LEARNING TO GENERALIZE ACROSS DOMAINS ON SINGLE TEST SAMPLES

Anonymous authors
Paper under double-blind review

ABSTRACT

We strive to learn a model from a set of source domains that generalizes well to unseen target domains. The main challenge in such a domain generalization scenario is the unavailability of any target domain data during training, resulting in the learned model not being explicitly adapted to the unseen target domains. We propose learning to generalize across domains on single test samples. We leverage a meta-learning paradigm to learn our model to acquire the ability of adaptation with single samples at training time so as to further adapt itself to each single test sample at test time. We formulate the adaptation to the single test sample as a variational Bayesian inference problem, which incorporates the test sample as a conditional into the generation of model parameters. The adaptation to each test sample requires only one feed-forward computation at test time without any fine-tuning or self-supervised training on additional data from the unseen domains. We conduct extensive ablation studies to demonstrate the effectiveness and advantages. Further, our model achieves at least comparable – and often better – performance than state-of-the-art methods on multiple benchmarks for domain generalization [1].

1 INTRODUCTION

Despite their widespread adoption and success in academia and industry alike, deep convolutional neural networks suffer from a fundamental flaw: they have insufficient generalizability to test data that is out of their training distributions (Recht et al., 2019). Improving the generalization of machine learning methods therefore remains a challenging problem (Moreno-Torres et al., 2012; Recht et al., 2019; Krueger et al., 2021). To deal with the distribution shift, domain adaptation (Saenko et al., 2010; Long et al., 2015; Lu et al., 2020; Li et al., 2021) and domain generalization (Muandet et al., 2013; Li et al., 2017; 2020) have been extensively investigated. In domain adaptation, the key assumption is that the target data, either labeled or unlabeled, is accessible during *training*, which allows the model to be adapted to the target domains. However, this assumption does not necessarily hold in realistic scenarios where the target domain data is not available. By contrast, domain generalization strives to learn a model on source domains that can generalize well to unseen target domains without any access to the target domain data during training. We aim for domain generalization.

Previous domain generalization approaches have successfully explored domain-invariant learning (Muandet et al., 2013; Motiian et al., 2017b; Dou et al., 2019; Chattopadhyay et al., 2020) or domain augmentation (Shankar et al., 2018; Qiao et al., 2020; Zhou et al., 2020a) to handle the unknown domain shift between the source domains and target domains. However, since those models are trained on source domains, there will always be an "adaptivity gap" when applying them to target domains without further adaptation (Dubey et al., 2021). Thus, it is necessary to train models that are able to further adapt to target domains without ever using target data during training.

Recently, further adapting a pre-trained model using target domain data by fine-tuning or self-supervised training at test time has shown effectiveness in improving performance on domain generalization (Sun et al., 2020; Wang et al., 2021). To achieve proper adaptation to the target domain, these methods typically rely on batches of samples from the target domain with extra fine-tuning operations or extra networks (Dubey et al., 2021) at test time. This prevents these methods from being able to effectively adapt to the target domain when provided with very few target samples or

[1] Our code is available in https://anonymous.4open.science/r/ICLR2022submissionSingleSampleGeneralization/.

Table 1: **Training and test-time settings of learning methods for generalization across domains.** (X_s, Y_s) and X_t denote the labeled source data and unlabeled target data. X indicates batches of samples, while $\mathbf{x}$ denotes just one sample. Our method does not need any target data during training, like (one-shot) domain adaptation methods, or for fine-tuning operations during inference. It just uses single test samples to do per-sample adaptation for domain generalization.

		Training	**Test-time**		
Task	**Method**	Data	Data	Fine-tune	Adaptive
Domain Adaptation	Common, e.g., (Long et al., 2015)	X_s, Y_s, X_t	$\mathbf{x}_t$	×	✓
	One-shot, e.g., (Luo et al., 2020)	$X_s, Y_s, \mathbf{x}_t$	$\mathbf{x}_t$	×	✓
Domain Generalization	Common, e.g., (Seo et al., 2020)	X_s, Y_s	$\mathbf{x}_t$	×	×
	Test-time adaptation, e.g., (Wang et al., 2021)	X_s, Y_s	$X_t, \mathbf{x}_t$	✓	✓
	Domain-adaptive method (Dubey et al., 2021)	X_s, Y_s	$X_t, \mathbf{x}_t$	×	✓
	Single test sample generalization (ours)	X_s, Y_s	$\mathbf{x}_t$	×	✓

when exposed to test samples from multiple target domains without domain identifiers. Different from these works, we aim to learn the ability of adaptation to a single sample such that the model can further adapt to each test sample from any target domain.

In this paper, we propose *learning to generalize on single test samples* for domain generalization. Using meta-learning (Li et al., 2017), we train our model to acquire the ability to adapt with each single sample on source domains such that it can then adapt to each individual sample in the target domain at test time. The rationale is that we can utilize the information contained in one sample to further adapt the trained model to this sample. Unlike previous methods, our model does not need to be fine-tuned using target data or train extra networks when generalized to the unseen target domains. Once it is trained on source domains, our model is able to adapt to each single target sample with just a feed-forward computation at test time, without the need for any extra target data. Note that not a single target sample is seen during training, which does not forgo the strict source and target separation setting of domain generalization. Moreover, in contrast to previous test-time adaptation methods (Sun et al., 2020; Wang et al., 2021; Dubey et al., 2021), our method does not need to fine-tune the model or generate a domain embedding for each target domain. Our model effortlessly handles test samples from multiple domains without knowing their domain identifiers.

To be specific, we build our model under the meta-learning framework and formulate the single test sample generalization as a variational inference problem. In the training stage, we divide the source domains into several meta-source domains and a meta-target domain and explore the adaptive model by incorporating the information of the meta-target sample into the generation of the model parameters. For any given meta-target sample, we propose a variational distribution generated by this sample and the meta-source data to approximate the model distribution obtained by the meta-target data. In the test phase, the adapted models for the samples from the real target domains are generated on the fly by the variational model distribution. The random splits of the meta-source and meta-target domains expose the model to domain shifts and mimic the real generalization process from source domains to the target domain. Thus, the model is endowed with the ability to adapt to any unseen sample by end-to-end training with the data from source domains only. By doing so, our method does not need to introduce any extra target data or fine-tuning operations on the target domain. A comparison of the training/test-time settings among different domain generalization (and domain adaptation) methods is summarized in Table 1.

We demonstrate the effectiveness of our single test sample generalization by conducting experiments on commonly-used benchmarks for domain generalization. Our ablation studies under different domain generalization settings demonstrate the advantages of our method in comparison to other recent adaptive methods.

2 Related Work

Domain adaptation To handle the domain shift between source and target domains, considerable efforts have been devoted to developing methods for domain adaptation (Long et al., 2015; Lu et al., 2020). These methods narrow the domain gap between source and target domains using supervised, semi-supervised or unsupervised target samples that are available during training (Hoffman et al., 2018; Kumar et al., 2010; Tzeng et al., 2017; Luo et al., 2019). However, the assumption that target data is accessible during training is often invalid. Recently, some even more challenging settings were proposed in domain adaptation, e.g., few-shot (Motiian et al., 2017a) and one-shot domain adaptation

(Luo et al., 2020), but there are still a few or one target sample available during training. The more strict domain generalization setting was proposed to address domain shift without access to any (labeled or unlabeled) target data during training. In this work, we focus on domain generalization.

Domain generalization The problem of domain generalization was introduced by Blanchard et al. (2011), and formally posed by Muandet et al. (2013). Muandet et al. (2013) and Ghifary et al. (2016) both considered domain-invariant learning by minimizing the dissimilarity of the moments of features across source domains. Following these works, many methods explored domain invariant learning for domain generalization (Motiian et al., 2017b; Rojas-Carulla et al., 2018; Xiao et al., 2021). Li et al. (2018b) further improved the model by learning conditional-invariant features. Seo et al. (2020) combined batch normalization and instance normalization to embed the samples from different domains into a domain-invariant feature space. Zhao et al. (2020) proposed an entropy regularization to learn domain-invariant features. Alternatively, augmenting the source domain data has been explored to simulate more domain shift during training (Shankar et al., 2018; Volpi et al., 2018; Zhou et al., 2020a). Qiao et al. (2020) introduced data augmentation with an adversarial method for single-domain generalization. Zhou et al. (2020b) mixed the feature statistics between instances from different domains and achieved feature-based data augmentation. Our method tackles domain shift by learning to adapt to each unseen target sample, which is orthogonal to domain-invariant and domain augmentation methods.

Domain meta-learning Meta-learning methods have also been studied for domain generalization (Xiao et al., 2021; Du et al., 2021). Li et al. (2018a) introduced the model agnostic meta-learning (Finn et al., 2017) into domain generalization. Dou et al. (2019) further enhanced the model agnostic meta-learning with global and local regularization. Balaji et al. (2018) meta-learned a regularization to make the model generalize well across domains. Du et al. (2020) proposed the meta-variational information bottleneck for domain-invariant learning. We also leverage meta-learning in our method. However, different from previous methods, we utilize the meta-learning framework to train our model to learn the capability of adaptation to each unseen target sample.

Test-time adaptation Recently, several methods have proposed to adapt a model trained on source data to unseen target data at test time (Sun et al., 2020; Wang et al., 2021; Dubey et al., 2021). Sun et al. (2020) trained their model on source domains with both supervised and self-supervised objective functions, and then fine-tuned it on target data with only the self-supervised one. Wang et al. (2021) further improved this method by removing the self-supervised module, while fine-tuning the model with an entropy minimization loss. Differently, Dubey et al. (2021) trained a domain embedding network with unsupervised source data, and then trained their domain-adaptive model with both the source features and the domain embedding. At test time, they utilized hundreds of unlabeled target samples to generate the target domain embedding to adapt the model, without fine-tuning or further training. These adaptive methods all need extra steps, e.g., fine-tuning the model or training embedding networks, and *batches* of target samples for the adaptation. In contrast, our method learns to adapt to each target sample using just its own information.

3 METHODOLOGY

In domain generalization, several domains are defined as different data distributions in the joint space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ denote the input space and the label space. The domains are divided into non-overlapping source domains, collectively denoted as $\mathcal{S}$ and the target domains $\mathcal{T}$. During training, only data from source domains is accessible. Data from target domains is never seen. Our goal is to learn a model on the source domains only that generalizes well to the target domains. As we focus on homogeneous domain generalization (Zhou et al., 2021) in this work, all domains share the same label space.

Since the target data is inaccessible during training, we design a meta-learning framework with episodic training (Li et al., 2019; Balaji et al., 2018) that endows the model with the capability of learning to generalize on single test samples from any unseen domain. To this end, we randomly divide the source domains $\mathcal{S}$ into several meta-source domains $\mathcal{S}'$ and a meta-target domain $\mathcal{T}'$ in each iteration to expose the model to domain shifts. After training, the learned model is applied to the target domain $\mathcal{T}$ for evaluation.

Variational learning We develop our model under the probabilistic framework. Assume we have a single sample $\mathbf{x}_{t'}$ from the meta-target domain $\mathcal{T}'$, and we would like to predict its class label $\mathbf{y}_{t'}$.

We consider the conditional predictive log-likelihood $\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \mathcal{T}')$. By incorporating the model $\boldsymbol{\theta}_{t'}$ into the prediction distribution, we have

$$\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \mathcal{T}') = \log \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'}) p(\boldsymbol{\theta}_{t'}|\mathcal{T}') d\boldsymbol{\theta}_{t'}, \tag{1}$$

where we condition the prediction on $\mathcal{T}'$ to leverage information from the meta-target data and $(\mathbf{x}_{t'}, \mathbf{y}_{t'})$ denotes the input-output pair of the meta-target sample. We refer to $p(\boldsymbol{\theta}_{t'}|\mathcal{T}')$ as the meta-prior distribution since it is conditioned on the meta-target data.

However, in practice, the labeled data from the target domain is inaccessible, leading to an intractable distribution $p(\boldsymbol{\theta}_t|\mathcal{T})$ during inference. To address the problem, we propose to approximate the model function $p(\boldsymbol{\theta}_t|\mathcal{T})$ with the accessible source data by the variational inference approach, which provides a framework to approximate intractable distributions via optimization. By training under the meta-learning framework, our method will learn the ability to approximate the model functions of unseen target domains by source domain data.

Variational domain generalization We introduce a variational distribution $q(\boldsymbol{\theta}_{t'}|\mathcal{S}')$ to approximate the true posterior over the model parameters $\boldsymbol{\theta}$. By incorporating $q(\boldsymbol{\theta}_{t'}|\mathcal{S}')$ into eq. (1), we derive a lower bound of the conditional predictive log-likelihood:

$$\begin{aligned} \log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \mathcal{T}') &= \log \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'}) p(\boldsymbol{\theta}_{t'}|\mathcal{T}') d\boldsymbol{\theta}_{t'} \\ &\geq \mathbb{E}_{q(\boldsymbol{\theta}_{t'})}[\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'})] - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{\theta}_{t'}|\mathcal{S}')||p(\boldsymbol{\theta}_{t'}|\mathcal{T}')]. \end{aligned} \tag{2}$$

The introduced Kullback-Leibler (KL) divergence term is minimized to reduce the distance between $q(\boldsymbol{\theta}_{t'}|\mathcal{S}')$ and $p(\boldsymbol{\theta}_{t'}|\mathcal{T}')$, which encourages the model $\boldsymbol{\theta}$ inferred by $q(\boldsymbol{\theta}_{t'}|\mathcal{S}')$ from meta-source domains to be generalizable to the meta-target domain.

Although eq. (2) reduces the difference between the variational distributions $q(\boldsymbol{\theta}_{t'}|\mathcal{S}')$ and $p(\boldsymbol{\theta}_{t'}|\mathcal{T}')$, $q(\boldsymbol{\theta}_{t'}|\mathcal{S}')$ still has limited capacity of handling the meta-target samples due to "adaptivity gap" (Dubey et al., 2021). Since only little information about the meta-target domain $\mathcal{T}'$ and the meta-target sample $\mathbf{x}_{t'}$ is available in $\mathcal{S}'$, there is no guarantee that the variational distribution $q(\boldsymbol{\theta}_{t'}|\mathcal{S}')$ will be adapted to the meta-target domain $\mathcal{T}'$. Thus, we further propose to learn to adapt the model $\boldsymbol{\theta}_{t'}$ by taking more information of the meta-target data into account.

Single test sample generalization In domain generalization, especially for real applications, when given data from an unseen target domain, the only accessible information of the target data is from the target sample, without any annotations. We therefore propose to utilize the information contained in the given target sample to construct the adapted model. Instead of utilizing $q(\boldsymbol{\theta}_{t'}|\mathcal{S}')$ as the variational distribution, we incorporate the meta-target sample $\mathbf{x}_{t'}$ into the variational distribution $q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'}, \mathcal{S}')$ to approximate the true posterior distribution. Given an unlabeled sample $\mathbf{x}_{t'}$ from the meta-target domain $\mathcal{T}'$, we have a new evidence lower bound (ELBO) as follows:

$$\begin{aligned} \log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \mathcal{T}') &= \log \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'}) p(\boldsymbol{\theta}_{t'}|\mathcal{T}') d\boldsymbol{\theta}_{t'} \\ &\geq \mathbb{E}_{q(\boldsymbol{\theta}_{t'})}[\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \boldsymbol{\theta}_{t'})] - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'}, \mathcal{S}')||p(\boldsymbol{\theta}_{t'}|\mathcal{T}')]. \end{aligned} \tag{3}$$

By conditioning on the given sample $\mathbf{x}_{t'}$ and the source data $\mathcal{S}'$, the adapted variational distribution $q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'}, \mathcal{S}')$ contains both the categorical information provided by the meta-source data and the domain information given by the meta-target sample. This to a large extent guarantees the inferred model to be discriminative and adapted to the given meta-target domain sample. Moreover, the KL divergence in eq. (3) further acts as a regularizer to guide the variational distribution to be similar to the parameter distribution $p(\boldsymbol{\theta}_{t'}|\mathcal{T}')$ obtained through the meta-target data, which is assumed to be the most suitable parameter for the meta-target sample $\mathbf{x}_{t'}$. This makes the model better adapted to the sample from the meta-target domain.

Based on the meta-learning framework, eq. (3) guides the model to learn the capability of adapting to each target sample from any unseen domain using the information from the given sample. During inference, our method directly uses the variational distributions $q(\boldsymbol{\theta}_t|\mathbf{x}_t, \mathcal{S})$ as the adapted model generated by the information of the given target sample $\mathbf{x}_t$ and the data from all source domains $\mathcal{S}$. Thus, the method is able to generalize to each unseen target sample without the requirement of any

extra data from the target domains or fine-tuning operations. The derivations of eq. (2) and eq. (3), as well as more discussions on their tightness are provided in Appendix A.

Learning to adapt classifiers The variational inference framework in eq. (3) indicates that we can learn to adapt all model parameters $\boldsymbol{\theta}$ to the target sample with its own information. This could, however, be computationally infeasible due to the large number of parameters in the network. Thus, for computational efficiency and feasibility we utilize the features extracted by the backbone network as the data information to construct an adapted classifier, which contains much fewer parameters.

To do so, we divide the model $\boldsymbol{\theta}$ into a feature extractor $\boldsymbol{\phi}$ and a classifier $\mathbf{w}$. The feature extractor is shared across domains, while the classifier is trained to be adapted to each single sample by eq. (3). Both the meta-prior distribution $p(\mathbf{w}_{t'}|\mathcal{T}')$ and the variational posterior distribution $q(\mathbf{w}_{t'}|\mathbf{x}_{t'}, \mathcal{S}')$ of the classifier $\mathbf{w}_{t'}$ are generated by amortized inference using the amortization technique (Kingma & Welling, 2013; Gershman & Goodman, 2014; Gordon et al., 2019). Specifically, we use the center features of samples in each class from the meta-target domain and meta-source domains as the representatives of $\mathcal{T}'$ and $\mathcal{S}'$. We generate the meta-prior distribution $p(\mathbf{w}_{t'}|\mathcal{T}')$ by taking the center features of $\mathcal{T}'$ as input to the amortized inference network $\boldsymbol{\theta}_a$. For the variational distribution $q(\mathbf{w}_{t'}|\mathbf{x}_{t'}, \mathcal{S}')$, the input to the amortized inference network is the concatenation of the center features of the meta-source domain $\mathcal{S}'$ with the features of the target sample $\mathbf{x}_{t'}$ for adaptation. The amortized inference network $\boldsymbol{\theta}_a$ is shared for the inference of both the meta-prior and posterior distributions.

However, the direct concatenation operation is not necessarily optimal as the input for the amortized inference network. We propose to generate the parameter distribution using the meta-source domains and then adapt it with the given meta-target sample. To do so, we first generate a classifier distribution $p(\mathbf{w}_{s'}|\mathcal{S}')$ using the center features of the meta-source data $\mathcal{S}'$. After that, we take $p(\mathbf{w}_{s'}|\mathcal{S}')$ as a prior and combine it with the meta-target sample to estimate the variational posterior, as follows:

$$q(\mathbf{w}_{t'}|\mathbf{x}_{t'}, \mathcal{S}') = \int p(\mathbf{w}_{t'}|\mathbf{x}_{t'}, \mathbf{w}_{s'}) p(\mathbf{w}_{s'}|\mathcal{S}') d\mathbf{w}_{s'}. \tag{4}$$

This establishes a hierarchical variational inference of the posterior distribution. We show in our experiments that the hierarchical inference yields better results than the direct concatenation. In practice, we approximate the integral by:

$$q(\mathbf{w}_{t'}|\mathbf{x}_{t'}, \mathcal{S}') = \sum_{\ell=1}^{L} p(\mathbf{w}_{t'}|\mathbf{x}_{t'}, \mathbf{w}_{s'}^{(\ell)}), \tag{5}$$

where $\mathbf{w}_{s'}^{(\ell)} \sim p(\mathbf{w}_{s'}|\mathcal{S}')$ and L is the number of Monte Carlo samples.

The negative KL divergence term in eq. (3) tries to close the gap between the variational adaptive distribution $q(\mathbf{w}_{t'}|\mathbf{x}_{t'}, \mathcal{S}')$ and the meta-prior distribution $p(\mathbf{w}_{t'}|\mathcal{T}')$, aiming to make the variational distribution adapt better to the target data. However, there is no guarantee that the classifier produced by the meta-prior is discriminative enough for classification. To address this issue, we introduce an intermediate supervision based on the meta-prior distribution to make it as predictive as possible, by maximizing:

$$\mathbb{E}_{p(\mathbf{w}_{t'}|\mathcal{T}')} \big[\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \mathbf{w}_{t'}) \big]. \tag{6}$$

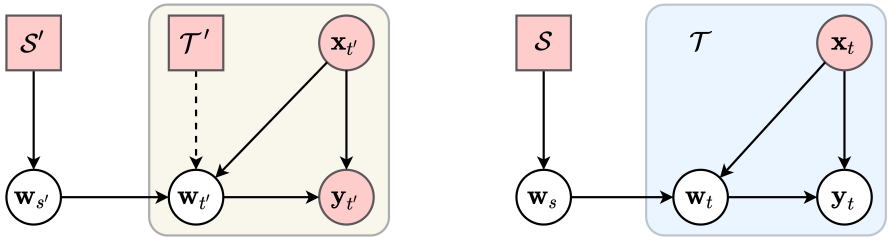
This is also implemented as a cross-entropy loss using Monte Carlo sampling from the meta-prior.

Empirical objective function By incorporating eq. (5) and eq. (6) into eq. (3), we obtain the final empirical objective function:

$$\mathcal{L} = \sum_{m=1}^{M} \log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \mathbf{w}_{t'}^{(m)}) + \sum_{n=1}^{N} \log p(\mathbf{y}_{t'}|\mathbf{x}_{t'}, \mathbf{w}_{t'}^{(n)}) - \mathbb{D}_{\mathrm{KL}}\Big(\sum_{\ell=1}^{L} p(\mathbf{w}_{t'}|\mathbf{x}_{t'}, \mathbf{w}_{s'}^{(\ell)}) || p(\mathbf{w}_{t'}|\mathcal{T}')\Big) \tag{7}$$

where we use Monte Carlo sampling: $\mathbf{w}_{t'}^{(m)} \sim \sum_{\ell=1}^{L} q(\mathbf{w}_{t'}|\mathbf{x}_{t'}, \mathbf{w}_{s'}^{(\ell)})$ and $\mathbf{w}_{t'}^{(n)} \sim p(\mathbf{w}_{t'}|\mathcal{T}')$.

As illustrated in Figure 1, at training time (a), we train the model episodically by randomly dividing the source domains $\mathcal{S}$ into meta-source domains $\mathcal{S}'$ and a meta-target domain $\mathcal{T}'$. At inference time (b), given any sample $\mathbf{x}_t$ from the target domain $\mathcal{T}$, we use this sample to adapt the $\mathbf{w}_s$ generated by the source domains. The adapted model on the single target sample is generated by combining $\mathbf{w}_s$ and the feature representation of the given target sample.



(a) Training on meta-source ($\mathcal{S}'$) and meta-target ($\mathcal{T}'$) domains (b) Testing on the unseen domain ($\mathcal{T}$)

Figure 1: **Illustration of single test sample generalization using meta-learning.** Through variational Bayesian inference, we incorporate the single sample as a conditional during the generation of model parameters. The dashed line indicates training only.

4 Experiments

4.1 Datasets and Implementation Details

We conduct our experiments on four widely used benchmarks for domain generalization and report the mean classification accuracy on target domains.

Datasets **PACS** (Li et al., 2017) is a widely used dataset consisting of 9,991 images of seven classes from four domains, i.e., *photo*, *art-painting*, *cartoon* and *sketch*. We use the same training and validation split as in (Li et al., 2017) and follow the "leave-one-out" protocol from (Li et al., 2017; 2019; Carlucci et al., 2019). **Office-Home** (Venkateswara et al., 2017) has 15,500 images of 65 categories. The images are also from four domains, i.e., *art*, *clipart*, *product* and *real-world*. We use the same experimental protocol as for PACS. **Rotated MNIST and Fashion-MNIST** are utilized in Piratla et al. (2020). We follow their recommended settings and randomly select a subset of 2,000 images from MNIST and 10,000 images from Fashion-MNIST. The subset of images is rotated by 0° through 90° in intervals of 15°, creating seven domains. We use the subsets with rotation angles from 15° to 75° as the source domains, and images rotated by 0° and 90° as the target domains. For comparison, we evaluate our method on both in-distribution and out-of-distribution data.

Implementation details We evaluate our method on PACS with both ResNet-18 and ResNet-50 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) as the backbone. During training, we use Adam optimization and train the model for 10,000 iterations. The learning rate of the backbone is set to 0.00005 for ResNet-18 and 0.00001 for ResNet-50, while the learning rate of the network for generating the classifier is set to 0.0001 consistently. The batch size is 128. To generate the classifier, in each iteration we select 10 samples from each category in each meta-source domain for generating the center features $\mathcal{S}'$, and the same number of samples from the meta-target domain for $\mathcal{T}'$. The model with the highest validation accuracy is utilized for evaluation on the target domain. Most of the experimental settings and hyperparameters on Office-Home are the same as on PACS. Since there are more categories in Office-Home, the number of samples per category per domain is set to 5 to fit the memory footprint. The learning rate of the backbone is set to 0.00001 for both ResNet-18 and ResNet-50. For fair comparisons, we evaluate the rotated MNIST and Fashion-MNIST with ResNet-18, following (Piratla et al., 2020). The other experimental settings are also the same as PACS. For the amortized inference, to make full use of the information of the given target sample, the features combined with $\mathcal{S}'$ are an aggregation of features from multiple levels of the feature extractor. More implementation details are provided in Appendix B.

4.2 Results

Benefit of single test sample generalization We first investigate the effectiveness of our single test sample generalization on the PACS dataset. To demonstrate the benefits of all components of our method, we consider four settings: (i) the "variational amortized classifier" with eq. (2) as the objective function, (ii) our single test sample generalization method without supervision on the meta-prior distribution (eq. 6), (iii) single test sample generalization without the hierarchical architecture (eq. 5), and (iv) the entire single test sample generalization method, with the objective function in eq. (7). Results based on the ResNet-18 backbone are shown in Table 2.

Table 2: **Benefits of single test sample generalization.** Ablation study on PACS using ResNet-18 averaged over five runs. Best runs within a 95% confidence margin bolded. Single test sample generalization is more effective than the invariant amortized classifier. The hierarchical architecture further improves the performance. Intermediate supervision on the prior distribution is also important.

Settings	Photo	Art-painting	Cartoon	Sketch	*Mean*
Invariant amortized classifier (eq. 2)	95.19 ±0.30	79.34 ±0.65	78.04 ±0.85	76.21 ±0.54	82.19 ±0.35
Single test sample generalization (w/o eq. 6)	**95.47** ±0.20	80.00 ±0.61	77.87 ±0.91	77.20 ±0.88	82.64 ±0.32
Single test sample generalization (w/o eq. 5)	**95.69** ±0.20	81.06 ±0.33	**79.20** ±0.49	77.10 ±0.64	83.26 ±0.31
Single test sample generalization (eq. 7)	**95.87** ±0.24	**82.02** ±0.36	**79.73** ±0.49	**78.96** ±0.67	**84.15** ± 0.21

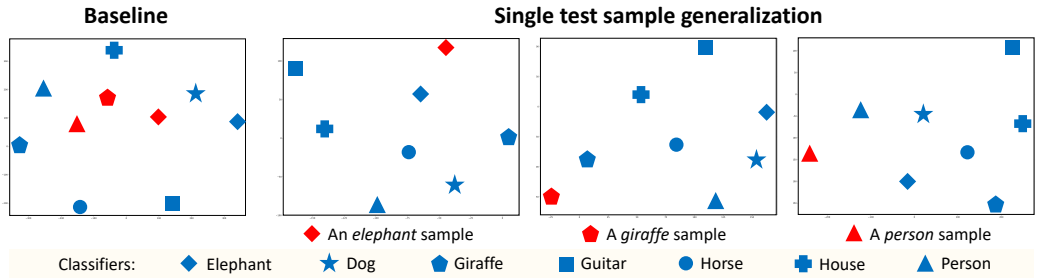


Figure 2: **Visualization of single test sample generalization** on the "cartoon" domain from PACS. We plot the test samples and classifiers in the same 2D plane. Different shapes denote different categories. Samples are in red, classifiers in blue. The baseline method utilizes the same classifier for all samples from each domain, while our method generates an adapted classifier for each target sample. The two methods use the same samples. In our single test domain generalization, the test sample is the closest to the adapted classifier of its own class, indicating better classification ability.

By comparing rows 1 and 3 of Table 2, it is clear that our single test sample generalization brings a considerable improvement in accuracy over the invariant amortized classifier, especially on the "art-painting" and "cartoon" domains. These results demonstrate the benefits of introducing the information of the given target sample into the model generation. The comparison also demonstrates that our model learns the capability to adapt to the target sample using its own information. Comparing rows 2 and 4 shows the benefits of the intermediate supervision in eq. (6), which provides a more discriminative meta-prior distribution during training to learn more adapted variational distribution, especially on the "art-painting", "cartoon" and "sketch" domains. Results in the last two rows of Table 2 show that the hierarchical architecture for the variational adaptive classifier $q(\mathbf{w}_{t'})$ provides a more adapted model for each target sample and achieves the best performance on all domains.

Visualization of single test sample generalization To illustrate the benefit of incorporating the test sample into classifier generation, we visualize the classifiers of the "variational amortized classifier" (baseline), our classifiers adapted by the given target sample, and the features of target samples in Figure 2. We use the same samples in the visualizations of the two methods. As shown in the left figure (baseline), the classifier parameters of the "invariant amortized classifier" are fixed for each target sample during inference. Since the samples are out of the training distributions, they are easy to misfit the fixed classifier trained on source domains, leading to failed predictions. In contrast, as shown in the three figures on the right, our method generates a specific, adapted classifier for each test target sample. For instance, the test sample (◆) has the shortest distance to the adapted classifier of its own class (◆), which produces the correct predictions. More visualizations, including failure cases, are provided in Appendix C.

Single test sample generalization vs. Tent[2] Next we conduct an experiment on rotated-MNIST to compare our method with Tent by Wang et al. (2021), which is a recent test-time adaptation method that fine-tunes the model with batches of target samples. We follow the setting in (Piratla et al., 2020) and generalize seven domains by rotating the samples by $0°$ to $90°$ in intervals of $15°$. To make the target domains more different from each other, we use images rotated by $0°$, $45°$ and $90°$ as target domains, $15°$, $30°$, $60°$ and $75°$ as source domains. The base model is trained on the source data without any extra operations. Then, Tent is utilized to adapt the model to different target domains. As shown in Figure 3, we design two settings for Tent for the comparisons: "single target domain" (left) and "multiple target domains" (right). In the "single target domain" setting, we adapt the base

[2] All Tent results are obtained by running the author-provided code accompanying (Wang et al., 2021).

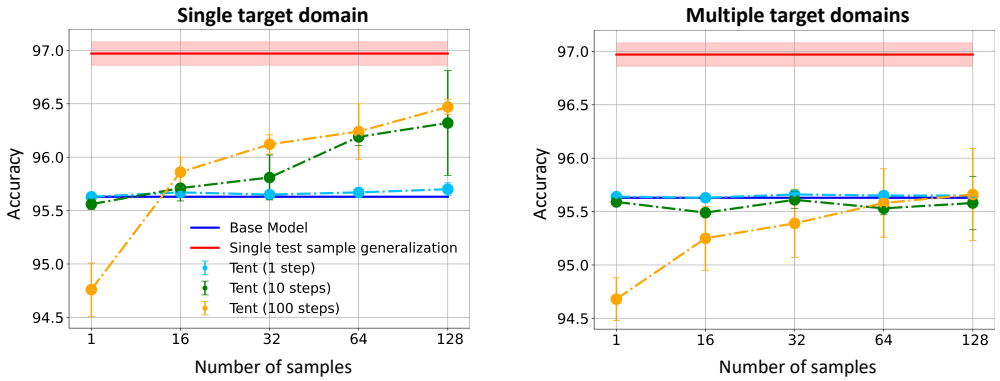


Figure 3: **Single test sample generalization vs. Tent** (Wang et al., 2021) for different settings on rotated-MNIST. Tent shows good performance with a large batch of samples from a single domain (left). However, when provided only one sample or given samples from different domains without the domain id (right), Tent suffers. More detailed comparisons are provided in Appendix D.

Table 3: **Comparison on PACS.** Our method achieves best mean accuracy with a ResNet-50 backbone and is competitive with ResNet-18. Notably, it surpasses the adaptive test-time methods by Wang et al. (2021) and Dubey et al. (2021), despite them using more data at test-time (see Table 1).

Backbone	Method	Photo	Art-painting	Cartoon	Sketch	*Mean*
ResNet-18	Carlucci et al. (2019)	96.03	79.42	75.25	71.35	80.51
	Dou et al. (2019)	94.99	80.29	77.17	71.69	81.04
	Zhao et al. (2020)	**96.65**	80.70	76.40	71.77	81.46
	Zhou et al. (2020a)	96.20	83.30	78.20	73.60	82.83
	Wang et al. (2021)	95.49 ±0.27	81.55 ±0.35	77.67 ±0.39	77.64 ±0.27	83.09 ±0.13
	Zhou et al. (2020b)	96.10	84.10	78.80	75.90	83.70
	Seo et al. (2020)	95.87	**84.67**	77.65	**82.23**	**85.11**
	This paper	95.87 ±0.24	82.02 ±0.36	**79.73** ±0.49	78.96 ±0.67	84.15 ± 0.21
ResNet-50	Dou et al. (2019)	95.01	82.89	80.49	72.29	82.67
	Dubey et al. (2021)	-	-	-	-	84.50
	Zhao et al. (2020)	**98.25**	87.51	79.31	76.30	85.34
	Gulrajani & Lopez-Paz (2020)	97.20	84.70	80.80	79.30	85.50
	Wang et al. (2021)	97.96 ±0.27	86.30 ±0.26	82.53 ±0.69	78.11 ±0.73	86.23 ±0.22
	Seo et al. (2020)	95.99	87.04	80.62	**82.90**	86.64
	This paper	97.88 ±0.15	**88.09** ±0.26	**83.83** ±0.68	80.21 ±0.66	**87.51** ±0.22

model to each of the target domains separately using Tent and take the average accuracy of different domains. For "multiple target domains", Tent is required to adapt to all three target domains jointly without the domain identifiers. Since our method is able to adapt to each sample, the performance is consistent under both settings.

As expected, Tent achieves good adaptation for large batch sizes (e.g., 128) from a single target domain. However, when there are only very few samples, and ultimately just one sample, the adaptation of Tent starts to suffer, and may even hurt accuracy. Tent, like other test-time adaptive domain generalization methods, relies on a batch of samples to adapt to the target domain. In other words, the test samples have known domain identifiers. However, in practical settings, the test samples might come from different domains and the domain identifiers may be inaccessible. Thus, we also explore the effectiveness under the "multiple target domains" setting, as shown in Figure 3 (right). In this scenario, the adaptability of test-time adaptation methods like Tent, starts to decline, even with more samples. In contrast, our method learns the capability to adapt to each sample from any target domain, overcoming the lack of target samples while being robust under multiple target domains. We also compare our method with Tent on PACS and Office-Home. The results are shown in Table 3 and 4. We perform Tent with a baseline model trained on all source domain data, which is then fine-tuned by 128 target samples for 100 iterations. The conclusion is the same as on rotated MNIST. Our method achieves better performance.

Comparisons with State of the Art Finally, we compare our method with the state-of-the-art on four datasets. The results on each of the datasets are reported in Tables 3, 4 and 5.

PACS In Table 3, we conduct experiments with both ResNet-18 and ResNet-50 backbones. Our method achieves good performance using the ResNet-18 backbone. For each individual domain,

Table 4: **Comparison on Office-Home.** Our method achieves the best mean accuracy using both a ResNet-18 and ResNet-50 backbone. Notably, it surpasses the adaptive test-time methods by Wang et al. (2021) and Dubey et al. (2021), despite them using more data at test-time (see Table 1).

Backbone	Method	Art	Clipart	Product	Real World	*Mean*
ResNet-18	Carlucci et al. (2019)	53.04	47.51	71.47	72.79	61.20
	Seo et al. (2020)	59.37	45.70	71.84	74.68	62.90
	Huang et al. (2020)	58.42	47.90	71.63	74.54	63.12
	Wang et al. (2021)	56.45 ±0.30	52.06 ±0.18	73.19 ±0.42	74.82 ±0.27	64.13 ±0.16
	Zhou et al. (2020a)	**60.60**	50.10	**74.80**	**77.00**	65.63
	This paper	59.39 ±0.43	**53.94** ±0.45	**74.68** ±0.33	76.07 ±0.17	**66.02** ±0.28
ResNet-50	Gulrajani & Lopez-Paz (2020)	61.30	52.40	75.80	76.60	66.50
	Wang et al. (2021)	62.12 ±0.32	56.65 ±0.49	75.61 ±0.57	77.58 ±0.42	67.99 ±0.22
	Dubey et al. (2021)	-	-	-	-	68.90
	Sun & Saenko (2016)	65.30	54.40	76.50	78.40	68.70
	This paper	**67.21** ±0.59	**57.97** ±0.37	**78.61** ±0.59	**80.47** ±0.16	**71.07** ±0.31

Table 5: **Comparison on rotated MNIST and Fashion-MNIST.** In-distribution performance is evaluated on the test sets of MNIST and Fashion-MNIST with rotation angles of $15°$, $30°$, $45°$, $60°$ and $75°$, while the out-of-distribution performance is evaluated on test sets with angles of $0°$ and $90°$. We achieve the best performance on both the in-distribution and out-of-distribution test sets.

	MNIST		**Fashion-MNIST**	
	In-distribution	Out-of-distribution	In-distribution	Out-of-distribution
Dou et al. (2019)	98.2	93.2	86.9	72.4
Piratla et al. (2020)	98.4	94.7	89.7	78.0
This paper	**98.9** ±0.1	**95.8** ±0.1	**90.9** ±0.2	**80.8** ±0.5

we are competitive with the state of the art and slightly better on the "cartoon" domain when using ResNet-18. On the "art-painting" and "cartoon" domains with ResNet-50, our method also achieves good performance. Although it delivers competitive, or sometimes even better, performance on most domains, our method is not as good on the "sketch" domain. One reason might be that the images and features from this domain carry less information than other domains, leading to less adaptive models.

Office-Home As shown in Table 4, our method again achieves good overall performance with both the ResNet-18 and ResNet-50 backbones. With ResNet-18 as the backbone, we outperform other methods by a good margin on the "clipart" domain, while delivering competitive performance on the other domains. When utilizing ResNet-50 as the backbone, our method achieves good performance on all four domains. Overall, compared with other methods, our method shows larger advantages with ResNet-50 than ResNet-18. This is reasonable since we generate the adaptive model using the features of the given target sample. The features extracted by ResNet-50 contain more information, leading to a more adapted model for the given target sample.

Rotated MNIST and Fashion-MNIST. The proposed single test sample generalization shows good generalizability to out-of-distribution data. Our method achieves the best performance on both datasets as shown in Table 5, especially for the out-of-distribution setting. Moreover, our method demonstrates less performance drop from in-distribution to out-of-distribution data in comparison to other methods. The results on all the datasets demonstrate the effectiveness of our method in domain generalization.

5 Conclusion and Future Work

We propose learning to generalize by single test samples for domain generalization. Our model is able to adapt to any unseen target sample by leveraging the information of the given sample to generate the model. We formulate the adaptation process as a variational inference problem and train the model under the meta-learning framework. After training on the source data, our method can adapt to each sample from the unseen target domains without any fine-tuning operations or extra target data. Ablation studies and further experiments on several benchmarks show the effectiveness of our method. Like other meta-learning-based domain generalization methods, our method utilizes the domain shift between source domains to mimic the shift between source and target domains, which requires at least two source domains during meta-training. This limits the generalization ability when only one source domain is available. An appealing solution is to augment the source domain at training time, which we leave as our future work.

ETHICS STATEMENT

With the capability to handle the domain shift without the accessibility of target data during training and fine-tuning operations, our method has potential positive impacts in the applications that often face unseen domains in practice, e.g., medical imaging and automatic driving. Accordingly, our method would also potentially face some negative social impacts accompanying with applications, e.g., lack of fairness with the model trained by incomplete data and the privacy of patients in medical imaging.

REPRODUCIBILITY STATEMENT

To reproduce our method, the details of the method and loss functions are provided in Section 3. We also provide the derivations and discussions about the evidence lower bound used in our method in Appendix A. The datasets used in this paper and the basic implementation details can be obtained by referencing Section 4.1. More detailed implementations of the method are shown in Appendix B together with an illustration of the architecture in Figure 4. We also provide a link for the code of our method in Abstract.

REFERENCES

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.

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.

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.

Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *European Conference on Computer Vision*, pp. 301–318. Springer, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

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 GM 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14340–14349, 2021.

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.

Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

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.

Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E Turner. Meta-learning probabilistic inference for prediction. In *International Conference on Learning Representations*, 2019.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pp. 1989–1998. PMLR, 2018.

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.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation. In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Abhishek Kumar, Avishek Saha, and Hal Daume. Co-regularization based semi-supervised domain adaptation. In *Advances in Neural Information Processing Systems*, volume 23, pp. 478–486, 2010.

Bo Li, Yezhen Wang, Shanghang Zhang, Dongsheng Li, Kurt Keutzer, Trevor Darrell, and Han Zhao. Learning invariant representations and risks for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1104–1113, 2021.

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/CVF International Conference on Computer Vision*, pp. 1446–1455, 2019.

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

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, 2018b.

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.

Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9111–9120, 2020.

Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2507–2516, 2019.

Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Adversarial style mining for one-shot unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 33, 2020.

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.

Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. *Advances in Neural Information Processing Systems*, 30:6670–6680, 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.

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/CVF 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.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision*, pp. 213–226, 2010.

Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *European Conference on Computer Vision*, pp. 68–83, 2020.

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.

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.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.

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.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, pp. 5339–5349, 2018.

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.

Zehao Xiao, Jiayi Shen, Xiantong Zhen, Ling Shao, and Cees G M Snoek. A bit more bayesian: Domain-invariant learning with uncertainty. In *International Conference on Machine Learning*. PMLR, 2021.

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.

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. Springer, 2020a.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2020b.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

A DERIVATIONS

Derivation of variational domain generalization In domain generalization, since the target domain data is inaccessible, the distribution $p(\boldsymbol{\theta}_t|\mathcal{T})$ is intractable. We leverage the meta-learning setting to mimic the domain generalization process. In particular, we episodically divide the source domains into meta-source domains $\mathcal{S}'$ and a meta-target domain $\mathcal{T}'$ and learn the ability to generalize to the target data. Thus, during training, we generate the parameter distribution by the meta-source data $\mathcal{S}'$ to approximate the distribution $p(\boldsymbol{\theta}_{t'}|\mathcal{T}')$ generated by the meta-target data. Then at inference, the model is able to approximate the distribution $p(\boldsymbol{\theta}_t|\mathcal{T})$ with the parameters generated by the source data $\mathcal{S}$. To do so, we introduce a variational distribution $q(\boldsymbol{\theta}_{t'}|\mathcal{S}')$ to approximate the true posterior over the model parameter $\boldsymbol{\theta}$. The full derivation of eq. (2) in the main paper is as follows:

$$\begin{aligned}
\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'},\mathcal{T}') &= \log \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'},\boldsymbol{\theta}_{t'})p(\boldsymbol{\theta}_{t'}|\mathcal{T}')d\boldsymbol{\theta}_{t'} \\
&= \log \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'},\boldsymbol{\theta}_{t'})\frac{p(\boldsymbol{\theta}_{t'}|\mathcal{T}')}{q(\boldsymbol{\theta}_{t'}|\mathcal{S}')}q(\boldsymbol{\theta}_{t'}|\mathcal{S}')d\boldsymbol{\theta}_{t'} \\
&= \log \Big[\mathbb{E}_{q(\boldsymbol{\theta}_{t'}|\mathcal{S}')}\big[p(\mathbf{y}_{t'}|\mathbf{x}_{t'},\boldsymbol{\theta}_{t'})\frac{p(\boldsymbol{\theta}_{t'}|\mathcal{T}')}{q(\boldsymbol{\theta}_{t'}|\mathcal{S}')}\big]\Big] \\
&\geq \mathbb{E}_{q(\boldsymbol{\theta}_{t'}|\mathcal{S}')}\Big[\log\big[p(\mathbf{y}_{t'}|\mathbf{x}_{t'},\boldsymbol{\theta}_{t'})\frac{p(\boldsymbol{\theta}_{t'}|\mathcal{T}')}{q(\boldsymbol{\theta}_{t'}|\mathcal{S}')}\big]\Big] \\
&= \mathbb{E}_{q(\boldsymbol{\theta}_{t'}|\mathcal{S}')}[\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'},\boldsymbol{\theta}_{t'})] - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{\theta}_{t'}|\mathcal{S}')||p(\boldsymbol{\theta}_{t'}|\mathcal{T}')].
\end{aligned} \tag{8}$$

Derivation of single test sample generalization Since there is little information available about the target domain or target data in the variational distribution $q(\boldsymbol{\theta}_{t'}|\mathcal{S}')$ in eq. (2), the resultant model is not adaptive to the target domain data. To generate a model adapted to the target sample, we introduce information of the target data, and incorporate the meta-target sample $\mathbf{x}_{t'}$ into the variational distribution $q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'},\mathcal{S}')$. The detailed derivation is:

$$\begin{aligned}
\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'},\mathcal{T}') &= \log \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'},\boldsymbol{\theta}_{t'})p(\boldsymbol{\theta}_{t'}|\mathcal{T}')d\boldsymbol{\theta}_{t'} \\
&= \log \int p(\mathbf{y}_{t'}|\mathbf{x}_{t'},\boldsymbol{\theta}_{t'})\frac{p(\boldsymbol{\theta}_{t'}|\mathcal{T}')}{q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'},\mathcal{S}')}q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'},\mathcal{S}')d\boldsymbol{\theta}_{t'} \\
&= \log \Big[\mathbb{E}_{q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'},\mathcal{S}')}\big[p(\mathbf{y}_{t'}|\mathbf{x}_{t'},\boldsymbol{\theta}_{t'})\frac{p(\boldsymbol{\theta}_{t'}|\mathcal{T}')}{q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'},\mathcal{S}')}\big]\Big] \\
&\geq \mathbb{E}_{q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'},\mathcal{S}')}\Big[\log\big[p(\mathbf{y}_{t'}|\mathbf{x}_{t'},\boldsymbol{\theta}_{t'})\frac{p(\boldsymbol{\theta}_{t'}|\mathcal{T}')}{q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'},\mathcal{S}')}\big]\Big] \\
&= \mathbb{E}_{q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'},\mathcal{S}'))}[\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'},\boldsymbol{\theta}_{t'})] - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'},\mathcal{S}')||p(\boldsymbol{\theta}_{t'}|\mathcal{T}')].
\end{aligned} \tag{9}$$

Comparison of the two evidence lower bounds The main difference between eq. (8) and eq. (9) is the variational posterior $q(\boldsymbol{\theta}_{t'})$. Compared with eq. (8), our method (eq. 9) introduces the target sample $\mathbf{x}_{t'}$ into the generation of the variational posterior, which therefore achieves more adaptive classifiers for the target sample. Moreover, the tightness of the bound is also related to the variational posterior. The approximation gap between our bound (eq. 3) and the objective log-likelihood corresponds to the KL divergence between the variational posterior $q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'},\mathcal{S}')$ and the true posterior $p(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'},\mathbf{y}_{t'},\mathcal{T}')$, as shown in the following equation:

$$\begin{aligned}
logp(\mathbf{y}_{t'}|\mathbf{x}_{t'},\mathcal{T}') = &\underbrace{\mathbb{E}_{q(\boldsymbol{\theta}_{t'})}[\log p(\mathbf{y}_{t'}|\mathbf{x}_{t'},\boldsymbol{\theta}_{t'})] - \mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{\theta}_{t'}|x_{t'},\mathcal{S}')||p(\boldsymbol{\theta}_{t'}|\mathcal{T}')]}_{ELBO} \\
&+ \underbrace{\mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'},\mathcal{S}')||p(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'},\mathbf{y}_{t'},\mathcal{T}')]}_{Approximation\ Gap}.
\end{aligned} \tag{10}$$

Compared with eq. (2), whose approximation gap is formulated as $\mathbb{D}_{\mathrm{KL}}[q(\boldsymbol{\theta}_{t'}|\mathcal{S}')||p(\boldsymbol{\theta}_{t'}|\mathbf{x}_{t'},\mathbf{y}_{t'},\mathcal{T}')]$, our method takes the given sample $\mathbf{x}_{t'}$ as one of the conditions, leading to a tighter lower bound.

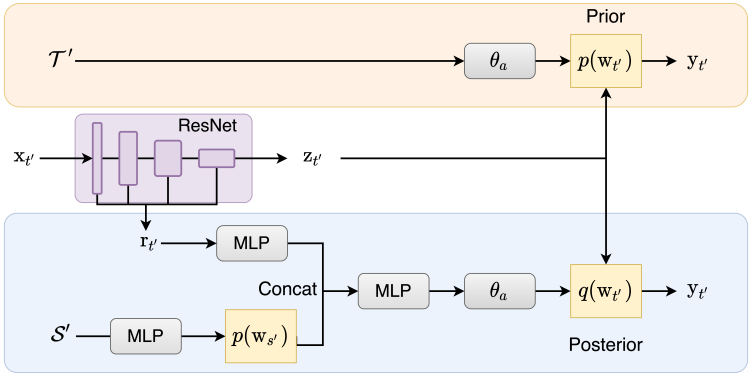


Figure 4: **Illustration of our model with the hierarchical architecture.** $\mathcal{S}'$ and $\mathcal{T}'$ denote the meta-source data and meta-target data. $\mathbf{x}_{t'}$, $\mathbf{z}_{t'}$, and $\mathbf{y}_{t'}$ denote the input image, feature representation and prediction of the given target sample. $\mathbf{r}_{t'}$ denotes the multi-level features of the input sample from which we generate the adapted posterior distribution of the classifier.

Table 6: **Implementation details of our method per dataset and backbone.** "Number of source samples" denotes the number of samples per class per source domain for generating the adapted classifier.

Dataset	Backbone	Classifier learning rate	Backbone learning rate	Number of source samples
PACS	ResNet-18	0.0001	0.00005	10
	ResNet-50	0.0001	0.00001	10
Office-Home	ResNet-18	0.0001	0.00001	5
	ResNet-50	0.0001	0.00001	5
Rotated MNIST	ResNet-18	0.0001	0.00005	10
Fashion-MNIST	ResNet-18	0.0001	0.00005	10

B Implementation Details

We train all models by the Adam optimizer on an NVIDIA Tesla V100 GPU. The models with different backbones for different datasets are trained for 10,000 iterations with a 128 batch size. The detailed information about other hyperparameters are summarized in Table 6. We also provide an illustration of our architecture in Figure 4. During training, we obtain the meta-prior distribution $p(\mathbf{w}_{t'}|\mathcal{T}')$ of the classifier by center features of the meta-target data $\mathcal{T}'$. To generate the adapted classifier, we first get the $p(\mathbf{w}_{s'})$ using the meta-source center features $\mathcal{S}'$, then we concatenate the $p(\mathbf{w}_{s'})$ and the multi-level features $\mathbf{r}_{t'}$ processed by a MLP to generate the posterior distribution $q(\mathbf{w}_{t'}|\mathbf{x}_{t'}, \mathcal{S}')$ as eq. (4) in the main paper. $\boldsymbol{\theta}_a$ in Figure 4 is an amortized network (Kingma & Welling, 2013; Gershman & Goodman, 2014), which is conducted also by a MLP module and shared by the generation network of both the meta-prior and posterior distributions. The amortized network takes meta-target domain center features $\mathcal{T}'$ and concatenated features of $\mathcal{S}'$ and $\mathbf{r}_{t'}$ as inputs, and returns the parameters (μ, σ) of the meta-prior distributions and the posterior distributions, respectively. There are three loss functions, cross-entropy loss functions on predictions of both $q(\mathbf{w}_{t'}|\mathbf{x}_{t'}, \mathcal{S}')$ and $p(\mathbf{w}_{t'}|\mathcal{T}')$, and the KL divergence between them, which correspond to the three items in eq. (7) in the main paper. During inference, the posterior distribution is utilized as the adapted classifier for each target sample $\mathbf{x}_t$, by replacing $\mathcal{S}'$ with $\mathcal{S}$.

C Visualizations

More comparisons with baseline To further demonstrate the effectiveness of our method, we provide more visualizations of the classifier parameters and feature representations from different domains in Figure. 5. The left column shows the visualizations of the "variational amortized classifier" (baseline), and the three columns on the right show our results. The same conclusion can be drawn here as in the main paper. Our method is able to generate adapted classifiers for each test sample, providing better predictions than the fixed classifiers in the baseline method.

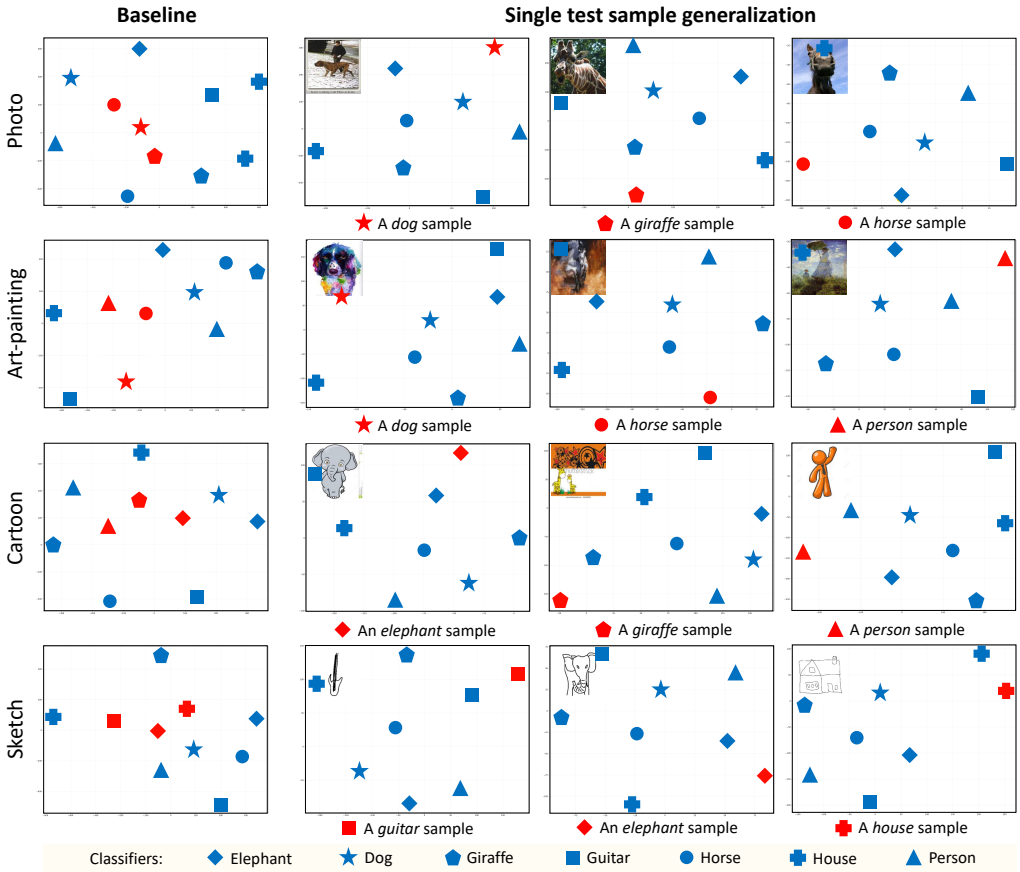


Figure 5: **Visualization of single test sample generalization on PACS.** We plot the test samples and classifiers in the same 2D plane. Different shapes denote different categories. Samples are in red, classifiers in blue. The baseline method produces the same classifier for all samples from each domain, while our method generates the classifier adapted to each target sample. The two methods use the same test samples from each domain. The test sample is shown in the left-top corner of each sub-figure. The visualization shows why our model performs better.

Failure cases To gain insights into our method, we also provide some failure cases in Figure 6. Our method gets confused when samples have objects of different categories in the same image, as shown in the first row. Although failing in these cases, our model provides correct predictions for the other objects in the image. In the visualizations, the classifier with the same category of the labeled object is also close to the feature, which demonstrates the effectiveness of our method. In the middle two rows, our method struggles with the samples that contain multiple objects or complex backgrounds. A possible reason is that to take full advantage of the given sample, we utilize multi-level features of the target sample to generate the adapted classifier. However, the complex background and multiple objects in the image bring too much noisy information, leading to less adapted classifiers for the given sample. A solution can be extracting information of the target sample selectively to reduce noise in the features. Moreover, in the "sketch" domain, as shown in the last row, the samples usually have less information than other domains. This makes our method more sensitive to the given information. When there is noisy information, e.g., the chair in the left figure, or object with unobvious features, e.g., the elephant in the right figure, the method fails to generate proper classifiers for the sample.

D More comparisons with Tent

We provide more detailed comparisons with Tent Wang et al. (2021) in Figure 7 to further show the effectiveness of our method. The three rows in Figure 7 show the accuracy of different settings on $0°$, $45°$, and $90°$ target domain, respectively. The mean accuracy is shown in Figure 3 in the main paper.

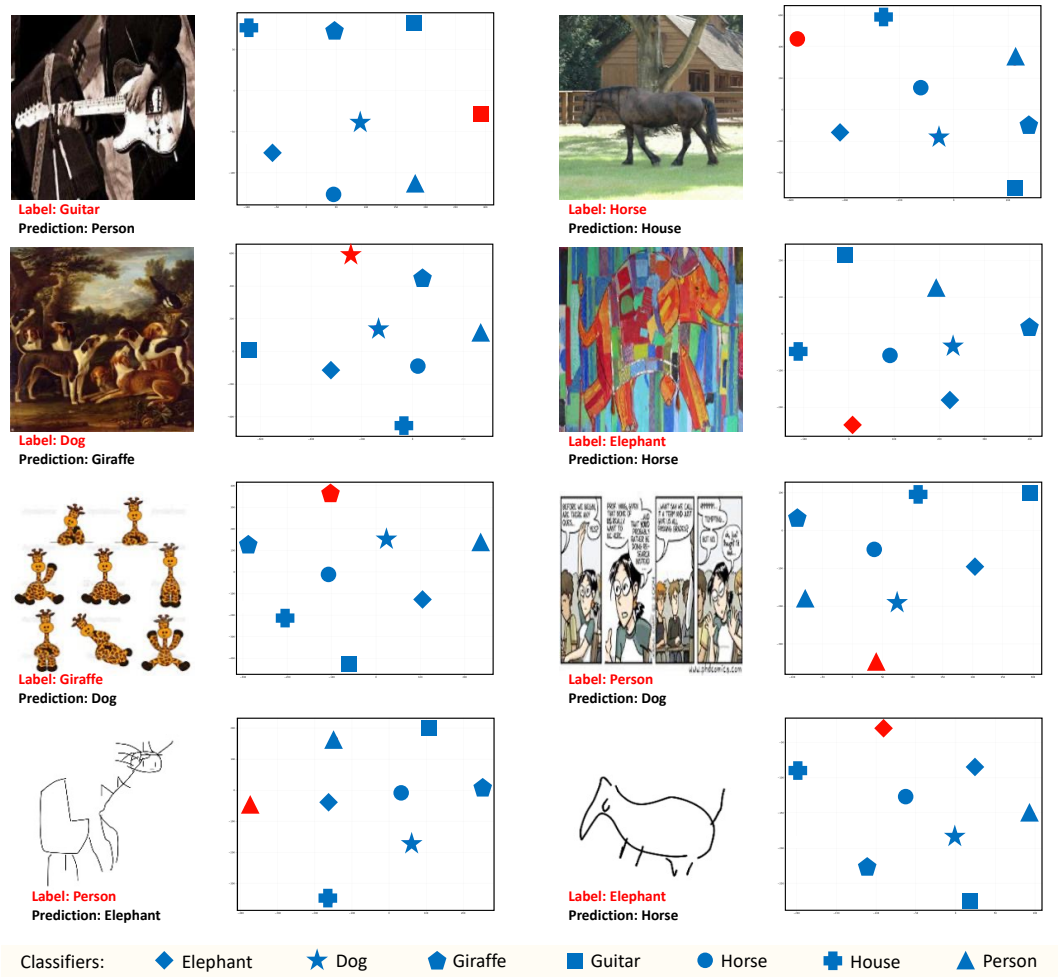


Figure 6: **Failure case visualizations of our method.** The visualization settings are the same as Figure 5. Our method makes wrong predictions, be it that the correct classifiers are also close to the samples, which shows that the classifiers generated by our method are still adapted to the given sample in these cases.

The same as in the main paper, the left column shows results under the "single target domain" setting, while the right column shows the "multiple target domains" setting. In the "single target domain" setting, Tent is utilized to adapt the base model trained on source domains to each target domain independently. By contrast, in the "multiple target domains" setting, the base model is adapted and evaluated on three target domains jointly, without the domain identifiers. The base model is adapted by Tent using different number of samples, e.g., 1, 32, 128, with different optimization steps, e.g., 1, 10, 100. The same conclusion can be drawn here as in the main paper. Compared with Tent, our method adapts the model to each target sample individually by a single and efficient feed-forward pass. Thus, our performance is independent to the settings and optimization steps.

Specifically, as shown in the left column, Tent achieves better performance with both larger numbers of samples and fine-tuning steps, especially on 0° and 90° domains. However, with fewer target samples being available, and ultimately just one sample, the adaptation of Tent starts to suffer, or even hurt the accuracy with more optimization steps. When adapted to three target domains jointly, as shown in the right column, the adapted performance is similar with the "single target domain" settings with few target samples. With more target samples, the performance degrades, worse than the left column. Accuracy on 0° domain improves slightly, while on 45° and 90° domain even drops as the adaptation is affected by samples from other domains. In contrast, since our method adapts the model to each target sample individually, we achieve consistently good performance regardless of

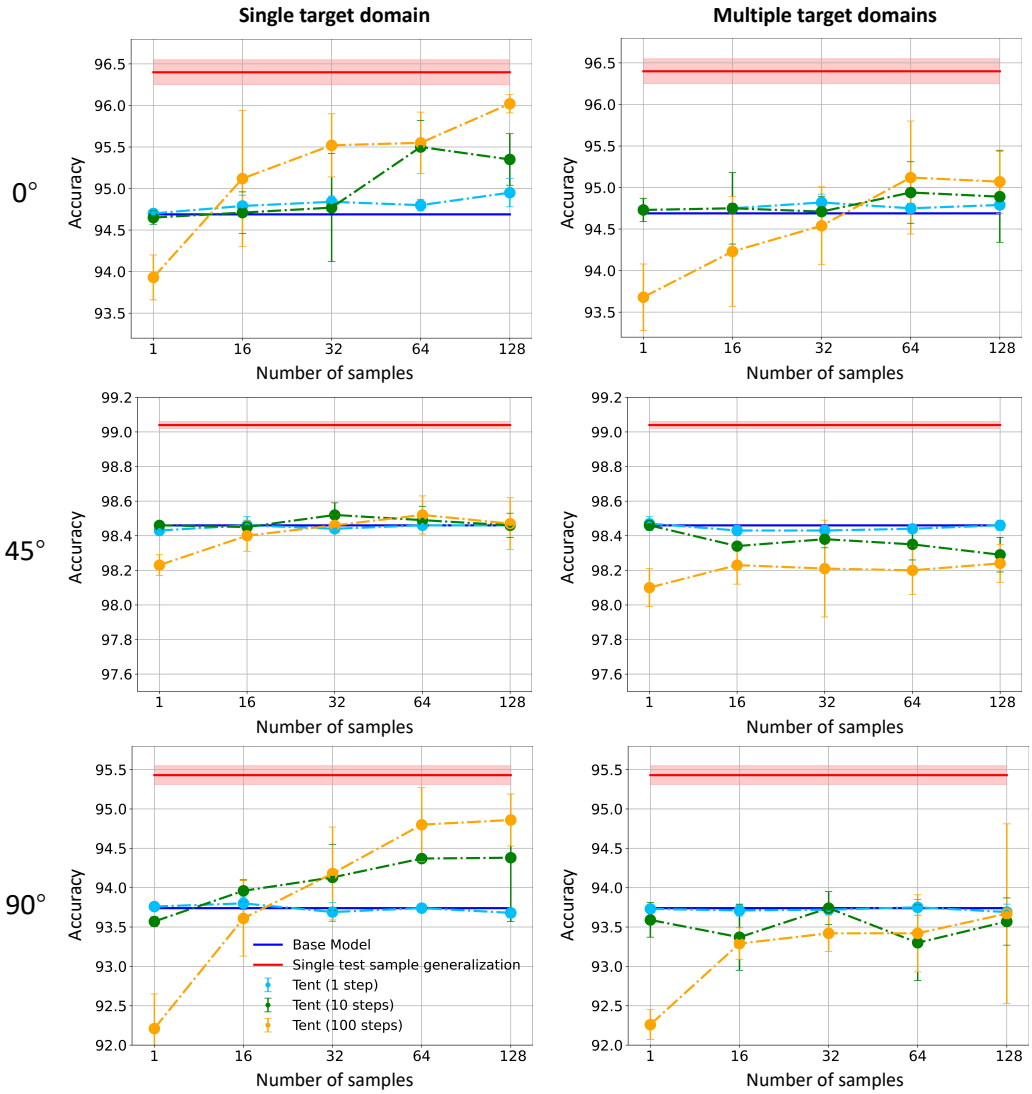


Figure 7: **Single test sample generalization vs. Tent for different settings on rotated-MNIST.** Tent shows good performance with a large batch of samples from a single domain (left). When provided only one sample or when given samples from different domains, without their domain id (right), Tent suffers. In contrast, our method is independent to the number of target samples and domains.

how many domains the test samples cover. This demonstrates the benefit of our single test sample generalization.