



## UvA-DARE (Digital Academic Repository)

### Probabilistic Test-Time Generalization by Variational Neighbor-Labeling

Ambekar, S.; Xiao, Z.; Shen, J.; Zhen, X.; Snoek, C.G.M.

**Publication date**

2024

**Document Version**

Final published version

**Published in**

Proceedings of Machine Learning Research

**License**

Other

[Link to publication](#)

**Citation for published version (APA):**

Ambekar, S., Xiao, Z., Shen, J., Zhen, X., & Snoek, C. G. M. (2024). Probabilistic Test-Time Generalization by Variational Neighbor-Labeling. *Proceedings of Machine Learning Research*, 274, 832-851. <https://proceedings.mlr.press/v274/ambekar25a.html>

**General rights**

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

# PROBABILISTIC TEST-TIME GENERALIZATION BY VARIATIONAL NEIGHBOR-LABELING

Sameer Ambekar<sup>\*◊1</sup>, Zehao Xiao<sup>\*1</sup>, Jiayi Shen<sup>1</sup>, Xiantong Zhen<sup>1,2‡</sup>, Cees G. M. Snoek<sup>1</sup>

<sup>1</sup>AIM Lab, University of Amsterdam <sup>2</sup>Core42

## ABSTRACT

This paper strives for domain generalization, where models are trained exclusively on source domains before being deployed on unseen target domains. We follow the strict separation of source training and target testing, but exploit the value of the unlabeled target data itself during inference. We make three contributions. First, we propose probabilistic pseudo-labeling of target samples to generalize the source-trained model to the target domain at test time. We formulate the generalization at test time as a variational inference problem, by modeling pseudo labels as distributions, to consider the uncertainty during generalization and alleviate the misleading signal of inaccurate pseudo labels. Second, we learn variational neighbor labels that incorporate the information of neighboring target samples to generate more robust pseudo labels. Third, to learn the ability to incorporate more representative target information and generate more precise and robust variational neighbor labels, we introduce a meta-generalization stage during training to simulate the generalization procedure. Experiments on seven 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 distribution shifts, domain generalization (Li et al., 2017; 2020; Motiian et al., 2017b; Muandet et al., 2013) has emerged as a promising tactic for generalizability to unseen target domains. However, as methods are only trained on source domains, this may still lead to overfitting and limited performance guarantees on unseen target domains.

To better adapt models to target domains – without relying on target data during training – test-time adaptation (Liang et al., 2023; 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 the model according to the unlabeled target data at test time. Different settings for test-time adaptation have emerged. Test-time training (Sun et al., 2020) and test-time adaptation (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 in an online manner. The paradigm is also employed under the domain generalization setting using multiple source domains during training (Dubey et al., 2021; Iwasawa & Matsuo, 2021; Jang et al., 2023; Xiao et al., 2022), where the domain shifts are typically manifested in varying image styles and scenes, rather than corruptions. In this paper, we focus on the 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 et al., 2023). 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 (Yi et al., 2023). As a result, the obtained model becomes unreliable and misspecified to the target data (Wilson & Izmailov, 2020). In this paper, we attack the unreliability of test-time domain generalization by pseudo labels and make the following three contributions.

First, we define pseudo labels as stochastic variables and estimate their distributions. By doing so, the uncertainty in predictions of the source-trained model is incorporated into the generalization to the target data at test time, alleviating the misleading effects of uncertain and inaccurate pseudo labels. Second, due to the proposed probabilistic formalism,

\* Equal contribution

◊ Currently with TU Munich, Germany

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

it is natural and convenient to utilize variational distributions to leverage extra information. By hinging on this benefit, we design variational neighbor labels that leverage the neighboring information of target samples into the inference of the pseudo-label distributions. This makes the variational 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. Third, to learn the ability to incorporate more representative target information in the variational neighbor labels, we simulate the test-time generalization procedure across domains by meta-learning. Beyond the well-known meta-source and meta-target stages (Alet et al., 2021; Dou et al., 2019; Xiao et al., 2022), we introduce a meta-generalization stage in between the meta-source and meta-target stages to mimic the target generalization procedure. Based on the multiple source domains seen during training, the model is exposed to different domain shifts iteratively and optimized to learn the ability to generalize to unseen domains. Our experiments on seven widely-used domain generalization benchmarks demonstrate the promise and effectiveness of our proposal.

## 2 RELATED WORK

**Domain generalization.** Domain generalization is introduced to learn a model on one or several source domains that can generalize well on any out-of-distribution target domain (Blanchard et al., 2011; Muandet et al., 2013; Zhou et al., 2022). Different from domain adaptation (Long et al., 2015; Luo et al., 2020; Wang & Deng, 2018), domain generalization methods do not access any target data during training. One of the most widely-used methods for domain generalization is domain-invariant learning (Arjovsky et al., 2019; Ghifary et al., 2016; Li et al., 2018c; Motiian et al., 2017a; Muandet et al., 2013; Zhao et al., 2020), which learns invariant feature representations across source domains. As an alternative, source domain augmentation methods (Li et al., 2018a; Qiao et al., 2020; Shankar et al., 2018; Zhou et al., 2020a;b) try to generate more source domains during training. Recently, meta-learning-based methods (Balaji et al., 2018; Chen et al., 2023a; Dou et al., 2019; Du et al., 2020; Li et al., 2018b) have been explored to learn the ability to handle domain shifts. We follow the meta-learning approach to domain generalization.

**Test-time adaptation.** Another approach to address distribution shifts without target data during training is adapting the model at test time. Source-free adaptation (Eastwood et al., 2021; Liang et al., 2020; Litrico et al., 2023) adapts the source-trained model to the entire target set. Differently, test-time adaptation achieves adaptation and prediction in an online manner, without halting inference. A common tactic is fine-tuning by entropy minimization (Wang et al., 2021; Goyal et al., 2022; Jang et al., 2023; Niu et al., 2022; Zhang et al., 2022). Since entropy minimization does not consider the uncertainty of source model predictions, probabilistic algorithms (Brahma & Rai, 2022; Zhou & Levine, 2021) based on Bayesian semi-supervised learning and models fine-tuned on soft pseudo labels (Rusak et al., 2021; Zou et al., 2019) have been proposed. Different from these works, we introduce the uncertainty by considering pseudo labels as latent variables and estimate their distributions by variational inference. Our models consider uncertainty within the same probabilistic framework, without introducing extra models or knowledge distillation operations.

**Test-time domain generalization.** Many test-time adaptation methods adjust models to corrupted data distributions with a single source distribution during training (Sun et al., 2020; Wang et al., 2021). The idea of adjusting the source-trained model at test time is further explored under the domain generalization setting to consider target information for better generalization (Dubey et al., 2021; Iwasawa & Matsuo, 2021; Xiao et al., 2023; Zhang et al., 2021). We refer to these methods as test-time domain generalization. Dubey et al. (2021) generate domain-specific classifiers for the target domain with the target domain embeddings. Iwasawa & Matsuo (2021) adjust their prototypical classifier online according to the pseudo labels of the target data. Some also investigated meta-learning for test-time domain generalization (Alet et al., 2021; Du et al., 2021; Xiao et al., 2022). These methods mimic domain shifts during training with multiple source domains. Du et al. (2021) meta-learn to estimate the batch normalization statistics from each target sample to adjust the source-trained model. Xiao et al. (2022) learn to adapt their classifier to each individual target sample by mimicking domain shifts during training. Our method also learns the ability to adjust the model by unseen data under the multi-source meta-learning setting. Differently, we design meta-generalization and meta-target stages during training to simulate both the generalization and inference procedures at test time. Our entire algorithm is explored under a probabilistic framework.

**Pseudo-label learning.** Pseudo-label learning relies on model predictions for retraining on downstream tasks. It is often applied for unlabeled data and self-training (Li et al., 2022; Miyato et al., 2018; Xie et al., 2020; Yalniz et al., 2019). To better utilize information from unlabeled target distributions, pseudo labels are also beneficial for unsupervised domain adaptation (Liu et al., 2021a; Shu et al., 2018; Zou et al., 2019), test-time adaptation (Chen et al., 2022; Rusak et al., 2021; Wang et al., 2022), and test-time domain generalization (Iwasawa & Matsuo, 2021; Jang et al., 2023; Wang et al., 2023). As pseudo labels can be noisy and overconfident (Zou et al., 2019), several studies focus on the appropriate selection and uncertainty of the pseudo labels. These works either select the pseudo labels with criteria such as the entropy consistency score of model predictions. (Liu et al., 2021a; Niu et al., 2022; Shin et al., 2022) or use soft pseudo

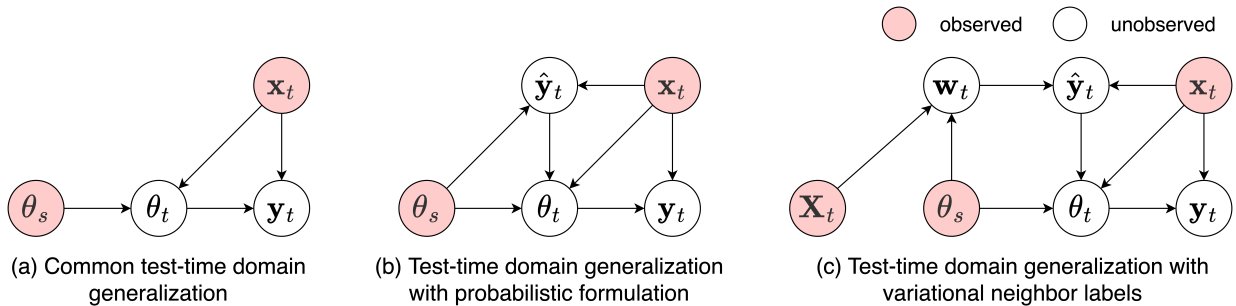


Figure 1: **Probabilistic modeling graph for test-time domain generalization.** (a) Common test-time domain generalization algorithm obtains the target model  $\theta_t$  by self-learning of the unlabeled target data  $\mathbf{x}_t$  on source-trained model  $\theta_s$  (Iwasawa & Matsuo, 2021; Jang et al., 2023). (b) We introduce pseudo labels  $p(\hat{\mathbf{y}}_t)$  as a latent variable to generate  $p(\theta_t)$  for more robust generalization. (c) We further propose variational neighbor labels to incorporate neighboring information into the generation of pseudo labels, where latent variable  $\mathbf{w}_t$  and  $\hat{\mathbf{y}}_t$  follow Gaussian and categorical distributions. We introduce a meta-generalization stage during training to optimize our model.

labels to take the uncertainty into account (Rusak et al., 2021; Yang et al., 2022; Zou et al., 2019). We also use pseudo labels to generalize the source-trained model to the target domain. Different from the previous methods, we are the first to introduce pseudo labels as latent variables in a probabilistic parameterized framework for test-time domain generalization, where we incorporate uncertainty and generate pseudo labels with neighboring information through variational inference and meta-learning.

### 3 METHODOLOGY

**Preliminary.** 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  $\mathcal{D}_t = \{(\mathbf{x}_t, \mathbf{y}_t)^i\}_{i=1}^{N_t}$ . Our 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  $\theta$  is first trained on source data  $\mathcal{D}_s$  by minimizing a supervised cross-entropy ( $L_{CE}$ ) loss  $\mathcal{L}_{train}(\theta) = \mathbb{E}_{(\mathbf{x}_s, \mathbf{y}_s)^i \in \mathcal{D}_s} [L_{CE}(\mathbf{x}_s, \mathbf{y}_s; \theta)]$ ; and then at test time the source-trained model  $\theta_s$  is generalized to the target domain by optimization with certain surrogate losses, e.g., entropy minimization ( $L_E$ ), based on the online unlabeled test data, which is formulated as:

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

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. Moreover, the entropy minimization tends to update the model with high confidence even for the wrong predictions, which would cause a misspecified model for the target domain. To solve those problems, we address test-time domain generalization from a probabilistic perspective and further propose variational neighbor labels to incorporate more target information. A graphical illustration to highlight the differences between common test-time domain generalization and our proposals is shown in Figure 1.

#### 3.1 PROBABILISTIC PSEUDO-LABELING

Given target sample  $\mathbf{x}_t$  and source-trained model  $\theta_s$ , we would like to make predictions on the target sample, formulated as  $p(\mathbf{y}_t | \mathbf{x}_t, \theta_s) = \int p(\mathbf{y}_t | \mathbf{x}_t, \theta_t) p(\theta_t | \mathbf{x}_t, \theta_s) d\theta_t$ . Since the distribution of  $p(\theta_t)$  is intractable, the common test-time adaptation and generalization methods usually optimize the source model to the target one by the maximum a posterior (MAP), which is an empirical Bayesian method and an approximation of the integration of  $p(\theta_t)$  (Finn et al., 2018). The predictive likelihood is then formulated as:

$$p(\mathbf{y}_t | \mathbf{x}_t, \theta_s) = \int p(\mathbf{y}_t | \mathbf{x}_t, \theta_t) p(\theta_t | \mathbf{x}_t, \theta_s) d\theta_t \approx p(\mathbf{y}_t | \mathbf{x}_t, \theta_t^*), \quad (2)$$

where  $\theta_t^*$  is the MAP value of the optimized target model. The MAP approximation is interpreted as inferring the posterior over  $\theta_t$ :  $p(\theta_t|\mathbf{x}_t, \theta_s) \approx \delta(\theta_t=\theta_t^*)$ , following a Dirac delta distribution.

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

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{x}_t, \theta_s) &= \int p(\mathbf{y}_t|\mathbf{x}_t, \theta_t) \left[ \int p(\theta_t|\hat{\mathbf{y}}_t, \mathbf{x}_t, \theta_s) p(\hat{\mathbf{y}}_t|\mathbf{x}_t, \theta_s) d\hat{\mathbf{y}}_t \right] d\theta_t \\ &\approx \mathbb{E}_{p(\hat{\mathbf{y}}_t|\mathbf{x}_t, \theta_s)} [p(\mathbf{y}_t|\mathbf{x}_t, \theta_t^*)], \end{aligned} \quad (3)$$

where  $\theta_t^*$  is the MAP value of  $p(\theta_t|\hat{\mathbf{y}}_t, \mathbf{x}_t, \theta_s)$ , obtained via gradient descent on the data  $\mathbf{x}_t$  and the corresponding pseudo labels  $\hat{\mathbf{y}}_t$  starting from  $\theta_s$ . Note that we only use MAP approximation with gradient descent to estimate the model parameter  $\theta_t$ , which will not hurt the generation of the probabilistic pseudo labels. This formulation allows us to sample different pseudo labels from the categorical distribution  $p(\hat{\mathbf{y}}_t)$  to update the model  $\theta_t^*$ , which takes into account the uncertainty of the source-trained predictions.

The common pseudo-labeling method can be treated as a specific case of eq. (3), which approximates the expectation of  $p(\hat{\mathbf{y}}_t)$  by utilizing the  $\text{argmax}$  function on  $p(\hat{\mathbf{y}}_t)$ , generating the hard pseudo labels.  $\theta_t^*$  is then obtained by a point estimation of the hard pseudo labels. However, due to domain shifts, the  $\text{argmax}$  value of  $p(\hat{\mathbf{y}}_t)$  is not guaranteed to always be correct. The optimization of the source-trained model then is similar to entropy minimization (eq. 1), where the updated model can achieve high confidence but wrong predictions of some target samples due to domain shifts. More analysis is provided in Appendix A.

### 3.2 VARIATIONAL NEIGHBOR LABELS

We rely on variational inference to approximate the true posterior of the probabilistic pseudo labels, in which we introduce more neighboring target information and categorical information during training. Introducing variational inference into pseudo-labeling is natural and convenient under the proposed probabilistic formulation. However, to generate pseudo labels that are more accurate and calibrated for more robust generalization, it is necessary to incorporate more target information. Assume that we have a mini-batch of target data  $\mathbf{X}_t = \{\mathbf{x}_t^i\}_{i=1}^M$ , we reformulate eq. (3) as:

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{x}_t, \theta_s, \mathbf{X}_t) &= \int p(\mathbf{y}_t|\mathbf{x}_t, \theta_t) \left[ \int \int p(\theta_t|\hat{\mathbf{y}}_t, \mathbf{x}_t, \theta_s) p(\hat{\mathbf{y}}_t, \mathbf{w}_t|\mathbf{x}_t, \theta_s, \mathbf{X}_t) d\hat{\mathbf{y}}_t d\mathbf{w}_t \right] d\theta_t \\ &= \int \int p(\mathbf{y}_t|\mathbf{x}_t, \theta_t^*) p(\hat{\mathbf{y}}_t, \mathbf{w}_t|\mathbf{x}_t, \theta_s, \mathbf{X}_t) d\hat{\mathbf{y}}_t d\mathbf{w}_t. \end{aligned} \quad (4)$$

As in eq. (3),  $\theta_t^*$  is the MAP value of  $p(\theta_t|\hat{\mathbf{y}}_t, \mathbf{x}_t, \theta_s)$ . 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 (c). To facilitate the estimation of the variational neighbor labels, we set the prior distribution as:

$$p(\hat{\mathbf{y}}_t, \mathbf{w}_t|\mathbf{x}_t, \theta_s, \mathbf{X}_t) = p(\hat{\mathbf{y}}_t|\mathbf{w}_t, \mathbf{x}_t) p_\phi(\mathbf{w}_t|\theta_s, \mathbf{X}_t), \quad (5)$$

where  $p_\phi(\mathbf{w}_t|\theta_s, \mathbf{X}_t)$  is generated by the features of  $\mathbf{X}_t$  together with their output values based on  $\theta_s$ . In detail, to explore the information of neighboring target samples, we first generate the predictions of  $\mathbf{X}_t$  by the source-trained model  $\theta_s$ . Then we estimate the averaged target features of each category according to the source-model predictions. The latent variable  $\mathbf{w}_t$  is obtained by the model  $\phi$  with the averaged features as the input. Therefore,  $\mathbf{w}_t$  contains the categorical information of the target features and can be treated as an updated classifier with more target information. The variational neighbor labels  $\hat{\mathbf{y}}_t$  are obtained by classifying the target samples using  $\mathbf{w}_t$ . Rather than directly using the source model  $\theta_s$ , we estimate  $\hat{\mathbf{y}}_t$  from the latent variable  $\mathbf{w}_t$ , which integrates the information of neighboring target samples to be more accurate and reliable.

To approximate the true posterior of the joint distribution  $p(\hat{\mathbf{y}}_t, \mathbf{w}_t)$  and incorporate more representative target information, we design a variational posterior  $q(\hat{\mathbf{y}}_t, \mathbf{w}_t|\mathbf{x}_t, \theta_s, \mathbf{X}_t, \mathbf{Y}_t)$  to supervise the prior distribution  $p(\hat{\mathbf{y}}_t, \mathbf{w}_t|\mathbf{x}_t, \theta_s, \mathbf{X}_t)$  during training:

$$q(\hat{\mathbf{y}}_t, \mathbf{w}_t|\mathbf{x}_t, \theta_s, \mathbf{X}_t, \mathbf{Y}_t) = p(\hat{\mathbf{y}}_t|\mathbf{w}_t, \mathbf{x}_t) q_\phi(\mathbf{w}_t|\theta_s, \mathbf{X}_t, \mathbf{Y}_t). \quad (6)$$

The variational posterior distribution is obtained similarly as the prior by generating  $\mathbf{w}_t$  through the categorical averaged features. The model  $\phi$  is shared by the prior and posterior distributions. The main difference is that the averaged features to generate  $\mathbf{w}_t$  are obtained with the actual target labels  $\mathbf{Y}_t$ . Since the target labels  $\mathbf{Y}_t$  are inaccessible, we can only utilize the prior distribution  $p(\hat{\mathbf{y}}_t, \mathbf{w}_t|\mathbf{x}_t, \theta_s, \mathbf{X}_t)$  at test time. Therefore, we introduce the variational posterior

under the meta-learning framework (Du et al., 2021; Finn et al., 2017; Xiao et al., 2022), where we mimic domain shifts and the test-time generalization procedure during training to learn the variational neighbor labels. In this case, the prior distribution  $p(\hat{\mathbf{y}}_t|\mathbf{w}_t, \mathbf{x}_t)p_\phi(\mathbf{w}_t|\boldsymbol{\theta}_s, \mathbf{X}_t)$  learns the ability to incorporate more representative target information and generate more accurate neighbor labels.

### 3.3 META-GENERALIZATION WITH VARIATIONAL NEIGHBOR LABELS.

We split the source domains  $\mathcal{D}_s$  into meta-source domains  $\mathcal{D}_{s'}$  and a meta-target domain  $\mathcal{D}_{t'}$  during training. The meta-target domain is selected randomly in each iteration to mimic diverse domain shifts. Moreover, we divide each iteration into meta-source, meta-generalization, and meta-target stages to simulate the training stage on source domains, test-time generalization, and test stage on target data, respectively.

**Meta-source.** We train the meta-source model  $\boldsymbol{\theta}_{s'}$  by minimizing the supervised loss  $L_{CE}(\mathbf{x}_{s'}, \mathbf{y}_{s'}; \boldsymbol{\theta})$ , where  $(\mathbf{x}_{s'}, \mathbf{y}_{s'})$  denotes the input-label sample pairs of the meta-source domains.

**Meta-generalization.** To mimic test-time generalization and prediction, our goal in the newly introduced meta-generalization stage is to optimize the meta-source model  $\boldsymbol{\theta}_{s'}$  by the meta-target data and make predictions with the generalized model. By introducing the variational neighbor labels, the log-likelihood of the meta-target prediction  $\mathbf{y}_{t'}$  is formulated as:

$$p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}) = \int \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'}^*)p(\hat{\mathbf{y}}_{t'}, \mathbf{w}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})d\hat{\mathbf{y}}_{t'}d\mathbf{w}_{t'}, \quad (7)$$

where  $\boldsymbol{\theta}_{t'}^*$  is the MAP value of  $p(\boldsymbol{\theta}_{t'}|\hat{\mathbf{y}}_{t'}, \mathbf{x}_{t'}, \boldsymbol{\theta}_{s'})$ , similar to eq. (4), and  $p(\hat{\mathbf{y}}_{t'}, \mathbf{w}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})=p(\hat{\mathbf{y}}_{t'}|\mathbf{w}_{t'}, \mathbf{x}_{t'})p_\phi(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})$  is the joint prior distribution of the meta-target neighbor labels  $\hat{\mathbf{y}}_{t'}$  and latent variable  $\mathbf{w}_{t'}$ . The joint variational posterior is designed as  $q(\hat{\mathbf{y}}_{t'}, \mathbf{w}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'})=p(\hat{\mathbf{y}}_{t'}|\mathbf{w}_{t'}, \mathbf{x}_{t'})q_\phi(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'})$  to learn more reliable neighbor labels by considering the actual labels  $\mathbf{Y}_{t'}$  of the meta-target data. Under this meta-learning setting, the actual labels  $\mathbf{Y}_{t'}$  of the meta-target data are accessible during source training. Thus, the variational distribution utilizes both the domain and categorical information of the neighboring samples and models the meta-target distribution more reliably, generating more accurate neighbor labels  $\hat{\mathbf{y}}_{t'}$  of the meta-target samples. With the variational neighbor labels  $\hat{\mathbf{y}}_{t'}$  the test-time domain generalization procedure is simulated by obtaining  $\boldsymbol{\theta}_{t'}^*$  from:

$$\boldsymbol{\theta}_{t'}^* = \boldsymbol{\theta}_{s'} - \lambda_1 \nabla_{\boldsymbol{\theta}} L_{CE}(\mathbf{x}_{t'}, \hat{\mathbf{y}}_{t'}; \boldsymbol{\theta}_{s'}), \quad \hat{\mathbf{y}}_{t'} \sim p(\hat{\mathbf{y}}_{t'}|\mathbf{w}_{t'}, \mathbf{x}_{t'}), \quad \mathbf{w}_{t'} \sim q_\phi(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'}), \quad (8)$$

where  $\lambda_1$  denotes the learning rate of the optimization in the meta-generalization stage.

**Meta-target.** Since our final goal is to obtain good performance on the target data after optimization, we further mimic the test-time inference on the meta-target domain and supervise the meta-target prediction on  $\boldsymbol{\theta}_{t'}^*$  by maximizing the log-likelihood of eq. (7):

$$\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}) \geq \mathbb{E}_{q_\phi(\mathbf{w}_{t'})} \mathbb{E}_{p(\hat{\mathbf{y}}_{t'}|\mathbf{w}_{t'}, \mathbf{x}_{t'})} [\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'}^*)] - \mathbb{D}_{KL}[q_\phi(\mathbf{w}_{t'})||p_\phi(\mathbf{w}_{t'})], \quad (9)$$

where  $p_\phi(\mathbf{w}_{t'})=p_\phi(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})$  generated by the features of  $\mathbf{X}_{t'}$  together with their output values based on  $\boldsymbol{\theta}_{s'}$ .  $q_\phi(\mathbf{w}_{t'})=q_\phi(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'})$  is obtained by the features of  $\mathbf{X}_{t'}$  considering the actual labels  $\mathbf{Y}_{t'}$ . The detailed formulation is provided in Appendix A.

As aforementioned, the actual labels  $\mathbf{y}_{t'}$  of the meta-target data are accessible during training. We can further supervise the updated model  $\boldsymbol{\theta}_{t'}^*$  on its meta-target predictions by the actual labels. Maximizing the log-likelihood  $\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})$  is equal to minimizing:

$$\mathcal{L}_{meta} = \mathbb{E}_{(\mathbf{x}_{t'}, \mathbf{y}_{t'})} [\mathbb{E}_{q_\phi(\mathbf{w}_{t'})} \mathbb{E}_{p(\hat{\mathbf{y}}_{t'}|\mathbf{w}_{t'}, \mathbf{x}_{t'})} L_{CE}(\mathbf{x}_{t'}, \mathbf{y}_{t'}; \boldsymbol{\theta}_{t'}^*)] + \mathbb{D}_{KL}[q_\phi(\mathbf{w}_{t'})||p_\phi(\mathbf{w}_{t'})]. \quad (10)$$

The source model  $\boldsymbol{\theta}_s$  in each iteration is finally updated by  $\boldsymbol{\theta}_s = \boldsymbol{\theta}_{s'} - \lambda_2 \nabla_{\boldsymbol{\theta}} \mathcal{L}_{meta}$ , where  $\lambda_2$  denotes the learning rate for the meta-target stage. Note that the loss in eq. (10) is computed on the  $\boldsymbol{\theta}_{t'}^*$  obtained by eq. (8), while the optimization is performed over the meta-source model  $\boldsymbol{\theta}_{s'}$ . Intuitively, the model updated by the meta-target neighbor labels is trained to achieve good performance on the meta-target data. Thus, the meta-generalization stage is further supervised to optimize the model well across domains and better generate and utilize the variational neighbor labels.

The variational inference model  $\phi$  is also optimized in the meta-target stage. To guarantee that the variational neighbor labels do extract the categorical neighboring information for classification, we add an extra cross-entropy loss ( $L_{CE}$ ) on the variational neighbor labels with actual labels during the meta-target stage. Thus,  $\phi$  is updated with a learning rate  $\lambda_3$  by  $\phi = \phi - \lambda_3 (\nabla_{\phi} L_{CE} + \nabla_{\phi} \mathcal{L}_{meta})$ . By simulating distribution shifts during training, the model learns the ability to generate more effective pseudo labels for fine-tuning the model across distribution shifts. The variational neighbor labels are further improved by considering more neighboring target information.

### 3.4 TEST-TIME GENERALIZATION.

At test time, the model trained on the source domains with the meta-learning strategy  $\theta_s$  is generalized to  $\theta_t^*$  by further optimization:

$$\theta_t^* = \theta_s - \lambda_1 \nabla_{\theta} L_{\text{CE}}(\mathbf{x}_t, \hat{\mathbf{y}}_t; \theta_s), \quad \hat{\mathbf{y}}_t \sim p(\hat{\mathbf{y}}_t | \mathbf{w}_t, \mathbf{x}_t), \quad \mathbf{w}_t \sim p_{\phi}(\mathbf{w}_t | \theta_s, \mathbf{X}_t). \quad (11)$$

Since the target labels  $\mathbf{Y}_t$  are inaccessible, we generate neighbor labels  $\hat{\mathbf{y}}_t$  and latent variables  $\mathbf{w}_t$  from the prior distribution  $p(\hat{\mathbf{y}}_t, \mathbf{w}_t | \mathbf{x}_t, \theta_s, \mathbf{X}_t) = p(\hat{\mathbf{y}}_t | \mathbf{w}_t, \mathbf{x}_t) p_{\phi}(\mathbf{w}_t | \theta_s, \mathbf{X}_t)$ . The distribution  $p(\mathbf{w}_t)$  is inferred as a Gaussian distribution by generating the mean  $\mu$  and variance  $\sigma$  using the target averaged features through  $\phi$ . Then we sample  $\mathbf{w}_t$  by Monte Carlo sampling and generate the categorical distribution  $p(\hat{\mathbf{y}}_t)$  with the input target features, which we utilize to obtain the MAP value  $\theta_t^*$ . From  $\theta_t^*$  we make predictions on the (unseen) target data  $\mathcal{D}_t$ , formulated as:

$$\begin{aligned} p(\mathbf{y}_t | \mathbf{x}_t, \theta_s, \mathbf{X}_t) &= \int p(\mathbf{y}_t | \mathbf{x}_t, \theta_t) \left[ \int p(\theta_t | \hat{\mathbf{y}}_t, \mathbf{x}_t, \theta_s) p(\hat{\mathbf{y}}_t, \mathbf{w}_t | \mathbf{x}_t, \theta_s, \mathbf{X}_t) d\hat{\mathbf{y}}_t d\mathbf{w}_t \right] d\theta_t \\ &= \mathbb{E}_{p_{\phi}(\mathbf{w}_t)} \mathbb{E}_{p(\hat{\mathbf{y}}_t | \mathbf{w}_t, \mathbf{x}_t)} [\log p(\mathbf{y}_t | \mathbf{x}_t, \theta_t^*)]. \end{aligned} \quad (12)$$

We provide both the training algorithm and test-time generalization algorithm in Appendix B.

## 4 EXPERIMENTS

### 4.1 DATASETS AND IMPLEMENTATION DETAILS

**Seven datasets.** We demonstrate the effectiveness of our method on seven widely used domain generalization datasets for image classification: *PACS* (Li et al., 2017), *VLCS* (Fang et al., 2013), *Office-Home* (Venkateswara et al., 2017), *TerraIncognita* (Beery et al., 2018), *Mini DomainNet* (Zhou et al., 2021), Rotated MNIST and Fashion MNIST (Piratla et al., 2020).

*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, and SUN with 10,729 samples. *Office-Home* (Venkateswara et al., 2017) contains 15,500 images of 65 categories from 4 domains, i.e., Art, Clipart, Product, and Real-World. *TerraIncognita* (Beery et al., 2018) has 4 domains and 4 locations: L100, L38, L43, and L46. The dataset includes 24,778 samples of 10 categories. *Mini DomainNet* (Zhou et al., 2021) is subset of DomainNet (Peng et al., 2019) with 140,000 samples, 4 domains and 126 classes. We follow training and validation split in (Li et al., 2017) and evaluate model according to “leave-one-out” protocol (Li et al., 2019; Carlucci et al., 2019). We also evaluate our method on *Rotated MNIST* and *Fashion-MNIST* following Piratla et al. (2020), where images are rotated by different angles as different domains. We use subsets with rotation angles from  $15^\circ$  to  $75^\circ$  in intervals of  $15^\circ$  as 5 source domains, images rotated by  $0^\circ$  and  $90^\circ$  as target domains.

**Implementation details.** We utilize ResNet-18 for all our experiments and ablation studies and report the accuracies on ResNet-50 for comparison as well. We evaluate the method on the online test-time domain generalization setting (Iwasawa & Matsuo, 2021), we increment the target data iteratively and keep updating and evaluating the model. When we report an ERM baseline, it means we directly evaluate the source-trained model without any adjustment at test time (Gulrajani & Lopez-Paz, 2020). Backbones are pretrained on ImageNet, like previous work. We also report two additional baselines: “hard pseudo label” is obtained using the *argmax* of model predictions and “soft pseudo label” which are defined in implementation details. Optimization with soft pseudo-labeling is similar to entropy minimization.

During training, we use a varied learning rate throughout and train for 10,000 iterations. The model is trained on all source domains. Note we utilize all available source data for both model training and simulating test-time generalization by randomly selecting and changing the meta-source and meta-target domains in each iteration, without sacrificing training data. More clarifications are in the appendix. In the meta-generalization procedure, we set the learning rate  $\lambda_1$  as  $1e - 4$  for all layers. During meta-target, we set the learning rate for the pretrained ResNet ( $\lambda_2$ ) to  $5e - 5$  and the learning rate of the variational module  $\phi$  ( $\lambda_3$ ) and classifiers as  $1e - 4$  for all datasets. The batch size is set to 70 during the training and set to 20 during test-time generalization. At test time, we use the learning rate of  $1e - 4$  for all the layers and update all parameters. We utilize test-time data augmentation Zhang et al. (2022) for our results. We choose the hyperparameters for model adjustment based on the training-domain validation set, following Gulrajani & Lopez-Paz (2020) and Iwasawa & Matsuo (2021). There are no additional hyperparameters involved in our method. We train and evaluate all our models on one NVIDIA Tesla 1080Ti GPU and utilize the PyTorch framework. We run all the experiments with 5 different random seeds. We use similar settings and hyperparameters for all domain generalization benchmarks. The method introduces a small computational cost for inference at test time and about 1% more parameters than the backbone model. The time cost for test-time generalization is competitive with other fine-tuning methods, with 5m 33s on PACS. We provide more implementation details and detailed computational costs in Appendix C.

Table 1: **State-of-the-art comparisons** for ResNet-18 (RN18) and ResNet-50 (RN50) backbones. Our results are averaged over five runs. Test-time adaptation results by Wang et al. (2021) and Liang et al. (2020) for domain generalization provided by Jang et al. (2023). Gray numbers for Xiao et al. (2022) based on our reimplementaion. Our method is either best (bold) or runner-up (underlined).

	PACS		VLCS		Office-Home		TerraIncognita	
	RN18	RN50	RN18	RN50	RN18	RN50	RN18	RN50
<b>Standard domain generalization</b>								
ERM baseline	80.3	85.7	75.8	77.4	61.0	67.5	35.8	47.2
Arjovsky et al. (2019)	80.9	83.5	75.1	78.5	58.0	64.3	38.4	47.6
Huang et al. (2020)	80.5	85.2	75.4	77.1	58.4	65.5	39.4	46.6
Shi et al. (2022)	82.0	85.5	76.9	77.8	62.0	68.6	40.2	45.1
Eastwood et al. (2022)	-	86.5	-	77.8	-	67.5	-	47.8
Li et al. (2023)	-	86.6	-	78.9	-	68.9	-	48.6
<b>Test-time adaptation on domain generalization</b>								
Wang et al. (2021)	83.9	85.2	72.9	73.0	60.9	66.3	33.7	37.1
Liang et al. (2020)	82.4	84.1	65.2	67.0	62.6	67.7	33.6	35.2
<b>Test-time domain generalization</b>								
Iwasawa & Matsuo (2021)	81.7	85.3	76.5	<b>80.0</b>	57.0	68.3	41.6	47.0
Dubey et al. (2021)	-	84.1	-	78.0	-	67.9	-	47.3
Jang et al. (2023)	81.9	84.1	77.3	77.6	63.7	68.6	42.6	47.4
Chen et al. (2023b)	83.8	-	76.9	-	62.0	-	43.2	-
Xiao et al. (2022)	<u>84.1</u>	<u>87.5</u>	<u>77.8</u>	78.6	<b>66.0</b>	<b>71.0</b>	<u>44.8</u>	<u>48.4</u>
<b>This paper</b>	<b>85.0</b> $\pm 0.4$	<b>87.9</b> $\pm 0.3$	<b>78.2</b> $\pm 0.3$	<u>79.1</u> $\pm 0.4$	<u>64.3</u> $\pm 0.3$	<u>69.1</u> $\pm 0.4$	<b>46.9</b> $\pm 0.4$	<b>49.4</b> $\pm 0.6$

## 4.2 RESULTS

**State-of-the-art comparisons.** We compare our proposal with state-of-the-art test-time domain generalization, as well as some standard domain generalization and test-time adaptation methods. Note the latter methods are designed for single-source image corruption settings, so we report the reimplemented results from Jang et al. (2023). Table 1 shows the results on PACS, VLCS, Office-Home, and TerraIncognita for both ResNet-18 and ResNet-50 backbones. Our method is competitive on most of the datasets, except for Office-Home where the sample-wise generalization of Xiao et al. (2022) performs better.

The reason can be that the representative neighboring information is more difficult to incorporate with a larger number of categories (e.g., 65 in Office-Home), which needs larger capacity models  $\phi$ . We have experimented with  $\phi$  values and obtained a mean accuracy of 57.1 with 2 layers and a mean accuracy of 64.3 with 3 layers in  $\phi$ . Note that our method still outperforms other recent methods (Chen et al., 2023b; Iwasawa & Matsuo, 2021; Jang et al., 2023; Wang et al., 2021) on Office-Home. Moreover, since we consider the uncertainty of the variational neighbor labels, the proposed method solves some hard cases of the single-sample approach reported in 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, showing good calibration of our method. With the proposed method, the model





Source: art; cartoon; sketch Target: photo	Source: art; cartoon; sketch Target: photo	Source: art; photo; sketch Target: cartoon	Source: art; cartoon; photo Target: sketch
			
Guitar	Horse	Person	Elephant
Person	House	Dog	Horse
Guitar [0.43]	House [0.39]	Person [0.89]	Elephant [0.53]
Guitar [0.58]	Horse [0.75]	Person [0.91]	Elephant [0.76]
Ground truth	Xiao et al.(2022)	This paper before generalization	This paper after generalization

Figure 2: **Comparison on hard examples from Xiao et al. (2022)** on PACS. Our proposal is more robust on samples with multiple objectives or complex scenes.

predicts these hard cases correctly, showing the effectiveness of test-time generalization with the meta-generalized variational neighbor labels in complex scenes. In addition, there are also some recent standard domain generalization methods achieving good performance. For instance, (Gao et al., 2022) achieved good results on PACS, VLCS, Office-Home, and TerraIncognita based on ResNet-50 by utilizing an extra dataset before training to meta-learn loss function. This implies that we can also improve by utilizing more datasets during training.

Table 4: **Benefit of method components.** Results on PACS and TerraIncognita with ResNet-18. Our probabilistic formulation performs better than the common pseudo-labeling baseline for test-time domain generalization by considering the uncertainty. Incorporating more target information by the variational neighbor labels improves results further, especially when used in concert with meta-generalization. We provide per-domain results in Appendix F.

		PACS	TerraIncognita
Pseudo-labeling baseline	(eq. 1)	81.3 $\pm$ 0.3	41.2 $\pm$ 0.4
Probabilistic pseudo-labeling	(eq. 3)	82.9 $\pm$ 0.3	43.5 $\pm$ 0.5
Variational neighbor-labeling	(eq. 4)	83.7 $\pm$ 0.4	44.8 $\pm$ 0.5
Meta-generalization with variational neighbor labels	(eq. 10)	85.0 $\pm$ 0.4	46.9 $\pm$ 0.4

**Comparisons on mini-DomainNet.** We also conduct experiments on mini-DomainNet with ResNet-18 to provide more comparisons as shown in Table 2. The conclusion is similar to the other datasets, we are at least competitive and sometimes better than alternative test-time adaptation and generalization methods. We provide more comparisons to the standard domain generalization methods in Appendix E, as well as detailed comparisons of each single domain. Our method also achieves competitive performance on these smaller- and larger-scale datasets.

Table 2: **Comparison on mini-DomainNet.** Our method performs best compared to state-of-the-art alternatives.

	Overall accuracy
ERM baseline	60.5
Peng et al. (2019)	<b>62.9</b>
Huang et al. (2020)	62.6
Blanchard et al. (2021)	62.1
Nam et al. (2021)	61.3
Scalbert et al. (2021)	61.9
Lee & Lee (2023)	62.0
<b>This paper</b>	<b>63.1 <math>\pm</math>0.3</b>

**Comparison on rotated MNIST and Fashion-MNIST.** We also conduct experiments on rotated MNIST, rotated Fashion-MNIST to provide another comparison as shown in Table 3. For rotated MNIST and Fashion-MNIST, we follow the settings in Piratla et al. (2020) and use ResNet-18 as the backbone. The conclusion is similar to the other datasets, we are at least competitive and sometimes better than alternative test-time adaptation and generalization methods.

Table 3: **Comparison on rotated MNIST and Fashion-MNIST.** The models are evaluated on the test sets of MNIST and Fashion-MNIST with rotation angles of 0° and 90°. Again, our method achieves the best performance (bold) compared to alternatives.

	MNIST	Fashion-MNIST
ERM Baseline	93.5	76.9
Wang et al. (2021)	95.3	78.9
Xiao et al. (2022)	<b>95.8</b>	<b>80.8</b>
<b>This paper</b>	<b>95.9 <math>\pm</math>0.1</b>	<b>82.4 <math>\pm</math>0.2</b>

### 4.3 ABLATIONS

**Benefit of method components.** To show the benefits of the individual components of our method, we conduct an ablation on PACS and TerraIncognita. We first compare the probabilistic pseudo-labeling (eq. 3) with the common one (eq. 1). As shown in the first two rows in Table 4, the probabilistic formulation performs better, which demonstrates the benefits of modeling uncertainty of the pseudo labels during generalization at test time. By incorporating more target information from the neighboring samples (eq. 4), the variational neighbor labels become more reliable, which benefits generalization on the target data. With the meta-generalization strategy (eq. 10), we learn the ability to incorporate more representative target information leading to further performance improvements. To show the benefits of meta-generalization, we conduct additional experiments for meta-generalization with pseudo-labeling and meta-generalization with probabilistic pseudo-labeling, which achieve mean accuracy of 82.2 and 83.9 on PACS, respectively. Meta-generalization improves all pseudo-labeling methods and the effect is most pronounced for our variational pseudo-labeling.

**Generalization in complex scenarios.** By considering the uncertainty and including more target information in the pseudo labels, our method can handle more complex test-time generalization scenarios. To demonstrate this ability, we conduct experiments with multiple target distributions on Rotated MNIST, as defined by Xiao et al. (2022). Specifically, we use 0°, 15°, 75° and 90° as source domains and 30°, 45° and 60° as target. As shown in Table 5, soft pseudo label achieves good results on the single target domains, while it is unable to outperform ERM baseline on the multiple target domains. The proposed method performs well under both settings and better than Xiao et al. (2022), which achieves generalization on each sample, demonstrating the generalization ability of our method in more complex scenarios.

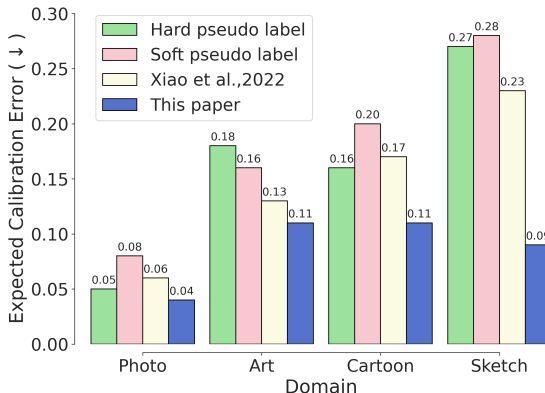
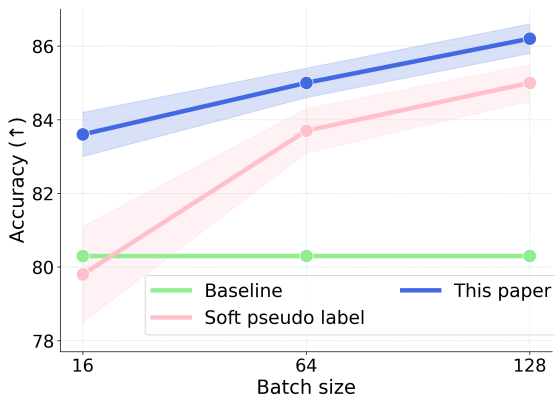
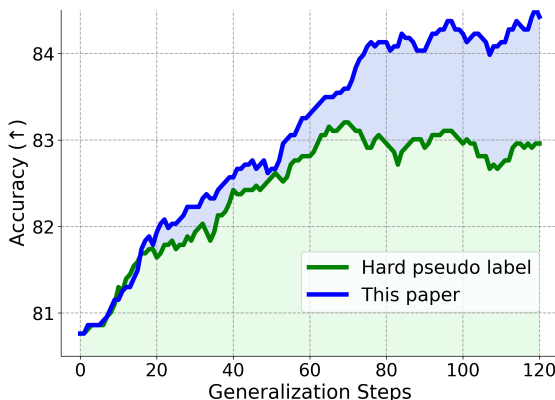
Table 5: **Generalization in complex scenarios.** Our method generalizes well on both single and multiple target distributions on rotated MNIST.

Settings	ERM baseline	Soft pseudo label	Xiao et al. (2022)	<i>This paper</i>
Single target distribution	95.6	96.5	96.9	<b>97.5 <math>\pm 0.3</math></b>
Multiple target distributions	95.6	95.6	96.9	<b>97.4 <math>\pm 0.3</math></b>

**Calibration ability.** We also investigate the calibration ability by measuring the Expected Calibration Error (Guo et al., 2017). We report hard and soft pseudo-labeling as baselines, as well as the results of Xiao et al. (2022) that considers uncertainty by variational inference. As shown in Figure 3, the error of our method is lower than the alternatives on all domains, demonstrating a better ability to model uncertainty at test time. By incorporating pseudo labels as latent variables with variational inference and considering neighboring target information, the proposed method models the uncertainty of the target samples more accurately. With the better-calibrated labels, the model achieves more robust generalization on the target domain at test time.

**Generalization with varying batch sizes.** Test-time generalization and adaptation methods usually require large batches of target samples to update their source-trained model. However, during real-world deployment, the number of available target samples may be limited. It is difficult to collect large batches of samples from the same target distribution during inference. In this case, the performance of common test-time generalization algorithms is constrained. In Figure 4 we compare with soft pseudo-labeling on PACS for varying batch sizes. Soft pseudo-labeling performs well with large batch sizes, but suffers with smaller batch sizes, e.g., 16, and is worse than the ERM baseline. By contrast, our method consistently achieves good results even with small target batch sizes. Demonstrating the benefit of incorporating the uncertainty and neighboring information. We provide more detailed results in Appendix F.

**Generalization along with inference.** For more insights into the variational neighbor labels, we provide the online performance along with generalization steps for the ‘art’ domain from PACS. As shown in Figure 5, starting from the same baseline accuracy, the gap between the results of variational neighbor labels and the hard pseudo labels becomes larger and larger along with the generalization steps. Variational neighbor labels achieve faster generalization of the source-trained model. After 50 iterations, the performance of the hard pseudo labels is saturated and even drops due to the error accumulation resulting from inaccurate pseudo labels during model updating. By considering the uncertainty and neighboring information, our variational neighbor labels improve performance and are less prone to saturation, leading to better accuracy.

Figure 3: **Calibration ability** on PACS. Variational neighbor labels consistently have a lower Expected Calibration Error.Figure 4: **Generalization with varying batch sizes.** Our method outperforms soft pseudo label on PACS, independent of batch size. Largest improvement for small batches.Figure 5: **Generalization along with inference.** Our method achieves faster generalization with less prone to saturation.

**Importance of the model adjustment at test time.** We also provide ablation studies on the pseudo-label generation and generalization with our variational neighbor labels. We directly make predictions using the variational neighbor labels generated by sampling from the pseudo label distributions at test time. The predictions based on the variational neighbor labels distributions (82.40%) are better than the ERM baseline (80.30%), demonstrating that our variational neighbor labels are better than the original prediction of the source model, i.e., the common pseudo labels. Moreover, after online adjusting the model parameters by our variational neighbor labels, the performance further improves (85.00%), demonstrating the effectiveness of the model adjustment at test time. Detailed results are shown in Table 12 in Appendix.

**Limitations.** Our approach follows common setup in test-time domain generalization, where multiple source domains and a small batch of target samples are available during training. We efficiently use the source training data for both model training and simulating test-time generalization through meta-learning and neighboring target information, which may be a limitation in some environments. We consider the single-source and single-target-sample variant of the method, with simulated or generated data, as an appealing direction for future work.

## 5 CONCLUSION

We cast test-time domain generalization as a probabilistic inference problem and model pseudo labels as latent variables in the formulation. By incorporating the uncertainty of the pseudo labels, the probabilistic formulation mitigates updating the source-trained model with inaccurate supervision, which arises due to domain shifts and leads to misspecified models. Based on the probabilistic formulation, we further propose variational neighbor labels under the designed meta-generalization setting, which estimates the pseudo labels by incorporating neighboring target information through variational inference and learns the ability to generalize the source-trained model. Ablation studies and further comparisons show the benefits, abilities, and effectiveness of our method on seven common domain generalization datasets.

## ACKNOWLEDGMENTS

This work is financially supported by Core42, 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

- Ferran Alet, Maria Bauza, Kenji Kawaguchi, Nurullah Giray Kuru, Tomás Lozano-Pérez, and Leslie Kaelbling. Tailoring: encoding inductive biases by optimizing unsupervised objectives at prediction time. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29206–29217, 2021.
- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.
- Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. MetaReg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems*, volume 31, pp. 998–1008, 2018.
- Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision*, pp. 456–473, 2018.
- Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2178–2186, 2011.
- Gilles Blanchard, Aniket Anand Deshmukh, Ürün Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22(1):46–100, 2021.
- Dhanajit Brahma and Piyush Rai. A probabilistic framework for lifelong test-time adaptation. *arXiv preprint arXiv:2212.09713*, 2022.
- 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.

- Jin Chen, Zhi Gao, Xinxiao Wu, and Jiebo Luo. Meta-causal learning for single domain generalization. *arXiv preprint arXiv:2304.03709*, 2023a.
- Liang Chen, Yong Zhang, Yibing Song, Ying Shan, and Lingqiao Liu. Improved test-time adaptation for domain generalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023b.
- Qi Dou, Daniel C Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Advances in Neural Information Processing Systems*, 2019.
- Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees G M Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. In *European Conference on Computer Vision*, pp. 200–216, 2020.
- Yingjun Du, Xiantong Zhen, Ling Shao, and Cees G M Snoek. MetaNorm: Learning to normalize few-shot batches across domains. In *International Conference on Learning Representations*, 2021.
- 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.
- Cian Eastwood, Ian Mason, Christopher KI Williams, and Bernhard Schölkopf. Source-free adaptation to measurement shift via bottom-up feature restoration. *arXiv preprint arXiv:2107.05446*, 2021.
- Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *arXiv preprint arXiv:2207.09944*, 2022.
- 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, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.
- Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.
- Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.
- Boyan Gao, Henry Gouk, Yongxin Yang, and Timothy Hospedales. Loss function learning for domain generalization by implicit gradient. In *International Conference on Machine Learning*, pp. 7002–7016. PMLR, 2022.
- 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.
- Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1414–1430, 2016.
- Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and Zico Kolter. Test-time adaptation via conjugate pseudo-labels. In *Advances in Neural Information Processing Systems*, 2022.
- Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.
- Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pp. 124–140, 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, Sae-Young Chung, and Hye Won Chung. Test-time adaptation via self-training with nearest neighbor information. In *International Conference on Learning Representations*, 2023.
- JoonHo Lee and Gyemin Lee. Feature alignment by uncertainty and self-training for source-free unsupervised domain adaptation. *Neural Networks*, 161:682–692, 2023.
- Chenming Li, Daoan Zhang, Wenjian Huang, and Jianguo Zhang. Cross contrasting feature perturbation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1327–1337, 2023.
- 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, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018a.
- 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, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409, 2018b.
- 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.
- Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry S Davis. Rethinking pseudo labels for semi-supervised object detection. In *AAAI Conference on Artificial Intelligence*, volume 36, 2022.
- Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018c.
- Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 624–639, 2018d.
- 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.
- Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *arXiv preprint arXiv:2303.15361*, 2023.
- Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7640–7650, 2023.
- Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22968–22981, 2021a.
- Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In *Advances in Neural Information Processing Systems*, volume 34, 2021b.
- Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pp. 97–105. PMLR, 2015.
- Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Adversarial style mining for one-shot unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.
- Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6670–6680, 2017a.

- 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, 2017b.
- Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.
- Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8690–8699, 2021.
- Shuaicheng Niu, Jiayang Wu, Yifan Zhang, Yaofu Chen, Shijian Zheng, Peilin Zhao, and Minghui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, pp. 16888–16905. PMLR, 2022.
- Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.
- Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *International Conference on Machine Learning*, pp. 7728–7738. PMLR, 2020.
- Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12556–12565, 2020.
- 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.
- Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.
- Marin Scalbert, Maria Vakalopoulou, and Florent Couzinié-Devy. Multi-source domain adaptation via supervised contrastive learning and confident consistency regularization. *arXiv preprint arXiv:2106.16093*, 2021.
- Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*, 2018.
- Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *International Conference on Learning Representations*, 2022.
- Inkyu Shin, Yi-Hsuan Tsai, Bingbing Zhuang, Samuel Schuster, Buyu Liu, Sparsh Garg, In So Kweon, and Kuk-Jin Yoon. Mm-tta: multi-modal test-time adaptation for 3d semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16928–16937, 2022.
- Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018.
- Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pp. 443–450, 2016.
- 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.

- Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.
- 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.
- Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.
- 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.
- Shuai Wang, Daoan Zhang, Zipei Yan, Jianguo Zhang, and Rui Li. Feature alignment and uniformity for test time adaptation. *arXiv preprint arXiv:2303.10902*, 2023.
- Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 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.
- Zehao Xiao, Xiantong Zhen, Shengcai Liao, and Cees G M Snoek. Energy-based test sample adaptation for domain generalization. In *International Conference on Learning Representations*, 2023.
- Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020.
- I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.
- Hongzheng Yang, Cheng Chen, Meirui Jiang, Quande Liu, Jianfeng Cao, Pheng Ann Heng, and Qi Dou. Dltta: Dynamic learning rate for test-time adaptation on cross-domain medical images. *IEEE Transactions on Medical Imaging*, 41(12):3575–3586, 2022.
- Li Yi, Gezheng Xu, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Charles Ling, A Ian McLeod, and Boyu Wang. When source-free domain adaptation meets learning with noisy labels. In *International Conference on Learning Representations*, 2023.
- 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, 2021.
- Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38629–38642, 2022.
- Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- Aurick Zhou and Sergey Levine. Bayesian adaptation for covariate shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 914–927, 2021.
- Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European Conference on Computer Vision*, pp. 561–578, 2020a.
- Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2020b.
- Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021.
- Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- 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 DERIVATIONS

### A.1 DETAILED DERIVATION OF META-GENERALIZED VARIATIONAL NEIGHBOR LABELS

We start the objective function from  $\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})$ . Here we provide the detailed generating process of the formulation:

$$\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}) = \log \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'})p(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})d\boldsymbol{\theta}_{t'}. \quad (13)$$

We then introduce the pseudo labels  $\hat{\mathbf{y}}_{t'}$  as the latent variable into eq. (13) and derive it as:

$$\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}) = \log \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'}) \int p(\boldsymbol{\theta}_{t'}|\hat{\mathbf{y}}_{t'}, \mathbf{x}_{t'}, \boldsymbol{\theta}_{s'})p(\hat{\mathbf{y}}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})d\hat{\mathbf{y}}_{t'}d\boldsymbol{\theta}_{t'}. \quad (14)$$

Theoretically, the distribution  $p(\boldsymbol{\theta}_{t'})$  is obtained by  $p(\boldsymbol{\theta}_{t'}|\mathbf{y}_{t'}, \mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}) \propto p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'})p(\boldsymbol{\theta}_{t'}|\boldsymbol{\theta}_{s'})$ , where  $p(\boldsymbol{\theta}_{t'}|\boldsymbol{\theta}_{s'})$  is the prior distribution. To simplify the formulation, we approximate the integration of  $p(\boldsymbol{\theta}_{t'})$  by the maximum a posteriori (MAP) value of  $\boldsymbol{\theta}_{t'}^*$ . We obtain the MAP value by training the model  $\boldsymbol{\theta}$  with inputs  $\mathbf{x}_{t'}$  and pseudo labels  $\mathbf{y}_{t'}$  starting from  $\boldsymbol{\theta}_{s'}$ . The formulation then is derived as:

$$\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}) = \log \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'}^*)p(\hat{\mathbf{y}}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})d\hat{\mathbf{y}}_{t'}. \quad (15)$$

To incorporate representative neighboring target information into the generation of  $\hat{\mathbf{y}}_{t'}$ , we further introduce the latent variable  $\mathbf{w}_{t'}$  into eq. (15) and a variational posterior of the joint distribution  $q(\hat{\mathbf{y}}_{t'}, \mathbf{w}_{t'})$ . The formulation is then derived as:

$$\begin{aligned} & \log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}) \\ &= \log \int \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'}^*)p(\hat{\mathbf{y}}_{t'}, \mathbf{w}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})d\hat{\mathbf{y}}_{t'}d\mathbf{w}_{t'} \\ &= \log \int \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'}^*) \frac{p(\hat{\mathbf{y}}_{t'}, \mathbf{w}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})}{q(\hat{\mathbf{y}}_{t'}, \mathbf{w}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'})} q(\hat{\mathbf{y}}_{t'}, \mathbf{w}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'})d\hat{\mathbf{y}}_{t'}d\mathbf{w}_{t'}, \end{aligned} \quad (16)$$

where  $p(\hat{\mathbf{y}}_{t'}, \mathbf{w}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}) = p(\hat{\mathbf{y}}_{t'}|\mathbf{w}_{t'}, \mathbf{x}_{t'})p_{\phi}(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})$  denote the prior distribution and  $q(\hat{\mathbf{y}}_{t'}, \mathbf{w}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'}) = p(\hat{\mathbf{y}}_{t'}|\mathbf{w}_{t'}, \mathbf{x}_{t'})q_{\phi}(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'})$  denotes the posterior one. By incorporating the prior and posterior distribution into eq. (16), the formulation is derived to:

$$\begin{aligned} & \log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}) \\ &= \log \int \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'}^*) \frac{p_{\phi}(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})}{q_{\phi}(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'})} p(\hat{\mathbf{y}}_{t'}|\mathbf{w}_{t'}, \mathbf{x}_{t'})q_{\phi}(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'})d\hat{\mathbf{y}}_{t'}d\mathbf{w}_{t'} \\ &\geq \int \int \log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'}^*)p(\hat{\mathbf{y}}_{t'}|\mathbf{w}_{t'}, \mathbf{x}_{t'})q_{\phi}(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'})d\hat{\mathbf{y}}_{t'}d\mathbf{w}_{t'} \\ &+ \int \log \frac{p_{\phi}(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})}{q_{\phi}(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'})} q_{\phi}(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'})d\mathbf{w}_{t'} \\ &= \mathbb{E}_{q_{\phi}(\mathbf{w}_{t'})} \mathbb{E}_{p(\hat{\mathbf{y}}_{t'}|\mathbf{w}_{t'}, \mathbf{x}_{t'})} [\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'}^*)] - \mathbb{D}_{KL}[q_{\phi}(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'})||p_{\phi}(\mathbf{w}_{t'}|\boldsymbol{\theta}_{s'}, \mathbf{X}_{t'})]. \end{aligned} \quad (17)$$

### A.2 FURTHER ANALYSES OF THE PROPOSED METHOD ON TEST-TIME GENERALIZATION

The goal of test-time generalization is to optimize for the expected classification performance during test time, i.e.,  $p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_t)$ . The model  $\boldsymbol{\theta}_t$  is obtained by  $p(\mathbf{x}_t)$  through  $p(\boldsymbol{\theta}_t|\mathbf{x}_t, \boldsymbol{\theta}_s)$  in common test-time adaptation or test-time generalization methods, which is achieved by entropy minimization or pseudo-labeling based on the prediction on  $\mathbf{x}_t$  of the source model  $\boldsymbol{\theta}_s$ . However, due to distribution shifts, the prediction of  $\mathbf{x}_t$  by  $\boldsymbol{\theta}_s$  can be overconfident and mispredicted.

By introducing probabilistic pseudo labels, we consider their uncertainty during adaptation, which mitigates overconfidence and introduces the discriminative information more reasonably. For example, consider a toy binary classification task in Figure 6, where the predicted probability is  $[0.4, 0.6]$  with the ground-truth label  $[1, 0]$ . The pseudo label generated by selecting the maximum probability is  $[0, 1]$ , which is inaccurate. Optimization based on these labels would give rise to a model misspecified to target data, failing to generalize to the target domain. In contrast, our 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  $\hat{\mathbf{y}}_t$  in a principled way. Continuing the example, the pseudo labels sampled from the predicted distribution have a probability of 40% to be the true label, which leads to the generalization of the model in

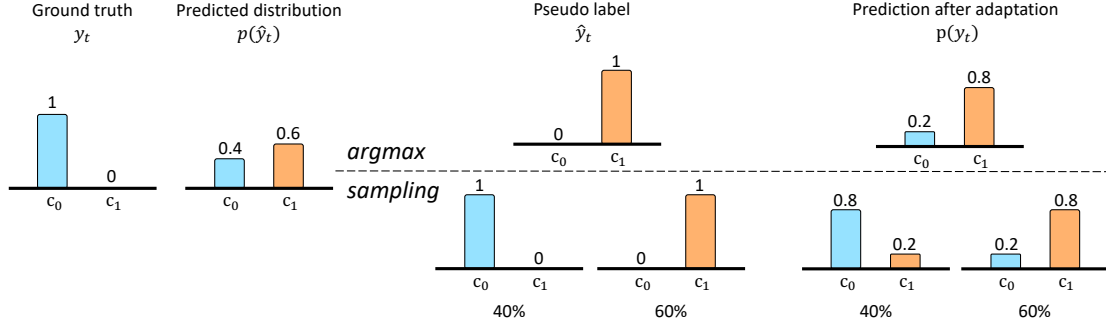


Figure 6: **Benefits of probabilistic pseudo labels.** Probabilistic pseudo labels consider the uncertainty to access correct supervision of the uncertain predictions, leading to more robust generalization than the common pseudo labels selected from the maximum probability.

the correct direction. Therefore, the formulation improves generalization by accessing accurate pseudo labels. The ablation studies in Table 4 (first two rows) and Figure 3 also demonstrate this.

We can also assume that the Oracle model for the target domain is obtained by  $p(\theta_t | \mathbf{x}_t, y_t, \theta_s)$  while we lack the categorical information in  $y_t$ . Theoretically, when the pseudo labels  $p(\hat{y}_t)$  carry richer and more accurate discriminatory information about  $p(\mathbf{x}_t, y_t)$ ,  $\theta_t$  adapts better to the target domain. Incorporating the neighboring information leverages category clusters from nearby target samples, enhancing target-specific cues in pseudo labels. Consequently, the model attained through variational neighbor labels yields more robust generalization of target data, which is also demonstrated in Table 4 (row 3). Moreover, with meta-learning, the model is trained to achieve good performance on target data with adaptation with variational pseudo labels, further improving the generalization ability. (Table 4 row 4).

## B ALGORITHMS

We provide the algorithms for source training and test-time generalization in Algorithm 1 and 2, respectively.

---

### Algorithm 1 Training for Meta-Generalized Variational Neighbor Labels

**Input:**  $\mathcal{S} = \{D_s\}_{s=1}^S$ : source domains with  $n$  sample pairs  $(\mathbf{x}_s, \mathbf{y}_s)$  for each;  $\theta$ : model parameters of backbone;  $\phi$ : model parameters of variational-neighbor-label generation;  $\lambda_{1,2,3}$ : learning rates;  $\mathcal{B}_{tr}$ : batch size during training;  $N_{iter}$ : the number of iterations.

**Output:** Learned  $\theta, \phi$

---

- 1: **for**  $iter$  in  $N_{iter}$  **do**
  - 2:  $\mathcal{T}' \leftarrow$  Randomly Sample  $(\{D_s\}_{s=1}^S, t')$ ;  
 $\mathcal{S}' \leftarrow \{D_s\}_{s=1}^S \setminus \mathcal{T}'$ ;
  - 3: Sample datapoints  $\{(\mathbf{x}_{s'}^{(k)}, \mathbf{y}_{s'}^{(k)})\}_{k=1}^{\mathcal{B}_{tr}} \sim \mathcal{S}'$ ,  $\{(\mathbf{x}_{t'}^{(k)}, \mathbf{y}_{t'}^{(k)})\}_{k=1}^{\mathcal{B}_{tr}} \sim \mathcal{T}'$ .
  - 4: **Meta-source stage:**
  - 5: Obtain meta-source model by training with the cross-entropy loss on meta-source labels and predictions  $\theta_{s'} = \min_{\theta} \mathbb{E}_{(\mathbf{x}_{s'}, \mathbf{y}_{s'}) \in \mathcal{D}_{s'}} [L_{CE}(\mathbf{x}_{s'}, \mathbf{y}_{s'}; \theta)]$ .
  - 6: **Meta-generalization stage:**
  - 7: Generate the prior  $p_{\phi}(\mathbf{w}_{t'} | \theta_{s'}, \mathbf{X}_{t'})$  and posterior distributions  $q_{\phi}(\mathbf{w}_{t'} | \theta_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'})$  of  $\mathbf{w}_{t'}$
  - 8: Sample the variational neighbor labels from the variational posterior distribution  
 $\hat{\mathbf{y}}_{t'} \sim p(\hat{\mathbf{y}}_{t'} | \mathbf{w}_{t'}, \mathbf{x}_{t'})$ ,  $\mathbf{w}_{t'} \sim q_{\phi}(\mathbf{w}_{t'} | \theta_{s'}, \mathbf{X}_{t'}, \mathbf{Y}_{t'})$ .
  - 9: Generalize the meta-source model to meta-target by cross-entropy loss with variational neighbor labels:  
 $\theta_{t'}^* = \theta_{s'} - \lambda_1 \nabla_{\theta} L_{CE}(\mathbf{x}_{t'}, \hat{\mathbf{y}}_{t'}; \theta_{s'})$ .
  - 10: **Meta-target stage:**
  - 11: Calculate meta-target loss on the generalized meta-target model  
 $\mathcal{L}_{meta} = \mathbb{E}_{(\mathbf{x}_{t'}, \mathbf{y}_{t'})} [\mathbb{E}_{q_{\phi}(\mathbf{w}_{t'})} \mathbb{E}_{p(\hat{\mathbf{y}}_{t'} | \mathbf{w}_{t'}, \mathbf{x}_{t'})} L_{CE}(\mathbf{x}_{t'}, \mathbf{y}_{t'}; \theta_{t'}^*)] + \mathbb{D}_{KL}[q_{\phi}(\mathbf{w}_{t'}) || p_{\phi}(\mathbf{w}_{t'})]$ .
  - 12: Calculate the cross-entropy loss on the variational neighbor labels and actual labels  $\mathcal{L}_{\hat{c}_e} = L_{CE}(\hat{\mathbf{y}}_{t'}, \mathbf{y}_{t'})$ .
  - 13: Update the parameters  $\theta$  and  $\phi$  by  $\theta = \theta_{s'} - \lambda_2 \nabla_{\theta} \mathcal{L}_{meta}$  and  $\phi = \phi - \lambda_3 (\nabla_{\phi} \mathcal{L}_{\hat{c}_e} - \nabla_{\phi} \mathcal{L}_{meta})$ , respectively. //Note the meta-target loss optimizes the meta-source model  $\theta_{s'}$ .
  - 14: **end for**
-

**Algorithm 2** Test-time Generalization by Meta-Generalized Variational Neighbor Labels

**Input:**  $\mathcal{T}$ : target domain with  $N_t$  samples  $\mathbf{x}_t$ ;  $\theta_s, \phi_s$ : source trained model parameters;  $\lambda_1$ : learning rate for test-time generalization;  $\mathcal{B}_{te}$ : batch size for each online step at test time.

---

```

1: Initialize  $\theta_t = \theta_s$ .
2: for  $iter$  in  $(N_t/\mathcal{B}_{te})$  do
3:   Sample one batch of target samples from the target domain  $\{\mathbf{x}_t^{(k)}\}_{k=1}^{\mathcal{B}_{te}} \sim \mathcal{T}$ .
4:   Sample the variational neighbor labels for the target batch from the prior distribution
    $\hat{\mathbf{y}}_t \sim p(\hat{\mathbf{y}}_t|\mathbf{w}_t, \mathbf{x}_t), \mathbf{w}_t \sim p_\phi(\mathbf{w}_t|\theta_t, \mathbf{X}_t)$ .
5:   Generalize the model parameters by cross-entropy loss with variational neighbor labels:
    $\theta_t^* = \theta_t - \lambda_1 \nabla_{\theta} L_{CE}(\mathbf{x}_t, \hat{\mathbf{y}}_t; \theta_t)$ .
6:   Make predictions of the target current target batch by
    $p(\mathbf{y}_t|\mathbf{x}_t, \theta_t, \mathbf{X}_t) = \mathbb{E}_{p_{\phi_s}(\mathbf{w}_t)} \mathbb{E}_{p(\hat{\mathbf{y}}_t|\mathbf{w}_t, \mathbf{x}_t)} [\log p(\mathbf{y}_t|\mathbf{x}_t, \theta_t^*)]$ .
7:   Update  $\theta_t = \theta_t^*$ 
8: end for

```

---

## C IMPLEMENTATION DETAILS

Our training setup follows [Iwasawa & Matsuo \(2021\)](#), including the dataset splits and hyperparameter selection methods. The backbones such as ResNet-18 and ResNet-50 are pretrained on ImageNet same as the previous methods. As discussed in the main paper, the ERM baseline means we directly evaluate the source-trained model without any adjustment at test time ([Gulrajani & Lopez-Paz, 2020](#)). We also have other two baselines in the main paper. ‘‘Hard pseudo label’’ is obtained using the *argmax* of model predictions and ‘‘soft pseudo label’’ refers to the original model predictions. Optimization with soft pseudo-labeling is similar to entropy minimization.

During training, we randomly select one source domain as the meta-target domain and the others as the meta-source domains in each iteration. The model is trained following the meta-source, meta-generalization, and meta-target stages as in Section 3 and Algorithm 1. We use a batch size of 70 for the model in the meta-training stage on source domains. To generate the variational neighbor labels, we implement the network ‘ $\phi$ ’ by a 3-layer MLP with 512 neurons per layer, which introduces approximately 1% more parameters than the backbone model in total. We use similar settings and hyperparameters for all domain generalization benchmarks that are reported in the main paper, i.e.,  $1e-4$ ,  $5e-5$ , and  $1e-4$  as  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$ , respectively for Adam optimizer. The source-trained model with the highest train validation accuracy is selected as the initial model for test-time generalization on the target domain as in [Iwasawa & Matsuo \(2021\)](#).

At test time, we sample the variational neighbor labels directly from the prior distribution (eq. 5) and utilize the labels for updating the source-trained model. The source model is generalized to target data in an online manner with 20 target samples per batch, which is a small number. We update all parameters of the model with the learning rate of  $1e-4$  in an online manner using Adam optimizer. We choose the hyperparameters for model adjustment based on the training-domain validation set as mentioned in [Gulrajani & Lopez-Paz \(2020\)](#) and [Iwasawa & Matsuo \(2021\)](#). There are no additional hyperparameters involved in our method. Our method is orthogonal to other deployment techniques, e.g., methods like data augmentation for test-time generalization [Zhang et al. \(2022\)](#). We utilize data augmentation methods at test-time for our results. We train and evaluate all our models on one NVIDIA Tesla 1080Ti GPU and utilize the PyTorch framework. We run all the experiments using 5 different random seeds and report the results. We will release the code.

## D COMPUTATIONAL COMPLEXITY AND RUNTIME COMPARISON

We provide the overall runtime comparison of our method in both the training (Table 6) and test-time stage (Table 7). As we utilize the meta-learning-based strategy to learn the ability to handle domain shifts during training and to generate variational neighbor labels, the runtime during training is larger than the ERM baseline. Moreover, compared with the ERM baseline, our variational neighbor-labeling and meta-learning framework only introduces a few more parameters (around 1%).

Since the meta-learning strategy is only deployed in the training time, our method has similar overall runtime for generalization at test time compared with the other test-time adaptation, e.g., Tent ([Wang et al., 2021](#)), and test-time domain generalization methods, e.g., TAST ([Jang et al., 2023](#)). Another factor for the runtime is the number of updated parameters at test time. The models only update the parameters of BN layers, e.g., Tent-BN ([Wang et al., 2021](#)), or classifiers, e.g., T3A ([Iwasawa & Matsuo, 2021](#)) have lower overall runtime than ours that update all model parameters. Compared with the methods that update all parameters with pseudo labels ([Liang et al., 2020](#); [Jang et al., 2023](#)), our overall computational runtime is competitive and even lower.

Table 6: **Runtime required for source training on PACS using ResNet-18 as a backbone network.** The proposed method has overall larger runtime during training due to the meta-learning strategy but introduces few extra parameters.

	Parameters	Time for 10000 iterations
ERM baseline	11.18 M	6.5 hours
<i>This paper</i>	11.96 M	14.6 hours

Table 7: **Runtime averaged for datasets using ResNet-18 as a backbone network.** The proposed method has similar or even better overall runtime at test time with the other test-time adaptation and test-time domain generalization methods.

	VLCS	PACS	Terra	OfficeHome
Wang et al. (2021)	7m 28s	3m 16s	10m 34s	7m 25s
Wang et al. (2021)	2m 8s	33s	2m 58s	1m 57s
Liang et al. (2020)	8m 09s	4m 22s	12m 40s	8m 38s
Jang et al. (2023)	10m 34s	9m 30s	26m 14s	22m 24s
Iwasawa & Matsuo (2021)	2m 09s	33s	2m 59s	2m 15s
<i>This paper</i>	2m 20s	5m 33s	14m 30s	7m 07s

## E ADDITIONAL RESULTS AND DISCUSSIONS

**Performance on single source image corruption datasets.** Apart from existing domain generalization datasets, we also conducted experiments on the CIFAR-10-C dataset. We train on the original data and evaluate the model on 15 types of corruption similar to Tent (Wang et al., 2021). We achieved 21.60% as the error rate. In comparison, the source model without adaptation achieves 29.14 %, and Tent achieves 14.30 %. The performance is not good enough since we cannot mimic distribution shifts by meta-generalization in this single-source setting. We consider single-source domain generalization at test time a worthwhile avenue for future work, as highlighted in the conclusion.

**Large number of categories and model capacity.** We have experimented with different numbers of layers in the MLP for Office-Home by utilizing ResNet-18 as the backbone. We obtain a mean accuracy of 57.1 with 2 layers and 64.3 with 3 layers in  $\phi$ .

**Comparisons with additional existing domain generalization algorithms.** We provide additional domain generalization comparisons from Gulrajani & Lopez-Paz (2020) in the following Table 8 based on ResNet-50. Our method achieves better results compared with these methods.

**Performance on DomainNet.** Apart from common domain generalization datasets, we also conducted experiments on DomainNet (Peng et al., 2019) dataset. We followed the common setup as in other datasets with ResNet-18 and achieved 41.8% mean accuracy in comparison to ERM, which achieves 40.9% mean accuracy. DomainNet dataset consists of 345 classes and 586,575 samples. Further performance on datasets with larger categories can be improved by utilizing a larger  $\phi$  model capacity.

**Impact of the extent of distribution shifts between source and target domains.** Since we changed the meta-target domain iteratively, the simulated domain shifts during training consider all domain shifts among source domains. To provide more insights into the effects of different extents of distribution shifts, we conduct experiments with different target domains on rotated MNIST. Specifically, we utilized  $0^\circ$ ,  $15^\circ$ ,  $30^\circ$ ,  $45^\circ$ , as source domains with 3 separate target domains  $60^\circ$ ,  $75^\circ$ ,  $90^\circ$  respectively. The difference in rotation degree between source and target simulates the amount of difference in domain shifts between source and target domains.

As shown in Table 9, we compared our method with ERM baseline and Tent (Wang et al., 2021). Indeed larger distribution shifts result in worse performance (e.g., all methods perform worse on the  $90^\circ$  domain). However, compared with ERM and Tent, our method consistently performs better on varied levels of rotation, demonstrating the robustness of our method on different extents of distribution shifts.

**Comparisons to additional Test-time adaptation methods with pre-training phase.** Compared with the fully test-time adaptation methods (e.g., Tent (Wang et al., 2021)), our method modifies the learning strategy during source training. To further show the effectiveness of our method, we also compare our method with some test-time training methods, TTT (Sun et al., 2020) and TTT+ (Liu et al., 2021b), who also modify the learning strategy during training. We re-implemented their methods on the PACS dataset with ResNet-18 for test-time domain generalization. TTT (Sun

Table 8: **Comparisons to additional existing domain generalization algorithms** for ResNet-50 backbones on different datasets. Our method performs better (bold) than the other methods.

	PACS	VLCS	Office-Home	Terra Incognita
ERM baseline	85.7	77.4	67.5	47.2
Arjovsky et al. (2019)	83.5	78.5	64.3	47.6
Sagawa et al. (2019)	84.4	76.7	66.0	43.2
Li et al. (2018a)	84.9	77.2	66.8	47.7
Sun & Saenko (2016)	86.2	78.8	68.7	47.6
Li et al. (2018b)	84.6	77.5	66.3	42.2
Ganin et al. (2016)	83.6	78.6	65.9	46.7
Li et al. (2018d)	82.6	77.5	65.8	45.8
Eastwood et al. (2022)	86.5	77.8	67.5	47.8
Li et al. (2023)	86.6	78.9	68.9	48.6
<b>This paper</b>	<b>87.9</b> $\pm 0.3$	<b>79.1</b> $\pm 0.4$	<b>69.1</b> $\pm 0.4$	<b>49.4</b> $\pm 0.6$

Table 9: **Impact of the extent of distribution shifts between source and target domains.** The source domains are kept the same, while the individual test domains vary. Our method also performs better with increased domain shift (difference in degrees) between source and target)

Method	60°	75°	90°
ERM baseline	96.8	91.6	78.9
Wang et al. (2021)	97.2	91.8	79.0
<b>This paper</b>	<b>97.7</b> $\pm 0.3$	<b>92.2</b> $\pm 0.3$	<b>79.4</b> $\pm 0.3$

et al., 2020) obtains 83.0% accuracy with  $\pm 0.2\%$  standard deviation and Liu et al. (2021b) obtains 83.8 accuracy with  $\pm 0.5\%$  standard deviation. Our method obtains 85.0 % accuracy with  $\pm 0.4\%$  standard deviation and outperforms these methods.

**Orthogonality.** Since the proposed meta-learned variational neighbor labels focus on generating pseudo labels at test time, the method is orthogonal to other deployment techniques, e.g., data augmentation for generalization at test time (Zhang et al., 2022). Achieving test-time domain generalization compounded with these methods will further improve the performance. To demonstrate this, we conduct test-time generalization by our method with augmented target samples on PACS without altering the source training strategy. When adding similar augmentation as in (Zhang et al., 2022), we increase our results on ResNet-18 from 83.5% to 85.0% overall accuracy. We provide the complete table including the per-domain results in Appendix F.

**Clarification about sacrificing training data.** We do not sacrifice training data rather we utilize all available source data for both model training and simulating test-time generalization. Specifically, for every iteration, we randomly select one training domain as the meta-test domain and others as meta-source domains. The model is first trained on meta-source domains and then generalized on the meta-test one. As a result, the meta-source and meta-target domains are changed for each iteration. Hence, the model is trained on all available source domains, without sacrificing training data.

## F DETAILED EXPERIMENTAL RESULTS

**Detailed results of different batch sizes at test time.** Test-time generalization and adaptation methods usually require updating the source-trained model with large online batches of target samples, which is not always available in real-world applications. To show the robustness of our method on limited data, we conduct experiments with different batch sizes during test-time generalization and compare the proposed method with Soft pseudo label in Figure 4 in the main paper. The experiments are conducted on PACS with ResNet-18. Here we provide detailed results of each domain in Table 10. The conclusion is similar to that in the main paper. Soft pseudo label performs well with large batch sizes, e.g., 128, but fails with small ones, e.g., 16. The performance with batch sizes of 16 is even lower than the baseline model. By contrast, our method performs consistently better than Soft pseudo label. Moreover, by incorporating the

Table 10: **Detailed comparisons of Soft pseudo label on PACS with different batch sizes.** Our method consistently performs better than Soft pseudo label with different batch sizes during test-time generalization. Moreover, the proposed method is more robust with small batch sizes.

	Photo	Art	Cartoon	Sketch	Mean
Baseline model	92.40 $\pm$ 0.2	78.70 $\pm$ 1.3	74.30 $\pm$ 0.6	75.60 $\pm$ 0.8	80.30 $\pm$ 0.4
<b>Generalization with 16 samples per step</b>					
Soft pseudo label	93.65 $\pm$ 0.3	80.20 $\pm$ 0.2	76.90 $\pm$ 0.5	68.49 $\pm$ 0.7	79.81 $\pm$ 0.3
<i>This paper</i>	95.45 $\pm$ 0.2	83.70 $\pm$ 0.4	80.42 $\pm$ 0.5	74.91 $\pm$ 0.7	83.62 $\pm$ 0.5
<b>Generalization with 64 samples per step</b>					
Soft pseudo label	96.04 $\pm$ 0.3	81.91 $\pm$ 0.4	80.81 $\pm$ 0.6	76.33 $\pm$ 0.7	83.77 $\pm$ 0.4
<i>This paper</i>	96.09 $\pm$ 0.2	84.37 $\pm$ 0.4	81.25 $\pm$ 0.5	78.12 $\pm$ 0.7	85.00 $\pm$ 0.5
<b>Generalization with 128 samples per step</b>					
Soft pseudo label	97.25 $\pm$ 0.2	84.91 $\pm$ 0.3	81.12 $\pm$ 0.5	76.80 $\pm$ 0.8	85.02 $\pm$ 0.5
<i>This paper</i>	96.75 $\pm$ 0.2	85.15 $\pm$ 0.4	85.70 $\pm$ 0.4	78.90 $\pm$ 0.6	86.62 $\pm$ 0.4

Table 11: **Variational neighbor labels combined with augmentation at test time on PACS.** Our method achieves better results in conjunction with augmentations.

Data augmentation	Photo	Art	Cartoon	Sketch	Mean
$\times$	95.50 $\pm$ 0.4	82.90 $\pm$ 0.3	81.28 $\pm$ 0.	74.11 $\pm$ 0.6	83.45 $\pm$ 0.4
$\checkmark$	<b>96.40</b> $\pm$ 0.2	<b>83.81</b> $\pm$ 0.4	<b>82.62</b> $\pm$ 0.4	<b>77.20</b> $\pm$ 0.6	<b>85.00</b> $\pm$ 0.4

Table 12: **Importance of the model adjustment at test time.** We conduct ablation studies on PACS using ResNet-18. Our method that updates the source-trained model with variational neighbor labels performs better than the prediction directly sampled from the variational neighbor distributions.

Settings	Photo	Art-painting	Cartoon	Sketch	Mean
ERM baseline	92.40 $\pm$ 0.2	78.70 $\pm$ 1.3	74.30 $\pm$ 0.6	75.60 $\pm$ 0.8	80.30 $\pm$ 0.4
Prediction by pseudo label distributions	95.97 $\pm$ 1.0	81.23 $\pm$ 0.5	79.19 $\pm$ 0.7	73.22 $\pm$ 0.5	82.40 $\pm$ 0.8
<i>This paper</i>	<b>96.40</b> $\pm$ 0.2	<b>83.81</b> $\pm$ 0.4	<b>82.62</b> $\pm$ 0.5	<b>77.20</b> $\pm$ 0.7	<b>85.00</b> $\pm$ 0.4

uncertainty and representative neighboring information, our method is more robust to small batch sizes, leading to a larger improvement than Soft pseudo label.

**Detailed results of the method with augmentation at test time.** As discussed in the main paper, our variational neighbor labeling is orthogonal to other deployment techniques for test-time domain generalization, e.g., data augmentation at test time. We provide detailed results in Table 11 where we compare the results of the proposed method on PACS with and without augmenting target samples during generalization at test time. Our method achieves better results on all domains in conjunction with augmentations.