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

**AAAI 2021 Submission**

Paper ID 1076

## Abstract

In this work, we address the problem of domain generalization (DG) which aims to learn from seen labeled source domains to yield the generalization on an "unseen" target domain. Essentially, the successful DG requires the alignment of the class conditional distribution $p(x|y)$ across domains. While the current invariant feature learning (IFL) methods largely rely on the assumption that $p(y)$ is invariant across domains, thereby attempting to align the marginal distribution $p(x)$ instead. This assumption is often violated due to the label shift in real-world applications. To address this, we propose a variational framework to explicitly align the conditional distributions using the latent space matching with the prior distribution. In addition, we propose to match the marginal label distribution via a concise posterior alignment. Furthermore, the adaptive target label refinement is investigated. Extensive experiments on various benchmarks demonstrate that our framework is robust to the label shift and the cross-domain proxy A-distance is significantly reduced, thereby achieving superior performance over the conventional IFL counterparts.

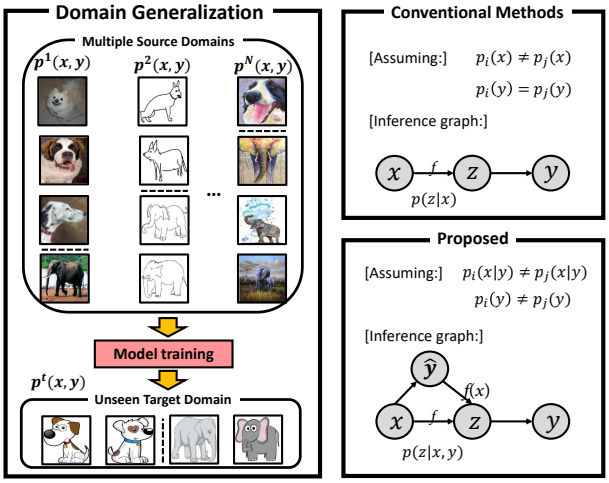



Figure 1: Illustration of the DG task and the comparison of coventional IFL methods and our solution w.r.t. the presumption and inference graph.

## Introduction

A basic assumption of deep learning is that the training and testing data are independently and identically distributed (Goodfellow et al. 2016). However, target tasks are usually significantly heterogeneous and diverse (Ghifary et al. 2017). This motivates many researchers to investigate domain adaptation (DA) and domain generalization (DG). With a variety of adaptation steps, the source and target domain shifts are expected to be compensated. 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 (Zou et al. 2019).

The recently aroused DG task (Matsuura and Harada 2020a) assumes that several labeled source domains are available in training without any 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 domain. Representative examples include hand-writing and speech recognition systems without any data from a new consumer (Hu et al. 2019). Also in the medical domain, the amount and quality of data that can be routinely acquired under different conditions across different places has increased significantly. Therefore, it is increasingly important to apply DG to yield improved diagnosis, therapeutic, and surgical procedures especially when using data collected from different places.

Essentially, we expect the alignment of conditional distribution $p(f(x)|y)$ across domains, where $f(\cdot)$ is a feature extractor (Zhang et al. 2013; Gong et al. 2016). A predominant stream in DG is the domain invariant feature learning (IFL) (Muandet et al. 2013; 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 (e.g., MMD) or adversarial training (Ghifary et al. 2017). The underlying assumption of IFL is that there is no concept shift (i.e., $p^i(y|f(x)) = p^j(y|f(x))$) and label shift (i.e., $p^i(y) = p^j(y)$)) for these domains. Given that $p(f(x)|y) = \frac{p(y|f(x))p(f(x))}{p(y)}$, the conditional distribution $p(f(x)|y)$ can be aligned if $p^i(f(x)) = p^j(f(x))$.

However, the label shift $p^i(y) \neq p^j(y)$, i.e., different

class proportions, 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. 2018c). Actually, both the conditional and label shifts are the realistic settings in most of DG tasks.

Recently, Li et al. (Li et al. 2018c) propose 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; Moreno-Torres et al. 2012). To mitigate this, both the $conditional$ and $label$ shifts are taken into account for DA tasks 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 tasks, and investigate the limitation of conventional IFL under different shift assumptions. Targeting to the covariate, conditional and label shifts, We propose to explicitly align $p(f(x))$, $p(f(x)|y)$, and $p(y)$ via conventional IFL, variational Bayesian inference, and posterior label alignment, respectively. Therefore, we can apply an unified representation classifier $p(y|f(x)) = \frac{p(f(x)|y)p(y)}{p(f(x))}$ for all of the domains.

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 and covariant shifts are aligned, we can then align the posterior classifier with the label distribution following a plug-and-play manner.

Although the unseen target setting in DG is realistic, we may expect a better performance on a target domain by accumulating the implementation experience. The posterior label alignment provides the flexibility to utilize the target sample without the need for target label or network retraining. Specifically, we can regard the inference result within a period of testing as pseudo-label and simply take its class proportion into account to refine our target class label distribution $p(y)$ prior.

The main contributions are be 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.

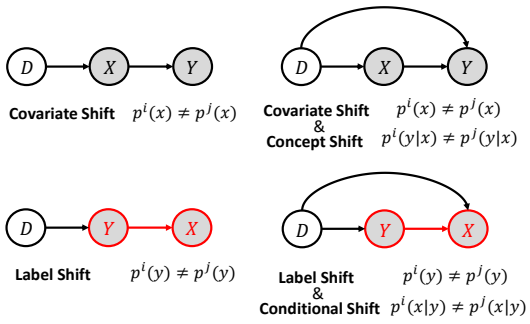



Figure 2: Summary of the possible shifts or shift combinations, and their inherent casual relationships.

• The label shift can be aligned by the posterior alignment operation. Moreover, the label distribution in testing can be adaptively updated based on the implementation experience following the self-refinement scheme.

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.

## 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 methods can be divided into two categories (Wang et al. 2019). The first strategy aims to extract the domain invariant features with IFL (Muandet et al. 2013; 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). In addition, Motiian et al. (Motiian et al. 2017) align the semantic probability distributions across domains by the point-wise distance instead of the distribution similarity. Kernel learning approaches to imposing the label-related constraints are proposed when learning domain-invariant feature representation (Ghifary et al. 2017; Li et al. 2018b). Khosla et al. (Khosla et al. 2012) ensure the feature representation to be stable across domains by applying the max-margin learning to eliminate the dataset-specific bias.

The other strategy focuses on the fusion of domain-specific feature representations (DFR-fusion). Mancini et al. (Mancini et al. 2018) build 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, Gong, and Schölkopf 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. (Li et al. 2018c) align the conditional shift assuming that no label shift occurs. We note that the concept shift usually has not been considered in DG tasks, since the prior works assume an object has different labels in different domains. Some recent works (Zhang et al. 2013; Gong et al. 2016) 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.

## 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 $\mathcal{Y} = \{1, 2, \ldots, K\}$ is the set of class label. 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, $p(f(x)|y)$, is also aligned. As stated in the previous section, the label shift is widely observed in DG (across different source domains and between source and target domains), which makes the covariant alignment not be able to replace the conditional alignment.

With the conditional and label shift assumption, the alignment is more challenging than the covariant shift which only requires to align the marginal distribution $p(f(x))$. Since our objective is to apply the same classifier to different domains. With the Bayes' rule, we can factorize the posterior of $f(x)$ as $p(y|f(x)) = \frac{p(f(x)|y)p(y)}{p(f(x))}$. Therefore, to build a domain-invariant classifier $p(y|f(x))$, we need to align $p(f(x)|y)$, $p(y)$, and $p(f(x))$ simultaneously. We note that the goal of conventional IFL is to align $p(f(x))$ and $p(y)$ for all source domains, which 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).

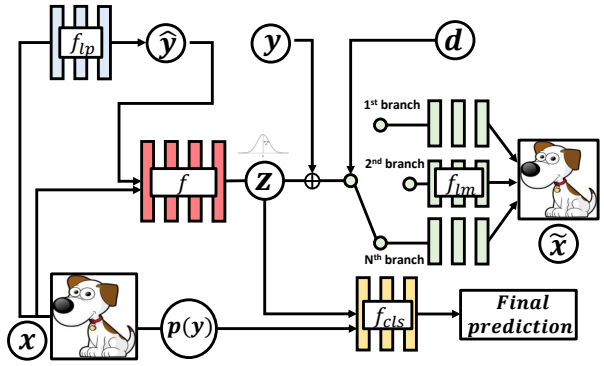



Figure 3: Illustration of our framework, which consists of conditional feature encoder $f$, domain-wise likelihood maximization network $f_{lm}$, label prior network $f_{lp}$, and classifier $f_{cls}$. Our latent representation $z$ is aligned with the Gaussian prior. $f$, $f_{lp}$, $f_{cls}$ are used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

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 $p(y)$.

### 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 scalable 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 (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.

Following the variational bound (Kingma and Welling 2013), minimizing the KL divergence between $p(z|x, y)$ and

[1]We use $f(x)$ and $z$ interlaced to align with the conventional IFL and variational autoencoder literature, respectively.

$q(z|x,y)$ is equivalent to maximizing the evidence lower bound (ELBO) (Domke and Sheldon 2018) of the likelihood $\log p(x,y,z)$, denoted by $\mathcal{L}$:

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

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

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

which can be reformulated as

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

where $\mathbb{E}$ 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, $p^i(z|y), i \in \{1,2,\ldots,N\}$ is the prior distribution in the variational model, e.g., 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., $p^1(z|x,y) = p^2(z|x,y) = \cdots = p^N(z|x,y)$.

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 Gaussian distribution, it is also natural to configure $q(z|x,y)$ as Gaussian distribution. 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 $\epsilon \in N(0,I)$, and

$$L_1 = \sum_{i=1}^{N}\sum_{j=1}^{M_i}[-\log\frac{\sigma_j^i}{\bar{\sigma}} + \frac{\bar{\sigma}^2 + (\bar{\mu}-\mu_j^i)^2}{2(\sigma_j^i)^2} - \frac{1}{2}], \tag{4}$$

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 Gaussian distribution $N(0,1)$, 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

**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
  $\hat{y} \leftarrow f_{lp}(x) z \leftarrow f(x,\hat{y}) \tilde{x} \leftarrow f_{lm}(z,\hat{y}) y \leftarrow f_{cls}(z,y)$
  //Calculate the loss functions: $L_1$, $L_2$, $L_{CE1}$, $L_{CE2}$
  //Update parameters according to gradients
  $f \leftarrow L_1 + \alpha L_2 + \beta L_{CE2} + \gamma L_{MMD} + \theta L_{\hat{y}}$;
  $f_{lp} \leftarrow L_{CE1}$; $f_{lm} \leftarrow L_2$; $f_{cls} \leftarrow L_{CE2}$
**until** deadline

can be formulated as

$$L_2 = \sum_{i=1}^{N}\sum_{j=1}^{M_i}||f_{lm}^i(z_j^i, y_j^i) - x_j^i||^2, \tag{5}$$

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.

## Label-prior

Inferencing the latent representation $z$ requires to know the label information $y$ in advance, since we are modeling the 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 the input of 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 $\hat{y}$.

## Marginal alignment for final prediction

To align $p(f(x))$, we follow the conventional IFL to minimize the multi-domain MMD (Li et al. 2018a), denoted by $L_{MMD}$. To avoid the condition of $y$, we simply sample the example to have the same class label $y$ from different domains in a batch, and only calculate $L_{MMD}$ among them. We note that MMD is shared for all classes.

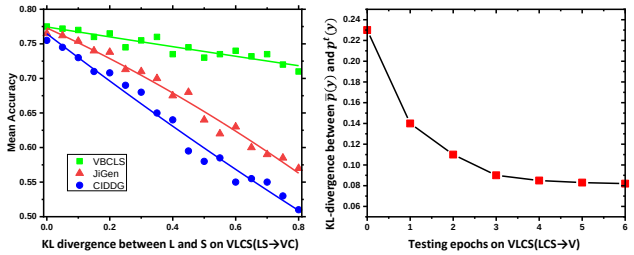



Figure 4: Left: Quantification of DG performance on VLCS (LS→VC) under different source label shifts ($p^L(y) \neq p^S(y)$ by sampling). Right: $p^t(y)$ refinement on LCS→V.

Finally, we propose to align $p(y)$ to obtain the 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 $\bar{p}(y|f(x)) = \frac{\bar{p}(f(x)|y)\bar{p}(y)}{\bar{p}(f(x))}$, 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 and conventional IFL, respectively, 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 $i$ can be

$$p^i(y=k|f(x)) = \frac{p^i(y=k)}{\bar{p}(y=k)}\bar{p}(y=k|f(x)), \qquad (6)$$

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 $L_1+\alpha L_2+\beta L_{CE2}+\gamma L_{MMD}+\theta L_{\hat{y}}$, update $f_{lp}$ with $L_{CE1}$, update $f_{lm}$ with $L_2$, and update $f_{cls}$ with $L_{CE2}$, respectively.

### Label distribution self-refinement in testing

In testing, we make the final prediction with $p^t(y=k|f(x)) = \frac{p^t(y=k)}{\bar{p}(y=k)}\bar{p}(y=k|f(x))$. Since $p^t(y)$ in the target domain is not available, a possible way to tackle this is to compute label proportion from the aggregated multi-source domains as prior (Zhang, David, and Gong 2017) by setting $p^t(y=k) = \bar{p}(y=k)$, i.e., using the classifier's prediction directly. However, $p(y)$ can be different across different target domains and therefore $\bar{p}(y)$ can be biased because of the differences in data collection. The conventional DG models are fixed after training and cannot be refined by the implementation experience.

The use of posterior label alignment offers the flexibility to refine the target label matching. Specifically, we initialize $p^t(y=k) = \bar{p}(y=k)$, and then update $p^t(y=k)$ by accounting for the network inference output, which is regarded as pseudo-label, within $t$ iterations. The underlying assumption is that the label shift of different period ($t$ iterations) in the target domain can be smaller than the label shift between the source and target domains.

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

| 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 |
| VBCLS+3T | 72.14 | 69.44 | 96.88 | 70.92 | 77.17±0.06 |

## Experiments

In this section, we demonstrate the effectiveness of our variational Bayesian inference framework under conditional and label shift (VBCLS) on the classic VLCS DG benchmark for image classification, Office+Caltech dataset, and the PACS benchmark for object recognition with domain shift.

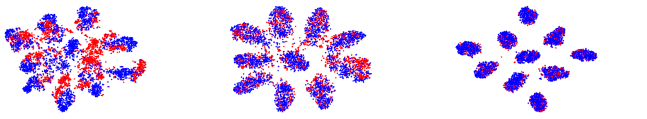
The domain invariant encoder $f$ and posterior alignment classifier $f_{pa}$ use the encoder and classifier structure as our compared models (e.g., AlexNet and ResNet18), and the label prior network is simply the concatenation of $f$ and $f_{pa}$, and the likelihood maximization network $f_{lm}$ uses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

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 $1e-3$, 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 it is seen that 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 $D_{KL}(q(z|x,y)||q(z|x,\hat{y}))$. VBCLS+3T indicates refining the target label distribution for 3 epochs.

### 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



**DANN** (Covariant Shift) | **CIDDG** (Conditional Shift) | **VBCLS** (Conditional and Label Shift)

Figure 5: The t-SNE visualization on Office+Caltech (A,W,D $\rightarrow$ C). The blue and red dots are the sampled source and target domain examples, respectively.

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. 2018c) is significant, which demonstrates the necessity of incorporating both conditional and label shifts. 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 a better performance in almost all cases. We note that JiGen (Carlucci et al. 2019) uses the Jigsaw Puzzles solving methods which are essentially different from the IFL DG.

The good performance of our strategy indicates the invariant feature learning works well and the conditional and shift assumption can be an inherent property which needs to be addressed for real-world DG challenges. The discrepancy between 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 a label shift. Our testing label distribution refinement can be a practical solution to utilize the testing data in the DG setting in which we do not need the source data to fine-tune the network. 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. It is seen that, in some cases, excluding a particular source from the training substantially degrades the target loss. However, it is seen that our VBCLS can still be more robust under these cases.

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

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

Table 3: Classification accuracy (mean$\pm$sd) on the Office + Caltech 256 Dataset (Gong et al. 2012).

| Target | A | C | A, C | W, D | W, C | D, C |
|---|---|---|---|---|---|---|
| DICA13 | 92.59 | 83.17 | 63.67 | 83.85 | 87.59 | 86.25 |
| SCA16 | 91.96 | 83.35 | 73.04 | 83.85 | 87.31 | 86.25 |
| CCSA17 | 90.98 | 83.37 | 77.56 | 80.04 | 85.80 | 84.91 |
| CIDG18 | 92.38 | 81.39 | 69.87 | 82.74 | 87.45 | 85.63 |
| MDA19 | 93.47 | 86.89 | 82.56 | 84.89 | 88.91 | 88.23 |
| VBCLS | 95.83±0.05 | 90.25±0.05 | 84.69±0.05 | 86.58±0.05 | 92.31±0.05 | 91.34±0.05 |
| VBCLS-$f_{pa}$ | 92.87±0.05 | 86.35±0.03 | 69.46±0.04 | 82.65±0.06 | 88.83±0.05 | 87.45±0.04 |
| VBCLS-$L_{\hat{y}}$ | 95.65±0.04 | 89.56±0.06 | 84.48±0.03 | 86.26±0.05 | 92.09±0.03 | 91.09±0.04 |
| VBCLS+3T | 95.64±0.05 | 90.39±0.05 | 84.55±0.03 | 86.33±0.06 | 93.09±0.04 | 91.92±0.05 |

## Office+Caltech dataset

The performance of different DG methods is also evaluated using Office+Caltech256 dataset (Gong et al. 2012), which is comprised of images from four different datasets including Amazon (A), Webcam (W), DSLR (D), and Caltech-256 (C). Because 10 shared classes are available in these datasets, images of these classes are selected and those from the same original dataset form one domain in Office+Caltech. Note that the domains within Office+Caltech correspond to the biases of different data collection procedures. For a fair comparison, the 4,096-dimensional DeCAF6 features (Donahue et al. 2014) are used, so that the input sample formats are consistent across all domains.

The accuracy is tabulated in Table 3. Note that the other four cases, such as A, D, C $\rightarrow$ W, are not reported, since the SVM baseline could achieve higher accuracy over 90%. It is observed that the improvement over the conditional shift only CIDG (Li et al. 2018b) is consistent on the VLCS task. Our VBCLS achieves better performance than previous SOTA MDA (Hu et al. 2019) and the previous DICA (Muandet et al. 2013), CCSA (Motiian et al. 2017) and SCA (Ghifary et al. 2017) by a large margin.

## 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). As shown in the results (c.f. Table 4), the Sketch domain produces the lowest accuracy when used as the target domain and therefore it is deemed the most challenging one. In light of this, we follow the previous work

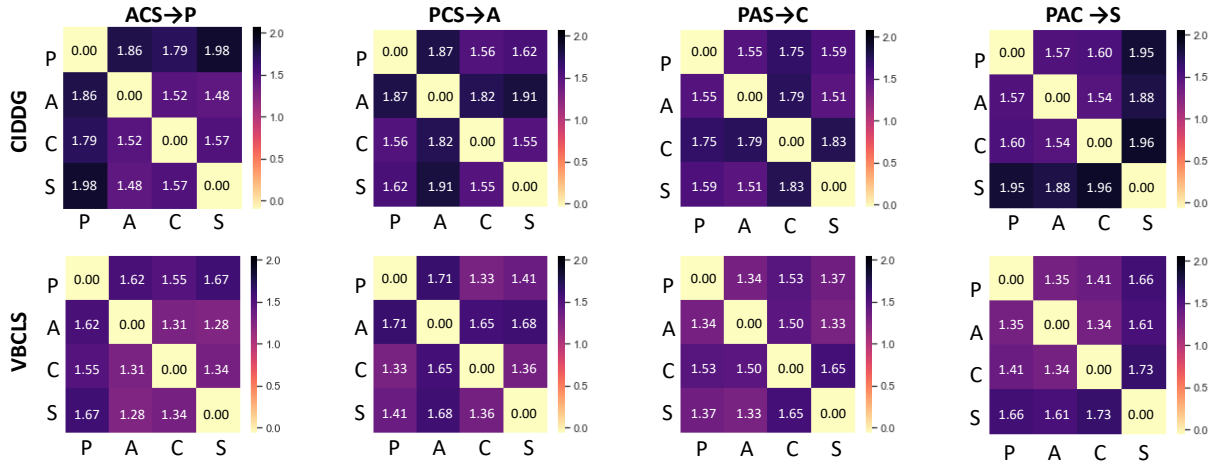



Figure 6: Cross-domain proxy $\mathcal{A}$-distance for models trained on the PACS benchmark using the leave-one-domain-out validation setting. The smaller proxy $\mathcal{A}$-distance, the better of source domains alignment (corresponding to lighter color). The proxy $\mathcal{A}$-distance of VBCLS is about 10% lower than CIDDG. Best viewed in color.

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

| 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-$L_{\tilde{y}}$ | 91.77 | 70.54 | 76.56 | 70.33 | 77.19±0.06 |
| VBCLS+3T | 92.98 | 70.68 | 77.41 | 71.02 | 77.92±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 |

to tune the model using 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, it is seen 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, it is observed that our method outperforms CIDDG (Li et al. 2018c), 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 ResNet18 backbone.

To further examine whether cross-domain divergences are decreased by the proposed VBCLS, and more importantly, in accordance to Proposition 1, we also estimate the pairwise proxy $\mathcal{A}$-distance (Ben-David et al. 2007). In addition, we investigate whether the divergence between the unseen target domain and the source domains reduces. Toward this goal, the divergence is approximated by computing the proxy $\mathcal{A}$-distance for all pairs of domains for the models using both CIDDG and our method. Specifically, the proxy $\mathcal{A}$-distance is computed by carrying out domain discrimination for each domain pair on the PACS benchmark by incorporating the representation space learned by the feature extractor from each trained model. The discriminators are implemented via tree ensemble classifiers with 100 estimators. The average accuracy is reported using 5-fold cross-validation independently run for each domain pair as depicted in Figure 6, where each domain is comprised of a random sample of size 500. Each entry of the matrix corresponds to the proxy $\mathcal{A}$-distance between a pair of domains in the row and column. Note that the diagonals always have the zero value resulting from the divergence of the same domains. It is shown from the experimental results that our method is effective in promoting invariance across pairs of source distributions compared with CIDDG. Note that CIDDG preserves more domain information, since it is able to distinguish data domains with higher accuracy.

## Conclusion

In this paper, we target to establish a more realistic assumption in DG that the conditional and label shifts can be independent and contemporary exist. 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 shift. Our system can also be flexible for adaptive target label refinement to utilize testing data in the DG setting. Extensive evaluations verify our analysis and revoke the IFL learning which can still be effective as the recent fast developed alternative methods.

## References


[Ben-David et al. 2007] Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2007. Analysis of representations for domain adaptation. In *NIPS*. 7

[Carlucci et al. 2019] Carlucci, F. M.; D'Innocente, A.; Bucci, S.; Caputo, B.; and Tommasi, T. 2019. Domain generalization by solving jigsaw puzzles. In *CVPR*. 6

[Ding and Fu 2017] Ding, Z., and Fu, Y. 2017. Deep domain generalization with structured low-rank constraint. *TIP*. 2

[Domke and Sheldon 2018] Domke, J., and Sheldon, D. R. 2018. Importance weighting and variational inference. In *NIPS*. 4

[Donahue et al. 2014] Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*. 5, 6

[Du et al. 2020] Du, Y.; Xu, J.; Xiong, H.; Qiu, Q.; Zhen, X.; Snoek, C. G.; and Shao, L. 2020. Learning to learn with variational information bottleneck for domain generalization. *ECCV*. 6, 7

[Ghifary et al. 2017] Ghifary, M.; Balduzzi, D.; Kleijn, W. B.; and Zhang, M. 2017. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE T-PAMI*. 1, 2, 5, 6

[Gong et al. 2012] Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*. 6

[Gong et al. 2016] Gong, M.; Zhang, K.; Liu, T.; Tao, D.; Glymour, C.; and Schölkopf, B. 2016. Domain adaptation with conditional transferable components. In *ICML*. 1, 2, 3

[Gong et al. 2018] Gong, M.; Zhang, K.; Huang, B.; Glymour, C.; Tao, D.; and Batmanghelich, K. 2018. Causal generative domain adaptation networks. *arXiv*. 3

[Goodfellow et al. 2016] Goodfellow, I.; Bengio, Y.; Courville, A.; and Bengio, Y. 2016. *Deep learning*. MIT. 1

[Hu et al. 2019] Hu, S.; Zhang, K.; Chen, Z.; and Chan, L. 2019. Domain generalization via multidomain discriminant analysis. In *UAI*. 1, 6

[Huang et al. 2020] Huang, Z.; Wang, H.; Xing, E. P.; and Huang, D. 2020. Self-challenging improves cross-domain generalization. *ECCV*. 6, 7

[Khosla et al. 2012] Khosla, A.; Zhou, T.; Malisiewicz, T.; Efros, A. A.; and Torralba, A. 2012. Undoing the damage of dataset bias. In *ECCV*. Springer. 2

[Kingma and Welling 2013] Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv*. 3, 4

[Kouw 2018] Kouw, W. M. 2018. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*. 2, 3

[Li et al. 2017] Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *ICCV*. 6

[Li et al. 2018a] Li, H.; Jialin Pan, S.; Wang, S.; and Kot, A. C. 2018a. Domain generalization with adversarial feature learning. In *CVPR*. 2, 3, 4, 6

[Li et al. 2018b] Li, Y.; Gong, M.; Tian, X.; Liu, T.; and Tao, D. 2018b. Domain generalization via conditional invariant representations. In *AAAI*. 2, 6

[Li et al. 2018c] Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; and Zhang, K. 2018c. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*. 2, 3, 6, 7

[Lian et al. 2019] Lian, R.; Xie, M.; Wang, F.; Peng, J.; and Wu, H. 2019. Learning to select knowledge for response generation in dialog systems. *IJCAI*. 4

[Mancini et al. 2018] Mancini, M.; Bulo, S. R.; Caputo, B.; and Ricci, E. 2018. Best sources forward: domain generalization through source-specific nets. In *ICIP*. 2

[Matsuura and Harada 2020a] Matsuura, T., and Harada, T. 2020a. Domain generalization using a mixture of multiple latent domains. *AAAI*. 1, 2

[Matsuura and Harada 2020b] Matsuura, T., and Harada, T. 2020b. Domain generalization using a mixture of multiple latent domains. In *AAAI*. 7

[Moreno-Torres et al. 2012] Moreno-Torres, J. G.; Raeder, T.; Alaiz-RodríGuez, R.; Chawla, N. V.; and Herrera, F. 2012. A unifying view on dataset shift in classification. *Pattern Recognition* 45(1):521–530. 2, 3

[Motiian et al. 2017] Motiian, S.; Piccirilli, M.; Adjeroh, D. A.; and Doretto, G. 2017. Unified deep supervised domain adaptation and generalization. In *ICCV*. 2, 6

[Muandet et al. 2013] Muandet, K.; Balduzzi, D.; Schölkopf; and Bernhard. 2013. Domain generalization via invariant feature representation. In *ICML*. 1, 2, 6

[Schölkopf et al. 2012] Schölkopf, B.; Janzing, D.; Peters, J.; Sgouritsa, E.; Zhang, K.; and Mooij, J. 2012. On causal and anticausal learning. *arXiv*. 3

[Shankar et al. 2018] Shankar, S.; Piratla, V.; Chakrabarti, S.; Chaudhuri, S.; and Sarawagi, S. 2018. Generalizing across domains via cross-gradient training. *arXiv*. 7

[Wang et al. 2019] Wang, H.; He, Z.; Lipton, Z. C.; and Xing, E. P. 2019. Learning robust representations by projecting superficial statistics out. *ICLR*. 2

[Wang et al. 2020] Wang, S.; Yu, L.; Li, C.; Fu, C.-W.; and Heng, P.-A. 2020. Learning from extrinsic and intrinsic supervisions for domain generalization. *ECCV*. 6, 7

[Zadrozny 2004] Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. In *ICML*. 6

[Zhang et al. 2013] Zhang, K.; Schölkopf, B.; Muandet, K.; and Wang, Z. 2013. Domain adaptation under target and conditional shift. In *ICML*. 1, 2, 3

[Zhang, David, and Gong 2017] Zhang, Y.; David, P.; and Gong, B. 2017. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*. 5

[Zhang, Gong, and Schölkopf 2015] Zhang, K.; Gong, M.; and Schölkopf, B. 2015. Multi-source domain adaptation: A causal view. In *AAAI*. 3

[Zou et al. 2019] Zou, Y.; Yu, Z.; Liu, X.; and Kumar, B. 2019. Confidence regularized self-training. *ICCV*. 1