Mutual Variational Inference: An Indirect Variational Inference Approach for Unsupervised Domain Adaptation

Jiahong Chen, *Member, IEEE*, Jing Wang, and Clarence W. de Silva, *Life Fellow, IEEE*

***Abstract*—In this article, the unsupervised domain adaptation problem, where an approximate inference model is to be learned from a labeled dataset and expected to generalize well on an unlabeled dataset, is considered. Unlike the existing work, we explicitly unveil the importance of the latent variables produced by the feature extractor, that is, encoder, where contains the most representative information about their input samples, for the knowledge transfer. We argue that an estimator of the representation of the two datasets can be used as an agent for knowledge transfer. To be specific, a novel variational inference approach is proposed to approximate a latent distribution from the unlabeled dataset that can be used to accurately predict its input samples. It is demonstrated that the discriminative knowledge of the latent distribution that is learned from the labeled dataset can be progressively transferred to that is learned from the unlabeled dataset by simultaneously optimizing the estimator via the variational inference and our proposed regularization for shifting the mean of the estimator. The experiments on several benchmark datasets demonstrate that the proposed method consistently outperforms state-of-the-art methods for both object classification and digit classification.**

***Index Terms*—Computer vision, deep learning, domain adaptation.**

I. Introduction

ONE OF the core problems of supervised learning is that its performance highly relies on a large amount of labeled data. Previous studies have demonstrated that the performance of an image classifier will drop significantly when the input data distributions vary due to some factors, for example, the different angles of the camera, the different noise conditions, the different background styles, etc. [1]–[4]. Therefore, there is a strong demand to design a learner that can produce domain-invariant representations, which allows the data distributions from the different but related data domains to have heterogeneous features with different dimensionalities [5]. More specifically, such a learner can transfer the knowledge learned from a labeled dataset (source domain) to an unlabeled dataset (target domain). Due to the lack of label information for the target input samples, the latent distribution, which is learned from the target domain, that can be used for predicting the categories of the target input samples is difficult to compute. Unsupervised domain adaptation (UDA) aims at solving the label-missing problem by mitigating the domain shift and ensuring that the learned classifier can generalize well to the target domain without using its labels. In this work, we utilize variational inference to efficiently approximate the latent representations of the target (label-missing) domain, which can be used to accurately predict the missing labels.

The existing UDA approaches attempt to mitigate the domain shift by regularizing the feature extractor to extract the features from both domains to construct a domain-invariant feature (latent) space. In their settings, a decision rule, which can be applied to the target domain, is able to be learned from the domain-invariant latent space with the support of the labels of the source input samples. Deep UDA approaches further improve the knowledge transferability and the model generalization by leveraging the better feature-extraction ability of deep neural networks [6]–[12]. Most of these methods quantify the domain shift between the source domain and the target domain using the intermediate features that are induced by classifiers, namely, the classifier-induced discrepancy, under the guidance of the convergence learning bound [13]. The main idea of these approaches is to minimize the classification error of the source domain and the classifier-induced discrepancy to capture both discriminative and domain-invariant representations. Nevertheless, the work based on the convergence learning bound does not explicitly consider the way to effectively infer a good latent distribution from the target domain for its classification task.

Deep generative models are widely applied to transform a simple distribution with some mapping functions into a more complicated one. An idea to achieve such transformation is to use an encoder, which can also be regarded as a feature extractor, to infer the latent variables that can be used to predict and construct the data distribution. Variational autoencoder (VAE) [14] utilizes a Kullback–Leibler (KL)-divergence penalty to impose a Gaussian distribution on the latent variables of the autoencoder such that the learning of the encoder's

Manuscript received 25 August 2020; revised 7 December 2020, 23 February 2021, and 13 June 2021; accepted 20 August 2021. Date of publication 3 September 2021; date of current version 17 October 2022. This work was supported by MITACS under Grant 11R11370. This article was recommended by Associate Editor H. Qiao. *(Jiahong Chen and Jing Wang contributed equally to this work.) (Corresponding author: Jiahong Chen.)*

Jiahong Chen and Clarence W. de Silva are with the Department of Mechanical Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: jhchen@mech.ubc.ca; desilva@mech.ubc.ca).

Jing Wang is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: jing@ece.ubc.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2021.3107292.

Digital Object Identifier 10.1109/TCYB.2021.3107292

parameters can be guided to infer a reasonable latent space whose sampled vectors can be decoded back to the input distribution. A similar work, called the importance weighted autoencoder [15], is also proposed to learn richer latent space representations using a strictly tighter log-likelihood lower bound derived from importance weighting. Their works indicate that the distribution transformation can be easily done by optimizing the latent space to cover the sufficient representations. Like generative modeling, latent space inference is also important for transferring knowledge across two data domains as the latent space is typically shared by the two domains. As the knowledge transfer itself is to regularize the feature extractor to extract the domain-invariant features so that the target input samples can make the best use of the discriminative features source features for the specific tasks in the target domain, we can argue that inferring a good latent space shared by the two data domains is of necessity for UDA.

In this article, we propose mutual variational inference (MVI), which utilizes the concepts of variational inference and mutual information (MI), to infer an estimator that can be used as an agent for knowledge transfer. We demonstrate that optimizing a variational approximation, which contains the information from the two domains, can maximize the variational lower bound for the knowledge transfer. We further propose to maximize a lower bound of the MI between the variational approximation and the target input samples to retain more information about the target domain in the variational approximation. The process of maximizing the lower bound of the MI shifts the mean of the variational approximation from the source to the target, which produces a more generalized model that benefits the knowledge transfer. Experimental results on several benchmark datasets indicate that MVI achieves state-of-the-art performance on different UDA tasks.

II. Related Work

A. Domain Adaptation

The cross-domain knowledge transfer requires the model to learn the discriminative features from a labeled (source) dataset and apply them to an unlabeled (target) dataset. However, the variations between the source data distributions and the target data distributions, that is, the domain shifts, could significantly degrade the generalization of the model trained on the labeled dataset and lead to poor performance on predicting those unlabeled data samples. To solve this issue, the UDA methods are proposed to mitigate the effect of the data variations on the model generalization by transferring the knowledge learned from the source domain to the target domain [5], [16]. In general, UDA mitigates domain shifts by extracting domain-invariant features from the two domains.

Adversarial UDAs are motivated by generative adversarial nets (GANs) [17], which is a cornerstone work in the generative model. The domain-adversarial neural network (DANN) is proposed to train the feature extractor to confuse the discriminative model, that is, the domain discriminator, into believing that the features it extracts come from the source domain [4]. Further introducing the adversarial learning strategies makes the training of the discriminative model more difficult to converge, which results in the extracted features being more domain invariant [18]. Joint distribution learning with GAN (JAN) further improves the existing adversarial UDA methods by learning the joint distribution of the marginals and the conditionals from multiple data domains [19]. Conditional adversarial domain adaptation (CDAN) optimizes the cross-covariance between the latent representations and their predictions, which improves the model discriminability and transferability of the previous adversarial UDA methods [20]. Lately, transferable adversarial training (TAT) is proposed to improve the adaptability of the source domain to improve the model transferability [21]. TAT could generate transferable features in the middle of the source domain and the target domain to avoid the distortion of the original feature distributions and the deterioration of adaptability. Besides adversarial UDAs, researchers also propose to measure the domain divergence by some statistical distances and minimize such distances to encourage the feature extractor to extract domain-invariant features [22], [23]. Instead of utilizing adversarial learning, correlation alignment evaluates the domain divergence by the difference of the mean and covariance between the source dataset and the target dataset [24], [25]. Methods based on maximum mean discrepancy (MMD), which measures the difference between the means of the two feature distributions, are also proposed [26], [27]. The deep adaptation network (DAN) is proposed to reduce the data bias and enhance the feature transferability by utilizing the task-specific layers [8]. Moreover, stepwise adaptive feature norm (SAFN) unveils that the domain divergence largely relies on some small-norm regions that are induced by the image classifier, so that the knowledge can be progressively transferred by putting features away from these regions [12]. Research also refines the architectures of DNN to improve the model transferability. Wang *et al.* [28] replaced the traditional normalization techniques (e.g., batch normalization) with transferable normalization (TransNorm) to investigate the domain-specific features to improve the transferability. TransNorm can be easily plugged into the existing domain adaption approaches and improve the classification accuracy without introducing any extra parameters.

B. Autoencoders and Mutual Information

Deep learning-based generative models have succeeded in modeling different data distributions. The denoising autoencoder (DAE) is proposed to extract robust features from the input samples and, therefore, exclude any image noise when reconstructing these data samples [29]. It is shown that minimizing the reconstruction error using an autoencoder can maximize the lower bound of the MI between the input space and its latent space, which can encourage the encoder to produce features that are robust to the variations of the input distribution [30]. Similarly, the VAE is proposed to retain the maximum amount of input information in their latent representations [14]. The autoencoder-based approaches are also applied in domain adaptation for better feature representation.

The deep reconstruction-classification network (DRCN) introduces an autoencoder-based architecture to reconstruct the target input samples. DRCN maximizes the MI between the target input space and its latent space to retain the maximum amount of the target information while transferring the knowledge from the source domain [31].

C. Variational Inference

Variational inference is widely used to approximate a conditional density of the latent variable in the Bayesian statistics and becomes increasingly important to computer vision. Deep learning-based variational inference methods can efficiently infer a latent distribution for a set of input samples through optimization [14]. Although variational inference underestimates the variance of its approximation, it can explore the model complexity efficiently. In this study, the task that we are focusing on is to transfer the knowledge learned by the source latent distribution to the target latent distribution instead of using the variational approximation directly. Thus, the underestimation of the variance would have less impact in preventing our proposed work from achieving its objective of knowledge transfer.

III. Preliminaries and Problem Formulation

A. Preliminaries

1) Variational Inference: Considering the computation of the conditional probability distribution of a latent variable $\mathbf{z}$ given its observation $\mathbf{x}$

$$p(\mathbf{z}|\mathbf{x}) = p(\mathbf{z})\frac{p(\mathbf{x}|\mathbf{z})}{p(\mathbf{x})}. \tag{1}$$

In the Bayesian models, the latent variable is drawn from a prior distribution $p(\mathbf{z})$ and the likelihood function is modeled as $([p(\mathbf{x}|\mathbf{z})]/[p(\mathbf{x})])$. The inference of a Bayesian model can be viewed as a process to compute the posterior $p(\mathbf{z}|\mathbf{x})$, which is shown in (1). When the Bayesian model becomes complex, the posterior can only be estimated by the inference approximation.

Traditional approaches to approximate the inference are based on the Markov Chain Monte Carlo (MCMC) whose sampling process is not efficient when the dataset is large or the model itself is complex [32]. In response to this, the variational inference approaches are developed to approximate the Bayesian inference efficiently by optimization instead of sampling. With the rapid improvement of computing power, VAE further extends the idea of variational inference to the deep learning model, which can effectively optimize the latent vectors to reproduce the input distributions.

Like VAE, we focus on the optimization of the latent space of a deep neural-network model via variational inference in this article. Defining a prior distribution $q(\mathbf{z})$ on a set of distribution densities $\mathcal{Q}$. Under our settings, the goal of variational inference is to find a member $q^*(\mathbf{z})$ from the set $\mathcal{Q}$ that can minimize the KL divergence between the prior distribution $q(\mathbf{z})$ and the latent distribution $p(\mathbf{z}|\mathbf{x})$

$$q^*(\mathbf{z}) = \arg\min_{q\in\mathcal{Q}} D_{\mathrm{KL}}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) \tag{2}$$

where the latent distribution $p(\mathbf{z}|\mathbf{x})$ is modeled by a deep neural-network-based encoding function.

In this manner, the latent distribution $p(\mathbf{z}|\mathbf{x})$ can be estimated by optimizing the model to find $q^*(\mathbf{z})$ within the set of distribution densities $\mathcal{Q}$.

2) Mutual Information: MI is widely applied to quantify the statistical dependencies between two random variables. In comparison to correlation, the dependencies that MI captures are nonlinear and are regarded as the true dependence. MI is a fundamental quantity that measures the amount of information that one random variable is shared with other random variables, and can be defined based on the Shannon entropy as

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned} \tag{3}$$

where $I(\cdot;\cdot)$ denotes the MI between two random variables, $H(\cdot)$ denotes the marginal Shannon entropy, and $H(\cdot|\cdot)$ denotes the conditional Shannon entropy.

The marginal Shannon entropy of a random variable is expressed as

$$H(X) = -\sum_{x\in X} P(x)\log P(x) \tag{4}$$

and the conditional Shannon entropy $H(X|Y)$ of the random variable X given the random variable Y can be expressed as

$$H(X|Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(y)}. \tag{5}$$

B. Problem Formulation

Under the settings of an UDA problem, we are given a source domain $D_S = \{(\mathbf{x}_s^{(i)}, \mathbf{y}_s^{(i)})\}_{i=1}^n$ consisting of n labeled images from the source input space $\{X_S, Y_S\}$, where $\mathbf{x}_s^{(i)} \in X_S$ represents an image sampled from the source input space and $\mathbf{y}_s^{(i)} \in Y_S$ is the label of this image. Likewise, a target domain $D_T = \{(\mathbf{x}_t^{(j)})\}_{j=1}^m$ is provided with m unlabeled images that are sampled from the target input space $\{X_T, Y_T\}$. Note that the probability distribution densities of the images across the two data domains, which are labeled as the same category, are significantly different. The statistical measurement of these density variations between the two domains is called the domain gap or the domain divergence, which should be minimized during the process of knowledge transfer.

To be specific, the formulation of a UDA problem is given as follows. The model should be trained to generalize well in the target domain without access to the labels of the target input samples. Due to the lack of correspondence between the target input samples and their labels, the domain gap should be statistically quantified and minimized during the training so that the target input samples can utilize the knowledge learned from the correspondence of the source domain for their own classification tasks. In other words, a regularization has to be explicitly designed to transfer the discriminative knowledge learned from the labeled source domain to the target domain. So, even without access to the target labels during the training,

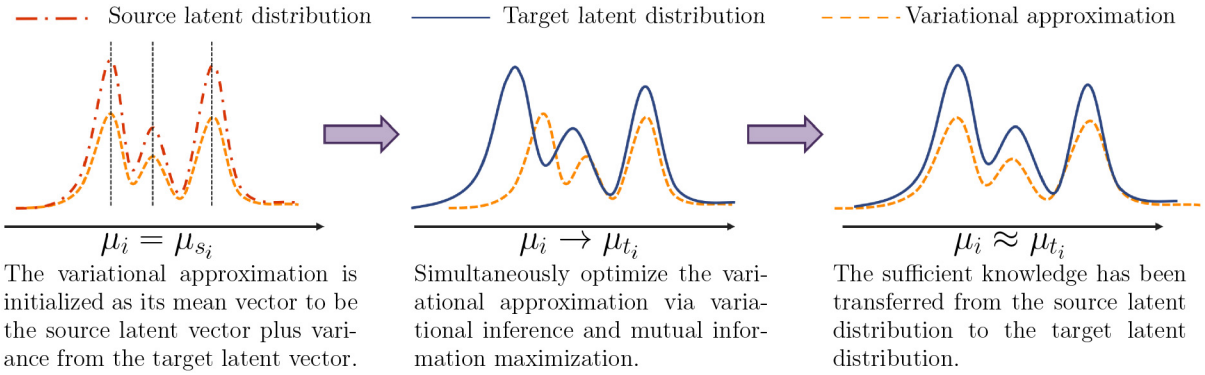


Fig. 1. (Best viewed in color.) Overall idea of the knowledge transfer via shifting the mean of the variational approximation. The variational approximation is initialized as its mean vectors to be sampled from the source latent distribution plus some variances sampled from the target latent distribution, that is, $\mu_i = \mu_{s_i}, \mu_{s_i} = \mathbf{z_s}, {\sigma_i}^2 = \mathbf{z_t}$. The KL-divergence between the variational approximation and the target latent distribution and the MI between the variational approximation and the target input space are optimized simultaneously to encourage the mean of the variational approximation to be shifted from the source domain to the target domain. Moreover, we argue that the knowledge learned by the source latent distribution can be sufficiently transferred to the target latent distribution when the variational approximation can retain the maximum information about the entire input sampling space. Note that the variational approximation serves as an agent for knowledge transfer in our proposed framework, and will not be used for the classification tasks.

the objective function of the target-domain classification can still be maximized during the phase of the testing

$$\max \delta\left(\hat{\mathbf{y}}_t, \mathbf{y}_t\right) \tag{6}$$

where $\hat{\mathbf{y}}_t$ denotes the predictions of the input samples from the target domain and $\delta(\cdot)$ represents the binary indicator that outputs 1 if the prediction $\hat{\mathbf{y}}_t$ matches its label $\mathbf{y}_t$.

IV. Methodology

The key to solve the UDA problem is to obtain a good latent (feature) space for both the source domain and the target domain. However, the existing methods that merely focus on minimizing a narrowly defined classification error may not effectively guide the learning of the transferable features. The proposed MVI aims at optimizing an estimator that contains the information about the two data domains, to guide the process of knowledge transfer.

A. Overall Idea

The objective of UDA is to obtain a feature extractor G so that an image classifier F can utilize its outputs, that is, latent vectors, to accurately predict the unlabeled target samples. In the absence of the target labels, the feature extraction of the target domain is not guided, which makes the target features not discriminative and meaningless to the target-domain specific classification tasks. Therefore, it is desirable to have an alignment mechanism that can effectively reduce the divergence between the source latent space and the target latent space and extract the target features under the guidance of the discriminative source features. To achieve this alignment, a multivariate Gaussian distribution is utilized as an agent to align the source latent distribution and the target latent distribution. The multivariate Gaussian is initialized as a variational approximation with its mean vectors are sampled from the source latent distribution and its log-variance vectors are sampled from the target latent distribution, like VAE [14]. By reshaping the latent distribution of MVI to a multivariate Gaussian, the features that are most important to the classification tasks (lead to the highest classification score) will be gathered to the mean of the multivariate Gaussian. Then, by optimizing the model using VAE and shifting the mean around the multivariate Gaussian from the source latent distribution to the target latent distribution, the features that are extracted from the target input samples will be guided by the discriminative source features gathering around the mean. As a result, the target features will gradually move to the mean of the multivariate Gaussian; and those target features that are important to the target-domain-specific classification tasks will be pushed to the mean and, therefore, be extracted by the feature extractor. To be specific, the mean shifting is achieved by maximizing the lower bound of the MI between the target sampling space and the space of the variational approximation. In this manner, the knowledge learned by the source latent distribution can be progressively transferred to the target latent distribution. The overall idea to transfer the knowledge through optimizing the variational approximation is shown in Fig. 1.

B. Knowledge Transfer by Optimization

As we have no access to the labels of the target input samples, it is hard to directly infer a good latent distribution for the classifier to make predictions on the target input samples. Recall that variational inference can estimate the conditional probability density distribution of the latent variables $\mathbf{z}$ given the observations $\mathbf{x}$, that is, latent distribution, through optimization [14], [32]. Unlike the existing work on variational inference, which is used to solve a single-domain problem, we utilize the variational approximation $q(\mathbf{z}^*)$ as an agent to indirectly transfer the knowledge from one latent distribution to another instead of using it to approximate these difficult-to-compute probability densities. Moreover, under the setting of UDA, the target latent distribution is supposed to be inferred from the entire sampling space not just from the target sampling space. Therefore, the optimization problem in our case becomes minimizing

$$D_{\mathrm{KL}}\left(q\left(\mathbf{z}^*\right)||p(\mathbf{z}_t|\mathbf{x})\right) = \mathbb{E}\left[\log q\left(\mathbf{z}^*\right)\right] - \mathbb{E}\left[\log p(\mathbf{z}_t|\mathbf{x})\right] \tag{7}$$

where expectations are taken with respect to the variational approximation $\mathbf{z}^*$ and the target latent distribution $\mathbf{z}_t$, where

$\mathbf{z}^* \sim q(\mathbf{z}^*) = N(\mu, \sigma^2)$ is applied with a reparameterization trick, that is, $\mathbf{z}^* = \mathbf{z}_s + e^{\frac{1}{2}\mathbf{z}_t} \odot \eta$ and $\eta \sim N(\mathbf{0}, \mathbf{I})$. Note that $\mathbf{z}_s$ indicates the source latent distribution, and $\mathbf{x}$ is sampled from the entire sampling space.

By further expanding the conditional probabilities in (7), we have

$$\begin{aligned} &D_{\mathrm{KL}}\big(q(\mathbf{z}^*)||p(\mathbf{z}_t|\mathbf{x})\big) \\ &\quad = \mathbb{E}\big[\log q(\mathbf{z}^*)\big] - \mathbb{E}\big[\log p(\mathbf{z}_t, \mathbf{x})\big] + \log p(\mathbf{x}) \end{aligned} \tag{8}$$

where $\log p(\mathbf{x})$ is a constant as $\mathbf{x}$ is not conditioned on either $\mathbf{z}^*$ or $\mathbf{z}_t$. Although we cannot directly optimize the KL-divergence penalty shown in (7) due to the lack of the labels of the target samples, we can differentiate and maximize a variational lower bound $\mathcal{B}(p)$ to minimize (7)

$$\mathcal{B}(p) = \mathbb{E}\big[\log p(\mathbf{z}_t, \mathbf{x})\big] - \mathbb{E}\big[\log q(\mathbf{z}^*)\big]. \tag{9}$$

Theorem 1: Given $\{\mathbf{x}|\mathbf{x} \in X\} = \{\mathbf{x}_s|\mathbf{x}_s \in X_S\} \cup \{\mathbf{x}_t|\mathbf{x}_t \in X_T\}$, the following bound holds:

$$\mathcal{B}(p) \geq \mathbb{E}\big[\log(p(\mathbf{x}_t|\mathbf{z}_t))\big] - D_{\mathrm{KL}}\big(q(\mathbf{z}^*)||p(\mathbf{z}_t)\big).$$

Proof: Equation (4) can be rewritten as the difference between the expected log likelihood of the reconstruction conditioned on the samples from the target latent distribution and the KL-divergence penalty between the target latent distribution $p(\mathbf{z}_t)$ and the variational approximation $q(\mathbf{z}^*)$

$$\begin{aligned} \mathcal{B}(p) &= \mathbb{E}\big[\log p(\mathbf{x}, \mathbf{z}_t)\big] - \mathbb{E}\big[\log q(\mathbf{z}^*)\big] \\ &= \mathbb{E}\big[\log(p(\mathbf{z}_t)p(\mathbf{x}|\mathbf{z}_t))\big] - \mathbb{E}\big[\log q(\mathbf{z}^*)\big] \\ &= \mathbb{E}\big[\log p(\mathbf{z}_t)\big] + \mathbb{E}\big[\log p(\mathbf{x}|\mathbf{z}_t)\big] - \mathbb{E}\big[\log q(\mathbf{z}^*)\big] \\ &= \mathbb{E}\big[\log p(\mathbf{x}|\mathbf{z}_t)\big] - D_{\mathrm{KL}}\big(q(\mathbf{z}^*)||p(\mathbf{z}_t)\big) \end{aligned} \tag{10}$$

where $-D_{\mathrm{KL}}(q(\mathbf{z}^*)||p(\mathbf{z}_t))$ is the negative KL-divergence penalty between the marginal density of the variational approximation $q(\mathbf{z}^*)$ and the marginal density of the target latent distribution $p(\mathbf{z}_t)$, which encourages the space of the variational approximation to be close to the latent space that is important to the task of the target-domain classification. Because of the lack of the correspondence between the samples from the target latent distribution and the samples from the entire input space, we further expand $\mathbb{E}[\log p(\mathbf{x}|\mathbf{z}_t)]$ from the entire input space to the source domain and the target domain, and derive a new variational lower bound that excludes the conditional density of $\mathbf{x}_s$ given $\mathbf{z}_t$. Specifically, given $\{\mathbf{x}|\mathbf{x} \in X\} = \{\mathbf{x}_s|\mathbf{x}_s \in X_S\} \cup \{\mathbf{x}_t|\mathbf{x}_t \in X_T\}$, we have

$$\begin{aligned} &\mathbb{E}\big[\log p(\mathbf{x} \in X|\mathbf{z}_t)\big] \\ &\quad = \mathbb{E}\big[\log p(\mathbf{x}_s \in X_S \cup \mathbf{x}_t \in X_T|\mathbf{z}_t)\big] \\ &\quad = \mathbb{E}\big[\log p(\mathbf{x}_s \cup \mathbf{x}_t|\mathbf{z}_t)\big] \end{aligned} \tag{11}$$

then, expanding the conditional probability with union evidence

$$\begin{aligned} &\mathbb{E}\big[\log p(\mathbf{x}_s \cup \mathbf{x}_t|\mathbf{z}_t)\big] \\ &= \mathbb{E}\left[\log \frac{p((\mathbf{x}_s \cup \mathbf{x}_t) \cap \mathbf{z}_t)}{p(\mathbf{z}_t)}\right] \\ &= \mathbb{E}\left[\log \frac{p(\mathbf{x}_s \cap \mathbf{z}_t) + p(\mathbf{x}_t \cap \mathbf{z}_t) - p(\mathbf{x}_s \cap \mathbf{x}_t \cap \mathbf{z}_t)}{p(\mathbf{z}_t)}\right] \\ &= \mathbb{E}\big[\log(p(\mathbf{x}_s|\mathbf{z}_t) + p(\mathbf{x}_t|\mathbf{z}_t) + p(\mathbf{x}_s \cap \mathbf{x}_t|\mathbf{z}_t))\big] \\ &\geq \mathbb{E}\big[\log(p(\mathbf{x}_s|\mathbf{z}_t) + p(\mathbf{x}_t|\mathbf{z}_t))\big] \\ &\geq \mathbb{E}\big[\log p(\mathbf{x}_t|\mathbf{z}_t)\big]. \end{aligned} \tag{12}$$

Therefore, combining (10), (11), and (12), we have the new variational lower bound

$$\mathcal{B}(p) \geq \mathbb{E}\big[\log(p(\mathbf{x}_t|\mathbf{z}_t))\big] - D_{\mathrm{KL}}\big(q(\mathbf{z}^*)||p(\mathbf{z}_t)\big). \tag{13}$$

■

Therefore, according to Theorem 1, minimizing (7) is equivalent to minimizing $D_{\mathrm{KL}}(q(\mathbf{z}^*)||p(\mathbf{z}_t))$ while maximizing $\mathbb{E}[\log(p(\mathbf{x}_t|\mathbf{z}_t))]$

$$\begin{aligned} &\min D_{\mathrm{KL}}\big(q(\mathbf{z}^*)||p(\mathbf{z}_t|\mathbf{x})\big) \\ &\quad = \max \mathbb{E}\big[\log(p(\mathbf{x}_t|\mathbf{z}_t))\big] - D_{\mathrm{KL}}\big(q(\mathbf{z}^*)||p(\mathbf{z}_t)\big) \end{aligned} \tag{14}$$

where the maximization of $\mathbb{E}[\log p(\mathbf{x}_t|\mathbf{z}_t)]$ can encourage the feature extractor G to configure the target latent space to still represent the target sampling space while approaching the variational approximation. As indicated and proved in [30], the maximization of $\mathbb{E}[\log(p(\mathbf{x}_t|\mathbf{z}_t))]$ can be achieved by minimizing the reconstruction loss using the formulation of an autoencoder

$$\max \mathbb{E}\big[\log(p(\mathbf{x}_t|\mathbf{z}_t))\big] \Rightarrow \min \big\|\mathbf{x}_t, \hat{\mathbf{x}}_t\big\| \tag{15}$$

where $\hat{\mathbf{x}}_t$ is the reconstruction of a target input sample $\mathbf{x}_t$.

C. Mean Shifting by Mutual Information Maximization

Inferring a variational approximation using the reparameterization trick that is introduced in Section IV-B makes an assumption that the mean of the variational approximation $\mathbf{z}^*$, that is, a multivariate Gaussian, totally depends on the source latent distribution. Nevertheless, in the case of missing labels for the target input samples, the mean of the variational approximation is supposed to be shifted to a critical point that could be a representation for both domains. In other words, we expect the variational approximation $\mathbf{z}^*$ to follow the *infomax* criterion: a good representation should retain a significant amount of input information [30], [33], [34]. Therefore, the information regarding the target sampling space should also be sufficiently retained in the variational approximation, that is, the mean of the variational approximation should be shifted from the source to the target. Specifically, this principle can be expressed in information-theory terms as the maximization of a lower bound of the MI between the target sampling space X_T and the space of the variational approximation Z^*, where $\mathbf{z}^* \sim Z^*$

$$\max \mathsf{I}\big(X_T; Z^*\big) = \max \mathsf{H}(X_T) - \mathsf{H}\big(X_T|Z^*\big) \tag{16}$$

where $\mathsf{I}(\cdot)$ denotes MI and $\mathsf{H}(\cdot)$ denotes the entropy.

The maximization of the lower bound of the MI between the two spaces is achieved by minimizing the L1-distance between the target input samples $\mathbf{x}_t$ and the decoded samples $\hat{\mathbf{x}}^*$ that are reconstructed from the variational approximation $\mathbf{z}^*$

$$\Delta(\mathbf{x}_t, \hat{\mathbf{x}}_*) = \big\|\mathbf{x}_t - \hat{\mathbf{x}}^*\big\| \tag{17}$$

where $\mathbf{x}_t \in X_T, \hat{\mathbf{x}}^* \in D(\mathbf{z}^* \sim q_\phi(\mathbf{z}^*))$; $\|\cdot\|$ is the ℓ_1-norm. The choice of the ℓ_1-norm is inspired by [35]. The authors

empirically verified that minimizing the ℓ_1-norm between two reconstructions that are decoded from two different latent distributions in an encoder–decoder setting will ultimately align the two latent distributions.

Therefore, we argue that minimizing (17) can maximize the lower bound of the MI between X_T and Z^*

$$\min \mathbb{E}_{\mathbf{z}^*\sim q_\phi(\mathbf{z}^*),\mathbf{x}_t\sim X_T}\left[\Delta\left(\mathbf{x}_t, D\left(\mathbf{z}^*\right)\right)\right] \Rightarrow \max \mathsf{I}\left(X_T; Z^*\right) \quad (18)$$

which encourages the mean of the variational approximation to be shifted from the source to the target; meanwhile, the knowledge learned by the source latent distribution, which is retained in the variational approximation, can be progressively transferred to the target latent distribution.

Theorem 2: For any distribution $p(\mathbf{x}_t|\mathbf{z}^*)$, if there exists

$$\mathbb{E}_{\mathbf{z}^*\sim q_\phi(\mathbf{z}^*)}\left[\log q_\phi\left(\mathbf{x}_t|\mathbf{z}^*\right)\right] \geq \mathbb{E}_{\mathbf{z}^*\sim q_\phi(\mathbf{z}^*)}\left[\log p\left(\mathbf{x}_t|\mathbf{z}^*\right)\right]$$

where $\mathsf{D}_{\mathrm{KL}}(q_\phi(\mathbf{x}_t|\mathbf{z}^*)||p(\mathbf{x}_t|\mathbf{z}^*)) \geq 0$. The following bound holds:

$$\exists k > 0 : \mathsf{I}(X_T; Z^*) \geq \sum_{\mathbf{x}_t\in X, \hat{\mathbf{x}}^*\in X^*} -\frac{1}{k}\cdot \Delta(\mathbf{x}_t, \hat{\mathbf{x}}^*).$$

Proof: To shift the mean of the variational approximation from the source to the target, we need to maximize the lower bound of the MI between the target input space and the space of the variational approximation. According to the *infomax* criterion, this process can retain the maximum information about the target domain in the space of the variational approximation Z^*. Therefore, we restrict the variational approximation using a conditional distribution $q_\phi(\mathbf{x}_t|\mathbf{z}^*;\theta')$ that is parameterized by the learning parameters θ' of the decoding network D. The MI maximization in information theory described in (7) can be expressed in terms of X_T and Z^*

$$\mathsf{I}(X_T; Z^*) = \mathsf{H}(X_T) - \mathsf{H}(X_T|Z^*) \quad (19)$$

where $\mathsf{H}(X_T)$ is a constant since the input space X_T will not be affected by the learning parameters θ of the encoding network G. Hence, the MI maximization shown in (19) can be simplified as

$$\begin{aligned}\arg\max_{\theta,\theta',\phi} \mathsf{I}\left(X_T; Z^*\right) &= \arg\max_{\theta,\theta',\phi} \mathsf{H}(X_T) - \mathsf{H}\left(X_T|Z^*\right)\\ &= \arg\max_{\theta,\theta',\phi} -\mathsf{H}\left(X_T|Z^*\right)\\ &= \arg\max_{\theta,\theta',\phi} \mathbb{E}_{\mathbf{z}^*\sim q_\phi(\mathbf{z}^*)}\left[\log q_\phi\left(\mathbf{x}_t|\mathbf{z}^*\right)\right].\end{aligned} \quad (20)$$

Now, for any distribution $p(\mathbf{x}_t|\mathbf{z}^*)$

$$\mathbb{E}_{\mathbf{z}^*\sim q_\phi(\mathbf{z}^*)}\left[\log q_\phi\left(\mathbf{x}_t|\mathbf{z}^*\right)\right] \geq \mathbb{E}[\log p(\mathbf{x}_t|\mathbf{z}^*)] \quad (21)$$

where $\mathsf{D}_{\mathrm{KL}}(q_\phi(\mathbf{x}_t|\mathbf{z}^*)||p(\mathbf{x}_t|\mathbf{z}^*)) \geq 0$.

The right-hand side of (21) can be regarded as the lower bound of the MI between the target input space and the space of the variational approximation. Considering a parametric conditional distribution $p(\mathbf{x}_t|\mathbf{z}^*;\theta')$, the lower bound of $\mathsf{I}(X_T; Z^*)$ can be represented as

$$\mathsf{I}\left(X_T; Z^*\right) \geq \mathbb{E}_{\mathbf{z}^*\sim q_\phi(\mathbf{z}^*)}\left[\log p\left(\mathbf{x}_t|\mathbf{z}^*;\theta'\right)\right]. \quad (22)$$

Normally, the reconstructed sample $\hat{\mathbf{x}}^* = D_{\theta'}(\mathbf{z}^*)$ is not exactly the same as a corresponding target sample $\mathbf{x}_t$. However, in probabilistic terms, the parameters of a distribution $p(\mathbf{x}_t|\hat{\mathbf{x}}^*)$ may produce $\mathbf{x}_t$ with high probability as they share the sufficient features for the same classification task [30]

$$p\left(\mathbf{x}_t|\mathbf{z}^*\right) \Rightarrow p\left(\mathbf{x}_t|\hat{\mathbf{x}}^* = D_{\theta'}\left(\mathbf{z}^*\right)\right) \quad (23)$$

where $p(\mathbf{x}_t|\hat{\mathbf{x}}^* = D_{\theta'}(\mathbf{z}^*))$ results in the associated L1-error

$$\begin{aligned}&\Delta\left(\mathbf{x}_t, \hat{\mathbf{x}}^*\right) \propto -\log p\left(\mathbf{x}_t|\hat{\mathbf{x}}^*\right)\\ &\Rightarrow \Delta\left(\mathbf{x}_t, \hat{\mathbf{x}}^*\right) = -k\cdot \log p\left(\mathbf{x}_t|\hat{\mathbf{x}}^*\right)\end{aligned} \quad (24)$$

where $k > 0$ is a constant.

Substituting (23) into (24), we have

$$\Delta(\mathbf{x}_t, \hat{\mathbf{x}}^*) = -k\cdot \log p(\mathbf{x}_t|\mathbf{z}^*). \quad (25)$$

Finally, combining (22) and (25), we have

$$\mathsf{I}(X_T; Z^*) \geq \sum_{\mathbf{x}_t\in X, \hat{\mathbf{x}}^*\in X^*} -\frac{1}{k}\cdot \Delta(\mathbf{x}_t, \hat{\mathbf{x}}^*). \quad (26)$$

■

Then, Corollary 1 can be derived.

Corollary 1: If there exists a set of learning parameters $\hat{\theta}'$, such that

$$\lim_{\theta'\to\hat{\theta}'} \mathsf{D}_{\mathrm{KL}}\left(q_\phi\left(\mathbf{x}_t|\mathbf{z}^*\right)||p\left(\mathbf{x}_t|\mathbf{z}^*;\hat{\theta}'\right)\right) = 0.$$

The exact value of $\mathsf{I}(X_T; Z^*)$ can be maximized by minimizing $\mathbb{E}_{\mathbf{z}^*\sim q_\phi(\mathbf{z}^*),\mathbf{x}_t\sim X_T}[\Delta(\mathbf{x}_t, D(\mathbf{z}^*))]$.

Proof: According to Theorem 2, $\mathsf{I}(X_T; Z)$ can be maximized by

$$\max_{\theta,\theta',\phi} \mathsf{I}\left(X_T; Z^*\right) \Rightarrow \max_{\theta,\theta'} \mathbb{E}_{\mathbf{z}^*\sim q_\phi(\mathbf{z}^*;\theta)}\left[\log p\left(\mathbf{x}_t|\mathbf{z}^*;\theta'\right)\right]. \quad (27)$$

The exact value of the MI $\mathsf{I}(X_T; Z^*)$ can be maximized when

$$\exists\hat{\theta}' : \lim_{\theta'\to\hat{\theta}'} \mathsf{D}_{\mathrm{KL}}(q_\phi(\mathbf{x}_t|\mathbf{z}^*)||p(\mathbf{x}_t|\mathbf{z}^*;\hat{\theta}')) = 0. \quad (28)$$

Then, rewriting (25) in Theorem 2 as the following optimization:

$$\begin{aligned}&\max_{\theta,\theta'} \mathbb{E}_{\mathbf{z}^*\sim q_\phi(\mathbf{z}^*;\theta)}\left[\log p\left(\mathbf{x}_t|\mathbf{z}^*;\theta'\right)\right]\\ &\Rightarrow \min_{\theta,\theta',\phi} \mathbb{E}_{\mathbf{z}^*\sim q_\phi(\mathbf{z}^*;\theta),\mathbf{x}_t\sim X_T}\left[\Delta\left(\mathbf{x}_t, D\left(\mathbf{z}^*\right)\right)\right].\end{aligned} \quad (29)$$

Thus, combining (27) and (29), we have

$$\begin{aligned}&\max_{\theta,\theta',\phi} \mathsf{I}\left(X_T; Z^*\right)\\ &\Rightarrow \min_{\theta,\theta',\phi} \mathbb{E}_{\mathbf{z}^*\sim q_\phi(\mathbf{z}^*;\theta),\mathbf{x}_t\sim X_T}\left[\Delta\left(\mathbf{x}_t, D\left(\mathbf{z}^*\right)\right)\right].\end{aligned} \quad (30)$$

■

At this point, we conclude that minimizing the proposed regularization for mean shifting in (18) can progressively shift the mean of the variational approximation $\mathbf{z}^*$ from the source to the target.

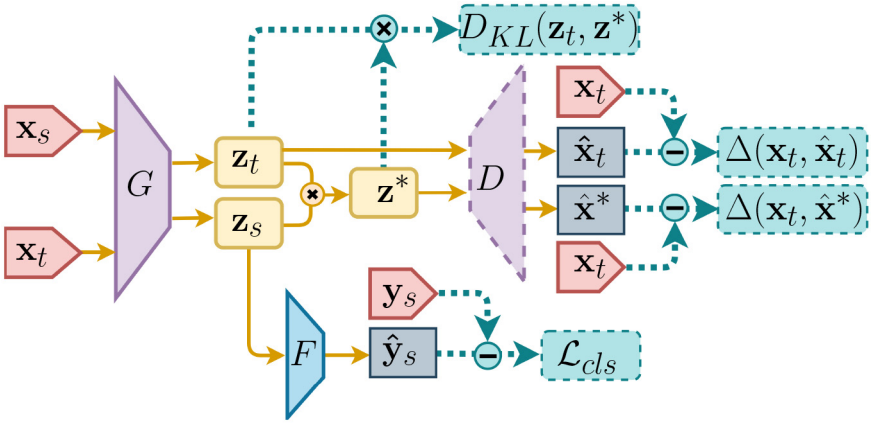


Fig. 2. Architectures of the proposed MVI. The red blocks are the model inputs and the dark-blue rectangles are the model outputs; the yellow-rounded rectangles are the latent vectors learned from the inputs; the trapezoids are the deep neural networks for the different purposes, that is, encoding, decoding, and classification; and the dashed-green rectangles are the objective functions to be optimized during the training.

D. Framework of MVI

Based on the theoretical deviations obtained in Sections IV-B and IV-C, the knowledge learned by the source latent distribution can be indirectly transferred to the target latent distribution by simultaneously optimizing the variational approximation via variational inference and the proposed regularization for the mean shifting. In this section, we present the implementation details of the proposed MVI. Fig. 2 illustrates the framework of MVI. The discriminative source latent vector $\mathbf{z}_s$ is produced by the feature extractor G under the support of the regularization of the classification task. To be specific, the source latent vector $\mathbf{z}_s$ is fed into the image classifier F to make the prediction on the corresponding input sample from the source domain, which is evaluated by the cross-entropy loss $\mathcal{L}_{\text{cls}}$

$$\mathcal{L}_{\text{cls}}(X_S, Y_S) = -\frac{1}{n}\sum_{i=1}^{n}\delta\left(\sigma \circ F \circ G\left(\mathbf{x}_s^{(i)}\right), \mathbf{y}_s^{(i)}\right)\log\left[\sigma \circ F \circ G\left(\mathbf{x}_s^{(i)}\right)\right] \tag{31}$$

where $\sigma(\cdot)$ is the softmax activation that interprets the model outputs as the non-negative probabilities that add up to 1, $\sigma \circ F \circ G(\cdot)$ is the mapping function, which is implemented by deep neural networks, that maps an input sample to its prediction, and $\delta(\cdot,\cdot)$ is the binary indicator that outputs 1 if the model prediction $\sigma \circ F \circ G(\mathbf{x}_s^{(i)})$ matches the class label $\mathbf{y_s}^{(i)}$ of the corresponding input sample $\mathbf{x}_s^{(i)}$.

Then, the target latent vector $\mathbf{z}_t$ is also produced by the feature extractor G with the same learning parameters without the support of its label. The variational approximation $\mathbf{z}^*$ is then produced with regard to $\mathbf{z}_s$ and $\mathbf{z}_t$. We utilize $\mathbf{z}^*$ in threefolds: 1) the KL-divergence penalty $\mathcal{L}_{\text{kld}}$ between $\mathbf{z}_t$ and $\mathbf{z}^*$ is computed to maximize the variational lower bound $\mathcal{B}^*(p)$; 2) the reconstruction error $\mathcal{L}_{\text{rec}}$ about the target domain, which is introduced by the variational inference, is minimized to produce the most representative $\mathbf{z}_t$; and 3) the lower bound of the MI between the target sampling space and the space of the variational approximation is maximized by minimizing the L1-loss $\mathcal{L}_{\text{mut}}$ between the target sampling space X_T and the decoded space $D(Z^*)$ of the variational approximation. The entire loss function is $\mathcal{L}_{\text{cls}} + \kappa(\mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{mut}}) + \lambda\mathcal{L}_{\text{kld}}$, and the details of MVI at each iteration of the training is summarized in Algorithm 1.

Algorithm 1: Derivation of MVI at the kth Iteration

Input: $\mathbf{x}_s$ with n minibatch source samples, $\mathbf{x}_t$ with n minibatch target samples, $\mathbf{y}_s$ with n minibatch source labels, and the hyper-parameters κ and λ.

Initialize the network parameters θ and θ', and a multivariate Gaussian distribution $\eta \sim N(\mathbf{0}, \mathbf{I})$;
for $\mathbf{x}_s \in X_S, \mathbf{x}_t \in X_T$ **do**
  $\mathbf{z}_s \leftarrow G(\mathbf{x}_s; \theta)$, $\mathbf{z}_t \leftarrow G(\mathbf{x}_t; \theta)$;
  $\hat{\mathbf{y}}_s \leftarrow F(\mathbf{z}_s)$;
  $\mathcal{L}_{cls} = \mathcal{L}_{cls}(X_S, Y_S)$;
  $\mathbf{z}^* \leftarrow \mathbf{z}_s + \eta \odot e^{\frac{1}{2}\mathbf{z}_t}$;
  $\hat{\mathbf{x}}^* \leftarrow D(\mathbf{z}^*; \theta')$, $\hat{\mathbf{x}}_t \leftarrow D(\mathbf{z}_t; \theta')$;
  $\mathcal{L}_{rec} = \frac{1}{n}\sum_{i=1}^{n}(\left\|\hat{\mathbf{x}}_t^{(i)} - \mathbf{x}_t^{(i)}\right\|)$;
  $\mathcal{L}_{mut} = \frac{1}{n}\sum_{i=1}^{n}(\left\|\hat{\mathbf{x}}^{*(i)} - \mathbf{x}_t^{(i)}\right\|)$;
  $\mathcal{L}_{kld} = -1.0 \times D_{KL}(\mathbf{z}_t || \mathbf{z}^*)$;
return $\mathcal{L}_{cls} + \kappa(\mathcal{L}_{rec} + \mathcal{L}_{mut}) + \lambda\mathcal{L}_{kld}$;

E. Computational Complexity Analysis

In our proposed framework, $\mathcal{L}_{\text{cls}}$ is the main objective to evaluate the classification error on the labeled source domain. In the meantime, the domain adaptation process happens when we minimize the objective $\kappa(\mathcal{L}_{\text{mut}} + \mathcal{L}_{\text{rec}}) + \lambda\mathcal{L}_{\text{kld}}$. Note that the process of $F \circ G$ can be regarded as a standard classification network. Denote the computational complexity of $F \circ G$ as $\mathcal{O}(G + F)$. The decoding process D can be regarded as a reverse process of G, which has the identical computational complexity as the feature extractor: $\mathcal{O}(G) = \mathcal{O}(D)$. Therefore, the computational complexity of our proposed framework is: $\mathcal{O}(G+D+F) = \mathcal{O}(G+F)$, which is the same as $F \circ G$. Thus, it can be said that our proposed framework can solve UDA problems efficiently.

V. Experiments

This section presents the experimental results of the proposed work. Our model is compared to state-of-the-art UDA algorithms on several benchmark datasets.

A. Datasets

We evaluate the proposed model using both object recognition datasets and digit recognition datasets for the task of UDA.

1) *Office-Home* contains 15 500 images of everyday objects [36]. It contains images from four different domains with different traits and background styles and each domain has 65 different object classes: a) Art (**Ar**); b) Clipart (**Cl**); c) Product (**Pr**); and d) Real-World (**Rw**).
2) *ImageCLEF-DA* is a dataset for the 2014 ImageCLEF domain adaptation challenge, which contains 12 object classes from three public datasets[1]: a) *Caltech-256* (**C**); b) *ImageNet ILSVRC2012* (**I**); and c) *Pascal VOC 2012* (**P**) and each domain contains 600 images with 50 images per class.
3) *Office-31* is a standard benchmark dataset for evaluating visual DA algorithms, which contains 31 object classes with images related to office environment [37]. This dataset has three different domains: a) *Amazon* (**A**); b) *Webcam* (**W**); and c) *DSLR* (**D**). *Amazon* consists of 2817 images from amazon.com. *Webcam* (795 images) and *DSLR* (498 images) contain images captured by a Web camera and a digital SLR camera, respectively.
4) *SVHN-MNIST-USPS*: The street view house numbers (SVHNs) dataset consists of images of digits from 0 to 9 [38]. It has significant variations in background, contrast, rotation, blurred figures, scale, etc. Both MNIST and USPS contain images of handwritten digits from 0 to 9 [39], [40].

All the datasets we used to evaluate the proposed MVI are the benchmark datasets for UDA. To be specific, Office-Home, ImageCELF-DA, and Office-31 are the benchmark datasets for the cross-domain object recognition [6], [10], [41]; SVHN, MNIST, and USPS are the benchmark datasets for the cross-domain digit recognition [9]. Some image examples for object recognition datasets and digit recognition datasets are presented in Figs. 3 and 4, respectively.

B. Implementation Details

In the experiments on Office-Home, ImageCLEF-DA, and Office-31 datasets, we follow the standard evaluation protocols for UDA as [6], [10], and [41] to utilize all labeled source samples and unlabeled target samples. For fair comparisons, ResNet-50 is selected as our backbone network, which is identical to the benchmark methods and is fine-tuned from the ImageNet [43] pretrained model. We used the SGD optimizer [44] with minibatch to train the model and repeat each transfer task five times to report the average accuracy as well as the standard deviation. We used unified hyperparameters for Office-Home, Office-31, and ImageCLEF-DA, with $\kappa = 0.1$, $\lambda = 0.1$, learning rate at 1e-3, and batch size at 32.

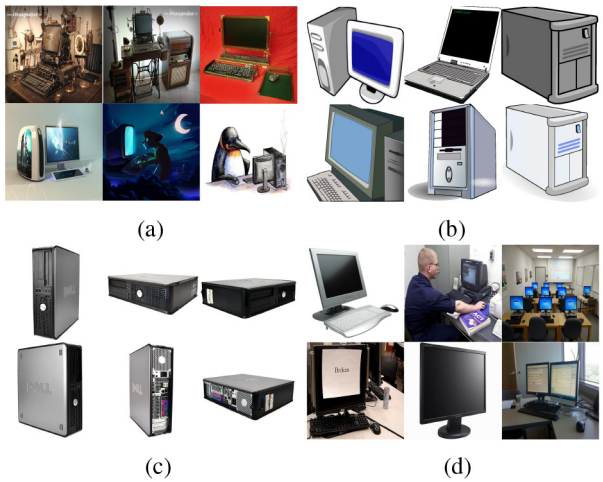


Fig. 3. Example images for computer from the four different domains of Office-Home. (a) Artistic. (b) Clipart. (c) Product. (d) Real World.

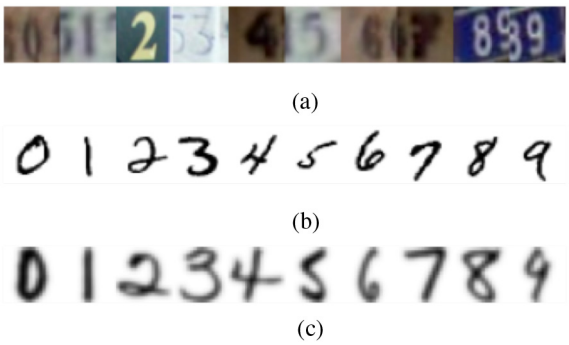


Fig. 4. Digits from the three different domains of digit recognition datasets. (a) SVHN. (b) MNIST. (c) USPS.

As for the experiments on SVHN-MNIST-USPS datasets, we utilized the SGD optimizer with a minibatch size of 128 in all experiments. We trained the model with the learning rate at 1e-2, λ at 0.1, and κ at 5e-2. We used the same network architecture as that used in MCD [9] in these digit recognition scenarios: two convolutional layers that are followed by max-pooling layers are used for the feature extraction, and three fully connected layers for calculating the classification scores are placed behind.

In this work, we tuned hyperparameters through a grid search. We essentially tune the three hyperparameters as follows: learning rate was tune from 1e-4 to 0.1, both κ and λ were tuned from 1e-3 to 1. All experiments were implemented on the *Pytorch* platform. Besides, we also minimize the conditional entropy of the softmax predictions for the target samples, which encourages the model to make more confident prediction on the unlabeled target samples [45]

$$\mathcal{L}_{\text{ent}} = \frac{1}{|X_T|} \sum_{\mathbf{x}_t \in X_T} -F(G(\mathbf{x}_t)) \log F(G(\mathbf{x}_t)). \tag{32}$$

[1]https://www.imageclef.org/2014/adaptation

TABLE I
ACCURACY(%) OF MVI ON THE *Office-Home*

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
ResNet50 [42]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [8]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
JAN [19]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN [20]	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
TransNorm [28]	50.2	71.4	77.4	59.3	72.7	73.1	61.0	**53.1**	79.5	71.9	**59.0**	82.9	67.6
TAT [21]	51.6	69.5	75.4	59.4	69.5	68.6	59.5	50.5	76.8	70.9	56.6	81.6	65.8
MVI (ours)	**51.9**	**74.5**	**79.2**	**66.5**	**74.3**	**74.5**	**63.7**	51.6	**81.4**	**74.2**	57.7	**83.5**	**69.4**
	±0.1	±0.2	±0.1	±0.2	±0.1	±0.1	±0.4	±0.1	±0.1	±0.3	±0.3	±0.1	

TABLE II
ACCURACY(%) OF MVI ON *ImageCLEF-DA*

Method	I→P	P→I	I→C	C→I	C→P	P→C	Avg
ResNet-50 [42]	74.8	83.9	91.5	78.0	65.5	91.2	80.7
DAN [8]	74.5	82.2	92.8	86.3	69.2	89.8	82.5
DANN [6]	75.0	86.0	96.2	87.0	74.3	91.5	85.0
JAN [19]	76.8	88.0	94.7	89.5	74.2	91.7	85.8
MSNT [23]	67.3±0.3	82.8±0.2	91.5±0.1	81.7±0.3	65.3±0.2	91.2±0.2	80.0
TransNorm [28]	78.3±0.3	90.8±0.2	96.7±0.4	92.3±0.2	**78.0**±0.1	94.8±0.3	88.5
MVI (ours)	**79.7** ± 0.1	**92.5**±0.3	**96.8**±0.1	**92.3**±0.5	76.5±0.3	**95.7**±0.5	**89.0**

The evaluation metric used in this article is the average classification accuracy of all input samples of the target domain, which is shown as follows:

$$\frac{1}{m}\sum_{i=1}^{m}\delta\left(\sigma\circ F\circ G\left(\mathbf{x}_t^{(i)}\right),\mathbf{y}_t^{(i)}\right) \tag{33}$$

where $\delta(\cdot,\cdot)$ is the binary indicator introduced in (31).

TABLE III
TWO-SAMPLE T-TESTS FOR TRANSNORM AND MVI

Dataset	Algorithm	Mean (%)	Std. Dev.	T-value	P-value
Office-Home	MVI	69.4	0.07	32.97	0.0001
	TransNorm	67.6	0.10		
ImageCLEF-DA	MVI	89.0	0.21	4.33	0.0025
	TransNorm	88.5	0.15		

C. Result Analysis

In the following experiments, we compared the proposed MVI to state-of-the-art methods on each dataset and reported the results of state-of-the-art methods from their original paper if similar performance could be replicated.

1) Office-Home: The evaluation results on Office-Home are shown in Table I. Our proposed framework outperforms the benchmark algorithms significantly in most transfer tasks except the transfer tasks that use Clipart as the target domain. The proposed framework also achieves the best average accuracy on the target-domain classification. The overall superior performance of the proposed framework suggests the effectiveness of the knowledge transfer via MVI. The reason for the less advanced performance in the adaptation scenarios that use Clipart as the target domain is due to the abstraction of the Clipart images. The objects from the Clipart dataset are intended for illustration, where far fewer features are included in their images compared to real-world objects. Therefore, maximizing the lower bound of the MI between two objects that are not correlated would prohibit the proposed MVI to retrieve sufficient information from the source domain for superior performance. Moreover, we notice that the improvement on Office-Home is more significant than that on other datasets, which is because Office-Home has more challenging transfer tasks by introducing the different object traits and the complex background types. The significant improvement on harder transfer tasks indicates the effectiveness of the proposed MVI and suggests that our method can transfer knowledge effectively despite the complexity of the adaptation scenario.

2) ImageCLEF-DA: The evaluation results on ImageCLEF-DA are presented in Table II. As shown in the table, our proposed framework outperforms all state-of-the-art methods except the transfer task from *Caltech-256* (**C**) to *Pascal VOC 2012* (**P**). Moreover, our proposed framework also achieves the highest overall classification accuracy. The significant improvement suggests the effectiveness of the knowledge transfer via the proposed MVI. Notably, our proposed framework significantly outperforms state-of-the-art methods except the transfer tasks that utilize *Caltech-256* (**C**) as the source domain. This is because the label pollution issue explained in [46]. Some image samples in *Caltech-256* (**C**) contain multiple objects (e.g., the images from the backpack class also contain the objects within the laptop class), which feeds the false-positive knowledge into the feature generator and the classifier and affects the effectiveness of the proposed MVI. Furthermore, as the average classification accuracies obtained from TransNorm and MVI are close, statistical tests are conducted and presented in Table III to further illustrate the superior performance of MVI. As shown in Table III, the two-sample *t*-test is conducted to determine whether the population means of the results of TransNorm and MVI on ImageCLEF-DA and Office-Home datasets are the same. The results show that the *p*-value for the average performances of MVI and TransNorm on Office-Home and ImageCLEF-DA

TABLE IV
ACCURACY(%) OF MVI ON *Office-31*

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
ResNet-50 [42]	68.4	96.7	99.3	68.9	62.5	60.7	76.1
ADDA [10]	86.2	96.2	98.4	77.8	69.5	68.9	82.9
JAN [19]	85.4	97.4	99.8	84.7	68.6	70.0	84.3
GTA [18]	89.5	97.9	99.8	87.7	72.8	71.4	86.5
DupGAN [47]	73.2±0.2	-	-	74.1±0.6	61.5±0.5	59.1±0.5	-
CCN [48]	78.2	97.4	98.6	73.5	62.8	60.6	78.5
SAFN [12]	90.1±0.8	98.6±0.2	99.8±0.0	90.7±0.5	73.0±0.2	**70.2**±0.3	87.1
MVI (ours)	**91.8**±0.3	**99.1**±0.2	**100.0**±0.0	**94.4**±0.5	**73.1**±0.2	68.5±0.3	**87.8**

TABLE V
ACCURACY(%) OF MVI ON DIGIT DATASETS

Method	SVHN→MNIST	MNIST*→USPS	MNIST→USPS	USPS→MNIST	Avg
Source Only	67.1	76.7	79.4	63.4	71.7
DANN [6]	71.1	77.1	85.1	73.0	76.6
ADDA [10]	76.0	89.4	-	90.1	-
DRCN [31]	82.0	91.8	-	73.7	-
MCD [9]	96.2± 0.4	94.2± 0.7	96.5± 0.3	94.1± 0.3	95.3
MSTN [23]	91.7± 1.5	-	92.9±1.1	-	-
GTA [18]	92.4 ± 0.9	92.8 ± 0.9	95.3 ± 0.7	90.8 ± 1.3	92.8
DEV [49]	93.2	-	92.5	96.9	-
MVI (Ours)	**98.5**± 0.5	**95.1**±1.0	**97.5**±0.3	**97.6**±0.1	**97.2**

TABLE VI
DISTANCE BETWEEN LATENT SPACES

	MNIST→USPS			USPS→ MNIST		
	$\|Z^*, Z_T\|_2$	$\|Z^*, Z_S\|_2$	$\|Z_S, Z_T\|_2$	$\|Z^*, Z_T\|_2$	$\|Z^*, Z_S\|_2$	$\|Z_S, Z_T\|_2$
Init.	0.319	0.277	0.160	0.320	0.280	0.154
Opt.	0.185	0.170	0.072	0.249	0.225	0.107
Reduce	42.0%	38.6%	55.0%	22.2%	19.6%	30.5%

datasets are much lower than 0.05. Therefore, we can conclude that the null hypothesis is rejected and the average performances of MVI and TransNorm on both datasets are significantly different. The advanced performance of MVI is not obtained due to the random chance.

3) Office-31: The evaluation results on Office-31 is shown in Table IV (the missing results were the adaptation scenarios that were not reported by the original work). Overall, our proposed framework achieves the highest classification accuracy in average and outperforms the benchmark methods in each transfer tasks except the adaptation scenario from *Webcam* to *Amazon* (**W→A**). The highest accuracy is achieved in the adaptation scenarios from *DSLR* to *Webcam* and from *Webcam* to *DSLR* (**D↔W**). This is because many image samples in *Webcam* and *DSLR* are taken from the same poses of the same object, which results in less significant domain gaps. The most significant improvements are seen in the adaptation scenarios that use *Amazon* as their source domain, which is because *Amazon* is a complex dataset with more training samples. Thus, more support could be obtained from the source domain by optimizing the proposed MVI when transferring the knowledge. In comparison, the adaptation scenarios that require transferring knowledge from the less complex datasets to *Amazon* obtain less significant improvements.

4) SVHN-MNIST-USPS: The evaluation results on digit classification using the SVHN dataset, MNIST dataset, and USPS dataset are shown in Table V (the missing results were the adaptation scenarios that did not reported by the original work). For the digit recognition tasks, we consider the domain adaptation scenarios: MNIST ↔ USPS and SVHN → MNIST. **Dataset*** denotes only a part of the dataset is used during the training phase, which follows the setting used in MCD [9]. The results show that our proposed framework outperforms state-of-the-art methods by a large margin in all transfer tasks, and achieves the highest average classification accuracy. This suggests the effectiveness of the proposed MVI in solving UDA problems.

VI. ABLATION STUDY

In the ablation study, we utilize two domain adaptation scenarios MNIST ↔ USPS as the performance indicators as they should be consistent with other evaluation protocols.

A. Effectiveness of Mutual Variational Inference

Table VI presents the distance between the feature (latent) spaces before (**init**) and after (**opt**) the training converges. Z^* denotes the space of the variational approximation. The distance is calculated as the average ℓ_2 distance between

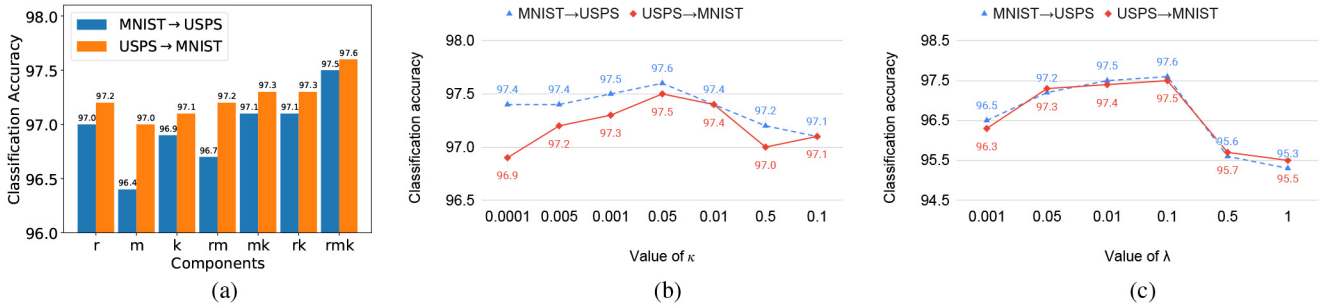


Fig. 5. Ablation study for MVI. (a) MVI components. (b) Sensitivity of κ. (c) Sensitivity of λ.

TABLE VII
COMPUTING TIME (SECONDS)

	MVI	Backbone Network	Diff.
MNIST→USPS	280.5(±1.6)	278.1(±1.1)	0.86%
USPS→MNIST	528.8(±6.5)	519.8(±5.2)	1.70%

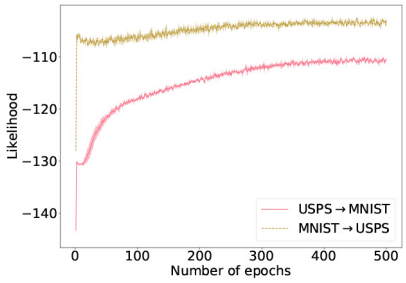


Fig. 6. Convergence of the subtraction of the likelihoods to the theoretical lower bound.

the latent spaces that are induced from the training with all source samples and target samples. The results show that the distance between the source latent space (Z_S) and the target latent space (Z_T) is reduced significantly after training by MVI. This indicates that the knowledge learned by the source latent distribution can be successfully transferred to the target latent distribution after the training of MVI converges. Note that the distance between the variational approximation and the target latent distribution reduces more significantly than the one between the variational approximation and the source latent distribution. This validates the argument in Section IV-C that minimizing (17) can encourage the mean of the variational approximation to be shifted from the source to the target through maximizing the lower bound of the MI. Furthermore, the average distance among all three distribution spaces reduces dramatically, which validates our argument that knowledge transfer can be achieved by optimizing the variational approximation. The computational complexity is compared as well in Table VII, where the average computing time for MVI and the backbone network is significantly different on both datasets, according to two sample's t-test. The results indicate that the MVI requires around 1% more computing time than the backbone network to achieve the performance improvement, which is in accordance with the complexity analysis in Section IV-E.

B. Component Analysis and Parameter Sensitivity

Fig. 5(a) presents the contribution of each component of MVI: the reconstruction of the target input samples, that is, $\hat{\mathbf{x}}_t$, the reconstruction of variational approximation, that is, $\hat{\mathbf{x}}^*$, and the KL-divergence penalty between $\mathbf{z}^*$ and $\mathbf{z}_t$, to the overall performance while keeping all hyperparameters the same. For $x \in \{\text{r}, \text{m}, \text{k}, \text{rm}, \text{mk}, \text{rk}, \text{rmk}\}$, MVI-$x$ denotes the proposed model with only the components x enabled. For this component analysis, r denotes the reconstruction of the target input samples; m denotes the reconstruction of variational approximation; and k denotes the minimization of the KL-divergence penalty. MVI-m performs worst, suggesting that simply shifting the mean of variational approximation without the optimization by variational inference cannot effectively transfer the knowledge from the source domain to the target domain. MVI-r, MVI-k, and MVI-rk improve the performance more significantly compared with MVI-m, indicating that maximizing the lower bound (or part of the lower bound) of the KL-divergence penalty between the variational approximation and the target latent distribution is useful for the knowledge transfer. The best performance is achieved by jointly using all components of MVI, that is, MVI-rmk. This suggests the importance of mean shifting (r) to optimization using variational inference (rk). Fig. 5 (b) and (c) illustrates the sensitivity of κ and λ, by varying κ from 0.001 to 0.1, and λ from 0.001 to 1. For parameter λ, the classification accuracy first steadily increases with larger values of λ and then decreases sharply. As for parameter κ, the accuracy stays almost the same as κ varies, which suggests that MVI works consistently well with different values of κ.

C. Tightness of the Variational Lower Bound

To validate the tightness of the variational lower bound, we conducted the ablation study on the adaptation scenarios MNIST ↔ USPS using an encoder with two convolutional layers with each followed by a max-pooling layer. There are 32 hidden units in the first convolutional layer and 48 hidden units in the second convolutional layer. The decoder simply deconvolves the encoder output to the input space. The architecture is identical to the prior experiment on validating the generalization of the proposed method. Fig. 6 shows the change of $\mathbb{E}[\log(p(\mathbf{x}_t|\mathbf{z}_t))] - D_{\text{KL}}(q(\mathbf{z}^*)||p(\mathbf{z}_t))$ as the training progresses, where $\mathbb{E}[\log(p(\mathbf{x}_t|\mathbf{z}_t))]$ is the likelihood of the conditional density of target input samples given their latent variables,

and $D_{\mathrm{KL}}(q(\mathbf{z}^*)||p(\mathbf{z}_t))$ is the likelihood of the KL-divergence. This subtraction of the likelihoods should converge to the variational lower bound $\mathcal{B}(p)$ after training. The results show that the subtraction of the likelihoods on the two adaptation scenarios converges to their variational lower bounds after 300 training epochs, which is identical to the theoretical derivation from Section IV-B.

VII. Conclusion

In this article, a novel method to indirectly transfer the knowledge learned by the source latent distribution to the target latent distribution via optimizing the variational approximation was proposed. We demonstrated that maximizing the lower bound of the MI between the target sampling space and the space of the variational approximation could shift the mean of the variational approximation from the source domain to the target domain. Experimental results demonstrated the importance of the proposed work.

References

[1] J. Liebelt and C. Schmid, "Multi-view object class detection with a 3D geometric model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1688–1695.

[2] D. Vazquez, A. M. López, J. Marín, D. Ponsa, and D. Gerónimo, "Virtual and real world adaptation for pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 797–809, Apr. 2014.

[3] B. Sun and K. Saenko, "From virtual to reality: Fast adaptation of virtual object detectors to real domains," in *Proc. BMVC*, vol. 1, 2014, pp. 1–12.

[4] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1–35, 2016.

[5] J. Hoffman, E. Rodner, J. Donahue, T. Darrell, and K. Saenko, "Efficient learning of domain-invariant image representations," 2013. [Online]. Available: arXiv:1301.3224.

[6] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," 2014. [Online]. Available: arXiv:1409.7495.

[7] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4068–4076.

[8] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.

[9] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3723–3732.

[10] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7167–7176.

[11] J. Hoffman *et al.*, "CyCADA: Cycle-consistent adversarial domain adaptation," 2017. [Online]. Available: arXiv:1711.03213.

[12] R. Xu, G. Li, J. Yang, and L. Lin, "Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1426–1435.

[13] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, nos. 1–2, pp. 151–175, 2010.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013. [Online]. Available: arXiv:1312.6114.

[15] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," 2015. [Online]. Available: arXiv:1509.00519.

[16] A. Taneja and A. Arora, "Cross domain recommendation using multidimensional tensor factorization," *Expert Syst. Appl.*, vol. 92, pp. 304–316, Feb. 2018.

[17] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., 2014, pp. 2672–2680.

[18] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8503–8512.

[19] Y. Pu *et al.*, "JointGAN: Multi-domain joint distribution learning with generative adversarial nets," 2018. [Online]. Available: arXiv:1806.02978.

[20] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., pp. 1640–1650, 2018.

[21] H. Liu, M. Long, J. Wang, and M. Jordan, "Transferable adversarial training: A general approach to adapting deep classifiers," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4013–4022.

[22] B. H. Nguyen, B. Xue, P. Andreae, and M. Zhang, "A hybrid evolutionary computation approach to inducing transfer classifiers for domain adaptation," *IEEE Trans. Cybern.*, early access, Apr. 8, 2020, doi: 10.1109/TCYB.2020.2980815.

[23] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5419–5428.

[24] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2058–2065.

[25] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–450.

[26] Y. Chen, S. Song, S. Li, L. Yang, and C. Wu, "Domain space transfer extreme learning machine for domain adaptation," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1909–1922, May 2018.

[27] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Transfer independently together: A generalized framework for domain adaptation," *IEEE Trans. Cybern.*, vol. 49, no. 6, pp. 2144–2155, Jun. 2019.

[28] X. Wang, Y. Jin, M. Long, J. Wang, and M. I. Jordan, "Transferable normalization: Towards improving transferability of deep neural networks," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., 2019, pp. 1951–1961.

[29] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.

[30] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.

[31] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 597–613.

[32] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, 2017.

[33] R. Linsker, "An application of the principle of maximum information preservation to linear systems," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., 1989, pp. 186–194.

[34] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.

[35] J. Wang, J. Chen, J. Lin, L. Sigal, and C. W. de Silva, "Discriminative feature alignment: Improving transferability of unsupervised domain adaptation by gaussian-guided latent alignment," 2020. [Online]. Available: arXiv:2006.12770.

[36] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5018–5027.

[37] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. Eur. Conf. Comput. Vision*, 2010, pp. 213–226.

[38] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn.*, 2011, pp. 1–9.

[39] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[40] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, May 1994.

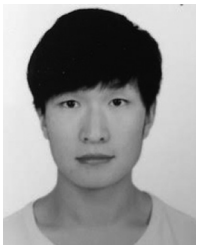
[41] Z. Cao, M. Long, J. Wang, and M. I. Jordan, "Partial transfer learning with selective adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2724–2732.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[44] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, 2010, pp. 177–186.
[45] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., 2005, pp. 529–536.
[46] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., 2016, pp. 343–351.
[47] L. Hu, M. Kan, S. Shan, and X. Chen, "Duplex generative adversarial network for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1498–1507.
[48] Y.-C. Hsu, Z. Lv, and Z. Kira, "Learning to cluster in order to transfer across domains and tasks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–20.
[49] K. You, X. Wang, M. Long, and M. Jordan, "Towards accurate model selection in deep unsupervised domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7124–7133.

Jiahong Chen (Member, IEEE) received the Ph.D. degree in mechanical engineering from the University of British Columbia, Vancouver, BC, Canada, in 2019.

He is currently a Postdoctoral Fellow with the Department of Mechanical Engineering, the University of British Columbia. His current research interests include signal processing, wireless-sensor networks, and machine learning.



Jing Wang received the B.A.Sc. degree in electrical engineering and the M.A.Sc. degree in mechanical engineering from The University of British Columbia, Vancouver, BC, Canada, in 2018 and 2020, respectively, where he is currently pursuing the Ph.D. degree with the Department of Electrical Engineering.

During his M.A.Sc studies, he worked on the robotics vision and transfer learning. His research interests include machine learning and silicon photonics.

Clarence W. de Silva (Life Fellow, IEEE) received the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1978, and the University of Cambridge, Cambridge, U.K., in 1998, and the Honorary D.Eng. degree from the University of Waterloo, Waterloo, ON, Canada, in 2008.

Since 1988, he has been a Professor of Mechanical Engineering, the Senior Canada Research Chair, and the NSERC-BC Packers Chair of Industrial Automation with the University of British Columbia, Vancouver, BC, Canada. He has authored 24 books and more than 550 papers, approximately half of which are in journals.

Prof. de Silva is a Fellow of the American Society of Mechanical Engineers, the Canadian Academy of Engineering, and the Royal Society of Canada.