Source-Free Domain-Invariant Performance Prediction

Ekaterina Khramtsova[1], Mahsa Baktashmotlagh[1], Guido Zuccon[1], Xi Wang[2], and Mathieu Salzmann[3]

[1] The University of Queensland, Australia
[2] Neusoft, China
[3] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
e.khramtsova@uq.edu.au
https://github.com/khramtsova/source_free_pp/

Abstract. Accurately estimating model performance poses a significant challenge, particularly in scenarios where the source and target domains follow different data distributions. Most existing performance prediction methods heavily rely on the source data in their estimation process, limiting their applicability in a more realistic setting where only the trained model is accessible. The few methods that do not require source data exhibit considerably inferior performance. In this work, we propose a source-free approach centred on uncertainty-based estimation, using a generative model for calibration in the absence of source data. We establish connections between our approach for unsupervised calibration and temperature scaling. We then employ a gradient-based strategy to evaluate the correctness of the calibrated predictions. Our experiments on benchmark object recognition datasets reveal that existing source-based methods fall short with limited source sample availability. Furthermore, our approach significantly outperforms the current state-of-the-art source-free and source-based methods, affirming its effectiveness in domain-invariant performance estimation.

1 Introduction

Accurately estimating the performance of a model on some target (i.e., test) data is a challenging problem, particularly in the presence of a discrepancy between the target and source (i.e., training) data distributions, since it invalidates the standard cross-validation strategy. This situation, however, occurs in many practical scenarios, as manually annotating data is expensive and cannot be achieved to cover all possible variations of the input space. Predicting a model's performance on new data is thus critical to assess its reliability in the real world.

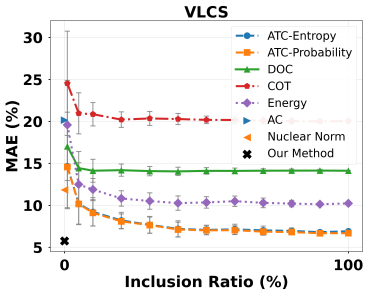


Fig. 1: MAE of estimation across different percentages of source data. Our source-free method outperforms source-based baselines.

The majority of existing approaches to performance prediction [5,7,12,16,30, 34] rely on the availability of the source data as a reference point. For instance, ATC [12] estimates an entropy threshold derived from the source data; COT [30] employs the Earth Mover's Distance between labels from the source domain and predictions from the target domain. However, the reliance on source data poses several challenges, especially in applications where data access is limited or the data is sensitive. For example, in critical areas such as medical diagnostics, privacy concerns limit data disclosure. In some scenarios, only a limited amount of validation data might be available. For instance, when the source dataset is small and has been used to fine-tune a network for a specific task, it is often preferable to hold out less data for validation.

Furthermore, many performance-estimation baselines [12, 16, 30] require the presence of incorrectly-predicted validation samples within the validation set. This requirement is essential to establish a threshold to differentiate between correctly and incorrectly predicted samples. Typically, the validation set acts as an indicator of convergence; thus, well-optimized networks should have a small number of incorrectly-predicted validation samples. Consequently, the effectiveness of existing methods significantly diminishes when the availability of source samples is limited. As illustrated in Fig.1, a notable decline in the performance of all source-based baselines is observed when the availability of source samples, defined by an inclusion ratio, decreases from 100% to 1%, which is evidenced by a spike in the mean absolute error at a 1% inclusion ratio.

In response to these challenges, the primary motivation of this paper is to overcome the constraints imposed by reliance on source data. We propose a more practical and realistic "source-free" setting in which we do not require access to the source data but only to the trained model. To this end, we introduce a method that relies exclusively on target data to predict model performance accurately in the absence of target labels. This approach faces the common challenge of neural networks exhibiting overconfidence in their predictions. To mitigate this, we introduce a novel calibration technique that diverges from traditional temperature scaling, prompted by the lack of source data in our setting. Our proposed unsupervised method leverages a generative model to reduce prediction certainty, and we show its connection to temperature scaling within the softmax function. Following our calibration, we evaluate the correctness of a prediction by measuring its similarity to a uniform distribution. Rather than a direct comparison, inspired by previous research [23], we rely on gradient norms derived from backpropagation of associated losses.

We conduct extensive experiments across a wide range of benchmark datasets, demonstrating that our method significantly surpasses existing state-of-the-art source-free approaches [6,20] and competes favorably with source-based methods, despite not utilizing any source data. Furthermore, we investigate the volume of data required for source-based methods to match our performance and show that our method outperforms other baselines on several datasets, even when all the source data is available. This highlights the effectiveness of our approach in producing reliable source-free domain-invariant accuracy predictions.

2 Related Work

Most existing performance-prediction methods depend on source data to analyze the network's behavior or properties. They can be divided in two categories: sample-wise and dataset-wide methods.

Sample-wise methods associate each sample with a score that correlates with the quality of its prediction. For example, the ATC method by Garg et al. [12] calculates a threshold from the source data's validation set, based on entropy or prediction probability. This threshold aligns with the validation accuracy, and the model's performance on the unlabeled target dataset is estimated by the proportion of samples exceeding this threshold. The DOC [16] approach calculates the difference in confidences between source and target data, derived from predicted probabilities. The Confidence Optimal Transport (COT) method, introduced by Lu et al. [30] leverages optimal transport theory for performance prediction by calculating the Wasserstein distance between predicted target class probabilities and the true source label distribution. Inspired by the work of Grathwohl et al [15], who reinterpret a classifier as an energy-based model, Peng et al. [34] propose using energy instead of the softmax output as a representative statistic for the model's predictions. The works of Baek et al. [1] and Jiang et al. [24] show that OOD accuracy can be estimated by examining the consensus among neural network classifiers. Rosenfeld et al. [35] devise a disagreement loss; the authors provide theoretical guarantees for their method to produce an upper bound on the error under distribution shift.

We have selected these methods as baselines for our approach, as they most closely align with our experimental setup.

In contrast to sample-wise methods, dataset-wide methods aim to derive a measure applicable to an entire dataset rather than individual samples. For example, Deng et al. uncover correlations between classification accuracy and both rotation prediction accuracy [5] and the Fréchet Inception distance [7]. Yu et al. [39] introduces Projection Norm to predict a model's performance by pseudo-labeling test samples and comparing the parameters of a new model trained on these pseudo-labels to an in-distribution model. Subsequent work by Xie et al. [37] shows that even without fine-tuning, gradients obtained after a single backpropagation step can serve as indicators of model performance. Another work by Xie et al. [38] uses feature dispersion to estimate the performance.

However, per-dataset methods are not scalable for scenarios with limited data availability. These approaches either report only the correlation coefficients with accuracy, without providing a means to estimate the actual accuracy value [37, 39], or they rely on generating a large collection of augmented source datasets and analyzing the model's behavior on these sets to predict its performance on the target dataset [5, 7, 34]. Furthermore, they yield a uniform prediction across all target samples without distinguishing between correctly and incorrectly predicted samples. For more details and experiments on dataset-wide methods, please refer to the Supplementary Material.

In addition to source-based methods, a few source-free approaches have been developed. Average Confidence (AC), proposed by Hendrycks et al. [20], calcu-

lates the model's performance by determining the maximum confidence value from softmax probabilities on the target data. Deng et al. [6] employ the nuclear norm, arguing that it captures the confidence and diversity of the predictions.

We show that, despite having fewer constraints, current source-free approaches are significantly less effective than their source-dependent counterparts, even when the latter use only a small amount of source data. By contrast, our source-free method not only competes with, but also surpasses the performance of source-dependent approaches on some datasets, thereby setting a new benchmark for efficiency in source-free methodologies.

Our work is also related to that of Huang et al. [23], who conducted an analysis of the network gradients, backpropagated from the KL divergence between the softmax output and a uniform probability distribution. Their research revealed that OOD samples exhibit a larger gradient norm than in-distribution samples. Although we draw inspiration from the work of Huang et al. [23] to make a decision of prediction correctness based on its closeness to the uniform distribution, our study fundamentally differs from theirs in terms of environmental assumptions, as they assume that the labels of the target dataset do not overlap with those of the source data, and access to data, as they assume access to both in-distribution and OOD samples.

Finally, a Gaussian distribution of the logits [25] or the features [31, 32] has also been used in uncertainty estimation task, e.g. for detecting samples from unknown classes or improving calibration. However, these methods do not estimate the actual performance value and require labeled source data to train the uncertainty quantifier, which is not available in our task.

3 Methodology

3.1 Problem Definition

Let $\mathcal{D}_s = \{(x_s, y_s)\}_{i=1}^{n_s}$ denote a labeled source dataset, where $x_s \in \mathcal{X}_s$ is a source sample, and $y_s \in \mathcal{Y}_s = \{1, ..., C\}$ is the associated label, distributed across C classes. Let $\mathcal{D}_t = \{(x_{t_j})\}_{j=1}^{n_t}$ denote an unlabeled target dataset. We assume that the marginal distributions of the source and target datasets are different, but the class distribution is the same: $p(y_s|x_s) \approx p(y_s|x_t)$, and $p(x_s) \neq p(x_t)$. The source dataset is used to train a classifier $\mathcal{G}_\theta : X_s \to Y_s$ - a function parameterized by θ that predicts a label for a given sample. The classification accuracy of $\mathcal{G}_\theta$ for n target samples is defined as $a_t = \sum_{i \in n}(\tilde{y_i} \equiv y_i)/n$.

The classical performance prediction task [7] aims to predict the performance of the classifier $\mathcal{G}_\theta$ on the target dataset, given the labeled source data and unlabeled target data. Formally, the goal is to create an accuracy predictor A, such that $A(\mathcal{G}_\theta, \mathcal{D}_s, \mathcal{D}_t) \to a_t$.

In this work, we consider a more realistic setting by adding a constraint to the classical performance estimation task. Specifically, we assume that the source data is only used for training and is unavailable during evaluation (hence the name "source-free"). Formally, the task of source-free performance prediction is defined as creating an accuracy predictor A, such that $A(\mathcal{G}_\theta, \mathcal{D}_t) \to a_t$.

3.2 Proposed Approach

In the absence of target labels, we propose to assess the correctness of predictions by evaluating their uncertainty. This involves determining whether the prediction for a sample is closer to a random prediction than to the one-hot class encoding.

While seemingly simple, this, however, suffers from the fact that, in practice, neural networks often exhibit a tendency towards overconfidence, resulting in overly certain estimates. To address this problem, several methods have been proposed. One such method is calibration via temperature scaling, as described in [18], where the network's certainty is reduced by dividing the logits z by a fixed temperature parameter T, which is typically learned from the validation set of the source data. Due to the unavailability of the source data in our setting, we propose to use an alternative approach to calibrate predictions.

Specifically, we start by reducing prediction certainty through adjusting the final probabilities according to the relative distance of the prediction to each class. To achieve this, we construct a probabilistic generative model that yields more accurate estimates of the likelihood of a target sample belonging to each class. It does so by leveraging statistics derived from the observed target distribution, as opposed to the original source distribution. We further show that it can be regarded as an unsupervised calibration method, and establish its connection to temperature scaling.

After calibrating the predictions, we evaluate their correctness by measuring their similarity to the uniform distribution. Instead of performing a direct comparison, we draw inspiration from [23], which showed that gradient norms have higher informativeness than the loss values themselves, and compare gradient norms derived from the backpropagation of associated losses.

Probabilistic Generative Model. The main idea behind our approach is to consider the classification task as a generative model that, given the features, uses posterior probabilities to make optimal class assignments for each sample. To account for the domain shift between source and target, we first model the class-conditional densities $p(x_t|c_i)$ and the class priors $p(c_i)$ from the observed target distribution. We then leverage these to calculate the posterior probabilities $p(c_i|x_t)$ using Bayes' theorem. Let us first define the softmax function from the perspective of a generative model.

Let x_s be the feature vector associated with a source sample, and $f(x_s) = z_s = Wx_s + b$ be a linear function of x_s that produces logits z_s.

We can define a probabilistic generative model where the posterior probability of a sample to belong to a class c_i is calculated as

$$p(c_i|x_s) = \frac{p(x_s|c_i)p(c_i)}{p(x_s)} = \frac{p(x_s|c_i)p(c_i)}{\sum_{j\in C} p(x_s|c_j)p(c_j)} = \frac{e^{f_i(x_s)}}{\sum_{j\in C} e^{f_j(x_s)}} \,.$$

Here, $f_i(x_s) = \ln p(x_s|c_i)p(c_i)$.

The primary limitation of the so-called softmax function is its inability to account for distribution shifts, specifically, the discrepancy in feature distribution generated by the model for the source and target datasets, i.e., $p(x_s) \neq p(x_t)$. Therefore, we aim to replace the softmax function to reflect the similarity (or the difference) in representations between the source and target datasets.

Assuming that the class-conditional densities follow a Gaussian distribution, the density for class i is defined as

$$p(x|c_i) = \mathcal{N}(x, \mu_{z_i}, \Sigma) = \frac{\exp(-\frac{1}{2}(z-\mu_i)^T \Sigma^{-1}(z-\mu_i))}{\sqrt{(2\pi)^C |\Sigma|}} ,$$

with a log likelihood

$$\log \mathcal{N}(x, \mu_{z_i}, \Sigma) = -\frac{1}{2}[\ln(|\Sigma|) + C\ln(2\pi) + (z-\mu_i)^T \Sigma^{-1}(z-\mu_i)] . \quad (1)$$

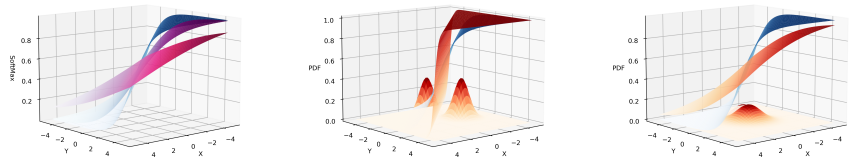
Then, the posterior probability becomes

$$p(c_i|x) = \frac{p(c_i)\mathcal{N}(x, \mu_{z_i}, \Sigma)}{\sum_{j \in C} p(c_j)\mathcal{N}(x, \mu_{z_j}, \Sigma)} . \quad (2)$$

The calculation of posterior probability in Equation 2 can be adapted to account for the feature distribution shift. This adaptation involves estimating the parameters of the generative model, which include Gaussian means μ_{z_j}, prior probabilities for each class $p(c_j)$, and the covariance matrix Σ, from the observed target distribution. However, accurately representing the target distribution poses a challenge due to the unavailability of the ground-truth label distribution in our setup.

To mitigate this, we propose to group the target samples into clusters based on their pseudo labels, and calculate the corresponding mean for each cluster, resulting in mean vectors $[\mu^t_{z_i}, \ldots, \mu^t_{z_k}]$. Note that, when the number of clusters (k) is less than the total number of classes (C), it indicates the presence of classes in the target data with no samples, according to the pseudo labels. For these unrepresented classes, we assign a mean value of 0. The prior probability, which represents the importance of each Gaussian, is defined by the probability of that Gaussian itself generated by other Gaussians, i.e., $p(c^t_i) = \frac{1}{\sum_{j \in C} p(c^t_i|j)}$. The covariance matrix Σ^t is calculated using all target logits.

Importantly, the decision boundaries resulting from our approach are inherently linear in the input space. This linearity is a direct consequence of the model assuming shared covariances among the different classes. We opt for a linear model as the work of Gaston et al. [14] has highlighted the risk of singularities in Gaussian Mixture Models, where one covariance matrix might shrink to zero when a few data points cluster very close to the center of a Gaussian distribution. Therefore, inspired by [14], we use a single (equal) covariance matrix across all Gaussian components, computed on the entire dataset.



(a) Softmax with Temperature **(b)** Generative Model **(c)** Generative Model

Fig. 2: Blue curve: Sotmax function; Orange curve: Generative Model.

Overall Objective. We calculate the posterior probability of the target sample by substituting the Gaussian parameters in Equation 2 with the parameters estimated from the observed target distribution, as discussed above. This approach yields calibrated predictions, where the calibration parameters are adjusted for each target dataset. The log likelihood of the target sample is thus computed as

$$\begin{aligned}\log p(c_i|x_t) = \log \mathcal{N}(x_t, \mu^t_{z_i}, \Sigma^t) - \log \sum_{j\in C, i\neq j} \mathcal{N}(\mu^t_{z_i}, \mu^t_{z_j}, \Sigma) \\ - \log \sum_{j\in C} \sum_{i\in C, i\neq j} \mathcal{N}(x_t, \mu^t_{z_j}, \Sigma^t)\mathcal{N}(\mu^t_{z_j}, \mu^t_{z_i}, \Sigma^t)\ .\end{aligned} \tag{3}$$

The resulting log likelihood can be directly integrated into the standard cross-entropy loss, forming our final objective function

$$L_{CE}(x_t) = -\sum_{i\in C} \mathbb{1}(z = c_i) \log p(c_i|x_t)\ , \tag{4}$$

where $\mathbb{1}$ is the indicator function, being 1 if the argument is true and 0 otherwise.

Connection to Temperature Scaling. Here, we explain the relationship between temperature scaling and our proposed generative model. According to Hinton et al. [21], one way to reduce the overconfidence of predictions is to introduce a temperature parameter to the logits. This yields a probability for class i given by $p(c_i|x) = \frac{\exp(z/T)}{\sum \exp(z/T)}$.

In Figure 2a, we visualize the impact of temperature on the softmax function, by considering only two variables, X and Y, and plotting the softmax values for different temperatures. As shown by the figure, increasing the temperature results in a flatter and more uniform probability distribution.

Similarly to the temperature in the softmax, the Gaussian parameters influence the slope of the probability density function. Well-separated Gaussians with small variances and distinct means lead to sharp slopes in the density function and, consequently, more confident predictions, as shown in Figure 2b. Conversely, similar Gaussians with close means or large variances produce flatter solutions, as depicted in Fig. 2c.

Addressing Numerical Instability. When the number of classes in a dataset is large, the calculations become unstable and lead to exploding gradients. This instability is primarily caused by the sparsity of the covariance matrix Σ as it becomes nearly singular when number of classes increases. Consequently, the determinant of Σ approaches zero, leading to instability when calculating $|\Sigma^{-1}|$ in Eq. 1.

To mitigate this, we normalize $|\Sigma^{-1}|$ using the L_2 norm when we have a large number of classes. Specifically, we stabilize the calculation of $\log \sum_{j \in C} f(z, \mu_j, \Sigma)$ in Eq. 3 by writing

$$\log \sum_{j \in C} \mathcal{N}(x, \mu_{z_j}, \Sigma) = -\log \sqrt{2\pi^C |\Sigma|} + \log \sum \exp(-\frac{1}{2}(z-\mu_i)^T \Sigma^{-1}(z-\mu_i)) \; .$$

The term $\log \sqrt{2\pi^C |\Sigma|}$ is constant with respect to z and, therefore, can be safely omitted as it does not impact the gradients. Note that the remaining value $(z-\mu_i)^T \Sigma^{-1}(z-\mu_i)$ represents the squared Mahalanobis distance, denoted as M^2. To prevent potential issues related to overflow during exponentiation, we adopt the LogSumExp trick used in softmax, which yields

$$\log \sum_{j \in C} \mathcal{N}(z, \mu_{z_j}, \Sigma) = \max(M^2) + \log \sum \exp(-\frac{1}{2}(M^2 - \max(M^2))) \; .$$

Gradient-based Performance Prediction. In the previous section, we adjusted the network's outputs by using the proposed probabilistic generative model. Here, we present our methodology to utilize those calibrated outputs to predict the performance on the target data. Our approach is inspired by [18], which employs the information from the gradient of the loss function to estimate prediction uncertainty. Note that our aim differs from that of [18], as we seek to determine whether a sample is closer to the uniform distribution or to the one-hot class label in order to derive the correctness of a prediction.

To achieve this, we compute two cross-entropy functions: $L_{CE}(s, PL)$ and $L_{CE}(s, U)$. Here PL represents the pseudo label, defined as $PL = [0, ..., 1_k, ..., 0]$, $k = argmax(s)$, u is a uniform distribution $u = [1/C, 1/C, ..., 1/C] \in \mathbb{R}^C$, and C is the number of classes.

To decide whether a sample has been predicted correctly, we compare the norms of the gradients with respect to the weights of the last linear layer, denoted as W. Specifically, if the gradient norm is smaller for the loss based on the uniform distribution (i.e., $L_{CE}(s, U)$), we consider the sample as predicted incorrectly. Conversely, if the gradient norm is smaller for the class-based loss (i.e., $L_{CE}(s, PL)$), the sample is considered as predicted correctly.

Finally, the performance for target dataset $\mathcal{D}_t = \{(x_{t_j})\}_{j=1}^{n_t}$ is computed as

$$A(\mathcal{D}_t, \mathcal{G}_\theta) = \sum_{i \in n_t} \frac{||\frac{\partial L_{CE}(s,U)}{\partial W}|| < ||\frac{\partial L_{CE}(s,PL)}{\partial W}||}{n_t} \; . \tag{5}$$

4 Experiments

4.1 Experimental Setup

We categorize setups into two groups based on the number of source datasets associated with each setup: single-source and multi-source. For the single-source setup, a single model is trained on the source data and then evaluated on one or multiple test sets. In the multi-source setups, each setup consists of multiple domains. We follow the leave-one-domain-out evaluation strategy: Each domain is used for training, while the remaining domains serve as test sets.

Single-Source Datasets:

Digits consists of one source domain, MNIST [28], with 60K training and 10K test images of handwritten digits distributed between 10 classes, and three target datasets: USPS [8], SVHN [33], and SYNTH [11]. The target datasets also consist of digit images, but they differ in terms of colors, styles, and backgrounds. USPS and SVHN represent natural shifts, while SYNTH represents a synthetic shift.

For a more extensive evaluation, we incorporate real-world datasets from the WILDs benchmark [26], namely fMoW [4] and Camelyon17 [2]. fMoW includes satellite images categorized into 62 building or land classes. Camelyon17, a medical dataset, includes patches from whole-slide images of lymph node sections, divided into 2 classes: benign patches and patches with metastatic breast cancer.

Multi-Source Datasets:

As multi-domain datasets, we selected the following 4 datasets from Domain Bed [17]: PACS [29], VLCS [9], Office Home [36], and Terra Incognita [3]. Terra Incognita features wild animal photographs captured across different locations, while the other datasets include images of generic categories, including people and various objects. Each dataset consists of four diverse domains that share the same class labels but significantly differ in style.

Networks. We utilize the LeNet architecture [27] for the Digits setup. For both the fMoW and Camelyon17 datasets, we adopt the Densenet-121 [22] architecture and follow the ERM training procedure from the WILDs benchmark [26]. For Pacs, VLCS, Office Home, and Terra Incognita, we use ResNet50 [19] and follow the default training procedure from DomainBed benchmark [17].

Methods and Evaluation Metrics. We conduct a comprehensive comparison between our method and the following state-of-the-art (SOTA) performance estimation techniques. In our source-based baselines, we include ATC [12], DOC [16], COT [30], Energy [34], and Agree Score [1]. For ATC, we use both probability and entropy as confidence scores. For source-free baselines, we evaluate AC [20] and Nuclear Norm [6]. Note that we also consider the calibrated variant of Nuclear Norm, which significantly enhances performance but requires source data for calibration. Finally, we show the significance of our generative model by reporting the performance of GradNorm without the proposed calibration.

The choice of the selected baseline methods is guided by specific criteria: They do not require any structural modifications to the network, avoid additional fine-tuning, and refrain from extensive data augmentation. These prerequisites align with the characteristics of our own method, ensuring a fair comparison.

Table 1: Absolute Error between predicted and Ground Truth accuracy for **Single-Source datasets**. Results for source-based baselines report the mean and standard deviation across 20 trials, with each trial randomly selecting 1% of source samples.

		Digits				Wilds		
	Source Free	SVHN	USPS	SYNTH	MAE	FMoW	Camelyon	MAE
ATC_e	✗	38.25 ± 1.76	20.54 ± 18.85	23.38 ± 9.67	27.39	3.88 ± 3.87	11.4 ± 3.01	7.64
ATC_e-C	✗	38.43 ± 1.83	22.63 ± 21.21	24.11 ± 10.19	28.39	4.10 ± 3.78	11.4 ± 3.01	7.75
ATC_p	✗	35.34 ± 4.9	18.09 ± 18.93	18.89 ± 12.86	24.10	3.85 ± 2.76	11.4 ± 3.01	7.63
ATC_p-C	✗	35.51 ± 4.99	19.75 ± 20.93	19.48 ± 13.22	24.91	3.46 ± 2.86	11.4 ± 3.01	7.43
DOC	✗	9.69 ± 0.82	4.75 ± 0.82	12.90 ± 0.82	9.11	3.81 ± 2.49	14.37 ± 1.28	9.09
DOC-C	✗	11.37 ± 3.81	5.54 ± 2.84	11.09 ± 3.91	9.33	3.50 ± 2.64	14.18 ± 1.51	8.84
COT	✗	10.84 ± 1.15	7.31 ± 3.86	8.05 ± 2.45	8.73	11.28 ± 4.57	3.14 ± 2.00	7.21
COT-C	✗	12.36 ± 3.72	9.02 ± 6.09	7.42 ± 3.12	9.60	9.43 ± 4.03	3.25 ± 1.95	6.34
Energy	✗	40.56 ± 0.72	47.46 ± 18.83	34.34 ± 6.61	40.79	4.44 ± 4.20	11.39 ± 3.01	7.92
Energy-C	✗	40.58 ± 0.70	47.82 ± 18.53	34.44 ± 6.51	40.95	4.34 ± 4.17	11.39 ± 3.01	7.86
Nuclear Norm-C	✗	2.29 ± 4.25	4.77 ± 2.29	17.86 ± 4.92	8.31	5.54 ± 2.42	14.48 ± 0.83	10.01
Agree Score	✗	23.49	10.31	22.17	18.66	6.2	8.06	7.13
AC	✓	9.38	5.05	13.2	9.21	13.46	16.15	14.81
Nuclear Norm	✓	0.05	4.48	20.4	8.31	12.1	15.16	13.63
GradNorm	✓	27.3	7.82	7.45	14.19	13.04	10.09	11.57
Our Approach	✓	8.77	3.64	1.39	4.60	1.93	3.29	2.61
GT Accuracy		41.59	81.46	50.67		52.91	72.91	

We report two key metrics for each dataset: AE (absolute error between observed and predicted accuracy on the target data) and MAE (mean absolute error across different target sets). For baselines requiring source data during evaluation, we report performance with limited data availability, specifically using 1% of source validation data (referred to as a 1% inclusion ratio). Due to the sensitivity of sample-wise baselines to subset sample selection, we randomly sample 1% of the data 20 times, reporting the mean and standard deviation for all methods across these samples. Unlike other baselines dependent on validation data, AgreeScore requires complete training data from the source domain to train a second classifier; we therefore use a 100% inclusion ratio for that baseline.

Furthermore, we investigate the amount of data required for the existing source-based methods to match the performance of our method. To that end, we report the percentage of inclusion ratio needed for these baselines to reach our method's performance. Notably, many source-based methods underperform compared to our approach, even when using all available validation data.

4.2 Results

We begin our exploration by examining single-domain datasets, followed by an in-depth analysis of multi-domain setups in the next section.

Single-Source Setups

Our experimental results are presented in Table 1 and Figure 3.

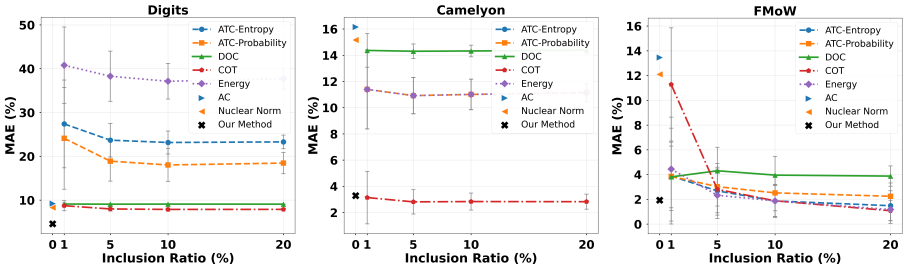


Fig. 3: Mean Absolute Error Across Various Inclusion Ratios for Single-Domain datasets. Points represent the mean, vertical lines represent the standard deviation over 20 trials, each randomly selecting a percentage of source samples corresponding to the Inclusion Ratio.

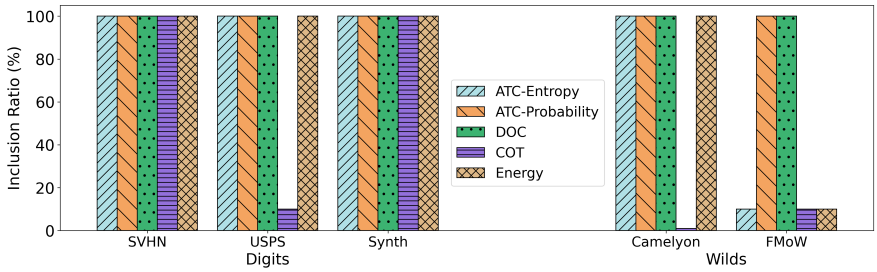


Fig. 4: Inclusion Ratio needed for baselines to match our method's performance for Single-Domain datasets. A 100% ratio indicates that baseline performance is inferior to ours, even when using all source data.

We first observe that the performance of source-based baselines diminishes with a limited number of available source samples, which highlights the importance for source-free methods. At the same time, the current source-free baselines exhibit poor performance. For example, in the Wilds benchmark, even with only 1% of source data, all source-based methods surpass source-free approaches in performance. In contrast, our method outperforms both source-based and source-free baselines by a large margin, with just 2.61% MAE on Wilds, and 4.6% MAE on Digits. In comparison, the nearest competitor in the Wilds setup is COT with 7.21% MAE, and in Digits setup - Nuclear Norm with 8.31% MAE.

Our second observation is that calibration does not uniformly enhance source-based methods. In the Digits setup, calibration degrades performance across all baseline methods; for the FMoW dataset, it enhences the DOC and COT methods but deteriorates ATC's performance. In contrast, for our method, the unsupervised calibration is paramount for achieving accurate performance predictions. By employing a probabilistic generative model, our approach adjusts the probabilities to align with the characteristics of the target dataset. Introducing this unsupervised calibration method allowed for a significant performance improvement compared to the standard GradNorm approach.

Table 2: Absolute Error between predicted and Ground Truth accuracy for **Multi-Source datasets**. Results for source-based baselines report the mean and standard deviation across 20 trials, with each trial randomly selecting 1% of source samples.

	Source-Free	PACS	VLCS	Office-Home	Terra Incognita	**MAE**
ATC_e	✗	$12.19_{\pm 4.60}$	$15.03_{\pm 5.15}$	$10.48_{\pm 3.54}$	$22.69_{\pm 6.60}$	15.10
ATC_e-C	✗	$15.84_{\pm 7.34}$	$19.5_{\pm 5.95}$	$10.82_{\pm 4.45}$	$22.43_{\pm 5.67}$	17.15
ATC_p	✗	$12.28_{\pm 4.56}$	$14.87_{\pm 4.89}$	$11.02_{\pm 4.01}$	$23.28_{\pm 5.85}$	15.36
ATC_p-C	✗	$12.47_{\pm 4.53}$	$17.24_{\pm 5.68}$	$10.90_{\pm 4.11}$	$23.40_{\pm 6.27}$	16.00
DOC	✗	$11.03_{\pm 1.97}$	$16.81_{\pm 4.90}$	$11.8_{\pm 3.31}$	$31.63_{\pm 3.02}$	17.82
DOC-C	✗	$10.68_{\pm 2.34}$	$17.4_{\pm 5.62}$	$8.88_{\pm 3.62}$	$31.60_{\pm 3.69}$	17.14
COT	✗	$22.87_{\pm 5.24}$	$24.12_{\pm 6.41}$	$48.34_{\pm 3.32}$	$11.54_{\pm 4.40}$	26.72
COT-C	✗	$22.58_{\pm 4.84}$	$19.73_{\pm 5.76}$	$42.22_{\pm 2.17}$	$11.25_{\pm 4.43}$	23.95
Energy	✗	$12.40_{\pm 4.62}$	$20.58_{\pm 5.23}$	$11.56_{\pm 4.70}$	$20.83_{\pm 6.92}$	16.34
Energy-C	✗	$12.53_{\pm 4.72}$	$22.29_{\pm 5.33}$	$11.38_{\pm 4.87}$	$20.74_{\pm 7.09}$	16.74
Nuclear Norm-C	✗	$9.66_{\pm 1.7}$	$9.75_{\pm 2.92}$	$10.93_{\pm 2.6}$	$17.92_{\pm 2.18}$	12.07
Agree Score	✗	3.65	10.27	2.34	15.61	7.97
AC	✓	12.82	20.15	22.1	33.88	22.24
Nuclear Norm	✓	11.11	11.84	21.72	19.12	15.95
GradNorm	✓	14.42	21.30	24.11	36.55	24.10
Our Approach	✓	4.27	5.75	7.08	8.72	6.46
GT Accuracy		82.00	74.83	63.31	50.07	

Next, we explore how many samples are required for the baseline methods to match our performance. Remarkably, for the SVHN and SYNTH datasets, our method surpasses all source-based methods, even with full source data availability, as illustrated in Figure 3. For other datasets, the COT baseline improves as more data becomes available, equalling our method's performance with a 10% inclusion ratio for USPS and FMoW, and 1% for Camelyon. Nonetheless, at lower inclusion ratios, COT and other source-based methods exhibit high variance in MAE, indicating instability. Figure 4 shows that while COT achieves less than 2% MAE at a 10% inclusion ratio, reducing this ratio to 1% causes a significant spike in MAE to 11.23%, with a standard deviation of 4.57%. In contrast, our approach is fully deterministic and thus maintains consistent and stable performance, even in the absence of source data.

Multi-Domain Setups

In the following section, we transition to multi-domain setups. The outcomes of our experiments are detailed in Table 2 and Figure 6.

Contrary to the previous setting in which a single domain was designated as the test set, the current datasets comprise four distinct domains. In each configuration, each domain is alternately used as the test set, with the remaining three domains employed for training. As a result, in this scenario, we evaluate the performance of four different models for each dataset, as opposed to the earlier setup where only one model was associated with each dataset.

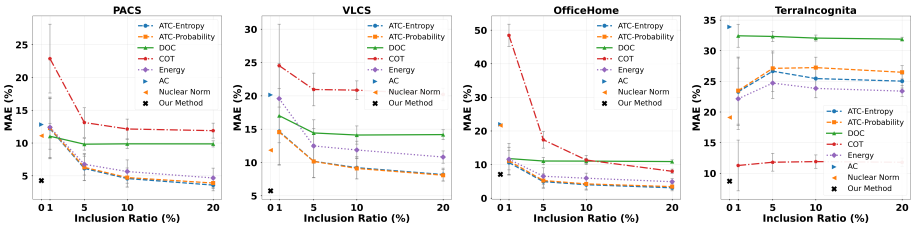


Fig. 5: Mean Absolute Error Across Various Inclusion Ratios for Multi-Domain datasets. Points represent the mean, and vertical lines represent the standard deviation over 20 trials, each randomly selecting a percentage of source samples corresponding to the Inclusion Ratio.

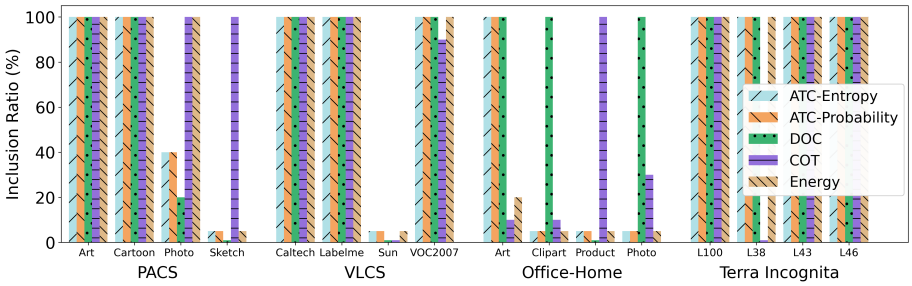


Fig. 6: Inclusion Ratio needed for baselines to match our method's performance for Multi-Domain datasets. A 100% ratio indicates that baseline performance is inferior to ours, even when using all source data.

Our initial observation reveals that the COT baseline, which exhibited superior performance in the single-domain setup, significantly underperforms in the multi-domain setup. Its performance is, on average, even inferior to that of source-free baselines, with an average MAE of 26.72%. Baselines used on PACS and VLCS benifit from more data to achieve competitve results (see Figure 5). This contrasts with the Terra Incognita dataset, where an increase in source data does not correspond to enhanced performance across baselines. Specifically, Figure 5 demonstrates that the MAE does not diminish with an increased inclusion ratio; in fact, it escalates for several baselines, including ATC, COT, and Energy. This outcome highlights the heterogeneity of the source dataset, which includes datasets from diverse domains.

In contrast, our proposed method consistently outperforms all other baselines, especially when little to no source data is available, as evidenced in Table 2. Specifically, on the Office Home dataset, our method surpasses all source-free baselines by a margin of at least 14% MAE. Additionally, most baseline methods on that dataset only begin to macth the performance of our approach at 10% inclusion ratio. Remarkably, across all test sets of Terra Incognita, our method maintains superiority over the baselines even at a 100% inclusion ratio, as depicted in Figure 6. This trend also extends to the Art and Cartoon test sets of PACS and the Caltech and LabelMe datasets of VLCS, demonstrating the robustness and adaptability of our method across diverse domains.

4.3 Ablation Study and Limitations

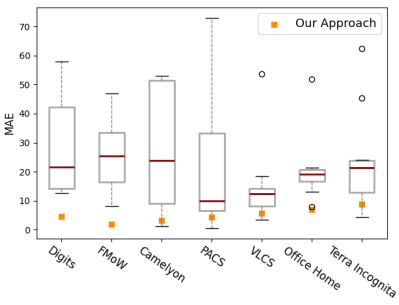


Fig. 7: MAE Variance (sensitivity) of GradNorm to temperature scaling averaged on all datasets. The MAE for our approach is shown in Orange.

Ablation study. In this section, we investigate the viability of our approach without the proposed calibration. This involves the use of GradNorm with the softmax function, parameterized by a constant temperature. This experiment aims to assess the effectiveness of scaling the softmax function with a fixed temperature, specifically to examine whether there exist an optimal threshold across all target datasets.

The results indicate that including the temperature parameter consistently improves the predictor's performance. This suggests that the original predictions for all datasets were overly confident. However, the optimal temperature value varies among datasets. For instance, for the Camelyon17 dataset, the best threshold value is 0.7, resulting in a 3.6% MAE, while the same threshold for the Digits datasets yields a 14.2% MAE.

In Figure 7, we further demonstrate the MAE variance of the GradNorm-based performance estimator across different temperatures. The large variance in MAE values highlights that accurate selection of temperature is critical for the GradNorm estimator. Without access to labeled test data, this choice cannot be made accurately. Therefore, our introduced unsupervised calibration method is crucial for obtaining reliable estimations.

For additional experiments, including comparisons of computational efficiency between methods and the effectiveness of our method versus dataset-wide methods, please refer to the Supplementary Material.

Limitations. Our method relies on the assumption that the logits follow a Gaussian distribution, as used in several other studies [10, 13]. Also, while our method can be sensitive to sparse datasets, making the accurate estimation of Gaussian parameters challenging with few samples per class, we address this by adopting a zero mean and unit variance approach, which works well in practice.

5 Conclusion

In conclusion, our work addresses the challenge of accurately estimating model performance in a source-free and domain-invariant setting, making it relevant for real-world scenarios where the source data is unavailable. Our novel unsupervised approach utilizes a generative model to reduce prediction certainty, addressing the issue of neural networks being overly confident. Then, the prediction correctness is evaluated using gradient norm derived from associated loss. Extensive experiments on diverse benchmark datasets demonstrate our approach's significant performance improvement over current state-of-the-art methods in source-free domain-invariant accuracy prediction.

Acknowledgements

This research is supported by the National Key Research and Development Program of China No. 2020AAA0109400 and the Shenyang Science and Technology Plan Fund (No. 21-102-0-09).

References

1. Baek, C., Jiang, Y., Raghunathan, A., Kolter, J.Z.: Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In: Advances in Neural Information Processing Systems (NeurIPS) (2022), https://openreview.net/forum?id=EZZsnke1kt
2. Bándi, P., Geessink, O.G.F., Manson, Q.F., van Dijk, M.C., Balkenhol, M.C.A., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F.G., Zinger, S., Fukuta, K., Komura, D., Ovtcharov, V., Cheng, S., Zeng, S., Thagaard, J., Dahl, A.B., Lin, H., Chen, H., Jacobsson, L., Hedlund, M., Çetin, M., Halici, E., Jackson, H., Chen, R., Both, F., Franke, J.K., Küsters-Vandevelde, H.V., Vreuls, W., Bult, P., van Ginneken, B., van der Laak, J.A., Litjens, G.J.S.: From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. IEEE Transactions on Medical Imaging **38**, 550–560 (2019), https://api.semanticscholar.org/CorpusID:59600431
3. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
4. Christie, G., Fendley, N., Wilson, J., Mukherjee, R.: Functional map of the world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6172–6180 (2018). https://doi.org/10.1109/CVPR.2018.00646
5. Deng, W., Gould, S., Zheng, L.: What does rotation prediction tell us about classifier accuracy under varying testing environments? In: Proceedings of the International Conference of Machine Learning (ICML) (2021), https://proceedings.mlr.press/v139/deng21a/deng21a.pdf
6. Deng, W., Suh, Y., Gould, S., Zheng, L.: Confidence and dispersity speak: Characterising prediction matrix for unsupervised accuracy estimation. In: Proceedings of the International Conference on Machine Learning (ICML) (2023), https://api.semanticscholar.org/CorpusID:256503627
7. Deng, W., Zheng, L.: Are labels always necessary for classifier accuracy evaluation? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021), https://openaccess.thecvf.com/content/CVPR2021/papers/Deng_Are_Labels_Always_Necessary_for_Classifier_Accuracy_Evaluation_CVPR_2021_paper.pdf
8. Denker, J., Gardner, W., Graf, H., Henderson, D., Howard, R., Hubbard, W., Jackel, L.D., Baird, H., Guyon, I.: Neural network recognizer for hand-written zip code digits. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 1 (1989)
9. Fang, C., Xu, Y., Rockmore, D.N.: Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. Proceedings of the IEEE International Conference on Computer Vision (ICCV) pp. 1657–1664 (2013), https://api.semanticscholar.org/CorpusID:722896

10. de G. Matthews, A.G., Hron, J., Rowland, M., Turner, R.E., Ghahramani, Z.: Gaussian process behaviour in wide deep neural networks. In: Proceedings of the International Conference on Learning Representations (ICLR) (2018), https://openreview.net/forum?id=H1-nGgWC-
11. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: Proceedings of the International Conference of Machine Learning (ICML) (2015), https://proceedings.mlr.press/v37/ganin15.pdf
12. Garg, S., Balakrishnan, S., Lipton, Z.C., Neyshabur, B., Sedghi, H.: Leveraging unlabeled data to predict out-of-distribution performance. In: Proceedings of the International Conference on Learning Representations, ICLR (2022), https://arxiv.org/abs/2201.04234
13. Garriga-Alonso, A., Rasmussen, C.E., Aitchison, L.: Deep convolutional networks as shallow gaussian processes. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019), https://openreview.net/forum?id=Bklfsi0cKm
14. Gaston, J., Das, S., of Computer Science, R.P.I.D.: Active Learning of Gaussian Mixture Models Using Direct Estimation of Error Reduction (2012), https://books.google.com.au/books?id=dPh8zQEACAAJ
15. Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Norouzi, M., Swersky, K.: Your classifier is secretly an energy based model and you should treat it like one. In: Proceedings of the International Conference on Learning Representations (ICLR) (2020), https://openreview.net/forum?id=HkxzxONtDB
16. Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., Schmidt, L.: Predicting with confidence on unseen distributions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1114–1124 (2021). https://doi.org/10.1109/ICCV48922.2021.00117
17. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021), https://openreview.net/forum?id=lQdXeXDoWtI
18. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the International Conference on Machine Learning (ICML). p. 1321–1330 (2017), https://proceedings.mlr.press/v70/guo17a/guo17a.pdf
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
20. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. ArXiv **abs/1610.02136** (2016), https://api.semanticscholar.org/CorpusID:13046179
21. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. ArXiv **abs/1503.02531** (2015), https://api.semanticscholar.org/CorpusID:7200347
22. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269 (2017). https://doi.org/10.1109/CVPR.2017.243
23. Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. In: Advances in Neural Information Processing Systems (NeurIPS) (2021), https://proceedings.nips.cc/paper_files/paper/2021/file/063e26c670d07bb7c4d30e6fc69fe056-Paper.pdf

24. Jiang, Y., Nagarajan, V., Baek, C., Kolter, J.Z.: Assessing generalization of SGD via disagreement. In: Proceedings of the International Conference on Learning Representations (ICLR) (2022), https://openreview.net/forum?id=WvOGCEAQhxl
25. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf
26. Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S.M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., Liang, P.: Wilds: A benchmark of in-the-wild distribution shifts. In: Proceedings of the International Conference on Machine Learning (ICML). vol. 139, pp. 5637–5664 (2021), https://proceedings.mlr.press/v139/koh21a.html
27. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998). https://doi.org/10.1109/5.726791
28. LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit database. ATT Labs [Online] **2** (2010), http://yann.lecun.com/exdb/mnist
29. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. Proceedings of the IEEE International Conference on Computer Vision (ICCV) pp. 5543–5551 (2017), https://api.semanticscholar.org/CorpusID:6037691
30. Lu, Y., Wang, Z., Zhai, R., Kolouri, S., Campbell, J., Sycara, K.P.: Predicting out-of-distribution error with confidence optimal transport. In: ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML (2023), https://openreview.net/pdf?id=dNGxmwRpFyG
31. Manh Bui, H., Liu, A.: Density-softmax: Efficient test-time model for uncertainty estimation and robustness under distribution shifts. In: Proceedings of the International Conference on Machine Learning (ICML) (2024)
32. Mukhoti, J., Kirsch, A., van Amersfoor t, J., Torr, P.H., Gal, Y.: Deep deterministic uncertainty: A new simple baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24384–24394 (2023), https://openaccess.thecvf.com/content/CVPR2023/papers/Mukhoti_Deep_Deterministic_Uncertainty_A_New_Simple_Baseline_CVPR_2023_paper.pdf
33. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2011), https://ai.stanford.edu/~twangcat/papers/nips2011_housenumbers.pdf
34. Peng, R., Zou, H., Wang, H., Zeng, Y., Huang, Z., Zhao, J.: Energy-based automated model evaluation. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024), https://openreview.net/forum?id=CHGcP6lVWd
35. Rosenfeld, E., Garg, S.: (almost) provable error bounds under distribution shift via disagreement discrepancy. ArXiv **abs/2306.00312** (2023), https://api.semanticscholar.org/CorpusID:258999608
36. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017), https://openaccess.thecvf.com/content_cvpr_2017/papers/Venkateswara_Deep_Hashing_Network_CVPR_2017_paper.pdf

37. Xie, R., Odonnat, A., Feofanov, V., Redko, I., Zhang, J., An, B.: Leveraging gradients for unsupervised accuracy estimation under distribution shift (2024)
38. Xie, R., Wei, H., Feng, L., Cao, Y., An, B.: On the importance of feature separability in predicting out-of-distribution error. In: Advances in Neural Information Processing Systems (NeurIPS) (2023), https://openreview.net/forum?id=A86JTXllHa
39. Yu, Y., Yang, Z., Wei, A., Ma, Y., Steinhardt, J.: Predicting out-of-distribution error with the projection norm. In: Proceedings of the International Conference on Machine Learning (ICML). vol. 162, pp. 25721–25746 (2022), https://proceedings.mlr.press/v162/yu22i/yu22i.pdf