

# Discriminative Feature Alignment: Improving Transferability of Unsupervised Domain Adaptation by Gaussian-guided Latent Alignment

Jing Wang<sup>a,\*</sup>, Jiahong Chen<sup>b</sup>, Jianzhe Lin<sup>a</sup>, Leonid Sigal<sup>c</sup> and Clarence W. de Silva<sup>b</sup>

<sup>a</sup>Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada

<sup>b</sup>Department of Mechanical Engineering, University of British Columbia, Vancouver, BC, Canada

<sup>c</sup>Department of Computer Science, University of British Columbia, Vancouver, BC, Canada

## ARTICLE INFO

### Keywords:

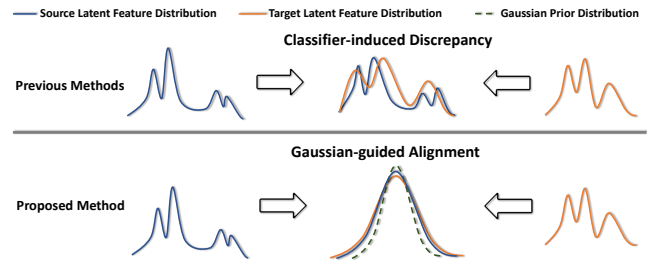
Domain adaptation  
Transfer learning  
Computer vision  
Distribution alignment  
Encoder-decoder  
Information theory

## ABSTRACT

In this paper, we focus on the unsupervised domain adaptation problem where an approximate inference model is to be learned from a labeled data domain and expected to generalize well to an unlabeled data domain. The success of unsupervised domain adaptation largely relies on the cross-domain feature alignment. Previous work has attempted to directly align latent features by the classifier-induced discrepancies. Nevertheless, a common feature space cannot always be learned via this direct feature alignment especially when a large domain gap exists. To solve this problem, we introduce a Gaussian-guided latent alignment approach to align the latent feature distributions of the two domains under the guidance of the prior distribution. In such an indirect way, the distributions over the samples from the two domains will be constructed on a common feature space, i.e., the space of the prior, which promotes better feature alignment. To effectively align the target latent distribution with this prior distribution, we also propose a novel unpaired L1-distance by taking advantage of the formulation of the encoder-decoder. The extensive evaluations on nine benchmark datasets validate the superior knowledge transferability through outperforming state-of-the-art methods and the versatility of the proposed method by improving the existing work significantly.

## 1. Introduction

The performance of computer vision models has been improved significantly by deep neural networks that take advantage of large quantities of labeled data. However, the models trained on one dataset typically perform poorly on another, different, but related, dataset [42, 30]. This shortcoming calls for adaptation strategies that help transfer knowledge from a label-rich source domain to a label-scarce target domain. Among such adaptation strategies, unsupervised domain adaptation (UDA) aims at mitigating domain shift in a way that does not use the target dataset labels, while attempting to maximize the performance of the classifier on them. Existing UDA algorithms attempt to mitigate domain shifts by only considering the classifier-induced discrepancy between the two domains, which can reduce the domain divergence [1]. Both adversarial [4, 10, 14, 37, 43] and non-adversarial domain adaptation (DA) [24, 48] methods work under the guidance of convergence learning bounds [1]. The main idea behind these bounds is that concurrently minimizing the source domain classification error and the classifier-induced discrepancy between the source domain and the target domain, inadvertently aligning the two latent feature spaces in which classification is done. In particular, adversarial DA attempts to align the feature spaces by minimizing the



**Figure 1: (Best viewed in color.)** Existing UDA methods try to align the feature distributions of the two domains by the classifier-induced discrepancies. However, it might be difficult for them to construct the two feature distributions in a single distribution space or align arbitrarily complex feature distributions in that space. Our method attempts to *indirectly* align the features of the two domains under the guidance of the Gaussian prior distribution. Our method can encourage the features of the two domains to be constructed in a common feature space, i.e., the space of the Gaussian prior, where the target samples can maximally take advantage of the discriminative source features for their own classification tasks.

classifier-induced discrepancy with adversarial objectives.

However, as shown in Figure 1, adaptation in this manner alone cannot effectively learn a common feature space for the classification in the two domains. This claim is empirically validated in Section 5.1. To address this problem, we propose a *discriminative feature alignment* (DFA) to align the two latent feature distributions of the source dataset and the target dataset under the guidance of the Gaussian prior (sim-

\*Code is available at <https://github.com/JingWang18/Discriminative-Feature-Alignment>

✉ jing@ece.ubc.ca (J. Wang); jhchen@mech.ubc.ca (J. Chen);  
jianzhelin@ece.ubc.ca (J. Lin); lsigal@cs.ubc.ca (L. Sigal);  
desilva@mech.ubc.ca (C.W. de Silva)

ORCID(s): 0000-0001-9417-1174 (J. Wang); 0000-0001-7152-8230 (J. Chen)

ilar to VAE [18]). Because the classification takes place in the latent space, the latent space itself is discriminative, in turn, making alignment focus on the discriminative feature distributions. Our approach is built on the encoder-decoder (autoencoder) formulation with an implicitly shared discriminative latent space (see Figure 3). Specifically, we define a feature extractor  $G$  which takes and encodes input samples into a latent space; similarly we define a decoder  $D$  which takes a latent feature vector, or a random vector sampled from a Gaussian prior, and decodes it back to the image. Both the encoder ( $G$ ) and the decoder ( $D$ ) are shared by the samples from the source domain and the target domain; and one can consider  $D$  as a form of regularization. We utilize a KL-divergence penalty to encourage the latent distribution over the source samples to be close to the Gaussian prior. While we can similarly encourage the target distribution in the feature space to be close to the Gaussian prior, thereby achieving the desired alignment, this turns out less effective in practice. Instead, the alignment between the source and target distributions in the latent space is achieved by a novel unpaired L1-distance between the reconstructed samples from the decoder, i.e., minimizing the distance between  $D(G(\mathbf{x}_s))$  and  $D(G(\mathbf{x}_t))$  among all pairs of samples from the source domain ( $s$ ) and the target domain ( $t$ ). The proposed regularization for the distribution alignment is named *distribution alignment loss*. We further find that instead of aligning the latent distributions directly, we get better results by aligning the target latent distribution to the Gaussian prior, i.e., minimizing the distance between  $D(G(\mathbf{x}_t))$  and the decoded samples from the prior in the feature space. The sampling also serves as data augmentation and could be useful in scenarios where the source dataset itself maybe limited.

Moreover, the proposed DFA can be incorporated into other UDA frameworks, either adversarial or non-adversarial, to improve results via a better feature alignment. To validate the versatility of DFA, we demonstrate it using an adversarial framework for the digit classification and a non-adversarial framework for the object classification. The two frameworks are developed based on the existing techniques, mainly: maximum classifier discrepancy (MCD) [37] and stepwise adaptive feature norm (SAFN) [48], since they are state-of-the-art for the digit classification and the object classification, respectively. In all settings, our DFA significantly improves the performance of the original frameworks and outperforms other existing frameworks by a large margin.

### Contributions:

- We propose a novel model for unsupervised domain adaptation, which utilizes an *indirect* latent alignment process to construct a common feature space under the guidance of a Gaussian prior.
- We introduce a new method to align two distributions, which, instead of minimizing discriminator error using a GAN, minimizes the direct L1-distance between the decoded samples.
- We evaluate the proposed frameworks and the versa-

tility of the proposed DFA on both digit and object classification tasks by adapting it into existing UDA approaches, and achieve state-of-the-art performance on the benchmark datasets.

## 2. Related Work

Existing UDA methods can be divided into two major types: adversarial and non-adversarial domain adaptation.

### 2.1. Adversarial Domain Adaptation

Motivated by generative adversarial nets (GANs) [11], adversarial DA methods, which stem from the technique proposed in [10], are widely explored by the DA community. The goal is for the latent feature distributions of the two domains to be aligned, such that domain classifier is unable to recognize domain from which the features originate. In early works, such alignment was realized by simple batch normalization statistics, which aligned the data distributions from the two domains to a canonical form [6, 21]. Introducing an adversarial loss makes it more difficult for the domain classifier to classify the domains correctly [38], producing better alignment. Further advances in adversarial DA can be found in recent works. Long et al. propose to measure the domain divergence by considering the distribution correlations for each class of objects [5, 25, 32]. Domain separation network [4] is also proposed to better preserve the component that is private to each domain before aligning the latent feature distributions.

However, the mechanism concerns constructing adversarial learning between the feature extractor and the domain classifier, which does not consider the relationship between the decision boundary and the target samples. Maximum classifier discrepancy (MCD), instead, involves an adversarial mechanism between its image classifiers and the feature extractor [37]. This method can align the latent feature distributions of the two domains by considering the decision divergence on predicting the target samples between the two image classifiers.

### 2.2. Non-adversarial Domain Adaptation

Existing non-adversarial DA methods attempt to quantify domain shifts by designing specific statistical distances between the two domains. Correlation alignment [40, 41] utilizes the difference of the mean and the covariance between the two datasets as the domain divergence, and attempts to match them during the training. The methods based on maximum mean discrepancy (MMD) [2] such as [24, 26] measure the variance between the latent feature distributions of the two domains. Some studies [8, 36, 50] also propose to learn the discriminative representations by pseudo-labels and aligning the output class distributions. However, they still consider classifier-induced discrepancies for the latent alignment, which cannot guarantee the safe transfer of the discriminative features across domains. Moreover, stepwise adaptive feature norm (SAFN) [48] identifies that domain shifts rely on the less-informative features with small norms

for the target-specific task, and the knowledge across domains can be safely transferred by placing the target features far away from these small-norm regions.

### 3. Method

In this section, the details of the proposed method are presented. First, we discuss the preliminary of the UDA problem in Section 3.1. Second, we explain about the way to achieve knowledge transfer by taking advantage of the formulation of the encoder-decoder in Section 3.2. Third, we discuss the overall idea of the proposed model in Section 3.3. Fourth, we give details about the loss functions that are used in the proposed method in Section 3.4. Finally, we demonstrate the versatility of the proposed method by incorporating it into the existing UDA methods.

#### 3.1. Preliminary

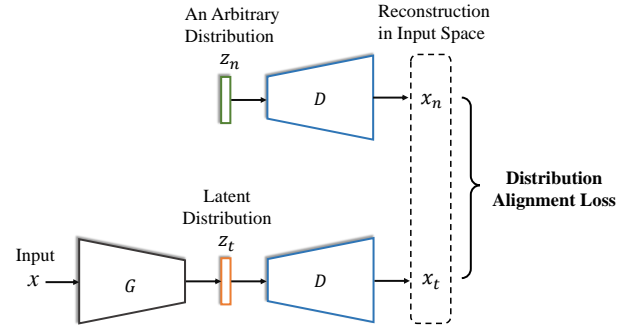
Under the setting of UDA, we sample  $n$  labeled images from the source space  $\{X_S, Y_S\}$  to form the source domain  $\mathfrak{D}_S = \{(\mathbf{x}_s^{(i)}, \mathbf{y}_s^{(i)})\}_{i=1}^n$ , as well as  $m$  unlabeled images from the target space  $\{X_T, Y_T\}$  to form the target domain  $\mathfrak{D}_T = \{(\mathbf{x}_t^{(j)})\}_{j=1}^m$ . The objective of UDA is to obtain a feature extractor  $G$  that generates a target distribution in the feature space that can maximize the performance of classifying  $\mathbf{x}_t$  without accessing its label.

#### 3.2. Knowledge Transfer via Encoder-Decoder

The proposed work is under the assumption that every neural-network-based UDA framework should consist of a feature extractor  $G$  and an image classifier  $F$ . The goal of the proposed method is not only to align the latent distributions of the two domains but also to make  $G$  learn the representation from the target samples under the guidance of the discriminative source representation. As illustrated in Figure 2, the decoder  $D$  is specifically used for the proposed *distribution alignment loss* to align the target latent distribution with the prior distribution. Thus,  $G$  is also an encoder that learns the hidden representations for both  $F$  and  $D$  in our setting. As  $G$  continuously shares its learning parameters with  $D$  during the training, our model can also be viewed as a weight-tied autoencoder. The proposed *distribution alignment loss*, which is different from the reconstruction loss used in the existing work on autoencoder, is an L1-distance between the reconstructed target samples and the decoded samples from the prior in the feature space.

##### 3.2.1. Knowledge Transfer via Distribution Alignment

The objective of unsupervised domain adaptation is to retain sufficient knowledge about the source domain in the target latent space. In a single-domain problem, the information about the input domain can be retained in its latent space by reconstructing the input samples [45]. Motivated by this, we argue that minimizing the difference between the reconstructed target samples and the source input samples can encourage the target latent space to cover sufficient information about the source domain. To be specific, minimizing the



**Figure 2:** Aligning two distributions by taking advantage of the formulation of the encoder-decoder. It contains an encoding function  $G$  and a decoding function  $D$ . The mapping function  $D(G(\circ))$  can be regarded as a weight-tied autoencoder that can put the less representative features into the nonlinear regime of  $G$ 's nonlinearity.

proposed *distribution alignment loss* on the premise of constructing the source feature space on the space of the prior is equivalent to maximizing the lower bound of the mutual information between the latent space of the target domain and the input space of the source domain.

In the setting of UDA, we are interested in learning the correspondence between the samples from the target latent space  $Z_T$  and the samples from the source input space  $X_S$ :

$$\begin{aligned} X_T &\xrightarrow{G_\theta} Z_T \xrightarrow{D_\theta} \hat{X}_T \\ X_S &\xrightarrow{G_\theta} Z_S \xrightarrow{D_\theta} \hat{X}_S, \end{aligned} \quad (1)$$

where the encoder  $G$  shares its learning parameters  $\theta$  with the decoder  $D$ .

The mutual information between the source input space and the target latent space can be expressed as

$$I(X_S; Z_T) = H(X_S) - H(X_S|Z_T), \quad (2)$$

where  $I(\cdot)$  is the mutual information;  $H(\cdot)$  is the entropy.  $H(X_S)$  is an unknown constant since the source input space  $X_S$  is from a fixed distribution that will not be affected by  $\theta$ . Hence, the information maximization process can be reduced according to Equation 2:

$$\begin{aligned} \max_{\theta} I(X_S; Z_T) &= \max_{\theta} -H(X_S|Z_T) \\ &= \max_{\theta} \mathbb{E}_{p(X_S, Z_T)} [\log p(X_S|Z_T; \theta)]. \end{aligned} \quad (3)$$

Normally, the reconstructed target sample  $\hat{x}_t = D_\theta(z_t)$  is not exactly the same as a corresponding source sample  $x_s$ . However, in probabilistic terms, the parameters of a distribution  $p(x_s|z_t)$  may produce  $\hat{x}_s$  with high probability as they share the same object feature. Therefore, the lower bound of the mutual information can be maximized by minimizing

$$L_1(x_s, \hat{x}_t) \propto -\log p(x_s|z_t), \quad (4)$$

where  $L_1$  is the L1 distance.

However, this objective cannot be achieved because of the lack of the correspondence between the reconstructed samples from the target domain and the input samples from the source domain.

To tackle this problem, we define a prior distribution  $q(\mathbf{z}_n)$  and construct the discriminative source features on the space of the prior  $Z_N$ . If there exists  $D_{KL}(q(\mathbf{z}_n)||p(\mathbf{z}_s)) = 0$ ,  $Z_S \approx Z_N$ , Equation 1 becomes

$$\begin{aligned} X_T &\xrightarrow{G_\theta} Z_T \xrightarrow{D_\theta} \hat{X}_T \\ X_S &\xrightarrow{G_\theta} Z_S \approx Z_N \xrightarrow{D_\theta} \hat{X}_N \approx \hat{X}_S, \end{aligned} \quad (5)$$

Now, we define a distribution  $q(\hat{X}_S|Z_T)$  for the following inequality:

$$\mathbb{E}_{p(X_S, Z_T)}[\log p(X_S|Z_T)] \geq \mathbb{E}_{q(\hat{X}_S, Z_T)}[\log q(\hat{X}_S|Z_T)], \quad (6)$$

where  $D_{KL}(q||p) \geq 0$ .

The left-hand side of Equation 6 is the lower bound of the mutual information between the source input space and the target latent space. We thus have a new lower bound for the mutual information:

$$\max_{\theta} \mathbb{I}(X_S; Z_T) \geq \max_{\theta} \mathbb{E}_{q(\hat{X}_S, Z_T)}[\log q(\hat{X}_S|Z_T; \theta)]. \quad (7)$$

Considering the parametric distribution  $q(\hat{X}_S|Z_T; \theta)$ , the lower bound shown in Equation 7 can be maximized by

$$\max_{\theta} \mathbb{E}_{q(\hat{X}_S, Z_T)}[\log q(\hat{X}_S|Z_T; \theta)]. \quad (8)$$

Therefore, the mutual information  $\mathbb{I}(X_S; Z_T)$  can be maximized when  $\exists \theta$  s.t.  $q(\hat{X}_S|Z_T; \theta) = p(X_S|Z_T; \theta)$ .

Combining Equation 5 and Equation 8, we have the lower bound of the mutual information between  $X_S$  and  $Z_T$  as maximizing

$$\mathbb{E}_{q(Z_N, X_T)}[\log q(\hat{X}_S \approx \hat{X}_N = D_\theta(Z_N)|Z_T = G_\theta(X_T))]. \quad (9)$$

Then, we consider the *distribution alignment error*:

$$L_1(\hat{x}_n, \hat{x}_t) \approx L_1(\hat{x}_s, \hat{x}_t) \propto -\log q(\hat{x}_s|z_t), \quad (10)$$

We thus have the following minimization that is equivalent to the maximization of the lower bound of the mutual information:

$$\begin{aligned} &\min_{\theta} \mathbb{E}_{q(\hat{X}_S, \hat{X}_T)}[L_1(\hat{X}_S, \hat{X}_T)] \\ &\Rightarrow \min_{\theta} \mathbb{E}_{q(Z_N, X_T)}[L_1(D_\theta(Z_N), D_\theta(G_\theta(X_T)))], \end{aligned} \quad (11)$$

which can be rewritten according to Equation 4 and Equation 10:

$$\begin{aligned} &\max_{\theta} \mathbb{I}(X_S; Z_T) \\ &\geq \max_{\theta} \mathbb{E}_{q(\hat{X}_S, Z_T)}[\log q(\hat{X}_S|Z_T; \theta)] \\ &\approx \max_{\theta} \mathbb{E}_{q(\hat{X}_N, Z_T)}[\log q(\hat{X}_N|Z_T; \theta)] \\ &= \max_{\theta} \mathbb{E}_{q(Z_N, X_T)}[\log q(D_\theta(Z_N)|G_\theta(X_T))] \\ &= \min_{\theta} \mathbb{E}_{q(Z_N, X_T)}[L_1(D_\theta(Z_N), D_\theta(G_\theta(X_T)))] \end{aligned} \quad (12)$$

At this point, we can conclude that the lower bound of the mutual information between the source input space  $X_S$  and the target latent space  $Z_T$  can be maximized by minimizing the proposed distribution alignment error  $L_1(\hat{x}_n, \hat{x}_t)$  on the premise that the source latent distribution is close enough to the prior.

### 3.2.2. Decoder

The proposed regularization has two functionalities in our model: 1) distribution alignment; 2) discriminative feature extraction. The distribution alignment mechanism alone cannot guarantee the produced latent distribution  $p(\mathbf{z}_t)$  is adequately discriminative for  $F$  to generalize well to the target domain. To further enforce  $G$  to focus on the cross-domain classification discriminative characteristics of the target samples, we let the weight matrices of  $G$  and  $D$  be symmetric. The choice of weight tying for the proposed encoder-decoder is motivated by the denoising autoencoder (DAE) [45]. DAE shows that the tying weight makes it more difficult for an encoder to stay in the linear regime of its nonlinearity.

We denote a mapping layer of  $G$  followed by a nonlinearity  $\sigma_i$  by

$$g_{\theta}(\mathbf{x}) = \sigma_i(\mathbf{W}_i \mathbf{x} + \mathbf{b}_i) \quad (13)$$

with learning parameters  $\theta = (\mathbf{W}_i, \mathbf{b}_i)$ , where  $\mathbf{W}_i$  is the weight matrix for the convolutional layer and  $\mathbf{b}_i$  is its bias matrix. Similarly, we define a mapping layer of  $D$  followed by the same nonlinearity  $\sigma_i$  as

$$d_{\theta^T}(\mathbf{y}) = \sigma_i(\mathbf{W}_i^T \mathbf{y} + \mathbf{b}_i^T) \quad (14)$$

with learning parameters  $\theta^T = (\mathbf{W}_i^T, \mathbf{b}_i^T)$ , where  $\mathbf{W}_i^T$  is the weight matrix for the 2-D transposed convolutional layer and  $\mathbf{b}_i^T$  is its bias matrix. Therefore, without considering the pooling, unpooling and batch normalization, our  $2L$ -layer autoencoder with tying weight can be denoted by

$$\hat{\mathbf{x}} = \sigma_1(\mathbf{W}_1^T (\dots \sigma_L(\mathbf{W}_L^T (\sigma_L(\mathbf{W}_L (\dots \sigma_1(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) + \dots) + \mathbf{b}_L) + \mathbf{b}_L^T) + \dots) + \mathbf{b}_1^T), \quad (15)$$

Then, with the support of a task-specific classifier, the less representative features can be placed in the nonlinear regime of the encoder  $G$  and, therefore, rejected. As our objective is to encourage  $p(\mathbf{z}_t)$  to be as discriminative as possible, it is straightforward to take advantage of this property of weight tying. The layers with different functionalities of the proposed decoder  $D$  are listed below:

**2-D Transposed Convolution** A convolutional layer can be represented as a sparse matrix  $\mathbf{W}$ , and has  $\mathbf{W}^T$  for its backward propagation. Thus for  $D$ , we have a transposed convolutional layer  $\mathbf{W}^T$  that utilizes  $\mathbf{W}^T$  and  $\mathbf{W}$  for its forward and backward propagations, respectively.

**Max Unpooling** The max unpooling used for  $D$  takes the output, i.e., the maximum value, of the corresponding max pooling of  $G$  and the indices of this output as its input. Then, the output of the max unpooling is appropriately sized by setting all non-maximal values to zero. While this type



of operation is not a good inverse of the max pooling, it is perfectly suitable for our objective. This is because we only want to retain the features extracted by  $G$  for the proposed *distribution alignment loss*.

**Average Unpooling** The average unpooling utilized for  $D$  takes the output of the corresponding average pooling of  $G$  as its input and sets other values to this average. Similar to the max unpooling, this operation only maintains the information of the features extracted by  $G$ .

**Nonlinearity** We observed from our experiments that the nonlinearity term retained a significant amount of features that were extracted by  $G$ . Therefore, we assume that the impact of the nonlinearity is limited to the reconstruction of the hidden representation extracted from the target domain to achieve the distribution alignment. In this study, we use the same activation function for  $D$  as that of  $G$ , i.e., ReLU activation, without considering the reversibility of the proposed encoder-decoder.

The average unpooling utilized for the decoder is the up-sampling using the nearest-neighbor interpolation. The max unpooling used for the decoder is `torch.nn.MaxUnpool2d`<sup>1</sup> implemented by **Pytorch**. The transposed convolution utilized for the decoder is `torch.nn.functional.conv_transpose2d`<sup>2</sup> implemented by **Pytorch**. The tying weight is achieved by sharing the weight matrix of the corresponding convolution with the transposed convolution. Our decoder for the object classification tasks can be viewed as an inverted version of the feature extractor of ResNet-50 with 2-D transposed convolution and upsampling. The detailed architecture and configuration of the proposed ResNet-50-based decoder are presented in the Appendix.

### 3.3. Framework of Discriminative Feature Alignment

In this section, we will discuss how to construct the latent distributions of the two domains on the space of the prior using the proposed regularization.

Our model, as illustrated in Figure 3, consists of a feature extractor  $G$  and a decoder  $D$  that share the learning parameters  $\theta_g$ . To predict the categories of the input samples, the framework developed based on our model should also have an image classifier  $F$ . We represent a mapping function from the input data, either  $\mathbf{x}_s$  or  $\mathbf{x}_t$ , to its latent feature vector  $\mathbf{z}_s$  or  $\mathbf{z}_t$  as  $G(\mathbf{x}; \theta_g)$ . Meanwhile, we denote a mapping function from a latent feature vector or the Gaussian prior vector to an image by  $D(\mathbf{z}; \theta_g)$ .

As the source dataset labels are accessible, we can make a reasonable assumption that the feature space of the source domain is discriminative. Therefore, the goal of our model is to learn a latent feature distribution  $p(\mathbf{z}_t)$  from the target domain that can maximally take advantage of the discriminative features of the source domain for its own classification. To achieve this, we need to design a feature alignment approach that can ultimately construct the two feature spaces in a common distribution space. The problem is how to define

such distribution space and effectively project the features of the two domains into this space.

For this objective, we propose to indirectly align the source features and the target features under the guidance of the Gaussian prior. As the first step of our model, we define the Gaussian prior distribution  $q(\mathbf{z}_n) \sim \mathcal{N}(0, 1)$  where we will construct the two feature spaces on. To encourage the discriminative feature space of the source domain to be constructed on the space of the prior, we regularize  $G$  and  $F$  by softmax cross-entropy loss on the labeled source samples, and enforce the distribution over the source samples  $p(\mathbf{z}_s)$  to be close to the Gaussian prior  $q(\mathbf{z}_n)$  via the KL-divergence penalty on  $G$ . Meanwhile, the latent feature distribution of the target domain  $p(\mathbf{z}_t)$  should be similarly close to the Gaussian prior. In preliminary experiments, we tried to use the same KL-divergence penalty to achieve such alignment, but it turned out to be not as effective as we expected. Therefore, to effectively align  $p(\mathbf{z}_t)$  with the prior distribution  $q(\mathbf{z}_n)$ , we propose a novel L1-distance between the reconstructed samples from the decoder, i.e., minimizing the distance between  $D(G(\mathbf{x}_t))$  and  $D(\mathbf{z}_n)$ , to regularize  $G$ . Once the training of our model converges, the three distributions, i.e., the source and the target distributions in the feature space and the Gaussian prior distribution, can be properly aligned. In other words, our method can effectively construct the feature spaces of the two domains in the same distribution space, i.e., the space of the Gaussian prior. We also include different ways to achieve such latent-space alignment in Section 5 and compare them with our proposed method.

### 3.4. Loss Functions

#### 3.4.1. Softmax Cross-entropy Loss

We use softmax cross-entropy loss to handle the classification task on the labeled source domain. This objective can ensure that the discriminative feature space of the source domain can be properly constructed on the space of the prior. We train both  $G$  and  $F$  to minimize the objective function:

$$\mathcal{L}_{cls}(X_S, Y_S) = -\frac{1}{M} \sum_{i=1}^M I(i = \mathbf{y}_s^{(i)}) \log p_s(\mathbf{x}_s^{(i)}), \quad (16)$$

where  $I(i = \mathbf{y}_s^{(i)})$  is a binary indicator which is 1 when  $i$  equals  $\mathbf{y}_s^{(i)}$ ;  $p_s$  is the mapping function for the classification scores, i.e.,  $p_s = \text{softmax} \circ F \circ G$ .

#### 3.4.2. Kullback-Leibler Divergence

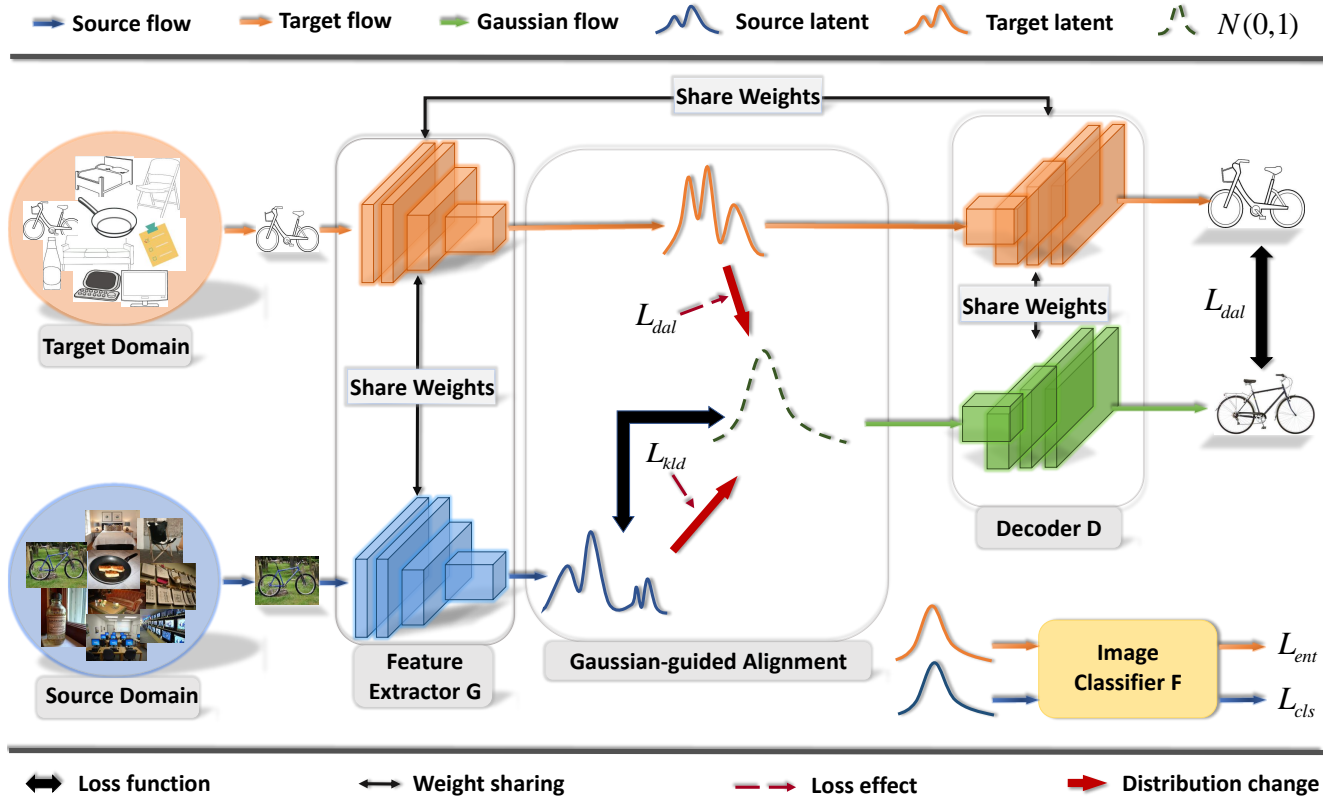
To encourage the latent feature distribution of the source domain to be close to the Gaussian prior, we apply the KL-divergence penalty between  $p(\mathbf{z}_s)$  and  $q(\mathbf{z}_n)$  to regularize  $G$ . We express this objective as:

$$\mathcal{L}_{kl}(X_S) = \frac{1}{M} \sum_{i=1}^M q(\mathbf{z}_n^{(i)}) \log \frac{q(\mathbf{z}_n^{(i)})}{G(\mathbf{x}_s^{(i)})}, \quad (17)$$

where  $G$  seeks to generate the discriminative features of the source domain in the space of the prior under the support of  $\mathcal{L}_{cls}$ .

<sup>1</sup><https://pytorch.org/docs/stable/nn.html>

<sup>2</sup><https://pytorch.org/docs/stable/nn.functional>



**Figure 3: (Best viewed in color.)** The overall architecture of the proposed framework. The feature extractor  $G$  maps the input data to their latent feature vectors. The decoder  $D$ , which can be viewed as an inverted version of  $G$ , maps a latent feature vector or Gaussian prior vector to an image that has the same dimensions as the input samples. Our model can encourage the discriminative features of the two domains to be projected into the space of the prior.

### 3.4.3. Distribution Alignment Loss

Regularizing  $G$  and  $F$  by  $\mathcal{L}_{cls}$  and  $\mathcal{L}_{kld}$ , respectively, makes the discriminative feature space of the source domain be constructed on the space of the prior. Therefore, by encouraging  $p(\mathbf{z}_t)$  to be defined in the same distribution space, tasks on the target domain can maximally take advantage of the knowledge learned from the source labels. To achieve this, we propose a simple yet effective method to align the target latent distribution with the prior distribution, namely, *distribution alignment loss* (DAL). DAL is applied to regularize both  $G$  and  $D$ . We utilize the absolute difference between the two data distributions produced by  $D$  and formulate the proposed DAL as:

$$\mathcal{L}_{dal}(X_T) = \frac{1}{M} \sum_{i=1}^M \|D(G(\mathbf{x}_t^{(i)}; \theta_g) - D(\mathbf{z}_n^{(i)}; \theta_g))\|_1, \quad (18)$$

where  $\|\cdot\|_1$  is the L1-norm. In Section 4.1, we present a detailed analysis of the proposed DAL, and empirically verify that it serves as a distribution alignment mechanism.

### 3.4.4. Entropy Loss

In the proposed framework DFA-ENT, the latent feature vector  $\mathbf{z}_t$  is fed into  $F$  to produce predictions for the target input samples. To control the contribution of the target pre-

dictions in the generalization of an image classifier, we employ a low-density separation technique *entropy minimization* (ENT) [12] to measure the class overlap of the target samples:

$$\mathcal{L}_{ent}(X_T) = \frac{1}{M} \sum_{i=1}^M -F(G(\mathbf{x}_t^{(i)})) \log F(G(\mathbf{x}_t^{(i)})). \quad (19)$$

### 3.4.5. Full Objective

The full objective function of the proposed framework DFA-ENT is a linear combination of softmax cross-entropy loss, KL-divergence penalty, *distribution alignment loss* and the entropy loss:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{ent} + \alpha \mathcal{L}_{kld} + \beta \mathcal{L}_{dal}, \quad (20)$$

where  $\alpha$  and  $\beta$  are the weights for the KL-divergence penalty and DAL, respectively, to control the relative importance of the proposed regularization.

## 3.5. Versatility

### 3.5.1. Adversarial Domain Adaptation

Maximum classifier discrepancy [37] achieves state-of-the-art on digit and traffic-sign classification. It has one feature extractor  $G$  and two image classifiers  $F_1$  and  $F_2$ . It regards the disagreement between  $F_1$  and  $F_2$  as its classifier-

induced discrepancy. It uses a three-step adversarial training strategy to avoid the input target samples that are outside the support of the source domain: first, minimizing softmax cross-entropy loss  $\mathcal{L}_{cls}$ ; second, minimizing the difference between  $\mathcal{L}_{cls}$  and the L1-loss between the outputs of the two image classifiers on the target samples  $\mathcal{L}_{adv}(X_T)$ ; and third, minimizing  $\mathcal{L}_{adv}(X_T)$ .

The proposed DFA-MCD is developed based on MCD. Our objective  $\mathcal{L}_{kld}$  is integrated into the first and the second training steps of MCD; and the proposed  $\mathcal{L}_{dal}$  is combined with the objective function of its last training step. To better clarify DFA-MCD, we include the details of the training procedures in Algorithm 1 and highlight our method in red.

#### Algorithm 1: DFA-MCD

```

1 Input image normalization; initialize the Gaussian prior  $q(\mathbf{z}_n) \sim \mathcal{N}(0, 1)$ ;
2 while  $epoch \leq \max\ epoch$  do
3   for  $batch \leftarrow 1$  to  $N$  do
4     Step 1: Sample minibatch of  $M$  samples from the Gaussian prior
        $q(\mathbf{z}_n)$ ;
5     Update  $G$ ,  $F_1$  and  $F_2$  to  $\min_{G, F_1, F_2} [\mathcal{L}_{cls}(X_S, Y_S) + \alpha \mathcal{L}_{kld}(X_S)]$ ;
6
7     Step 2: Fix  $G$ ; and update  $F_1$  and  $F_2$  to
        $\min_{F_1, F_2} [\mathcal{L}_{cls}(X_S, Y_S) - \mathcal{L}_{adv}(X_T) + \alpha \mathcal{L}_{kld}(X_S)]$ ;
8
9     Step 3: Fix  $F_1$  and  $F_2$ . Calculate  $\mathcal{L}_{dal}(X_T)$  using the current  $\theta_g$ .
       Then update  $G$  and  $D$  to  $\min_{G, D} [\mathcal{L}_{adv}(X_T) + \beta \mathcal{L}_{dal}(X_T)]$ .
10   end
11 end

```

### 3.5.2. Non-adversarial Domain Adaptation

Stepwise adaptive feature norm [48] is state-of-the-art approach on non-adversarial DA and object classification. It follows the standard DA setting with a feature extractor  $G$  and a  $l$ -layer image classifier  $F$ . It denotes the first  $l-1$  layers of its image classifier as  $F_f$ , and utilizes the intermediate features from  $F_f$  to calculate its classifier-induced discrepancy:

$$L_d(x_i) = L_2(h(x_i; \theta_p) + \delta r, h(x_i; \theta_c)), \quad (21)$$

where  $L_2$  is the L2-distance;  $h(x)$  is the L2-norm of  $F_f(G(x))$ ;  $\theta_p$  and  $\theta_c$  represent the learning parameters in the previous and the current iterations, respectively; and  $\delta r$  is a constant to control the feature-norm enlargement. Thus, SAFN can mitigate domain shifts by minimizing the following loss:

$$\begin{aligned} & \mathcal{L}_{safn}(X_S, Y_S, X_T) \\ &= \mathcal{L}_{cls}(X_S, Y_S) + \mathcal{L}_{ent}(X_T) + \kappa \mathbb{E}_{x_i \in (X_S \cup X_T)} [L_d(x_i)], \end{aligned} \quad (22)$$

where  $\kappa$  is a trade-off among the objectives.

Our DFA-SAFN is developed based on SAFN. We implement a ResNet-50-based decoder to generate  $D(\mathbf{z}_t)$  and  $D(\mathbf{z}_n)$  for the proposed DAL. We integrate all of our objective functions into the final loss of SAFN. The details of DFA-SAFN are shown in Algorithm 2.

## 4. Experiments

We implemented all experiments on the **PyTorch**<sup>3</sup> platform. We reported the results of the benchmark algorithms

<sup>3</sup><https://pytorch.org/>

#### Algorithm 2: DFA-SAFN

```

1 Input image normalization; initialize tensors for storing  $h(x_i; \theta_p)$ ,  $D(\mathbf{z}_t)$  and
   $D(\mathbf{z}_n)$ ; initialize the Gaussian prior  $q(\mathbf{z}_n) \sim \mathcal{N}(0, 1)$ ;
2 while  $epoch \leq \max\ epoch$  do
3   for  $batch \leftarrow 1$  to  $N$  do
4     Sample minibatch of  $M$  samples from the Gaussian prior  $q(\mathbf{z}_n)$ ;
5     Calculate  $L_d(X_S \cup X_T)$  using  $h(x_i; \theta_p)$  and  $h(x_i; \theta_c)$ ;
6     Calculate  $\mathcal{L}_{dal}$  using  $D(\mathbf{z}_t)$  and  $D(\mathbf{z}_n)$  from the previous iteration;
7     Update  $G$ ,  $D$  and  $F$  to minimize  $[\mathcal{L}_{safn} + \alpha \mathcal{L}_{kld} + \beta \mathcal{L}_{dal}]$ ;
8     Calculate  $h(x_i; \theta_c)$  and store it as  $h(x_i; \theta_p)$  for the next iteration;
9     Get  $D(\mathbf{z}_t)$  and  $D(\mathbf{z}_n)$  using the current  $\theta_g$  for the next iteration;
10   end
11 end

```

Table 1

Network Architectures of the encoder and the decoder for the synthetic experiments to validate the distribution alignment mechanism of the proposed regularization. FC- $x$  represents fully-connected layer with  $x$  hidden neurons. ReLU denotes the ReLU activation. BatchNorm represents the batch normalization.

Model	Architecture
Encoder $G$	FC-56, ReLU, FC-128, ReLU, FC-256, ReLU, BatchNorm
Decoder $D$	FC-128, ReLU, BatchNorm, FC-56, ReLU, FC-2, ReLU

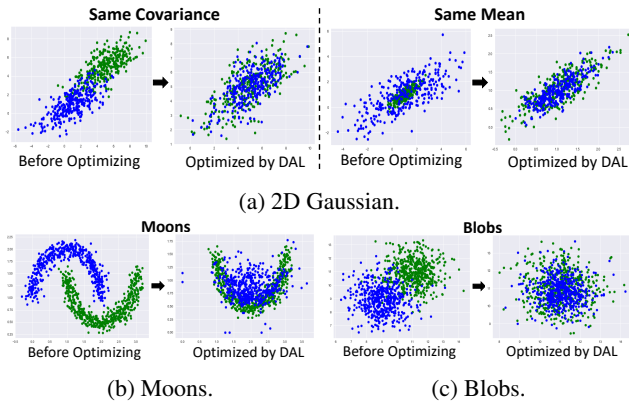
under their optimal hyper-parameter settings. To better validate the versatility of our model, we followed the same settings and the hyper-parameters that were utilized in MCD [37] and SAFN [48] for evaluating DFA-MCD and DFA-SAFN, and did not fine-tune the two frameworks. To be specific, we used Adam [17] optimizer, and set the learning rate and the batch size to  $2.0 \times 10^{-4}$  and 128, respectively, in all experiments for the evaluation on the digit and traffic-sign recognition datasets; we utilized SGD optimizer, and set the learning rate and the batch size to  $1.0 \times 10^{-3}$  and 32, respectively, in all experiments for the evaluation on the object recognition benchmark datasets.

### 4.1. Experiments on Synthetic Datasets

In this section, we empirically verified the distribution alignment mechanism of the proposed *distribution alignment loss* (DAL) on three synthetic datasets, namely, 2D Gaussian distributions with different mean or covariance, *moons* dataset and *blobs* dataset. For each experiment, we generated 500 samples for each domain. We employed the same networks  $G$  and  $D$  for all synthetic experiments. The encoder  $G$  is a 3-layer MLP that maps a 2D distribution to a higher dimensional space. The decoder  $D$ , which is also a 3-layer MLP, maps the higher dimensional latent distribution back to the input distribution space. The architectures for the two MLPs are shown in Table 1.

The samples from the target input distribution are fed into the encoder  $G$  and the decoder  $D$  to generate their predictions  $D(G(x_i))$ . The outputs of  $D$ , which are the predicted target samples, and the samples from the source input distribution are utilized for the proposed DAL. We tested the same

covariance case and the same mean case for the 2D Gaussian distributions. For the same covariance case, the **green** points (source) were sampled from a 2D Gaussian with mean  $(5 \ 5)$  and covariance  $\begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix}$ ; and the **blue** points (target) indicate the samples from a 2D Gaussian with the same covariance but different mean  $(1 \ 1)$ . For the same mean case, the two 2D Gaussian distributions have the same mean  $(1 \ 1)$  but different covariance, i.e.,  $\begin{pmatrix} 0.3 & 0.2 \\ 0.2 & 0.2 \end{pmatrix}$  for the source input distribution and  $\begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix}$  for the target input distribution. We used *scikit-learn* [31] to generate *moons* and *blobs* datasets. For *moons* dataset, we made two interleaving half circles for the two domains and add a Gaussian noise with standard deviation 0.1 to the data. For *blobs* dataset, we generated two isotropic Gaussian blobs with centers at  $(11 \ 11)$  and  $(9 \ 9)$  for the source input distribution and the target input distribution, respectively. As shown in Figure 4, the predicted target samples (**blue** points) successfully align with the source samples (**green** points) after optimizing by DAL alone in all synthetic experiments. Therefore, we can claim that the proposed DAL serves as the distribution alignment mechanism in our model.



**Figure 4: (Best viewed in color.)** Green and blue points indicate the samples from the source distribution and the target distribution, respectively. The predicted target distribution well aligns with the source distribution after the proposed *distribution alignment loss* converges, which validates the distribution alignment mechanism of *distribution alignment loss*.

## 4.2. Digit Classification

### 4.2.1. Setup

In this section, we evaluated the adaptation of our two frameworks DFA-ENT and DFA-MCD on five digit and traffic-sign recognition datasets. For each adaptation scenario, we employed the same network architectures utilized in [3, 10, 37], and implemented the decoder  $D$  accordingly. To evaluate DFA-ENT, we used the SGD optimizer with a mini-batch size of 256 in all digit and traffic-sign recognition experiments. We set the learning rate to 0.1 in the adaptation from SVHN to MNIST and 0.02 in other adaptation scenarios for evaluating DFA-ENT. Our hyper-parameters  $\alpha$  and  $\beta$  were

**Table 2**

Accuracy(%) of the proposed frameworks on the benchmark datasets for digit and traffic-sign recognition.

Method	SV→MN	SY→GT	MN→US	MN*→US*	US→MN
Source Only	67.1	85.1	76.7	79.4	63.4
DANN[10]	71.1	88.7	77.1	85.1	73.0
DSN[4]	82.7	93.1	91.3	-	-
ADDA[43]	76.0	-	89.4	-	90.1
MSTN[47]	91.7	-	-	92.9	-
GTA[38]	92.4	-	92.8	95.3	90.8
DEV[49]	93.2	-	-	92.5	<b>96.9</b>
GPDA*[16]	98.2	96.2	96.4	98.1	96.4
MCD[37]	96.2	94.4	94.2	96.5	94.1
(n = 4)	± 0.4	± 0.3	± 0.7	± 0.3	± 0.3
<b>DFA-ENT</b>	98.2	96.8	96.5	97.9	96.2
<b>(Ours)</b>	± 0.3	± 0.2	± 0.4	± 0.2	± 0.1
<b>DFA-MCD</b>	<b>98.9</b>	<b>97.5</b>	<b>97.3</b>	<b>98.6</b>	96.6
<b>(Ours)</b>	± 0.2	± 0.2	± 0.1	± 0.1	± 0.2

set to 0.01 and 10, respectively, in all adaptation scenarios for both frameworks.

**SVHN (SV) → MNIST (MN):** Street-View House Number (SVHN) [29] and MNIST [20] datasets were used as the source domain and the target domain, respectively. The two datasets consist of images of digit from 0 to 9. However, SVHN [29] has significant variations in the colored background, contrast, rotation, scale, etc.

**MNIST (MN) ↔ USPS (US):** We evaluated two adaptation scenarios on USPS [15] and MNIST [20] datasets. We used the same setup provided by [37] for the two adaptation scenarios.

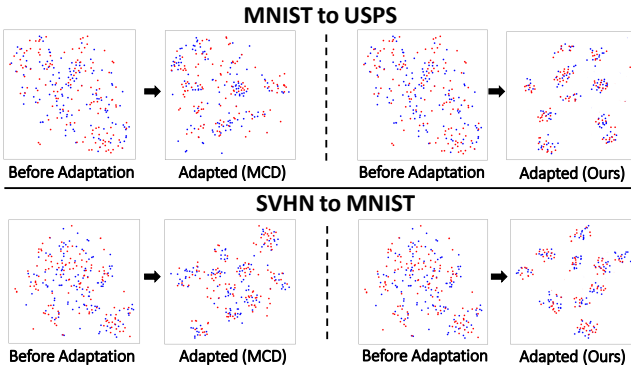
**SYN SIGNS (SY) → GTSRB (GT):** We also evaluated the proposed frameworks on a more complex scenario, from synthetic traffic signs dataset (SYN SIGNS) [28] to the real-world German Traffic Signs Recognition Benchmark (GTSRB) [39]. This domain adaptation scenario has 43 different traffic signs (classes). We split the datasets based on [37].

### 4.2.2. Results

Table 2 lists the results for the target domain classification.  $\{dataset\}^*$  denotes that all of the training samples are used for training the frameworks. We used the same networks for the source only evaluation. The average and the standard deviation of the accuracy on each DA scenario are reported by repeating each experiment 5 times. The results indicate that our model significantly improves the adaptation performance of MCD on all digit and traffic-sign datasets. The standard deviations of DFA-MCD are much lower than those of MCD, which indicates that our model can result in more robust performance. The visualizations of the learned feature representations are shown in Figure 5. The comparison is conducted between DFA-MCD and MCD. The better feature clustering indicates that our model significantly improves the adaptation performance of MCD through better

<sup>4</sup>This framework is developed based on MCD.





**Figure 5: (Best viewed in color.)** t-SNE [27] visualizations of the learned feature representations for two different adaptation scenarios. **Blue** and **red** points indicate the latent features from the source domain and the target domain, respectively.

feature alignment.

### 4.3. Object Classification

#### 4.3.1. Setup

We extensively evaluated the adaptation performance of DFA-ENT and DFA-SAFN on five benchmark datasets for object recognition, namely, *VisDA2017*, *Office-31*, *ImageCLEF-DA* and *Office-Home*. For each adaptation scenario, we employed ResNet-50 [13] that was fine-tuned from the ImageNet [9] pre-trained model. We implemented our decoder  $D$  as an inverted version of the feature extractor of ResNet-50. To evaluate DFA-ENT, we used the SGD optimizer with a learning rate of  $1 \times 10^{-3}$ , and set the batch size to 32 on all benchmark datasets. Our hyper-parameters  $\alpha$  and  $\beta$  were set to 0.1 and 10, respectively, for both frameworks.

**VisDA2017** [33] is a large-scale benchmark dataset used for the 2017 visual domain adaptation challenge. The goal of the dataset is trying to bridge the domain gap between the synthetic objects and the real objects. It has over 280K images across 12 object categories. The source domain consists of 152,397 synthetic images that are generated by rendering the 3D models of a certain object categories. The target domain contains 55,388 images of the real objects, which are collected from Microsoft COCO dataset [22]. This could be the most challenging benchmark dataset for UDA.

**Office-Home** [44] has images of everyday objects from four different domains: Artistic (**Ar**), Clipart (**Cl**), Product (**Pr**) and Real-World (**Rw**). The dataset has around 15,500 images. Each domain contains 65 object classes. Notably, **Ar** consists of the images from the different forms of artistic depictions of objects, while a regular camera takes the images of **Rw**. Some image samples from this dataset are shown in Figure 6.

**ImageCLEF-DA**<sup>5</sup> is a dataset used for the 2014 ImageCLEF domain adaptation challenge. This dataset selects 12 common object classes from three public datasets: *Caltech-256* (**C**), *ImageNet ILSVRC2012* (**I**) and *Pascal VOC 2012*

<sup>5</sup><https://www.imageclef.org/2014/adaptation>



**Figure 6:** Example images for alarm clock from the four different domains of Office-Home.

(**P**). The dataset organizers selected 50 images per class and 600 images in total for each domain.

**Office-31** [35] is a standard benchmark dataset for evaluating visual DA algorithms. It has three different domains: *Amazon* (**A**), *Webcam* (**W**), and *DSLR* (**D**). *Amazon* consists of images from amazon.com. *Webcam* and *DSLR* contain images for the office environment captured by a web camera and a digital SLR camera, respectively. It consists of 4,652 images of 31 object categories.

#### 4.3.2. Results

The results of DFA-ENT and DFA-SAFN on *VisDA2017*, *ImageCLEF-DA*, *Office-31* and *Office-Home* are listed in Table 3, 4, 5 and 6, respectively.  $\{Method\}^*$  indicates that ten-crop images are used in the evaluation phase with its best-performing models. We repeated each experiment 3 times and reported the average and the standard deviation of the accuracy for evaluating the datasets *Office-Home*, *ImageCLEF-DA* and *Office-31*. We reported the accuracy of the evaluation on *VisDA2017* after 20 epochs with no repeated experiments. The results illustrate that the proposed frameworks significantly outperform the benchmark algorithms on object classification. The robustness of SAFN is also improved by DFA with lower variance among each repeated experiments.

Results on *VisDA2017* show that the proposed DFA can significantly help the existing methods to better bridge the synthetic-to-real domain gap, which improves the performance of the baseline methods by at least 3.9% (6.2% for SAFN and 3.9% for MCD). Notably, the proposed DFA-MCD achieves state-of-the-art performance on this large-scale dataset. Besides, our simplified framework DFA-ENT achieves the competitive performance in all four benchmark datasets for the object recognition task, which suggests the effectiveness of the latent alignment in transfer learning. Moreover, the outstanding improvement on the adaptation scenarios (*Office-31*, *Office-Home*) with significant nuisance image variations suggests that our model can improve other frameworks' knowledge transferability remarkably in the adaptation scenario

**Table 3**Accuracy(%) of the proposed frameworks on *VisDA2017* (ResNet-50).

Method	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Per-class
ResNet-50 [13]	60.2	10.3	54.7	54.5	42.9	2.1	78.9	4.5	45.5	29.5	89.0	12.4	40.4
SAFN [48]	90.5	55.9	80.3	64.6	88.8	31.8	92.7	70.4	<b>93.2</b>	49.6	87.7	23.2	69.1
MCD [37]	90.3	62.6	84.8	71.7	85.9	72.9	<b>93.7</b>	71.9	86.8	79.1	81.6	14.3	74.6
<b>DFA-ENT (Ours)</b>	88.3	55.1	81.0	<b>72.9</b>	<b>91.4</b>	94.4	91.1	75.1	80.6	45.7	88.2	15.8	73.3
<b>DFA-SAFN (Ours)</b>	<b>93.1</b>	58.4	<b>85.8</b>	69.9	89.8	<b>96.1</b>	90.3	77.5	87.4	48.9	85.1	21.1	75.3
<b>DFA-MCD (Ours)</b>	91.2	<b>77.4</b>	80.5	63.3	87.1	85.4	86.4	<b>79.5</b>	90.3	<b>79.7</b>	<b>89.2</b>	<b>31.6</b>	<b>78.5</b>

**Table 4**Accuracy(%) of the proposed frameworks on *ImageCLEF-DA* (ResNet-50).

Method	I→P	P→I	I→C	C→I	C→P	P→C	Avg
ResNet-50 [13]	74.8	83.9	91.5	78.0	65.5	91.2	80.7
DANN [10]	75.0	86.0	96.2	87.0	74.3	91.5	85.0
CDAN*[25]	76.7	90.6	97.0	90.5	74.5	93.5	87.1
CADA[19]	78.0	90.5	96.7	92.0	77.2	95.5	88.3
CDAN+TN [46]	78.3	90.8	96.7	92.3	78.0	94.8	88.5
HAFN [48]	76.9	89.0	94.4	89.6	74.9	92.9	86.3
SAFN [48]	79.3	93.3	96.3	91.7	77.6	95.3	88.9
	± 0.1	± 0.4	± 0.4	± 0.0	± 0.1	± 0.1	
<b>DFA-ENT (Ours)</b>	79.5	93.0	96.4	92.5	77.2	95.8	89.1
	± 0.0	± 0.3	± 0.2	± 0.2	± 0.1	± 0.3	
<b>DFA-SAFN (Ours)</b>	<b>80.0</b>	<b>94.2</b>	<b>97.5</b>	<b>93.8</b>	<b>78.7</b>	<b>96.7</b>	<b>90.2</b>
	± 0.1	± 0.3	± 0.2	± 0.0	± 0.1	± 0.0	

**Table 5**Accuracy(%) of the proposed frameworks on *Office-31* (ResNet-50).

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
ResNet-50 [13]	68.4	96.7	99.3	68.9	62.5	60.7	76.1
DANN [10]	82.0	96.9	99.1	79.7	68.2	67.4	82.2
GTA [38]	89.5	97.9	99.8	87.7	72.8	71.4	86.5
CDAN*[25]	93.1	98.2	100.0	89.8	70.1	68.0	86.6
DSBN[7]	93.3	99.1	100.0	90.8	72.7	<b>73.9</b>	88.3
TAT[23]	92.5	99.3	100.0	93.2	73.1	72.1	88.4
HAFN [48]	83.4	98.3	99.7	84.4	69.4	68.5	83.9
SAFN [48]	90.1	98.6	99.8	90.7	73.0	70.2	87.1
	± 0.8	± 0.2	± 0.0	± 0.5	± 0.2	± 0.3	
<b>DFA-ENT (Ours)</b>	90.5	99.0	100.0	94.3	72.1	67.8	87.3
	± 0.7	± 0.1	± 0.0	± 0.4	± 0.2	± 0.4	
<b>DFA-SAFN (Ours)</b>	<b>93.5</b>	<b>99.4</b>	<b>100.0</b>	<b>94.8</b>	<b>73.8</b>	71.0	<b>88.8</b>
	± 0.5	± 0.1	± 0.0	± 0.3	± 0.1	± 0.2	

with significant variations.

One interesting observation can be revealed from these results that the transfer gains of the existing approaches, which mitigate the domain gap by classifier-induced discrepancies, can be further improved by improving the alignment in the

feature spaces. One limitation of our research is that we only consider the way to better construct the feature spaces for the DA problem and directly incorporate the proposed method into the classifier-induced discrepancy based methods. Therefore, we believe that the transfer gains can be more significantly improved by explicitly considering the relationship between the features induced from the feature extractor and the feature induced from the classifier. But how to trade off the alignment of the latent distributions against the alignment of the output class distributions is still a big challenge for the DA community.

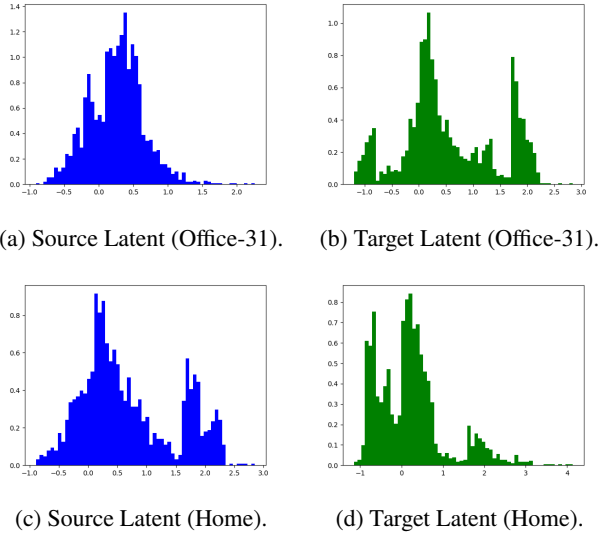
## 5. Ablation Study

### 5.1. The Shape of The Latent Distribution

In this ablation study, we validated the claim that the proposed regularization could construct the latent distributions of the two domains on a common distribution space. In our setting, the common distribution space is the space of the Gaussian prior. The best-performing models that were trained previously were used in the study. We selected a vector from the source latent distribution and one corresponding vector from the target latent distribution, and plotted their histograms for demonstration. Note that the selected vectors from the source latent distribution and the target latent distribution fall under the same category so that they should share the discriminative features. Figure 7 demonstrates that the existing UDA methods (take SAFN [48] as an example) cannot effectively construct the feature spaces of the two domains on a common distribution space. This could make the classification tasks on the target samples hard to make the most use of the discriminative source features. By contrast, as shown in Figure 8 and Figure 9, the proposed regularization can encourage the source discriminative features to be projected into the space of the Gaussian prior, and construct the target feature space on this prior distribution space. This indicates that the proposed DFA can encourage the latent distributions of the two domains to be closed to a common distribution in the feature space, i.e., the Gaussian prior, which promotes better feature alignment. Note that, the latent vectors are observed from the layer before the last ReLU activation of the encoder for better demonstration.

**Table 6**Accuracy(%) of the proposed frameworks on *Office-Home* (ResNet-50).

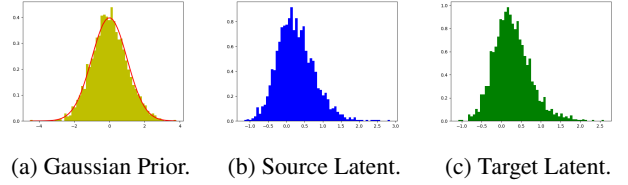
Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
ResNet-50 [13]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DANN[10]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
CDAN*[25]	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
DWT-MEC[34]	50.3	72.1	77.0	59.6	69.3	70.2	58.3	48.1	77.3	69.3	53.6	82.0	65.6
TAT[23]	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
CDAN+TN [46]	50.2	71.4	77.4	59.3	72.7	73.1	61.0	<b>53.1</b>	79.5	71.9	<b>59.0</b>	82.9	67.6
HAFN [48]	50.2	70.1	76.6	61.1	68.0	70.7	59.5	48.4	77.3	69.4	53.0	80.2	65.4
SAFN [48]	52.0	71.7	76.3	64.2	69.9	71.9	63.7	51.4	77.1	70.9	57.1	81.5	67.3
	± 0.1	± 0.6	± 0.3	± 0.3	± 0.6	± 0.6	± 0.4	± 0.2	± 0.0	± 0.4	± 0.1	± 0.0	
<b>DFA-ENT</b>	<b>50.6</b>	<b>74.8</b>	<b>79.3</b>	<b>65.2</b>	<b>73.8</b>	<b>74.5</b>	<b>63.5</b>	<b>51.4</b>	<b>81.4</b>	<b>73.9</b>	<b>58.2</b>	<b>83.3</b>	<b>69.2</b>
(Ours)	± 0.1	± 0.3	± 0.2	± 0.2	± 0.3	± 0.4	± 0.4	± 0.3	± 0.0	± 0.4	± 0.0	± 0.0	
<b>DFA-SAFN</b>	<b>52.8</b>	<b>73.9</b>	<b>77.4</b>	<b>66.5</b>	<b>72.9</b>	<b>73.6</b>	<b>64.9</b>	<b>53.1</b>	<b>78.7</b>	<b>74.5</b>	<b>58.1</b>	<b>82.4</b>	<b>69.1</b>
(Ours)	± 0.1	± 0.4	± 0.2	± 0.1	± 0.3	± 0.3	± 0.2	± 0.1	± 0.0	± 0.3	± 0.0	± 0.0	



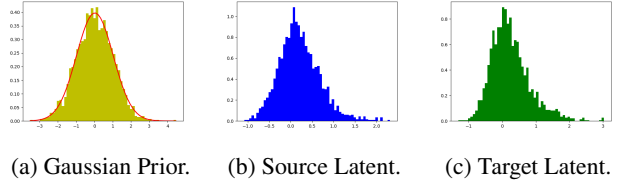
**Figure 7:** Histograms of the source latent distribution and the target latent distribution after the training of SAFN converges. **Top:** the adaptation scenario from *Amazon* to *DSLR* (*Office-31*). **Bottom:** the adaptation scenario from *Clipart* to *Product* (*Office-Home*).

## 5.2. Effectiveness of the Proposed Regularization

In this ablation study, we validated that our method could effectively align the feature spaces of the two domains. We conducted a case study on the adaptation scenario from SVHN to MNIST as its significant domain variation. We randomly selected 100 images per class from both domains and 2000 images in total. We utilized the best-performing models that were trained in the previous experiments. By measuring the distance between the feature spaces, the effectiveness of the feature alignment can be examined. We computed the average L2-distances between the feature space of SVHN and the feature space of MNIST after the adaptation with and without our model, as shown in Table 7. As expected, the



**Figure 8:** Histograms of the source latent distribution and the target latent distribution after the training of the proposed DFA-SAFN on the adaptation scenario from *Amazon* to *DSLR* (*Office-31*) converges.



**Figure 9:** Histograms of the source latent distribution and the target latent distribution after the training of the proposed DFA-SAFN on the adaptation scenario from *Clipart* to *Product* (*Office-Home*) converges.

feature-space distance of DFA-MCD is much shorter than that of MCD.

## 5.3. How to Effectively Align Feature Spaces

We investigated the most effective method for the latent alignment in this ablation study. We conducted a case study on the adaptation scenario from MNIST to UPSP. To better illustrate this study, we first define some loss functions. We formulate the paired reconstruction loss of an autoencoder as:

$$\mathcal{L}_{recon}(X) = \frac{1}{M} \sum_{i=1}^M [||D(G(\mathbf{x}^{(i)}); \theta_g) - \mathbf{x}^{(i)}||_1]. \quad (23)$$

**Table 7**

Average L2-distance between the SVHN feature space and the MNIST feature space. The numbers (0-9) denote the digit labels, and **All** indicates evaluating by all samples.

Method	0	1	2	3	4	5
MCD	0.1658	0.1433	0.1585	0.1539	0.1544	0.1598
DFA-MCD	<b>0.0644</b>	<b>0.0797</b>	<b>0.0867</b>	<b>0.0879</b>	<b>0.0871</b>	<b>0.0783</b>

Method	6	7	8	9	All
MCD	0.1529	0.1596	0.1472	0.1517	0.0564
DFA-MCD	<b>0.0800</b>	<b>0.0829</b>	<b>0.0692</b>	<b>0.0756</b>	<b>0.0266</b>

**Table 8**

Accuracy(%) of different latent-alignment methods on the adaptation scenario from MNIST to USPS. Note that all methods utilize  $\mathcal{L}_{ent}$  and  $\mathcal{L}_{cls}$  for classification.

	$\mathcal{L}_{kld} + \mathcal{L}_{dal}$ (Ours)	$\mathcal{L}_{daldir}$	$\mathcal{L}_{kld}$
<b>Accuracy</b>	<b>97.3</b>	93.1	87.9

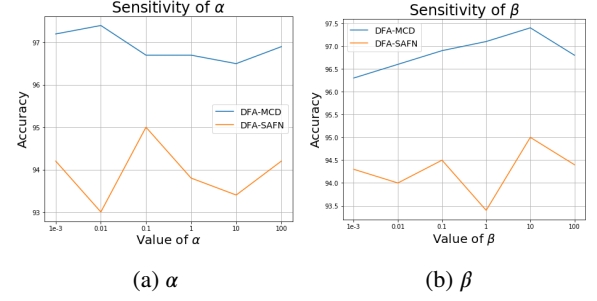
  

	$\mathcal{L}_{kld} + \mathcal{L}_{dal}, \theta_d \neq \theta_g$	$\mathcal{L}_{klddir}$	$\mathcal{L}_{klddir} + \mathcal{L}_{recon}$
<b>Accuracy</b>	95.8	89.2	83.6

We define a KL-divergence penalty to encourage  $p(\mathbf{z}_t)$  to be close to  $p(\mathbf{z}_s)$  as  $\mathcal{L}_{klddir}$ . To validate the effect of weight tying, we further define the learning parameters  $\theta_d$  for the decoder  $D$  in the case where the tying weight is not applied. We explored six different ways to align the two latent feature distributions  $p(\mathbf{z}_s)$  and  $p(\mathbf{z}_t)$ : **1)** the proposed **DFA-ENT** framework; **2)** **DFA-ENT** but the encoder  $G$  and the decoder  $D$  do not share their weights ( $\theta_d \neq \theta_g$ ); **3)** instead of using our DAL to align the target latent distribution with the Gaussian prior, utilizing a KL-divergence to make  $p(\mathbf{z}_t)$  close to the prior; **4)** the direct latent alignment via an unpaired L1-distance between the reconstructed samples from the two domains, i.e., minimizing the distance between  $D(G(\mathbf{x}_s))$  and  $D(G(\mathbf{x}_t))$  ( $\mathcal{L}_{daldir}$ ); **5)** the direct latent alignment using  $\mathcal{L}_{klddir}$ ; and **6)** further regularizing Case **5)** by two reconstruction losses  $\mathcal{L}_{recon}(X_S) + \mathcal{L}_{recon}(X_T)$  ( $\mathcal{L}_{recon}$ ) with our weight-tied encoder-decoder formulation. The results, which are shown in Table 8, indicate that the proposed DFA is the most effective approach to align the latent distributions of the two domains. The ablation study validates that all of the Gaussian-guided alignment, unpaired L1-distance and weight tying are of necessity for the proposed model.

#### 5.4. Parameter Sensitivity

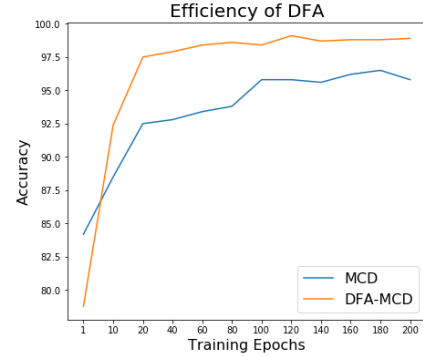
To quantify the impact of our *discriminative feature alignment* (DFA) on the UDA frameworks, we investigated the sensitivity of our hyper-parameters, i.e.,  $\alpha$  and  $\beta$ , in DFA-MCD and DFA-SAFN. We selected adaptation scenarios from MNIST to USPS and from Amazon to DSLR for demonstration. The results are shown in Figure 10(a)(b). For each case study,  $\alpha$  and  $\beta$  were varied from 0.001 to 100. As shown in both figures, DFA can stably improve the performance of adversarial and non-adversarial UDA frameworks with different values of  $\alpha$  and  $\beta$ .



**Figure 10:** Sensitivity analysis of the hyper-parameters  $\alpha$  and  $\beta$  for DFA-MCD and DFA-SAFN (orange lines indicate DFA-SAFN; blue lines indicate DFA-MCD).  $\alpha$  was set to 0.1 when evaluating  $\beta$ .  $\beta$  was set to 10 when evaluating  $\alpha$ .

#### 5.5. Computational Complexity Analysis

We investigated the computational efficiency of our model as it could be combined with other UDA frameworks. We conducted a case study on the adaptation scenario from SVHN to MNIST. Although the time spent on training one epoch for DFA-MCD is 1.21 times MCD (NVIDIA GeForce RTX 2070), DFA-MCD requires fewer epochs to converge, as shown in Figure 11. Therefore, we can say that our model can efficiently improve the performance of various UDA frameworks.



**Figure 11:** Relationship between the training epoch and the accuracy (orange line indicates the proposed DFA-MCD; blue line indicates MCD).

#### 6. Conclusion

In this paper, we introduced a novel model for UDA to better align the source and the target features, which could improve the adaptation performance of the UDA framework. We proposed an indirect latent alignment process to encourage the features of the two domains to be constructed on a common feature space, i.e., the space of the Gaussian prior. To better align two distributions, we also proposed a novel unpaired L1-distance in the decoder space, and empirically confirmed that it served as a distribution alignment mechanism. Our frameworks outperformed state-of-the-arts in most experiments. The results of the extensive experiments have validated the importance and the versatility of our research.



## References

- [1] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W., 2010. A theory of learning from different domains. *Machine Learning* 79, 151–175.
- [2] Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B., Smola, A.J., 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22, e49–e57.
- [3] Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D., 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In the *IEEE Conference on Computer Vision and Pattern Recognition*, 95–104.
- [4] Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D., 2016. Domain separation networks. *Advances in Neural Information Processing Systems*, 343–351.
- [5] Cao, Z., Long, M., Wang, J., Jordan, M.I., 2018. Partial transfer learning with selective adversarial networks. In the *IEEE Conference on Computer Vision and Pattern Recognition*, 2724–2732.
- [6] Carlucci, F.M., Porzi, L., Caputo, B., Ricci, E., Bulò, S.R., 2017. Autodial: Automatic domain alignment layers. In the *IEEE Conference on Computer Vision and Pattern Recognition*, 5067–5075.
- [7] Chang, W.G., You, T., Seo, S., Kwak, S., Han, B., 2019. Domain-specific batch normalization for unsupervised domain adaptation. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7354–7362.
- [8] Chen, C., Xie, W., Huang, W., Rong, Y., Ding, X., Huang, Y., Xu, T., Huang, J., 2019. Progressive feature alignment for unsupervised domain adaptation. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 627–636.
- [9] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In the *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.
- [10] Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 1180–1189.
- [11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2672–2680.
- [12] Grandvalet, Y., Bengio, Y., 2005. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 529–536.
- [13] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In *proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- [14] Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T., 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *proceedings of the 35th International Conference on Machine Learning*.
- [15] Hull, J., 1994. A dataset for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 550–554.
- [16] Kim, M., Sahu, P., Gholami, B., Pavlovic, V., 2019. Unsupervised visual domain adaptation: A deep max-margin gaussian process approach. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4380–4390.
- [17] Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [18] Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- [19] Kurmi, V.K., Kumar, S., Nambodiri, V.P., 2019. Attending to discriminative certainty for domain adaptation. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 491–500.
- [20] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient based learning applied to document recognition. In *proceeding of the IEEE*, 2278–2324.
- [21] Li, Y., Wang, N., Shi, J., Liu, J., Hou, X., 2016. Revisiting batch normalization for practical domain adaptation. *arXiv:1603.04779*.
- [22] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: *European conference on computer vision*, Springer. pp. 740–755.
- [23] Liu, H., Long, M., Wang, J., Jordan, M., 2019. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, 4013–4022.
- [24] Long, M., Cao, Y., Wang, J., Jordan, M.I., 2015. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, 97–105.
- [25] Long, M., Cao, Z., Wang, J., Jordan, M., 2018. Conditional adversarial domain adaptation. *Advances in Neural Information Processing Systems*, 1640–1650.
- [26] Long, M., Zhu, H., Wang, J., Jordan, M.I., 2017. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning*, 2208–2217.
- [27] Maaten, L.V.D., Hinton, G., 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2579–2605.
- [28] Moiseev, B., Konev, A., Chigorin, A., Konushin, A., 2013. Evaluation of traffic sign recognition methods trained on synthetically generated data. In *International Conference on Advanced Concepts for Intelligent Vision Systems*.
- [29] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A., 2011. Reading digits in neural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.
- [30] Pan, S.J., Yang, Q., 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 1345–1359.
- [31] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research* 12, 2825–2830.
- [32] Pei, Z., Cao, Z., Long, M., Wang, J., 2018. Multi-adversarial domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [33] Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., Saenko, K., 2017. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*.
- [34] Roy, S., Sirohin, A., Sangineto, E., Bulò, S.R., Sebe, N., Ricci, E., 2019. Unsupervised domain adaptation using feature-whitening and consensus loss. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9471–9480.
- [35] Saenko, K., Kulis, B., Fritz, M., Darrell, T., 2010. Adapting visual category models to new domains. In *European Conference on Computer Vision*, 213–226.
- [36] Saito, K., Ushiku, Y., Harada, T., 2017. Asymmetric tri-training for unsupervised domain adaptation. In *proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2988–2997.
- [37] Saito, K., Watanabe, K., Ushiku, Y., Harada, T., 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In the *IEEE Conference on Computer Vision and Pattern Recognition*, 3723–3732.
- [38] Sankaranarayanan, S., Balaji, Y., Castillo, C.D., Chellappa, R., 2018. Generate to adapt: Aligning domains using generative adversarial networks. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8503–8512.
- [39] Stallkamp, J., Schlipsing, M., Saleman, J., Igel, C., 2011. The german traffic sign recognition benchmark: A multi-class classification competition. In *International Joint Conference on Neural Networks*.
- [40] Sun, B., Feng, J., Saenko, K., 2016. Return of frustratingly easy domain adaptation. In the *Thirtieth AAAI Conference on Artificial Intelligence*.
- [41] Sun, B., Saenko, K., 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision Workshops*, 443–450.
- [42] Torralba, A., Efros, A.A., 2011. Unbiased look at dataset bias. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1521–1528.

- [43] Tzeng, E., Hoffman, J., Saenko, K., Darrell, T., 2017. Adversarial discriminative domain adaptation. In the IEEE Conference on Computer Vision and Pattern Recognition , 7167–7176.
- [44] Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S., 2017. Deep hashing network for unsupervised domain adaptation. In the IEEE Conference on Computer Vision and Pattern Recognition , 5385–5394.
- [45] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* 11, 3371–3408.
- [46] Wang, X., Jin, Y., Long, M., Wang, J., Jordan, M.I., 2019. Transferable normalization: Towards improving transferability of deep neural networks. *Advances in Neural Information Processing Systems* , 1951–1961.
- [47] Xie, S., Zheng, Z., Chen, L., Chen, C., 2018. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning* , 5423–5432.
- [48] Xu, R., Li, G., Yang, J., Lin, L., 2019. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In the *IEEE International Conference on Computer Vision* .
- [49] You, K., Wang, X., Long, M., Jordan, M., 2019. Towards accurate model selection in deep unsupervised domain adaptation. In *International Conference on Machine Learning* , 7124–7133.
- [50] Zhang, W., Ouyang, W., Li, W., Xu, D., 2018. Collaborative and adversarial network for unsupervised domain adaptation. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* , 3801–3809.