Mind the Discriminability: Asymmetric Adversarial Domain Adaptation

Jianfei Yang[1][0000−0002−8075−0439], Han Zou[2], Yuxun Zhou[2], Zhaoyang Zeng[3], and Lihua Xie[1] (✉)

[1] Nanyang Technological University, Singapore
{yang0478,elhxie}@ntu.edu.sg
[2] University of California, Berkeley, USA
[3] Sun Yat-sen University, China

Abstract. Adversarial domain adaptation has made tremendous success by learning domain-invariant feature representations. However, conventional adversarial training pushes two domains together and brings uncertainty to feature learning, which deteriorates the discriminability in the target domain. In this paper, we tackle this problem by designing a simple yet effective scheme, namely Asymmetric Adversarial Domain Adaptation (AADA). We notice that source features preserve great feature discriminability due to full supervision, and therefore a novel asymmetric training scheme is designed to keep the source features fixed and encourage the target features approaching to the source features, which best preserves the feature discriminability learned from source labeled data. This is achieved by an autoencoder-based domain discriminator that only embeds the source domain, while the feature extractor learns to deceive the autoencoder by embedding the target domain. Theoretical justifications corroborate that our method minimizes the domain discrepancy and spectral analysis is employed to quantize the improved feature discriminability. Extensive experiments on several benchmarks validate that our method outperforms existing adversarial domain adaptation methods significantly and demonstrates robustness with respect to hyper-parameter sensitivity.

Keywords: adversarial domain adaptation; asymmetric training;

1 Introduction

Learning robust representations from large-scale labeled datasets, deep neural networks have achieved huge success in kinds of applications, such as visual recognition and neural language processing [16, 7]. Nevertheless, well-trained deep models are sensitive to cross-domain distribution shift (*domain shift*) that exists when applying them to a new domain, which usually requires tremendous efforts on annotating new labels. To render data-hungry model strong representation ability like data-adequate model, domain adaptation is proposed to transfer the knowledge from a label-rich domain (*source domain*) to a label-scarce or unlabeled domain (*target domain*) [26]. It alleviates the negative effect of *domain shift* in transfer learning and reduces the manual overhead for labeling.

Prevailing domain adaptation [25] tackles the problem of *domain shift* by enhancing the transferability of feature learning, *i.e.* aligning the marginal distributions across domains. For deep neural networks, aligning the deep features mainly relies on two categories of approaches. One category reduces the distribution discrepancy by measuring the statistics [36, 33], which is simple to implement and usually possesses stable convergence. Another category is adversarial domain adaptation, inspired by the Generative Adversarial Network (GAN) [12]. A binary domain discriminator is introduced to distinguish the domain labels while the feature extractor learns to fool the discriminator [9]. Adversarial domain adaptation approaches have achieved prominent performance on many challenging tasks including semantic segmentation [38], 3D estimation [43], sentiment classification [18] and wireless sensing [44, 41].

Though adversarial domain adaptation methods yield superior performance, they bring "uncertainty" to feature learning [5]. Such uncertainty is due to the side effect of domain adversarial training (DAT). Specifically, any features that can deceive the domain discriminator and perform well in the source domain conform with the goal of DAT. The worst consequence could be overfitting to the source domain and generating meaningless features in the target domain as long as they can fool the domain discriminator. Therefore, this uncertainty in adversarial training is severely detrimental to learning discriminative features in the target domain. This explains why many subsequent works aim to preserve semantic information [39] or adjust the boundary of classifier during DAT [29, 32]. These solutions ameliorate the traditional DAT yet by adding more learning steps, which either increases the computational overhead or requires a sophisticated hyper-parameter tuning process for multiple objectives.

In this paper, we address the uncertainty problem by proposing an Asymmetric Adversarial Domain Adaptation (AADA). The key problem of DAT consists in the symmetric objective which is to equally push two domains as close as possible, which deteriorates feature learning and neglects the decision boundary. Inspired by the fact that the source domain has great discriminability due to full supervision, AADA aims to fix the source domain and only adapts the target domain. To this end, we design an autoencoder serving as a domain discriminator to embed the source features, while a feature extractor is trained to deceive it — to embed the target features. Such adversarial process can be realized by a minimax game with a margin loss. As shown in Fig 1, DAT employs a binary domain discriminator to align two domains together, while we only force the target domain to approach to the source domain that possesses good discriminability. Furthermore, it is acknowledged that the autoencoder is an energy function that learns to map the observed sample to the low-energy space [17]. Therefore, energy function can cluster similar data to form a high density manifold, which helps to preserve more semantic information. We leverage the energy function to fix the source domain by associating lower energies to it while pushing the target domain to the low-energy space via adversarial training in an innovative manner. *The proposed method is a novel and new fundamental domain alignment tech-*

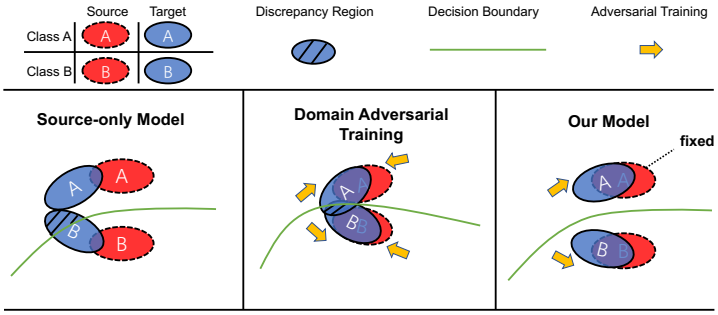


Fig. 1: Comparison between previous Domain Adversarial Training (DAT) method and ours. **Left:** The discrepancy region is large before adaptation. **Middle:** DAT based on a domain classifier aligns two domains together and pushes them as close as possible, which hurts the feature discriminability in the target domain. **Right:** AADA fixes the source domain that is regarded as the low-energy space, and pushes the target domain to approach the source one, which makes use of the well-trained classifier in the labeled source domain.

nique which can be easily integrated with other domain adaptation approaches. Our contribution in this paper is threefold:

- We propose a novel asymmetric adversarial scheme that replaces the conventional domain classifier with an autoencoder, which incorporates only the target domain into the adversarial feature training, circumventing the loss of discriminability in traditional domain adversarial training.
- The autoencoder is an energy function that maps the two domains to the low-energy space, which encourages the feature clusters to be tight and thus further benefits the classification task in an unsupervised manner.
- AADA is a generic domain alignment approach that can be used as an ingredient in existing domain adaptation approaches. The experiment validates that AADA outperforms other domain alignment approaches significantly. We further demonstrate the boosted discriminability by spectral analysis.

The paper is organized as follows. We firstly revisit adversarial domain adaptation while highlighting its limitations in Section 2. Then the AADA method is detailed in Section 3. Section 4 demonstrates the effectiveness and superiority of AADA. Section 5 compares AADA with other relevant approaches and Section 6 concludes the paper.

2 Revisiting Transferability and Discriminability in Symmetric Adversarial Domain Adaptation

We first revisit the learning theory of unsupervised domain adaptation (UDA) [1] where we analyze how the target expected error $\epsilon_{\mathcal{T}}(h)$ is connected to transferability and discriminability in representation learning. Then we highlight the

problem of forfeiting discriminability due to symmetric adversarial training, and excogitate the insights of our approach that conceals behind.

2.1 The Theory of Domain Adaptation

In unsupervised domain adaptation, we have access to N_s labeled examples from a source domain $\mathcal{D}_S = \{\mathbf{x}_i^s, y_i^s | \mathbf{x}_i^s \in \mathbf{X}_s, y_i^s \in Y_s\}$ and N_t unlabeled examples from a target domain $\mathcal{D}_T = \{x_i^t | \mathbf{x}_i^t \in \mathbf{X}_t\}$, which are sampled from distinct distributions $\mathbb{P}$ and $\mathbb{Q}$, respectively. The objective of UDA is to learn a model that performs well for the target domain. The learning theory of UDA was proposed by Ben-David [1].

Theorem 1. *Let $\mathcal{H}$ be the common hypothesis class for source and target. Let ϵ_s and ϵ_t be the source and target generalization error functions, respectively. The expected error for the target domain is upper bounded as*

$$\epsilon_t(h) \leq \epsilon_s(h) + \frac{1}{2} d_{\mathcal{H}\triangle\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda, \forall h \in \mathcal{H}, \tag{1}$$

where $d_{\mathcal{H}\triangle\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 2 \sup_{h_1, h_2 \in \mathcal{H}} |\Pr_{x\sim\mathcal{D}_S}[h_1(x) \neq h_2(x)] - \Pr_{x\sim\mathcal{D}_T}[h_1(x) \neq h_2(x)]|$ and $\lambda = \min_h[\epsilon_s(h^*) + \epsilon_t(h^*)]$.

As source data is annotated, the source error $\epsilon_s(h)$ can be simply minimized via supervised learning. To minimize $\epsilon_t(h)$, UDA focuses on reducing the domain discrepancy term $\frac{1}{2} d_{\mathcal{H}\triangle\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ and the ideal risk λ. In representation learning for UDA, minimizing the domain discrepancy is able to improve **transferability** of features. This can be achieved by domain adversarial training [9] or the minimization of statistical measures of such discrepancy [36]. Another criterion that plays a vital role in feature representations is **discriminability**. It refers to the capacity of clustering in the feature manifold, and therefore controls the easiness of separating categories. For UDA, we pursue good discriminability in both source and target domains simultaneously. As we typically use a shared feature extractor for two domains, enhancing discriminability is equivalent to seeking for a better ideal joint hypothesis $h^* = \min_h[\epsilon_s(h) + \epsilon_t(h)]$ [5].

2.2 Limitations and Insights

From the analyses above, the feature learning of UDA should guarantee both transferability and discriminability, which inspires us to investigate the existing domain alignment methods from these two perspectives. Adversarial domain adaptation methods have shown prominent performance and become increasingly popular. Ganin *et al.* [9] pioneered it by proposing Domain Adversarial Neural Network (**DANN**) that learns domain-invariant features using the reverse gradients from a domain classifier. Typically, adversarial UDA approaches consist of a shared feature extractor $\mathbf{f} = G_f(x)$, a label predictor $\mathbf{y} = G_y(x)$ and a domain discriminator $\mathbf{d} = G_d(x)$. Apart from standard supervised learning on the source domain, a minimax game between $\mathbf{f}$ and $\mathbf{d}$ is designed. The domain

discriminator **d** is trained to distinguish the domain label between the source domain and the target domain, while the feature extractor **f** learns to deceive the domain classifier **d**. In this manner, the domain adversarial training enables the model to learn transferable features across domains when the Nash Equilibrium is achieved. The whole process can be formulated as

$$\min_{G_f, G_y} \mathcal{L}_c(\mathbf{X}_s, Y_s) - \gamma \mathcal{L}_c(\mathbf{X}_s, \mathbf{X}_t), \tag{2}$$

$$\min_{G_d} \mathcal{L}_c(\mathbf{X}_s, \mathbf{X}_t), \tag{3}$$

where $\mathcal{L}_c$ is the classification loss such as cross-entropy loss, and the hyper-parameter γ decides the importance of transferability in feature learning.

Symmetric DAT does improve the transferability across domains but sacrifices the discriminability in the target domain. Let us analyze the objectives of training feature extractor **f** that are two-folds: (1) good discriminability in the source domain and (2) learning representations that are indistinguishable to the domain discriminator. There is no constraint on the discriminability in the target domain. As depicted in Fig 1, DAT is symmetric and makes two domains as close as possible. Theoretically, the worst case is that the feature extractor generates meaningless representations on the target domain, as long as they can deceive the domain classifier. Hence, a good decision boundary on the source domain cannot perform well on the target domain. Previous works quantified the discriminability by spectral analysis and drew a similar conclusion [5, 20].

We believe that *the decreasing discriminability is caused by symmetric adversarial training that involves both source domain and target domain in adversarial feature learning* as shown in the second term of Eq(2). Specifically, to deceive the binary domain discriminator, DAT aims to push two domains close. In this process, DAT cannot control how the domains are aligned and cannot guarantee whether the decision boundary separates the categorical clusters in the target domain. This motivates us to fix the source domain and only render the target domain to approach to the source, as depicted in Fig 1. In this fashion, the feature discriminability is preserved and a good classifier is easily obtained. To achieve symmetric adversarial training, we innovatively propose to leverage the autoencoder as the domain classifier.

3 Asymmetric Adversarial Domain Adaptation

Maintaining the source manifolds during DAT is not a trivial task with the consideration of the complexity of network architecture. This requires us to design a simple yet effective asymmetric adversarial mechanism. In this section, Asymmetric Adversarial Domain Adaptation (AADA) is detailed and we theoretically justify how it reduces domain discrepancy.

3.1 The Learning Framework

As shown in Fig 2, our model consists of a shared feature extractor G_f parameterized by θ_f, a label predictor G_y parameterized by θ_y and an autoencoder

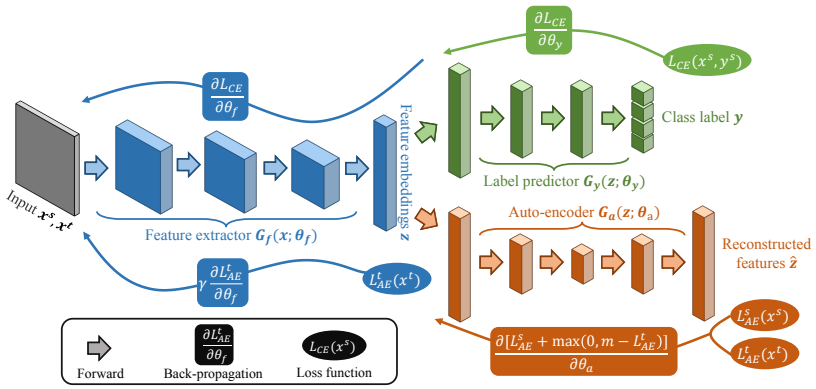


Fig. 2: AADA model constitutes a shared feature extract G_f, a classifier G_y and an autoencoder G_a. Except for supervised learning on the source domain, the autoencoder plays a domain discriminator role that learns to embed the source features and push the target features away, while the feature extractor learns to generate target features that can deceive the autoencoder. Such process is an asymmetric adversarial game that pushes the target domain to the source domain in the feature space.

G_a parameterized by θ_a. The feature extractor G_f, typically composed of multiple convolutional layers, embeds an input sample to a feature embedding z, and then the label predictor G_y that usually consists of several fully connected layers maps the feature embedding to the predicted label $\hat{y}$. The autoencoder G_a reconstructs an embedding z to $\hat{z}$.

In the learning phase, the first objective of the model is to learn feature discriminability in the source domain. As we have access to the labeled source data $(\mathbf{X}_s, Y_s)$, it is simply achieved by minimizing the cross-entropy loss via back-propagation:

$$\begin{aligned}&\min_{G_f,G_y} \mathcal{L}_{CE}(\mathbf{X}_s, Y_s) = \\ &-\mathbb{E}_{(\mathbf{x}_s,y_s)\sim(\mathbf{X}_s,Y_s)} \sum_{n=1}^{N_s} [\mathbb{I}_{[l=y_s]} \log G_y(G_f(\mathbf{x}_s))].\end{aligned} \tag{4}$$

With robust feature learning in the source domain, the next objective is to learn transferable representations using the unlabeled data $(\mathbf{X}_t)$ in the target domain. To this end, we propose an asymmetric adversarial training scheme that involves an autoencoder G_a with a margin Mean Squared Error (MSE) loss. The autoencoder G_a, which plays a domain discriminator role, only learns to embed features from the source domain, but not to embed features from the target domain. The objective of the autoencoder-based domain discriminator is

formulated as:

$$\min_{G_a} \mathcal{L}_{AE}(\mathbf{X_s}) + \max(0, m - \mathcal{L}_{AE}(\mathbf{X_t})), \tag{5}$$

where m is the margin between two domains in the feature space. Here, the MSE loss of the autoencoder is defined as:

$$\mathcal{L}_{AE}(\mathbf{x}_i) = ||G_a(G_f(\mathbf{x}; \theta_f); \theta_a) - \mathbf{x}_i||_2^2, \tag{6}$$

where $|| \cdot ||_2^2$ denotes the squared L_2-norm. Such unsupervised loss introduces a cycle-consistent constraint that improves feature discriminability.

To play the adversarial game, the feature extractor G_f learns to fool the autoencoder G_a by generating source-like features for the samples in the target domain. When the feature extractor succeeds, the representations of the target domain can inherit good discriminability from the source domain, and the label predictor G_y trained in the source domain applies equally. The adversarial training of the feature extractor G_f is formulated by:

$$\min_{G_f} \mathcal{L}_{AE}(\mathbf{X_t}). \tag{7}$$

The overall optimization of the proposed AADA model is formally defined by:

$$\begin{aligned} &\min_{G_f, G_y} \mathcal{L}_{CE}(\mathbf{X}_s, Y_s) + \gamma \mathcal{L}_{AE}(\mathbf{X_t}), \\ &\min_{G_a} \mathcal{L}_{AE}(\mathbf{X_s}) + \max(0, m - \mathcal{L}_{AE}(\mathbf{X_t})), \end{aligned} \tag{8}$$

where γ is a hyper-parameter that controls the importance of transferability. In this fashion, our approach only incorporates the $\mathcal{L}_{AE}(\mathbf{X_t})$ term into the training of G_f, which pushes the target domain to the source domain. Oppositely, the objective of G_a serves as a domain discriminator that pushes two domains away from a margin m.

3.2 Discussions and Theories

The autoencoder used in our approach is an energy function, which associates lower energies (*i.e.* MSE) to the observed samples in a binary classification problem [17]. For UDA, the autoencoder in our model associates low energies to the source features, and AADA compels the target features to approach to low-energy space. The design of such adversarial scheme is inspired by the Energy-based GAN which theoretically proves that using an energy function in GAN, the true distribution can be simulated by the generator at *Nash Equilibrium* [42]. Similarly, in AADA, the feature extractor G_f can mimic the source distribution for the samples in the target domain when the model achieves convergence.

As theoretically justified in [9], domain adversarial training using a domain classifier effectively reduces the domain discrepancy term $d_{\mathcal{H}\triangle\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ in **Theorem 1**. In our approach, we notice that the autoencoder can be treated as a form of a domain classifier with a margin. The MSE loss of the autoencoder

in our framework, *i.e.* $\mathcal{L}_{AE}(\cdot)$, works as the same way of the domain classifier. The training objective of $\mathcal{L}_{AE}(\cdot)$ is $\mathcal{L}_{AE}(\mathbf{X_s}) = 0$ and $\mathcal{L}_{AE}(\mathbf{X_t}) = m$, which is equivalent to the functionality of the standard domain classifier. Therefore, following the theorem proved in [9], the proposed autoencoder can maximize the domain divergence while the adversarial feature learning minimizes the divergence by deceiving the autoencoder. Moreover, as the autoencoder memorizes more domain information than a binary classifier, it transfers more knowledge during asymmetric adversarial training.

4 Experiments

We evaluate the proposed AADA on three UDA benchmarks and compare our model with the prevailing approaches. Then we validate the improved feature discriminability by spectral analysis. In the discussions, we verify our motivations by quantitatively analyzing the transferability and discriminability.

4.1 Experimental Setup

Digits [10]. We use five digits datasets *MNIST, MNIST-M, USPS, SVHN* and *Synthetic Digits* (*SYN-DIGIT*)), all of which consist of 32×32 images. We assess five types of adaptation scenarios with distinct levels of domain shift. We follow the experimental settings of **DANN** [9] that uses the official training splits in two domains for training and evaluates the model on the testing split.

Image-CLEF[4] is a domain adaptation dataset for ImageCLEF Challenge. It constitutes three domains: *Caltech-256* (**C**), *ImageNet ILSVRC 2012* (**I**) and *Pascal VOC 2012* (**P**), which form six transfer tasks.

Office-Home [37] consists of 65 categories in office and home settings, and has more than 15500 images. It is a very challenging dataset with four extremely distinct domains: *Artistic images* (**Ar**), *Clip Art* (**Cl**), *Product images* (**Pr**) and *Real-World images* (**Rw**), which forms 12 transfer tasks. For Image-CLEF and Office-Home, we employ the full training protocol in [22] that employs all images from the source domain and the target domain for training.

Baselines. We compare our AADA with the state-of-the-art UDA methods: **MMD** [36], Deep Adaptation Network (**DAN**) [22], Deep **CORAL** [33], **DANN** [9], Self-ensembling **SE** [8], Deep Reconstruction Classification Network (**DRCN**) [11], Domain Separate Network (**DSN**) [3], **ADDA** [35], **CoGAN** [21], **CyCADA** [14], Maximum Classifier Discrepancy (**MCD**) [29], Conditional Domain Adversarial Network (**CDAN**) [23], Batch Spectral Penalization (**BSP**) [5], **CCN** [15], **GTA** [31] and **MCS** [19]. To further prove that our method can be integrated with other UDA methods to achieve better adaptation, we integrate it with the Constrained Clustering Network (**CCN**) [15] that employs self-training to improve feature transferability.

Implementation Details. For the task with 32×32 images, we use the same network architecture as DANN, while for Image-CLEF and Office-Home,

[4] http://imageclef.org/2014/adaptation

Table 1: Accuracy (%) of domain adaptation tasks on *Digit*.

Source Target	MNIST USPS	USPS MNIST	SVHN MNIST	SYN-DIGIT SVHN	MNIST MNIST-M
Source-only	78.2	63.4	54.9	86.7	56.3
Domain Alignment Methods					
MMD [36]	81.1	-	71.1	88.0	76.9
CORAL [33]	80.7	-	63.1	85.2	57.7
Adversarial Training based Methods					
DANN [9]	85.1	73.0	74.7	90.3	76.8
CoGAN [21]	91.2	89.1	-	-	62.0
ADDA [35]	89.4	90.1	76.0	-	-
CDAN [23]	93.9	96.9	88.5	-	-
GTA [31]	95.3	90.8	92.4	-	-
CyCADA [14]	95.6	96.5	90.4	-	-
Other State-of-the-Art Methods					
DRCN [11]	91.8	73.7	82.0	87.5	68.3
DSN [3]	91.3	-	82.7	91.2	83.2
MCD [29]	96.5	94.1	96.2	-	-
BSP+DANN [5]	94.5	97.7	89.4	-	-
BSP+CDAN [5]	95.0	98.1	92.1	-	-
MCS+GTA [19]	97.8	98.2	91.7	-	-
AADA_{opt}	95.6±0.3	92.7±0.5	74.8±0.8	88.8±0.2	47.1±1.2
AADA (Ours)	**98.4±0.3**	**98.6±0.3**	**98.1±0.5**	**92.2±0.4**	**95.5±0.2**

we use **ResNet-50** with pretrained parameters on ImageNet [6]. In AADA, we use an autoencoder with only fully connected layers, and the detailed network architectures are in the appendix. We use Adam optimizer with the constant learning rate $\mu = 5e^{-4}$ for digit adaptation, and SGD with the decaying learning rate in DANN for object recognition. For hyperparameter m and γ, we set $m = 0.5, \gamma = 1e^{-2}$ for 32×32 images, and $m = 1, \gamma = 1e^{-1}$ for object recognition datasets, which are empirically obtained by cross-validation on MNIST→USPS. The whole experiment is implemented by **PyTorch** framework.

4.2 Overall Results

We first compare our methods with MMD, CORAL, DANN, DAN and JAN that are only based on domain alignment. In Table 1, our approach shows significant improvement against the standard DAT (DANN), even outperforming the state-of-the-art methods that require much higher computation overhead such as CyCADA and GTA due to the training of cyclic or generative networks. For pixel-level domain shift in MNIST→MNIST-M, the traditional domain alignment methods such as MMD and DANN cannot effectively deal with them, but AADA achieves 95.5% accuracy. Moreover, for SVHN→MNIST with larger domain shift, AADA surpasses DANN by 23.4% and BSP+DANN by 8.7%. For more challenging tasks in Table 2 and 3, the proposed AADA surpasses all the methods that are purely based on domain alignment. AADA outperforms DANN

Table 2: Classification accuracy (%) on *Image-CLEF* (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
DAN [22]	74.5	82.2	92.8	86.3	69.2	89.8	82.5
DANN [9]	75.0	86.0	96.2	87.0	74.3	91.5	85.0
CCN [15]	77.1	87.5	94.0	86.0	74.5	91.7	85.1
CDAN [23]	76.7	90.6	**97.0**	90.5	74.5	93.5	87.1
AADA	78.0	90.3	94.0	87.8	75.2	93.5	86.5
AADA+CCN	**79.2**	**92.5**	96.2	**91.4**	**76.1**	**94.7**	**88.4**

Table 3: Classification accuracy (%) 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
DAN [22]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN [9]	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
CCN [15]	47.3	65.2	70.1	51.3	60.5	60.9	48.1	45.5	71.3	65.1	53.5	77.0	59.7
SE [8]	48.8	61.8	72.8	54.1	63.2	65.1	50.6	49.2	72.3	66.1	55.9	78.7	61.5
CDAN [23]	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
BSP+DANN [5]	51.4	68.3	75.9	56.0	67.8	68.8	57.0	49.6	75.8	70.4	57.1	80.6	64.9
AADA	52.3	69.5	76.3	59.7	68.2	70.2	58.2	48.9	75.9	69.1	54.3	80.5	65.3
AADA+CCN	**54.0**	**71.3**	**77.5**	**60.8**	**70.8**	**71.2**	**59.1**	**51.8**	76.9	**71.0**	**57.4**	**81.8**	**67.0**

by 1.5% on Image-CLEF and 7.7% on OfficeHome. Since the domain shift is small in Image-CLEF, the improvement margin is not large.

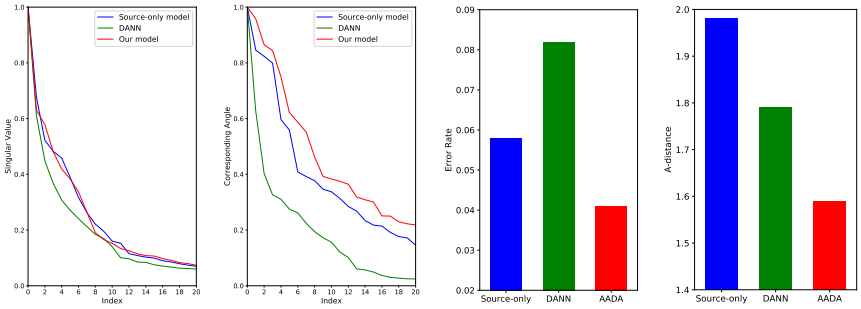
AADA is a generic domain alignment approach that can be integrated to other novel UDA frameworks, achieving better performance. We integrate AADA with CCN [15] and evaluate it on challenging tasks. As shown in Table 2 and 3, the proposed AADA+CCN achieves the state-of-the-art accuracies, outperforming CDAN and BSP that also aim to improve discriminability. We can see that AADA improves CCN by 2.9% on Image-CLEF and 7.3% on Office-Home, which implies that AADA can bring much improvement to existing UDA methods by preserving more discriminability in domain alignment.

4.3 Spectral Analysis using SVD

The intuition and the theoretical analysis validate that the proposed model achieves a good trade-off between transferability and discriminability. Here we employ the quantitative method to further demonstrate it. The prior research proposes to apply Singular Value Decomposition (SVD) to the representation z, and then infer transferability and discriminability by the Singular Values (SV) and the Corresponding Angles (CA) of eigenvectors, respectively [5]. Motivated by this, we conducted an experiment on a digit adaptation task SVHN→MNIST. Using the feature extractor G_f, we obtained the target feature matrix $\mathbf{F_t} = [\mathbf{f_t^1} \dots \mathbf{f_t^b}]$ where b is the batch size. Then we apply SVD to the target feature matrix as follows:

$$\mathbf{F_t} = \mathbf{U_t}\mathbf{\Sigma_t}\mathbf{V_t^T}, \tag{9}$$

where $\mathbf{\Sigma_t}$ denotes the eigenvalue, $\mathbf{U_t}$ denotes the eigenvector and $\mathbf{V_t}$ is an unitary matrix. SVD is also applied to the source feature matrix to obtain $\mathbf{\Sigma_s}$ and $\mathbf{U_s}$. In



(a) SV (b) CA (c) Hypothesis λ (d) A-distance d_A

Fig. 3: Measures of discriminability and transferability.

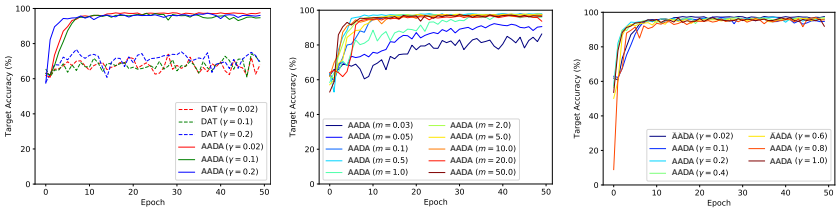
Fig 3(a), we plot the top-20 normalized singular values $\mathbf{\Sigma_t}$ *w.r.t* three models including the source-only model, DANN and our AADA model. It is observed that the largest singular value of the DANN target feature matrix is greater than the other values, which impairs the semantic information included in other smaller singular values. In comparison, the distribution of singular values for AADA is similar to that for the source-only model that preserves more discriminability in feature learning. In Fig 3(b), we show the normalized corresponding angles of singular values. The corresponding angle depicts the commonality between the source eigenvectors $\mathbf{U_s}$ and the target eigenvectors $\mathbf{U_t}$, which indicates the transferability of the features. For DANN, the sharp distribution of the angles indicates that DANN only utilizes several peak transferable features. This deteriorates the informative representation in the target domain. Whereas, AADA obtains more transferable features that also show better discriminability, which has good repercussion for learning a common decision boundary.

4.4 Analytics and Discussions

Opposite Direction of AADA. The proposed AADA fixes the source domain and then learns to force the target domain to approach to it, which utilizes the good classifier of the source domain. What if we fix the target domain and force the source domain to approach to it? We denote this situation as the opposite form of AADA, namely AADA_{opt}. The optimization procedure of AADA_{opt} is written as:

$$\begin{aligned} &\min_{G_f, G_y} \mathcal{L}_{CE}(\mathbf{X}_s, Y_s) + \gamma \mathcal{L}_{AE}(\mathbf{X_s}), \\ &\min_{G_a} \mathcal{L}_{AE}(\mathbf{X_t}) + \max(0, m - \mathcal{L}_{AE}(\mathbf{X_s})). \end{aligned} \tag{10}$$

Intuitively, as the decision boundary of the target domain keeps changing during training, this may weaken the discriminability in domain adversarial training, which is similar to DANN. The result in Table 1 proves our analysis. AADA_{opt}



(a) Comparison (b) Sensitivity of m (c) Sensitivity of γ

Fig. 4: The training procedures *w.r.t.* the hyper-parameters m and γ.

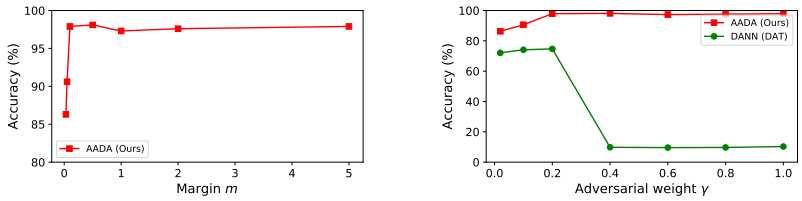
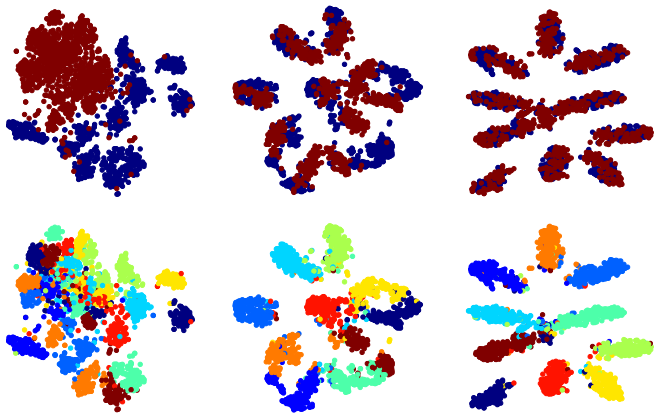


Fig. 5: Accuracy by varying m (left) and γ (right).

only produces similar results as DANN, which implies that it is infeasible to align two domains and the boundaries of classifier simultaneously.

Sensitivity Study. To demonstrate that AADA is not sensitive to the hyperparameters, we conduct the experiments on SVHN→MNIST across multiple m and γ. Fig 5 illustrates the sensitivity results in terms of the margin m. As the margin increases, the accuracy increases until 0.5, which conforms with our intuition. When the margin is small, the degree of transferability is limited. If the margin is greater than 0.1, we observe that the results are very stable in Fig 4(b). As to γ, we compare our approach with Domain Adversarial Training (DAT) that is the reproduced DANN where γ is the weight of the adversarial loss in Eq 2. In Fig 5 and Fig 4(c), the results of AADA are robust given large γ but DANN cannot converge with large γ. In Fig 4(a), we can observe that the target accuracies of DAT fluctuate a lot during training while AADA provides more stable training procedure. The sensitivity study is consistent with the insights provided earlier, and moreover the results show more robustness *w.r.t* training procedure, which makes it easier to apply our method in an unsupervised manner.

Ideal Joint Hypothesis. We estimate the ideal joint hypothesis λ in **Theorem 1** to show the **discriminability** of feature embeddings. To this end, we train an MLP classifier on all source and target data with labels. As shown in Fig 3(c), AADA has the lowest λ in the feature space on task SVHN→MNIST. This demonstrates that AADA preserves more feature discriminability in two domains by asymmetric and cycle-consistent objectives. It helps learn a good decision boundary that separates the data from two domains.



(a) Source-only (b) DANN (c) AADA

Fig. 6: The t-SNE visualization of the embeddings z on task **SVHN→MNIST**. The top figures are visualization with domain labels (blue: source, red: target). The bottom figures are visualization with category labels (10 classes).

Distribution Discrepancy. As proposed in the theory of domain adaptation [1], the A-distance is a measure of domain discrepancy that quantifies the **transferability** of feature embeddings. It is defined as $d_A = 2(1 - 2\epsilon)$ where ϵ is the error of a domain classifier. We train an MLP classifier to discriminate source and target domains on task SVHN→MNIST. Results are shown in Fig 3(d), and it is observed that AADA has better transferability than DANN.

Visualization Representation. We visualize learned features on the *Digit* task SVHN→MNIST via t-SNE [24] and present it in Fig 6. The visualization validates the insights of AADA. DANN aligns the features of two domains, but we can see that due to the lack of discriminability, some categories of data are confused. Hence many approaches proposed to adjust the boundary for DANN features [29]. In comparison, the AADA features are domain-invariant and preserve the good decision boundary of the source domain simultaneously. This further proves that AADA learns better discriminability in domain alignment.

5 Related Work

Domain adaptation tackles the problem of *domain shift* in statistical learning [26]. Massive works on UDA were developed recently, and here we discuss and compare the related domain adaptation progress.

Adversarial Domain Adaptation. Inspired by GAN [12], adversarial domain adaptation (ADA) methods yield remarkable results by learning representations that cannot be distinguished by a domain discriminator [9, 34, 35, 32]. ADA can act on both feature-level and pixel-level alignment [2, 14]. Adversarial training can also maximize classifier discrepancy to learn adapted classifier

[29]. ADA is generalized to new scenarios including partial adversarial domain adaptation [4] and open set domain adaptation [30]. As such, ADA becomes a necessary ingredient for many subsequent UDA approaches.

Enhancing Discriminability in UDA. Though ADA methods contribute to learning domain-invariant features, adversarial learning hinders the feature discriminability in the target domain [5]. Conditional adversarial domain adaptation captures the cross-covariance between features and predictions to improve the discriminability [23]. More methods propose to learn semantic features by clustering and self-training [39]. BSP penalizes the largest singular values of features [5], and Xu *et al.* [40] propose that larger feature norm boosts the discriminability. These methods effectively increase the discriminability by adding extra regularization terms or building self-training algorithms, which can lead to much complexity or more difficult hyper-parameter tuning process.

Asymmetric Training. Asymmetric training means an unequal training process for multiple networks or parts of networks. Satio *et al.* pioneered an asymmetric tri-training for domain adaptation, which leverages two networks for generating pseudo labels and one network to learn target representations [28]. Asymmetric training between two feature extractors is developed in ADDA with untied sharing weights [35, 45]. Our approach proposes an asymmetric training between the two players of adversarial game.

Autoencoder in Domain Adaptation. Autoencoder can learn representations in an unsupervised manner, and it has been directly utilized for learning target domain in DRCN [11] and DSN [3]. It also enables cyclic methods such as CyCADA [14]. All these methods use auto-encoder to learn target domain features in a straightforward way. In AADA, we employ autoencoder as a domain classifier which empowers asymmetric adversarial training and hence improve the discriminability. From the perspective of energy-based model [17], autoencoder is an energy function [27, 42] that maps the correct variables to low energies. In this paper, it is designed to assign low energies to the source domain, and encourage the target domain to approach to the low-energy space.

6 Conclusion

In this paper, we propose a novel asymmetric adversarial regime for unsupervised domain adaptation. As the conventional adversarial UDA methods affect the discriminability while improving the transferability of feature representations, our method aims to preserve the discriminability by encouraging the target domain to approach to the source domain in the feature space, which is achieved by an autoencoder with an asymmetric adversarial training scheme. Spectral analysis is utilized to justify the improved discriminability and transferability. The experimental results demonstrate its robustness and superiority on several public UDA datasets.

References

1. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**(1-2), 151–175 (2010)
2. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3722–3731 (2017)
3. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D.: Domain separation networks. In: Advances in Neural Information Processing Systems. pp. 343–351 (2016)
4. Cao, Z., Ma, L., Long, M., Wang, J.: Partial adversarial domain adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 135–150 (2018)
5. Chen, X., Wang, S., Long, M., Wang, J.: Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 1081–1090. PMLR, Long Beach, California, USA (09–15 Jun 2019)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. French, G., Mackiewicz, M., Fisher, M.: Self-ensembling for visual domain adaptation. In: International Conference on Learning Representations. No. 6 (2018)
9. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 1180–1189. PMLR, Lille, France (07–09 Jul 2015)
10. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research **17**(1), 2096–2030 (2016)
11. Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W.: Deep reconstruction-classification networks for unsupervised domain adaptation. In: European Conference on Computer Vision. pp. 597–613. Springer (2016)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680 (2014)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: CyCADA: Cycle-consistent adversarial domain adaptation. In: Proceedings of the 35th International Conference on Machine Learning. pp. 1989–1998 (2018)
15. Hsu, Y.C., Lv, Z., Kira, Z.: Learning to cluster in order to transfer across domains and tasks. In: International Conference on Learning Representations (ICLR) (2018)
16. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436 (2015)

17. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F.: A tutorial on energy-based learning. Predicting structured data **1**(0) (2006)
18. Li, Z., Zhang, Y., Wei, Y., Wu, Y., Yang, Q.: End-to-end adversarial memory network for cross-domain sentiment classification. In: IJCAI. pp. 2237–2243 (2017)
19. Liang, J., He, R., Sun, Z., Tan, T.: Distant supervised centroid shift: A simple and efficient approach to visual domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2975–2984 (2019)
20. Liu, H., Long, M., Wang, J., Jordan, M.: Transferable adversarial training: A general approach to adapting deep classifiers. In: International Conference on Machine Learning. pp. 4013–4022 (2019)
21. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: Advances in Neural Information Processing Systems. pp. 469–477 (2016)
22. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37. pp. 97–105. ICML'15 (2015)
23. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: Advances in Neural Information Processing Systems (2018)
24. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
25. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks **22**(2), 199–210 (2011)
26. Pan, S.J., Yang, Q., et al.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2010)
27. Ranzato, M., Boureau, Y.L., Chopra, S., LeCun, Y.: A unified energy-based framework for unsupervised learning. In: Artificial Intelligence and Statistics. pp. 371–379 (2007)
28. Saito, K., Ushiku, Y., Harada, T.: Asymmetric tri-training for unsupervised domain adaptation. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 2988–2997. JMLR. org (2017)
29. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3723–3732 (2018)
30. Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open set domain adaptation by backpropagation. In: The European Conference on Computer Vision (ECCV) (September 2018)
31. Sankaranarayanan, S., Balaji, Y., Castillo, C.D., Chellappa, R.: Generate to adapt: Aligning domains using generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8503–8512 (2018)
32. Shu, R., Bui, H.H., Narui, H., Ermon, S.: A dirt-t approach to unsupervised domain adaptation. In: Proc. 6th International Conference on Learning Representations (2018)
33. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: European Conference on Computer Vision. pp. 443–450. Springer (2016)
34. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4068–4076 (2015)
35. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Computer Vision and Pattern Recognition (CVPR). vol. 1, p. 4 (2017)

36. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. CoRR **abs/1412.3474** (2014), http://arxiv.org/abs/1412.3474
37. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proc. CVPR. pp. 5018–5027 (2017)
38. Vu, T.H., Jain, H., Bucher, M., Cord, M., Perez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
39. Xie, S., Zheng, Z., Chen, L., Chen, C.: Learning semantic representations for unsupervised domain adaptation. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 5423–5432. PMLR, Stockholmsmässan, Stockholm Sweden (10–15 Jul 2018)
40. Xu, R., Li, G., Yang, J., Lin, L.: Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1426–1435 (2019)
41. Yang, J., Zou, H., Cao, S., Chen, Z., Xie, L.: Mobileda: Towards edge domain adaptation. IEEE Internet of Things Journal (2020)
42. Zhao, J., Mathieu, M., LeCun, Y.: Energy-based generative adversarial network. In: Proc. 5th International Conference on Learning Representations (2017)
43. Zhou, X., Karpur, A., Gan, C., Luo, L., Huang, Q.: Unsupervised domain adaptation for 3d keypoint estimation via view consistency. In: ECCV. pp. 137–153 (2018)
44. Zou, H., Yang, J., Zhou, Y., Xie, L., Spanos, C.J.: Robust wifi-enabled device-free gesture recognition via unsupervised adversarial domain adaptation. In: 2018 27th International Conference on Computer Communication and Networks (ICCCN). pp. 1–8. IEEE (2018)
45. Zou, H., Zhou, Y., Yang, J., Liu, H., Das, H.P., Spanos, C.J.: Consensus adversarial domain adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 5997–6004 (2019)