8	65.9	49.8	35.8
		Cos	63.7	37.3	58.8	30.2	50.5	46.9	16.5	14.5	6.6	18.5	17.1	4.2	2.6	83.0	71.5	70.2	21.6	74.7	**29.1**	83.9	65.4	48.5	33.5
		MSP	61.5	39.9	67.0	45.3	65.2	62.5	52.2	50.4	42.4	36.8	36.7	19.5	17.5	85.5	62.6	76.6	40.3	76.7	34.4	83.9	54.8	51.7	41.0
		MaxL	63.2	45.8	68.8	45.9	61.5	58.6	52.8	50.4	41.7	35.0	34.8	20.8	19.1	86.1	**59.1**	77.4	34.3	77.1	35.8	86.5	**51.9**	50.2	40.6
		ViM	63.2	39.2	65.8	34.3	47.2	42.1	**21.8**	**19.0**	**10.8**	**20.5**	**19.0**	**3.8**	**2.3**	87.9	86.8	69.4	34.3	73.5	23.2	**81.8**	76.0	47.8	38.7
		Maha	69.1	47.1	68.2	39.0	58.0	53.1	27.5	25.5	16.0	28.7	28.0	7.8	6.0	87.9	88.9	**66.9**	41.9	73.5	31.1	83.9	82.2	56.8	47.2
		E+R	94.3	93.5	93.2	90.7	89.2	89.3	87.2	88.7	89.2	87.5	87.0	89.2	89.3	95.8	93.6	91.9	80.1	90.0	90.1	94.8	86.5	90.0	89.6
EffNetv2-M-21k	85.6	Ener	69.6	54.9	75.0	54.7	64.5	63.1	65.5	63.5	56.9	40.2	39.4	26.2	24.5	90.3	63.8	79.8	39.0	83.9	47.7	90.1	56.7	55.2	45.8
		KL-M	65.9	45.8	70.2	47.1	64.5	63.1	51.7	50.4	42.4	37.8	38.6	23.0	21.4	84.2	69.4	76.6	42.4	75.5	32.5	83.9	63.0	54.2	43.4
		KNN	83.5	71.9	75.5	52.9	45.5	41.4	26.0	24.7	14.6	40.8	38.3	13.2	11.5	95.2	95.3	74.2	40.3	88.0	47.7	93.8	86.1	65.0	52.4
		RMaha	70.1	53.6	61.5	34.3	59.8	56.0	34.8	33.0	24.0	27.0	26.6	9.0	7.0	84.8	75.3	**66.9**	35.2	77.5	29.8	82.3	71.2	53.0	39.6
		RCos	63.2	39.9	63.0	38.4	57.2	53.7	39.0	36.7	27.1	25.0	25.3	6.8	5.5	84.2	71.5	68.5	33.9	**72.7**	25.8	85.4	66.3	47.5	35.8
		Cos	**60.5**	**37.3**	**59.8**	**29.7**	**41.8**	**37.5**	23.8	22.0	12.8	22.0	21.5	5.2	3.9	**83.6**	71.1	67.7	**26.3**	73.9	**18.5**	83.9	63.9	**45.2**	**34.0**
		MSP	54.6	34.0	69.5	53.5	69.2	67.3	56.0	54.2	48.3	42.5	41.6	38.8	37.1	84.8	69.4	75.8	62.3	73.9	53.6	83.3	69.7	63.0	55.2
		MaxL	**46.9**	**31.4**	**67.8**	49.4	66.0	64.4	53.2	51.2	44.8	**33.0**	**33.2**	35.2	33.2	81.8	**64.3**	77.4	58.5	71.5	55.6	79.7	**62.5**	59.8	51.4
		ViM	96.5	93.5	92.0	85.5	77.5	77.0	87.5	86.9	85.4	91.8	91.8	83.0	82.8	90.9	95.3	73.4	69.9	80.7	60.3	87.0	88.5	75.5	76.9
		Maha	95.8	92.2	89.0	80.2	76.5	75.7	82.8	82.0	79.2	88.2	88.3	76.5	76.2	90.9	93.6	75.8	67.4	79.9	59.6	87.5	88.5	73.5	73.1
		E+R	57.8	37.9	76.2	61.6	67.5	67.6	59.8	57.9	51.7	41.5	42.1	50.7	49.3	81.8	71.9	80.6	65.7	71.9	55.6	81.2	66.8	60.8	56.6
EffNetb7-ns	86.8	Ener	52.3	39.9	75.5	62.2	77.2	77.3	64.8	63.0	58.0	48.0	48.1	56.8	55.6	**79.4**	67.2	83.9	77.5	79.9	71.5	82.3	64.4	70.2	66.0
		KL-M	62.5	37.9	74.0	54.7	62.3	59.9	53.2	51.7	45.5	40.8	40.5	32.8	30.8	87.3	78.3	69.4	55.1	69.9	51.0	85.4	73.1	62.5	53.3
		KNN	63.7	41.8	83.5	67.4	65.2	63.8	43.2	41.0	29.5	45.5	46.5	36.5	35.0	90.3	91.9	77.4	55.9	67.5	**36.4**	83.3	79.3	59.5	57.1
		RMaha	80.5	63.4	70.2	**48.3**	71.5	70.6	64.8	63.5	57.3	47.2	46.2	**20.0**	**17.8**	82.4	71.1	73.4	**45.3**	73.9	45.7	83.3	69.2	59.2	50.0
		RCos	57.8	32.0	73.8	54.7	**61.3**	59.5	41.0	38.9	27.4	36.5	36.1	22.8	20.9	88.5	80.0	**66.1**	46.2	**65.1**	**36.4**	79.7	71.6	**54.2**	**46.7**
		Cos	57.0	**31.4**	75.2	55.2	61.5	**58.9**	**40.0**	**37.8**	**26.0**	35.0	35.3	24.2	22.5	89.1	83.0	68.5	47.0	65.5	37.1	**78.1**	72.6	54.5	48.1

Table 17: 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	IN1K-f	IN1K-c	OS-f	OS-c	SBe-f	SBe-c	SBh-f	SBh-c	CO-f	CO-c
		MSP	71.6	58.8	78.2	61.6	82.8	81.9	62.5	60.9	54.2	54.0	54.1	46.8	45.7	92.7	81.3	76.6	64.0	77.1	55.6	87.0	78.4	66.0	63.2
		MaxL	68.9	**52.9**	75.5	54.7	80.8	80.6	55.5	54.2	46.2	44.5	44.8	37.8	36.8	90.9	85.5	75.8	57.2	75.1	51.7	85.9	82.2	63.2	61.3
		ViM	73.3	62.1	73.5	55.8	82.8	81.9	54.2	54.2	46.9	49.0	49.2	39.5	38.9	84.8	80.4	**73.4**	59.3	76.7	61.6	84.9	77.4	67.2	62.3
		Maha	70.4	56.2	62.5	35.5	79.2	78.0	**47.0**	**45.8**	**39.6**	**36.0**	**35.1**	16.5	14.9	78.2	65.1	77.4	44.1	77.5	57.0	81.8	63.9	61.3	**50.5**
		E+R	**68.6**	**52.9**	70.8	46.5	80.8	79.6	50.5	49.3	41.7	39.0	38.9	27.8	26.9	87.3	79.1	**73.4**	53.8	75.1	55.0	82.3	77.9	60.8	57.5
ViT-B-384	81.1	Ener	69.6	53.6	75.2	55.2	80.0	78.6	50.5	49.3	41.3	42.2	42.4	35.5	34.2	89.1	86.8	75.0	56.4	75.1	53.0	83.3	82.7	63.5	62.7
		KL-M	72.8	60.8	72.2	52.3	**77.8**	**76.4**	59.5	58.2	52.8	51.0	51.4	36.2	35.0	86.7	76.6	76.6	58.1	76.3	59.6	83.3	75.5	63.7	57.5
		KNN	74.8	64.7	81.5	64.0	78.5	77.0	48.8	47.2	41.3	49.0	49.7	55.2	54.3	92.7	88.5	75.8	66.9	**72.7**	**45.7**	89.1	85.6	66.5	66.0
		RMaha	68.9	53.6	**59.5**	**33.1**	79.2	78.3	48.8	47.5	41.3	38.8	38.0	**13.0**	**11.2**	**77.6**	**59.1**	78.2	**41.1**	79.5	60.3	**78.6**	**62.0**	**60.5**	51.9
		RCos	73.3	60.8	77.2	58.7	79.8	79.3	51.7	50.7	42.7	46.0	46.5	45.2	43.9	92.7	86.8	75.8	57.6	74.7	47.7	87.0	82.2	64.5	63.2
		Cos	71.4	58.8	76.0	56.4	79.5	79.0	49.8	48.5	41.0	44.8	45.1	41.8	40.2	91.5	86.0	75.0	55.9	73.5	**45.7**	86.5	80.8	63.0	61.3
		MSP	68.6	52.3	78.0	64.0	83.8	84.1	66.0	64.9	58.0	58.2	57.3	47.8	46.7	91.5	77.9	80.6	59.7	75.5	57.0	88.0	77.4	67.8	63.7
		MaxL	69.4	54.9	76.5	62.2	79.2	79.3	56.2	55.8	48.6	58.5	57.3	45.5	44.6	90.9	79.6	77.4	59.3	79.5	62.3	85.9	78.8	69.5	65.6
		ViM	63.2	45.1	76.5	54.7	**75.8**	**75.1**	48.5	46.9	38.2	31.8	32.3	22.5	21.1	90.9	87.7	73.4	44.1	67.5	41.1	81.8	83.2	55.2	50.0
		Maha	59.8	41.2	73.2	50.0	77.5	77.0	51.7	50.1	41.0	31.2	31.0	20.2	18.5	90.3	77.9	74.2	43.6	67.9	39.7	80.7	75.0	55.5	48.1
		E+R	67.2	49.0	81.5	66.3	76.0	75.7	47.8	46.6	38.9	49.8	49.7	44.8	43.9	90.9	84.7	75.8	57.2	75.9	53.6	82.8	81.2	68.2	65.6
Swinv2-B-256	84.6	Ener	76.8	66.7	82.5	73.8	76.2	76.4	56.2	55.2	50.7	64.0	63.0	54.8	54.6	**87.9**	84.3	83.9	66.1	84.3	67.5	86.5	80.8	74.0	71.2
		KL-M	72.6	56.9	75.0	58.7	79.5	80.3	63.0	62.2	55.6	51.7	50.8	45.0	44.1	91.5	77.0	77.4	56.8	72.3	53.0	83.9	76.4	65.2	60.4
		KNN	64.7	47.7	79.5	60.5	79.2	78.3	49.5	47.7	39.2	37.5	38.6	40.5	39.2	93.3	90.6	78.2	53.0	**66.7**	39.7	83.3	84.6	59.8	55.2
		RMaha	**58.8**	**39.2**	**69.5**	**44.2**	77.0	76.4	50.7	49.1	38.9	**30.0**	**29.6**	**17.5**	**15.7**	**87.9**	**69.4**	**72.6**	**38.6**	67.9	**39.1**	**79.2**	**70.2**	**53.2**	**45.3**
		RCos	62.7	42.5	75.5	53.5	76.8	75.7	**44.8**	**42.9**	**33.7**	32.5	33.2	24.8	23.2	90.3	81.3	74.2	44.1	67.1	41.1	81.2	77.9	55.8	48.1
		Cos	63.5	43.1	76.2	55.2	77.2	76.7	47.0	45.0	36.1	33.5	34.0	28.7	27.4	90.9	83.4	75.0	47.0	67.5	**39.1**	81.2	80.3	56.8	50.9
		MSP	64.7	52.3	76.5	60.5	80.8	79.9	58.0	55.8	50.0	57.2	57.6	41.5	40.7	90.3	75.7	77.4	61.4	76.7	55.0	86.5	72.6	59.2	54.2
		MaxL	70.4	58.2	80.0	62.8	82.2	82.5	54.8	52.5	47.9	65.0	64.7	50.0	49.9	91.5	78.3	82.3	69.1	81.9	60.3	85.4	75.0	65.5	63.7
		ViM	**58.0**	**34.6**	70.5	44.2	71.2	69.3	42.0	39.9	31.9	34.2	35.1	**16.8**	**15.1**	**82.4**	70.2	**71.0**	41.9	**65.5**	**35.8**	**77.1**	70.2	**51.5**	45.3
		Maha	63.5	46.4	71.0	45.9	76.5	74.1	59.0	57.6	50.7	37.2	39.1	20.8	19.3	88.5	80.9	72.6	45.8	67.5	**35.8**	83.3	77.4	55.5	48.1
		E+R	93.6	90.8	94.5	93.6	90.8	92.6	81.5	81.0	78.5	91.5	90.8	91.0	91.4	91.5	85.5	94.4	91.5	94.0	94.0	90.1	84.6	89.2	87.7
Deit3-B-384	85.1	Ener	88.4	81.7	91.5	86.0	86.0	87.1	66.5	65.4	61.5	83.8	82.3	83.2	83.0	94.5	91.9	88.7	85.2	92.8	84.8	90.1	87.0	84.5	84.9
		KL-M	69.9	54.9	77.2	58.1	74.5	73.1	56.5	54.4	49.3	51.5	51.9	39.2	38.6	89.7	77.0	76.6	55.9	73.5	51.0	84.9	72.1	57.8	53.8
		KNN	66.4	51.0	84.8	69.8	82.5	81.6	64.2	62.7	55.2	50.5	53.3	59.8	59.5	93.3	93.6	75.0	63.6	67.5	43.0	83.9	85.1	65.5	66.0
		RMaha	61.5	43.8	70.2	44.2	76.0	73.8	57.0	55.5	48.3	35.0	36.7	17.2	15.7	85.5	74.0	71.8	42.8	67.1	38.4	82.3	70.2	53.2	45.8
		RCos	70.4	51.6	**69.0**	**41.9**	**66.2**	**63.1**	**39.5**	**37.5**	**29.2**	**29.0**	**29.6**	**16.8**	15.4	85.5	**69.8**	71.8	**37.3**	67.1	**35.8**	82.3	**66.8**	**51.5**	**41.5**
		Cos	64.0	47.1	79.0	58.7	79.5	77.3	57.0	55.2	46.2	40.8	42.7	38.2	37.3	91.5	89.8	73.4	54.2	66.3	40.4	83.3	80.3	59.5	55.7
		MSP	65.2	47.7	73.2	58.1	88.8	88.0	58.0	56.0	51.0	52.5	51.9	41.5	40.2	87.9	74.5	78.2	56.4	75.1	57.6	86.5	76.4	63.5	62.7
		MaxL	64.4	46.4	74.0	58.1	86.8	85.4	55.0	52.8	47.2	56.0	55.7	46.8	45.7	92.7	79.1	78.2	61.9	79.5	60.9	86.5	78.8	66.0	65.1
		ViM	**57.8**	**36.6**	**67.2**	**40.7**	80.8	79.3	**44.0**	**42.6**	**35.8**	**35.5**	**35.1**	**18.8**	**17.5**	**81.2**	72.3	73.4	47.5	72.3	47.7	81.8	74.0	58.2	51.9
		Maha	67.4	49.0	75.2	52.3	85.8	84.1	65.2	64.6	60.8	41.5	42.4	28.5	27.2	86.1	81.3	72.6	50.8	72.7	47.0	82.3	80.3	62.7	58.0
		E+R	83.2	75.8	84.5	72.7	84.8	86.1	69.5	68.1	63.5	79.8	80.2	78.2	78.1	95.8	94.0	87.9	81.8	88.8	80.1	92.7	89.9	81.2	82.5
Deit3-B-224	83.8	Ener	75.3	63.4	82.5	74.4	85.8	85.8	63.7	62.5	58.3	74.2	74.5	69.8	69.5	95.2	87.7	91.1	79.7	88.4	78.8	90.6	86.5	81.2	78.8
		KL-M	73.1	53.6	72.8	55.2	83.2	81.6	58.2	56.8	51.7	50.5	50.5	38.5	37.1	87.9	77.0	71.8	52.5	75.1	50.3	85.4	76.0	63.2	62.7
		KNN	78.5	66.0	89.2	79.1	87.8	86.7	63.2	61.7	57.3	63.2	65.2	74.2	74.2	93.3	94.9	76.6	75.8	73.9	51.7	86.5	86.1	73.8	74.1
		RMaha	64.7	45.8	68.0	43.6	85.5	84.1	62.3	61.7	58.3	37.8	38.3	21.8	20.1	84.2	74.5	**71.0**	**44.5**	**71.5**	**43.7**	**81.2**	73.6	59.8	54.2
		RCos	68.4	49.7	70.0	45.9	**74.0**	**71.2**	45.0	43.4	37.2	38.8	39.4	25.0	23.8	83.0	**68.9**	73.4	47.0	73.1	44.4	81.8	**68.3**	**56.2**	**49.1**
		Cos	72.8	53.6	82.2	66.3	84.2	82.8	57.0	55.2	49.7	52.0	53.3	51.0	50.4	92.1	92.3	73.4	64.8	72.3	48.3	83.3	84.6	64.8	62.7

Table 18: 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	IN1K-f	IN1K-c	OS-f	OS-c	SBe-f	SBe-c	SBh-f	SBh-c	CO-f	CO-c
		MSP	70.1	58.8	79.8	69.2	86.0	84.1	61.5	59.8	54.2	59.0	59.8	49.0	48.0	92.1	84.3	79.0	64.8	75.9	58.3	88.5	**78.4**	64.8	63.7
		MaxL	70.9	60.1	80.0	68.6	**82.8**	81.6	58.0	56.3	50.0	58.5	59.2	47.2	46.0	90.9	85.1	79.0	62.3	75.9	59.6	87.0	80.8	66.2	63.7
		ViM	64.7	48.4	74.0	52.9	84.0	**81.2**	58.2	56.8	49.3	43.0	**42.9**	26.8	25.6	89.7	84.3	71.8	**47.0**	71.9	**47.7**	83.9	79.8	**56.5**	**50.5**
		Maha	63.7	47.7	75.0	54.7	85.8	83.5	65.2	64.1	57.3	46.8	47.3	29.2	28.2	87.9	85.5	72.6	50.0	71.5	51.0	**82.8**	82.2	60.0	53.8
		E+R	70.6	60.1	82.0	67.4	83.0	82.2	61.5	60.1	53.5	59.0	59.8	49.8	49.1	92.7	87.7	78.2	64.4	76.3	60.9	85.4	82.2	69.2	68.9
XCiT-M-224	82.6	Ener	77.8	70.6	82.2	72.7	85.0	84.5	62.5	61.7	55.9	64.0	64.7	59.0	57.7	92.1	87.7	83.9	69.9	82.3	66.9	87.5	83.7	74.0	71.7
		KL-M	72.3	59.5	78.8	70.3	83.8	82.2	61.5	60.1	55.2	60.5	60.6	52.0	51.2	89.1	82.6	77.4	64.0	79.5	56.3	88.0	**78.4**	64.2	59.0
		KNN	65.9	51.0	82.5	65.7	84.2	82.8	55.0	53.6	45.5	50.7	51.6	49.5	48.8	93.3	93.2	74.2	61.0	71.5	48.3	83.3	83.7	63.5	62.7
		RMaha	64.4	45.8	**72.2**	**50.6**	85.5	83.2	64.2	63.0	55.9	**42.2**	**42.9**	**25.2**	**24.0**	**87.3**	**81.3**	**71.0**	47.9	71.9	51.0	83.3	79.3	58.0	50.9
		RCos	62.5	**45.1**	76.2	57.0	83.2	**81.2**	55.0	53.6	45.8	42.8	43.8	32.0	31.1	90.9	86.0	**71.0**	52.1	71.5	**47.7**	85.4	78.8	58.2	51.4
		Cos	**62.0**	**45.1**	77.5	58.1	83.2	**81.2**	**54.2**	**52.5**	**44.8**	44.5	45.4	33.8	32.9	91.5	87.7	71.8	53.0	**71.1**	**47.7**	84.4	81.2	58.8	53.3
		MSP	72.8	58.2	74.2	61.6	87.5	87.7	60.2	58.7	53.1	55.5	55.2	49.8	49.1	91.5	79.6	79.0	64.4	75.5	55.6	83.3	76.9	65.2	61.3
		MaxL	70.1	56.2	73.5	59.9	86.2	85.8	52.2	50.7	44.4	55.5	55.4	40.0	38.9	88.5	80.9	78.2	61.9	75.9	58.3	83.3	75.5	64.2	60.8
		ViM	69.4	49.0	70.2	**43.6**	84.0	**82.5**	49.0	47.5	38.2	**33.5**	**32.9**	**18.8**	**17.0**	87.3	77.9	**72.6**	**47.5**	70.3	41.1	**82.8**	73.1	59.2	51.9
		Maha	71.4	52.3	73.2	51.2	88.2	87.1	55.5	53.9	45.5	41.5	41.3	26.0	24.5	87.9	79.6	74.2	53.0	71.5	44.4	83.9	74.0	62.5	56.6
		E+R	74.1	58.8	81.8	69.8	85.5	84.8	54.8	53.4	45.5	58.5	58.4	48.8	47.5	91.5	87.7	78.2	68.6	75.5	59.6	86.5	81.7	67.8	67.9
XCiT-M-224-d	84.3	Ener	77.0	63.4	80.8	73.8	86.8	86.1	58.8	57.1	50.7	66.5	66.3	50.0	49.1	90.9	86.4	81.5	68.2	85.5	72.2	88.5	81.7	74.2	70.8
		KL-M	75.3	63.4	76.0	60.5	**83.0**	83.5	56.2	54.4	46.9	56.2	56.0	45.2	44.6	90.3	76.6	75.8	58.9	73.5	49.0	**82.8**	74.5	63.2	57.5
		KNN	73.1	54.9	82.2	65.7	86.0	84.8	**47.0**	**45.0**	**35.4**	41.5	42.4	41.8	40.7	93.9	90.2	79.8	58.5	69.1	40.4	84.4	81.7	62.3	59.9
		RMaha	71.6	51.0	**69.8**	46.5	87.8	86.4	54.8	53.1	45.1	39.2	38.9	22.5	20.9	**86.7**	**72.3**	74.2	50.4	71.1	43.0	84.4	**71.2**	61.0	54.7
		RCos	**67.2**	**44.4**	73.0	51.2	84.8	84.1	48.0	46.1	36.8	36.5	36.7	25.2	23.8	92.1	80.9	75.8	51.3	69.1	41.1	83.3	75.0	**58.2**	**50.5**
		Cos	68.6	48.4	75.0	53.5	84.8	83.2	48.0	46.1	36.1	39.0	38.9	28.5	26.9	91.5	82.6	76.6	50.4	**67.5**	**39.7**	83.9	76.9	58.8	52.8
		MSP	69.4	55.6	**66.8**	50.6	91.5	91.3	69.5	67.3	61.8	62.7	63.0	39.2	37.9	85.5	74.9	79.8	62.3	78.7	59.6	83.3	72.6	68.2	64.2
		MaxL	77.8	66.7	72.2	60.5	94.0	93.9	75.8	75.1	69.8	78.0	78.3	49.8	49.1	**84.8**	74.0	88.7	74.6	88.8	79.5	87.5	75.5	77.0	73.1
		ViM	66.7	44.4	74.2	51.7	84.8	83.8	51.2	49.3	39.9	37.2	37.8	32.8	31.3	87.9	80.0	**70.2**	48.3	73.5	43.7	85.4	77.4	59.0	52.4
		Maha	66.2	43.8	74.8	55.2	87.8	87.1	54.5	52.8	44.1	37.8	39.1	39.5	37.6	87.9	80.4	71.8	50.4	71.5	41.7	84.4	76.0	60.2	55.2
		E+R	92.1	88.9	84.2	86.0	95.8	95.5	89.8	89.8	87.8	92.5	92.7	76.8	76.5	87.3	86.0	97.6	89.8	94.4	96.7	89.6	83.2	92.8	92.5
CnvNxt-B	84.4	Ener	94.3	94.8	86.2	90.7	96.2	96.1	93.2	93.6	92.4	95.8	95.9	85.8	85.6	89.1	85.5	98.4	93.6	96.8	98.0	90.6	85.1	96.2	95.8
		KL-M	74.3	57.5	79.0	62.8	85.2	85.8	61.3	59.5	52.4	53.5	54.1	48.0	46.0	88.5	82.6	73.4	58.1	71.9	51.0	**82.8**	79.8	63.7	60.8
		KNN	74.8	56.9	82.0	65.1	84.5	**83.2**	**48.8**	**46.6**	**36.1**	43.8	44.8	46.8	46.2	93.3	88.9	76.6	60.6	71.9	44.4	90.1	83.2	62.3	59.0
		RMaha	**64.7**	42.5	70.8	50.6	87.8	87.1	54.2	52.5	43.8	35.2	36.4	29.8	27.7	**84.8**	**71.5**	71.8	47.0	70.3	42.4	83.9	**71.2**	57.5	52.8
		RCos	65.9	**41.8**	70.0	**48.8**	**84.2**	**83.2**	49.0	46.9	36.5	**34.8**	**35.9**	**26.5**	**25.1**	89.7	79.1	75.8	**44.1**	**69.9**	**40.4**	84.4	73.1	**55.2**	**49.1**
		Cos	67.4	43.1	72.8	53.5	85.5	84.8	49.2	47.2	37.2	35.8	36.4	33.0	31.9	89.7	83.4	76.6	44.5	**69.9**	42.4	85.9	78.8	58.0	53.3
		MSP	78.8	65.4	84.2	74.4	97.0	96.8	76.0	74.8	71.9	71.8	72.8	60.8	59.8	89.1	90.6	**79.8**	73.3	80.7	60.3	87.0	88.9	75.2	73.6
		MaxL	75.3	60.1	85.0	74.4	98.0	98.1	72.2	70.8	68.4	71.0	72.3	75.0	75.2	89.7	92.8	82.3	77.5	**79.1**	64.9	88.0	91.3	77.8	80.2
		ViM	69.1	58.2	80.8	66.3	**64.8**	**62.8**	**6.2**	**5.6**	**3.5**	44.5	42.4	53.5	53.5	94.5	94.9	87.1	62.7	87.1	**57.6**	97.4	91.3	75.5	67.5
		Maha	80.5	72.5	83.2	70.3	70.8	67.6	17.0	16.1	13.9	62.5	60.3	61.3	60.6	93.9	97.0	96.0	71.6	93.6	79.5	95.8	90.4	82.0	74.1
		E+R	**56.0**	**39.2**	76.8	61.6	90.8	90.6	51.0	49.3	45.1	53.8	53.3	48.2	48.6	83.6	87.2	83.9	66.1	81.9	70.9	86.5	88.9	73.2	71.7
BiT-s	78.0	Ener	75.6	61.4	85.2	75.6	97.8	97.7	73.0	71.6	68.1	70.2	71.2	79.0	80.2	89.7	92.8	84.7	78.4	79.9	76.2	90.1	94.2	79.5	81.6
		KL-M	76.5	62.7	70.0	52.9	87.8	87.7	52.2	50.9	46.2	51.0	51.6	31.5	29.8	80.6	72.3	**79.8**	58.9	82.3	62.3	**83.3**	72.1	69.2	60.8
		KNN	84.0	74.5	85.2	73.3	65.5	63.1	13.0	13.1	9.4	61.3	59.2	79.2	79.9	94.5	96.2	91.9	75.8	93.6	78.8	97.9	96.2	84.5	77.4
		RMaha	80.0	69.3	**56.0**	**34.3**	77.8	77.7	24.2	22.5	15.3	**42.2**	**40.8**	**24.2**	**22.5**	**77.0**	**68.9**	83.1	**47.9**	87.6	67.5	84.9	**63.5**	70.5	**56.1**
		RCos	90.1	85.6	87.8	83.7	93.5	93.5	66.2	65.1	60.8	75.5	77.2	87.8	88.3	86.7	92.8	87.9	88.6	89.6	76.2	89.1	93.8	85.8	84.0
		Cos	68.9	53.6	74.5	57.6	67.2	64.4	11.8	10.5	7.6	45.0	42.1	39.0	38.6	87.3	90.2	87.1	58.5	90.4	**57.6**	91.1	89.4	**68.2**	**56.1**
		MSP	66.9	50.3	72.0	52.9	87.5	87.4	54.8	53.4	47.2	53.2	54.3	45.2	44.4	85.5	77.0	77.4	56.4	74.3	53.6	84.9	71.2	59.2	54.2
		MaxL	69.1	52.3	73.0	54.1	86.5	87.1	53.8	52.0	45.5	56.2	56.8	45.2	44.4	87.3	77.4	79.8	56.4	75.1	51.0	85.9	73.1	60.5	55.7
		ViM	68.1	51.6	74.8	52.9	80.2	79.0	47.8	45.6	37.5	37.8	38.3	33.0	32.4	90.9	90.2	69.4	55.1	71.9	48.3	82.3	85.1	61.8	61.3
		Maha	66.9	48.4	68.2	41.9	75.5	73.8	40.2	38.1	29.5	30.8	31.0	18.2	17.0	86.1	79.1	69.4	45.3	69.1	41.1	**76.0**	74.5	53.2	48.6
		E+R	69.9	49.7	78.5	59.3	79.8	79.0	45.5	43.4	35.1	47.8	47.8	43.2	42.8	88.5	88.9	74.2	58.9	70.7	50.3	85.9	82.2	65.8	64.6
EffNetv2-M	85.1	Ener	76.8	64.7	80.8	69.8	86.5	87.7	61.5	60.3	54.9	66.0	66.8	63.5	63.4	87.9	83.0	82.3	72.5	78.7	62.3	89.1	78.4	71.2	72.6
		KL-M	71.1	56.9	70.8	50.6	81.5	81.6	54.5	53.6	46.9	49.8	49.5	41.0	40.2	84.8	71.5	73.4	51.7	71.1	51.0	82.8	71.2	61.8	53.8
		KNN	69.9	51.6	74.8	51.7	80.5	79.3	43.5	41.0	32.6	38.5	39.9	35.5	34.5	89.7	84.7	75.8	51.3	69.1	44.4	83.3	79.3	55.8	52.4
		RMaha	64.4	43.8	**62.7**	**34.9**	75.5	74.1	38.8	36.5	27.8	**27.8**	**28.3**	**15.0**	**13.6**	**81.2**	**65.5**	**67.7**	**36.9**	**67.5**	39.1	76.6	**65.4**	50.5	**42.9**
		RCos	**59.3**	**35.9**	65.8	39.5	**74.8**	**72.8**	**34.5**	**32.2**	**22.9**	29.5	29.6	18.2	16.7	86.7	72.3	70.2	39.4	67.9	**38.4**	80.2	67.3	**50.2**	43.4
		Cos	65.9	43.8	71.2	47.7	80.2	79.3	43.0	40.8	31.6	33.8	34.8	27.0	25.8	87.9	77.4	73.4	44.9	68.7	44.4	81.8	72.6	54.5	49.5
		MSP	66.7	59.5	68.5	50.6	86.5	86.4	55.5	53.6	46.5	52.2	52.7	44.2	43.1	84.2	77.9	77.4	57.6	73.9	57.6	83.9	73.1	60.8	56.6
		MaxL	69.4	60.8	76.8	60.5	85.8	85.8	57.5	55.5	47.6	57.0	57.6	50.5	49.6	84.8	79.1	82.3	63.1	78.3	60.3	85.4	73.6	62.3	59.0
		ViM	74.6	63.4	79.8	64.5	73.5	72.5	61.5	60.6	55.6	49.8	50.3	41.2	40.7	89.1	91.5	**67.7**	61.0	77.5	49.7	84.4	84.6	66.5	63.7
		Maha	73.3	59.5	73.5	52.3	73.5	73.5	59.8	58.7	53.1	41.8	42.7	29.2	28.2	87.9	87.7	71.0	54.2	77.5	45.0	86.5	83.2	63.2	57.1
		E+R	76.5	65.4	82.0	65.7	80.2	80.6	54.2	52.3	42.4	58.0	59.2	56.8	56.1	91.5	88.9	76.6	69.1	79.1	58.3	85.4	78.8	70.2	68.4
EffNetb7	84.9	Ener	83.5	77.1	87.0	76.2	87.2	88.0	65.8	64.1	56.6	73.8	74.2	72.2	71.8	89.1	85.5	87.1	77.5	87.1	72.2	89.6	77.4	76.8	76.4
		KL-M	69.4	57.5	70.2	48.8	79.0	78.3	50.2	48.3	42.4	44.0	45.1	38.0	36.3	86.1	75.7	69.4	50.8	73.5	54.3	**82.8**	73.6	61.0	53.3
		KNN	75.1	59.5	77.8	58.1	77.0	76.7	47.5	45.0	35.4	42.2	44.3	44.2	43.1	87.9	86.0	71.8	58.1	**69.5**	43.0	84.9	81.2	59.8	56.6
		RMaha	71.4	56.9	**63.2**	**37.2**	74.8	73.5	51.0	49.3	42.4	31.0	32.1	20.8	19.6	**82.4**	**64.3**	71.0	42.4	75.5	43.7	**82.8**	**66.3**	56.2	**45.3**
		RCos	**63.0**	**39.9**	65.5	39.5	**72.0**	**70.2**	**39.0**	**37.0**	**27.8**	**29.5**	**30.7**	**18.2**	**17.0**	83.0	74.0	73.4	**39.8**	69.9	**35.1**	84.4	68.8	**53.2**	**45.3**
		Cos	70.4	51.6	71.8	50.0	76.5	74.4	51.0	48.8	39.2	34.5	36.1	30.8	29.5	84.8	78.7	72.6	47.0	**69.5**	48.3	84.4	73.1	56.5	51.9
		MSP	77.0	67.3	80.2	69.8	95.2	94.5	66.8	65.4	62.5	65.0	64.9	54.0	53.3	92.1	79.6	**79.8**	67.8	**79.1**	64.2	89.1	75.5	68.0	63.2
		MaxL	80.7	70.6	83.5	73.8	93.8	93.2	66.5	65.1	60.1	67.0	66.8	61.5	60.8	88.5	**73.2**	83.9	69.9	80.7	66.9	**85.9**	71.2	69.2	62.7
		ViM	84.2	75.2	79.5	65.7	74.5	71.5	24.8	23.6	17.0	56.8	55.2	54.8	53.8	90.3	93.2	81.5	74.6	83.5	**53.0**	91.1	90.9	77.2	73.1
		Maha	90.1	85.0	83.0	72.7	77.8	76.1	44.2	44.2	37.8	70.5	70.9	65.2	64.0	92.7	95.3	89.5	87.3	90.0	69.5	94.8	92.8	84.5	83.5
		E+R	82.5	72.5	83.5	71.5	84.0	82.5	50.0	49.6	44.1	64.8	64.1	63.5	62.9	90.9	74.5	86.3	68.6	85.9	71.5	**85.9**	72.1	72.0	61.8
EffNet-B0	77.7	Ener	86.9	77.8	86.2	79.1	91.5	91.9	72.8	72.4	69.4	77.2	76.9	75.8	75.7	87.9	74.9	89.5	78.0	85.9	85.4	87.0	**70.7**	76.8	70.3
		KL-M	80.7	68.0	77.0	64.5	90.8	90.9	61.8	61.1	58.0	59.8	59.5	45.2	44.1	**86.7**	83.0	81.5	66.1	82.7	70.9	88.0	82.2	70.5	63.2
		KNN	94.1	90.8	87.8	83.1	**70.2**	**67.0**	30.2	29.0	22.2	74.0	72.8	75.5	75.2	98.2	96.6	96.8	82.6	96.0	84.1	97.9	92.8	86.8	79.7
		RMaha	83.2	69.3	71.8	57.6	89.0	88.0	55.0	54.2	50.3	61.8	61.7	46.5	45.2	87.9	83.0	86.3	71.2	83.1	69.5	90.1	88.9	77.0	70.3
		RCos	77.5	62.1	64.8	48.8	89.5	89.3	53.0	51.7	45.5	56.2	56.0	38.2	37.1	**86.7**	80.0	83.1	66.9	82.7	67.5	88.0	84.1	71.8	64.2
		Cos	**71.6**	**54.2**	**59.0**	**41.9**	73.2	72.2	**23.0**	**21.7**	**14.2**	**41.5**	**38.6**	**24.2**	**22.5**	88.5	79.1	80.6	**49.2**	83.5	61.6	**85.9**	80.8	**65.0**	**54.7**
		MSP	**73.1**	**58.2**	79.5	69.2	96.0	95.5	63.7	62.5	58.3	62.3	62.5	51.0	49.9	92.7	83.0	83.1	67.4	**83.1**	60.9	90.1	81.2	66.5	65.1
		MaxL	75.8	64.1	79.2	67.4	96.5	96.1	67.5	66.0	62.2	65.2	64.9	54.0	53.0	92.1	83.4	83.1	67.4	83.5	62.3	90.1	80.3	68.8	68.9
		ViM	86.9	79.7	84.8	75.6	89.0	88.3	40.8	39.9	33.7	64.8	66.6	74.0	73.6	88.5	87.7	**81.5**	77.1	83.9	55.6	88.0	89.9	75.2	75.0
		Maha	94.1	90.2	89.8	84.3	79.8	78.3	53.2	53.1	49.7	77.5	78.5	85.8	85.4	92.7	89.8	87.1	83.5	89.2	73.5	90.1	92.3	86.2	85.8
		E+R	99.8	100.0	99.2	99.4	92.0	92.2	98.2	98.1	98.3	96.5	97.0	99.8	99.7	98.2	99.1	92.7	97.9	91.6	92.7	96.9	98.6	95.8	98.1
ResNet50	80.4	Ener	83.7	78.4	81.5	72.1	96.0	95.5	76.0	75.3	74.0	72.0	72.6	68.5	67.6	90.9	84.3	88.7	74.6	85.5	68.2	88.0	81.7	74.5	74.1
		KL-M	75.3	62.1	75.0	62.8	93.2	92.2	58.8	57.1	51.7	57.2	57.1	44.5	43.6	91.5	77.4	82.3	62.3	**83.1**	60.3	84.9	77.4	66.5	62.3
		KNN	84.2	73.9	82.2	69.2	**68.5**	**65.0**	**24.0**	**23.1**	**16.0**	58.2	57.3	58.8	58.7	89.1	88.9	84.7	70.3	89.2	67.5	89.6	87.5	78.8	70.8
		RMaha	93.1	89.5	70.8	62.2	79.5	79.3	75.5	75.6	73.6	75.0	76.4	83.2	82.8	86.1	**66.4**	85.5	76.7	88.4	79.5	**79.7**	**73.1**	81.8	75.5
		RCos	79.8	64.7	**65.2**	**47.1**	82.5	79.9	39.0	36.7	27.8	47.5	48.6	30.8	29.2	**80.0**	70.6	82.3	49.2	84.3	60.9	80.2	**73.1**	65.8	53.3
		Cos	76.3	60.8	67.2	**47.1**	78.0	75.7	32.2	30.0	21.2	**43.0**	**42.9**	**25.2**	**23.8**	**80.0**	74.0	83.1	**44.9**	**83.1**	**52.3**	81.2	73.6	**63.0**	**52.8**

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 (9 vs. 55 classes) and difficulty of the individual classes. Most models and methods perform better on the classes *with* IN-21k overlap, and ViT+Maha is the best OOD-detector in both cases. While RMaha and (Relative) Cosine yield 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 19: Mean FPR for classes without 21k overlap.

pre	acc.	model	MSP	MaxL	Ener	KL-M	Maha	RMaha	ViM	E+R	KNN	Cos	MCM/RCos
	86.0	ViT-B-384	56.5	41.8 −15	39.6 −17	51.7 −5	**31.7** −25	36.9 −20	32.2 −24	40.9 −16	67.3 +11	46.7 −10	42.2 −14
	84.5	ViT-B-224	64.8	50.6 −14	48.3 −17	60.2 −5	**34.1** −31	43.7 −21	34.8 −30	50.2 −15	68.5 +4	54.8 −10	53.5 −11
	86.3	Swinv2-B-256	66.3	58.7 −8	59.1 −7	62.0 −4	40.1 −26	42.7 −24	**34.3** −32	50.3 −16	54.8 −11	47.5 −19	47.3 −19
	86.7	Deit3-B-384	72.9	71.1 −2	73.3 +0	68.6 −4	**43.0** −30	43.6 −29	44.1 −29	64.4 −9	49.3 −24	47.2 −26	46.8 −26
21k	85.7	Deit3-B-224	75.1	72.8 −2	72.6 −3	69.5 −6	47.7 −27	48.7 −26	**47.1** −28	67.5 −8	56.3 −19	52.9 −22	53.5 −22
	86.3	CnvNxt-B	61.4	60.0 −1	67.0 +6	57.6 −4	31.0 −30	37.4 −24	**27.5** −34	61.6 +0	47.0 −14	40.6 −21	39.7 −22
	84.1	CnvNxt-T	62.9	57.2 −6	54.4 −9	61.6 −1	34.7 −28	42.2 −21	**30.6** −32	52.9 −10	53.3 −10	49.1 −14	46.2 −17
	82.3	BiT-m	69.7	62.2 −7	63.9 −6	62.6 −7	40.9 −29	42.1 −28	**31.5** −38	60.2 −10	39.1 −31	36.0 −34	42.1 −28
	85.6	EffNetv2-M	55.9	51.8 −4	56.3 +0	55.7 −0	48.6 −7	46.9 −9	40.9 −15	96.5 +41	55.3 −1	**33.8** −22	42.4 −14
	81.1	ViT-B-384	70.0	64.5 −5	61.1 −9	65.0 −5	56.6 −13	**56.2** −14	62.8 −7	59.7 −10	66.3 −4	63.0 −7	63.5 −6
	84.6	Swinv2-B-256	72.4	67.7 −5	68.2 −4	68.2 −4	58.9 −13	**56.9** −15	57.6 −15	65.8 −7	67.8 −5	62.2 −10	60.5 −12
	85.1	Deit3-B-384	70.4	75.1 +5	85.4 +15	64.4 −6	59.3 −11	57.4 −13	51.5 −19	91.2 +21	70.7 +0	65.1 −5	**49.2** −21
	83.8	Deit3-B-224	76.4	77.1 +1	83.3 +7	69.5 −7	62.3 −14	60.0 −16	57.9 −18	83.9 +8	75.7 −1	69.4 −7	**55.8** −21
	82.6	XCiT-M-224	79.5	79.1 −0	82.4 +3	76.1 −3	71.6 −8	69.7 −10	**69.2** −10	78.5 −1	76.6 −3	73.3 −6	73.0 −7
none	84.3	XCiT-M-224-d	72.6	71.7 −1	78.8 +6	66.6 −6	63.4 −9	60.8 −12	**60.0** −13	75.3 +3	69.6 −3	62.7 −10	60.9 −12
	84.4	CnvNxt-B	74.1	82.3 +8	94.5 +20	63.9 −10	59.3 −15	56.8 −17	**56.2** −18	90.8 +17	65.7 −8	59.2 −15	58.0 −16
	78.0	BiT-s	74.2	74.5 +0	76.5 +2	58.2 −16	83.2 +9	**56.8** −17	64.4 −10	71.3 −3	81.3 +7	66.8 −7	77.2 +3
	85.1	EffNetv2-M	70.0	69.5 −1	74.4 +4	65.3 −5	52.1 −18	**51.4** −19	59.6 −10	61.7 −8	60.3 −10	56.6 −13	53.0 −17
	84.9	EffNetb7	69.0	70.5 +2	81.3 +12	62.5 −7	55.5 −14	**50.4** −19	59.2 −10	71.0 +2	61.7 −7	58.0 −11	50.4 −19
	77.7	EffNet-B0	75.0	75.9 +1	84.0 +9	68.7 −6	71.0 −4	66.8 −8	62.2 −13	75.0 +0	85.8 +11	**58.7** −16	62.8 −12
	80.4	ResNet50	76.0	76.6 +1	77.5 +1	69.0 −7	77.0 +1	66.4 −10	75.1 −1	94.8 +19	64.0 −12	57.6 −18	**56.6** −19
JFT	86.8	EffNetb7-ns	71.3	64.8 −7	67.5 −4	66.5 −5	83.7 +12	72.0 +1	85.2 +14	65.8 −6	70.3 −1	64.2 −7	**63.8** −7
clip	87.2	ViT-B-384-l2b	53.7	51.1 −3	55.9 +2	52.7 −1	37.8 −16	40.2 −14	**31.7** −22	47.3 −6	41.1 −13	37.3 −16	37.0 −17
+12k	87.0	ViT-B-384-oai	56.0	51.8 −4	54.6 −1	53.6 −2	40.9 −15	39.8 −16	36.9 −19	50.4 −6	36.6 −19	**33.8** −22	34.1 −22
clip	86.6	ViT-B-384-l2b	65.8	63.5 −2	62.5 −3	59.0 −7	49.6 −16	50.4 −15	**46.1** −20	61.0 −5	53.8 −12	49.5 −16	48.3 −18
	86.2	ViT-B-384-oai	65.8	64.1 −2	67.7 +2	62.4 −3	52.4 −13	54.7 −11	**48.1** −18	65.4 −0	57.1 −9	53.9 −12	53.4 −12
clip	74.3	clip-ViT-L-336	—	—	—	—	—	—	—	—	—	**55.7**	55.8
z. shot	66.6	clip-ViT-B-224	—	—	—	—	—	—	—	—	—	**56.9**	62.8

Table 20: Mean FPR for classes with 21k overlap.

pre	acc.	model	MSP	MaxL	Ener	KL-M	Maha	RMaha	ViM	E+R	KNN	Cos	MCM/RCos
	86.0	ViT-B-384	51.1	37.2 −14	36.5 −15	50.1 −1	**26.8** −24	30.2 −21	32.7 −18	38.1 −13	61.9 +11	45.9 −5	45.5 −6
	84.5	ViT-B-224	56.8	45.8 −11	45.7 −11	56.7 −0	**31.6** −25	35.7 −21	39.0 −18	49.3 −8	68.9 +12	54.6 −2	54.4 −2
	86.3	Swinv2-B-256	48.6	38.2 −10	36.9 −12	55.0 +6	66.5 +18	55.7 +7	58.2 +10	**35.3** −13	63.1 +15	52.0 +3	48.3 −0
	86.7	Deit3-B-384	60.0	53.5 −6	53.6 −6	59.0 −1	55.7 −4	49.6 −10	59.0 −1	49.5 −10	54.1 −6	48.6 −11	**47.8** −12
21k	85.7	Deit3-B-224	63.1	57.0 −6	55.8 −7	64.5 +1	62.0 −1	54.7 −8	65.0 +2	53.1 −10	59.1 −4	54.4 −9	**53.1** −10
	86.3	CnvNxt-B	44.9	38.0 −7	39.4 −5	54.4 +10	52.6 +8	43.2 −2	43.8 −1	**37.0** −8	52.6 +8	44.8 −0	43.0 −2
	84.1	CnvNxt-T	53.3	46.4 −7	44.0 −9	60.6 +7	48.9 −4	46.4 −7	**38.5** −15	42.7 −11	57.1 +4	51.5 −2	49.7 −4
	82.3	BiT-m	67.5	62.0 −6	63.0 −4	65.3 −2	51.5 −16	45.6 −22	**42.2** −25	56.6 −11	61.1 −6	54.2 −13	56.5 −11
	85.6	EffNetv2-M	49.9	47.8 −2	53.8 +4	54.4 +4	65.3 +15	52.4 +2	55.6 +6	88.7 +39	69.5 +20	**47.3** −3	51.9 +2
	81.1	ViT-B-384	69.4	68.2 −1	69.3 −0	67.0 −2	60.6 −9	**57.2** −12	70.4 +1	66.8 −3	74.8 +5	69.6 +0	70.8 +1
	84.6	Swinv2-B-256	69.5	67.6 −2	72.9 +3	67.4 −2	64.7 −5	**60.6** −9	67.9 −2	69.3 −0	69.5 −0	63.7 −6	62.3 −7
	85.1	Deit3-B-384	66.8	72.4 +6	87.9 +21	64.6 −2	64.8 −2	59.7 −7	61.3 −5	90.0 +23	75.0 +8	67.5 +1	**58.1** −9
	83.8	Deit3-B-224	69.3	71.1 +2	82.1 +13	68.2 −1	70.1 +1	64.9 −4	64.4 −5	83.0 +14	81.2 +12	73.6 +4	**62.9** −6
	82.6	XCiT-M-224	71.5	72.3 +1	78.6 +7	71.1 −0	65.4 −6	**62.5** −9	64.2 −7	76.0 +4	71.1 −0	66.1 −5	64.9 −7
none	84.3	XCiT-M-224-d	67.6	65.2 −2	72.2 +5	66.9 −1	66.9 −1	**62.1** −6	62.7 −5	72.0 +4	70.6 +3	64.9 −3	62.9 −5
	84.4	CnvNxt-B	63.2	69.7 +7	88.2 +25	68.7 +6	66.8 +4	61.2 −2	67.0 +4	85.1 +22	71.2 +8	61.7 −2	**58.7** −5
	78.0	BiT-s	79.6	82.3 +3	83.9 +4	70.0 −10	83.6 +4	**65.3** −14	75.0 −5	78.9 −1	83.5 +4	73.0 −7	85.3 +6
	85.1	EffNetv2-M	64.5	64.6 +0	74.6 +10	62.4 −2	64.2 −0	55.5 −9	74.7 +10	70.9 +6	65.1 +1	60.0 −4	**54.6** −10
	84.9	EffNetb7	66.4	68.7 +2	81.6 +15	62.7 −4	70.2 +4	55.3 −11	74.9 +8	77.2 +11	67.7 +1	61.0 −5	**54.3** −12
	77.7	EffNet-B0	71.6	71.9 +0	78.9 +7	72.8 +1	85.3 +14	75.2 +4	77.3 +6	75.1 +4	87.1 +16	**61.8** −10	70.9 −1
	80.4	ResNet50	71.8	73.9 +2	78.0 +6	69.0 −3	87.3 +16	70.0 −2	79.2 +7	97.9 +26	80.2 +8	63.8 −8	**63.0** −9
JFT	86.8	EffNetb7-ns	61.8	**54.2** −8	60.5 −1	64.1 +2	88.0 +26	68.2 +6	89.8 +28	61.1 −1	74.3 +13	65.4 +4	63.7 +2
clip	87.2	ViT-B-384-l2b	49.6	46.8 −3	49.4 −0	52.1 +3	55.0 +5	48.5 −1	48.1 −2	44.5 −5	46.2 −3	**40.6** −9	40.7 −9
+12k	87.0	ViT-B-384-oai	47.7	42.3 −5	42.3 −5	48.9 +1	60.4 +13	49.8 +2	55.1 +7	40.9 −7	46.3 −1	40.2 −7	**39.9** −8
clip	86.6	ViT-B-384-l2b	61.2	61.3 +0	66.4 +5	57.3 −4	53.2 −8	50.5 −11	52.6 −9	63.6 +2	57.5 −4	51.0 −10	**49.2** −12
	86.2	ViT-B-384-oai	64.7	65.1 +0	70.1 +5	61.7 −3	56.3 −8	**53.6** −11	58.3 −6	67.7 +3	62.0 −3	57.0 −8	54.4 −10
clip	74.3	clip-ViT-L-336	—	—	—	—	—	—	—	—	—	75.2	**68.9**
z. shot	66.6	clip-ViT-B-224	—	—	—	—	—	—	—	—	—	82.8	**82.6**