Source-Free Domain-Invariant Performance Prediction

Supplementary Material

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/

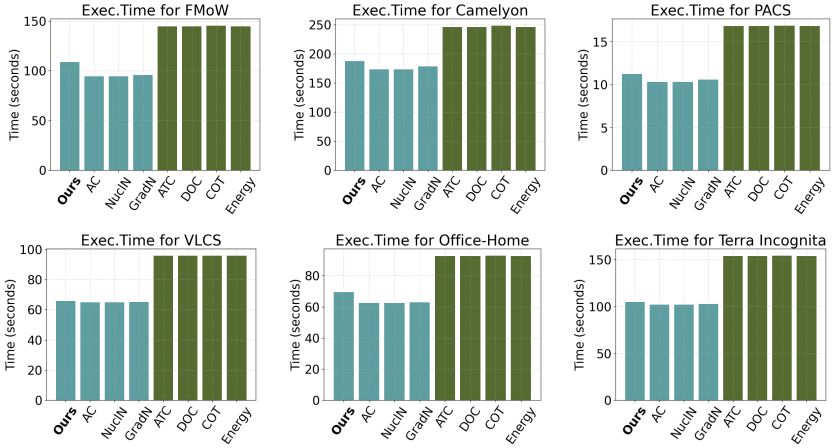


Fig. 1: Execution Time. Source-Free (Blue), Source-Based (Green)

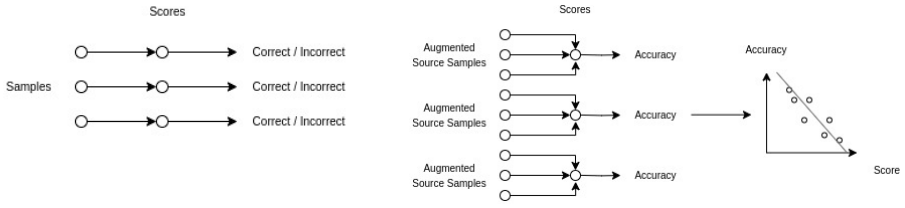
This supplementary material consists of two main parts. The first part discusses the efficiency of various sample-based methods used in the main paper. The second part explores dataset-wide methods, which were not covered in the main paper due to their fundamentally different experimental assumptions.

Efficiency Analysis of Sample-Based Methods

The computational complexity of each method is determined by the number of forward and backward passes of the dataset through the network. While our method requires a backward pass through the network's last layer, all source-based approaches involve a forward pass of the source data. With a 100% inclusion rate, this takes more time than a backward pass, as illustrated in Figure 1. Additionally, the Agree Score requires an additional full training on the source data. Consequently, only source-free methods are more computationally efficient

Table 1: Complexity Comparison wrt Network Pass Requirement

	Source		Target	
	Forward	Backward	Forward	Backward
AC, Nuclear Norm	✓	✗	✗	✗
ATC, DOC, COT, Energy	✓	✗	✓	✗
Agree Score	✓✓	✓✓	✓	✓
Our Approach, GradNorm	✗	✗	✓	✓



(a) Sample-wise methods **(b)** Dataset-wide methods

Fig. 2: Conceptual difference between Sample-wise and Dataset-wide methods

than our approach; however, they are significantly less effective. The backward and forward passes required for each method are summarized in Table 1.

Note that pseudo-labeling does not introduce any computational overhead: it is performed by forwarding the data through the network under evaluation and considering the most probable prediction of that network as the label.

Exploration of Dataset-wide Methods

In this section, we extend our analysis to include additional performance prediction methods, providing an alternative viewpoint on the performance prediction problem. We outline the primary differences and challenges associated with this alternative perspective in Section 0.1, followed by the presentation of experimental results in Section 0.2 and its analysis in Section 0.3.

0.1 Sample-wise vs. Dataset-wide methods

This study introduces a novel approach for unsupervised source-free performance prediction. All the considered baselines represent Sample-wise methods (see Fig.2a) and share a common structure. In essence, each sample is forwarded through the network and assigned a score based on either the network's output, or the network's parameters (e.g., the gradient, as per our method). These scores are further used to predict the overall accuracy, for example by taking the average confidence across all samples (AC [4]). Other methods, including ours, extend this by predicting the correctness of each sample based on specified criteria. For instance, in the ATC method [3], a sample is predicted as correct if its score is below a certain threshold, estimated from the source set. Our method considers a sample to be predicted correctly if the gradient of its most probable prediction is smaller than that towards a uniform prediction.

Table 2: Absolute Error between predicted and Ground Truth accuracy. Results for **Dataset-wide baselines** report the mean and standard deviation across 3 trials, with each trial randomly selecting **1%** of source samples.

	Type	FMoW	Camelyon	PACS	VLCS	Office-Home	Terra Incognita	**MAE**
GT Accuracy		52.91	72.91	82.00	74.83	63.31	50.07	
AC	S-W	13.46	16.15	12.82	20.15	22.1	33.88	19.76
Nuclear Norm	S-W	12.1	15.16	11.11	11.84	21.72	19.12	15.17
GradNorm	S-W	13.04	10.09	14.42	21.30	24.11	36.55	19.92
AC	D-W	3.8 ± 2.09	8.8 ± 2.67	11.11 ± 2.19	18.8 ± 6.64	10.59 ± 3.56	23.57 ± 2.09	13.57
Nuclear Norm	D-W	4.2 ± 2.7	5.78 ± 1.8	10.43 ± 2.65	21.28 ± 13.74	11.07 ± 4.58	8.68 ± 3.35	10.24
GradNorm	D-W	2.0 ± 0.82	8.51 ± 1.37	11.33 ± 2.72	19.51 ± 6.61	11.66 ± 3.76	19.95 ± 2.56	12.16
OT cost	D-W	8.99 ± 2.37	2.91 ± 0.27	25.9 ± 5.02	22.69 ± 3.29	44.57 ± 5.02	13.31 ± 2.39	19.73
Energy	D-W	5.7 ± 2.86	11.57 ± 2.19	11.0 ± 3.88	17.81 ± 8.25	9.73 ± 2.5	23.84 ± 2.02	13.27
FID	D-W	13.53 ± 3.75	0.47 ± 0.4	31.17 ± 11.25	28.52 ± 16.78	40.12 ± 18.32	15.80 ± 3.06	21.60
Dispersion	D-W	4.84 ± 3.2	4.67 ± 1.76	13.17 ± 4.8	20.1 ± 7.55	11.49 ± 3.29	31.30 ± 2.64	14.26
Our Approach	S-W	1.93	3.29	4.27	5.75	7.08	8.72	5.17

However, we distinguish another class of performance prediction methods, namely Dataset-wide methods (see Fig.2b) , which we excluded from the main body of the paper due to their significantly different computational complexity and experimental setup.

The main difference of Dataset-wide methods is that they assign a single score to the entire dataset instead of individual samples, requiring learning the correlation between this score and dataset accuracy, typically through Linear Regression. This approach requires generating training data with varied accuracies, often by using corrupted or augmented versions of the source dataset. The trained model then predicts performance based on the dataset's score.

In contrast, our method predicts performance with a single pass of the test data through the network, whereas Dataset-wide methods need augmented data versions and multiple network passes for training. With most existing methods requiring at least 500 training data points, this translates to 500 additional network passes with source validation data. Despite the computational overhead, we adapted Dataset-wide methods to our setup for experimental completeness.

0.2 Experimental Results

We adapted the following existing Sample-wide baselines: AC [4], OT Cost [5], NuclearNorm [1], and Energy [6]. To convert them into Dataset-wide methods, we calculated the average corresponding score across the dataset. In addition, we include the following Dataset-wide baselines: GradNorm [7], FID [2], Dispersion Score [8]. Note that differently from our setup, where we rely on the gradient of each sample, the Dataset-wide variant by Xie et al. [7] uses dataset-level gradient.

The results of the experiment are presented in Table 2 and Table 3 for 1% and 5% openness, respectively.

In our analysis, we first observe that when adapted to a Dataset-wide format, Source-free baselines such as AC, NuclearNorm, and GradNorm show significant

Table 3: Absolute Error between predicted and Ground Truth accuracy. Results for **Dataset-wide baselines** report the mean and standard deviation across 3 trials, with each trial randomly selecting **5%** of source samples.

	Type	FMoW	Camelyon	PACS	VLCS	Office-Home	Terra Incognita	**MAE**
GT Accuracy		52.91	72.91	82.00	74.83	63.31	50.07	
AC	S-W	13.46	16.15	12.82	20.15	22.1	33.88	19.76
NuclearNorm	S-W	12.1	15.16	11.11	11.84	21.72	19.12	15.17
GradNorm	S-W	13.04	10.09	14.42	21.30	24.11	36.55	19.92
AC	D-W	4.12 ± 0.88	8.48 ± 0.44	8.65 ± 1.23	12.92 ± 3.28	7.55 ± 1.49	22.98 ± 1.33	10.78
NuclearNorm	D-W	3.67 ± 1.89	4.16 ± 0.47	6.32 ± 2.38	11.84 ± 3.76	5.57 ± 3.01	8.16 ± 1.27	6.62
GradNorm	D-W	3.1 ± 1.86	5.81 ± 0.17	10.61 ± 1.31	14.28 ± 3.18	9.61 ± 1.04	17.32 ± 1.92	10.12
OT cost	D-W	0.91 ± 0.73	3.3 ± 0.49	15.34 ± 2.47	14.01 ± 3.57	16.41 ± 2.0	12.81 ± 0.9	10.46
Energy	D-W	5.65 ± 1.59	10.21 ± 0.15	9.21 ± 2.11	14.63 ± 3.96	9.51 ± 2.95	22.48 ± 1.17	11.95
FID	D-W	2.41 ± 1.9	0.68 ± 0.57	14.0 ± 6.05	18.31 ± 7.57	19.12 ± 3.39	11.36 ± 3.38	10.98
Dispersion	D-W	5.41 ± 2.66	3.07 ± 0.64	8.45 ± 2.57	13.32 ± 3.52	14.58 ± 3.63	25.0 ± 0.88	11.64
Our Approach	S-W	1.93	3.29	4.27	5.75	7.08	8.72	5.17

performance improvements. Note, however, that transitioning to a Dataset-wide variant renders them no longer source-free.

Next we notice that our method consistently outperforms both Sample-Wise and Dataset-wide variants of GradNorm across all datasets examined, proving the importance of the proposed unsupervised calibration with generative model. Although the performance of Dataset-wide baselines increases with more data, our method remains superior in scenarios with limited data availability.

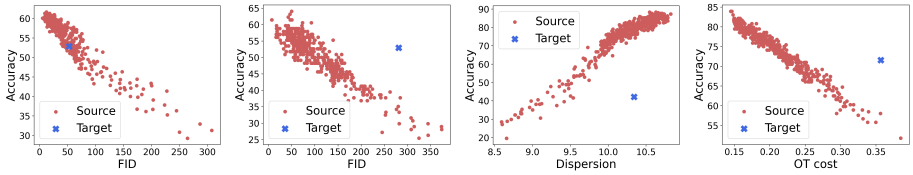
Among the Dataset-wide baselines, there is no consistent leader; FID, for example, performs best Camelyon dataset, but worst on OfficeHome dataset, as shown in Table 3. In the next section, we analyse and visualize different aspects that affect the performance of Dataset-wide baselines.

0.3 Performance Analysis

In this section, we discuss the challenges faced by the Dataset-wide methods when access to source data is limited.

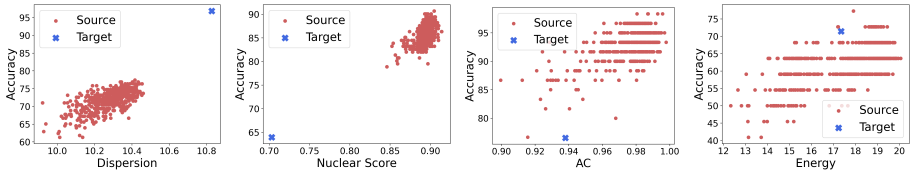
Poor Score Distribution We first notice that the score distributions between target and source do not always match. Moreover, the sensitivity of some scores to the sample size significantly influences their representativeness. For instance, FID, which relies on a covariance matrix, becomes less representative with a smaller number of samples. Consequently, in scenarios where openness is large, the correlation between target and source data closely aligns, as illustrated in Fig.3a. However, when openness is reduced, the discrepancy between these distributions becomes more pronounced, negatively affecting the accuracy of the performance predictor, as seen in Fig.3b.

This discrepancy is not exclusive to scenarios with varying openness. For other datasets, even when openness is large, the score distributions between



(a) FMoW, Corrup. "FMoW-Test" 10% openness
(b) FMoW, Corrup. "FMoW-Test" 1% openness
(c) Terra Incognita, Corruption "L38" 10% openness
(d) Office Home, Corruption "Photo" 10% openness

Fig. 3: Examples of Poor Score Distribution



(a) VLCS, Corrup. "Caltech101" 10% openness
(b) VLCS, Corrup. "LabelMe" 10% openness
(c) PACS, Corrup. "Photo" 5% openness
(d) OfficeHome, Corrup. "Product" 1% openness

Fig. 4: Examples of Poor Accuracy Distribution

source and target significantly diverge. Importantly, in this work we focus on natural shifts rather than artificial shifts. We highlight that a strong correlation coefficient obtained from synthetic shifts does not necessarily translate to effective predictor performance in natural settings, as evidenced Figure 3c and Figure 3d.

Given these observations, we stress the importance of utilizing Mean Absolute Error instead of solely relying on correlation coefficients for evaluating performance prediction methods. This approach provides a more reliable measure of a predictor's quality, especially in the face of natural shifts, ensuring a more accurate assessment of its effectiveness.

Poor Accuracy Distribution Dataset-wide methods depend on data augmentations to mimic the distribution shift between source and target data. However, when the network demonstrates robustness to these augmentations and corruptions, the accuracy of augmented source samples remains relatively unchanged. Consequently, these corrupted versions fail to accurately represent the real distribution shift. This phenomenon is depicted in Figure 4a and Figure 4b, showing all source samples are clustered together far from the target samples.

Another challenge arises with a very small sample size. In such cases, the limited range of possible accuracies restricts the quality of data available for the Linear Regression model. For instance, having only 20 source samples means there are just 20 possible accuracy values. This limitation is visualized in Figure 4c and Figure 4d, where the points are binned across Y axis, indicating a constrained variability in accuracy due to the small number of samples.

References

1. 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
2. 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
3. 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
4. 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
5. 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
6. 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
7. Xie, R., Odonnat, A., Feofanov, V., Redko, I., Zhang, J., An, B.: Leveraging gradients for unsupervised accuracy estimation under distribution shift (2024)
8. 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