Distance Matters For Improving Performance Estimation Under Covariate Shift

Mélanie Roschewitz
Imperial College London
mb121@ic.ac.uk

Ben Glocker
Imperial College London
b.glocker@imperial.ac.uk

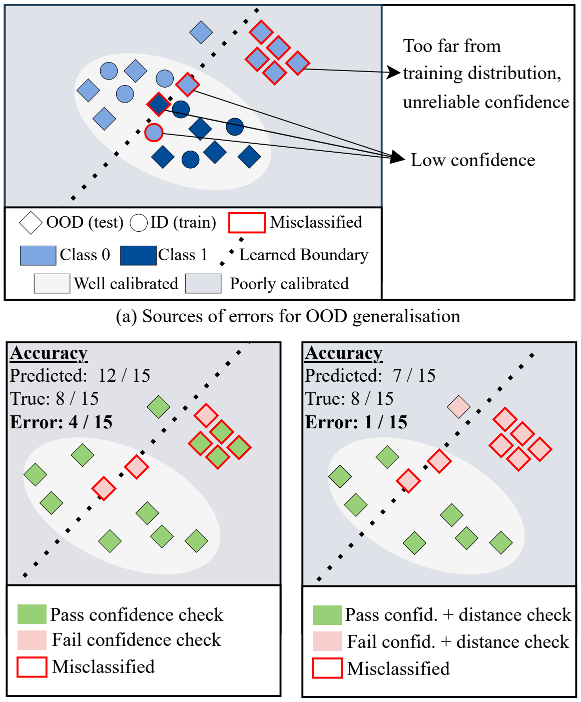
Abstract

Performance estimation under covariate shift is a crucial component of safe AI model deployment, especially for sensitive use-cases. Recently, several solutions were proposed to tackle this problem, most leveraging model predictions or softmax confidence to derive accuracy estimates. However, under dataset shifts confidence scores may become ill-calibrated if samples are too far from the training distribution. In this work, we show that taking into account distances of test samples to their expected training distribution can significantly improve performance estimation under covariate shift. Precisely, we introduce a "distance-check" to flag samples that lie too far from the expected distribution, to avoid relying on their untrustworthy model outputs in the accuracy estimation step. We demonstrate the effectiveness of this method on 13 image classification tasks, across a wide-range of natural and synthetic distribution shifts and hundreds of models, with a median relative MAE improvement of 27% over the best baseline across all tasks, and SOTA performance on 10 out of 13 tasks. Our code is publicly available at https://github.com/melanibe/distance_matters_performance_estimation.

1. Introduction

Machine learning models are sensitive to variations in their deployment environments [17, 49, 35, 23, 45, 33]. Due to the unavailability of ground truth labels for continuous performance monitoring at deployment time, real-time and accurate performance estimation is crucial to detect any unexpected behavior or model failure, particularly in distribution-shifted settings. This is especially important for sensitive use cases such as clinical decision making.

The difficulty in estimating model performance arises from the lack of reliability of model outputs under covariate shift [35, 25]. Recently, several attempts have been made at addressing this problem, many of them based on confidence estimates [15, 14, 29]. For example, Average Thresholded Confidence (ATC) [14] leverages softmax outputs for estimating classification accuracy, considering that all outputs whose confidence do not reach a certain threshold are incorrectly classified. While this method has shown to be effective at estimating the performance under mild shifts (e.g. on synthetically corrupted images), experiments show that the method under-performs in more substantial shifts such as natural sub-population shifts. In particular, current approaches tend to *overestimate* accuracy in natural real-world distribution shift settings [14, 20]. This can notably be explained by a deterioration of model calibration when going further from the training distribution [22], with softmax outputs becoming over-confident and unreliable [35].



(a) Sources of errors for OOD generalisation

(b) Confidence-based only accuracy estimation (e.g. [14])

(c) Confidence-based + distance-based accuracy estimation (ours)

Figure 1. **Performance estimation under covariate shift needs to take into account different sources of errors.** Distance to the source distribution in the embedding space matters as confidence estimates become unreliable with increased distance.

If test samples are too far from training samples, relying on the output of the classification layer for performance estimation is insufficient. From an uncertainty point of view, softmax outputs can been seen as capturing aleatoric uncertainty, arising from overlapping class boundaries [21, 9]. However, under dataset shifts, errors may also arise from the fact that the model has never seen this type of input data and does not know how to respond to such inputs. This is referred to as epistemic uncertainty [9] and is not well captured by softmax outputs [21, 43], as demonstrated by its poor performance on the related out-of-distribution (OOD) detection task [35, 25, 41, 43, 27]. Note that in OOD detection, the goal is to separate on separating ID from OOD inputs, regardless of the downstream classification performance, often considering inputs completely unrelated to the task. This differs from performance estimation under covariate shift, where we assume that the classification task still applies to the shifted inputs and we focus on estimating performance, not on detecting shifts.

Methodological contributions In this paper, we argue that performance estimators should identify samples far away from the training set in the embedding space, for which softmax estimates are most likely unreliable. By measuring the distance in the embedding space, we are able to measure how well the model "understood" the sample when projecting the input to the classification space. This idea is illustrated in fig. 1. Following this intuition, we propose a simple yet effective method to improve the quality of current SOTA performance estimators. Specifically, we use nearest-neighbours distance in the embedding space to reject samples that lay too far from the training distribution. We then only use confidence-based performance estimators on the remaining samples, considering all previously rejected samples as mis-classified. Our distance check approach is versatile and can be used to improve the quality of various existing performance estimators (e.g. [14, 20]).

Main results We evaluate our approach on 13 classification tasks ranging from cancer cell to animal classification. The nature of the distribution shifts studied covers a wide-range of shifts: from synthetic corruption, acquisition shift, real-world population shift to representation shift. For each task we evaluate between 18 and 259 models, covering various training strategies and network architectures. These experiments demonstrate that integrating distance into accuracy estimators significantly improves the quality of the estimation. For example, our proposed estimator ATC-DistCS is significantly better than previous SOTA ATC [14] on all but one task, with a median relative MAE improvement of 30% across all tasks. Furthermore, comparing to the most recent COT method [30], we demonstrate a 27% median relative performance improvement across all tasks, with new SOTA performances on 10 out of 13 tasks. We also demonstrate significant improvements across all datasets for agreement based accuracy estimation when integrating our distance check. Ablation studies yield further insights in the method and its limitations. Finally, to the best of our knowledge, we provide the first comprehensive publicly available codebase of current SOTA baselines for accuracy estimation, along with the complete code to reproduce our experiments.

2. Background

2.1. Performance estimation without ground truth

Current methods for performance estimation under covariate shift can be broadly grouped in 4 categories:

Estimating performance via auxiliary task performance Modifies the main classification model to incorporate a (sufficiently informative) auxiliary task for which ground truth labels are available at test time: accuracy on the main task is then approximated by the computed accuracy on the auxiliary task. For example, [11] trains a multi-task model for predicting the class at hand as well as the rotation applied to the input image. The main limitation of this line of work is the requirement to build a multi-task model, making it unusable as a post-hoc tool.

Training a regressor between ID and OOD accuracy This regressor can be trained based on model outputs or on measures of distance between datasets [13, 40, 15, 31, 12]. One major drawback of this class of estimators is their requirement for having access to labelled OOD data for training the regression model. This is not always available in practice, in particular in data-scarce domains such that healthcare. In absence of such OOD datasets, regressor are sometimes trained using corrupted versions of the original validation set as "OOD" sets. However, this can not guarantee the robustness of this estimator against other shifts e.g. natural subpopulation shift [39].

Agreement-based estimators Are based on the idea that agreement between member of model ensembles correlate with model accuracy. For example, generalised disagreement equality (GDE) [20] use pairs of models trained with different random seeds to compute disagreement. Others use more intricate methods for training specialised models to align disagreement and accuracy further [10, 6]. However, these procedures often require expensive additional training steps to derive the siblings models and are not applicable to post-hoc scenarios where only the final model is available to the end user. In [1], the authors go as far as training dozens models to fit a regressor between agreement and accuracy, whereas [6] requires training a new ensemble for every single test set requiring performance estimation.

Confidence-based estimators These methods, contrarily to the ones above, only require the final model's outputs to perform accuracy estimation and do not require any OOD data for calibration. As such, they are versatile and can be used with any classification model. For example, Difference of Confidence (DOC) [15] approximates the difference in accuracy between the evaluation set and the in-distribution (ID) validation set by the difference in average model confidence. ATC [14] introduces a confidence threshold such that all test samples for which the confidence is lower than this threshold are considered wrong and all samples meeting the minimum confidence requirement are considered correct (see Methods). Finally, concurrently to our work, COT [30] proposed to estimate accuracy based on based on optimal transport of model confidences. Precisely, they measure the Wasserstein distance between source label distribution and target softmax distribution to estimate the test error. Note, that this is expected to perform well if the source label distribution matches the target label distribution but might fail if this assumption breaks.

2.2. Distance-based out-of-distribution detection

The idea that OOD samples should lie far from the training samples in the embedding space is at the core of distance-based methods for OOD detection. For example, [27] propose to fit multi-variate Gaussians on the training embedding distribution of each class and use the Mahalanobis distance [32] to characterise how far test samples are from this expected distribution. If a sample is far from all class clusters, it is considered OOD. This method has shown some success at various OOD detection tasks [4, 27] and extensions of this work have since further improved its capabilities [37]. However, this method suffers from one major limitation: it has a strong assumption that the class embeddings clusters can be accurately modelled by a Gaussian Multivariate distribution. Without any constraints on the training procedure or the embedding space at training time, this assumption may not hold [41]. This is the motivation for the work of [41] who proposed a non-parametric alternative OOD detection method. The authors still focus on the idea of using distances in the embedding space to detect OOD samples, but they leverage nearest-neighbours distances instead of the Mahalanobis distances, removing the normality assumption on the embedding. Precisely, they use the distance to the Kth nearest-neighbour to classify samples as OOD. They derive the classification threshold for OOD versus ID task such that 95-99% of the training samples are classified as ID.

3. Methods

In this section, we begin by reminding the reader of the core principles of two base performance estimators which we build on top of: ATC [14] and GDE [20]. We then introduce our plug-in distance checker designed to flag untrustworthy samples, and discuss how to incorporate this distance check into these performance estimators to yield our proposed estimators "ATC-DistCS" and "GDE-DistCS".

3.1. Base estimators

Average Thresholded Confidence (ATC) [14] approximates accuracy by the proportion of OOD predictions that do not exceed a certain confidence threshold (derived from the ID validation set, where confidence is defined as temperature-scaled [16] softmax confidence). Precisely, the threshold ATC is defined such that on source data D_s the expected number of points that obtain a confidence less than ATC matches the error of the model on D_s. This method has been further refined by [29], where the authors propose to apply class-wise temperature scaling and to define class-wise confidence thresholds to improve the quality of the estimation, in particular for class-imbalanced problems.

Generalised Disagreement Equality (GDE) [20] Assuming access to two models g and g' trained with different random seeds (but identical architecture and training paradigms), GDE estimates model accuracy by $\frac{1}{N}\sum_{i\in\text{test set}}[g(x_i) = g'(x_i)]$, where N is the size of the OOD test set, x_i the inputs to the model, and $g(x_i)$ denotes the model prediction.

3.2. Integrating distance to training set

Average Distance Check Inspired by the OOD detection work in [41], we propose to improve standard performance estimators with a "distance checker". Instead of simply rejecting samples with low confidence or model disagreement, we argue that the distance of any given sample to the in-distribution training set should also be taken into account to determine whether its confidence (and prediction) is likely to be trustworthy for estimation purposes or not. In simple terms, we "reject" samples whose penultimate-layer embeddings lie in a region "far" from the ID embedding space. The distance from a sample to the in-distribution set is determined by the average distance between the sample and all of its K-nearest-neighbours:

$$\text{AD}_i = \frac{1}{K}\sum_k \left\| f_i - n_i^{(k)} \right\|_2, \qquad (1)$$

where K is the number of nearest-neighbours to consider, f_i is the embedding of the i^{th} test sample and $n_i^{(k)}$ the k^{th} nearest neighbours to f_i in the embedding space, nearest neighbours are searched for in the training set. The acceptable threshold is determined on the in-distribution validation set as the 99^{th}-percentile of the average distances observed on this set i.e.

$$\text{DistThreshold} = \text{quantile}_{.99}\left\{AD_i, \forall i \in \text{val set}\right\}. \quad (2)$$

Note, that our distance criterion differs from [41], in that (i) we use the average of all K distances instead of the distance to the K^{th} neighbour only (to be less sensitive to outliers); (ii) we do not normalise the embeddings (see ablation study in section 4); (iii) we do not use a contrastive loss for training our models as this assumption may restrict the scope of application of the method. The fitting procedure for the distance checker can be found in algorithm 1.

Using distance to improve the quality of performance estimators Our proposed "distance-checker" can be used as a plug-in method to improve the estimation results of different existing accuracy estimators. Specifically, first we propose "ATC-Dist", where we combine both criteria to estimate the accuracy under shift: a sample is estimated as being correct if it is (i) of high enough confidence, (ii) not too far from the in-distribution embeddings. Similarly we extend GDE with our distance criterion to get "GDE-Dist". There the correctness of a sample is estimated by (i) agreement between both models, and (ii) distance to the in-distribution embeddings. The estimation procedure for ATC-Dist and GDE-Dist is shown in algorithm 1.

Class-wise distance thresholds As the tightness of class clusters may differ for different classes, we argue that the quality of the distance threshold can be further improved by defining class-wise distance thresholds. Concretely, for each class c we compute DistThreshold_c by taking the 99^{th} percentile of the average distance distribution of the subset of cases labelled as c in the validation set. At test time, we use the distance threshold associated with the predicted class to determine the validity of a given sample prediction. In cases where less than 20 samples were present in the validation set for any given class, we use the global threshold for this class. Replacing the global distance threshold by the classwise thresholds in the procedure described above, yields our proposed "ATC-DistCS" and "GDE-DistCS" estimators.

4. Results

Motivating example Prior to diving into quantitative analysis of our results, let's start with an illustrative example of our main idea: *in the embedding space, regions of the shifted test set not covered by the ID set are likely to be regions of very low accuracy*. This pattern appears distinctively in the example in fig. 2, where we show the T-SNE [44] representation of the embeddings of the ID validation set as well as the OOD test set, on a model trained on the WILDS-CameLyon [23, 2] dataset. We can clearly see how the region in black – which is not well represented in the ID validation set – contains an extremely high proportion of errors in the OOD test set.

Algorithm 1

```
procedure FIT DISTANCECHECKER($X_{train}$, $X_{val}$)
    $f_{train} \leftarrow$ GET FEATURES($X_{train}$)
    $f_{val} \leftarrow$ GET FEATURES($X_{val}$)
    KNN $\leftarrow$ FIT NEAREST NEIGHBORS($f_{train}$)
    $\text{AD}_{val} \leftarrow$ AVERAGE NN DISTANCES(KNN, $f_{val}$)
    DistThreshold $\leftarrow$ QUANTILE($\text{AD}_{val}$, .99)
    return DistThreshold, KNN
end procedure

procedure GET ATC-DIST($X_{\text{test}}$, ATC, KNN, DistThreshold)
    $f_{\text{test}} \leftarrow$ GET FEATURES($X_{\text{test}}$)
    $\text{AD}_{\text{test}} \leftarrow$ AVERAGE NN DISTANCES(KNN, $f_{\text{test}}$)
    $\text{kept}_{\text{Dist}} \leftarrow \text{AD}_{\text{test}} <$ DistThreshold    ▷ Distance check
    $\text{c}_{\text{test}} \leftarrow$ SOFTMAX CONFIDENCE($X_{\text{test}}$)
    $\text{kept}_{\text{ATC}} \leftarrow \text{c}_{\text{test}} >$ ATC    ▷ Confidence check
    ATC-DIST $\leftarrow \frac{|\text{kept}_{ATC} \cap \text{kept}_{\text{Dist}}|}{|X_{\text{test}}|}$
    return ATC-DIST
end procedure

procedure GET GDE-DIST($X_{\text{test}}$, $g_1$, $g_2$, DistThreshold, KNN)    ▷ $g_1$, $g_2$ two models trained with different seeds
    $f_{\text{test}} \leftarrow$ GET FEATURES($g_1$, $X_{\text{test}}$)
    $\text{AD}_{\text{test}} \leftarrow$ AVERAGE NN DISTANCES(KNN, $f_{\text{test}}$)
    $\text{kept}_{\text{Dist}} \leftarrow \text{AD}_{\text{test}} <$ DistThreshold    ▷ Distance check
    $\text{agree}_{\text{test}} \leftarrow g_1(X_{\text{test}}) = g_2(X_{\text{test}})$    ▷ Agreement
    GDE-DIST $\leftarrow \frac{|\text{agree}_{\text{test}} \cap \text{kept}_{\text{Dist}}|}{|X_{\text{test}}|}$
    return GDE-DIST
end procedure
```

Datasets We validate our proposed method on a wide range of tasks and covering various natural and synthetic distribution shifts (more details in supplement):

- ImageNet [38] to ImageNet-Sketch [45] where the distribution shifts from photographs to sketch; ImageNet-A [19], where the distribution shifts adversarially; ImageNet-V2 [36] a setting with only mild shifts, designed to mimic ImageNet test set.

- CIFAR10 [24] to CIFAR10-C [18] covering various synthetic corruptions, yielding 95 OOD datasets.

- MNIST [26] to SVHN [34], classic digit classification, shifting from binary digit images to house numbers.
- WILDS [23] benchmark, designed to study natural shifts occuring "in the wild". WILDS-Camelyon17 [2] defines a histopathology binary task, with staining protocol shifts. WILDS-iCam [3] is a 182-classes animal classification task from camera traps, with shifts in camera location. WILDS-FMoW [8] is a satellite image 62-class task, with temporal and geographical shifts. WILDS-RxRx1 [42] is a 1,139 genetic treatment classification task on fluorescent microscopy images, where the shift occurs from so-called experimental "batch-effect".
- The BREEDS [39] benchmark defines various tasks (Entity30, Entity13, Living17, NonLiving26) based on ImageNet subsets and superclasses. The main task consisting of predicting the super-class and the train-val-test split defined such that the subpopulations covered by the OOD test set are disjoint from the ones represented in the training and validation set.
- PACS [28] a 7-class task, where models are trained and validated on photographs and tested on 3 other domains (painting, sketches and cartoon).
- PathMNIST [48] histopathology 9-class task, where training and test splits are taken from different sites.

Experimental setup and models For each evaluated model, we fit our nearest neighbours algorithm on the training set, using K=25 neighbours for the distance check. Distance thresholds are computed on the in-distribution validation set. For each task, we evaluate our accuracy estimator on all available OOD sets, as described above, and measure estimation quality in terms of Mean Absolute Error (MAE) between predicted and true accuracy across all models. Note that if there were more than N=50,000 training samples, we randomly subsampled N samples in the K-NN fitting step to speed up inference. Moreover, for ImageNet to avoid doing a full inference pass on the extremely large training set, we fitted the K-NN algorithm directly on the validation set (discarding distance to self to get the distance threshold). For each task, we evaluate the quality of performance estimation on a large variety of models. For ImageNet, we test on 259 pretrained models from the `timm` [46] package, covering a range of 14 family of model architectures. For all other datasets, we trained models ourselves using various architectures, training setups, random seeds and initialising models both from ImageNet and random weights (except for BREEDS datasets as they are build from ImageNet images, hence pretrained weights would violate the OOD assumptions of the testing subpopulations); amounting to 18 models for BREEDS tasks and 30 models for all other tasks. More details can be found in Supp Note 2 and in our codebase.

Choice of baselines Our first analysis focuses on single-model accuracy estimation. We compare our method to established ATC [14] and DoC [15] baselines as well as their improved class-wise version [29]. For class-wise estimation, if any given target class was not present in the validation set, or if less than 20 samples were predicted for that class, we used the global temperature and ATC-threshold for that particular class. This may happen for some classes in imbalanced datasets or with an extremely high number of classes (e.g. WILDS RxRx1 or WILS iCam). We also compare our method to the recently proposed COT estimator [30]. Note that, to date, in their pre-print, the authors only tested their method on a very limited set of tasks, as such our evaluation considerably extends the assessment of COT's capabilities. Regression-like methods are not included as we assume that no OOD dataset is available at training and validation time, similarly we do not include methods that require external metadata such as Mandoline [7] as it was not available. Methods such as self-training ensembles [6] which require model retraining for every single test set, were also not considered as they were computationally much more heavy (it would require training over 3,000 ensembles in our experimental setting) and do not allow for real time monitoring. Weaker baselines such that simply using the average softmax confidence as accuracy estimation are not included as the extensive analysis in the ATC paper [14] already clearly demonstrates the superiority of ATC as a baseline. In a second analysis, we place ourselves in the scenario where we have access to two models for each task for accuracy estimation and compare agreement-based estimator GDE [20] to our improved version GDE-DistCS.

Results for single-model performance estimation In table 1, we compare DoC, ATC, COT and our method in two different settings, one where temperature scaling (TS) [16] and ATC threshold are optimised globally for the entire dataset (left column group) and the second where we apply class-wise TS and ATC thresholds (rightmost columns). Temperature scaling is applied as previous studies have shown better results over raw model ouputs [29, 14]. Results are presented in terms of MAE over all shifted test sets and all models for each task. We can see that our method ATC-Dist achieves lower MAE than its counterpart ATC across all but one dataset (Wilds iCam, see discussion section). Furthermore, on all these datasets, ATC-DistCS using class-wise distance thresholds further improves the results over ATC-Dist. The overall median relative MAE reduction of ATC-DistCS over all datasets is of 30% compared to standard ATC in the global setting and 13% in

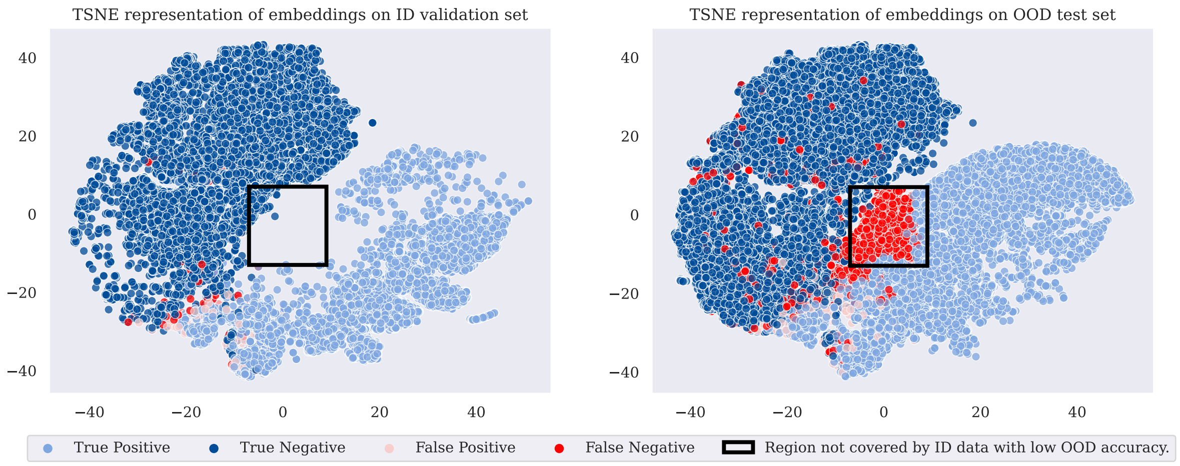


Figure 2. **Why distance matters, an example.** Joint TSNE [44] representation of the ID validation set and OOD test set plotted separately for a ResNet18 model on the WILDS CameLyon dataset. We can clearly distinguish a region with low density on the validation set and high density on the OOD set, where most points are misclassified.

	Global TS & ATC thresholds					Classwise TS & ATC thresholds		
Dataset family	DoC [15]	COT [30]	ATC [14]	ATC-Dist (ours)	ATC-DistCS (ours)	COT [30]	ATC [29]	ATC-DistCS (ours)
ImageNet-Sketch	19.26**	4.62**	6.05**	4.75**	**3.28**	5.31**	5.36**	**3.42**
ImageNet-A	38.66**	31.73**	26.81**	23.20**	17.95**	21.77**	35.67**	**15.10**
ImageNet-V2	4.94**	5.70**	1.90**	1.43**	**0.64**	1.83**	5.39**	3.51**
Living17	21.08**	18.95**	17.98**	15.86**	14.45*	20.18**	15.02**	**11.82**
NonLiving26	24.31**	21.38*	16.71**	15.65**	**14.53**	21.85*	15.84**	**13.87**
Entity13	13.55**	12.78**	8.96**	8.30*	**8.15**	12.99**	8.64**	**7.84**
Entity30	17.75**	15.45**	12.31**	11.65*	**11.31**	15.98**	12.15**	**11.15**
WILDS CameLyon	7.57**	3.07**	6.86**	4.90**	4.71**	**2.99**	6.82**	4.69**
WILDS iCam	**8.14**	7.72*	**7.15**	7.92*	9.13**	**6.48**	**5.39**	**6.95**
WILDS FMoW	3.54**	2.04**	2.72**	2.06*	**1.91**	1.94*	**1.36**	**1.58**
WILDS RxRx1	7.47**	**2.36**	6.02**	5.01**	3.86**	**2.54**	8.87**	9.62**
MNIST	61.41**	**15.17**	49.52**	**17.41**	**15.96**	**15.33**	41.44**	**16.12**
PACS	55.38**	**12.61**	45.98**	26.25**	26.21**	13.23*	49.45**	26.65**
PathMNIST	3.68**	9.90**	2.67**	**1.31**	**1.14**	9.92**	2.37**	**1.09**
CIFAR10	2.73**	1.53**	1.20**	1.11*	**1.08**	1.59**	1.24**	**1.07**

Table 1. **Improving confidence-based accuracy estimation - summary table.** Results are reported in terms of Mean Absolute Error (in %) across all models and OOD datasets. * denotes a p-value after Bonferroni correction <0.05, ** p-value <1e-3 for Wilcoxon signed-rank test [47] to test for difference between the best method versus all the others, for each dataset. Bold denotes the best model, all methods not significantly different from the best are highlighted.

the class-wise setting. Additionally, our experimental results confirm the preliminary findings of [29] i.e. ATC with class-wise optimisation of temperature and thresholds outperforms ATC with global optimisation for the majority of datasets (only performance on heavily imbalanced CIFAR10-C had been reported so far). To measure statistical significance, for each dataset, we use the Wilcoxon signed-rank test [47] to compare the best method (i.e. with lowest MAE on that task) against all other methods, with Bonferroni [5] correction to account for multiple testing. We highlight in bold the method with lowest MAE and all methods that are not significantly different to this method (at the level 0.05 after correction). We can see that ATC-DistCS achieves SOTA results on 10 out of 13 tasks, with a median relative MAE improvement of 27% for ATC-DistCS over COT with global TS and 30% with class-wise TS. We discuss differences between COT and ATC-DistCS in more details in section 5. Finally, we detail the performance comparison on CIFAR10-C in fig. 3, we can see that our method outperforms the baselines at all levels of corruption

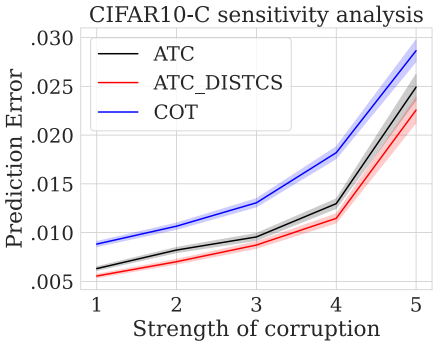


Figure 3. **Ablation study MSE in function of corruption strength for CIFAR10-C** across all models, shaded area depicts +/- one standard deviation.

Dataset family	GDE [20]	+DistCS (ours)
Living17	19.92**	**16.60**
NonLiving26	23.49**	**21.26**
Entity13	12.62**	**11.78**
Entity30	16.57**	**15.67**
WILDS CameLyon	4.96**	**3.62**
WILDS iCam	**6.44**	**5.93**
WILDS FMoW	9.39**	**8.56**
WILDS RxRx1	9.14**	**7.78**
MNIST	34.63**	**15.63**
PACS	42.41**	**27.76**
PathMNIST	2.81**	**1.62**
CIFAR10	4.73**	**4.19**

Table 2. **Results for improving agreement-based estimates.** Best in bold, * denotes a p-value <0.05, ** <1e-5 for Wilcoxon-signed-rank test [47] for GDE-DistCS against GDE.

strength. Additional scatterplots detailing predicted versus true accuracy can be found in Supp. Note 3.

Results for agreement-based accuracy estimation. Our distance check is not only tailored for improving ATC but rather is a general addition that can be "plugged-in" to various estimators. We demonstrate this by showing how our method improves the quality of agreement-based estimator GDE [20], another well established baseline. For every training configuration, we repeat training with 3 different seeds. To estimate the accuracy for model g_1, we use another model g_2 trained with a different seed to compute disagreement and deduce the predicted accuracy for g_1. We then further improve the estimation with our proposed distance check i.e. we fit our distance checker to the validation features on g_1 and use it on the corresponding OOD features to discard distant samples. We evaluate the error for every model using all possible pairs. Results are summarised in table 2. The proposed GDE-DistCS shows statistically significant improvements across all tasks (expect for one task where it is equivalent), with a median relative MAE improvement of 13% over the standard GDE method. However, it is worth noting that results obtained with the accuracy estimators from the previous paragraph are systematically better than these disagreement estimates.

Ablation studies: choice of distance measure and K-NN hyperparameters We ran additional experiments to justify our choice to use K-NN distance for detecting unreliable samples. As mentioned in section 2.2, other methods have been proposed to perform distance-based OOD detection. Most famous is the Mahalanobis criteria proposed by [27]. Hence, we compare the performance of the proposed ATC-DistCS to ATC-Maha where we use Mahalanobis to compute the distance (all other steps the same). Results in fig. 4 show that (i) for most datasets adding the distance check helps, regardless of the distance choice; (ii) the K-NN distance performs better than Mahalanobis distance (and is often computationally faster). Secondly, in Supp Note 5, we also investigate the impact of the number of neighbours and the effect of features normalisation (as it improved OOD detection in [41]), showing that our method is robust to the choice of number of neighbours and does not require normalisation. Similarly, Supp. Note 6 shows that our 99^{th}-percentile distance threshold choice for rejecting samples is generalisable across all datasets, alleviating the need for cumbersome hyperparameter tuning and allowing us to all parameters fixed across all experiments.

5. Discussion & Conclusion

Main take-aways The proposed "distance-check" significantly improves accuracy estimation results across datasets and tasks both for ATC and GDE; with a median relative MAE improvement of 30% for ATC versus ATC-DistCS in the global setting (resp. 13% in the class-wise setting) and 13% for GDE versus GDE-DistCS. Importantly, our method is versatile, can be applied to any model and does not require any OOD data to tune the performance estimator. In particular, we can apply the method even if we only have access to the final model at deployment time, enabling external performance monitoring (e.g. by regulators or local auditing teams). This contrasts with some recent methods that require dozens of models to improve upon ATC results [1]. Moreover, we would like to underline the demonstrated plug-in aspect of the proposed method, i.e. its ability to improve estimation quality across several "base" accuracy estimators. Indeed, this attests that distance to the expected distribution has to be taken into account for improved performance estimation and that estimators should not solely rely on model outputs. This is further corroborated by our ablation study comparing the use of K-NN versus Mahalanobis distances for the distance check step

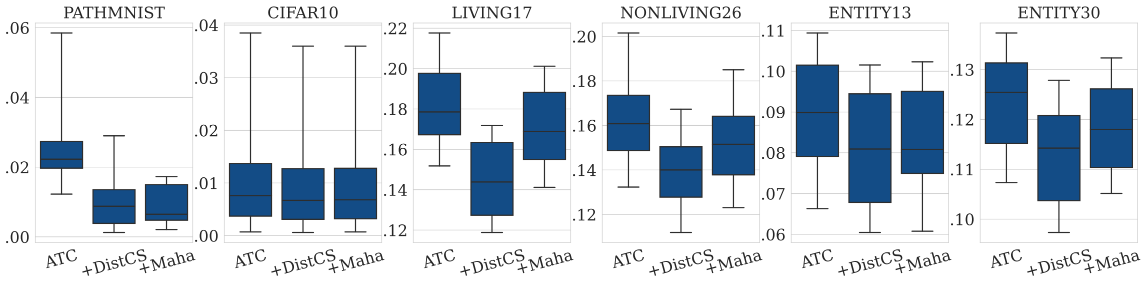


Figure 4. **Ablation study for the choice of distance estimation method: K-NN (DistCS) versus Mahalanobis distance (Maha).** Each boxplot shows the distribution of the Mean Absolute Error for accuracy estimation. Whiskers denote the [5%;95%]-percentiles of the distribution, outliers omitted for readability. Using distance improves the results for all but one dataset, no matter if K-NN or Mahalanobis distance. However, K-NN distance is better than Mahalanobis overall. For additional datasets, see Supp. Note 4.

in the proposed accuracy estimation flow. Indeed, results show that regardless of the distance measure choice, our proposed ATC with distance check outperforms the standard ATC baseline for all but one dataset. This work is, to the best of our knowledge, the first proposing to combine confidence and distance based performance estimation, without requiring access to OOD data at calibration time.

ATC-DistCS, COT and computational considerations Results show that our method performs significantly better (or equivalent) to the concurrently proposed COT method on 10 out of 13 datasets. Our extensive evaluation not only justifies our method but also allows to gain more insight into COT, as it had only be evaluated on a few tasks in the original work [30]. Another important consideration is that COT's runtime increases in $\mathcal{O}(n^3)$ with the number of test samples and linearly with the number of classes, whereas K-NN distance increases linearly with the number of training samples (here limited to 50,000). The authors in [30] propose to alleviate this problem by splitting the test set in batches and averaging accuracy estimates. Despite following this and limiting the number of test samples to 25,000, we still observed a runtime penalty of approximately one order of magnitude compared to ATC-DistCS for datasets with a high number of classes and where the transport optimisation problem needed a large number of iterations before convergence (e.g. on our CPU for a ImageNet ResNet150 model it took over 3500s to get one COT estimate versus 300s for ATC-Dist, 450s vs 50s for Wilds-RR1). Finally, the lightweight aspect of ATC-DistCS (and the ATC baseline) is in start contrast with other proposed methods such as e.g. self-training ensembles [6] which requires training new ensembles for every single evaluation set, highly impractical in real-world monitoring scenarios.

Limitations The proposed method relies on the representativeness of the in-distribution validation set to calibrate the distance threshold. In other words, the validation set should cover the expected set of possibilities encountered in-distribution. If the validation data is not sufficiently representative of the ID setting e.g. not all classes are represented in the validation set, then the distance check is expected to be sub-optimal. This is what is happening with the WILDS-iCam results in the section above. For this heavily data imbalanced task, not all possible targets are present in the validation set. This led to the distance check not improving the results due to a sub-optimal distance threshold choice. Moreover, because classes were missing in the validation set we were not able to compute class-wise thresholds for many classes and had to use the global threshold for these classes, which are especially important in heavily imbalanced settings, as argued by [29]. Similarly, in Wilds-RxRx1 many classes had only a few samples in the given in-distribution validation set leading to sub-optimal class-wise thresholds for this equally imbalanced task.

Finally, our method, by design, generates more conservative performance estimates than their counterparts without the distance check. As it considers that any point that lies too far from the expected embedding space is wrong, it will reduce the estimated accuracy. In most cases, this assumption holds in practice as shown by our experimental results. However, in some settings with extremely heavy distributional shift such as PACS, this assumption may lead to rejecting a lot of samples that appear “too” OOD. This may in turn yield excessively conservative performance estimates. Nevertheless, we argue that in practice if the input data is very far from the expected training distribution, having a low accuracy estimate triggering an auditing alert is a more desirable behaviour than having over-confident accuracy estimates which may mislead the user and generate unsafe AI use. Once the system is validated on the new data, it can easily be included in the calibration set.

Conclusion Taking into account distance to the training distribution substantially improves performance estimation on a wide-range of tasks. Our proposed estimators implementing a distance check demonstrate SOTA performance on a large variety of tasks and significant improvement over previous SOTA baselines. Our method offers a practical and versatile approach to performance estimation on new data distributions, and thus, enables important safety checks for AI model deployment in critical applications. Importantly, our work clearly demonstrates the need to bridge the gap between performance estimation and traditional OOD detection literature and proposes a first step towards this end.

6. Acknowledgements

M.R. is funded through an Imperial College London President's PhD Scholarship. B.G. received support from the Royal Academy of Engineering as part of his Kheiron/RAEng Research Chair.

References

[1] Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2, 7

[2] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018. 4, 5

[3] Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020. 5

[4] Christoph Berger, Magdalini Paschali, Ben Glocker, and Konstantinos Kamnitsas. Confidence-based out-of-distribution detection: a comparative study and analysis. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*, pages 122–132. Springer, 2021. 3

[5] C.E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber, 1936. 6

[6] Jiefeng Chen, Frederick Liu, Besim Avci, Xi Wu, Yingyu Liang, and Somesh Jha. Detecting errors and estimating accuracy on unlabeled data with self-training ensembles. *Advances in Neural Information Processing Systems*, 34:14980–14992, 2021. 2, 3, 5, 8

[7] Mayee Chen, Karan Goel, Nimit S Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Ré. Mandoline: Model evaluation under distribution shift. In *International Conference on Machine Learning*, pages 1617–1629. PMLR, 2021. 5

[8] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 5

[9] Michelle Chua, Doyun Kim, Jongmun Choi, Nahyoung G Lee, Vikram Deshpande, Joseph Schwab, Michael H Lev, Ramon G Gonzalez, Michael S Gee, and Synho Do. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, pages 1–8, 2022. 2

[10] Ching-Yao Chuang, Antonio Torralba, and Stefanie Jegelka. Estimating generalization under distribution shifts via domain-invariant representations. *arXiv preprint arXiv:2007.03511*, 2020. 2

[11] Weijian Deng, Stephen Gould, and Liang Zheng. What does rotation prediction tell us about classifier accuracy under varying testing environments? In *International Conference on Machine Learning*, pages 2579–2589. PMLR, 2021. 2

[12] Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15069–15078, 2021. 2

[13] Hady Elsahar and Matthias Gallé. To annotate or not? predicting performance drop under domain shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173, 2019. 2

[14] Saurabh Garg, Sivaraman Balakrishnan, Zachary Chase Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2022. 1, 2, 3, 5, 6

[15] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1134–1144, 2021. 1, 2, 3, 5, 6

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 3, 5

[17] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, October 2021. 1

[18] Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*, 2018. 4

[19] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 4

[20] Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of SGD via disagree-

ment. In *International Conference on Learning Representations*, 2022. 1, 2, 3, 5, 7

[21] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 2

[22] Andreas Kirsch and Yarin Gal. A note on "assessing generalization of SGD via disagreement". *Transactions on Machine Learning Research*, 2022. 1

[23] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. 1, 4, 5

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[25] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016. 1, 2

[26] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5

[27] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 2, 3, 7

[28] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 5

[29] Zeju Li, Konstantinos Kamnitsas, Mobarakol Islam, Chen Chen, and Ben Glocker. Estimating model performance under domain shifts with class-specific confidence scores. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 693–703, Cham, 2022. Springer Nature Switzerland. 1, 3, 5, 6, 8

[30] Yuzhe Lu, Zhenlin Wang, Runtian Zhai, Soheil Kolouri, Joseph Campbell, and Katia P Sycara. Predicting out-of-distribution error with confidence optimal transport. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*, 2023. 2, 3, 5, 6, 8

[31] Simona Maggio, Victor Bouvier, and Léo Dreyfus-Schmidt. Performance prediction under dataset shift. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2466–2474. IEEE, 2022. 2

[32] Prasanta Chandra Mahalanobis. On the generalised distance in statistics. In *Proceedings of the national Institute of Science of India*, volume 12, pages 49–55, 1936. 3

[33] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021. 1

[34] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5

[35] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019. 1, 2

[36] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 4

[37] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021. 3

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 4

[39] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. {BREEDS}: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*, 2021. 2, 5

[40] Sebastian Schelter, Tammo Rukat, and Felix Bießmann. Learning to validate the predictions of black box classifiers on unseen data. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 1289–1299, 2020. 2

[41] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. 2, 3, 4, 7

[42] J. Taylor, B. Earnshaw, B. Mabey, M. Victors, and J. Yosinski. Rxrx1: An image set for cellular morphological variation across many experimental batches. In *International Conference on Learning Representations (ICLR)*, 2019. 5

[43] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR, 2020. 2

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 4, 6

[45] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 4

[46] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 5

[47] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. 6, 7

[48] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d

biomedical image classification. *Scientific Data*, 10(1):41, 2023. 5

[49] Yaodong Yu, Heinrich Jiang, Dara Bahri, Hossein Mobahi, Seungyeon Kim, Ankit Singh Rawat, Andreas Veit, and Yi Ma. An empirical study of pre-trained vision models on out-of-distribution generalization. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. 1