Exploring the Limits of Out-of-Distribution Detection

Stanislav Fort [*1] **Jie Ren** [*2] **Balaji Lakshminarayanan** [2]

Abstract

Near out-of-distribution (OOD) detection is a major challenge for deep neural networks. We demonstrate that large-scale pre-trained transformers can significantly improve the state-of-the-art on a range of near OOD tasks across different data modalities. For instance, on CIFAR-100 vs CIFAR-10 OOD detection, we improve the AUROC from 85% to more than 96% using Vision Transformers pre-trained on ImageNet21k. On a challenging genomics OOD detection benchmark, we improve the AUROC from 66% to 77% using transformer and unsupervised pre-training. To further improve performance, we explore the few-shot outlier exposure setting where a few examples from outlier classes may be available; we show that the AUROC of OOD detection on CIFAR-100 vs CIFAR-10 can be improved to 98.7% with just 1 image per OOD class.For multi-modal image-text pre-trained transformers such as CLIP, we explore a new way of using just the names of outlier classes as a sole source of information without any accompanying images, and show that this outperforms previous SOTA on standard OOD benchmark tasks. *This is a condensed version of https://arxiv.org/abs/2106.03004.*

1. Introduction

Deep neural networks are increasingly used in high-stakes applications such as healthcare [26; 25]. Safe deployment of models requires that models not only be accurate but also be robust to distribution shift [1]. Neural networks can assign high-confidence predictions to mis-classified inputs [9; 15] as well as test inputs that do not belong to one of the training classes [21]. This motivates the need for methods that can reliably detect out-of-distribution (OOD) inputs. There has been a lot of progress in detecting OOD inputs including methods based on discriminative models [11; 16; 17; 18] as well as methods based on deep generative models [19; 30].

[*]Equal contribution [1]Stanford University [2]Google Research, Brain Team. Correspondence to: Balaji Lakshminarayanan <balajiln@google.com>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning., Copyright 2021 by the author(s).

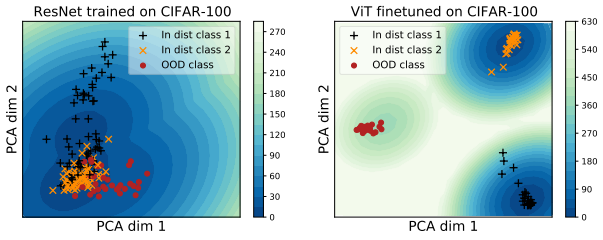


Figure 1: A two-dimensional PCA projection of the space of embedding vectors for 2 models, with examples of 2 in-distribution (from CIFAR-100, "sunflowers", "turtle") and 1 out-of-distribution class (from CIFAR-10, "automobile"). The color coding shows the Mahalanobis outlier score. The left panel shows a ResNet-20 trained on CIFAR-100, which assigns low Mahalanobis distance to OOD inputs and leads to overlapping clusters of class embeddings. ViT fine-tuned on CIFAR-100 (right panel) is great at clustering embeddings based on class, as well as assigning high Mahalanobis distance to OOD inputs (red).

The difficulty of the OOD detection task depends on how semantically close the outliers are to the inlier classes. Winkens et al. [29] distinguish between *near-OOD* tasks which are harder and *far-OOD* tasks which are easier, as evidenced by the difference in state-of-the-art (SOTA) for area under the receiver operating characteristic curve (AUROC). For instance, for a model trained on CIFAR-100 (which consists of classes such as mammals, fish, flowers, fruits, household devices, trees, vehicles, insects, etc), a far-OOD task would be detecting digits from street-view house numbers (SVHN) dataset as outliers. For the same model, detecting images from CIFAR-10 datasets (which consists of the following ten classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck) would be considered a near-OOD task which is more difficult as the classes are semantically similar. There has been impressive progress on far-OOD detection, for instance there are several approaches which can achieve AUROC close to 99% on CIFAR-100 (in) vs SVHN (out) task, cf. [27]. However, the state-of-the-art for near-OOD detection is much lower, for instance the SOTA AUROC for CIFAR-100 (in) vs CIFAR-10 (out) task is around 85% [30] which is much lower than the SOTA for far-OOD tasks. Similar trends are observed in other modalities such as genomics where the SOTA AUROC of near-OOD detection for a classification model is only 66% [25]. Improving the SOTA for these near-OOD detection

tasks and closing the performance gap between near-OOD detection and far-OOD detection is one of the key challenges in ensuring the safe deployment of models.

Large-scale pre-trained transformers have led to significant accuracy improvements in multiple domains, cf. Bidirectional Encoder Representations from Transformers (BERT) for text [7], Vision Transformers (ViT) for images [8], Contrastive Language–Image Pre-training (CLIP) trained on image-text pairs [22]. We show that classifiers obtained by fine-tuning large-scale pre-trained transformers are significantly better at near-OOD detection.Figure 1 visualizes two-dimensional PCA projections of representations from residual networks (ResNet) [10] trained on CIFAR-100 and ViT model pre-trained and fine-tuned on CIFAR-100; we can observe that representations obtained by fine-tuning pre-trained transformers are better suited at near-OOD detection than representations from ResNet just trained on CIFAR-100.

Motivated by real-world applications which demand very high level of OOD detection for safe deployment, we explore variants of outlier exposure to further improve OOD detection. We show that pre-trained transformers are particularly well-suited at leveraging known outliers due to their high-quality representations (see Figure 1). We systematically vary the number of outlier examples per class, and show that even a handful of known outliers can significantly improve OOD detection, we refer to this setting as *few shot outlier exposure*. For multi-modal pre-trained transformers, we explore a new form of outlier exposure that leverages names of outlier classes without any accompanying images, and show that this can significantly improve OOD detection for zero-shot classification.

In summary, our contributions are the following:

- We show that pre-trained transformers lead to significant improvements on near-OOD benchmarks. Concretely, we improve the AUROC of OOD detection on CIFAR-100 $\rightarrow$ CIFAR-10 from 85% (current SOTA) to more than 96% using ViT pre-trained on ImageNet-21k, and improve the AUROC on a genomics OOD detection benchmark from 66% (current SOTA) to 77% using BERT.
- We show that pre-trained transformers are well-suited for few-shot outlier exposure. With just 10 labeled examples per class, we can improve the AUROC of OOD detection on CIFAR-100 vs CIFAR-10 to 99%, and improve the AUROC of OOD detection on genomics to 86%.
- We explore OOD detection for pre-trained multi-modal image-text transformers in the zero-shot classification setting, and show that just using the names of outlier classes as candidate text labels for CLIP can achieve AUROC of 94.8% on CIFAR-100 vs CIFAR-10 task. On easier far-OOD tasks such as CIFAR-$\{100, 10\}$ $\rightarrow$ SVHN, we achieve AUROC of 99.6% and 99.9% respectively.

2. Background and Related work

Notation We assume that we have an in-distribution dataset $\mathcal{D}^{\text{in}}$ of $(\boldsymbol{x}^{\text{in}}, y^{\text{in}})$ pairs where $\boldsymbol{x}$ denotes the input feature vector, and $y^{\text{in}} \in \mathcal{Y}^{\text{in}} := \{1, \ldots, K\}$ denotes the class label. Let $\mathcal{D}^{\text{out}}$ denote an out-of-distribution dataset of $(\boldsymbol{x}^{\text{out}}, y^{\text{out}})$ pairs where $y^{\text{out}} \in \mathcal{Y}^{\text{out}} := \{K+1, \ldots, K+O\}, \mathcal{Y}^{\text{out}} \cap \mathcal{Y}^{\text{in}} = \emptyset$. Depending on how different $\mathcal{D}^{\text{out}}$ is from $\mathcal{D}^{\text{in}}$, we categorize the OOD detection tasks into near-OOD and far-OOD. We first study the scenario where the model is fine-tuned only on the training set $\mathcal{D}^{\text{in}}_{\text{train}}$ without any access to OOD data. The test set contains $\mathcal{D}^{\text{in}}_{\text{test}}$ and $\mathcal{D}^{\text{out}}_{\text{test}}$ for evaluating OOD performance using AUROC. Next, we explore the scenario where a small number of OOD examples are available for training, i.e. the few-shot outlier exposure setting. In this setting, the training set contains $\mathcal{D}^{\text{in}}_{\text{train}} \bigcup \mathcal{D}^{\text{out}}_{\text{few-shot}}$, where $|\mathcal{D}^{\text{out}}_{\text{few-shot}}|$ is smaller than 100 per OOD class. We describe a few popular techniques for detecting OOD inputs using neural networks.

Maximum over softmax probabilities (MSP) A simple and popular baseline method for OOD detection is to use the maximum softmax probability as the confidence score, i.e. $\text{score}_{\text{msp}}(\boldsymbol{x}) = \max_{c=1,\ldots,K} p(y=c|\boldsymbol{x})$ [11].

Mahalanobis distance Lee et al. [16] proposed to fit a Gaussian distribution to the class-conditional embeddings and use the Mahalanobis distance for OOD detection. Let $f(\boldsymbol{x})$ denote the embedding (e.g. the penultimate layer before computing the logits) of an input $\boldsymbol{x}$. We fit a Gaussian distribution to the embeddings, computing per-class mean $\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{i:y_i=c} f(\boldsymbol{x}_i)$ and a shared covariance matrix $\boldsymbol{\Sigma} = \frac{1}{N} \sum_{c=1}^{K} \sum_{i:y_i=c} \big(f(\boldsymbol{x}_i) - \boldsymbol{\mu}_c\big)\big(f(\boldsymbol{x}_i) - \boldsymbol{\mu}_c\big)^\top$. The Mahalanobis score (negative of the distance) is then computed as: $\text{score}_{\text{Maha}}(\boldsymbol{x}) = -\min_c \left(\big(f(\boldsymbol{x}) - \boldsymbol{\mu}_c\big) \boldsymbol{\Sigma}^{-1} \big(f(\boldsymbol{x}) - \boldsymbol{\mu}_c\big)^\top \right)$.

Outlier exposure Hendrycks et al. [12] proposed *outlier exposure* which leverages a large dataset of known outliers. For classification problems, the model is trained to predict uniform distribution over labels for these inputs. Thulasidasan et al. [28] proposed to use a single outlier class as the $(K+1)^{\text{th}}$ class for a K-way classification problem. Roy et al. [26] showed that leveraging the labels of known outliers (rather than assigning all known outliers to a single class) can further improve OOD detection performance.

3. Near-OOD detection on vision benchmarks

3.1. Fine-tuning the Vision Transformer

We use the Vision Transformer (ViT) architecture [8] and its pre-trained model checkpoints.[1] The checkpoints are pre-trained on ImageNet-21k [6]. We fine-tune the full ViT architecture on a downstream task that is either the CIFAR-

[1] https://github.com/google-research/vision_transformer

10 or CIFAR-100 classification problem (using a TPU in Google Colab). We found that we get better generalization and higher quality embeddings when we do not use data augmentation for fine-tuning. Once the model is fine-tuned, we get its pre-logit embeddings (the layer immediately preceding the final layer) for the train and test sets of CIFAR-10 and CIFAR-100 to use for out-of-distribution tasks. For OOD detection, we use the maximum over softmax probabilities (labeled as `MSP`) and the Mahalanobis distance (labeled as `Maha`).

Table 1: ImageNet-21k pre-trained ViT/ BiT fine-tuned on the in-distribution training set.

Model	In-distribution	fine-tuned test accuracy	Out-distribution	Mahalanobis AUROC	MSP AUROC
BiT-M R50x1	CIFAR-100	86.89%	CIFAR-10	81.38%	81.04%
ViT-B_16		90.95%	CIFAR-10	95.53%	91.89%
R50+ViT-B_16		91.71%	CIFAR-10	96.23%	92.08%
ViT-L_16		94.73%	CIFAR-10	**97.98%**	94.28%
ViT ensemble		—	CIFAR-10	**98.11%**	95.15%
BiT-M R50x1	CIFAR-10	97.66%	CIFAR-100	94.57%	85.65%
ViT-B_16		98.10%	CIFAR-100	98.42%	97.68%
R50+ViT-B_16		98.70%	CIFAR-100	**98.52%**	97.75%

The results are summarized in Table 1 and Figure 6. We observe that the MSP baseline yields surprisingly good results when used on top of a large pre-trained transformer that has been fine-tuned on the in-distribution training set. Applying Mahalanobis distance to a pre-trained ViT fine-tuned on CIFAR-100, we achieve AUROC of 96% on CIFAR-100 vs CIFAR-10, significantly improving over the previous SOTA of 85% using a hybrid model [30]. Since ViT models are pre-trained using a large labeled set, Section B.1 presents additional ablations to understand how much of the improvement is due to supervision vs transformers. We refer to the full version for more details https://arxiv.org/abs/2106.03004.

Due to semantic similarity between classes in CIFAR, this task is hard for humans as well.[2] We mainly focus on the difficult near-OOD problem as this is a more challenging and realistic problem; many methods can achieve high AUROC on the easier far-OOD benchmarks, but do not perform as well as near-OOD tasks, cf. [29, Table 1] which compares many methods on both OOD tasks. For completeness, we also evaluate the performance of our approach on some far-OOD benchmarks such as CIFAR-* vs SVHN and CIFAR-* Textures, and achieve very high AUROC values of around 98% or higher, see Table 6 in Appendix B.2.

3.2. Few-shot outlier exposure using ViT

In Section 3.1, we demonstrated that fine-tuning a pre-trained ViT model can improve near-OOD detection (with relatively simple OOD detection techniques such as MSP and Mahalanobis distance). Figure 1 shows that the representations from fine-tuned ViT models are well-clustered.

[2]To approximately estimate the human performance, we performed the CIFAR-100 vs CIFAR-10 near-OOD detection ourselves. Further details are available in Appendix A.

This motivates *few-shot outlier exposure*, where we assume access to just a handful of known outliers (and optionally their labels). This setting is motivated by real-world applications which require high-quality OOD detection and teams are willing to collect a handful of known outlier examples (rather than just rely on modeling approaches) to improve safety.

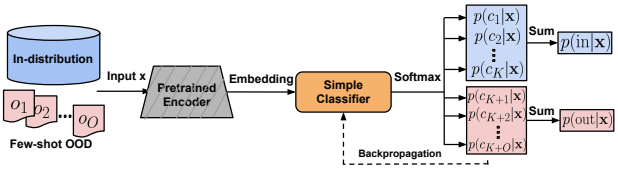


Figure 2: Few-shot outlier exposure with pre-trained transformers. The OOD samples are used during fine-tuning.

The general approach is shown in Figure 2. By using the in-distribution training set $D^{\text{in}}_{\text{train}}$ with K classes and a small number of known OOD examples from $D^{\text{out}}_{\text{few-shot}}$ with O classes. We train a simple classifier $h(\cdot)$ that maps an embedding vector $\boldsymbol{z}$ to a probability vector $\boldsymbol{p} \in \mathbb{R}^{K+O}$, which concatenates the in- and out-of-distribution classes. We considered two types of outlier exposure, one which assumes access to outlier labels (similar to [26]) and one where all the outlier examples are collapsed to a single $K+1$th class (similar to [28]). For models pre-trained with labels (such as ViT), we use a linear classifier. For models that use unsupervised pre-training (e.g. genomics in Section 4), we use a shallow multi-layer perceptron (MLP) with a single hidden layer so that fine-tuning can learn discriminative features. We use the sum of the probabilities of all K in-distribution classes as the confidence score for the OOD detection task, $\text{score}_{\text{oe}}(\boldsymbol{x}) = p(\text{in}|\boldsymbol{x}) = \sum_{c=1,\ldots,K} p(y=c|\boldsymbol{x})$. When training the MLP $h(\cdot)$ with few-shot OOD examples, there could be many more examples of the in-distribution data than the small number of OOD data. We therefore oversample the OOD inputs by a factor that we calculate as $(|\mathcal{D}^{\text{in}}_{\text{train}}|/|\mathcal{D}^{\text{out}}_{\text{oe}}|)(O/K)$. Algorithm 1 and Algorithm 2 describe the details of training and scoring.

Figure 3 and Table 7 show the few-shot outlier exposure results for CIFAR-100 vs CIFAR-10. We observe that even with 1-10 known outliers per class, we can achieve 99% AUROC for near-OOD detection on CIFAR-100 vs CIFAR-10. Interestingly, we observe that having labels for outliers is less important when the transformer is fine-tuned on in-distribution (dashed vs solid red lines) than the scenario where the transformer is not fine-tuned (dashed vs solid blue lines), as illustrated in Figure 9 in Appendix C.2.

4. Near OOD detection of genomic sequences

We investigate OOD detection in genomics as another input modality for near-OOD detection. Ren et al. [25] proposed a benchmark dataset[3] for OOD detection in genomics, mo-

[3]https://www.tensorflow.org/datasets/catalog/genomics_ood

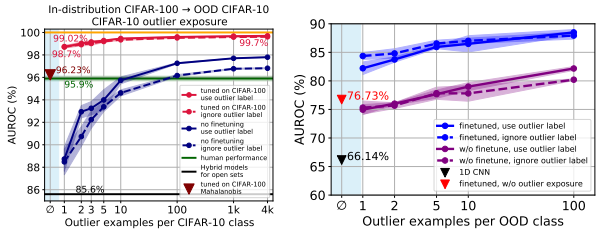


Figure 3: **Left:** The effect of few-shot outlier exposure and fine-tuning on CIFAR-100 vs CIFAR-10 using a R50+ViT-B_16 pre-trained on ImageNet-21k. **Right:** Few-shot outlier exposure for genomics OOD. The x-axis shows the number of outliers per class that the model was exposed to. See Table 9 for exact numbers.

tivated by the real-world problem of bacteria identification based on genomic sequences. Real bacteria sequencing data can contain approximately 60-80% of sequences from unknown classes that have not been studied before. Hence, a classifier trained on all known classes so far will be inevitably asked to predict on genomes that do not belong to one of the known classes. Since different bacteria classes are discovered gradually over the years, Ren et al. [25] use a set of 10 bacteria classes that were discovered before the year 2011 as in-distribution classes, use a set of 60 bacteria classes discovered between 2011-2016 as the validation OOD, and use a set of 60 different bacteria classes discovered after 2016 as the test OOD. The best AUROC for MSP was only 66.14%, and 62.41% for Mahalanobis distance [25]. We also explore the usefulness of pre-trained transformers and fine-tuning for near-OOD detection.

Table 2: Genomics OOD BERT pre-trained and fine-tuned on the in-distribution training set.

Model	Test Accuracy	Mahalanobis AUROC	MSP AUROC
1D CNN [25]	85.93±0.11%	64.75±0.73%	65.84±0.46%
BERT pre-train and fine-tune	89.84±0.00%	**77.49±0.04%**	73.53±0.03%

The results are reported in Table 2. It can be seen that using the approach of pre-training transformer and fine-tuning, the OOD detection performance is significantly improved, from 64.75% to 77.49% for Mahalanobis distance, and from 65.84% to 73.53% for MSP. The in-distribution accuracy also improves a bit, from 85.93% to 89.84%.

Few-shot outlier exposure Results are shown in Figure 3. We observe that exposing to just a small number of OOD examples significantly improves the OOD performance, increasing AUROC from 76.73% to 88.48%. As expected, using the embeddings from the fine-tuned model (blue lines) is better than that from the model without fine-tuning (purple lines). Also, using the outlier labels (purple solid line) has a slightly better performance than collapsing the OOD classes into a single class (purple dashed line) using the pre-trained embeddings without fine-tuning.

5. Using candidate labels with multi-modal text-image models such as CLIP

Multi-modal transformers such as CLIP [22], which are pre-trained on image-text pairs, have been shown to perform well on zero-shot classification tasks. *We show that such multi-modal transformers open the door to new forms of outlier exposure which can significantly improve out-of-distribution (OOD) detection in the zero-shot classification setting.* Our goal is to show that multi-modal transformers can leverage a weaker form of outlier exposure than the few-shot outlier exposure assumption in previous sections, and improve their safety for zero-shot classification.

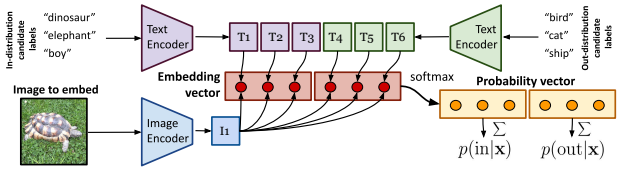


Figure 4: Using candidate text labels and an image-text multi-modal model (CLIP) to produce an embedding vector for OOD detection. Note that this is zero-shot classification and the model is not fine-tuned and does not leverage any in-distribution or OOD images/labels. It only uses the names of the classes (or other informative words) as candidate labels and works well due to the strong pre-training of CLIP.

We use the pre-trained CLIP model[4] (specifically ViT-B/32) that was trained on 400 million (text,image) pairs from the internet. Its image encoder can map an image I into an embedding vector $\boldsymbol{z}_{\text{image}}(I)$, while its text encoder can do the same for a string T as $\boldsymbol{z}_{\text{text}}(T)$. By choosing a set of D *candidate labels* for an image, the similarity between the embedding of the candidate label T_i and an image I can be used as the i^{th} component of the image's embedding vector $\boldsymbol{z}$ as $z_i = \boldsymbol{z}_{\text{text}}(T_i) \cdot \boldsymbol{z}_{\text{image}}(I)$. The procedure is shown in Figure 4. We choose two groups of candidate labels, in-distribution and out-of-distribution labels (e.g. CIFAR-100 and CIFAR-10 class names). Results shown in Table 3.

Table 3: Zero-shot OOD detection using image-text multi-modal models. Even without any fine-tuning (on in-distribution or OOD data), we outperform previous SOTA.

Distribution 1	Distribution 2	Labels 1	Labels 2	AUROC
CIFAR-100	CIFAR-10	CIFAR-100 names	—	69.49%
CIFAR-100	CIFAR-10	CIFAR-100 names	CIFAR-10 names	94.68%
CIFAR-10	CIFAR-100	CIFAR-10 names	—	89.17%
CIFAR-10	CIFAR-100	CIFAR-10 names	CIFAR-100 names	94.68%
CIFAR-100	SVHN	CIFAR-100 names	—	93.05%
CIFAR-100	SVHN	CIFAR-100 names	["number"]	99.67%
CIFAR-10	SVHN	CIFAR-10 names	—	96.90%
CIFAR-10	SVHN	CIFAR-10 names	["number"]	99.95%

[4] https://github.com/openai/CLIP

Acknowledgements

We thank Abhijit Guha Roy, Jim Winkens, Jeremiah Liu and the anonymous reviewers for helpful feedback. We thank David Dohan and Andreea Gane for the helpful advice on BERT genomics model pre-training. We thank Basil Mustafa for providing the BiT model checkpoints. We thank Matthias Minderer for his helpful advice on ViT. We thank Winston Pouse for useful discussions on human performance.

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[2] Guillaume Bernard, Cheong Xin Chan, and Mark A Ragan. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. *Scientific reports*, 6(1):1–12, 2016.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[4] Cheong Xin Chan, Guillaume Bernard, Olivier Poirion, James M Hogan, and Mark A Ragan. Inferring phylogenies of evolving sequences without multiple sequence alignment. *Scientific reports*, 4:6504, 2014.

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

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[9] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[11] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[12] Dan Hendrycks, Mantas Mazeika, and Thomas G Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6(2): 8, 2019.

[15] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

[16] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018.

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

[18] Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *NeurIPS*, 2020.

[19] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Hybrid models with deep and invertible features. In *ICML*, 2019.

[20] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[21] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pages 427–436, 2015.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[23] Gesine Reinert, David Chew, Fengzhu Sun, and Michael S Waterman. Alignment-free sequence comparison (I): statistics and power. *Journal of Computational Biology*, 16(12):1615–1634, 2009.

[24] Jie Ren, Xin Bai, Yang Young Lu, Kujin Tang, Ying Wang, Gesine Reinert, and Fengzhu Sun. Alignment-free sequence analysis and applications. *Annual Review of Biomedical Data Science*, 1:93–114, 2018.

[25] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A DePristo, Joshua V Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *NeurIPS*, 2019.

[26] Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, Nam Vo, Peggy Bui, Samantha Winter, Patricia MacWilliams, Greg S. Corrado, Umesh Telang, Yun Liu, Taylan Cemgil, Alan Karthikesalingam, Balaji Lakshminarayanan, and Jim Winkens. Does your dermatology classifier know what it doesn't know? Detecting the long-tail of unseen conditions. *arXiv preprint arXiv:2104.03829*, 2021.

[27] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In *ICML*, 2020.

[28] Sunil Thulasidasan, Sushil Thapa, Sayera Dhaubhadel, Gopinath Chennupati, Tanmoy Bhattacharya, and Jeff Bilmes. A simple and effective baseline for out-of-distribution detection using abstention. 2021. URL https://openreview.net/forum?id=q_Q9MMGwSQu.

[29] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R. Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, Taylan Cemgil, S. M. Ali Eslami, and Olaf Ronneberger. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

[30] Hongjie Zhang, Ang Li, Jie Guo, and Yanwen Guo. Hybrid models for open set recognition. *European Conference on Computer Vision*, 2020.

[31] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

A. Measuring human performance on CIFAR-100 vs CIFAR-10 OOD task

While the typical accuracy a human reaches is typically known for classification tasks, there is a lack of such benchmark for near-OOD detection. We decided to measure human performance on the task of distinguishing CIFAR-100 and CIFAR-10. To do that, we wrote a simple graphical user interface (GUI) where a user is presented with a fixed number of images randomly chosen from the in-distribution and out-of-distribution test sets (CIFAR-10 and 100 in our case). The user would then click on the images they believe belong to the in-distribution. To make this easier, we allow the user to choose the images belonging to the individual classes of the in-distribution. An example of our GUI is shown in Figure 5.

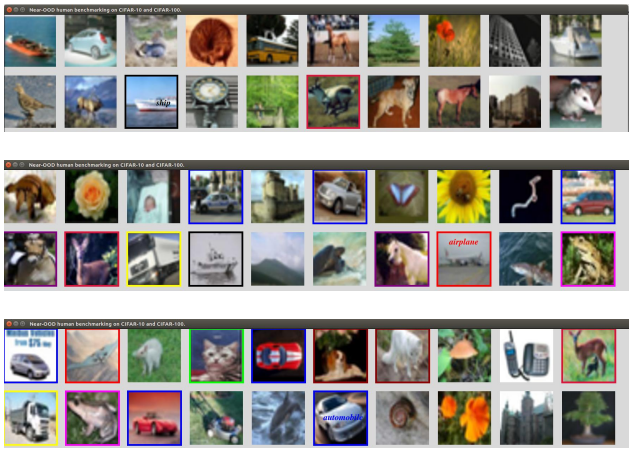


Figure 5: Graphical User Interface (GUI) for human benchmarking of near-OOD detection. The user clicks on images belonging to any of the 10 in-distribution CIFAR-10 classes. The user is shown 20 images at a time, and once they are done with their selection, they press a key and a new group of randomly chosen images (equal probability of CIFAR-10 and CIFAR-100) is shown. The images show 3 randomly chosen groups of 20 images, together with user-selected CIFAR-10 images framed by the color corresponding to their selected class.

For the case of CIFAR-10 and CIFAR-100 distributions (selecting the classes of CIFAR-10), the user would click on all images belonging to each of the 10 classes *airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck*. This was done without any exposure to the training set or examples of any of the classes, based only on the class names and using the fact that people are generally very familiar with the semantic concept of these labels. Coincidentally, this is quite similar to the kind of familiarity large pre-trained transformers gain by the breath of their pre-training.

We calculated the AUROC for the CIFAR-10 / CIFAR-100 task on their test sets, where choosing any of the 10 classes was acceptable as a valid CIFAR-10 selection by the user. The results are shown in Table 4. The average AUROC weighted by the number of images in each trial is AUROC 95.90%.

Table 4: Author's results on CIFAR-100 vs CIFAR-10 distinguishing task.

Date	Number of images	AUROC
April 7 2021	1140	96.14%
April 25 2021	1700	95.67%
April 27 2021	940	96.03%

B. Additional vision OOD details

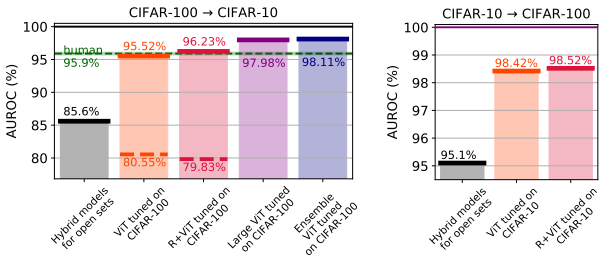


Figure 6: Left: CIFAR-100 vs CIFAR-10 OOD AUROC for previous state-of-the-art [30], and our fine-tuned ViT with two different backbones, ViT-B_16 and R50+ViT-B_16. Right: CIFAR-10 vs CIFAR-100 OOD task.

To study the effect of model architecture, we also evaluate OOD performance on another large-scaled pre-trained model, Big Transfer (BiT) [14][5], as a comparison to ViT. We use model BiT-M R50x1 pre-trained on ImageNet-21k, and fine-tuned the full model architecture on CIFAR-10 and CIFAR-100 respectively. The results are shown in Table 1. For both directions, the AUROCs for BiT are lower than that for ViT. Note that the fine-tuned test accuracy for BiT is slightly lower than that for ViT, which may affect the performance. More importantly, BiT uses a different model architecture ResNet instead of transformer, which may explains the large difference on the OOD performance.

B.1. Additional ablations to measure the effect of supervised pre-training

Since ViT models are typically pre-trained using a large labeled set, we ran additional ablations to understand how much of the improvement is due to supervision vs transformers. To assess the role of supervised pre-training, we compared the results to a ViT pre-trained on ImageNet-21k in a self-supervised way that does not use any labels. We took a pre-trained checkpoint from Caron et al. [3], and fine-tuned it on CIFAR-100. The results are shown as DINO ViT-B_16 in Table 5. Since the fine-tuned test accuracy is

[5] https://github.com/google-research/big_transfer

lower for DINO ViT-B_16, we also include an ViT-B_16 that was fine-tuned for fewer steps to achieve comparable test accuracy as DINO ViT-B_16. Note that even though DINO ViT-B_16 is pre-trained without labels, the AUROC is significantly higher than the current SOTA for CIFAR-100 vs CIFAR-10. We believe that better unsupervised pre-training and/or better fine-tuning strategies could improve this number even further. The difference between DINO ViT-B_16 vs early stopped ViT-B_16 shows the difference due to supervision during pre-training. In Section 4, we explore unsupervised pre-training for transformers to improve near-OOD detection in genomics.

Table 5: Additional ablations to measure the effect of supervised pre-training. * indicates self-supervised pre-training without labels.

Model	In-distribution	fine-tuned test accuracy	Out-distribution	Mahalanobis AUROC	MSP AUROC
DINO ViT-B_16*	CIFAR-100	88.95%	CIFAR-10	88.78%	81.25%
ViT-B_16 (early stop)	CIFAR-100	88.71%	CIFAR-10	93.05%	88.82%
ViT-B_16	CIFAR-100	90.95%	CIFAR-10	95.53%	91.89%
R50+ViT-B_16	CIFAR-100	91.71%	CIFAR-10	**96.23**%	92.08%

B.2. Additional datasets and metrics

We tested our OOD detection on several more tasks in addition to CIFAR-$\{10, 100\}$ $\rightarrow$ CIFAR-$\{100, 10\}$ in Table 1. We used the *Describable Textures dataset* (DTD, or Textures in our tables) [5], *Places365-Standard* (Places365 in our tables) [31] and SVHN Netzer et al. [20] datasets as provided by `tensorflow` datasets.[6] In all cases we use the test set of each dataset.

Besides the area under the receiver operating characteristic curve (AUROC), we also evaluate OOD performance using the area under the precision-recall curve (AUPRC) and the false positive rate at N% true positive rate (FPRN), as in [12]. Since our goal is to detect OOD, we treat OOD test set as the positive set and in-distribution test set as the negative set. AUROC and AUPRC are threshold independent, evaluating the overall OOD performance across multiple thresholds. AUROC is also sample size independent, while AUPRC is sensitive to detect imbalance between positive and negative sets. FPRN computes the false positive rate at which $N\%$ of OOD data is recalled. As a common practice, we set $N = 95$.

The additional results are shown in Table 6. SVHN and Textures are far-OOD tasks and our methods achieve AUROC of around 98% or higher. Places365 is a harder OOD task, see the discussion in [29] where they compute Confusion Log Probability (CLP) score for different OOD datasets and demonstrate that CIFAR-100 vs CIFAR-10 and CIFAR-* vs Places365 are more difficult OOD detection tasks than CIFAR-* vs SVHN and CIFAR-* vs Textures. On CIFAR-100 vs Places365, we achieve 93.9% (Winkens et al. [29] report 82%) and on CIFAR-10 vs Places365, we achieve 98.5% (Winkens et al. [29] report 95%).

[6] https://www.tensorflow.org/datasets

B.3. Qualitative analysis of OOD detection using ViT

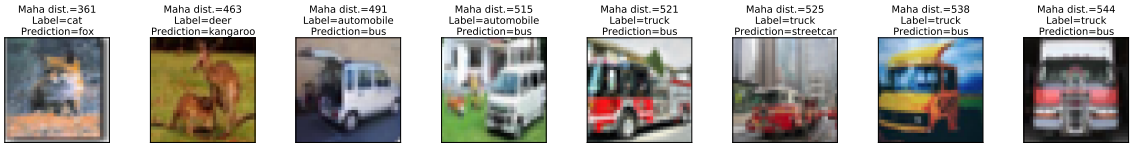
In this section we present some qualitative failure cases of OOD detection. The results for the CIFAR-100 (in-distribution) vs CIFAR-10 (out-distribution) experiment are shown in Figure 7. The most CIFAR-10-like images are photos of birds, as CIFAR-100 does not include any bird classes at all. The CIFAR-10 images that are mistakenly classified as in-distribution (CIFAR-100), the top images are actually mislabeled images from the CIFAR-10 test set (a fox labeled as a *cat*, and a kangaroo labeled as a *deer*). They are followed by *automobiles* and *trucks* (both CIFAR-10 classes) that are seen as *buses* and *streetcars* (similar CIFAR-100 classes). In those cases, the distinction could be unclear even to a human. We show several qualitative failure cases for the CIFAR-100 (in-distribution) vs SVHN (out-distribution) experiment in Figure 8. We show the SVHN images least similar to CIFAR-100. The out-distribution images that are most like CIFAR-100 (Figure 8a) are numbers "2", "5" and "6" written in an especially wiggly font, sometimes with plastic-like filler, giving them a very worm-like quality that the model perceives. The last image we show gets perceived as a cloud, likely due to its deep blue background and low resolution. In Figure 8b we show the images from the in-distribution that have the closest embedding vector to the SVHN mistakes. They do indeed look very similar.

Overall the mistakes our models make are semantically meaningful to a human. For CIFAR-10, they contain mislabeled examples of the CIFAR-10 test set that should not have been included in the first place, or genuine imperfect categories such as *automobile*, *truck*, *bus*, and *streetcar*, which are even harder to distinguish in the 32×32 resolution. For SVHN, the mistakes we get are also very plausible to a human.

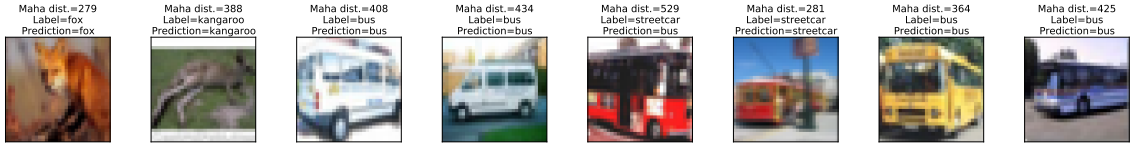
Figure 11 and Figure 12 show examples of the most confusing pairs of (OOD class (CIFAR-10), in-distribution class (CIFAR-100)). We looked at the OOD examples with the lowest Mahalanobis distances, e.g. the ones that are the most in-distribution-like to the network at hand (ViT fine-tuned on the CIFAR-100 train set). Going from the smallest distance to the highest, we sorted these images based on the tuple of which CIFAR-10 class the OOD image came from, and which class the CIFAR-100 test set image that has the most similar embedding vector to it. We show examples that come from classes related to **vehicles** in Figure 11, and to **animals** in Figure 12. Some of the least-OOD-like CIFAR-10 examples shown are literal mistakes, and they should not have been included in CIFAR-10 at all (e.g. the first column "truck" being actually a tractor in Figure 11,

Table 6: ImageNet-21k pre-trained Vision Transformer fine-tuned on the in-distribution training set. More datasets in addition to Table 1.

In-distribution	fine-tuned test accuracy	Out-distribution	Maha. AUROC ↑	Maha. AUPRC↑	Maha. FPR95↓	MSP AUROC↑	MSP AUPRC↑	MSP FPR95↓
CIFAR-100	91.71%	CIFAR-10	96.23%	96.32%	18.73%	92.13%	92.57%	37.73%
		Textures	99.03%	96.53%	4.27%	96.28%	99.22%	17.30%
		SVHN	97.80%	98.87%	8.42%	97.15%	93.61%	13.05%
		Places365	93.95%	99.78%	29.22%	88.37%	39.57%	51.02%
CIFAR-10	98.70%	CIFAR-100	98.52%	98.70%	6.89%	97.79%	97.72%	9.76%
		Textures	99.97%	99.83%	0.05%	99.59%	99.92%	1.73%
		SVHN	99.58%	99.82%	1.90%	98.77%	97.75%	4.03%
		Places365	98.51%	99.95%	4.54%	97.14%	50.92%	10.13%

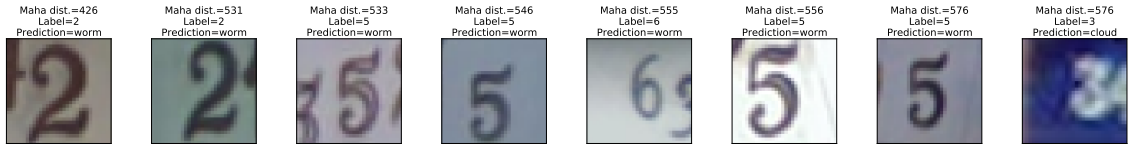


(a) OOD images (CIFAR-10) closest to the in-distribution (CIFAR-100).

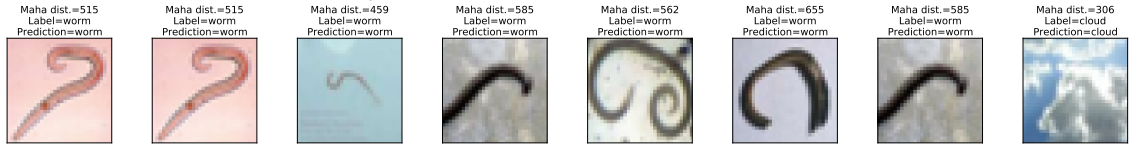


(b) The in-distribution (CIFAR-100) images with the closest embedding vector to images in Figure 7a.

Figure 7: CIFAR-100 (in-distribution) vs CIFAR-10 (out-distribution): Images from the out-distribution that reached the lowest Mahalanobis distances from the in-distribution (Figure 7a) and the in-distribution images with the most similar embedding vectors (Figure 7b). The comparison to the closest in-distribution images demonstrates that OOD detection in these particular cases might be failing due to the limitations of the labelling in the original dataset and genuine semantic ambiguity of some classes. The first two images in Figure 7a do not belong to CIFAR-10 and have likely been included by accident.



(a) OOD images (SVHN) closest to the in-distribution (CIFAR-100).



(b) The in-distribution (CIFAR-100) images with the closest embedding vector to images in Figure 8a.

Figure 8: CIFAR-100 (in-distribution) vs SVHN (out-distribution): Images from the out-distribution that reached the lowest Mahalanobis distances from the in-distribution (Figure 8a) and the in-distribution images with the most similar embedding vectors (Figure 8b). The most similar images demonstrate that the some of the digits genuinely look like worms.

or the first column "cat" in Figure 12 being actually a fox). Many of the classes are have a genuine semantic overlap, especially among the vehicles. For example, the CIFAR-10 class "automobile" has a large number of vans, that are also included in the CIFAR-100 class "bus". The CIFAR-10 class "truck" and the CIFAR-100 class "pickup truck" share equally mutually similar vehicles.

C. Few shot Outlier exposure

Table 7 show the few-shot outlier exposure results for CIFAR-100 vs CIFAR-10.

C.1. Pseudocode for few-shot outlier exposure

Algorithms 1 and 2 describe pseudocode for few-shot outlier exposure training and evaluation.

Algorithm 1 Few-shot outlier exposure training

Input: In-distribution train set $\mathcal{D}^{\text{in}}_{\text{train}} = \{(\boldsymbol{x}, y)\}$ with K classes, out-of-distribution train subset $\mathcal{D}^{\text{out}}_{\text{few-shot}} = \{(\boldsymbol{x}, y)\}$ with O classes, oversampling factor Γ, a pre-trained feature extractor $f(\cdot) : \boldsymbol{x} \to \boldsymbol{z}$, a simple classification head $h(\cdot) : \boldsymbol{z} \to \boldsymbol{p} \in \mathbb{R}^{K+O}$.
Initialize: Initialize $h(\cdot)$ randomly, generate random batches from $\mathcal{D}^{\text{in}}_{\text{train}} \bigcup \mathcal{D}^{\text{out}}_{\text{train}}$, oversampling $\mathcal{D}^{\text{out}}_{\text{train}}$ by Γ.
for $\mathsf{train\_step} = 1$ **to** $\mathsf{max\_step}$ **do**
    loss $= \mathsf{CrossEntropy}(h(f(\boldsymbol{x})), y)$
    SGD update of $h(\cdot)$ w.r.t loss
end for

Algorithm 2 Few-shot outlier inference

Input: In-distribution test set $\mathcal{D}^{\text{in}}_{\text{test}} = \{(\boldsymbol{x}, y)\}$ with K classes, out-of-distribution test subset $\mathcal{D}^{\text{out}}_{\text{test}} = \{(X, y)\}$ with O classes, a pre-trained $f(\cdot) : \boldsymbol{x} \to \boldsymbol{z}$ from inputs to embedding vectors, a trained classification head $h(\cdot) : \boldsymbol{z} \to \mathbf{p} \in \mathbb{R}^{K+O}$.
Compute $\text{score}^{\text{in}}_{\text{oe}}(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{D}^{\text{in}}_{\text{test}}$
Compute $\text{score}^{\text{out}}_{\text{oe}}(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{D}^{\text{out}}_{\text{test}}$
Compute AUROC based on the scores.

C.2. Visualizing the effect of outlier exposure on fine-tuned pre-trained models

Figure 9 shows the outlier score on near-OOD task. Intuitively, the embeddings obtained by fine-tuning a pre-trained transformer are well-clustered, so just a handful of known outliers can significantly improve OOD detection.

D. Pre-training for genomics OOD

We first pre-train the transformer model in an unsupervised fashion as in BERT to capture biologically relevant properties. For unlabeled in-distribution sequences in the training set, we randomly mask the characters in the sequence at the rate of 0.15, feed the masked sequence into transformer-based model of 8 heads and 6 layers and embedding dimension 512, and predict the masked characters. To boost the performance, we add the unlabeled validation data to the training set for pre-training. In the fine-tuning stage, we load the pre-trained transformer model, mean pool the embeddings over the positions, and add a single linear projection classification head for 10 in-distribution classes on top of the embeddings. The setup is shown in Figure 10a. All the parameters in the model including those in the pre-trained transformer and those in the classification head are fine-tuned using the labeled training data. The model is pre-trained for 300,000 steps using learning rate of 0.001 and Adam optimizer [13] on TPU, and the accuracy for predicting the masked token is 48.35%. The model is fine-tuned for 100,000 steps at the learning rate of 0.0001, and the classification accuracy is 89.84%. We use the validation in-distribution and validation OOD data to select the best model checkpoint for each of the two methods and evaluate on test set.

The training set only contains genomic sequences of in-distribution classes. The validation and test sets contain sequences from both in-distribution and OOD classes. The genomic sequence is of fixed length of 250 base pairs, composed by characters of A, C, G, T.

We also study the relationship between the genetic distance and the AUROC of OOD detection for the 60 test OOD classes. We compute the genetic distance using the popular alignment-free method d_2^S which is based on the similarity between the word frequencies of the two genomes [24; 23]. Studies have shown that this genetic distance reflects true evolutionary distances [4; 2]. For each of the 60 OOD test classes, we use the minimum genetic distance between this OOD class to any of the 10 in-distribution classes as the final distance measure. Figure 10 shows the AUROC and the minimum distance for each of the 60 OOD classes.

Table 7: ImageNet-21k pre-trained ViT (some of them fine-tuned on in-distribution), with an additional final layer that was trained using the in-distribution train set and a small number of examples of the OOD train set (including the O OOD class labels, corresponding to red curves in Figure 3).

(a) CIFAR-100 → CIFAR-10 AUROC results.

Number of OOD (CIFAR-10) examples per class	R+ViT (without fine-tuning)	R+ViT fine-tuned on CIFAR-100
1	88.73 ± 1.08%	98.70 ± 0.08%
2	92.94 ± 0.55%	99.02 ± 0.15%
3	93.25 ± 0.59%	99.16 ± 0.11%
10	95.73 ± 0.31%	99.46 ± 0.01%
100	97.70 ± 0.01%	99.67 ± 0.01%

(b) CIFAR-10 → CIFAR-100 AUROC results.

Number of OOD (CIFAR-100) examples per class	R+ViT (without fine-tuning)	R+ViT fine-tuned on CIFAR-10
1	94.35 ± 0.05%	98.96 ± 0.05%
2	95.10 ± 0.30%	99.11 ± 0.04%
3	95.60 ± 0.01%	99.17 ± 0.03%
10	96.42 ± 0.02%	99.29 ± 0.02%
100	97.38 ± 0.01%	99.50 ± 0.01%

Table 8: Genomics OOD detection using pre-trained BERT and fine-tuned on the in-distribution training set. The error bars shown are standard deviations over 3 runs.

Model	Test Accuracy	Mahalanobis			MSP		
		AUROC↑	AUPRC↑	FPR95↓	AUROC↑	AUPRC↑	FPR95↓
1D CNN [25]	85.93±0.11%	64.75±0.73%	60.25±0.82%	77.76±0.84%	65.84±0.46%	62.24±0.31%	89.79±0.18%
BERT pre-train and fine-tune	89.84±0.00%	77.49±0.04%	78.79±0.06%	68.22±0.13%	73.53±0.03%	73.86±0.03	85.39±0.07%

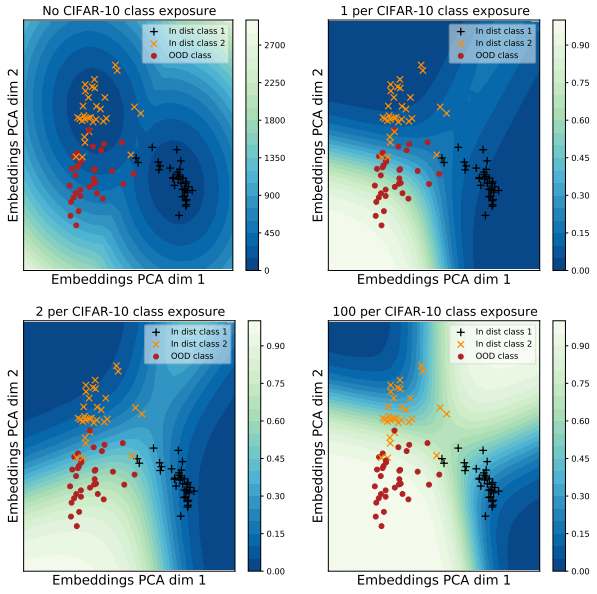


Figure 9: The effect of outlier exposure (CIFAR-10) on ViT a model fine-tuned on CIFAR-100. Each panel shows the same two-dimensional PCA cut of the space of embeddings. Examples of 2 in-distribution (CIFAR-100) classes (black plus signs = "bus", yellow crosses = "pickup truck"), and 1 out-of-distribution (CIFAR-10) class (red dot = "automobile"). The first plot shows Mahalanobis OOD score as the color coding. The more OOD outlier exposure (second to left to right), the better aligned the OOD probability contours with the underlying classes.

We expect the AUROC is higher as the distance is greater. Using the baseline model 1D CNN, we did not see obvious correlation between the AUROC and the minimum distance, with $r^2 = 0.0000$, based on Mahalanobis distance. The AUROC based on MSP method has positive correlation to the minimum distance, with $r^2 = 0.1190$. After we use the pre-trained+fine-tuned transformer, both MSP and Mahalanobis distance methods have significantly higher AUROC overall, and the positive correlation between the minimum distance and the AUROC are more prominent than the baseline model. We report additional results for genomics in Table 8.

Outlier exposure for genomics. Given that pre-trained and fine-tuned model improves the OOD performance, we next explore the idea of few shot outlier exposure to further boost the performance. We randomly select 1, 2, 5, 10, 100 examples per test OOD class and add them to the training set respectively. For each input $\boldsymbol{x}$ in the training set, we extract its corresponding embedding vector $\boldsymbol{z}$ from the above pre-trained and fine-tuned model (or alternatively the model without fine-tuned). We construct a single layer perceptron network of 1024 units for classifying each individual in-distribution classes and OOD classes, as shown in Figure 2. At inference time, we use the sum of the probability of in-distribution classes as the final confidence score for OOD detection. Additionally, we also tried the idea of collapsing all OOD classes into one single class (as in [28]) for comparison. The model is trained for 10,000 steps with the learning rate of 0.001. The best model checkpoint is selected based on the highest AUROC on a small set of validation dataset disjoint from the test set. The results are shown in Table 9.

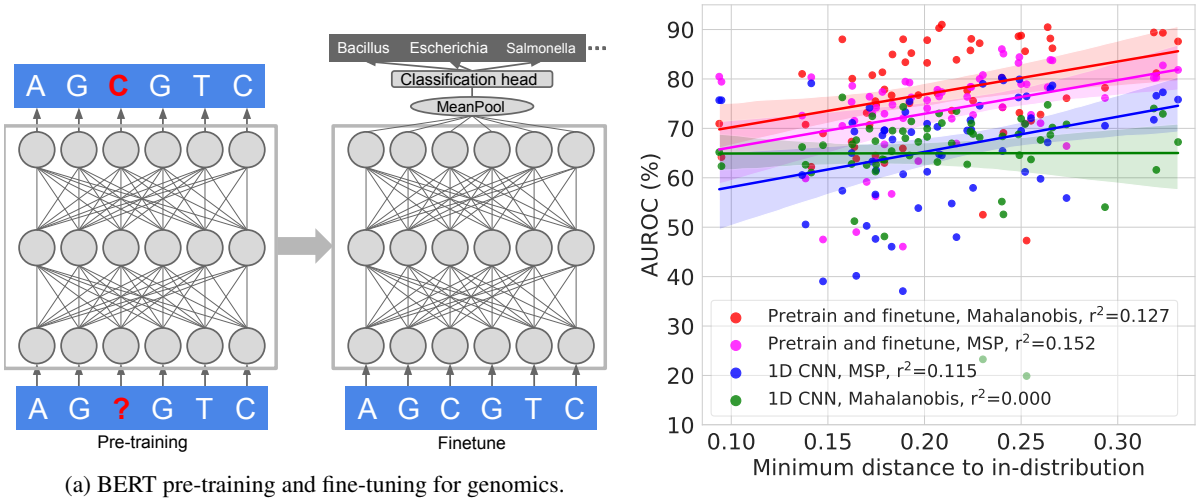
Table 9: Few-shot outlier exposure for genomics OOD. The numbers here correspond to Figure 3. The error bar is the standard deviation of 3 runs.

# OOD examples per class	w/o fine-tuning ignore outlier label	w/o fine-tuning use outlier label	w/ fine-tuning ignore outlier label	w/ fine-tuning use outlier label
1	75.32±1.14%	74.95±0.87%	84.36±0.60%	82.21±1.10%
2	75.67±0.50%	76.01±0.19%	84.81±0.83%	83.74±0.47%
5	77.84±1.06%	77.67±0.71%	86.54±0.58%	85.94±0.28%
10	77.81±1.37%	79.00±0.69%	87.02±0.45%	86.49±1.43%
100	80.21±0.14%	82.17±0.33%	87.94±0.62%	88.49±0.51%

E. Multi-modal zero-shot OOD detection

In the zero-shot classification setting, the candidate labels are chosen to describe the semantic content of the in-distribution classes (e.g. names of the classes). We propose to include the candidate labels related to the out-of-distribution classes, and utilize this knowledge as a very weak form of outlier exposure in multi-modal models. This could be relevant in applications, where we might not actually have any outlier images for fine-tuning but we might know the *names* or *descriptions* of outlier classes.

We produce an embedding vector $\boldsymbol{z}$ for each image I, apply softmax to get probabilities as $\mathbf{p} = \mathrm{softmax}(\boldsymbol{z})$. Those get split to $p(\mathrm{in}|\boldsymbol{x}) = \sum_{i\in\mathrm{in}} \mathbf{p}_i$, and $p(\mathrm{out}|\boldsymbol{x}) = \sum_{i\in\mathrm{out}} \mathbf{p}_i$, where $p(\mathrm{in}|\boldsymbol{x}) + p(\mathrm{out}|\boldsymbol{x}) = 1$. Similar to Figure 2, we use $\mathrm{score}_{\mathrm{oe}}(\boldsymbol{x}) = p(\mathrm{in}|\boldsymbol{x})$ as the confidence score. By choosing the candidate labels to represent the in-distribution and OOD dataset we would like to distinguish (e.g. CIFAR-100 and 10), we can get a very informative score that leads to AUROC above previous SOTA, despite no exposure to the training set of the in-distribution (zero-shot).



(a) BERT pre-training and fine-tuning for genomics.

(b) AUROC of near-OOD detection

Figure 10: (a) Model architecture for BERT pre-training and fine-tuning. The unsupervised pre-training model uses a transformer encoder to predict the masked token (shown in red). The fine-tuned model adds a simple classification head (MLP with one hidden layer) to predict in-distribution classes. (b) The relationship between the minimum genetic distance and the AUROC scores for the 60 OOD test classes. The pre-train+fine-tune based MSP and Mahalanobis distance methods have significantly higher AUROC overall, and the positive correlation between the minimum distance and the AUROC are more prominent than the baseline model.

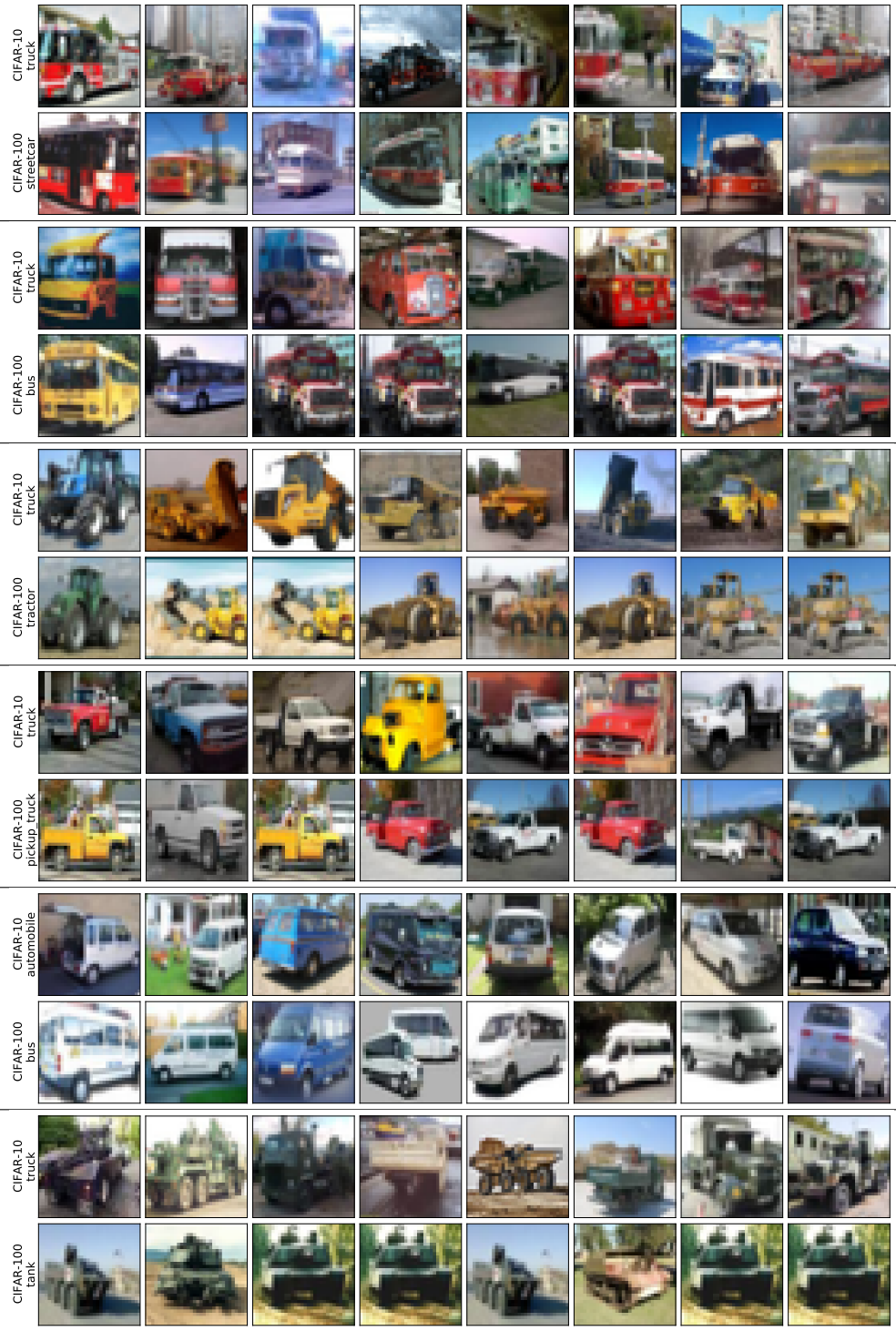


Figure 11: CIFAR-100 (in-distribution) vs CIFAR-10 (out-distribution), hard to distinguish pairs of class involving **vehicles**: The plots shows pairs of images, top row from the out-distribution (CIFAR-10) and the bottom row having the closest embedding vector to it from the in-distribution (CIFAR-100), for pairs of classes for which OOD images are the closest to the in-distribution (=hardest to distinguish). Our examples show that there is a genuine semantic overlap between some CIFAR-10 and CIFAR-100 classes that limits the OOD performance of any model.

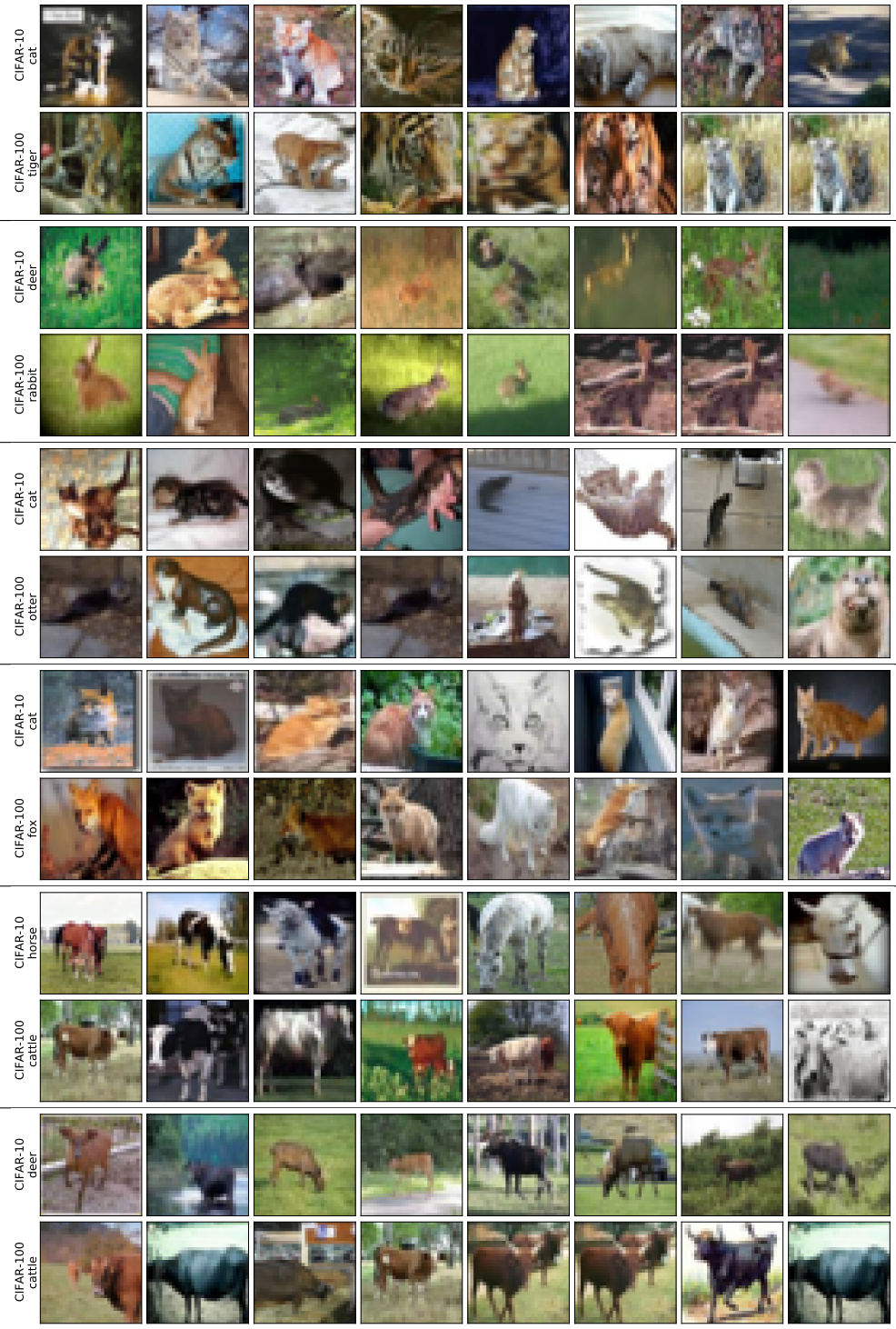


Figure 12: CIFAR-100 (in-distribution) vs CIFAR-10 (out-distribution), hard to distinguish pairs of class involving **animals**: The plots shows pairs of images, top row from the out-distribution (CIFAR-10) and the bottom row having the closest embedding vector to it from the in-distribution (CIFAR-100), for pairs of classes for which OOD images are the closest to the in-distribution (=hardest to distinguish). Our examples show that there is a genuine semantic overlap between some CIFAR-10 and CIFAR-100 classes that limits the OOD performance of any model.