Predicting with Confidence on Unseen Distributions

Devin Guillory
UC Berkeley
dguillory@berkeley.edu

Vaishaal Shankar
Amazon
vaishaal@amazon.com

Sayna Ebrahimi
UC Berkeley
sayna@berkeley.edu

Trevor Darrell *
UC Berkeley
trevor@eecs.berkeley.edu

Ludwig Schmidt *
Toyota Research Institute
ludwigschmidt2@gmail.com

Abstract

Recent work has shown that the accuracy of machine learning models can vary substantially when evaluated on a distribution that even slightly differs from that of the training data. As a result, predicting model performance on previously unseen distributions without access to labeled data is an important challenge with implications for increasing the reliability of machine learning models. In the context of distribution shift, distance measures are often used to adapt models and improve their performance on new domains, however accuracy estimation is seldom explored in these investigations. Our investigation determines that common distributional distances such as Frechet distance or Maximum Mean Discrepancy, fail to induce reliable estimates of performance under distribution shift. On the other hand, we find that our proposed difference of confidences (DoC) approach yields successful estimates of a classifier's performance over a variety of shifts and model architectures. Despite its simplicity, we observe that DoC outperforms other methods across synthetic, natural, and adversarial distribution shifts, reducing error by ($> 46\%$) on several realistic and challenging datasets such as ImageNet-Vid-Robust and ImageNet-Rendition.

1. Introduction

Even under the best of conditions, machine learning models are still susceptible to large variations in performance whenever the test data does not come from the same distribution as the training data. For instance, recent dataset replication studies have shown that despite concerted efforts to closely mirror the data generating process of the training set, model accuracy still changed substantially on the new test sets [43, 63, 37]. Machine learning models deployed in real-world environments invariably encounter different data distributions. Hence reliable estimates of how well a model will perform on a new test set are critical. For instance, if a practitioner wants to use an X-ray classifier on data from a new hospital, they require a good estimate of their classifier's performance on this previously unseen distribution to have confidence in the results. In some settings it maybe

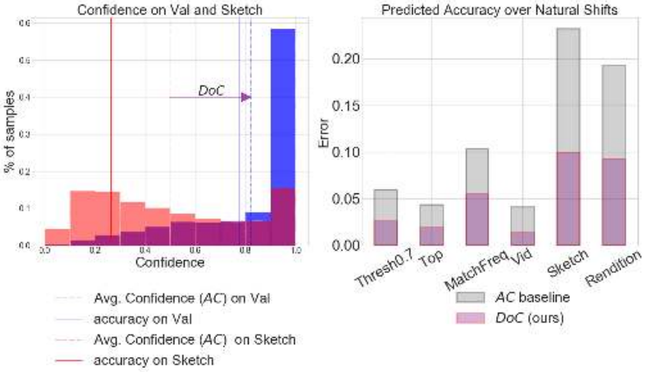

Figure 1: The average confidence (AC) over a distribution is a natural predictor of a model's accuracy. Models are often poorly calibrated, and while AC provides an overly optimistic prediction of accuracy, the difference of confidences (DoC) between distributions provides a useful estimate of accuracy changes. (**Left**) Confidence histograms of a ResNet-101 over the ImageNet validation set and ImageNet-Sketch. Average confidences (AC) are higher than the corresponding accuracies. Our method, difference of confidences (DoC) can be computed from AC over the validation set and ImageNet-Sketch. (**Right**) shows the average confidences (AC) as predictors of accuracy over various natural distribution shifts, contrasted with difference of confidences (DoC). DoC consistently improves over the baseline.

[*] equal contribution

prohibitively costly to acquire new labeled data each time a model encounters a distribution shift. In order to mitigate the consequences of unexpected performance changes, we explore how to best predict changes in a model's accuracy when we only have access to unlabeled data.

We investigate the underexplored problem of Automatic model Evaluation [8] over various model architectures and forms of natural and synthetic distribution shift. Distributional distances such as Frechet distance [24, 65, 8], Maximum Mean Discrepancy (MMD) [2] , and discriminative discrepancy [1, 11, 13], which are common tools for alignment in domain adaptation, are evaluated as features in a regression model for predicting accuracy on unseen distributions. Recognizing the relationship of this problem with that of predictive uncertainty, we also evaluate the utility of model confidences and entropy at predicting accuracy.

Learning our regression models over a set of synthetic distribution shifts, we show that common distributional distances fail to reliably predict accuracy changes on natural distribution shifts. While most approaches are able to encode useful information about held-out forms of synthetic distribution shift, they do not produce useful encodings for natural distribution shifts, where a regression-free average confidence baseline AC outperforms them.

Surprisingly, however, we discover that a substantial amount of information about both synthetic and natural distribution shifts is encoded in the difference of confidences (DoC) of the classifier's predictions between the base (i.e., training) distribution and the previously unseen target distribution. In Figure 1, we show how DoC can be directly used to estimate the accuracy gap between base and target distributions. Treating DoC as a feature, we obtain regression models which substantially outperform all other methods and reduce predictive error by nearly half (46%) across all challenging natural distribution shifts such as the ImageNet-VidRobust [49] and ImageNet-Rendition [18] datasets. Our work demonstrates that it is possible to attain high quality estimates of a model's accuracy across a variety of model architectures and types of distributions shifts.

2. Related Work

Our work touches on multiple related lines of research that have seen substantial activity over the past few years. Hence we only summarize the most closely related work in this section.

Domain Adaptation. In the domain adaptation literature, distributional distances are frequently defined and subsequently minimized to improve model performance. Notable works in this vein define discriminative distances [11, 56], transport distances [5, 6, 35], confusion distances [57], and entropic minimizations [58, 48]. Some of these works attempt to bound the degradation of model performance due to distribution shift [1, 4], based on an assumption of covariate shift [50]. These approaches focus on measures of distribution difference, yet they seldom investigate questions of predicting accuracy or calibration. Through test-time adaptation, [38] have noted calibration improvements, but this comes at the expense of degraded model performance on natural distribution shifts.

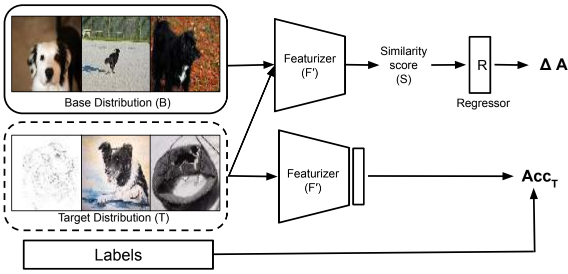

Figure 2: **Illustration of our proposed method.** In order to estimate a model F's accuracy on a previously unseen and unlabeled target distribution T, we feed examples from the base distribution B (usually the training distribution) and the target distribution T into a featurizer F'. In the calibration stage, we first compute a distributional difference, S, from the outputs of F' on various calibration sets and learn a regressor R which predicts the difference in model accuracies from the distributional difference, i.e., $\Delta\text{Acc} = R(S)$. When we later encounter a new unlabeled target set, we again compute the distributional difference and then use the learned regressor to estimate the accuracy difference. Adding the estimated accuracy difference to the accuracy on the base distribution B then yields the estimated accuracy on the target distribution T. This setup enables us to both accurately predict performance over unseen distribution shifts, and to better understand the quality of various distributional differences.

Calibration. Calibrating large DNNs is an active area of research with methods focused on post-hoc calibration [15], encoding uncertainty [10, 34, 60, 25, 46], and training interventions [22, 20, 18] dominating the field. However, a large-scale evaluation of calibration over natural distribution shifts showed that these methods seldom exhibit desired calibration performance in the presence of distribution shifts [39]. Research in calibration focuses on instance-level evaluation of models' predictive uncertainty, while our accuracy prediction task is focused on aggregate performance over an entire test set.

Natural Distribution Shift. Several works in recent years have examined the problem of natural distribution shift, roughly divided into two camps. One line of work, Domain Generalization, has focused on the problem of learning models that adapt well to various forms of distribution shift [55, 31, 40, 14, 54]. These works establish datasets consisting of multiple distinct distributions where some of the distributions could be used for training a model and others for testing it.

Another line of work, Robust Classification, has focused on assuming one training distribution and evaluating or improving performance over a set of specific types of distribution shifts [49, 43, 18, 23, 53]. We blend both lines of work, as we assume training happens with only one training distribution, yet we allow for multiple subsets of the distributions to be used for calibrating the model and evaluating performance. Several recent works have explored the question of model evaluation with [65, 8] using Frechet Distance and [7] using rotation prediction to encode shift, however these approaches struggle to predict generalization gaps on natural distribution shifts. [3] explores the problem from an importance weighting and density estimation perspective, in contrast to the distribution distance and calibration based investigation done in this work. We plan to incorporate experiments based in this work into an updated version of this paper and evaluate over the variety of shifts explored in this paper.

Predicting Generalization Gaps in Deep Learning. Several recent works [28, 29] have explored empirical and theoretical methods to predict the generalization gap of deep learning models, i.e., the gap between training and test accuracy. These methods focus on identifying characteristics of the model and relationship between the model and training distribution, which can be used to predict the generalization gap on unseen examples in the test distribution. Since these methods assume that the training and test distributions are the same, they do not incorporate characteristics of the test distribution and do not have a mechanism for adjusting their predicted generalization gaps under distribution shifts. So when used in our setting with distribution shift, these methods would only predict a single accuracy regardless of the specific test set and hence underperform on the task of predicting accuracy under distribution shift.

3. Baseline Distance Measures

We let F represent a model and $F(x)$ represent the output probabilities of F over instance x. We are interested in estimating the accuracy of F over an arbitrary target distribution $\mathcal{T}$ as computed over test samples comprising dataset T. We also assume the model F is trained over distribution $\mathcal{B}$, whose accuracy can be computed over samples comprising the held-out test dataset B. As some of the distribution shift settings also contain changes in the label space, we define K_B as the set of labels present in $\mathcal{B}$ and $K_{B\cap T}$ as the intersection of labels present in $\mathcal{B}$ and $\mathcal{T}$. $B_{B\cap T}$ is the set of points in B whose labels are in $K_{B\cap T}$. We are interested in predicting either the accuracy, $A_T^B = \mathbf{Acc}(T_{T\cap B})$, or the accuracy gap, $\Delta\mathbf{Acc}(B,T) = \mathbf{Acc}(B_{B\cap T}) - A_T^B$, on unseen distributions. Given access to our models accuracy over the base distribution, A_B^T, predicting either A_T^B or $\Delta\mathbf{Acc}(B,T)$ would allow us to recover the other term. Some methods directly predict accuracy while others predict the accuracy gap; this is further specified in supplementary materials. To enable fair comparisons, we evaluate all approaches on their ability to predict accuracy and apply the $A_B^T + \Delta\mathbf{Acc}(B,T)$ transformation before evaluation. For readablity, B' and T' will be used to represent $B_{B\cap T}$ and $T_{T\cap B}$ respectively.

We focus on computing a measure of distance $S_{B',T'}$ between B' and T' which would allow us to use a regression model R to predict $\hat{A}_T^B = R(S_{B',T'})$. Our measures of distance, S, are computed based on a featurizer, F', associated with the classification model F. F' may represent any features that can be extracted from the model; most commonly these are activations from the penultimate layer of the network, but alternatively they may be logits, probabilities, or the activations from convolutional layers. Unless explicitly stated, all F' in this work deal with the penultimate layers of a deep neural network (DNN). Given base B and target T datasets, we let $S_{B,T} = M(B,T,F')$ where M can be any function that returns a scalar $S_{B,T}$ based on $B, T,$ and, F'.

The measures of distance we explore in this work are commonly used in domain adaptation literature. A large number of works in domain adaption look into the theoretical and practical benefits of minimizing discriminative distances on new domains [11, 1, 4, 56]. We train domain discriminators and look at both their final AUC, and A-proxy [1] distances on held-out test samples. Prior works explore Frechet distance [65, 8] as a useful measure for predicting accuracy and as such we include it in our analysis. We also evaluate Maximum Mean Discrepancy (MMD) [2], which is commonly used in domain adaptation methods [57] and has even been shown to be correlated with target domain accuracy [57, 36]. The pretext task of predicting which of 4 $(0°, 90°, 180°, 270°)$ rotations has been applied to an image has been useful in self-supervised learning, adaptation, generalization, and even predicting generalization on unseen shifts. Our Rotation prediction approach mirrors that of concurrent work [7], however we treat the model representations as fixed and do not update weights to help solve the rotation task. More details behind each of these approaches are provided in the supplementary material.

Predicting Accuracy. In order to learn our regression model R and evaluate its performance over novel shifts, we establish data groupings C and V for calibration and validation, respectively. Our regression model, R, is trained over the accuracy gaps $\Delta\mathbf{Acc}(B,T)$ and measures of difference S present in the calibration grouping C. We train R to minimize Mean Squared Error (MSE) with the accuracy gap as the target, as specified below:

$$\sum_{T\in C} \|R(S_{B,T}) - \Delta\mathbf{Acc}(B,T)\|_2^2 \ . \tag{1}$$

We assess the quality of our regression model over the validation data grouping V to understand how it generalizes to unseen shifts. Our primary experiments are done using linear regression to learn R, and the experiments with non-linear regression models in the supplementary materials follow the same trends as the linear models. The entire prediction pipeline is illustrated in Figure 2.

4. Predicting Performance with Difference of Confidences

A perfectly calibrated model is defined as follows with $\hat{P} = \max(F(x))$ corresponding to the predicted confidence and $\hat{Y} = \arg\max(F(x))$ corresponding to predicted label:

$$P(\hat{Y} = Y|\hat{P} = p) = p, \ \forall\, p \in [0,1] \ . \tag{2}$$

Thus for a perfectly-calibrated model, the expected value of the model's confidence over a distribution should correspond to the model's accuracy over this distribution:

$$\mathbb{E}[\hat{P}] = \mathbb{E}[\hat{Y} = Y] = \mathbf{Acc} \tag{3}$$

However, prior work establishes that modern neural networks are often miscalibrated [15, 39]; they also show that the average confidences of these models differ from the model's accuracy over the respective distributions. The average confidence (AC) for F on B is defined as follows:

$$AC_B = \frac{1}{|B|}\sum_{x\in B} \max_{\{K_B\}}(F(x)) \ . \tag{4}$$

Average confidence is used as a baseline to directly estimate accuracy on an unseen distribution. While average confidence alone proves an unreliable estimate of accuracy, we show here that the difference of confidences (DoC) is informative over various forms of distribution shift. This allows for accurate predictions of model performance on challenging domains. As the average confidence is directly related to calibration, we also report results average confidence after performing temperature scaling on the ImageNet-Validation dataset as described in [15] as *AC TempScaling*.

Given the definition of M above, difference of confidences DoC and average confidences AC can be computed based on a featurization F' of the probabilities of the model. As some of the distribution shifts we explore also change the label space, we need to consider the average confidence with respect to classes present in both T and B, as $K_{B\cap T}$. We calculate average confidence over instances of B' which contains classes present in the $K_{B\cap T}$:

$$AC_B^T = \frac{1}{|B'|}\sum_{x\in B'} \max_{\{K_{B\cap T}\}}(F(x)) \ . \tag{5}$$

We propose difference of confidences DoC as a way to quantify distribution shifts:

$$DoC_{B,T} = AC_B^T - AC_T^B \ . \tag{6}$$

We show that this simple strategy encodes useful information about distribution shift that can lead to successful predictions of accuracy changes.

As our proposed measure, DoC, is a summarizing function of the probabilities of our model F, we also report a variant of our method with difference of average entropy of a model's output probabilities. Entropy may be used as a measure of uncertainty in its own right and [58] show a positive correlation between entropy and accuracy of batches of data. We define average entropy as:

$$\text{Ent}_B^T = -\frac{1}{|B'|}\sum_{x\in B'}\big(\sum_{\{K_{B\cap T}\}} F(x)log(F(x))\big) \tag{7}$$

and difference of average entropy (DoE) as:

$$\text{DoE}_{B,T} = \text{Ent}_B^T - \text{Ent}_T^B \ . \tag{8}$$

As seen below, our empirical results suggest that DoE shares characteristics with DoC which allows it outperform other baselines on predicting natural shifts, yet it does not perform as well as DoC reducing prediction error. We evaluate both DoE and DoC as features encoding the magnitude of distribution shift and show they both exceed performance of other well-known approaches described in Section 3.

5. Experiments

Using DNNs trained on ImageNet [47], we evaluate our ability to predict performance on unseen distributions. This popular large-scale dataset is frequently used as a source for models to be used on different distributions [32, 27], as such it is important to better understand how these models perform in the presence of distribution shift. Using ImageNet as the base dataset, B, allows us to easily access several pre-trained models with various architectures, training curriculum, data augmentations, and calibrations. Additionally, there has been significant recent work [43, 19, 53] on assessing model performance on related datasets which introduce distribution shifts and share class labels with ImageNet.

Natural Shifts. Through our empirical evaluation, we examine the impact of both natural and synthetic distribution shifts. For the purposes of this study, we define natural shifts to be shifts caused by how the data was collected and synthetic shifts as those that can be induced by programmatic alterations applied to the input images. Specifically, we look at ImageNet-V2 [43], ImageNet-VidRobust [49], ImageNet-Rendition [18], ImageNet-Sketch [59]. These

Table 1: Mean Accuracy Gaps of Distribution Groupings

Model	Natural	Synthetic	ImageNet-A
AlexNet	0.25	0.36	0.74
VGG-19	0.26	0.42	0.87
RN-18	0.24	0.37	0.86
RN-34	0.24	0.35	0.88
RN-50	0.25	0.37	0.918
RN-101	0.23	0.33	0.88
RN-152	0.22	0.32	0.87
RNxt-101	0.22	0.32	0.84
WRN-101	0.22	0.33	0.86
Deepaug	0.217	0.26	0.89
AM	0.23	0.29	0.89
AM-Deepaug	0.20	0.20	0.88
All Models	0.23	0.33	0.87

Shows the predicted accuracy error of the models if we rely on accuracy on the base distribution (Base Acc) to evaluate model performance. This optimistic form of model evaluation can be referenced as a naive baseline of the error we seek to reduce throughout this work.

settings introduce various types and intensities of distribution shifts, over which it is desirable for our models to remain well-behaved. In addition to the above stated natural distribution shifts, we also examine performance over ImageNet-Adversarial [23], a distribution that contains natural images, yet they were collected in manner that is explicitly designed to hurt performance on ImageNet classification models. This dataset blends adversarial robustness and natural distribution shift, as such it is examined separately from the other natural datasets which are explained in more detail in the supplementary material.

Synthetic Shifts. Synthetic distribution shifts provide a compelling way for us to investigate model performance under distribution shift as they are controllable in both style of perturbation and intensity of corruption. Yet, since the cause of the shift in these distributions does not naturally arise from our data collection processes, there is a risk that methods which successfully tackle synthetic distribution shift may not prove useful in the more challenging and realistic natural distribution shift setting. We examine the common synthetic corruptions and perturbations present in ImageNet-C [19].

Approaches. Throughout this work we title our approaches based upon the function they use to compute distribution similarity. Frechet, Disc. AUC, Disc A-proxy, Rotation, MMD, DoE, and DoE are the titles for approaches which predict accuracy gap, $\Delta(\mathbf{Acc})$, by training a linear regression model R on top of their difference quantification S. The approaches AC, DoC-Feat, and AC TempScaling directly estimate accuracy themselves and do not use a regression model.

Models. In order to ensure that our approach is not sensitive to certain model architectures, data augmentations, or training schemes we evaluate over a range of models with distinct training pipelines and various accuracy and robustness characteristics. Models explored include AlexNet [33], VGG [51], ResNet [16], ResNext [62], WideResNet [64], DenseNet [26], and Deep Ensembles [34], AugMix [22], DeepAugment [18]. In Table 1 we present the accuracies of these models on the natural distribution shifts described above. The results we present when comparing approaches are average results across all of the aforementioned model architectures, unless specified. Details on the exact experimental setup can be found in supplementary material; including architectural variants and training schemes for models and calibration.

Data Groupings. We take special care to ensure that our ability to predict performance on unseen distributions is not corrupted by any form of information leakage [30]. Ordering our distribution shifts into three distinct groups, two synthetic and one natural, allows us to minimize shared information between data groups and ensure that the unseen shifts in our validation group are unseen in both style and magnitude of shift. Some of our target distribution shifts share similarities with one another which could advantage an accuracy predictor which was exposed to similar distribution shift at calibration time. To mitigate this we place all shifts with known similarities into the same data group, so that our validation shifts share little in common with our calibration shifts and gives a more faithful measure of an approach's ability to generalize.

Our work's chief concern is the ability to predict natural distribution shifts, and as such focuses on settings with synthetic calibration and natural validation datasets. We detail the types of shifts and corruptions contained in each grouping in the supplementary material and present ablation studies on the impact of different calibration groupings there as well.

5.1. Results

Learning models that transfer well from synthetic to natural environments (Syn2Real) [45, 41, 9] is appealing due to the relative ease of creating more synthetic data. As such we conduct the majority of our experiments in challenging setting of calibrating over synthetic shifts and predicting performance on natural distribution shifts.

Synthetic to Natural. In Figure 3, we present the results of 10 different approaches for predicting model accuracy over 6 different natural distribution shifts with 12 distinct neural network architectures. We visualize the results with mean absolute error (MAE) and confidence intervals in the main paper, and present scatter plots of the data points in

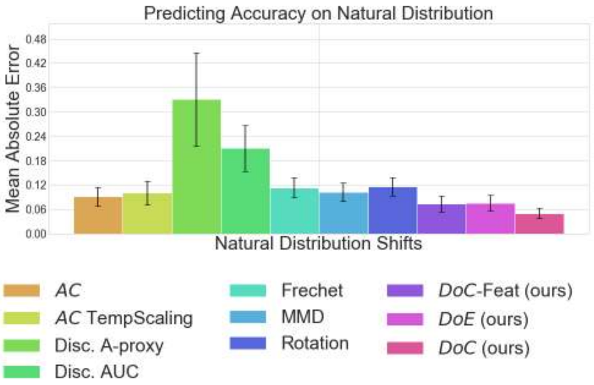

Figure 3: On the challenging and important task of predicting performance over natural distribution shifts with exposure only to synthetic shifts, our approach DoC outperforms all alternatives on predicting accuracy. Traditional distance metrics, such as Frechet distance and A-Proxy distance perform worse than the AC baseline when calibrated over synthetic shifts.

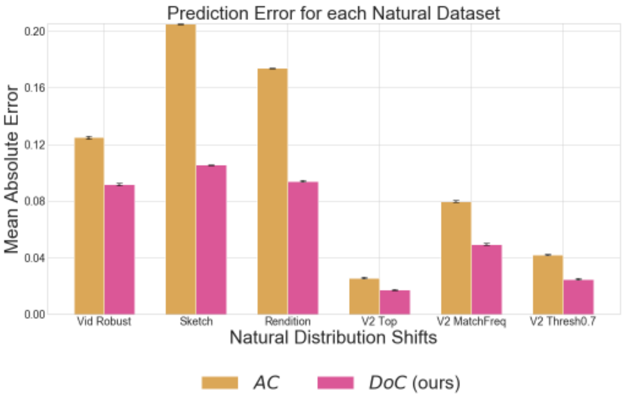

Figure 4: Expanding the results of figure 3 we observe that baseline accuracy AC varies significantly over different natural distribution shifts, with datasets not exclusively composed of natural photography images producing the highest error. However, DoC significantly reduces error across each shift by nearly 50%.

the supplementary materials. Across these various distribution shifts and ImageNet models, we find that the three approaches proposed in our work, (DoC, DoC-Feat, and DoE), are not only the best performing approaches but are also the only approaches to outperform the AC baseline. With a MAE of 0.092 ± 0.019, AC outperforms most prior measures of difference explored in this work with only MMD (0.10 ± 0.018) having substantial overlap. Our best performing approach, DoC, reduces MAE by $> 45\%$, with an average error in accuracy prediction of only 5.0 ± 0.010. DoE and DoC-Feat perform comparably to one another, 0.075 ± 0.015 and 0.072 ± 0.016 respectively, despite DoE leveraging the additional information from the calibration datasets and DoC-Feat not using any regression model. Interestingly, though DoE and DoC operate over the same featurization F', output probabilities, DoC significantly outperforms its entropy-based counterpart through discarding all non-maximum probabilities. Table 1 shows a MAE of 0.23 over the natural data grouping if one assumes there was no accuracy gap or equivalently estimates target accuracy from Base Acc. When calibrated over synthetic distribution shifts, A-proxy [11] and Discriminative AUC do not significantly improve on this naive baseline.

In addition to understanding which approaches work best overall, we are also interested in understanding which, if any, situations our approach fails to improve performance.

In Figure 4, we examine average error over all models for the baseline AC and our best performing approach DoC over each dataset of natural distribution shift. We see that for some distributions, *Vid Robust* and *V2 Top*, AC does a good job at predicting model accuracy with $< 5\%$ error. Other distribution shifts, ImageNet-Rendition and ImageNet-Sketch, the baseline has a much higher magnitude of error. Encouragingly our approach DoC reduces error by close to 50% across each form of natural distribution shift. The results from Figure 4 indicate that it is harder to predict accuracy in the face of certain forms of distribution shift; even after calibration the best average error for ImageNet-Sketch, and ImageNet-Rendition datasets is still higher than the baseline AC error on the remaining distributions. It is worth noting that not all images in ImageNet-Rendition and Sketch are natural photos which may present a more challenging form of shift than the other datasets which are comprised exclusively of natural photography.

Model Specific Performance. In Figure 5, we examine how our approach performs on natural distributions for each model architecture and see that DoC improves performance for every model variant except Augmix-DeepAugment (AM-DeepAug). Furthermore, we observe a steep decline in the magnitude of improvement offered from our DoC approach as it is evaluated over models which incorporate synthetic corruptions into the training process (Augmix, AM-DeepAugment). As we do not replicated the corruptions present in the Augmix training process when comparing between the B base dataset and the calibration datasets C, it is likely that our measured distances with respect to these models are off as a result. For all models not trained with these robustness interventions, our DoC significantly improves the ability to predict accuracy over unseen natural distribution shifts.

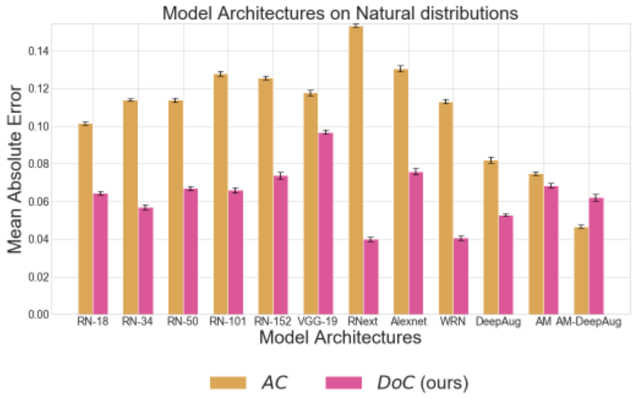

Figure 5: When evaluated over natural distribution shifts DoC consistently improve performance over each specific model type, with the exception of those trained using Augmix. Models trained with Augmix are designed to combat synthetic corruptions thus calibrating on synthetic corruptions may harm results.

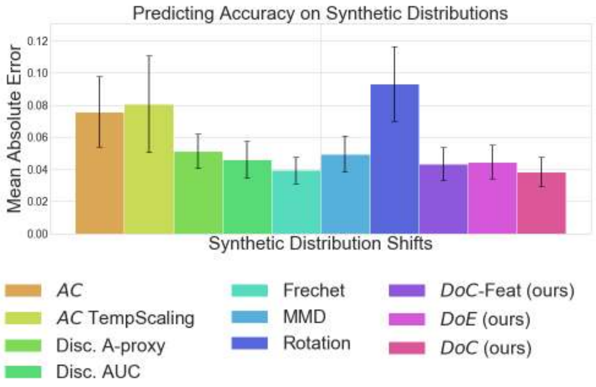

Figure 6: When calibrated with exposure to one set of synthetic shifts, all of the approaches substantially reduce the base accuracy MAE of 0.33 shown in Table 1, when predicting on a held out set of synthetic shifts. Furthermore, with the exception of rotation prediction, all approaches outperform the AC baseline for predicting performance under synthetic distribution shifts. While, DoC and Frechet distance are the best performing approaches in this setting, most approaches are similarly able to predict accuracy over these synthetic shifts.

Synthetic to Synthetic. Using the same setup as in the natural to synthetic analysis, in Figure 6 we evaluate our approaches over novel synthetic shifts. Unlike the synthetic to natural setting all approaches studied substantially improve on the 0.33 MAE of the base accuracy method shown in Table 1. Furthermore, most approaches outperform the baseline of AC, MAE (0.076 ± 0.017), with the sole exception being Rotation Prediction, MAE (0.093 ± 0.019). Frechet Distance and DoC are the best performing approaches with near identical MAE of 0.039 ± 0.007. These best performing approaches yield a 49% reduction in relative error from the AC baseline and an 88% reduction in error from the base accuracy method.

As the majority of the distance metrics explored in this work overlap significantly on the synthetic to synthetic prediction task and noticeably reduce errors in predicted accuracy, studying this form of shift alone may not lead to the insight that DoC methods generalize much better than their competitors when the style of distribution shift changes radically. The observation that metrics which reduce error for synthetic shifts, may have an inverse relationship with natural distribution shifts is worth highlighting and better understanding, as it could either indicate that these methods primarily encode information useful for synthetic shifts, or that the calibrating on synthetic data does not allow them to learn useful mappings over natural distribution settings.

While DoC improves upon the AC baseline for each form of natural distribution shift, we note that in 2 of the 8 synthetic shifts explored (Defocus Blur and Gaussian Blur) DoC actually decreases the accuracy of the predictions. These results are presented in the supplementary material and merit further investigation.

Adversarial Distribution Shift. ImageNet-A [23] produces a uniquely challenging scenario for our approaches. The dataset was designed in an adversarial manner with knowledge of the predictions of an ImageNet classifier, however, unlike many other adversarial tasks, it is composed of non-perturbed images. In Figure 7, we observe that this distribution shift produces the highest predicted accuracy error of shifts we have studied across all approaches. In this setting, our approach DoC, MAE 0.389 ± 0.027, is the only approach able to noticeably reduce error from the AC baseline, MAE 0.476 ± 0.024. Whereas DoE and DoC-Feat perform comparably with the AC baseline, all of the other approaches significantly increase predicted error on this task.

In Figure 8, we look at the results of the various approaches when calibrated over the set of natural distribution shifts. In this setting, our results show DoE and DoC are the only encodings which outperform the AC baseline, with errors of 0.371 ± 0.033 and 0.293 ± 0.024 respectively. When comparing these results to those shown in Figure 7, we see that both DoC and DoE noticeably improve in their ability to predict accuracy on this distinctly challenging form of distribution shift. While all approaches which use the natural calibration make gains in their prediction accuracy, they still are significantly worse than the AC baseline. This both illustrates the importance of calibrating and evaluating over natural distribution shifts and highlights some limitations of these approaches at encoding this difference.

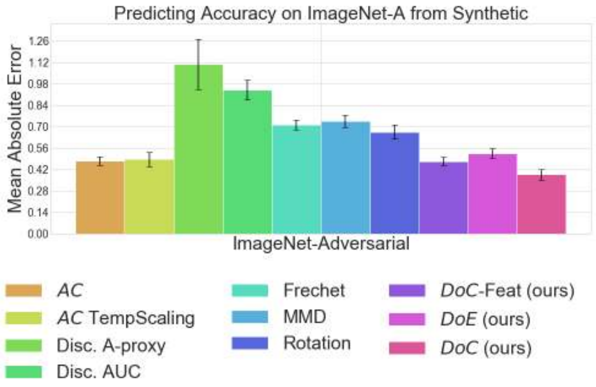

Figure 7: As ImageNet-A was adversarially designed, it presents a distinctly challenging scenario for predicting model accuracy. DoC is the only approach to significantly improve upon the AC baseline when calibrated over synthetic shifts, yet it still much room to improve with a high MAE of 0.389 ± 0.027.

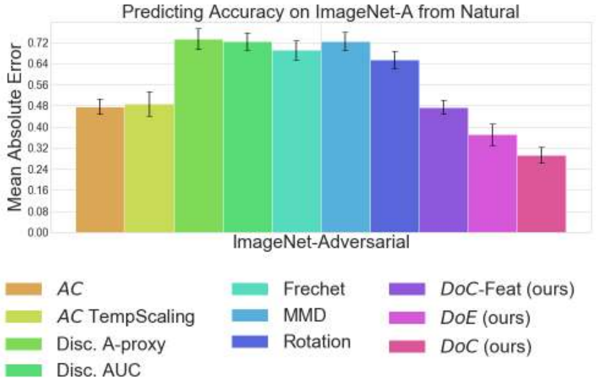

Figure 8: Calibrating over natural distribution shifts improves the performance of all approaches of their ability to predict ImageNet-A accuracy compared to Figure 7. DoE and DoC now improve upon baseline AC performance and improve on their prior performance by 29% and 25% respectively, highlighting the impact of calibration groupings on predicting accuracy.

Resnets vs Deep Ensembles. Large-scale studies on predictive uncertainty with regards to distribution shifts [39] have identified Deep Ensembles [34] as the state of the art in this space. In Figure 9, we show that our post-hoc calibration approach, DoC, can enable a single ResNet-101 model to outperform the more costly Res-Ensemble model, defined in supplementary materials, at predicting accuracy on natural and synthetic unseen distributions. We also show that DoC can be applied to Deep Ensembles to further improve their performance.

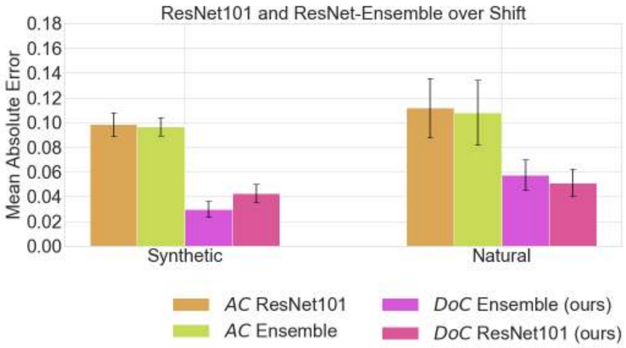

Figure 9: We ompare our method DoC approaches with ResNet-Ensemble, established in [39] as best performing over distribution shifts. ResNet101 is compared with ResNet-Ensemble over natural and synthetic distribution shifts and we show that DoC based approaches on a simple Resnet-101 outperform a deep ensemble.

6. Conclusion

We study the problem of predicting performance change under distribution shift. We find that a simple method, difference of classifier confidences (DoC), accurately predicts the performance change on a wide array of both natural and synthetic distribution shifts. The simplicity of DoC and variants introduced in this paper presents a sobering view on the problem of accuracy prediction, echoing the findings of [39]: many methods explicitly designed for uncertainty prediction seem to fall short of simpler baselines.

While our results present a promising step forward for *detecting* performance drop under distribution shift, we note that the problem is far from solved. Furthermore, the issue of *reducing* the performance drop remains completely open. In this vein we highlight one avenue of future research:

Robust Dataset Construction. Recent work has shown that models trained on orders of magnitude more data can make significant gains in robustness to distribution shift [42, 61, 53]. Constructing such large datasets can be difficult and expensive, but DoC could act as a potential filtering mechanism for focusing data collection on difficult sub-distributions.

We hope that our findings can lay the groundwork for future research on both detecting and reducing performance drops induced by distribution shifts.

Acknowledgements This work was supported in part by DoD including DARPA's XAI, and LwLL programs, as well as Berkeley's BAIR industrial alliance programs.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, pages 137–144, 2006.

[2] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[3] Mayee Chen, Karan Goel, Nimit Sohoni, Fait Poms, Kayvon Fatahalian, and Christopher Ré. Mandoline: Model evaluation under distribution shift. *arXiv preprint arXiv:2107.00643*, 2021.

[4] Corinna Cortes, Mehryar Mohri, and Andrés Munoz Medina. Adaptation based on generalized discrepancy. *The Journal of Machine Learning Research*, 20(1):1–30, 2019.

[5] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3733–3742, 2017.

[6] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018.

[7] Weijian Deng, Stephen Gould, and Liang Zheng. What does rotation prediction tell us about classifier accuracy under varying testing environments? *arXiv preprint arXiv:2106.05961*, 2021.

[8] 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.

[9] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[10] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[12] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[13] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011.

[14] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[15] 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.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020.

[18] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

[19] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

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

[21] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *arXiv preprint arXiv:1906.12340*, 2019.

[22] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[23] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6629–6640, 2017.

[25] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

[26] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[27] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.

[28] Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the generalization gap in deep networks with margin distributions. *arXiv preprint arXiv:1810.00113*, 2018.

[29] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

[30] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):1–21, 2012.

[31] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020.

[32] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2661–2671, 2019.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[34] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6405–6416, 2017.

[35] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019.

[36] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

[37] John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning*, pages 6905–6916. PMLR, 2020.

[38] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

[39] 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.

[40] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[41] Xingchao Peng, Ben Usman, Kuniaki Saito, Neela Kaushik, Judy Hoffman, and Kate Saenko. Syn2real: A new benchmark forsynthetic-to-real visual domain adaptation. *arXiv preprint arXiv:1806.09755*, 2018.

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

[43] 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.

[44] Colorado J Reed, Sean Metzger, Aravind Srinivas, Trevor Darrell, and Kurt Keutzer. Selfaugment: Automatic augmentation policies for self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2674–2683, 2021.

[45] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016.

[46] Carlos Riquelme, George Tucker, and Jasper Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. *arXiv preprint arXiv:1802.09127*, 2018.

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[48] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019.

[49] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? *arXiv preprint arXiv:1906.02168*, 2019.

[50] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[51] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[52] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, pages 9229–9248. PMLR, 2020.

[53] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *arXiv preprint arXiv:2007.00644*, 2020.

[54] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

[55] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[56] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[57] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[58] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[59] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

[60] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter ensembles for robustness and uncertainty quantification. *arXiv preprint arXiv:2006.13570*, 2020.

[61] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

[62] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[63] Chhavi Yadav and Léon Bottou. Cold case: The lost mnist digits. *arXiv preprint arXiv:1905.10498*, 2019.

[64] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[65] Julian Zilly, Hannes Zilly, Oliver Richter, Roger Wattenhofer, Andrea Censi, and Emilio Frazzoli. The frechet distance of training and test distribution predicts the generalization gap. 2019.