SoFA: Source-data-free Feature Alignment for Unsupervised Domain Adaptation

Hao-Wei Yeh[1], Baoyao Yang[3], Pong C. Yuen[3], Tatsuya Harada[1,2]
[1]The University of Tokyo [2]RIKEN [3]Hong Kong Baptist University
yeh@mi.t.u-tokyo.ac.jp, {byyang, pcyuen}@comp.hkbu.edu.hk, harada@mi.t.u-tokyo.ac.jp

Abstract

Applying a trained model on a new scenario may suffer from domain shift. Unsupervised domain adaptation (UDA) has been proven to be an effective approach to solve the problem of domain shift by leveraging both data from the scenario that the model was trained on (source) and the new scenario (target). Although the source data are available for training the source model, there is no guarantee that the source data will still be available when applying UDA in the future due to emerging regulations on privacy of data. This results in the in-applicability of most existing UDA methods in the absence of source data. This paper proposes a source-data-free feature alignment (SoFA) method to address this problem by only using the trained source model and unlabeled target data. The source model is used to predict the labels for target data, and we model the generation process from predicted classes to input data to infer the latent features for alignment. Specifically, a mixture of Gaussian distributions is induced from the predicted classes as the reference distribution. The encoded target features are then aligned to the reference distribution via variational inference to extract class semantics without accessing source data. Relationship of the proposed method and the theory of domain adaptation is provided to verify the performance. Experimental results show the proposed method achieves higher or comparable accuracy compared to the existing methods in several cross-dataset classification tasks. Ablation studies are also conducted to confirm the importance of latent feature alignment to adaptation performance.

1. Introduction

Machine learning models are widely applied in practical scenarios with varying environmental conditions. Therefore, the test data are commonly derived from a distribution different from that of the training data. This domain shift problem [28] leads to the performance degradation when applying a model trained on one scenario (source) to a different scenario (target). Unsupervised domain adaptation [8] (UDA) has become one of the popular and effective approaches to tackle this problem. Over the years, many UDA methods [25] have been proposed, and encouraging results have been achieved in recent researches [29, 33, 14, 19] by employing deep-learning networks.

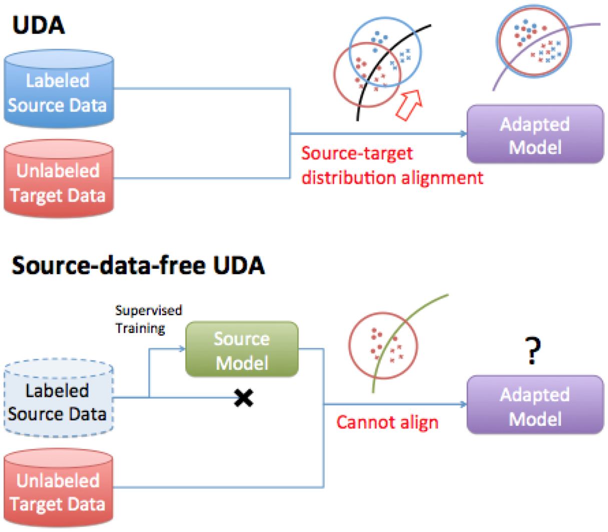

Figure 1. Comparison of the problem settings in traditional UDA and source-data-free UDA: ***Upper***: In traditional UDA, both labeled source data and unlabeled target data are available during adaptation. ***Lower***: In source-data-free UDA, labeled source data are inaccessible during adaptation. Instead, the source model, a model that has been trained on the labeled source data, and the unlabeled target data are available during adaptation. This setting is more challenging, because without the source data, the source-target feature alignment strategy that was commonly adopted in UDA methods becomes inapplicable.

Among these researches, most of them achieved successful adaptation performance with the help of a set of labeled source data. In real-world applications, however, the assumption on the availability of the source data is not always true due to emerging regulations on privacy of data. As an example, the General Data Protection Regulation (GDPR) have been promulgated and implemented in Europe to restrict the use and transfer of personal data. As a consequence, some training data or even the whole training dataset have been deleted. For instance, Microsoft announced[1] the deletion of the Ms-celeb-1m dataset [10], which is the largest publicly available dataset for facial recognition, since the publication of many images was not authorized by the owners. In such scenario, the source facial images are available when training the source facial recognition model. However, by the time we acquire new facial images and want to apply facial recognition on them, the source facial images have been deleted due to containing private information, and only the source facial recognition model is available.

In this work, we consider a more practical setting that is a variant of the traditional UDA: in addition to the unlabeled target data, the source model that has been trained on the labeled source data, rather than the labeled source data themselves, are available for domain adaptation. The problem setting in this paper is illustrated in Figure 1. The source model is more privacy-protected than the source data in the sense that it is much harder to decode private information from a trained model than inspecting actual training data. Thus, the overall setting in this paper is more suitable for the real-world applications, and is more challenging because without the source data, the effective source-target feature alignment strategy that was commonly adopted in UDA methods becomes inapplicable.

To tackle the lack of source data, some domain adaptation methods [27, 3, 24, 31] are proposed in the absence of source data. One of the solutions is to refine the source model with a few labeled target data [36, 18, 27], some generated target data [20], or through self-supervised pseudo-labeling and information maximization between target data and predictions [21]. However, these methods either initialize or regularize the adapted model by the parameters of the source model, which assumes that the parameters of the source model are available for adaptation.

Source-model prediction adjustment is another way to address domain adaptation without the source data, which does not require the availability assumption of source model parameters. The existing methods tries to either denoise [3, 24] or stabilize [31] the predictions made by the source model on target data. However, these methods are lack of encouraging the adapted model to extract class semantic information, which is desirable for classification, from target data. As the example shown in the upper half of Figure 2, without any constraints on the extracted information, the adapted recognition model extracts information such as background and product color, which are less helpful for classifying objects.

To overcome the aforementioned limitation, this paper proposes a source-data-free feature alignment method, named as SoFA, to guide the latent feature in the adapted model to extract the class semantic information from target data. In addition to the target data and the source-model predictions of target data, a mixture of Gaussian distributions is induced from the predicted classes as the reference distribution for feature alignment. Each Gaussian distribution in the mixture corresponds to a predicted class, respectively. The latent features of target data are thus required to form the same number of clusters in order to be aligned with the reference distribution. The described mechanism is illustrated in the lower half of Figure 2. In this manner, the adapted model can extract latent features with class semantic information, which is desirable for classification, from target data more efficiently. The alignment is obtained through variational inference. Specifically, we developed a Latent Alignment Variational Auto-Encoder (LA-VAE), whose decoding process models the generation process of target data. In the LA-VAE, the encoded features of the target data are used for data reconstruction to learn discriminative information from target data. In addition, the encoded target features are aligned to the mixture of Gaussian distributions in the generation process. After learning the LA-VAE, the encoded target features will tend to contain desirable class semantics. To further verify the performance, we also analyze the relationship between the proposed method and the theory of domain adaptation [1].

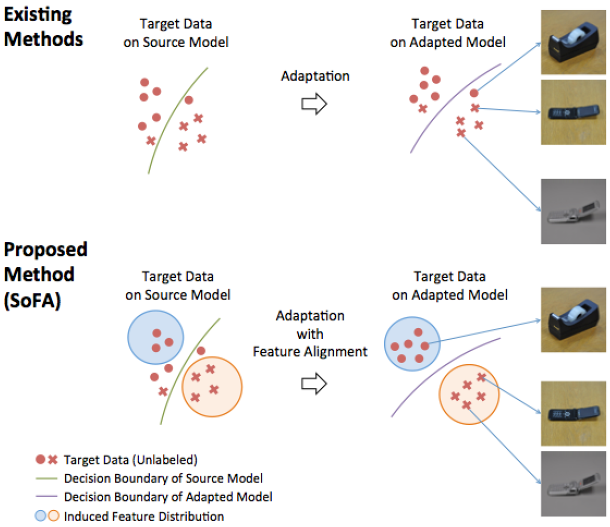
Existing Methods
Target Data on Source Model
Adaptation
Target Data on Adapted Model
Proposed Method (SoFA)
Target Data on Source Model
Adaptation with Feature Alignment
Target Data on Adapted Model
Target Data (Unlabeled)
Decision Boundary of Source Model
Decision Boundary of Adapted Model
Induced Feature Distribution

Figure 2. Comparison of SoFA with existing UDA methods without accessing source data : ***Upper***: Without constraining the extracted information from target data, the existing methods might capture ineffective information for classification (for example, object *tape dispenser* erroneously classified as *mobile phone* due to the similarity of background (wooden) and product color (black).) ***Lower***: In SoFA, the reference distribution induced from predicted classes is able to facilitate the adapted model to extract class semantic information, which is helpful for classification.

We summarize the contributions of this paper as follows,

- We propose an idea of inducing a reference feature distribution from predicted classes, and propose the

[1] https://www.engadget.com/2019/06/06/microsoft-discreetly-wiped-its-massive-facial-recognition-databa

method of source-data-free feature alignment (SoFA) to extract features with class semantics, thus realize UDA without accessing the source data. To verify the performance, we also connect the proposed method with the theory of domain adaptation.

- We show that the proposed method achieves higher or comparable accuracy when compared to existing methods on a wide range of cross-dataset classification tasks.

- We confirm the importance of feature alignment by conducting ablation studies on the proposed method.

2. Related Works

Unsupervised Domain Adaptation (UDA) is an effective approach to adapt a source model to an unlabeled target dataset whose distribution is different from the source domain. In traditional UDA, both a set of labeled source data and a set of unlabeled target data are available. Learning embedded features that are invariant across domains from the source and target data is a main approach to achieve the adaptation. In the domain-invariant feature space, the model learned by the source features can be directly applied to the target features. In recent years, many UDA algorithms [25] have been proposed to obtain the domain-invariant features with different distribution alignment criteria [9, 19] based on the techniques of manifold learning [8, 7], sparse coding [34, 35] and deep learning [22, 23, 2]. Adversarial learning techniques [6, 30, 12] are also used to explore the domain-invariant feature space. Features are regarded to be domain-invariant if they could not be correctly classified by the domain classifier.

Unsupervised Domain Adaptation without Source Data In some practical applications, source data are non-reproducible due to the current data protection regulations. To tackle this challenge, some methods are proposed to improve the performance in the target domain by refining the source model with few labeled target data [36, 18, 27] ,generated target data [20], or through self-supervised pseudo-labeling and information maximization between target data and predictions [21]. However, these methods either initialize or regularize the adapted model by the parameters of the source model, which assumes that the parameters of the source model are available for adaptation.

On the other hand, some methods are proposed to improve the performance through adjusting the predictions made by the source model on target data [3, 31, 24]. In particular, the performance in the target domain is improved by either denoise [3, 24] or stabilize [31] the source-model predictions. However, these methods are lack of encouraging the adapted model to extract class semantic information from target data. This might result in the adapted model extracting information that is less helpful for classification, such as background or product color.

3. Source-data-free Feature Alignment

This section introduces the proposed Source-data-free Feature Alignment (SoFA) method. We use the uppercase X and Y to denote the set of data and label samples, respectively. The lowercase x and y represent the data and the label of a sample, respectively. The inferred features of data x is written as z. In this paper, we focus on solving the problem of Closed Set UDA [37], where the predicted classes are identical between the source and target domains. Given the labeled source dataset $\{X_s, Y_s\}$ and the unlabeled target dataset X_t, the goal of UDA is to learn a model that can correctly classify X_t. As discussed in Section 1, this paper aims to solve the problem of unsupervised domain adaptation without source data, in which only the source model that has been well trained by the source data is given. Thus, the existing methods that extract features from source data are no longer applicable. Instead of inferring latent features with class semantics from the source data, this paper proposes inferring features from the predicted classes.

3.1. Overview

The overview of the proposed SoFA method is illustrated in Figure 3. The proposed method includes two main processes, generation process and inference process. As the trained source model and a set of target data are available, source-model predictions of target data can be obtained by inputting the target data into the source model. The generation process provides the reference feature distribution induced from the predicted classes. This reference feature distribution, also known as the prior distribution, is modeled as a mixture of Gaussian distributions. Each Gaussian distribution represents the latent feature distribution of one predicted class, respectively. On the other hand, given target data, the inference process approximates a posterior feature distribution with the assumption that the latent features of each target data sample are Gaussian distributed. Details of the generation and inference processes will be presented in Sections 3.2 and 3.3, respectively. With the inferred prior and the approximated posterior distributions, we derive the objective function which maximizes the Evidence Lower Bound (ELBO) to train our network for source-data-free feature alignment. The criteria for matching source-model predictions and data inputs, and the alignment of latent features are derived with a framework of variational inference, which will be introduced in Section 3.4.

3.2. Prior Distribution Induction from Predicted Classes

We first introduce the generation process for data x. Data x can be either the source or target data. That is, $x \in$

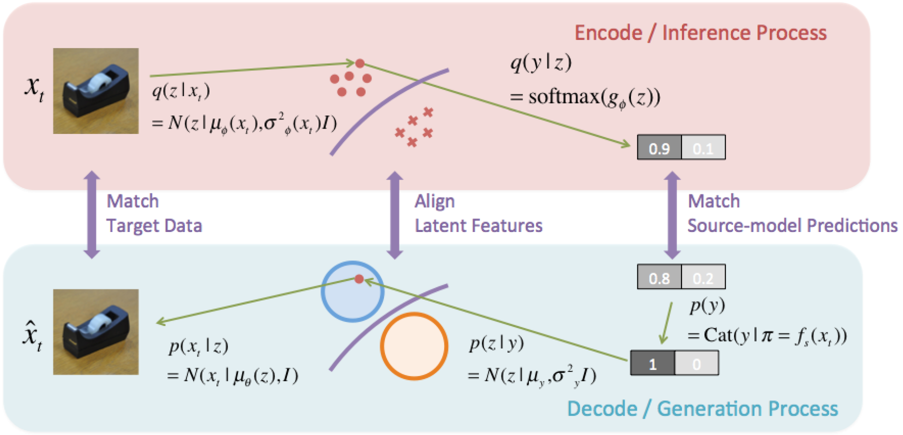
Encode / Inference Process
x_t
$q(z|x_t)$
$= N(z|\mu_\phi(x_t),\sigma^2{}_\phi(x_t)I)$
$q(y|z)$
$= \text{softmax}(g_\phi(z))$
0.9
0.1
Match
Target Data
Align
Latent Features
Match
Source-model Predictions
0.8
0.2
$p(y)$
$= \text{Cat}(y|\pi = f_s(x_t))$
1
0
$\hat{x}_t$
$p(x_t|z)$
$= N(x_t|\mu_\theta(z),I)$
$p(z|y)$
$= N(z|\mu_y,\sigma^2{}_y I)$
Decode / Generation Process

Figure 3. Illustration of the inference process from data to label (upper), and the generation process from label to data (lower) of SoFA. The directions of the processes are depicted with green arrows. With variational inference, in addition to matching the source-model predictions and target data, we can also apply distribution alignment for latent features.

$\{x_s, x_t\}$. Regardless of the domain label, a well-trained model should assign distinct modes to different classes in the classification tasks. In other words, an ideal latent feature space should cluster data with the same class label to the same mode, and assign data with different class labels to different modes. Based on this assumption, a reference distribution in the latent feature space can be modeled by a mixture of Gaussian distributions. Each Gaussian distribution in the mixture represents the latent feature distribution of one predicted class, respectively. The latent features are then mapped to the data space for data reconstruction. The entire generation process can be written as follows.

$$p(y) = \text{Cat}(y|\pi) \tag{1}$$

$$p(z|y) = \mathcal{N}(z|\mu_y, \sigma_y^2 I) \tag{2}$$

$$p(x|z) = \mathcal{N}(x|\mu_\theta(z), I) \tag{3}$$

where $\text{Cat}(\cdot)$ represents a categorical distribution with parameters π. μ_y and σ_y are the means and standard deviations of the estimated Gaussian distribution given label y. Gaussian distribution is chosen to model the generation of data given the latent features, with $\mu_\theta(z)$ denotes the neural network parametrized by a set of weights θ. To ensure the label y captures the same class semantics as in the source domain, the source-model predictions are used as the parameters for $p(y)$, i.e., $\pi = f_s(x)$, where $f_s(x)$ is the source-model predictions for data x.

3.3. Posterior Distribution Approximation via Variational Inference

We then introduce the inference process for posterior distribution approximation. In other words, we aim at calculating the posterior distribution $p(y, z|x = x_t)$. Inspired by Variational Auto-Encoder (VAE) [16], we approximate the posterior with $q(y, z|x = x_t)$. Mathematically,

$$q(y, z|x = x_t) = q(z|x = x_t)q(y|z) \tag{4}$$

$$q(z|x = x_t) = \mathcal{N}(z|\mu_\phi(x_t), \sigma_\phi^2(x_t)I) \tag{5}$$

$$q(y|z) = \text{softmax}(g_\phi(z)), \tag{6}$$

where $\mu_\phi(x)$, $\sigma_\phi(x)$ and $g_\phi(z)$ are neural networks parametrized by a set of weights ϕ, and $\text{softmax}(\cdot)$ represents the softmax function. As we wish our method can be used in most image recognition tasks, the posterior distribution is modeled by a feed-forward network, which we assume x and y are conditionally independent given z *(as shown in Equation (4))*. In order to learn smooth latent features, we formulate $q(z|x = x_t)$ as a sample-wise Gaussian distribution. $q(y|z)$ is formulated as a classifier, which takes the latent features as the inputs to predict labels.

3.4. Objective Function for Maximizing the Evidence Lower Bound

After defining the prior and posterior distributions, we achieve domain adaptation by aligning the prior and posterior distributions using variational inference. The generated target data are matched to the real target data in the autoencoder framework. The adapted predictions are matched to the source-model predictions to capture the meaning of the same set of classes as in the source domain. Finally, the encoded target features are aligned to the mixture of Gaussian distributions in the generation process. The concept is illustrated in Figure 3. The objective function that maximizes the Evidence Lower Bound (ELBO) in variational

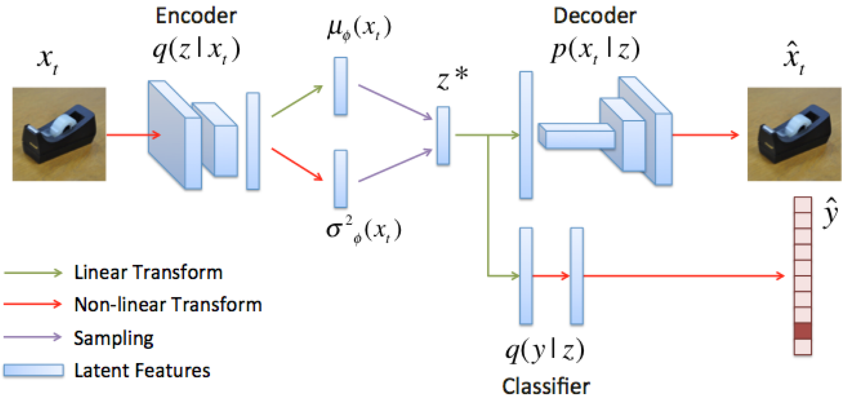

Figure 4. Overall pipeline of Latent Alignment Variational Auto-Encoder (LA-VAE). Each target data x_t is first encoded to a Gaussian distribution $\mathcal{N}(z|\mu_\phi(x_t), \sigma^2_\phi(x_t)I)$ in the latent space. We then sample the distribution to get a latent feature sample z^* of x_t, and pass it through the decoder and the classifier for data reconstruction and label prediction, respectively. *(Note that x_t can be either raw images, or features extracted from images, for example, DeCAF-fc7 features [5], features from ImageNet-pre-trained [4] ResNet50 [11], and etc.)*

inference is derived as follows.

$$
\begin{aligned}
L &= -E_{q(z,y|x_t)}\Big[\log(\frac{p(x_t,z,y)}{q(z,y|x_t)})\Big] \\
&= E_{q(z|x_t)}\Big[-\log p(x_t|z) + \sum_y [q(y|z)\log(\frac{q(z|x_t)}{p(z|y)})] \\
&\quad + KL(q(y|z)||p(y))\Big] \\
&= -\log p(x_t|z^*) + \sum_y [q(y|z^*)\log(\frac{q(z^*|x_t)}{p(z^*|y)})] \\
&\quad + KL(q(y|z^*)||p(y)), z^* \sim q(z|x_t),
\end{aligned} \tag{7}
$$

where z^* is a random sample from $q(z|x_t)$ and $KL(p||q)$ is the Kullback-Leibler divergence between distributions p and q. The set of parameters $\{\theta, \phi, \mu_y, \sigma_y\}$ are then learned by minimizing Equation (7). The objective function of Equation (7) consists of three terms: The first term is a reconstruction loss that constrains the latent features to be informative for reconstructing the target data. The second term aims to align the posterior latent features and the reference Gaussian-mixture-based latent features. The third term regularizes the posterior predictions to be close to the source-model predictions. In summary, the derived objective function produces constraints for data, latent features and predict labels, as shown in Figure 3. The first and third terms in Equation (7) encode the information from the target domain (i.e. target data) and the source domain (i.e. source-model predictions) into the latent features, respectively. The second term in Equation (7) constrains the latent features by aligning features to the mixture of Gaussian distributions, one Gaussian distribution per predicted class, helping the latent features extract information of class semantics.

Table 1. Measure of $\epsilon_{D_T}(l_S, l_T)$ training with/without Reconstruction Loss in the Dslr→Amazon setting of Office31[28]. The $\epsilon_{D_T}(l_S, l_T)$ is measured by the error rate of l_S on target data, averaging over 30 samples of $q(z|x_t)$ for every target data x_t

	$\epsilon_{D_T}(l_S, l_T)$
Without Reconstruction Loss	55.45
With Reconstruction Loss	**45.97**

3.5. Network Architecture

We realize the idea of SoFA by developing a Latent Alignment Variational Auto-Encoder (LA-VAE). The network architecture is illustrated in Figure 4. LA-VAE consists of a VAE structure and a classifier. The target data are first encoded by an encoder. The encoded outputs are passed through a batch-normalization layer [13], whose outputs are passed through a fully-connected layer and two fully-connected layers with ReLU activation to obtain the feature means $\mu_\phi(x_t)$ and feature variances $\sigma^2_\phi(x_t)$, respectively. The feature means and variances are then used as the parameters of a Gaussian distribution $\mathcal{N}(z|\mu_\phi(x_t), \sigma^2_\phi(x_t)I)$ to get a latent feature sample z^* of x_t. This sample is then passed through a decoder and a classifier for data reconstruction and label prediction, respectively. The decoder corresponds to the prior distribution inference, while the encoder and classifier correspond to the posterior distribution approximation. According to the objective function in Section 3.4, the target data reconstruction is computed by matching the decoder output to the target data. The match of label predictions across domains is computed by matching the classifier outputs to the source-model predictions. Finally, the latent feature alignment is achieved by maximizing the probability of encoded features in the mixture of Gaussian distributions, weighted by the outputs of the classifier.

3.6. Theoretical Insights

We now relate the proposed method with the theory of domain adaptation [1]. A domain S is defined as (D_S, l_S), where D_S is a data distribution and l_S is a labeling function on D_S. The disagreement between a hypothesis h and a labeling function l on distribution D can be defined as

$$\epsilon_D(h, l) = E_{z \sim D}[|h(z) - l(z)|] \tag{8}$$

Based on theory proposed by Ben-David et al.[1], given the source domain $S = (D_S, l_S)$, the target domain $T = (D_T, l_T)$, and a hypothesis space H, the error of a given hypothesis $h \in H$ on T, $\epsilon_{D_T}(h, l_T)$, is upper-bounded by the following inequality :

$$\epsilon_{D_T}(h, l_T) \leq \epsilon_{D_S}(h, l_S) + \frac{1}{2} d_{H \Delta H}(D_S, D_T) + \lambda^* \tag{9}$$

, where

$$d_{H \Delta H}(D_S, D_T) = 2 \sup_{h, h' \in H} |\epsilon_{D_S}(h, h') - \epsilon_{D_T}(h, h')|$$

$$\lambda^* = \min_{h \in H} [\epsilon_{D_S}(h, l_S) + \epsilon_{D_T}(h, l_T)]$$

$\epsilon_{D_S}(h, l_S)$ is the error of h on S, $d_{H \Delta H}(D_S, D_T)$ is the $H \Delta H$-divergence [1] between the two data distributions, and λ^* is the optimal joint error on both domains.

On the other hand, our objective function can be rewritten as follows :

$$\begin{aligned} L = E_{q(z|x_t)} \Big[-\log p(x_t|z) \Big] + KL(q(z|x_t)||p(z)) \\ + E_{q(z|x_t)} \Big[KL(q(y|z)||p(y|z)) \Big] \end{aligned} \tag{10}$$

By considering D_S, D_T, h, and l_S as $p(z)$, $q(z|x_t)$, $\text{argmax}_y(q(y|z))$, and $\text{argmax}_y(p(y|z))$, respectively, we can see the relationship between Equation (9) and Equation (10): $KL(q(z|x_t)||p(z))$ measures $d_{H \Delta H}(D_S, D_T)$, and $E_{q(z|x_t)}[KL(q(y|z)||p(y|z))]$ can measure $\epsilon_{D_S}(h, l_S)$ since the distance between D_S and D_T is reduced by minimizing $KL(q(z|x_t)||p(z))$.

We finally show the reconstruction loss, $E_{q(z|x_t)}[-\log p(x_t|z)]$, can reduce λ^*. We first notice that λ^* can be reduced if l_S and l_T have less disagreement, which can be measured by $\epsilon_{D_T}(l_S, l_T)$ since the distance between D_S and D_T is reduced during training. Table 1 summarizes the effect on the error rate of l_S on target data training with and without the reconstruction loss. The results indicate that by adding the reconstruction loss, the disagreement of the two labeling functions can be reduced. Therefore, λ^* can be reduced with the reconstruction loss.

In summary, the above shows the relationship of the proposed objective function with theory of domain adaptation, which implies minimizing the proposed objective function can reduce the bound of error on target domain and realize successful adaptation.

4. Experiments

In this section, we evaluate the proposed method on two unsupervised domain adaptation tasks: cross-dataset real-world object recognition and cross-dataset creation-to-real object recognition. Results of these tasks are reported and analyzed in Section 4.3 and Section 4.4, respectively. To evaluate the proposed method on large-scale dataset, we also conduct experiments on the VisDA-C dataset [26] and report the results in the supplementary materials. In all the experiments, ***the labeled source data are only used for training the source model and are not used during the adaptation.*** Our results of SoFA is the expected value of the predictions on latent distribution, i.e., $E_{q(z|x_t)}[q(y|z)]$, which is estimated by averaging the predictions over 30 latent samples of $q(z|x_t)$. To fairly compare with the existing linear UDA methods (for example, sMDA [3]), we also train a linear network with the latent features to mimic the final predictions of SoFA. Though we didn't observe significant difference in performance, the results of the linear model, named as SoFA student, are also provided for the purpose of fair comparison.

4.1. Implementation Details

Cross-dataset Real-world Object Recognition: As most existing methods trained a linear classifier on top of the pretrained features for adaptation, we apply a similar setting for the source model, where a linear classifier is added on top of the DeCAF-fc7 [5] features, the deep features extracted from the ImageNet-pre-tained [4] AlexNet [17]. For LA-VAE, the DECAF-fc7 features are used to infer the latent features, and a fully-connected layer with dropout is added on top of the latent features as the classifier. We set the dimension of latent features z as 1024. The decoder consists of 2 layers of the "fully-connected + batch normalization + Leaky ReLU (alpha=0.2)" module and a final fully-connected layer to reconstruct the DECAF-fc7 features. The number of channels in the fully-connected layers of the decoder are set to 4096. With the source model learned by the labeled source DeCAF-fc7 features, we trained the overall pipeline of LA-VAE for 5000 epochs until convergence. The batch size is set to 256, and the ADAM [15] optimizer with learning rate of 1e-4 is used for optimization. During the training process, we apply a "kl annealing"-like scheduling, in which the weight of the alignment term in the objective function, $\sum_y \Big[q(y|z^*) \log(\frac{q(z^*|x_t)}{p(z^*|y)}) \Big]$, is set to zero for the first 1000 epochs, and gradually ramps up from 0 to 1 over the subsequent 1500 epochs. We find this scheduling strategy prevents the network from arriving at poor local minima in the early training stages.

Cross-dataset Creation-to-real Object Recognition: Similar to the previous experiments of cross-dataset real-world object recognition, we also consider the linear classification

Table 2. Accuracies (%) of Cross-dataset Real-world Object Recognition

Method	D→A	W→A	A→D	W→D	A→W	D→W	Average
Source Only	43.66	45.83	59.84	97.59	57.99	94.09	66.50
sMDA [3]	44.34	47.00	63.86	98.19	60.75	95.09	68.21
RWA [31]	47.35±0.15	50.15±0.12	**74.34±0.99**	97.19±0.40	**72.20±1.07**	96.20±0.18	72.90±0.61
SHOT [21]	48.07±0.50	51.85±0.52	57.23±0.72	84.58±2.98	71.95±0.62	82.79±0.81	66.08±1.28
SHOT-IM [21]	**54.34±1.54**	54.20±0.32	52.57±1.23	**98.27±0.33**	66.42±1.63	95.50±0.29	70.22±1.11
SoFA (Ours)	53.71±0.53	54.63±0.56	73.90±0.44	98.19±0.18	71.72±0.56	**96.68±0.43**	**74.81±0.56**
SoFA student (Ours)	53.72±0.54	**54.64±0.55**	73.90±0.44	98.19±0.18	71.72±0.56	**96.68±0.43**	**74.81±0.56**

Table 3. Accuracies (%) of Cross-dataset Creation-to-real Object Recognition

Method	Ar→Pr	Ar→Rw	Cl→Pr	Cl→Rw	Average
Source Only	61.18	70.60	58.80	60.82	62.85
sMDA [3]	64.72	72.00	61.14	63.21	65.27
RWA [31]	73.68±0.18	76.90±0.11	71.21±0.32	69.89±0.11	72.92±0.23
SHOT [21]	69.77±0.41	74.31±0.28	70.11±0.27	72.46±0.31	71.66±0.40
SHOT-IM [21]	70.95±0.88	74.59±0.33	63.94±1.26	65.59±0.57	68.77±0.95
SoFA (Ours)	**74.14±0.10**	**77.63±0.15**	71.86±0.26	**75.09±0.37**	**74.68±0.27**
SoFA student (Ours)	74.13±0.10	77.62±0.16	**71.87±0.25**	75.08±0.36	**74.68±0.27**

setting. The features before the final linear classifier from the ImageNet-pre-trained ResNet50 [11] are used for experiment. The architectures of the source model and the LA-VAE are the same as the one in the experiments of cross-dataset real-world object recognition, with only the number of channels in the fully-connected layers of the decoder changed to 2048. The overall pipeline is trained for 5000 epochs until convergence, with batch size of 256 and ADAM optimizer with learning rate of 1e-4. The ”kl annealing”-like scheduling is also applied in this experiment, in which the weight for the alignment term is set to zero in the first 1000 epochs, and gradually ramps up from 0 to 1 over the subsequent 1500 epochs.

4.2. Comparison Methods

The results of the proposed method are compared to three existing methods that also tackle UDA in the absence of source data. 1) ***Stacked Marginalized Denoising Autoencoder (sMDA)*** [3]: A denoising auto-encoder framework is applied to marginalize the corrupted target data and the source predictions. 2) ***Random Walk based Adaptation (RWA)*** [31]: To increase the label stability, RWA repeatedly trains the network from scratch for several episodes, and re-samples the target dataset in each episode[2]. We follow the settings in the paper of RWA [31] and set the number of episodes $K = 500$ for all the experiments. 3) ***Source Hypothesis Transfer (SHOT)*** [21]: While keeping the classifier frozen, SHOT fine-tunes the feature extractor of the source model by self-supervised pseudo-labeling and information maximization between the input target data and the predictions made by the adapted model. As there are no trainable parameters in our feature extractors, we add a layer of ”fully-connected + batch normalization” module between the input features and the classifier. The module is initialized to produce identity mapping and is trained with SHOT during adaptation. We also provide the results of SHOT-IM that applies SHOT without self-supervised pseudo-labeling, as we found self-supervised pseudo-labeling worsen the performance in some experiments. In the two experiments above, except for the source model that was run once, all the compared methods and the proposed SoFA method, are conducted 5 different runs. The means and standard deviations of the results are reported in Table 2 and 3. [3]

4.3. Cross-dataset Object Recognition

To evaluate the proposed method, we conduct experiments on the Office31 dataset [28], which consists of 31 classes of real-world object images from three domains: images downloaded from amazon.com (A) and images in the office environment taken by webcams (W) and DSLR cameras (D), respectively. In each experiment, these three domains take turns to be either source or target domains. The results are summarized in Table 2, showing the proposed method achieves higher or comparable accuracy to the existing methods in each of the adaptation directions, and outperforms the existing methods in overall average accuracy.

4.4. Cross-dataset Creation-to-real Object Recognition

As the example mentioned in Section 1, the absence of source data may occur due to privacy issues. In the practical

[2] In this paper, ”episode” refers to the ”iteration” in the RWA method [31], in order to distinguish it from the term ”iteration” within each epoch in the training process.

[3] Note that as sMDA conducts deterministic computation, the standard deviations are 0 over the 5 runs and thus were not shown in the tables.

Table 4. Accuracies (%) of the Ablation Study on Cross-dataset Object Recognitions.

			Real-world	Creation-to-real
Source Only			66.50	62.85
Recon	Feature	KL		
		✓	65.72	62.29
✓		✓	66.09	62.27
	✓	✓	64.67	66.05
✓	✓	✓	**74.95**	**74.70**

cases, artworks created by individuals are one of the categories of images that are vulnerable to privacy and copyright issues. To evaluate our method in the scenario prone to privacy issues, we consider the situation where the absent source domain is the domain of drawings or creations. We conduct experiments using the Office-Home dataset [32], which consists of images of 65 object classes in four domains: artworks (Ar), clipart images (Cl), product images (Pr), and real-world images taken from camera (Rw). We select Art (Ar) and Clipart (Cl) as the source domains and the rest two domains as the target domains. The results in accuracy are summarized in Table 3. It is shown that the proposed method outperforms the existing methods in all 4 adaptation directions. We also observed that the standard deviations of the proposed method are within an acceptable range, since the proposed method still outperforms the existing methods even subtracting one standard deviations from the means of the accuracy. The results indicate that the proposed method not only outperforms the existing methods in traditional cross-domain scenarios, but is also suitable for practical cases prone to the absent of source data.

5. Discussions : Reasons for the improvements

In this section, we discuss the reasons that SoFA improves adaptation results by conducting ablation studies on the proposed method. The first two terms in the objective function, namely, the reconstruction loss, $-\log p(x_t|z^*)$, and the latent feature alignment loss, $\sum_y \left[q(y|z^*)\log(\frac{q(z^*|x_t)}{p(z^*|y)})\right]$, are partially removed for experiment. We keep the KL divergence, $KL(q(y|z^*)||p(y))$, activated to ensure the adapted predictions hold the same class semantics as in the source domain. The average results of the two cross-dataset object recognition experiments are summarized in Table 4.

As shown in Table 4, we find that the improvement achieved by the proposed method comes in two-folds: 1) Reconstruction loss that learns the discriminative information from the target data; 2) Latent feature alignment loss that facilitates the latent features to extract classification-related semantics. In the first row, the models are trained to match the source-model predictions with KL Divergence, thus generally achieve accuracy closed to the source model. In the second row, in addition to matching the source-model predictions, the models are also trained to reconstruct target data. Compared to the first row, as the reconstruction loss provides an additional training signal from the target domain, the discriminative information learned by the latent features is less affected by the domain gap. However, without the latent feature alignment loss, we can not be sure if such discriminative information learned from the reconstruction loss extracts class semantics or other information like background. Hence, such unconstrained latent features give little improvement to the performance. In the third row, the latent feature alignment loss is activated. However, without the reconstruction loss, the latent features only contain the information learned by matching the source-model predictions. Such noisy information still cannot give significant improvement to the performance. Finally, the highest accuracy occurs in the last row where all terms are included, where the reconstruction loss learns discriminative information from target data, and the latent feature alignment loss facilitates latent features to extract classification-related semantics.

6. Conclusion

In this paper, we proposed a novel method of source-data-free feature alignment (SoFA) to tackle the problem of unsupervised domain adaptation in the absence of source data. We eliminate the need of source data for unsupervised domain adaptation by inducing a reference distribution of latent features, which facilitates the model to extract semantics useful for classification. This idea is realized with variational inference that builds a path to align the encoded information across the source and target domains. We also provide theoretical insights, connecting the proposed method with the theory of domain adaptation to verify the performance. We conduct experiments on multiple classification tasks to show the effectiveness and practicality of the proposed method. In addition, effectiveness of latent feature alignment is further confirmed through ablation studies, which highlights the importance of aligning encoded information between the source and target domains in the source-data-free unsupervised domain adaptation.

Acknowledgements

We want to thank Antonio Tejero de Pablos and Adrien Bitton for inspiring discussions, and our reviewers for helpful feedback. This project was partially supported by JST AIP Acceleration Research Grant Number JPMJCR20U3, JSPS KAKENHI Grant Number JP19H01115, Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo, as well as Hong Kong Research Grant Council GRF project RGC/HKBU12200518, and the scholarship organized by Japan-Taiwan Exchange Association.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

[2] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 627–636, 2019.

[3] Boris Chidlovskii, Stéphane Clinchant, and Gabriela Csurka. Domain adaptation in the absence of source domain data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 451–460. ACM, 2016.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[7] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[8] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 999–1006, 2011.

[9] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, pages 513–520, 2007.

[10] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Duplex generative adversarial network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[14] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[18] Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pages 942–950, 2013.

[19] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019.

[20] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2020.

[21] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. *arXiv preprint arXiv:2002.08546*, 2020.

[22] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of International Conference on Machine Learning*, pages 97–105, 2015.

[23] Mingsheng Long, Jianmin Wang, Yue Cao, Jiaguang Sun, and S Yu Philip. Deep learning of transferable representation for scalable domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, pages 2027–2040, 2016.

[24] Pietro Morerio, Riccardo Volpi, Ruggero Ragonesi, and Vittorio Murino. Generative pseudo-label refinement for unsupervised domain adaptation. *arXiv preprint arXiv:2001.02950*, 2020.

[25] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *Signal Processing Magazine, IEEE*, 32(3):53–69, 2015.

[26] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[27] Michaël Perrot and Amaury Habrard. A theoretical analysis of metric hypothesis transfer learning. In *International Conference on Machine Learning*, pages 1708–1717, 2015.

[28] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In

European conference on computer vision, pages 213–226. Springer, 2010.

[29] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018.

[30] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[31] Twan van Laarhoven and Elena Marchiori. Unsupervised domain adaptation with random walks on target labelings. *arXiv preprint arXiv:1706.05335*, 2017.

[32] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *(IEEE) Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[33] Riccardo Volpi, Pietro Morerio, Silvio Savarese, and Vittorio Murino. Adversarial feature augmentation for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5495–5504, 2018.

[34] Songsong Wu, Xiao-Yuan Jing, Dong Yue, Jian Zhang, K Jian Yang, and Jingyu Yang. Unsupervised visual domain adaptation via dictionary evolution. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1–6, 2016.

[35] Baoyao Yang, Andy J Ma, and Pong C Yuen. Learning domain-shared group-sparse representation for unsupervised domain adaptation. *Pattern Recognition*, 81:615–632, 2018.

[36] Jun Yang, Rong Yan, and Alexander G Hauptmann. Cross-domain video concept detection using adaptive svms. In *Proceedings of the International Conference on Multimedia*, pages 188–197, 2007.

[37] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2720–2729, 2019.