

# Rethinking the Invariant Feature Learning: Variational Bayesian Inference for Domain Generalization

Xiaofeng Liu<sup>1,2</sup> Bo Hu<sup>2,3</sup> Linghao Jin<sup>2,4</sup> Xu Han<sup>2,4</sup> Fangxu Xing<sup>1</sup> Jinsong Ouyang<sup>1</sup> Jun Lu<sup>2</sup>  
 Georges El Fakhri<sup>1</sup> Jonghye Woo<sup>1</sup>

<sup>1</sup>Dept. of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

<sup>2</sup>Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA

<sup>3</sup>National University of Singapore, Singapore

<sup>4</sup>Johns Hopkins University, Baltimore, MD, USA

liuxiaofengcmu@gmail.com, huboniccolo@bupt.edu.cn, ljin23@jhu.edu, xhan32@jhu.edu, jlu@bidmc.harvard.edu, {fxing1, ouyang.jinsong, elfakhri, jwoo}@mgh.harvard.edu

## Abstract

In this work, we propose a domain generalization (DG) approach to learn on several labeled source domains and transfer knowledge to a target domain that is inaccessible in training. Considering the inherent conditional and label shifts, we would expect the alignment of  $p(x|y)$  and  $p(y)$ . However, the widely used domain invariant feature learning (IFL) methods relies on aligning the marginal concept shift w.r.t.  $p(x)$ , which rests on an unrealistic assumption that  $p(y)$  is invariant across domains. We thereby propose a novel variational Bayesian inference framework to enforce the conditional distribution alignment w.r.t.  $p(x|y)$  via the prior distribution matching in a latent space, which also takes the marginal label shift w.r.t.  $p(y)$  into consideration with the posterior alignment. Extensive experiments on various benchmarks demonstrate that our framework is robust to the label shift and the cross-domain accuracy is significantly improved, thereby achieving superior performance over the conventional IFL counterparts.

## 1 Introduction

Deep learning usually relies on the independent, identically distributed (i.i.d.) assumption of training and testing datasets, while target tasks are usually significantly heterogeneous and diverse [Che *et al.*, 2021; Liu *et al.*, 2021a]. This motivates many researchers to investigate domain adaptation (DA) [Liu *et al.*, 2021b] and domain generalization (DG) [Matsuura and Harada, 2020a]. The source and target domain shifts are expected to be compensated for by a variety of adaptation steps. Despite the success of DA in several tasks, much of the prior work relies on utilizing the massive labeled/unlabeled target samples for its training.

The recently aroused DG task [Matsuura and Harada, 2020a] assumes that several labeled source domains are available in training without access to the target sample/label. Collectively exploiting these source domains can potentially lead to a trained system that can be generalized well on a target

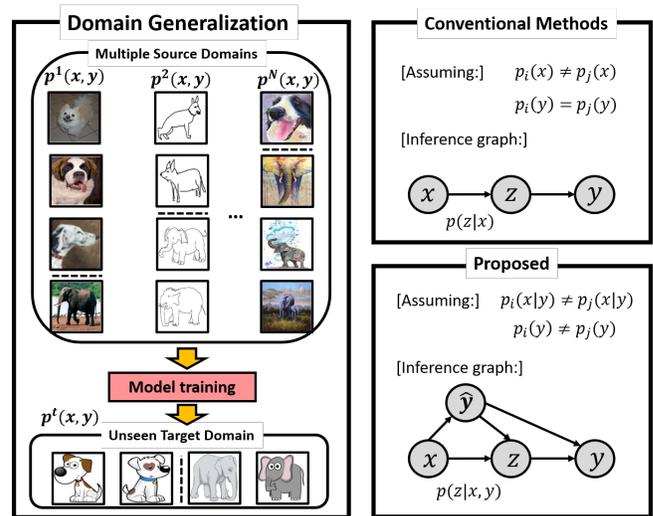


Figure 1: Illustration of the DG and the comparison of conventional IFL and our solution w.r.t. the presumption and inference graph.

domain [Hu *et al.*, 2019]. A predominant stream in DG is the domain invariant feature learning (IFL) [Ghifary *et al.*, 2017], which attempts to enforce  $p^i(f(x)) = p^j(f(x))$ , where  $i$  and  $j$  index the two different source domains. The typical solution can be obtained via momentum or adversarial training [Ghifary *et al.*, 2017].

However, the different classes can follow different domain shift protocols, e.g., the street lamp will be sparkly at night, while the pedestrian is shrouded in darkness. Therefore, we would expect the fine-grained class-wise alignment of the condition shift w.r.t.  $p(f(x)|y)$ , where  $f(\cdot)$  is a feature extractor [Zhang *et al.*, 2013; Gong *et al.*, 2016; Combes *et al.*, 2020].

Assuming there is no concept shift (i.e.,  $p^i(y|f(x)) = p^j(y|f(x))$ ) and label shift (i.e.,  $p^i(y) = p^j(y)$ ), and given the Bayes' theorem,  $p(f(x)|y) = \frac{p(y|f(x))p(f(x))}{p(y)}$ , IFL is able to align  $p(f(x)|y)$  if  $p^i(f(x)) = p^j(f(x))$ .

However, the label shift  $p^i(y) \neq p^j(y)$  [Liu *et al.*, 2021b], i.e., class imbalance, is quite common in DG, as illustrated

in Fig. 1. Since  $f(\cdot)$  is a deterministic mapping function, IFL is able to encode the domain invariant representation under the *covariate shift* assumption (i.e., only  $p^i(x) \neq p^j(x)$ ) [Moreno-Torres *et al.*, 2012]. Under the label shift, the *covariate* alignment cannot be used as an alternative of *conditional* alignment (i.e.,  $p^i(x|y) = p^j(x|y)$ ) [Li *et al.*, 2018b]. Actually, both the conditional and label shifts are the realistic setting in most of DG tasks.

Recently, [Li *et al.*, 2018b] proposes to align the conditional shift, assuming that there is no label shift. However, it is ill-posed to only consider one of conditional or label shift [Zhang *et al.*, 2013; Kouw, 2018]. To mitigate this, both the *conditional* and *label* shifts are taken into account for DA from a causal interpretation view [Zhang *et al.*, 2013; Gong *et al.*, 2016]. However, its linearity assumption might be too restrictive for real-world challenges.

In this work, we first analyze the different shift conditions in real-world DG, and rethink the limitation of conventional IFL under different shift assumptions. Targeting the conditional and label shifts, we propose to explicitly align  $p(f(x)|y)$  and  $p(y)$  via variational Bayesian inference and posterior label alignment. Note that the fine-grained class-wise  $p(f(x)|y)$  alignment can lead to  $p(f(x))$  alignment following the law of total probability [Zhang *et al.*, 2013].

Aligning the conditional distribution  $p(f(x)|y)$  across source domains under the label shift is usually intractable. Thus, we propose to infer the domain-specific variational approximations of these conditional distributions, and reduce the divergence among these approximations.

Specifically, we enforce the conditional domain invariance by optimizing two objectives. The first one enforces the approximate conditional distributions indistinguishable across domains by matching their reparameterized formulations (i.e., the mean and variance of Gaussian distribution). The second one maximizes the probability of observing the input  $x$  given the latent representation and domain label, which is achieved by a domain-wise likelihood learning network. Assuming that the conditional shift is aligned, we can then align the posterior classifier with the label distribution following a plug-and-play manner.

The main contributions are summarized as follows:

- We explore both the conditional and label shifts in various DG tasks, and investigate the limitation of conventional IFL methods under different shift assumptions.
- We propose a practical and scalable method to align the conditional shift via the variational Bayesian inference.
- The label shift can be explicitly aligned by the posterior alignment operation.

We empirically validate its effectiveness and generality of our framework on multiple challenging benchmarks with different backbone models and demonstrate superior performance over the comparison methods.

## 2 Related Work

DG assumes that we do not have prior knowledge about the target domain in that we do not have access to labeled/unlabeled samples of the target domain at the training stage [Matsuura and Harada, 2020a]. The conventional DG

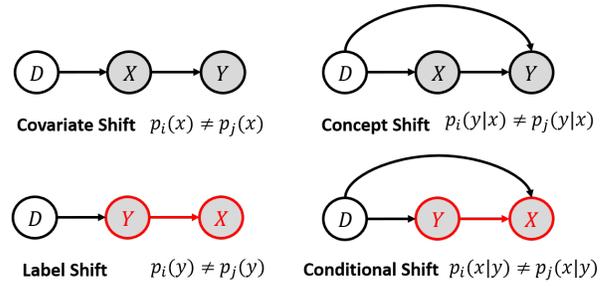


Figure 2: Summary of the possible shifts/combinations, and their inherent causal relationships. We target the last one, and  $p(x|y)$  is the fine-grained class-wise distribution of  $p(x)$  [Zhang *et al.*, 2013]. The observed variables are shaded. The red circle is used to emphasize the  $y$  to  $x$  causality.

methods can be divided into two categories. The first strategy aims to extract the domain invariant features with IFL [Ghifary *et al.*, 2017]. A typical solution is based on adversarial learning, which reduces the inter-domain divergence between the feature representation distributions [Li *et al.*, 2018a; Liu *et al.*, 2021a].

The other strategy focuses on the fusion of domain-specific feature representations (DFR-fusion). [Mancini *et al.*, 2018] builds the domain-specific classifiers by multiple independent convolutional networks. Then, it uses a domain agnostic component to fuse the probabilities of a target sample belonging to different source domains. [Ding and Fu, 2017] infers the domain-invariant feature by matching its low-rank structure with domain-specific features. Typically, these DG methods assume that  $p(y)$  is invariant across domains. Therefore, aligning  $p(f(x))$  can be a good alternative to align the conditional shift. However, this assumption is often violated due to the label shift in real-world applications. Therefore, independent conditional and label shift assumptions are more realistic in real-world applications.

Domain shifts in DA can be categorized into covariant, label, conditional, and concept shifts [Moreno-Torres *et al.*, 2012; Zhang *et al.*, 2015]. In this work, we examine these concepts and adapt their causal relationships to DG, as summarized in Fig. 2. Conventionally, each shift is studied independently, by assuming that the other shifts are invariant [Kouw, 2018]. For example, [Li *et al.*, 2018b] aligns the conditional shift assuming that no label shift occurs. We note that the concept shift usually has not been considered in DG tasks, since most of the prior work assumes that an object has different labels in different domains. Some recent works [Zhang *et al.*, 2013; Gong *et al.*, 2016; Combes *et al.*, 2020] assume that both conditional and label shifts exist in DA tasks and tackle the problem with a causal inference framework. However, its linearity assumption and location-scale transform are too restrictive to be applied in many real-world applications. It is worth noting that under the conditional and label shift assumption,  $y$  is the cause of  $x$ , and therefore it is natural to infer  $p(x|y)$  of different domains directly as in [Schölkopf *et al.*, 2012; Gong *et al.*, 2018] as a likelihood maximization network.

In this work, we propose a novel inference graph as shown

in Fig. 1 to explicitly incorporate the conditional dependence, which is trained via variational Bayesian inference.

### 3 Methodology

We denote the input sample, class label, and domain spaces as  $\mathcal{X}$ ,  $\mathcal{Y}$ , and  $\mathcal{D}$ , respectively. With random variables  $x \in \mathcal{X}$ ,  $y \in \mathcal{Y}$ , and  $d \in \mathcal{D}$ , we can define the probability distribution of each domain as  $p(x, y|d)$ . For the sake of simplicity, we assume  $y$  and  $d$  are the discrete variables for which  $\mathcal{Y} = \{1, 2, \dots, K\}$  is the set of class labels. In DG, we are given  $N$  source domains  $\{p^i(x, y)\}_{i=1}^N$  to train the latent representation encoder  $f(x)$  and the representation classifier  $p(y|f(x))$  [Li *et al.*, 2018a]. The trained and fixed encoder and classifier are used to predict the labels of samples drawn from the marginal distribution  $\{p^t(x)\}$  of an “unseen” target domain  $\{p^t(x, y)\}$ .

The conventional IFL assumes that  $p(y)$  and  $p(y|x)$  are invariant across domains. Since  $f(\cdot)$  is a deterministic mapping function,  $p(y|f(x))$  should also be invariant across domains. Therefore, if  $p(f(x))$  of different domains are aligned, the conditional shift,  $p(f(x)|y) = \frac{p(y|f(x))p(f(x))}{p(y)}$ , is also aligned. However, with the conditional and label shift assumption, the alignment of  $p(f(x)|y)$  and  $p(y)$  is more challenging than the covariant shift which only requires to align the marginal distribution  $p(f(x))$ .

We note that with the law of total probability,  $p(f(x))$  can be aligned if the fine-grained  $p(f(x)|y)$  is aligned [Zhang *et al.*, 2013]. Moreover,  $p(y)$  for all source domains can be calculated by simply using the class label proportion in each domain. Besides, modeling  $p(f(x)|y)$  is natural, since it follows the inherent causal relation under the conditional and label shift (see Fig. 2).

For the simplicity and consistency with the notation of autoencoder, we denote  $f(x)$  as  $z^1$ , which is the latent variable encoded from  $x$  by  $f$ . We note that  $z = f(x)$  is dependent on its corresponding input sample  $x$ . The class conditional distribution  $p(f(x)|y)$  can be reformulated as  $p(z|x, y)$ .

The corresponding inference graph and detailed framework are shown in Fig. 1 and Fig. 3, respectively. When inferring the latent representation  $z$ , we explicitly concatenate  $x$  and  $y$  as input. Moreover, the final class prediction is made by a posterior alignment of the label shift, which also depends on label distribution  $p(y)$ .

#### 3.1 Variational Bayesian Conditional Alignment

Although  $p(f(x))$  and  $p(y)$  for all source domains can be modeled by IFL and class label proportion, respectively,  $p(z|x, y)$  is usually intractable for moderately complicated likelihood functions, e.g., neural networks with nonlinear hidden layers. While this could be solved by the Markov chain Monte Carlo simulation (MCMC), this requires expensive iterative inference schemes per data point and does not scale to the large-scale high-dimensional data.

To align the class-dependent  $p(z|x, y)$  across different domains, we first construct its approximate Gaussian distribution  $q(z|x, y)$ , and resort to the variational Bayesian inference

<sup>1</sup>We use  $f(x)$  and  $z$  interlaced to align with the conventional IFL and variational autoencoder literature, respectively.

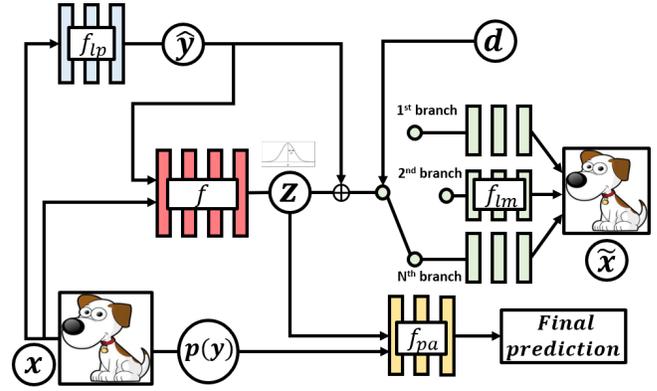


Figure 3: Illustration of our framework, which consists of a conditional feature encoder  $f$ , a domain-wise likelihood maximization network  $f_{lm}$ , a label prior network  $f_{lp}$ , and a classifier  $f_{cls}$ . Our latent representation  $z$  is aligned with the Gaussian prior.  $f$ ,  $f_{lp}$ ,  $f_{cls}$  are used in testing,  $f_{lm}$  is only for training.

[Kingma and Welling, 2013] to bridge it with a simple Gaussian family for which the inference is tractable. Specifically, we have the following proposition:

**Proposition 1.** The minimization of the inter-domain conditional shift  $p(z|x, y)$  is achieved when its approximate distribution  $q(z|x, y)$  is invariant across domains, and the KL divergence between  $p(z|x, y)$  and  $q(z|x, y)$  is minimized.

**Proof 1.** A simple justification can be:

if we have  $KL(q_1(z|x, y)||p_1(z|x, y)) = 0$ ,

$$KL(q_2(z|x, y)||p_2(z|x, y)) = 0,$$

$$KL(q_1(z|x, y)||q_2(z|x, y)) = 0,$$

then, we have  $KL(p_1(z|x, y)||p_2(z|x, y)) = 0$ .

Following the variational bound [Kingma and Welling, 2013], minimizing the KL divergence between  $p(z|x, y)$  and  $q(z|x, y)$  is equivalent to maximizing the evidence lower bound (ELBO) [Domke and Sheldon, 2018] of the likelihood  $\log p(x, y, z)$ , denoted by  $\mathcal{L}$ :

$$\max \mathcal{L} = \min D_{KL}(q(z|x, y)||p(z|x, y)), \quad (1)$$

where the  $D_{KL}$  term is the KL divergence of the approximate from the true posterior and the ELBO of the likelihood  $\log p(x, y, z)$ , i.e.,  $\mathcal{L}$  can be re-written as

$$\mathcal{L} = \int q(z|x, y) \log \frac{p(x, y, z)}{q(z|x, y)} dz, \quad (2)$$

which can be reformulated as

$$\mathcal{L} = -D_{KL}(q(z|x, y)||p(z|y)) + E_{z \sim q(z|x, y)}[\log p(x|y, z)], \quad (3)$$

where  $E$  denotes the expectation. Therefore, approximating  $p(z|x, y)$  with  $q(z|x, y)$  requires two objectives, i.e., minimizing  $D_{KL}(q(z|x, y)||p(z|y))$ , while maximizing the expectation of  $\log p(x|y, z)$ .

For  $N$  domains,  $p^i(z|y)$ ,  $i \in \{1, 2, \dots, N\}$  is the prior distribution in the variational model, e.g., multivariate Gaussian distribution. When  $p^i(z|y)$  is sampled from the same Gaussian distribution and is invariant across the source domains, the first objective in Eq. (3), i.e.,  $D_{KL}(q(z|x, y)||p(z|y))$ , can explicitly enforce  $q^i(z|x, y)$  to be invariant across domains.

By further incorporating the second objective into Eq. (3), we attempt to minimize the KL divergence of  $p(z|x, y)$  and  $q(z|x, y)$  as in Eq. (1). Then,  $p^i(z|x, y)$  should be invariant across the source domains, i.e.,

$$p^1(z|x, y) = p^2(z|x, y) = \dots = p^N(z|x, y) \quad (4)$$

Actually, if we have  $p_1(y|x) = p_1(x|y)p_1(y)/p_1(x)$ ,  $p_2(y|x) = p_2(x|y)p_2(y)/p_2(x)$ , and  $p_1(x|y) = p_2(x|y)$ ,  $p_1(x) = p_2(x)$ , then  $p_1(y|x)/p_1(y) = p_2(y|x)/p_2(y)$ .

The first optimization objective of Eq. (3) targets to align the conditional distribution  $q(z|x, y)$  across the source domains. Since the prior distribution  $p(z|y)$  is the multivariate Gaussian distribution, it is also natural to configure  $q(z|x, y)$  as multivariate Gaussian. Practically, we follow the reparametric trick of variational autoencoder [Kingma and Welling, 2013] in such a way that the inference model has two outputs, i.e.,  $\mu_j^i$  and  $\sigma_j^i$  which are the mean and variance of the Gaussian distribution  $q(z_j^i|x_j^i, y_j^i)$ . Then,  $z_j^i = \mu_j^i + \sigma_j^i \odot \epsilon$ , where  $\epsilon \in N(0, I)$ . Without loss generality, we have:

$$L_1 = \sum_{i=1}^N \sum_{j=1}^{M_i} \left[ -\log \frac{\sigma_j^i}{\bar{\sigma}} + \frac{\bar{\sigma}^2 + (\bar{\mu} - \mu_j^i)^2}{2(\sigma_j^i)^2} - \frac{1}{2} \right], \quad (5)$$

where  $M_i$  is the number of input in a batch from domain  $i$ . Usually, we set the prior  $p^i(z|y)$  to be the standard multivariate Gaussian distribution, where the mean and variance are  $\bar{\mu} = 0$  and  $\bar{\sigma} = 1$ , respectively.

The second optimization objective of Eq. (3) aims to maximize the probability of observing the input  $x$  given  $y$  and  $z$ . We propose to configure a likelihood maximization network  $f_{lm}$ . It maximizes the likelihood that the latent feature representation of images in a specific domain can effectively represent the same-class images in this domain. Practically, our  $f_{lm}$  contains  $N$  sub-branches, each of which corresponds to a domain. At the training stage, we choose the corresponding branch according to the source domain label  $d$ . Its loss can be formulated as

$$L_2 = \sum_{i=1}^N \sum_{j=1}^{M_i} \|f_{lm}^i(z_j^i, y_j^i) - x_j^i\|^2, \quad (6)$$

which solves the maximum likelihood problem by minimizing the difference between the input data and the generated data in the corresponding domain. We note that  $f_{lm}$  is only used in training.

Posterior collapses is a long-lasting problem of VAEs using continuous Gaussian prior. The recent progress of discrete priors or flow model can be potentially added on our model. However, our training does not suffer from it significantly. Note that each of the  $N$  decoders is only trained with about  $1/N$  data of encoder  $f$  and it is likely that the weak decoder can also be helpful. Using an approximation with multivariate Gaussian prior offers much better tractability as in variational autoencoder (VAE), which is good for posterior modeling.

### 3.2 Label-prior

Inferencing the latent representation  $z$  requires to know the label information  $y$  in advance, since we are modeling the

---

### Algorithm 1 VB inference under conditional and label shifts

---

Initialize network parameters

**repeat**

//Construct the mini-batch for the training

$\{x, y, d\} \leftarrow$  random sampling from the dataset

//Forward pass to compute the inference output

$\hat{y} \leftarrow f_{lp}(x); z \leftarrow f(x, \hat{y});$

$\hat{x} \leftarrow f_{lm}(z, \hat{y}); y \leftarrow f_{cls}(z, y)$

//Calculate the loss functions:  $L_1, L_2, L_{CE1}, L_{CE2}$

//Update parameters according to gradients

$f \leftarrow L_1 + \alpha L_2 + \beta L_{CE2} + \theta L_{\hat{y}};$

$f_{lp} \leftarrow L_{CE1};$

$f_{lm} \leftarrow L_2;$

$f_{cls} \leftarrow L_{CE2}$

**until** deadline

---

approximate conditional distribution  $q(z|x, y)$ . Although  $y$  is always available in training, the ground truth  $y$  is not available in testing. We note that  $f_{lm}$  is only used in training which always has  $y$ .

To alleviate this limitation, we infer the label from the input image  $x$  as a prior to control the behavior of the conditional distribution matching module. Specifically, we configure a label-prior network  $f_{lp} : x \rightarrow \hat{y}$  to infer the pseudo-label, and use  $\hat{y}$  as input to the posterior label alignment classifier  $f_{cls}$  in both training and testing. Our label prior network  $f_{lp}$  is trained by the cross-entropy loss  $L_{CE1}$  with the ground-truth label  $y$ .

Moreover, at the training stage, we can further utilize the ground-truth  $y$  and minimizing  $D_{KL}(q(z|x, y)||q(z|x, \hat{y}))$  to update the encoder  $f$ . We denote the to be minimized KL divergence as  $L_{\hat{y}}$ . Minimizing  $L_{\hat{y}}$  is not mandatory, while  $L_{\hat{y}}$  can encourage the encoder to be familiar with the noisy  $\hat{y}$  [Lian *et al.*, 2019] and learn to compensate for the noisy prediction. We note that assigning  $\hat{y}$  to an uniform histogram as the dialog system [Lian *et al.*, 2019] to fill the missing variate  $y$  can degenerate the modeling of  $p(f(x)|y)$  to  $p(f(x))$  in our DG task. Therefore, the pseudo-label will be post-processed by both encoder and classifier, which may adjust the unreliable  $\hat{y}$ .

### 3.3 Posterior Alignment with Label Shift

Finally, we align  $p(y)$  to obtain the final classifier  $f_{cls}$ . Since the classifier is deployed on all of the source domains, we can regard all of the source domains as a single domain, and denote the classifier as

$$\bar{p}(y|f(x)) = \frac{\bar{p}(f(x)|y)\bar{p}(y)}{\bar{p}(f(x))}, \quad (7)$$

where  $\bar{p}(f(x)|y)$ ,  $\bar{p}(f(x))$ , and  $\bar{p}(y)$  are its class-conditional, latent representation, and label distribution, respectively.

Suppose that all the conditional distribution  $p^i(f(x)|y)$  and the latent distribution  $p^i(f(x))$  are aligned to each other using variational Bayesian inference, they should also be aligned with  $\bar{p}(f(x)|y)$  and  $\bar{p}(f(x))$ . Therefore, the posterior alignment and the final prediction of the sample  $f(x)$  from domain

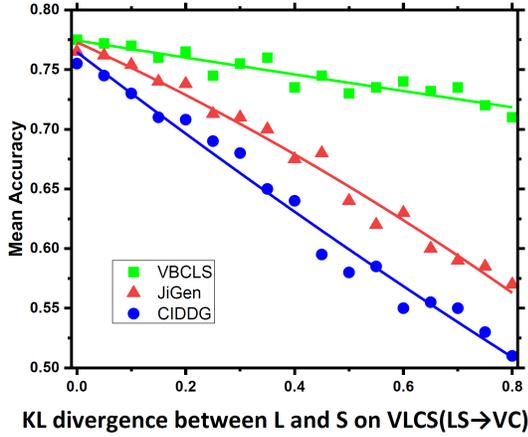


Figure 4: Left: Quantification of DG performance on VLCS (LS→VC) under different source label shifts ( $p^L(y) \neq p^S(y)$  by sampling).

$i$  can be formulated as

$$p^i(y = k|f(x)) = \frac{p^i(y = k)}{\bar{p}(y = k)} \bar{p}(y = k|f(x)), \quad (8)$$

where  $\bar{p}(y = k|f(x))$  is the  $k$ -th element value of the classifier’s softmax prediction. Here, we also calculate the cross entropy loss  $L_{CE2}$  between  $p^i(y = k|f(x))$  and the ground truth label  $y$ .

As detailed in Algorithm 1, we update  $f$  with  $L_1 + \alpha L_2 + \beta L_{CE2} + \theta L_{\hat{y}}$ , update  $f_{lp}$  with  $L_{CE1}$ , update  $f_{lm}$  with  $L_2$ , and update  $f_{cls}$  with  $L_{CE2}$ , respectively.

## 4 Experiments

This section demonstrates the effectiveness of our variational Bayesian inference framework under conditional and label shift (VBCLS) on the classic VLCS DG benchmark for image classification and the PACS benchmark for object recognition with domain shift.

### 4.1 Implementation Details

The domain invariant encoder  $f$  and posterior alignment classifier  $f_{pa}$  use the encoder and classifier structure as our compared models (e.g., AlexNet and ResNet18), and the label prior network is simply the concatenation of  $f$  and  $f_{pa}$ , and the likelihood maximization network  $f_{lm}$  uses the reversed CNN decoder structure of  $f$ . We implement our methods using PyTorch, we empirically set  $\alpha = 0.5$ ,  $\beta = 1$ ,  $\theta = 0.1$  and  $t = 100$  via grid searching. In our experiments, the performance does not sensitive to these hyperparameters for a relatively large range.

Following previous work on domain generalization [Li *et al.*, 2017; Li *et al.*, 2019], we use models pre-trained on the ILSVRC dataset and label smoothing on the task classifier in order to prevent overfitting. The model is trained for 30 epochs via the mini-batch stochastic gradient descent (SGD) with a batch size of 128, a momentum of 0.9, and a weight decay of  $5e - 4$ . The initial learning rate is set to  $1e - 3$ ,

Target (→)	V	L	C	S	Average
CCSA 2017	67.10	62.10	92.30	59.10	70.15
MMD-AAE 2018	67.70	62.60	94.40	64.40	72.28
CIDDG 2018	64.38	63.06	88.83	62.10	69.59
Epi-FCR 2019	67.10	64.30	94.10	65.90	72.90
JiGen 2019	70.62	60.90	96.93	64.30	73.19
MetaVIB 2020	70.28	62.66	97.37	67.85	74.54
RCS 2020	73.93	61.86	97.61	68.32	75.43
VBCLS	72.16	68.63	96.52	70.37	76.92±0.06
VBCLS- $f_{pa}$	69.40	65.00	94.60	65.60	73.70±0.08
VBCLS- $L_{\hat{y}}$	71.74	68.18	96.20	70.02	76.56±0.07

Table 1: Classification accuracy (mean±sd) using leave-one-domain-out validation on the VLCS benchmark.

which is scaled by a factor of 0.1 after 80% of the epochs. For the VLCS dataset, the initial learning rate is set to  $1e - 4$ , since a high learning rate leads to early convergence due to the overfitting in the source domain. In addition, the learning rate of the domain discriminator and the classifier is set to be larger (i.e., 10 times) than that of the feature extractor.

For the ablation study, VBCLS- $f_{pa}$  denotes without posterior alignment and VBCLS- $L_{\hat{y}}$  denotes without minimizing  $D_{KL}(q(z|x, y)||q(z|x, \hat{y}))$ .

### 4.2 VLCS Dataset

VLCS [Ghifary *et al.*, 2017] contains images from four different datasets including PASCAL VOC2007 (V), LabelMe (L), Caltech-101 (C), and SUN09 (S). Different from PACS, VLCS offers photo images taken under different camera types or composition bias. The domain V, L, C, and S have 3,376, 2,656, 1,415, and 3,282 instances, respectively. Five shared classes are collected to form the label space, including bird, car, chair, dog, and person. We follow the previous works to exploit the publicly available pre-extracted DeCAF6 features (4,096-dim vector) [Donahue *et al.*, 2014] for leave-one-domain-out validation by randomly splitting each domain into 70% training and 30% testing. We report the mean over five independent runs for our results.

The mean accuracy on the test partition using leave-one-domain-out validation on VLCS is reported in Table 1. We also compare our method with the DG models trained using the same architecture and source domains.

Our results indicate that our VBCLS outperforms the covariant shift setting methods (e.g., CCSA [Motiian *et al.*, 2017], MMD-AAE [Li *et al.*, 2018a]) by a large margin. The improvement over the conditional shift only method CIDDG [Li *et al.*, 2018b] is significant, demonstrating the necessity of incorporating both conditional and label shift. When compared with the recent SOTA methods, e.g., self-challenging based RSC [Huang *et al.*, 2020], information bottleneck based MetaVIB [Du *et al.*, 2020], and self-training based EISNet [Wang *et al.*, 2020], we can observe that our VBCLS yields better performance in almost all cases. We note that JiGen [Carlucci *et al.*, 2019] uses the Jigsaw Puzzles solving, which are essentially different from the IFL.

The good performance of our strategy indicates the invariant feature learning works well, and the conditional and shift assumption can be an inherent property that needs to be addressed for real-world DG challenges. The discrepancy be-

Target	Method	Source					
		VC	VL	VS	LC	LS	CS
2*V	CIDDG	-	-	-	60.42	62.21	59.56
	VBCLS	-	-	-	65.82	68.45	66.76
2*L	CIDDG	53.24	-	52.27	-	-	49.58
	VBCLS	60.75	-	60.76	-	-	58.82
2*C	CIDDG	-	78.82	78.58	-	74.67	-
	VBCLS	-	85.56	86.68	-	81.47	-
2*S	CIDDG	59.04	56.29	-	59.80	-	-
	VBCLS	62.34	60.35	-	61.52	-	-

Table 2: Impact of reducing the number of source domains on VLCS. Note that the rows denote the two source domains.

	Guitar	House	Giraffe	Person	Horse	Dog	Elephant
Art Paint	184	295	285	449	201	379	255
Cartoon	135	288	346	405	324	389	457
Photo	186	280	182	432	199	189	202
Sketch	608	80	753	160	816	772	740

Table 3: Sample sizes for each <domain, class> pair in PACS dataset. The column shows category name.

tween marginal distributions  $\bar{p}(y)$  and  $p^t(y)$  is measured via the KL-divergence as the semi-supervised learning with the selective bias problem [Zadrozny, 2004]. The impact of the label shift is empirically illustrated in Fig. 4 left. In Fig. 4 right, we show the label refinement can effectively estimate the testing label distribution without tuning the network. The label alignment can be relatively accurate after 3 epochs and is almost stable after 5 epochs.

In the ablation study, we can see that posterior alignment is necessary if there is label shift. Besides, the performance of the label prior network can be improved by minimizing  $D_{KL}(q(z|x, y)||q(z|x, \hat{y}))$  on encoder  $f$ .

Another evaluation protocol on VLCS is to examine whether eliminating examples from one source domain impacts the performance on the target domain. This protocol is designed to evaluate the impact of the diversity of the source domains on the target accuracy. In this experiment, each target domain on models by training all combinations of the remaining domains as the source domains is evaluated, as shown in Table 2. In addition, the CIDDG baseline is included for reference. Results in Table 2 demonstrate that for all target domains, reducing the number of source domains from 3 (see Table 1) to 2 degrades the performance for all combinations of the source domains. We can see that, in some cases, excluding a particular source from the training substantially degrades the target loss. However, we can see that our VBCLS can still be more robust under these cases.

### 4.3 PACS Dataset

The object recognition benchmark PACS [Li *et al.*, 2017] consists of images divided into 7 classes from four different datasets, including Photo (P), Art painting (A), Cartoon (C), and Sketch (S). In Tab. 3, we provide the detailed statistics of the number of samples in each domain. It is clear that the class ratio is different across domains, which indicates the significant label shift.

As shown in the results (c.f. Table 4), the Sketch domain produces the lowest accuracy when used as the target domain,

Target (→)	P	A	C	S	Average
CrossGrad 2018	87.6	61.0	67.2	55.9	67.9
CIDDG 2018	78.7	62.7	69.7	64.5	68.9
Epi-FCR 2019	86.1	64.7	72.3	65.0	72.0
JiGen 2019	89.0	67.6	71.7	65.2	73.4
MMLD 2020	88.98	69.27	72.83	66.44	74.38
RSC 2020	90.88	71.62	66.62	75.11	76.05
EISNet 2020	91.20	70.38	71.59	70.25	75.86
MetaVIB 2020	91.93	71.94	73.17	65.94	75.74
VBCLS	92.12	70.60	77.36	70.19	77.55±0.07
VBCLS- $f_{pa}$	91.02	68.86	74.18	65.40	74.61±0.04
VBCLS- $L_{\hat{y}}$	91.77	70.54	76.56	70.33	77.19±0.06
JiGen 2019 [Res18]	96.03	79.42	75.25	71.35	80.51
MMLD 2020 [Res18]	96.09	81.28	77.16	72.29	81.83
EISNet 2020 [Res18]	95.93	81.89	76.44	74.33	82.15
RSC 2020 [Res18]	95.99	83.43	80.85	80.31	85.15
VBCLS [Res18]	97.21	84.63	82.06	79.25	86.73±0.05

Table 4: Classification accuracy (mean±sd) using leave-one-domain-out validation on PACS.

and therefore it is deemed the most challenging one. In light of this, we follow the previous work to tune the model using the S domain as the target domain and reuse the same hyperparameters for the experiments with the remaining domains. In Table 4, we show the result using our method, which was averaged over 5 different initializations alongside all the other comparison methods. Overall, we can see that our method yields better average performance over all source domains compared to previous SOTA methods. Among them, CrossGrad [Shankar *et al.*, 2018] synthesizes data for a new domain, MMLD [Matsuura and Harada, 2020b] using a mixture of multiple latent domains. It is also comparable to the recent self-challenging based RSC [Huang *et al.*, 2020], information bottleneck based MetaVIB [Du *et al.*, 2020], and self-training based EISNet [Wang *et al.*, 2020]. More importantly, we can observe that our method outperforms CIDDG [Li *et al.*, 2018b], an adversarial conditional IFL strategy, by a large margin. The ablation studies are also consistent with the results in VLCS and PACS. We also provide the results using the ResNet18 backbone.

## 5 Conclusion

In this work, we target to establish a more realistic assumption in DG that both conditional and label shifts arise independently and contemporarily. We theoretically analyze the inequilibrium of conventional IFL under the different shift assumptions. Motivated by that, a concise yet effective VBCLS framework based on variational Bayesian inference with the posterior alignment is proposed to reduce both the conditional shift and label shifts. Extensive evaluations verify our analysis and demonstrate the effectiveness of our approach, compared with IFL approaches.

## Acknowledgments

This work was supported in part by the Hong Kong GRF 152202/14E, PolyU Central Research Grant G-YBJW, and Jiangsu NSF (BK20200238).

## References

- [Carlucci *et al.*, 2019] Fabio M Carlucci, Antonio D’Innocente, Silvia Bucci, Barbara, and Tatiana. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019.
- [Che *et al.*, 2021] Tong Che, Xiaofeng Liu, Site Li, Yubin Ge, Ruixiang Zhang, Caiming Xiong, and Yoshua Bengio. Deep verifier networks: Verification of deep discriminative models with deep generative models. *AAAI*, 2021.
- [Combes *et al.*, 2020] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoff Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *arXiv*, 2020.
- [Ding and Fu, 2017] Zhengming Ding and Yun Fu. Deep domain generalization with structured low-rank constraint. *TIP*, 2017.
- [Domke and Sheldon, 2018] Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. In *NIPS*, 2018.
- [Donahue *et al.*, 2014] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.
- [Du *et al.*, 2020] Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. *ECCV*, 2020.
- [Ghifary *et al.*, 2017] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE T-PAMI*, 2017.
- [Gong *et al.*, 2016] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *ICML*, 2016.
- [Gong *et al.*, 2018] Mingming Gong, Kun Zhang, Biwei Huang, Clark Glymour, Dacheng, and Kayhan. Causal generative domain adaptation networks. *arXiv*, 2018.
- [Hu *et al.*, 2019] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. Domain generalization via multidomain discriminant analysis. In *UAI*, 2019.
- [Huang *et al.*, 2020] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. *ECCV*, 2020.
- [Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv*, 2013.
- [Kouw, 2018] Wouter M Kouw. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018.
- [Li *et al.*, 2017] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017.
- [Li *et al.*, 2018a] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018.
- [Li *et al.*, 2018b] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, and Kun Zhang. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018.
- [Li *et al.*, 2019] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *ICCV*, 2019.
- [Lian *et al.*, 2019] Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. Learning to select knowledge for response generation in dialog systems. *IJCAI*, 2019.
- [Liu *et al.*, 2021a] Xiaofeng Liu, Yang Chao, Jane J You, C-C Jay Kuo, and Bhagavatula Vijayakumar. Mutual information regularized feature-level frankenstein for discriminative recognition. *PAMI*, 2021.
- [Liu *et al.*, 2021b] Xiaofeng Liu, Xiongchang Liu, Bo Hu, Wenxuan Ji, Fangxu Xing, Jun Lu, Jane You, C-C Jay Kuo, Georges El Fakhri, and Jonghye Woo. Subtype-aware unsupervised domain adaptation for medical diagnosis. *AAAI*, 2021.
- [Mancini *et al.*, 2018] Massimiliano Mancini, Samuel Rota Bulo, Barbara Caputo, and Elisa Ricci. Best sources forward: domain generalization through source-specific nets. In *ICIP*, 2018.
- [Matsuura and Harada, 2020a] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. *AAAI*, 2020.
- [Matsuura and Harada, 2020b] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *AAAI*, 2020.
- [Moreno-Torres *et al.*, 2012] 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.
- [Motiian *et al.*, 2017] Saeid Motiian, Marco Piccirilli, Donald Adjeroh, and Gianfranco. Unified deep supervised domain adaptation and generalization. In *ICCV*, 2017.
- [Schölkopf *et al.*, 2012] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv*, 2012.
- [Shankar *et al.*, 2018] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha, and Sunita. Generalizing across domains via cross-gradient training. *arXiv*, 2018.
- [Wang *et al.*, 2020] Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. Learning from extrinsic and intrinsic supervisions for domain generalization. *ECCV*, 2020.
- [Zadrozny, 2004] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *ICML*, 2004.
- [Zhang *et al.*, 2013] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun. Domain adaptation under target and conditional shift. In *ICML*, 2013.
- [Zhang *et al.*, 2015] Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *AAAI*, 2015.