VARIATIONAL PSEUDO LABELING FOR TEST TIME DOMAIN GENERALIZATION

Sameer Ambekar[1], Zehao Xiao[1], Jiayi Shen[1], Xiantong Zhen[1,2]*, Cees G. M. Snoek[1]
[1]AIM Lab, University of Amsterdam [2]Inception Institute of Artificial Intelligence

ABSTRACT

This paper strives for domain generalization, where models are trained exclusively on source domains before being deployed at unseen target domains. We propose probabilistic pseudo labels of target samples for fine-tuning the source-trained model at test time, to generalize the model to the target domain. To do so, we formulate the adaptation as a variational inference problem by modeling pseudo labels as distributions. Variational pseudo labels are more robust to achieve a model better specified to the target domain. We learn the ability to generate better pseudo labels through simulating domain shifts during training. Experiments on widely-used datasets demonstrate the benefits, abilities and effectiveness of our proposal.

1 INTRODUCTION

As soon as test data distributions differ from the ones experienced during training, deep neural networks start to exhibit generalizability problems and accompanying performance degradation (Geirhos et al., 2018; Recht et al., 2019). To deal with the distribution shift, domain generalization (Muandet et al., 2013; Motiian et al., 2017; Li et al., 2017; 2020) has emerged as a promising tactic for generalizability to unseen target domains by learning more robust distributions from several source domains. However, as the methods are only trained on source domains, this may still lead to overfitting and limited guarantees for good performance on unseen target domains.

To better adapt models to target domains, without relying on target data during training, test-time adaptation, e.g., (Sun et al., 2020; Varsavsky et al., 2020; Wang et al., 2021) was introduced. It provides an alternative learning paradigm, by training a model on source data and further adjusting model according to the unlabeled target data at test time. Different settings for test-time adaptation have emerged. Test-time training by Sun et al. (2020) and test-time adaptation by Wang et al. (2021) attack image corruptions with a model trained on the original uncorrupted image distribution. The trained model is fine-tuned with self-supervised learning or entropy minimization to adapt to different corruptions. The paradigm is also employed under the domain generalization setting using multiple source domains during training (Iwasawa & Matsuo, 2021; Dubey et al., 2021; Xiao et al., 2022; Jang & Chung, 2022), where the domain shifts are typically manifested in varying image styles and scenes. In this paper, we focus on latter setting and refer to it as test-time domain generalization.

One widely applied strategy for updating models at test time is by optimizing or adjusting the model with target pseudo labels based on the source-trained model (Iwasawa & Matsuo, 2021; Jang & Chung, 2022). However, due to domain shifts, the source-model predictions of the target samples can be uncertain and inaccurate, leading to updated models that are overconfident on mispredictions. As a result, the obtained model becomes unreliable and misspecified to the target data (Wilson & Izmailov, 2020). To attack the unreliability of test-time domain generalization by pseudo labels we formulate it as a variational inference problem.

In this paper, we define pseudo labels as stochastic variables and estimate a distribution over them by variational inference. By doing so, the uncertainty in source-trained model predictions is incorporated into the generalization to the target data at test time, alleviating the misleading effects of uncertain and inaccurate pseudo labels. Thanks to the proposed probabilistic formalism, it is natural and convenient to utilize variational distributions to leverage extra information. By hinging on this

*Currently with United Imaging Healthcare, Co., Ltd., China.

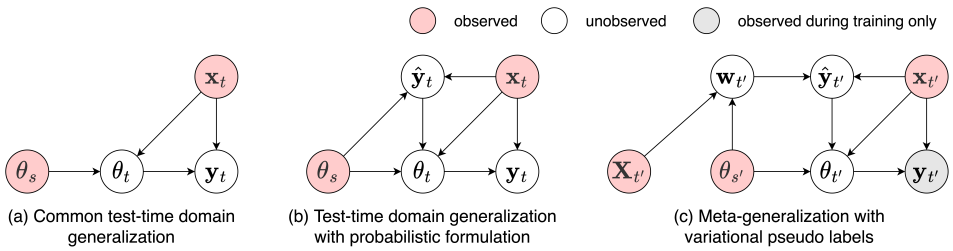


Figure 1: **Test-time domain generalization.** (a) The original test-time domain generalization algorithm (Iwasawa & Matsuo, 2021; Jang & Chung, 2022) obtains $\boldsymbol{\theta}_t$ by self learning of the unlabeled target data $\mathbf{x}_t$ on source-trained model $\boldsymbol{\theta}_s$. (b) Our probabilistic formulation models pseudo labels $p(\hat{\mathbf{y}}_t)$ for more robust generalization on the target data. (c) Furthermore, we propose variational pseudo labels to incorporate neighboring target information.

benefit, we design the variational pseudo labels to explore the neighboring information of target samples into the inference of the pseudo-label distributions. This makes the variational pseudo labels more accurate, which enables the source-trained model to be better specified to target data and therefore conducive to model generalization on the target domain. We expose the model domain shifts iteratively that facilitate learning the ability to generalize to new domains.

We conduct experiments on three widely-used domain generalization benchmarks to demonstrate the promise and effectiveness of our method.

2 METHODOLOGY

We are given data from different domains defined on the joint space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ denote the data space and label space, respectively. The domains are split into several source domains $\mathcal{D}_s = \{(\mathbf{x}_s, \mathbf{y}_s)^i\}_{i=1}^{N_s}$ and the target domain $\mathcal{D}_t = \{(\mathbf{x}_t, \mathbf{y}_t)^i\}_{i=1}^{N_t}$. The goal is to train a model on source domains that is expected to generalize well on the (unseen) target domain.

We follow the test-time domain generalization setting (Dubey et al., 2021; Iwasawa & Matsuo, 2021; Xiao et al., 2022), where a source-trained model is generalized to target domains by adjusting the model parameters at test time. A common strategy for adjusting the model parameters is that the model $\boldsymbol{\theta}$ is first trained on source data $\mathcal{D}_s$ and then at test time the source-trained model $\boldsymbol{\theta}_s$ is generalized to the target domain by optimization with certain surrogate losses, e.g., entropy minimization, based on unlabeled test data, which is formulated as:

$$\mathcal{L}_{test}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}_t \in \mathcal{D}_t}[L_E(\mathbf{x}_t; \boldsymbol{\theta}_s)], \tag{1}$$

where the entropy is calculated on the source model predictions. However, test samples from the target domain could be largely misclassified by the source model due to the domain shift, resulting in large uncertainty in the predictions. To solve those problems, in this work we address test-time domain generalization from a probabilistic perspective.

Probabilistic formulation. We first provide a probabilistic formulation for test-time domain generalization based on pseudo labels. Given the target sample $\mathbf{x}_t$ and the source-trained model $\boldsymbol{\theta}_s$, we would like to make predictions on the target sample. To this end, we formulate the predictive likelihood as follows:

$$p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_s) = \int p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t|\mathbf{x}_t, \boldsymbol{\theta}_s) d\boldsymbol{\theta}_t \approx p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_t^*), \tag{2}$$

where we use the value $\boldsymbol{\theta}_t^*$ obtained by the maximum a posterior (MAP) to approximate the integration (Finn et al., 2018). Intuitively, the MAP approximation is interpreted as inferring the posterior over $\boldsymbol{\theta}_t$: $p(\boldsymbol{\theta}_t|\mathbf{x}_t, \boldsymbol{\theta}_s) \approx \delta(\boldsymbol{\theta}_t = \boldsymbol{\theta}_t^*)$, which we obtain by fine-tuning $\boldsymbol{\theta}_s$ using the target data $\mathbf{x}_t$.

Pseudo labels as stochastic variables. To model the uncertainty of predictions for more robust generalization at test time, we treat pseudo labels as stochastic variables in the probabilistic framework as shown in Figure 1 (b). The pseudo labels are obtained from the source model predictions, which follows categorical distributions. Then we reformulate eq. (2) as follows:

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_s) &= \int p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_t) \Big[\int p(\boldsymbol{\theta}_t|\hat{\mathbf{y}}_t, \mathbf{x}_t, \boldsymbol{\theta}_s) p(\hat{\mathbf{y}}_t|\mathbf{x}_t, \boldsymbol{\theta}_s) d\hat{\mathbf{y}}_t \Big] d\boldsymbol{\theta}_t \\ &\approx \mathbb{E}_{p(\hat{\mathbf{y}}_t|\mathbf{x}_t, \boldsymbol{\theta}_s)}[p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_t^*)], \end{aligned} \tag{3}$$

where $\boldsymbol{\theta}_t^*$ is the MAP value of $p(\boldsymbol{\theta}_t|\hat{\mathbf{y}}_t, \mathbf{x}_t, \boldsymbol{\theta}_s)$, obtained via gradient descent on the data $\mathbf{x}_t$ and the corresponding pseudo labels $\hat{\mathbf{y}}_t$ starting from $\boldsymbol{\theta}_s$. The formulation allows us to sample different pseudo labels from the categorical distribution $p(\hat{\mathbf{y}}_t)$ to update the model $\boldsymbol{\theta}_t^*$, which takes into account the uncertainty of the source-trained predictions.

By approximating the expectation of $p(\hat{\mathbf{y}}_t)$ with the `argmax` function on $p(\hat{\mathbf{y}}_t)$, $\boldsymbol{\theta}_t^*$ is obtained by gradient descent based on only a point estimation of the pseudo labels $p(\hat{\mathbf{y}}_t)$. However, due to domain shifts, the `argmax` value of $p(\hat{\mathbf{y}}_t)$ is not guaranteed to be always correct. The optimization of the source-trained model then is similar to entropy minimization, where the updated model can achieve high confidence but wrong predictions of some target samples due to domain shifts. In contrast, the probabilistic formulation allows us to sample pseudo labels from the categorical distribution $p(\hat{\mathbf{y}}_t|\mathbf{x}_t, \boldsymbol{\theta}_s)$, which incorporates the uncertainty of the pseudo label in a principled way.

Variational pseudo labels. Under the probabilistic formalism, we derive variational inference of pseudo labels. We propose variational pseudo labels that incorporate information of the neighboring target samples to estimate pseudo-label distributions that are more robust against domain shifts. The variational pseudo labels are natural and convenient to deploy under the probabilistic formulation. Assume that we have a batch of meta-target data $\mathbf{X}_t = \{\mathbf{x}_t^i\}_{i=1}^M$, we derive the log-likelihood of the target samples as:

$$\begin{aligned}\log p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_s, \mathbf{X}_t) &= \log \int p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_t)\Big[\int\int p(\boldsymbol{\theta}_t|\hat{\mathbf{y}}_t, \mathbf{x}_t, \boldsymbol{\theta}_s)p(\hat{\mathbf{y}}_t, \mathbf{w}_t|\mathbf{x}_t, \boldsymbol{\theta}_s, \mathbf{X}_{t'})d\hat{\mathbf{y}}_t d\mathbf{w}_t\Big]d\boldsymbol{\theta}_t \\ &= \log \int\int p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_t^*)p(\hat{\mathbf{y}}_t, \mathbf{w}_t|\mathbf{x}_t, \boldsymbol{\theta}_s, \mathbf{X}_t)d\hat{\mathbf{y}}_t d\mathbf{w}_t,\end{aligned} \tag{4}$$

where $\boldsymbol{\theta}_t^*$ is the MAP value of $p(\boldsymbol{\theta}_t|\hat{\mathbf{y}}_t, \mathbf{x}_t, \boldsymbol{\theta}_s)$. We introduce the latent variable $\mathbf{w}_t$ to integrate the information of the neighboring target samples $\mathbf{X}_t$ as shown in Figure 1.

To approximate the true posterior of the joint distribution $p(\hat{\mathbf{y}}_t, \mathbf{w}_t)$, we design a variational posterior $q(\hat{\mathbf{y}}_t, \mathbf{w}_t|\mathbf{x}_t, \boldsymbol{\theta}_s, \mathbf{X}_t, \mathbf{Y}_t)$, where $\mathbf{Y}_t = \{\mathbf{y}_t^i\}_{i=1}^M$ denotes the actual labels of the meta-target data $\mathbf{X}_t$. To facilitate the estimation of pseudo labels, we set the prior distribution as:

$$p(\hat{\mathbf{y}}_t, \mathbf{w}_t|\mathbf{x}_t, \boldsymbol{\theta}_s, \mathbf{X}_t) = p(\hat{\mathbf{y}}_t|\mathbf{w}_t, \mathbf{x}_t)p_{\boldsymbol{\phi}}(\mathbf{w}_t|\boldsymbol{\theta}_s, \mathbf{X}_t) \tag{5}$$

where $p_{\boldsymbol{\phi}}(\mathbf{w}_t|\boldsymbol{\theta}_s, \mathbf{X}_t)$ is generated by the features of $\mathbf{X}_t$ together with their output values based on $\boldsymbol{\theta}_s$. Similarly, we define the variational posterior distribution as:

$$q(\hat{\mathbf{y}}_t, \mathbf{w}_t|\mathbf{x}_t, \boldsymbol{\theta}_s, \mathbf{X}_t, \mathbf{Y}_t) = p(\hat{\mathbf{y}}_t|\mathbf{w}_t, \mathbf{x}_t)q_{\boldsymbol{\phi}}(\mathbf{w}_t|\boldsymbol{\theta}_s, \mathbf{X}_t, \mathbf{Y}_t), \tag{6}$$

where $q_{\boldsymbol{\phi}}(\mathbf{w}_t|\boldsymbol{\theta}_s, \mathbf{X}_t, \mathbf{Y}_t)$ is obtained by the features of $\mathbf{X}_t$ and the actual labels $\mathbf{Y}_t$based on $\boldsymbol{\theta}_s$.

By introducing eqs. (5) and (6) into (4), we derive the evidence lower bound (ELBO) of the log-likelihood in eq. (4) as follows:

$$\begin{aligned}\log p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_s, \mathbf{X}_t) \geq\ & \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{w}_t)}\mathbb{E}_{p(\hat{\mathbf{y}}_t|\mathbf{w}_t, \mathbf{x}_t)}[\log p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_t^*)] \\ & - \mathbb{D}_{KL}[q_{\boldsymbol{\phi}}(\mathbf{w}_t|\boldsymbol{\theta}_s, \mathbf{X}_t, \mathbf{Y}_t)||p_{\boldsymbol{\phi}}(\mathbf{w}_t|\boldsymbol{\theta}_s, \mathbf{X}_t)].\end{aligned} \tag{7}$$

Rather than directly using the meta-source model $\boldsymbol{\theta}_s$, we estimate the pseudo labels $\mathbf{y}_t$ from the latent variable $\mathbf{w}_t$, which integrates the features of neighboring target samples. By considering the actual labels $\mathbf{Y}_t$, the variational distribution utilizes both the target information and categorical information of the neighboring samples. Thus, the variational posterior models the distribution of different categories in the target domain more reliably and produces more accurate pseudo labels to improve model generalization. To learn the ability to generate variational pseudo labels and achieve robust generalization across domains, we utilize meta-learning to mimic the test-time generalization procedure during training.

Test-time prediction. At test time, the model trained on the source domains $\boldsymbol{\theta}_s$ is generalized by further optimization on the target data.

The updated model is then evaluated on the (unseen) target data $\mathcal{D}_t$. We formulate the prediction as:

$$\begin{aligned}p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_s, \mathbf{X}_t) &= \int p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_t)\Big[\int p(\boldsymbol{\theta}_t|\hat{\mathbf{y}}_t, \mathbf{x}_t, \boldsymbol{\theta}_s)p(\hat{\mathbf{y}}_t, \mathbf{w}_t|\mathbf{x}_t, \boldsymbol{\theta}_s, \mathbf{X}_t)d\hat{\mathbf{y}}_t d\mathbf{w}_t\Big]d\boldsymbol{\theta}_t \\ &= \mathbb{E}_{p_{\boldsymbol{\phi}}(\mathbf{w}_t)}\mathbb{E}_{p(\hat{\mathbf{y}}_t|\mathbf{w}_t, \mathbf{x}_t)}[\log p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_t^*)],\end{aligned} \tag{8}$$

where $\boldsymbol{\theta}_t^*$ is the MAP value of $p(\boldsymbol{\theta}_t|\mathbf{x}_t, \hat{\mathbf{y}}_t, \boldsymbol{\theta}_s)$. $p_{\boldsymbol{\phi}}(\mathbf{w}_t) = p_{\boldsymbol{\phi}}(\mathbf{w}_t|\boldsymbol{\theta}_s, \mathbf{X}_t)$ is generated by the features of $\mathbf{X}_t$ according to the outputs or common pseudo labels based on $\boldsymbol{\theta}_s$.

3 RELATED WORK

Test-time domain generalization. Recent ways of mitigating domain shifts when deployed is test-time adaptation (Sun et al., 2020; Wang et al., 2021; Zhou & Levine, 2021). For domain generalization, (Iwasawa & Matsuo, 2021; Zhang et al., 2021b; Dubey et al., 2021; Xiao et al., 2022) adjust source trained model at test time for better generalization. We refer to these methods as test-time domain generalization. We explore test-time generalization under a probabilistic framework with the variational pseudo labels and mimic domain shifts to simulate domain generalization procedure based on our probabilistic formulation.

Pseudo-label learning. Pseudo labels for unsupervised domain adaptation (Shu et al., 2018; Zou et al., 2019) and test-time adaptation (Rusak et al., 2021; Chen et al., 2022; Wang et al., 2022) have been used for unlabelled target data. Our work is related to these works since we also use pseudo labels to adapt the source-trained model to the target domain. To make full use of pseudo labels under domain shifts, we model pseudo labels as distributions that have been previously unexplored. We draw samples from each pseudo label distribution and also obtain better pseudo labels by using neighboring target samples.

4 EXPERIMENTS

4.1 SETTINGS

We demonstrate the effectiveness of our method on image classification problems. We evaluate our method on three widely used domain generalization datasets: *PACS* (Li et al., 2017), *VLCS* (Fang et al., 2013), *TerraIncognita* (Beery et al., 2018). The details of the datasets are provided in Appendix A. We utilize ResNet-18 for all our experiments and ablation studies and report the accuracies on ResNet-50 for comparison as well that are pre-trained on Imagenet. We follow the training and validation split in (Li et al., 2017) and evaluate the model according to the "leave-one-out" protocol (Li et al., 2019; Carlucci et al., 2019).

In real-world applications, we usually obtain unlabeled target data in an online manner. To achieve continuous generalization and improvement of the model on target data, we increment the target data iteratively and keep updating and evaluating the model on the online target data akin to (Iwasawa & Matsuo, 2021).At test-time we use the learning rate of 0.0001 for all the layers and update all parameters. We provide more implementation details and computational cost in Appendix A.

4.2 ABLATION STUDY

Benefit of the probabilistic formulation. We first investigate the effectiveness of our probabilistic formulation of test-time domain generalization and the meta-learned variational pseudo labels. To demonstrate the benefits of probabilistic formulation, we conduct test-time domain generalization with eq. (3) and compare it with a common test-time domain generalization tactic (eq. 1). As shown in Table 4.2, our probabilistic test-time domain generalization already performs slightly better than the common one. When we incorporate the distribution of pseudo labels into the probabilistic formulation, and further propose the variational pseudo labels on the pseudo-label distributions (eq. 8) results improve further. We provide more ablation studies in Appendix B.

Table 1: **Benefit of the probabilistic formulation.** The experiments are conducted on PACS with ResNet-18. Our probabilistic formulation performs better than the common test-time domain generalization. Based on the probabilistic formulation, the variational pseudo labels further improve the performance. We provide per domain results in Appendix B.

Method	Equation	**Online**
Baseline	(1)	81.28
Probabilistic baseline	(3)	82.01
This paper	(8)	**83.45**

Table 2: **Comparisons on domain generalization datasets.** We provide the results for the online setting, averaged over five runs. Note that the test-time adaptation methods (with †) do not provide the results on domain generalization datasets in their paper. We report reimplemented results from Jang & Chung (2022). For all settings, our method performs at least competitive and sometimes better compared to state-of-the-art alternatives.

	PACS		VLCS		TerraIncognita	
	ResNet-18	ResNet-50	ResNet-18	ResNet-50	ResNet-18	ResNet-50
ERM baseline	79.3	83.2	74.9	75.5	40.6	45.4
Test-time adaptation on domain generalization						
Wang et al. (2021) †	80.8	83.7	67.0	69.7	39.9	43.9
Liang et al. (2020) †	82.4	84.1	65.2	67.0	33.6	35.2
Test-time domain generalization						
Iwasawa & Matsuo (2021)	81.7	84.5	76.5	**78.3**	41.6	45.9
Dubey et al. (2021)	-	84.1	-	78.0	-	47.3
Jang & Chung (2022)	81.9	84.1	**77.3**	77.6	42.6	47.4
Xiao et al. (2022)	84.1	**87.5**	-	-	-	-
This paper	**84.9** ±0.4	86.0 ±0.3	**77.8** ±0.5	**78.7** ±0.5	**46.2** ±0.4	**49.4** ±0.6

4.3 Comparisons

To further demonstrate the effectiveness of our method, we compare with state-of-the-art test-time domain generalization, as well as test-time adaptation methods. Note the latter methods are designed for single-source image corruption settings, so we report the reimplemented results from Jang & Chung (2022). Table 2 shows the results on PACS, VLCS, and TerraIncognita for both ResNet-18 and ResNet-50 backbones. We report the results of the proposed method with data augmentation Zhang et al. (2021a) at test time. Our method is competitive on all datasets, except for PACS with a ResNet-50 where the single sample generalization of Xiao et al. (2022) performs better.

Comparisons to existing DG method Our method is competitive on all datasets. We also demonstrate the efficacy of this paper with further comparison to Xiao et al. (2022). As shown in Figure 2, our method has low confidence in the uncertain samples, e.g., with different objectives or limited information, which shows that it's better at addressing unknown environments. After generalization at test time, the confidence of the correct category improves and the model predicts correctly, showing the effectiveness of test-time generalization with variational pseudo labels in complex environments. However, since we do generalization online and use variational labels, the proposed method solves some hard cases of the single sample generalization as provided in Xiao et al. (2022).

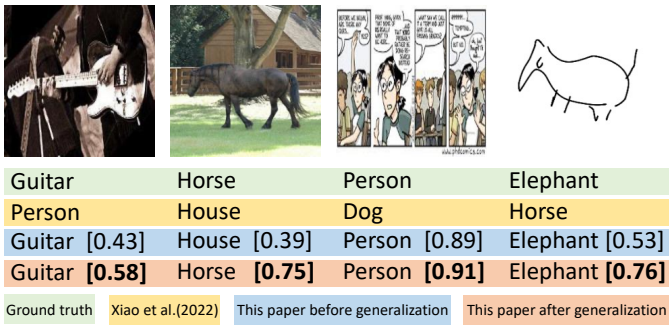


Figure 2: **Comparison on hard examples from Xiao et al. (2022).** The proposed method is more robust on samples with multiple objectives or complex environments.

5 Conclusion

We propose to cast test-time domain generalization as a probabilistic inference problem and model pseudo labels as distributions in the formulation. The probabilistic formulation mitigates the problem of updating the source model with incorrect pseudo labels or predictions. Based on the probabilistic formulation we propose variational pseudo labels which exposes the model to domain shifts and learn the ability to generalize.

ACKNOWLEDGMENTS

This work is financially supported by the Inception Institute of Artificial Intelligence, the University of Amsterdam and the allowance Top consortia for Knowledge and Innovation (TKIs) from the Netherlands Ministry of Economic Affairs and Climate Policy.

REFERENCES

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision*, pp. 456–473, 2018.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2229–2238, 2019.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 295–305, 2022.

Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain generalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14340–14349, 2021.

Yuming Fang, Weisi Lin, Zhenzhong Chen, Chia-Ming Tsai, and Chia-Wen Lin. A video saliency detection model in compressed domain. *IEEE transactions on circuits and systems for video technology*, 24(1):27–38, 2013.

Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Minguk Jang and Sae-Young Chung. Test-time adaptation via self-training with nearest neighbor information. *arXiv preprint arXiv:2207.10792*, 2022.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision*, pp. 5542–5550, 2017.

Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *IEEE International Conference on Computer Vision*, pp. 1446–1455, 2019.

Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex C Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *arXiv preprint arXiv:2009.12829*, 2020.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 6028–6039. PMLR, 2020.

Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *IEEE International Conference on Computer Vision*, pp. 5715–5725, 2017.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Evgenia Rusak, Steffen Schneider, George Pachitariu, Luisa Eck, Peter Vincent Gehler, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. If your data distribution shifts, use self-learning. *Transactions of Machine Learning Research*, 2021.

Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018.

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*, pp. 9229–9248. PMLR, 2020.

Thomas Varsavsky, Mauricio Orbes-Arteaga, Carole H Sudre, Mark S Graham, Parashkev Nachev, and M Jorge Cardoso. Test-time unsupervised domain adaptation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 428–436. Springer, 2020.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in Neural Information Processing Systems*, 33:4697–4708, 2020.

Zehao Xiao, Xiantong Zhen, Ling Shao, and Cees G M Snoek. Learning to generalize across domains on single test samples. In *International Conference on Learning Representations*, 2022.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *arXiv preprint arXiv:2110.09506*, 2021a.

Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. In *Advances in Neural Information Processing Systems*, volume 34, 2021b.

Aurick Zhou and Sergey Levine. Bayesian adaptation for covariate shift. *Advances in Neural Information Processing Systems*, 34:914–927, 2021.

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *IEEE International Conference on Computer Vision*, pp. 5982–5991, 2019.

A IMPLEMENTATION DETAILS AND COMPUTATIONAL COST

Implementation details Our train setup follows Iwasawa & Matsuo (2021). We use a batch size of 70 and train our method using the ERM algorithm Gulrajani & Lopez-Paz (2020). As stated, our backbones such as ResNet-18 and ResNet-50 are pretrained on ImageNet same as the previous methods. During training, the source-trained model with the highest validation accuracy is selected for adjustment on the target domain. We use similar settings and hyperparameters for all domain generalization benchmarks that have been reported in the paper. At test time, we choose the hyperparameters for model adjustment based on the validation set as mentioned in "In Search of lost domain generalization" Gulrajani & Lopez-Paz (2020) and T3A Iwasawa & Matsuo (2021). We will release the code. We train all our models on NVIDIA Tesla 1080Ti GPU.

We also provide the time cost of our method for inference stage (Table 3). Moreover, compared with the ERM baseline, our variational pseudo-label learning and meta-learning framework only introduce a few parameters. Since the meta-learning strategy only complicates the training process, our method has similar runtime during inference compared with the other test-time adaptation, e.g., Tent (Wang et al., 2021) and test-time domain generalization methods, e.g., T3A (Iwasawa & Matsuo, 2021) and TAST (Jang & Chung, 2022).

Datasets We provide a detailed information of datasets that we have utilized. *PACS* (Li et al., 2017) consists of 7 classes and 4 domains: Photo, Art painting, Cartoon, and Sketch with 9,991 samples. *VLCS* (Fang et al., 2013) consists of 5 classes from 4 different datasets: Pascal, LabelMe, Caltech, SUN with 10,729 samples. *TerraIncognita* (Beery et al., 2018) has 4 domains that taken by camera from 4 different locations: Location 100, Location 38, Location 43 and Location 46. The datasets includes 24,778 samples of 10 categories.

Table 3: **Runtime averaged for datasets using ResNet-18 as a backbone network.** The proposed method has similar or even better time costs at test time with the other test-time adaptation and test-time domain generalization methods.

	VLCS	PACS	Terra-Incognita
Tent (Wang et al., 2021)	7m 28s	3m 16s	10m 34s
Tent-BN (Wang et al., 2021)	2m 8s	33s	2m 58s
SHOT (Liang et al., 2020)	8m 9s	4m 22s	12m 40s
T3A (Iwasawa & Matsuo, 2021)	2m 9s	33s	2m 59s
TAST (Jang & Chung, 2022)	10m 34s	9m 30s	26m 14s
This paper	2m 20s	5m 33s	14m 30s

B DETAILED EXPERIMENTAL RESULTS AND ABLATIONS

Detailed results of the ablation study on the probabilistic formulation. We provide the detailed results of Table 4.2 on PACS in Table 5. The conclusion is similar to that of Table 4.2. The probabilistic formulation improves the performance of the common test-time domain generalization on most of the domains for online setting. Moreover, based on the probabilistic formulation, the proposed meta-learned variational pseudo labels further improves the performance obviously.

Benefit of meta-learning. We also investigate the importance of meta-learning in our method. We conduct the experiments with and without the meta-learning strategy under the offline test-time generalization setting. As shown in Table 4, the results with meta-generalization are better than that without the meta-generalization procedure, showing that the meta-learning strategy indeed helps in the proposed method. Without meta-learning, it is difficult for the model to learn the ability to handle domain shifts and generate better pseudo labels. Thus, there is a considerable decrease in accuracy.

Effectiveness of variational pseudo labels. To demonstrate the effectiveness of the variational pseudo label, we compare it with the normal pseudo labels drawn directly from the prediction distributions of source-trained models. We evaluate the methods in an offline test-time generalization setting with different amounts of target data. As shown in Figure 3(a), independent of the amount of

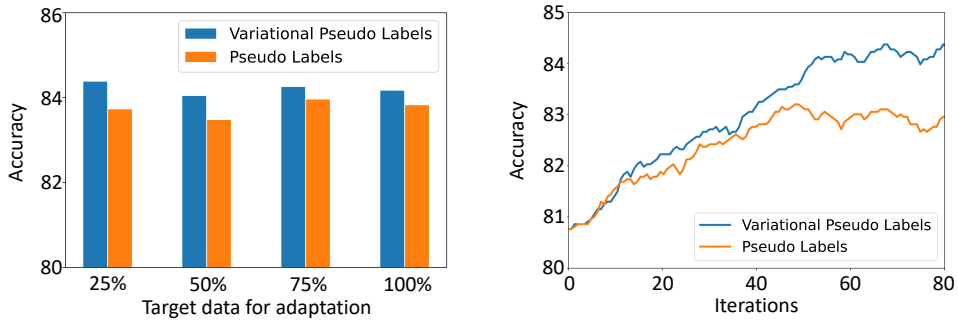
Table 4: **Benefit of meta-learning** on PACS. The meta-learning strategy mimics domain shifts during training, learning the ability to generate better variational pseudo labels and handle the domain shifts with the offline setting.

Meta-learning	**Photo**	**Art**	**Cartoon**	**Sketch**	*Mean*
✗	94.76	80.70	78.87	68.39	80.68
✓	**95.80**	**84.32**	**83.44**	**74.57**	**84.78**

Table 5: **Detailed ablations of the variational pseudo labels.** The experiments are conducted on PACS with ResNet-18. Our probabilistic formulation performs better than the common test-time domain generalization on most of the domains for online setting. Based on the probabilistic formulation, the variational pseudo labels with meta-learning strategy further improve the overall performance.

Method	Equation	**Photo**	**Art-painting**	**Cartoon**	**Sketch**	*Mean*
Test-time domain generalization	(1)	93.91	78.52	78.33	**74.37**	81.28
Probabilistic test-time domain generalization	(3)	94.55	80.07	79.14	74.29	82.01
This paper	(8)	**95.50** ±0.2	**82.90** ±0.4	**81.28** ±0.5	74.11 ±0.7	**83.45**

target data available for adaptation, the model updated by the variational pseudo labels achieves better overall results than the normal pseudo labels. We also provide the evaluation results along with optimization steps in Figure 3(b). Starting from the same baseline accuracy, the variational pseudo labels achieve faster generalization than the normal pseudo labels. Generalization with variational pseudo labels at test time is less prone to saturating in performance, leading to better final accuracy.



(a) (b)

Figure 3: **Effectiveness and calibration of the variational pseudo labels.** The experiments are conducted on PACS with the offline test-time generalization setting. (a) The model updated by our variational pseudo labels achieves better results, independent of the amount of target data. (b) A model with variational pseudo labels generalizes faster than one with common pseudo labels (shown for the *art-painting domain*).

Detailed results of the method with augmentation at test time. We provide detailed results in Table 6 where we compare the results of the paper on PACS dataset with and without augmented target samples at test time.

Table 6: **Variational pseudo labels combined with augmentation at test time** on PACS. We use data augmentation only at test time with Variational labels and do not use it during source training.

Data augmentation	**Photo**	**Art**	**Cartoon**	**Sketch**	*Mean*
✗	95.50	82.90	81.28	74.11	83.45
✓	**96.22**	**83.81**	**82.43**	**77.20**	**84.91**

Offline test-time domain generalization with limited target data. In addition to the experiments in the main paper, we also investigate the performance of our method with limited data for offline test-time domain generalization. As shown in Figure 4, the accuracy increases obviously with small numbers of target samples, e.g., 10% and 25%. However, the overall accuracy tends to saturate with an increase in target data for optimizing the source-trained model. This indicates that our method is able to achieve good generalization under the offline test-time domain generalization setting with even small amounts of target data, showing the applicability of the proposed method in practice.

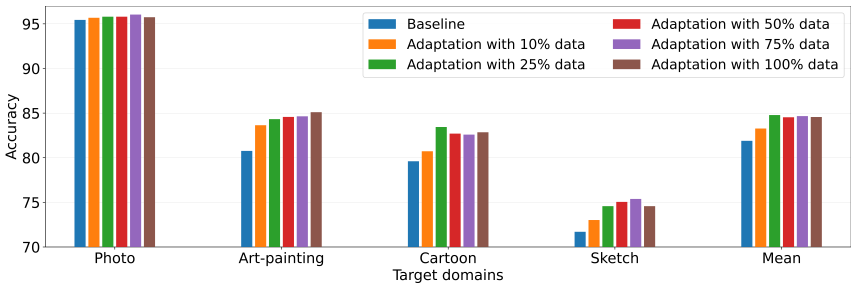


Figure 4: **Offine test-time domain generalization with different amounts of target data.** The experiments are conducted on PACS using ResNet-18 averaged over five runs. Under the offline setting, we observe that with increments in the amount of test data samples, the accuracy increases for each domain. Our method achieves good generalization with even small amounts of target data.