A. Detailed Information of Models (Sec. 6)

In the experiment, we benchmarked a collection of deep classification models. Their detailed information, including the specification, the architecture, the pre-train information, and the top-1 accuracy, is listed in Tab. 5. To summarize, half of them are CNN-based, and half are transformer-based. Vision Transformer and Swin Transformer are pre-trained on ImageNet-21K before training on the ImageNet-1K.

B. Detailed Results of Four Models (Sec. 6.3)

In Sec. 6.3 we gave the average AUROC and FPR95 for RepVGG, ResNet-50d, Swin Transformer and DeiT. We provided the detailed AUROC and FPR95 on OpenImage-O, Texture, iNaturalist, and ImageNet-O in Tab. 6.

C. Details on OpenImage-O (Sec. 5)

An illustrative software interface for labelers is shown in Fig. 6. For each candidate OOD image to be labeled, we find the top 10 classes in ImageNet-1K predicted by a classification model. Then we gather the most similar images in those top 10 classes by cosine similarity in the feature space. Next, we patch them as well as their labels with the corresponding OpenImage samples, and let the labelers distinguish whether the OpenImage sample belongs to any of the top 10 categories. We also set a choice called difficult, so that labelers can put the undistinguishable hard samples into the difficult category. To reduce annotation noises, each image is labeled twice from different group of labelers. Then we take the set of OOD images having consensus from the two groups, resulting in an OOD dataset with 17,632 unique images. In the end, a random inspection process is performed to guarantee the quality of the OOD dataset.

The OpenImage-O follows a natural image distribution as both the source dataset and the labeling process do not involve any filtration based on pre-defined list of labels. To get a sense of its distribution, we use the BiT model to find the most similar ID class in ImageNet for each OOD image. Then the histogram is illustrated in Fig. 7. It shows that the coverage of OpenImage-O is broader compared to the other three OOD datasets.

D. Details on Grouping (Sec. 6.5)

MOS [7] is trained using the officially released code and its default parameter setting. For all experiments in Sec. 6.5, the grouping strategy follows the taxonomy grouping defined in [7].

Grouping Results on ViT

The grouping strategy is less effective for the ViT model, as seen from results in Tab. 7. Comparing MSP with its group version, MaxGroup, we can see that the improvement on AUROC is very small, while FPRs become even worse. Examining ViM with its group variant ViM+Group, we can see that their difference is very small, and the original version of ViM is slightly better than ViM+Group.

E. Details on Baselines (Sec. 6)

Mahalanobis

On the BiT model, when including lower level features, the performance of Mahalanobis degrades a lot. The average AUROC on the four OOD datasets is 56%, which is much worse than the baseline MSP. Similar results is also found in [7, Table 1]. In this paper, we implement the
<table>
<thead>
<tr>
<th>Model</th>
<th>Specification</th>
<th>Architecture</th>
<th>Pre-Trained Dataset</th>
<th>Top1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiT [8]</td>
<td>BiT-S-R101x1 CNN</td>
<td>CNN</td>
<td>—</td>
<td>81.30</td>
</tr>
<tr>
<td>ViT [2]</td>
<td>ViT-B/16 Transformer</td>
<td>Transformer</td>
<td>ImageNet-21K</td>
<td>85.43</td>
</tr>
<tr>
<td>RepVGG [1]</td>
<td>RepVGG-b3 CNN</td>
<td>CNN</td>
<td>—</td>
<td>80.52</td>
</tr>
<tr>
<td>Swin [12]</td>
<td>Swin-base-patch4-window7-224 Transformer</td>
<td>ImageNet-21K</td>
<td>85.27</td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Detailed information on the used models. The detailed specification and the top-1 accuracy of the model are provided. Three of them are CNN-based, and the other three are transformer-based. Both ViT and Swin Transformer are pre-trained on ImageNet-21K before training on ImageNet-1K, so their general OOD performances are much better than alternatives.

Figure 7. The diversity of four OOD datasets shown by how they look similar to the ImageNet-1K classes. We use the BiT model to predict which ID class the image most resembles, and count the number of such OOD images for each class. Results are shown above. Due to space limitation, the y-axis is clipped at 155. Our newly created OpenImage-O has a wider coverage on ImageNet ID classes.
<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>Source</th>
<th>OpenImage-O</th>
<th>Texture</th>
<th>iNaturalist</th>
<th>ImageNet-O</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>AUROC</td>
<td>FPR95↑</td>
<td>AUROC</td>
<td>FPR95↑</td>
<td>AUROC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MODEL</td>
<td>METHOD</td>
<td>SOURCE</td>
<td>OOD</td>
<td>Detection</td>
</tr>
</tbody>
</table>
Mahalanobis score using the feature vector before the final classification fc layer, as in [3]. The precision matrix and the class-wise average vector are estimated using 200,000 random training samples. The ground-truth class label is used during computation.

**KL Matching** We estimate the class-wise average probability using 200,000 random training samples. Following the practice of [5], the predicted class is used instead of ground-truth labels. We would like to note that the hyperparameter selection for OOD methods should not base on the ID set that is used for computing FPR95 and AUROC (in our case, its the validation set of ImageNet), because once the OOD method overfits the validation set, the evaluation result can be higher than the actual performance.

**ReAct** For ReAct, we use the Energy+ReAct setting, which is the most effective settings in [13]. In the original paper, they recommended the 90-th percentile of activations estimated on the ID data for the clipping threshold. However, for BiT and ViT, we found that the rectification percentile $p = 99$ works much better than 90. So we report results using $p = 99$.

**F. OOD Examples Detected by KL Matching and Residual (Sec. 3)**

In Sec. 3, we showed that feature-based OOD scores (e.g., Residual) and logit/softmax-based OOD scores (e.g., KL Matching) have different performances on the Texture OOD dataset. Here we visualize the OOD examples found by the two methods in Fig. 8.

### G. Running Time of Four Methods (Sec. 6.2)

From Tab. 2 and Tab. 6, it is clear that the four most competitive methods are ViM, Mahalanobis, KL Matching, and Residual. Our ViM is the fastest among all four methods. We show their inference time on the four datasets in Tab. 8.

### References


Figure 8. OOD examples detected by Residual and KL Matching. There are three rows for each OOD dataset. The first row shows images from the top 5% OODs detected by Residual, with overlapping images in the top 50% list of KL Matching removed. The second row displays images from the intersection of the top 5% OODs detected by Residual and the top 5% OODs detected by KL Matching. The third row shows images from the top 5% OODs detected by KL Matching, with overlapping ones in the top 50% list of Residual removed.


