SUPPLEMENTARY MATERIAL

7. Non-Linear Regression

As referenced in section 3, we run non-linear regression experiments in which we train a Deep Neural Network to predict model accuracy given measures of distributional difference S. We use fully-connected architectures with a 3-layer fully connected architecture, with layer sizes of 512, 256, and 128 units respectively. The models are trained using stochastic gradient descent for 20k epochs or until convergence, with a learning rate of $1e-4$, weight decay of $1e-3$, and momentum of 0.9. We evaluate over the same models as used in our linear analysis (Resnet-18, Resnet-34, Resnet-50, Resnet-101, Resnet-152, VGG-19, AlexNet, Resnext-101, WideResnet-101, Augmix, DeepAugment, and AM-DeepAugment). The results reported in Figures 10, 11, and 12, mirror those of the linear regression experiments. One noticeable difference is that the prediction error for DoE on synthetic shifts decreases to equivalence with DoC.

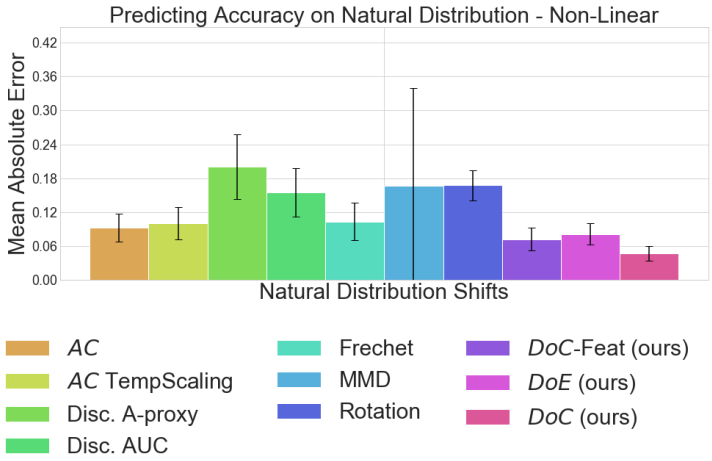

Figure 10: Our approach, DoC, is the best performing approach when calibrating on synthetic distributions and evaluating on natural distribution shifts. We show that our approaches DoC, DoC-Feat, and DoE are the only ones to outperform the baseline of AC. These are consistent with our findings for linear regression.

8. Distribution Shifts and Grouping

Among the distribution shifts, we differentiate between natural and synthetic distribution shifts. Synthetic distribution shifts can be generated in an automated fashion from existing data. We explore the synthetic shifts from the ImageNet-C dataset [19] and combine results from all five intensities. We separate the synthetic data shifts into two groupings $Syn1$ and $Syn2$. $Syn1$ is comprised of Digital shifts (Jpeg, Elastic, Contrast, and Pixelate), Noise shifts (Gaussian, Impulse, Shot), and Weather shifts (Frost, Fog, Snow, and Brightness). $Syn2$ is comprised of Blur shifts (Motion, Defocus, Glass, and Zoom), and Extra shifts (Gaussian Blur, Speckle, Spatter, and Saturate). We look into how predictive performance changes for each type of synthetic shift in Figure 20. While on aggregate our approach DoC outperforms all other explored approaches, we see that for certain forms of synthetic shift our approach decreases the accuracy of the predictions. Further exploration is merited in order to best understand the cause of degradation on these specific synthetic corruptions which are both forms of image blurring (Defocus Blur, and Gaussian Blur).

Throughout the main text all results with synthetic calibration were calibrated over $Syn1$. For completeness we present the results of calibration on $Syn2$ in these Figures 22, 21, and 23. The results we observe follow the same patterns as those in the main text when calibrated over $Syn1$.

The natural distribution shifts that we explore include the three variations of ImageNet-V2 [43] (Top Images, Threshold-0.7, and Matched Frequency) and ImageNet-Vid-Robust [49], both of which consist entirely of natural photography images, with ImageNet-V2 comprising the entire 1k classes of the original ImageNet and ImageNet-Vid-Robust representing a subset of 30 classes. We also examine ImageNet-Sketch and ImageNet-Rendition, both of which contain object representations that are not derived from natural photography and are over a set of 200 classes contained in ImageNet. The aforementioned

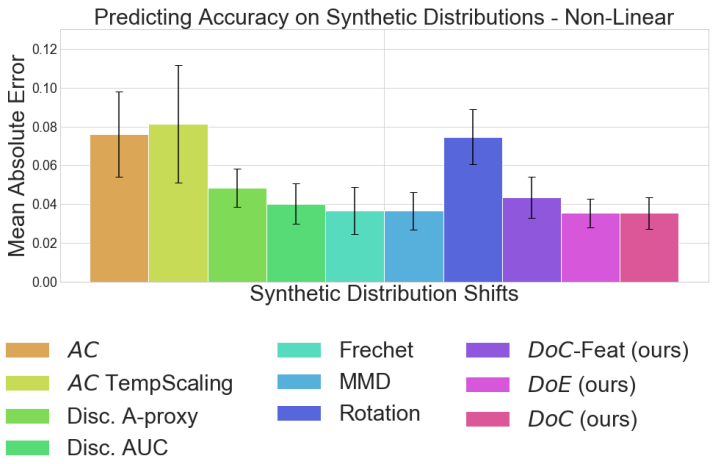

Figure 11: Using the non-linear regression approach described in section 7, we see that all of the approaches outperform the AC baseline for predicting performance under synthetic distribution shifts. When predicting accuracy with non-linear regression, DoC, DoE, Frechet, and MMD all perform comparably at predicting accuracy under synthetic shifts.

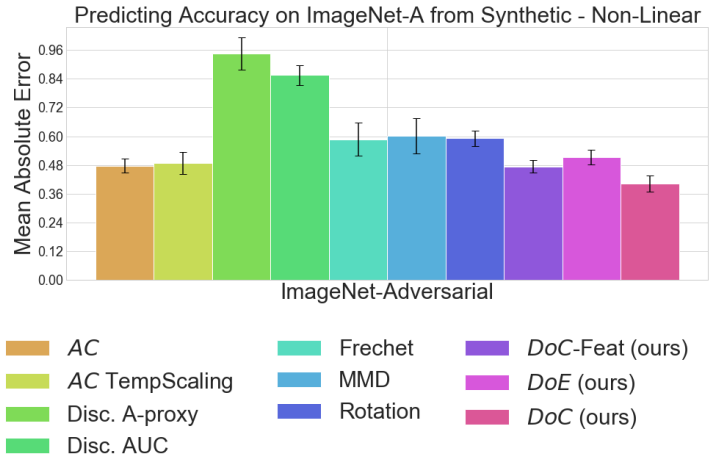

Figure 12: Predicting on ImageNet-A from synthetic shifts with non-linear regression follows the same trends as with linear regression, with only DoC outperforming the AC baseline.

datasets represent our natural distribution data grouping, which is used for both evaluating and calibrating our accuracy predictors. We also evaluate over ImageNet-Adversarial ($Adv.$), but do not include this distribution in our calibrations, as it is collected in an adversarial fashion and may display different characteristics than model-agnostic shifts.

As referenced in section 5, the results over all algorithmic approaches and all calibration and evaluation splits are reported in Table 2. We observe that our approaches (DoC, DoE) outperforms all other approaches in all synthetic to natural calibration settings ($Syn1 \rightarrow Natural, Syn1 \rightarrow Adv., Syn2 \rightarrow Natural, Syn2 \rightarrow Adv.$). The results also show that (DoC, DoE) outperform other approaches in natural to natural calibration settings ($Natural \rightarrow Adv.$).

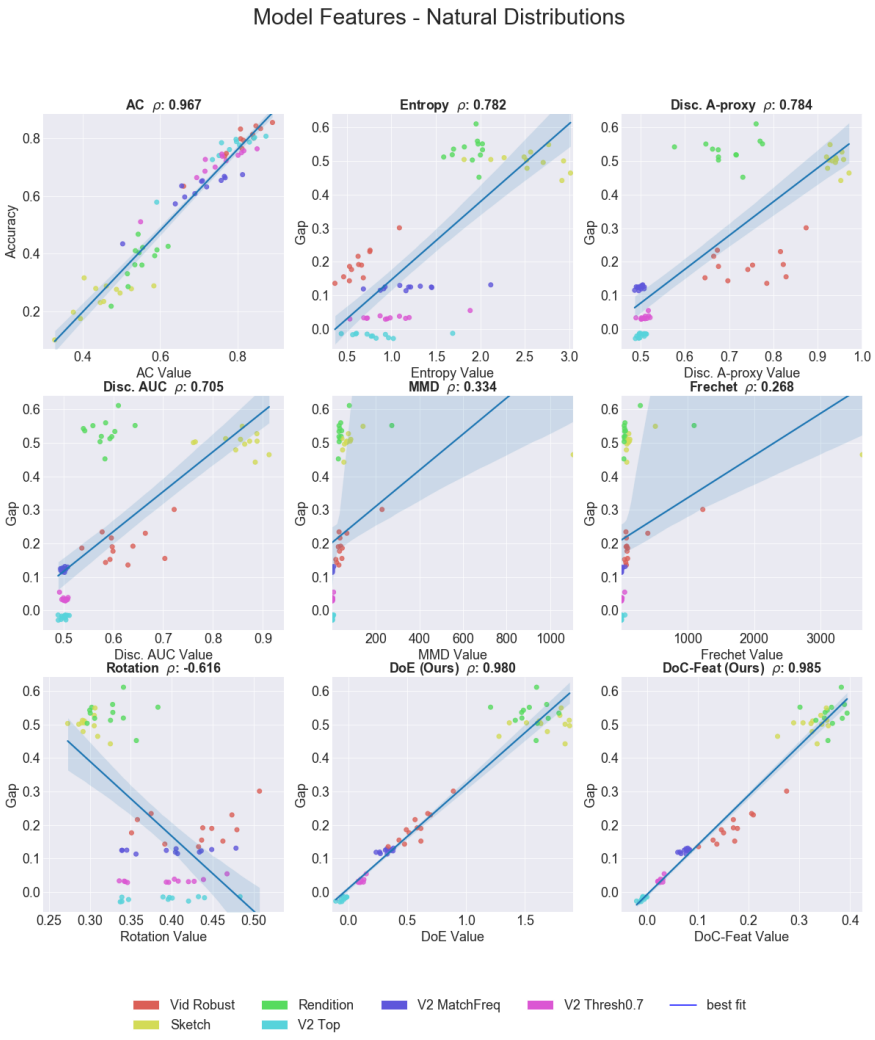

Figure 13: We plot the featurization value (S) and the corresponding accuracy gap for each approach explored in this work, as well as the line of best fit for all natural distribution shifts. We also include Pearson correlation coefficient (ρ), for each plot to capture a measure of how informative each feature is in a global setting. For AC, we plot the accuracy instead of the accuracy gap because it is a direct estimate.

9. Visualizing Features and Predictions

In figure 13, we plot the values of the featurization, S, of each of the approaches explored in this work and the actual accuracy gap for each model and natural distribution shift. We also show the Pearson correlation coefficient ρ, and the line of best for these data points to illustrate how well each of these approaches might be able to encode shifts. Confidence-based measures (AC, DoC-Feat, DoE, Entropy) display the highest levels of correlation and perhaps counter-intuitively, the discriminative discrepancy-based measures (Disc. AUC, Disc. A-proxy) show the next highest levels of correlation. Despite, showing the second highest levels of correlation, the discriminative discrepancy-based measures were the worst performing on predicting accuracy changes over natural distribution shifts. As reason for this that the lines of best fit and Pearson correlation coefficient are determined with knowledge of the every models' accuracy on each target distribution and best-case setting for regression model based on these features. As our regression models are fit with exposure only to synthetic shifts and the accuracy data of one model at a time, we see that our discriminative distance predictions overfit to the synthetic shifts.

Figure 14, shows the value of the featurization, S, of models across all test distributions ($Syn2$, $Natural$, Adv). We

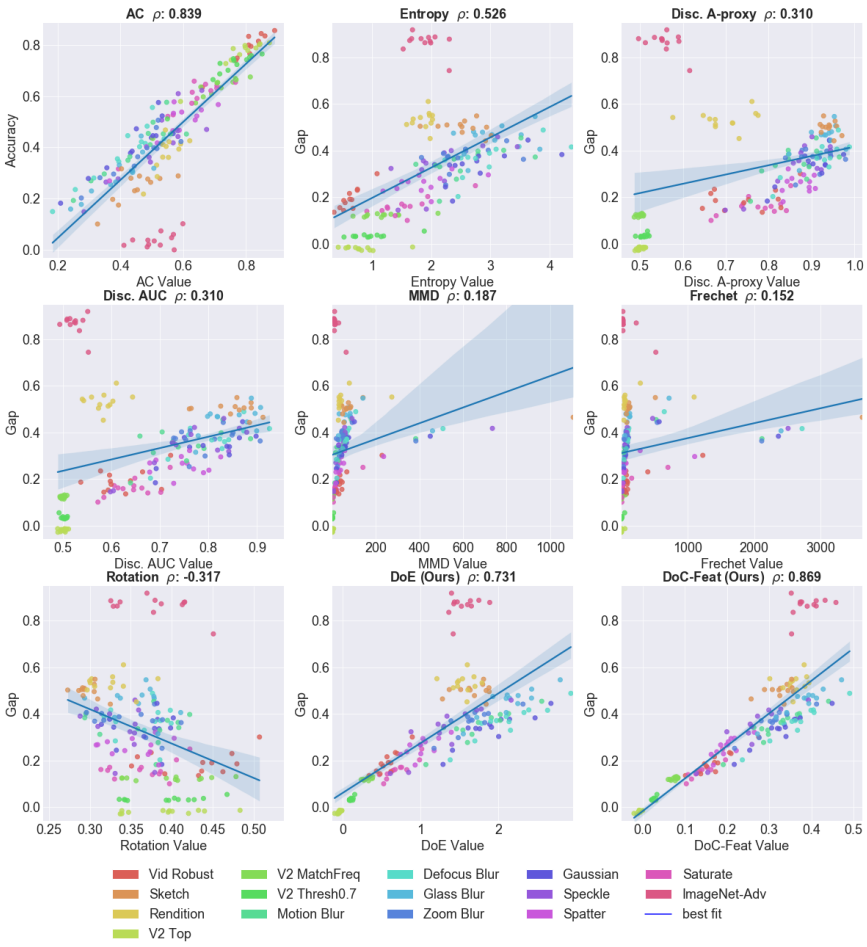

Figure 14: We plot the featurization value (S) and the corresponding accuracy gap for each approach explored in this work, as well as the line of best fit for all test distribution shifts ($Natural$, $Syn2$, $Adv.$). We also include Pearson correlation coefficient (ρ), for each plot to capture a measure of how informative each feature is in a global setting. For AC, we plot the accuracy instead of the accuracy gap because it is a direct estimate.

note many of the same trends as in 13, and the consistent clustering of ImageNet-A as an outlier group. Based upon these visualizations, it appears that measures such as MMD and Frechet distance are non-informative, with regards to predicting accuracy under distribution shifts. However, it is important to note that we learn model-specific regressors for predicting accuracy, and while these measures may not be informative for a global predictor, we see in Figure 16 that when we observe these values for a single model (ResNet-18) , the correlations significantly increase. We observe that the trends on the single model plot, mirror those observed in our regression experiments and highlight the ability of these approaches to learn an informative prediction model that works for various types of shift.

Figure 16, plots the predicted gaps and actual gaps over natural distribution shifts for each approach presented in this work and includes the line of best fit, MAE, and the coefficent of determination (R^2) to summarize how well each approach is doing. Additionally, the line $x = y$ is included on the plot for reference to where the points would land if the predictions were perfect. We note that our confidence-based measures have the lowest MAE and highest R^2, and that predicted error tends to grow as the actual gaps increase. For natural distribution shifts, we see that confidence-based measures tend to produce optimistic estimates with $y > x$ for nearly all points.

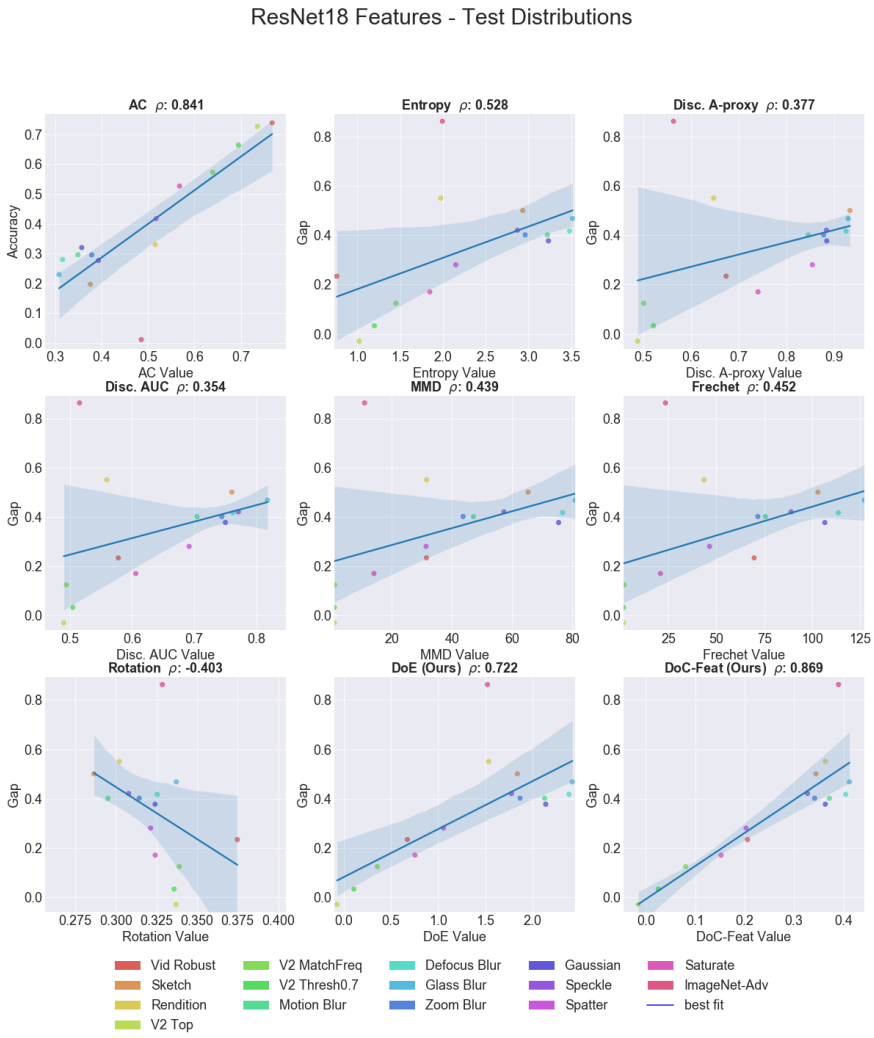

Figure 15: We plot the featurization value (S) and the corresponding accuracy gap for each approach explored in this work, as well as the line of best fit over all test set distribution on a ResNet-18 model. We also include Pearson correlation coefficient (ρ), for each plot to capture a measure of how informative each feature could be with a perfect oracle. For AC, we plot the accuracy instead of the accuracy gap because it is a direct estimate.

In Figure 17, we visualize the predicted gaps of all approach over the held-out synthetic shifts, $Syn2$. It is important to note how well most all of the approaches fare at predicting accuracy on held-out synthetic shifts, with the exception of rotation prediction. The drastic difference in the performance of approaches on Figure 16 and 17, highlight a limitation of prior work studying distribution shift solely on synthetic or natural shifts and highlights the importance of understanding how these forms of shift relate to each other.

Figure 18 show the predicted gaps and actual accuracy gaps over all held-out distribution shifts ($Syn2$, $Natural$, $Adv.$). While confidence-based approaches still provide the best estimates of performance, it is important to note that despite having the lowest correlations in Figure 14, Frechet distance and MMD yield more accurate predictions on unseen shifts than Disc. AUC and Disc. A-proxy. While some of the approaches capture natural and synthetic shifts near equally well, ImageNet-A instances universally present themselves as outliers for all approaches. In order to illustrate how well, the correlation present in a single model (ResNet-18) is captured based on our problem paradigm, Figure 16 shows the results for the predicted gaps of this single model. When we compare with the raw features in 15, we see that our regression models do a relatively poor job of capturing the correlation present in discriminative distances, but do well in the other settings.

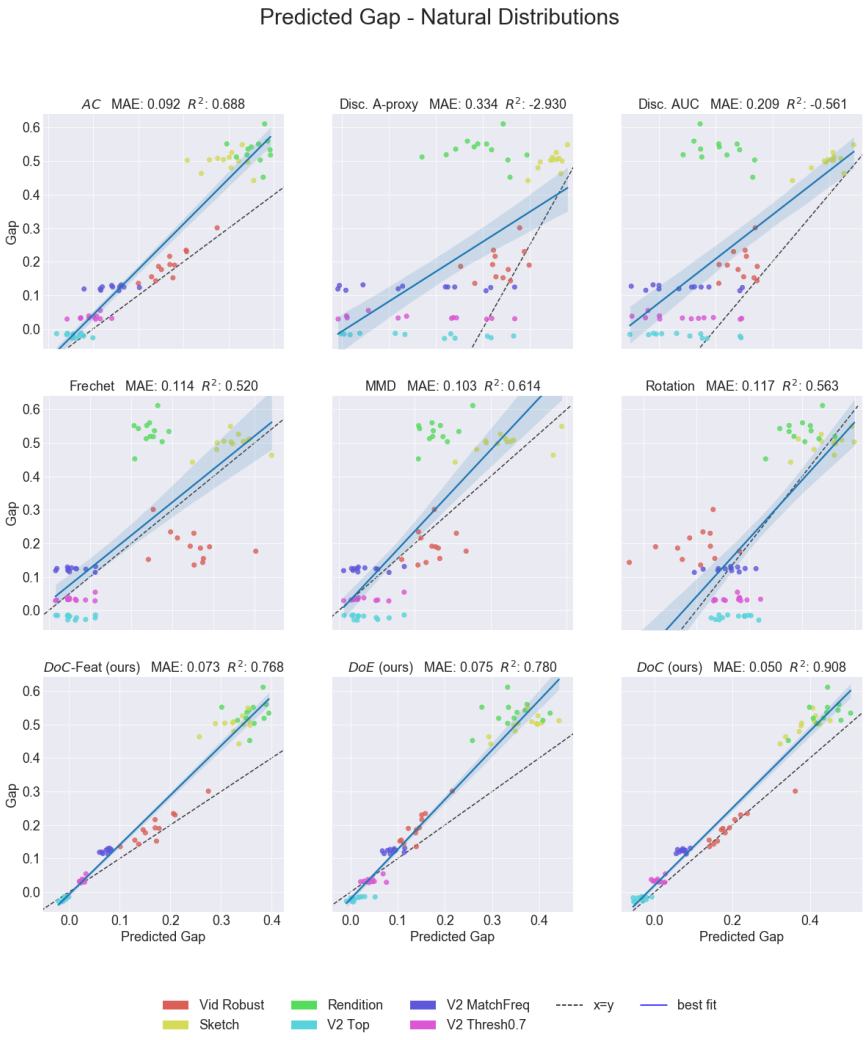

Figure 16: For each approach explored in this work, we plot the actual accuracy gap over the predicted accuracy gap across all twelve architectures. We include MAE and R^2 score to give a since of aggregate fit. Additionally, we plot the line of best fit in blue and the $x = y$ line which would represent perfect predictions as a dashed black line. We see that even without any exposure to natural distribution shifts during calibration our approach DoC does a good job of aligning predictions with the $x = y$ line.

10. Baseline Algorithms

In this section we elaborate on and present formulas used to compute the baseline algorithms discussed in Section 3.

Maximum Mean Discrepancy. The distribution distance metric, Maximum Mean Discrepancy (MMD) [2] is commonly used in domain adaptation methods [57, 36] and has even been shown to be correlated with target domain accuracy [57]. We use the following empirical estimate of MMD between base and target datasets.

$$\text{MMD}(B', T') = \left\| \frac{1}{|B'|} \sum_{x \in B'} F'(x) - \frac{1}{|T'|} \sum_{x \in T'} F'(x) \right\| \tag{9}$$

Rotation Prediction. Several self-supervised learning approaches [12, 17] take advantage of rotation prediction as a useful auxiliary (pretext) task. Recent work in adaptation [52] and generalization [44, 21] has shown rotation prediction to be an

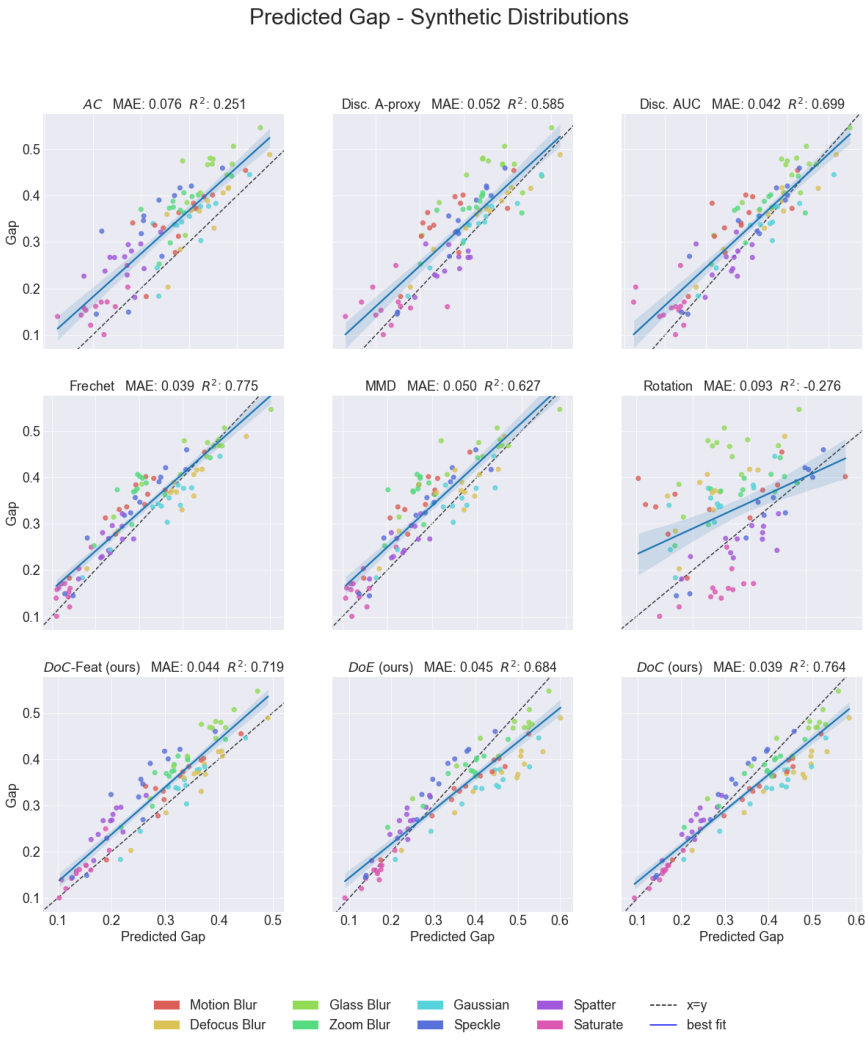

Figure 17: For each approach explored in this work, we plot the actual accuracy gap over the predicted accuracy gap across all twelve architectures over synthetic distributions $Syn2$. We include MAE and R^2 score to give a since of aggregate fit. Additionally, we plot the line of best fit in blue and the $x = y$ line which would represent perfect predictions as a dashed black line. We see that even most approaches reliably predict accuracy over synthetic shifts when calibrated over a disjoint set of synthetic perturbations.

informative task in predicting generalization gaps and improving model performance on shifted domains. As such we train linear models to predict which of the four rotations (0, 90, 180, 270) have been applied to an image, based on the model's featurization, F'. These models are trained over the base dataset B and the accuracy and AUC are reported over the target datasets T.

Frechet Distance. Frechet distance is a popular method of comparing high-dimensional image distributions. Frechet Inception Distance [24, 65] was popularized in the computer vision community as a method of evaluating the quality of generative models. Since Frechet Distance can be computed from summary statistics of the features extracted over the base and target domains, we compute it over the entirety of the datasets and ignore the train / validation / test data splits.

$$\text{Frechet} = \|\bar{F}'(B) - \bar{F}'(T)\| + tr\Big(\Sigma_{F'(B)} + \Sigma_{F'(T)} - 2(\Sigma_{F'(B)}\Sigma_{F'(T)})^{1/2}\Big) \tag{10}$$

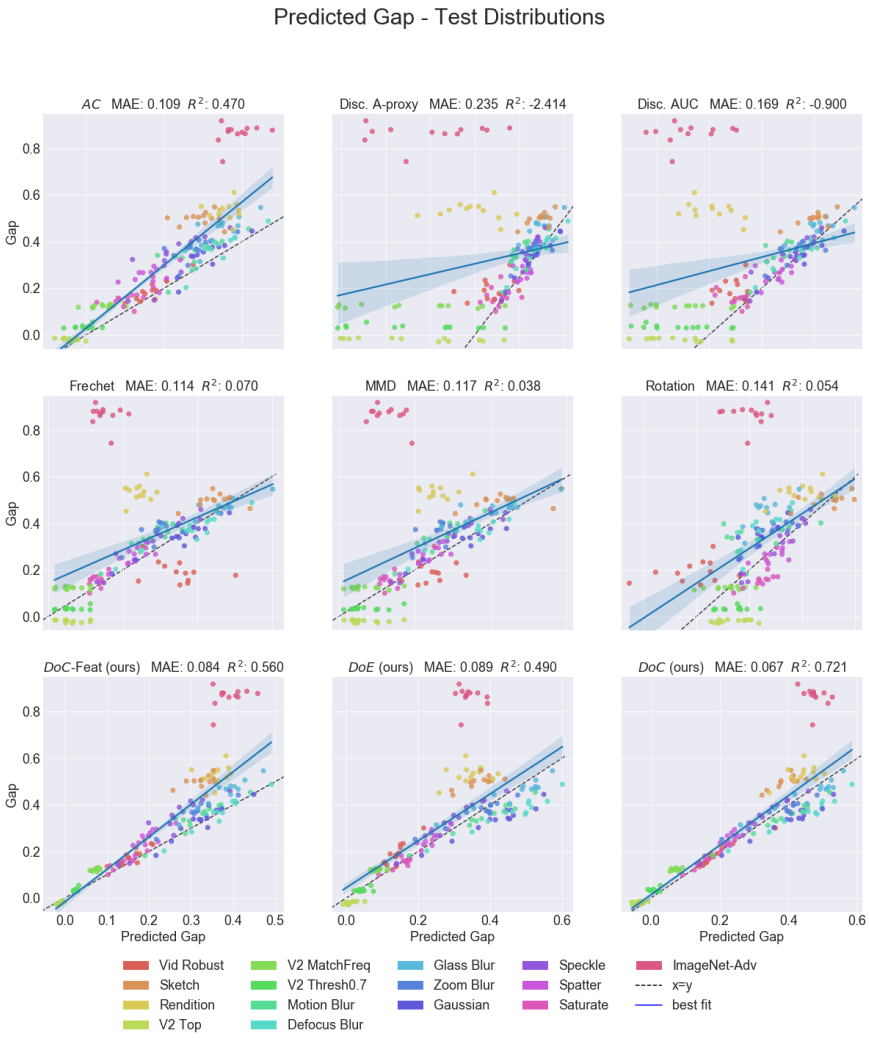

Figure 18: For each approach explored in this work, we plot the actual accuracy gap over the predicted accuracy gap across all twelve architectures over all distribution shifts in the test set ($Natural$, $Syn2$ $Adv.$). We include MAE and R^2 score to give a since of aggregate fit. Additionally, we plot the line of best fit in blue and the $x = y$ line which would represent perfect predictions as a dashed black line. We note that while some approaches are able to capture both synthetic and natural distributions well, the adversarial examples in ImageNet-A continue to present a unique challenge.

where $\bar{F}'(X)$ denotes the average of the featurizations over the dataset X and Σ represents the covariance matrix.

Discriminative distances. In theory and practice, a large number of domain adaptation [11, 1, 4, 56] reduce the error from domain shift by minimizing the ability to discriminate between the base distribution $\mathcal{B}$ and the target distributions $\mathcal{T}$, often dependent on a featurization F' of the classifier. In order to adequately estimate discriminative capacity, we create train, validation, and test splits on B' and T'. We train linear classifiers to discriminate between the base dataset and target dataset based on train and validation splits. Multilayer perceptrons (MLPs) are explored in the supplemental materials. The quality of this discriminator is evaluated over the separate test split to avoid overfitting. We capture several metrics from this discriminator such as accuracy, AUC, and A-proxy [1] defined as:

$$\text{A-proxy} = 2\left(1 - 2 \times \text{error}\right) . \tag{11}$$

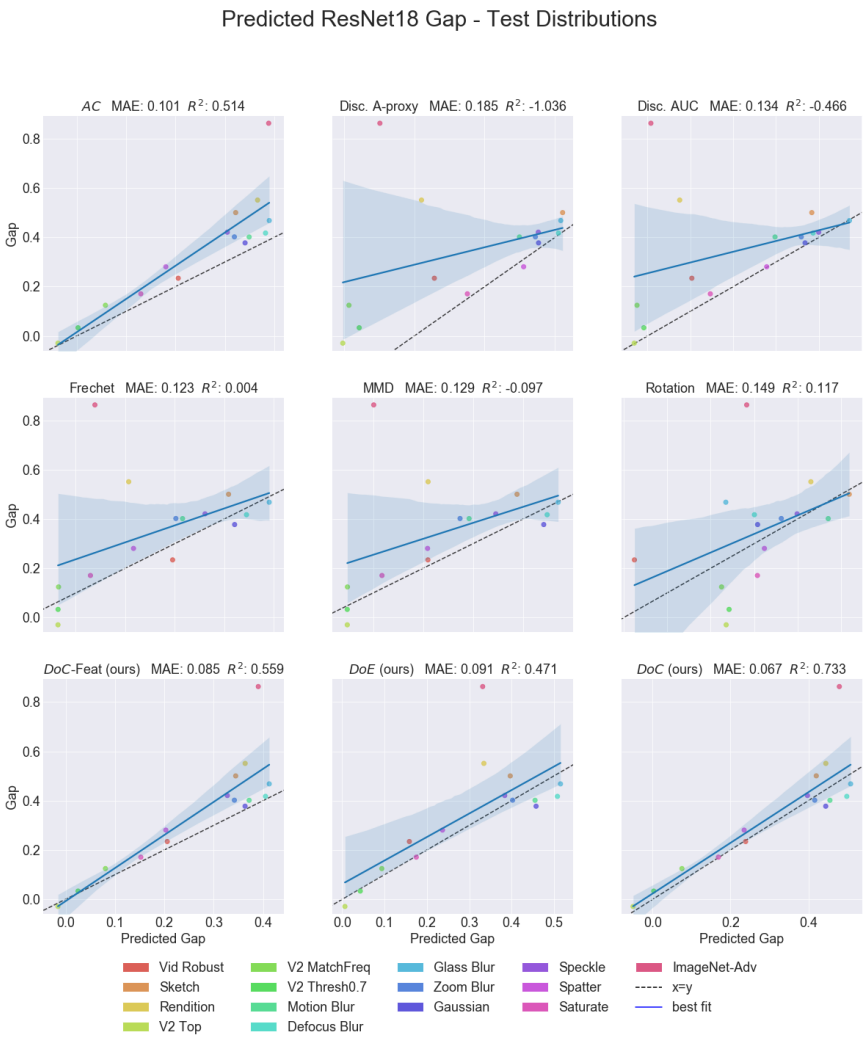

Figure 19: For each approach explored in this work, we plot the actual accuracy gap over the predicted accuracy gap on ResNet-18 over all distribution shifts in the test set ($Natural$, $Syn2$ $Adv.$). We include MAE and R^2 score to give a since of aggregate fit. Additionally, we plot the line of best fit in blue and the $x = y$ line which would represent perfect predictions as a dashed black line. Comparing with 15, we can see how well our models were able to fit the feature distributions.

Temperature Scaling As we are are using the average confidence (AC)of uncalibrated models as our baseline, we also include results of these models calibrated over the base dataset, B, with Temperature Scaling [15]. We use a single temperature parameter and optimize with negative log likelihood as in [15].

11. Experimental Details

To account for the different subsets of labels that our distribution shifts operate over, we compute the accuracy on ImageNet-Val for the subset of classes used by each shifted distribution and attempt to predict the change in accuracy with respect to this number. Because certain measures of distributional difference will be significantly impacted by the classes present, we compute our distributional difference, S, only over the instances that contain the classes present in our target distribution T. For some measures, we seek to optimize an auxilary model (discriminator) to estimate distribution distances. To prevent overfitting on these estimates, we train these models over 40% of the base B and target T distributions. We tune the hyperparameters of these models based on an additional 10% of these distributions, and report distance measures based on the remaining 50% of the distributions. Table 3 presents methods used in this work as well as the specific data splits the

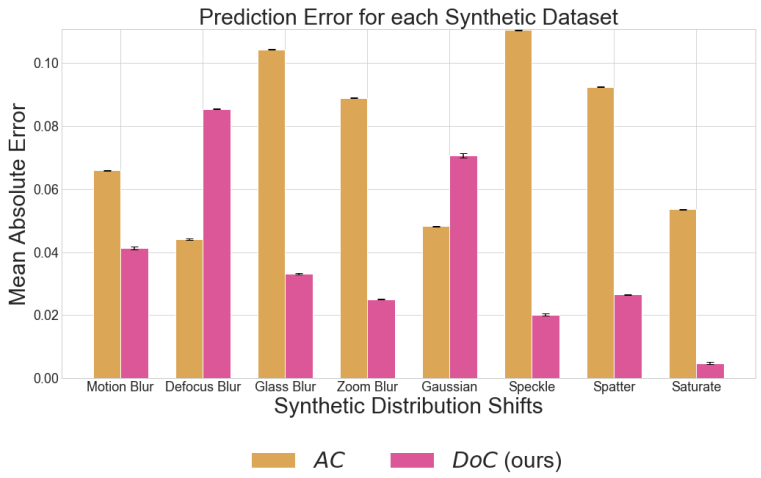

Figure 20: As with distributions, some synthetic distributions present a harder challenge than others. DoC improves performance on 6 of the 8 corruptions but degrades performance on Defocus Blur and Gaussian Blur shifts.

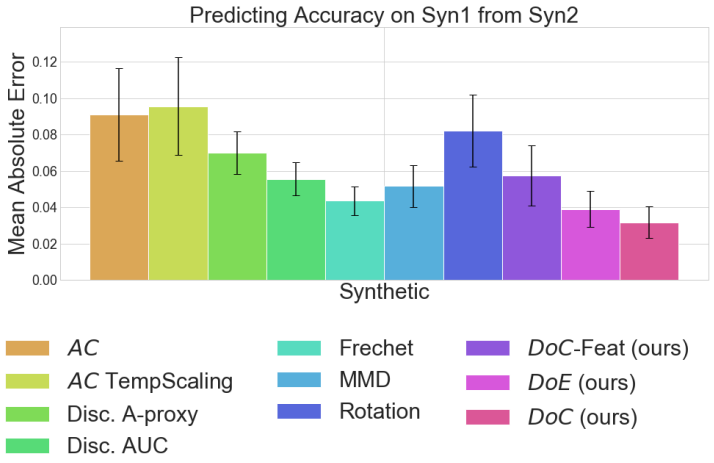

Figure 21: All of the approaches outperform the AC baseline for predicting performance under synthetic distribution shifts when calibrated with $Syn2$ grouping. Our DoC and DoE approaches outperform the other encodings of distributional difference, with DoC producing the best estimates.

measures were computed over and what values they are used to predict.

12. Res-Ensemble

Deep Ensembles were identified as the calibration approach with the best performance under distribution shift in [39]. As such we used an ensemble of Resnet Architectures to serve as a calibration baseline. Our Res-Ensemble model is comprised of pretrained Resnet-18, Resnet-34, Resnet-50, Resnet-101, and Resnet-152 architectures.

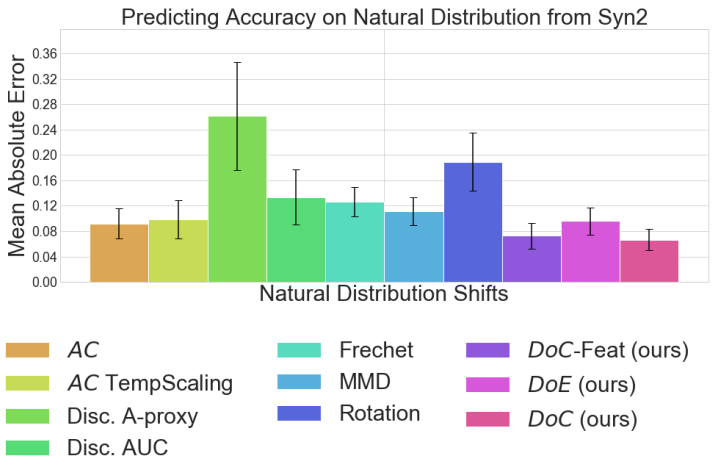

Figure 22: When calibrating over $Syn2$ and evaluating on natural distributions, DoC is still the best performing approach. DoE performs near parity with our AC baseline and all prior approaches performing worse than the baseline.

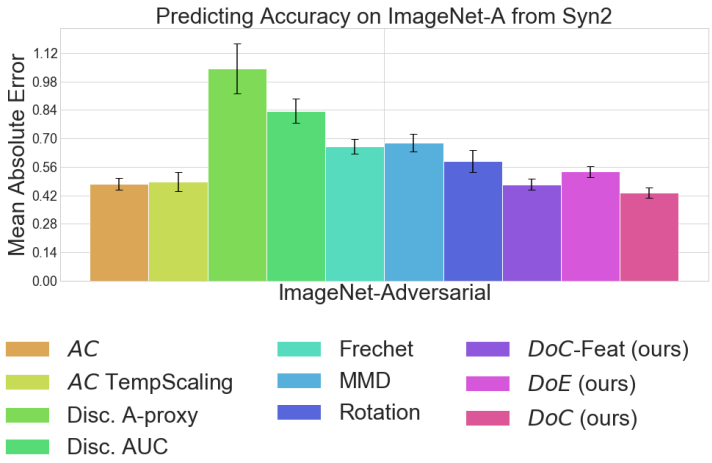

Figure 23: DoC is the best performing approach over ImageNet-A when calibrated with $Syn2$. Unlike prior synthetic calibration, DoE performs worse than the baseline in this instance.

		Syn1	Syn2	Natural	Adversarial
cal	Algorithm	MAE (std)	MAE (std)	MAE (std)	MAE (std)
N/A	Base Acc	0.32 (0.069)	0.328 (0.063)	0.234 (0.087)	0.865 (0.040)
	AC	0.092 (0.042)	0.075 (0.035)	0.087	0.476 (0.047)
	DoC-Feat	0.058 (0.026)	0.043 (0.017)	0.072 (0.033)	0.474 (0.043)
Syn1	Disc. A-proxy	-	0.052 (0.017)	0.335 (0.19)	1.109 (0.266)
	Disc. AUC	-	0.042 (0.016)	0.207 (0.094)	0.94 (0.10)
	Frechet	-	0.0384 (0.013)	0.113 (0.040)	0.713 (0.054)
	MMD	-	0.049 (0.0178)	0.104 (0.037)	0.738 (0.066)
	Rotation	-	0.094 (0.039)	0.115 (0.036)	0.665 (0.075)
	DoE	-	0.045 (0.017)	0.075 (0.032)	0.524 (0.051)
	DoC	-	**0.039** (0.015)	**0.050** (0.021)	0.388 (0.055)
Syn2	Disc. A-proxy	0.070 (0.019)	-	0.261 (0.137)	1.04 (0.196)
	Disc. AUC	0.056 (0.015)	-	0.133 (0.070)	0.836 (0.098)
	Frechet	0.044 (0.013)	-	0.126 (0.038)	0.661 (0.059)
	MMD	0.052 (0.019)	-	0.111 (0.035)	0.678 (0.072)
	Rotation	0.082 (0.032)	-	0.189 (0.074)	0.588 (0.089)
	DoE	0.039 (0.016)	-	0.096 (0.034)	0.537 (0.043)
	DoC	**0.032** (0.014)	-	0.066 (0.027)	0.432 (0.040)
Natural	Disc. A-proxy	0.163 (0.038)	0.134 (0.030)	-	0.734 (0.063)
	Disc. AUC	0.145 (0.029)	0.118 (0.030)	-	0.723 (0.052)
	Frechet	0.056 (0.025)	0.045 (0.026)	-	0.691 (0.057)
	MMD	0.083 (0.046)	0.0596 (0.031)	-	0.724 (0.056)
	Rotation	0.059 (0.022)	0.098 (0.045)	-	0.654 (0.052)
	DoE	0.146 (0.068)	0.186 (0.087)	-	0.371 (0.065)
	DoC	0.070 (0.025)	0.096 (0.036)	-	**0.292** (0.049)

Table 2: We show the mean absolute error (MAE) and standard deviation (std) in predicting accuracy over each approach given various calibration settings. We observe that DoC produces the best overall prediction for each target distribution. DoC also consistently produces the best predictions for natural distribution shifts regardless of calibration setting.

	Eval Data	Estimated Value
Base Acc	B'	Target Acc (A_T^B)
AC	T'	Target Acc (A_T^B)
AC TempScaling	T'	Target Acc (A_T^B)
Frechet	B', T'	$\mathbf{\Delta Acc}(B, T)$
Disc. A-Proxy	B'_{test}, T'_{test}	$\mathbf{\Delta Acc}(B, T)$
Disc. AUC	B'_{test}, T'_{test}	$\mathbf{\Delta Acc}(B, T)$
MMD	B', T'	$\mathbf{\Delta Acc}(B, T)$
Rotation	T'	$\mathbf{\Delta Acc}(B, T)$
DoE	B', T'	$\mathbf{\Delta Acc}(B, T)$
DoC	B', T'	$\mathbf{\Delta Acc}(B, T)$
DoC-Feat	B', T'	$\mathbf{\Delta Acc}(B, T)$

Table 3: Summary table showing all methods of accuracy prediction used in this work along with what data is used to compute measures of difference and what accuracy value they are used to predict.