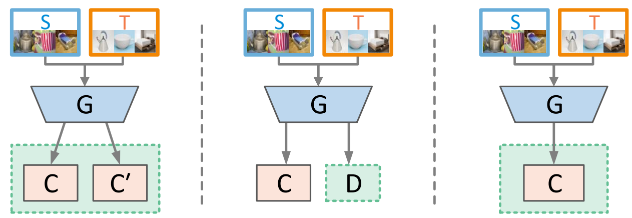
Reusing the Task-specific Classifier as a Discriminator: Discriminator-free Adversarial Domain Adaptation

Lin Chen* Huaian Chen* Zhixiang Wei Xin Jin Xiao Tan Yi Jin† Enhong Chen
University of Science and Technology of China
{chlin, anchen, zhixiangwei, jinxustc, tx2015}@mail.ustc.edu.cn
{jinyi08, cheneh}@ustc.edu.cn

Abstract

Adversarial learning has achieved remarkable performances for unsupervised domain adaptation (UDA). Existing adversarial UDA methods typically adopt an additional discriminator to play the min-max game with a feature extractor. However, most of these methods failed to effectively leverage the predicted discriminative information, and thus cause mode collapse for generator. In this work, we address this problem from a different perspective and design a simple yet effective adversarial paradigm in the form of a discriminator-free adversarial learning network (DALN), wherein the category classifier is reused as a discriminator, which achieves explicit domain alignment and category distinguishment through a unified objective, enabling the DALN to leverage the predicted discriminative information for sufficient feature alignment. Basically, we introduce a Nuclear-norm Wasserstein discrepancy (NWD) that has definite guidance meaning for performing discrimination. Such NWD can be coupled with the classifier to serve as a discriminator satisfying the K-Lipschitz constraint without the requirements of additional weight clipping or gradient penalty strategy. Without bells and whistles, DALN compares favorably against the existing state-of-the-art (SOTA) methods on a variety of public datasets. Moreover, as a plug-and-play technique, NWD can be directly used as a generic regularizer to benefit existing UDA algorithms. Code is available at https://github.com/xiaoachen98/DALN.



(a) Bi-classifier (b) Extra Discriminator (c) Ours

Figure 1. Illustration of different adversarial paradigms, in which G, C, and D denote the feature extractor, task-specific classifier, and discriminator, respectively. Different from typical paradigms that adopt an (a) additional classifier C' (called bi-classifier) or (b) additional discriminator D, we present a different perspective for UDA and introduce a simple but effective adversarial paradigm illustrated in (c), in which the original task-specific classifier C is reused as a implicit discriminator, achieving explicit domain alignment and category distinguishment via a unified objective.

1. Introduction

Deep neural networks (DNNs) have achieved a significant progress in many computer vision tasks [4, 5, 16, 37]. However, the success of these methods highly depends on large amounts of annotated data [13, 47, 51], which is extremely time-consuming and expensive to obtain. Moreover, due to the discrepancy [31, 32] between training data and real-world testing data, the DNN model trained on annotated data may suffer from a dramatic performance decline in testing set despite extensive annotation efforts. To address this problem, unsupervised domain adaptation (UDA) [6, 9, 30, 48], which aims to transfer knowledge from a labeled source domain to an unlabeled target domain in the presence of a domain shift, has been deeply explored.

Inspired by the theoretical analysis of Ben-David *et al.* [2], the existing UDA methods usually explore the idea of learning domain-invariant feature representations. Generally, these methods can be categorized into two branches, i.e., moment matching methods [20, 24, 25, 43, 49] and adversarial learning methods [11, 12, 25, 39]. Moment matching methods explicitly reduce the domain shift by matching a well-defined distribution discrepancy of the source and target domain features. Adversarial learning methods implicitly mitigate the domain shift by playing an adversarial min-max two-player game, which drives the generator to extract indistinguishable features to fool the discriminator.

*indicates equal contribution.
†Corresponding author.

Encouraged by the remarkable performance achieved by adversarial learning, increasingly more researchers have been devoted to developing a UDA method based on an adversarial paradigm [9, 10, 21, 23, 27, 42].

Basically, adversarial learning-based UDA methods usually follow two lines of adversarial paradigms. One line [10, 19, 21, 27, 39] leverages the disparity of two task-specific classifiers C and C' (as shown in Fig. 1(a)), which can be deemed as a discriminator, to implicitly achieve adversarial learning and improve feature transferability. This paradigm enables UDA methods to reduce the class-level domain discrepancy. However, the methods following this paradigm are prone to be affected by ambiguous predictions and thus hinder the adaption optimization. The other line [9, 11, 12, 25] directly constructs an additional domain discriminator D as shown in Fig. 1(b), which improves the feature transferability by sufficiently confusing the cross-domain feature representations. However, the methods following this paradigm usually focus on the domain-level feature confusion, which may hurt the category-level information and thus cause mode collapse problem [18, 42].

To address these problems, we present a different perspective for UDA and introduce a simple but effective adversarial paradigm illustrated in Fig. 1(c). In this paradigm, the original task-specific classifier is coupled with a novel discrepancy to serve as a discriminator/critic, which simultaneously achieves domain alignment and category distinguishment through a unified objective, enabling the model to leverage the predicted discriminative information to capture the multi-modal structures [12, 25] of the feature distributions. Particularly, when classifier C is used for classification, it helps achieve category-level distinguishment; furthermore, when C serves as a discriminator, it achieves feature-level alignment. The novel discrepancy, called Nuclear-norm Wasserstein discrepancy (NWD), leverages the advantages of the Nuclear norm and 1-Wasserstein distance to encourage the prediction determinacy and diversity. Different from the discrepancy metrics used in existing adversarial methods [11, 42, 50], the NWD not only has a promising theoretical generalization bound but also has definite guidance meaning for performing discrimination, i.e., naturally giving high scores to the source domain samples and low scores to the target domain samples due to the supervised training on the source domain. Such guidance encourages the intra-class and inter-class correlations of the target domain to approach those of the source domain. Moreover, in contrast to the existing Wasserstein discrepancy used in recent work [40], the NWD enables the adversarial UDA paradigm to satisfy the K-Lipschitz constraint without the need to set up an additional weight clipping [1] or gradient penalty [15].

Based on the introduced paradigm, we propose a discriminator-free adversarial learning network (DALN), which achieves adversarial UDA classification without explicit domain discriminator. Benefiting from the definite guidance of the NWD, the proposed DALN converges rapidly and achieves competitive prediction determinacy and diversity. Note that, the DALN is considerably different from recent approaches [42, 50] that integrate the discriminator into the classifier. DALN directly reuses the original task-specific classifier without requiring any additional components, making it quite simple and efficient. Extensive experiments on a variety of datasets demonstrate that the proposed DALN outperforms the existing state-of-the-art (SOTA) methods. Moreover, we show that the proposed NWD is general and plug-and-play, which can be used as a regularizer to benefit the existing methods, which helps them achieve more competitive performance. The main contributions of this work are summarized as follows:

- We present a different perspective for UDA by introducing a simple yet effective adversarial paradigm, in which the original task-specific classifier is reused as a discriminator. Based on this, we propose a new UDA method, namely DALN, which can leverage the predicted discriminative information for sufficient feature alignment.
- We introduce a new discrepancy, termed NWD, which has a theoretical generalization bound and definite guidance meaning. Such discrepancy enables the implicitly constructed discriminator to satisfy the K-Lipschitz constraint without the requirements of additional weight clipping and gradient penalty strategies.
- Without bells and whistles but only a few lines of code, the proposed method achieves highly competitive performance on various public datasets. By taking the proposed NWD as a regularizer for existing methods, these methods can achieve more competitive performance.

2. Related Works

The existing UDA methods can be mainly divided into two categories, i.e., moment matching methods [24, 26, 43, 49] and adversarial learning methods [10, 11, 25, 39, 50].

Moment Matching Methods. Moment matching methods learn domain-invariant feature representations by matching a well-defined moment-based distribution discrepancy [51] across domains. Typically, DDC [43] attempted to explicitly align the learned feature distributions across domains by minimizing the maximum mean discrepancy (MMD). Later, methods in [24, 26] improved DDC by performing alignment with multi-kernel maximum mean discrepancy (MK-MMD) and joint maximum mean discrepancy (JMMD), respectively. In addition, MDD [49] proposed margin disparity discrepancy (MDD) to reduce the distribution discrepancy.

Adversarial Learning Methods. Inspired by generative adversarial network (GAN) [14], adversarial learning methods learn domain-invariant features via a min-max two-player

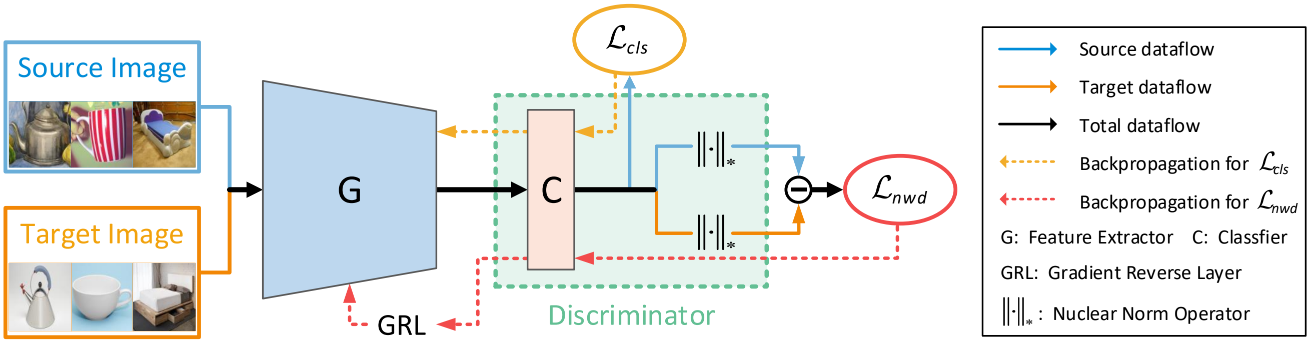


Figure 2. An overview of the adversarial paradigm in the form of DALN, which consists of a feature extractor G and a task-specific classifier C. $\mathcal{L}_{cls}$ is used to guarantee a low source risk for the source domain, and $\mathcal{L}_{nwd}$ is used to empirically estimate the NWD that can be coupled with classifier C to implicitly serve as a discriminator. The gradient reverse layer is used to help perform the adversarial learning.

game. As one of the earliest attempts, DANN [11] introduced an additional discriminator to distinguish the features generated by the feature extractor, which successfully achieves the domain-level adaptation. The success of the DANN exhibits the ability to improve UDA with the GAN model. Later, FGDA [12] leveraged a discriminator to distinguish the gradient distribution of features, which achieved better performance for reducing domain discrepancy. Inspired by the conditional GAN [29], methods in [25, 34] combined the predicted discriminative information with learned features to improve feature alignment. Additionally, DADA [42] attempted to couple the task-specific classifier with the domain discriminator to align the joint distributions of two domains. Although these methods successfully learn domain-invariant features, they cannot guarantee an appropriate divergence used for the discriminator when the support sets of two distributions do not overlap with each other [1].

In addition to the methods adopting an additional discriminator, some studies attempted to use two task-specific classifiers (called bi-classifier), in which the disparity of two task-specific classifiers can be deemed as a discriminator [10, 19, 27, 39, 50], to implicitly achieve the adversarial learning. Representatively, MCD [39] simply used the L1 distance to measure the intra-class divergence of two classifiers. SWD [19] proposed using sliced Wasserstein discrepancy instead of L1 distance to obtain a more geometrically meaningful intra-class divergence. CGDM [10] additionally introduced the cross-domain gradient discrepancy to further alleviate the domain discrepancy. Although these methods have achieved considerable improvements in reducing domain discrepancy, most of them consider only the intra-class divergence between predictions, which may result in ambiguous predictions.

Different from the aforementioned methods adopting an additional discriminator or classifier, we reuse the original task-specific classifier by coupling it with the designed NWD, implicitly constructing a discriminator/critic satisfying the K-Lipschitz constraint without the requirements of additional weight clipping or gradient penalty strategy.

3. Method

3.1. Recap of Preliminary Knowledge

Given a labeled source domain set $\{(x_i^s, y_i^s)\}_{i=1}^{N^s}$ with N_s samples drawn from source domain $\mathcal{D}_\mathcal{S}$, where $x_i^s \in \mathcal{X}_s$, $y_i^s \in \mathcal{Y}_s$, and label y^s covers k classes, and an unlabeled domain target set $\{x_i^t\}_{i=1}^{N^t}$ with N_t samples drawn from target domain $\mathcal{D}_\mathcal{T}$, where $x_i^t \in \mathcal{X}_t$, the goal of this work is to learn a deep UDA model for learning domain-invariant representations and achieving reliable predictions on the target domain. This model consists of a feature generator $G(\cdot)$ that maps the input data to the features $f \in \mathbb{R}^d$, i.e., $f^s = G(x^s)$ and $f^t = G(x^t)$, and a task-specific classifier $C(\cdot)$ that generates corresponding predictions $p \in \mathbb{R}^k$, i.e., $p^s = C(f^s)$ and $p^t = C(f^t)$. To this end, the existing adversarial UDA approaches usually take an additional discriminator or classifier. Typically, many popular methods [11, 25] use an additional discriminator $D(\cdot)$ to achieve adversarial UDA by optimizing object classification loss $\mathcal{L}_{cls}$ and domain adversarial loss $\mathcal{L}_{adv}$:

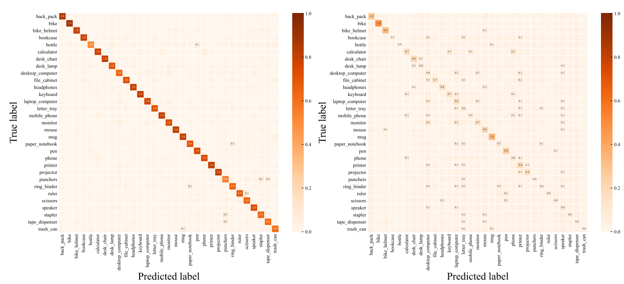
$$\mathcal{L}_{cls} = \mathbb{E}_{(x_i^s, y_i^s)\sim\mathcal{D}_\mathcal{S}} \mathcal{L}_{\text{ce}}\left(C\left(G\left(x_i^s\right)\right), y_i^s\right), \tag{1}$$

$$\begin{aligned}\mathcal{L}_{adv} = &\ \mathbb{E}_{G(x_i^s)\sim\tilde{\mathcal{D}}_s} \log\left[D\left(G\left(x_i^s\right)\right)\right] \\ &+ \mathbb{E}_{G(x_i^t)\sim\tilde{\mathcal{D}}_t} \log\left[1 - D\left(G\left(x_i^t\right)\right)\right],\end{aligned} \tag{2}$$

where $\tilde{\mathcal{D}}_s$ and $\tilde{\mathcal{D}}_t$ denote the induced feature distributions of $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$, respectively, and $\mathcal{L}_{\text{ce}}(\cdot, \cdot)$ is the cross-entropy loss function. However, we find that the original task-specific classifier C has an implicit discriminative ability for the source domain and target domain, and can be directly used as a discriminator (see Sec. 3.2). Inspired by this observation, as shown in Fig. 2, we propose a simple yet effective adversarial paradigm for adversarial UDA: reusing the task-specific classifier as a discriminator.

3.2. Reusing the Classifier as a Discriminator

Motivation Re-clarification. As we claimed before, the original task-specific classifier has an implicit discriminative ability for the source domain and the target domain. Fig. 3 presents the self-correlation matrices of the predictions on the source and target domains based on a model trained with the source-only data. For the source domain, benefiting from the supervised training, the values of the self-correlation matrix are concentrated on the main diagonal. In contrast, for the target domain, the prediction generates larger values on the off-diagonal elements due to the lack of supervision. Therefore, the intra-class and inter-class correlations represented in the self-correlation matrix are capable of constructing the adversarial critic.



(a) Source domain (b) Target domain

Figure 3. The self-correlation matrices of the predictions on the source and target domains based on a DNN model trained only with the source domain data on task A→W of Office-31. (Zoom in for a clear visualization.)

Rethinking the Intra-class and Inter-class Correlations. Given a prediction matrix $Z \in \mathbb{R}^{b\times k}$ predicted by C that contains the prediction probabilities of k categories multiplied by b samples, the self-correlation matrix $R \in \mathbb{R}^{k\times k}$ can be calculated by $R = Z^T Z$, where the prediction matrix $Z = C(f)$ satisfies

$$\sum_{j=1}^{k} Z_{i,j} = 1 \quad \forall i \in 1 \dots b$$
$$Z_{i,j} \geq 0 \quad \forall i \in i \dots b, j \in 1 \dots k. \tag{3}$$

For a self-correlation matrix R, the main diagonal elements represent the intra-class correlation and the off-diagonal elements denote the inter-class correlation or confusion [17]. For convenient, in this work, we define the overall intra-class correlation as I_a and the overall inter-class correlation as I_e:

$$I_a = \sum_{i,j=1}^{k} R_{ij} \quad I_e = \sum_{i\neq j}^{k} R_{ij}. \tag{4}$$

For the source domain, the prediction contributes to a large I_a and a small I_e; while for the target domain, the prediction generally produces a relatively small I_a and large I_e due to the lack of supervised training. Thus, $I_a - I_e$ can be used to represent the domain discrepancy. According to Equation 3, I_a and I_e satisfy $I_a + I_e = b$. Meanwhile, I_a is equal to the Frobenius norm of prediction matrix Z, i.e, $I_a = \|Z\|_F$. Thus, we have $I_a - I_e = 2\|Z\|_F - b$. Z is predicted via the classifier C, so we can use $2\|C\|_F - b$ as a correlation critic function, which naturally gives high scores for the source domain samples and low scores for the target domain samples due to the supervised training on source domain. Moreover, considering weight 2 and bias b are both constants, the $\|C\|_F$ can be directly used as a correlation critic function.

From Correlations Critic to 1-Wasserstein Distance. Inspired by the WGAN [1], a straightforward idea is to introduce an additional discriminator D to learn a K-Lipschitz critic function h expected to give high scores to source representations $f \in \tilde{\mathcal{D}}_s$ and low scores to the target representations $f \in \tilde{\mathcal{D}}_t$, and measure the 1-Wasserstein distance $W_1\left(\tilde{\mathcal{D}}_s, \tilde{\mathcal{D}}_t\right)$ between two feature distributions $\tilde{\mathcal{D}}_s$, $\tilde{\mathcal{D}}_t$ by

$$W_1\left(\tilde{\mathcal{D}}_s, \tilde{\mathcal{D}}_t\right) = \sup_{\|h\|_L \leq K} \mathbb{E}_{f\sim\tilde{\mathcal{D}}_s}\left[h\left(f\right)\right] - \mathbb{E}_{f\sim\tilde{\mathcal{D}}_t}\left[h\left(f\right)\right], \tag{5}$$

where $\|\cdot\|_L$ denotes the Lipschitz semi-norm [46], and K denotes the Lipschitz constant. But, as we claimed above, $\|C\|_F$ has exactly definite critic meaning to serve as D. Then, the domain discrepancy can be written as

$$W_F = \sup_{\left\|\|C\|_F\right\|_L \leq K} \mathbb{E}_{\tilde{\mathcal{D}}_s}\left[\|C\left(f\right)\|_F\right] - \mathbb{E}_{\tilde{\mathcal{D}}_t}\left[\|C\left(f\right)\|_F\right], \tag{6}$$

where W_F is short for $W_F\left(\tilde{\mathcal{D}}_s, \tilde{\mathcal{D}}_t\right)$, which denotes the Frobenius norm-based 1-Wasserstein distance of two domain distributions. In this way, we can achieve explicit domain alignment and category distinguishment through a unified objective, contributing to leveraging the predicted discriminative information for capturing the multi-modal structures of the feature distributions.

3.3. Adversarial Learning with the NWD

From Frobenius Norm to Nuclear Norm. The constructed discriminator/critic $D = \|C\|_F$ can perform adversarial training with the generator G, which helps achieve transferable and discriminative representations while improving the prediction determinacy. However, adversarial learning based on the Frobenius-norm 1-Wasserstein distance may reduce the prediction diversity because it tends to push the category with a small number of samples to the neighbouring category containing large amounts of samples far from the decision boundary [8]. Inspired by recent works on the Nuclear norm [7, 8, 36, 41], which has been demonstrated to be bound with the Frobenius norm, we attempt to replace the Frobenius norm $\|\cdot\|_F$ with the Nuclear norm $\|\cdot\|_*$ because maximizing $\|Z\|_*$ means maximizing the rank of Z when the $\|\cdot\|_F$ is nearby $\sqrt{b}$ [7, 8], which improves the prediction diversity. Therefore, the domain discrepancy can be rewritten as

$$W_N = \sup_{\left\|\|C\|_*\right\|_L \leq K} \mathbb{E}_{\tilde{\mathcal{D}}_s}\left[\|C(f)\|_*\right] - \mathbb{E}_{\tilde{\mathcal{D}}_t}\left[\|C(f)\|_*\right], \tag{7}$$

where W_N is short for $W_N\left(\tilde{\mathcal{D}}_s, \tilde{\mathcal{D}}_t\right)$, which denotes the Nuclear-norm 1-Wasserstein discrepancy (NWD) of two domain distributions. Then, our discriminator can be rewritten as $D = \|C\|_*$. When classifier C is used for classification, it helps achieve category-level distinguishment, but when C serves as a discriminator, it achieves feature-level alignment. Note that our classifier consists of a fully connected layer and a softmax activation function. It can be demonstrated that all the components of our implicit discriminator satisfy the K-Lipschitz constraint (**see supplementary material for the proof**), which enables the proposed model to be trained without the requirements of additional weight clipping and gradient penalty strategies. Therefore, we can approximately estimate the empirical NWD $\hat{W}_N$ by maximizing the domain critic loss $\mathcal{L}_{nwd}$:

$$\mathcal{L}_{nwd}\left(x^s, x^t\right) = \frac{1}{N_s}\sum_{i=1}^{N_s} D\left(G\left(x_i^s\right)\right) - \frac{1}{N_t}\sum_{j=1}^{N_t} D\left(G\left(x_i^t\right)\right), \tag{8}$$

$$\hat{W}_N = \max_D \mathcal{L}_{nwd}\left(x^s, x^t\right). \tag{9}$$

Adversarial Learning for DALN. In this work, we build a DALN consisting of a generator G based on a pretrained ResNet and a classifier C constructed with a fully connected layer and a softmax layer. To avoid tedious alternating updates for the DALN, a gradient reverse layer (GRL) [11], which does not include the above mentioned gradient penalty or weight clipping, is used to help achieve updating within one back propagation. In this way, DALN can be trained by playing the min-max game as

$$\min_G \max_C \mathcal{L}_{nwd}\left(x^s, x^t\right). \tag{10}$$

Moreover, to ensure the fidelity of UDA classification, we need to guarantee a low source risk for the source domain. Therefore, generator G and classifier C should also be optimized by minimizing the supervised classification loss $\mathcal{L}_{cls}$ for the source domain as

$$\mathcal{L}_{cls}\left(x^s, y^s\right) = \frac{1}{N_s}\sum_{i=1}^{n_s} \mathcal{L}_{ce}\left(C\left(G\left(x_i^s\right)\right), y_i^s\right). \tag{11}$$

In short, the overall loss used to optimize the classification model can be written as

$$\min_{C,G}\left\{\mathcal{L}_{cls}\left(x^s, y^s\right) + \lambda \max_C \mathcal{L}_{nwd}\left(x^s, x^t\right)\right\}, \tag{12}$$

where λ is used to balance $\mathcal{L}_{cls}$ and $\mathcal{L}_{nwd}$. In this work, λ is set to 1. With the help of adversarial learning, the DALN learns transferable and discriminative representations while promising the prediction determinacy and diversity.

Generalization Bound. Here, we present the theoretical guarantees for the proposed method. Following [2], we consider a binary classification instance. Then, let $\mathcal{F}$ $(f \in \mathcal{F})$ denote a fixed representation space and $C : \mathcal{F} \rightarrow [0, 1]$ be a family of source classifiers, where C belongs to hypothesis space $\mathcal{H}$. We assume that the risk of C on the source domain is described as $\varepsilon_s(C) = \mathbb{E}_{f\sim\tilde{\mathcal{D}}_s}\left[C(f) \neq y\right]$, where $\tilde{\mathcal{D}}_s$ is the feature distribution induced by the data distribution of source domain $\mathcal{D}_S$ and y is the label corresponding to the induced feature f. Moreover, given two classifiers $C_1, C_2 \in \mathcal{H}$, we define the risk of these two classifiers on the source domain as $\varepsilon_s(C_1, C_2) = \mathbb{E}_{f\sim\tilde{\mathcal{D}}_s}\left[C_1(f) \neq C_2(f)\right]$. In the same way, we define the risk on the target domain, i.e, $\varepsilon_t(C)$ and $\varepsilon_t(C_1, C_2)$. Then, the ideal joint hypothesis is written as $C^* = \arg\min_C \varepsilon_s(C) + \varepsilon_t(C)$, which can be used to minimize the combined risk on the source and target domains. Therefore, according to [2], the probabilistic bound of $\varepsilon_t(C)$ can be written as

$$\varepsilon_t(C) \leq \varepsilon_s(C) + \left|\varepsilon_s(C, C^*) - \varepsilon_t(C, C^*)\right| + \eta^*, \tag{13}$$

where $\eta^* = \varepsilon_s(C^*) + \varepsilon_t(C^*)$ is a sufficiently small constant representing the ideal combined risk. Thus, the goal of UDA classification is to reduce the domain discrepancy term $\left|\varepsilon_s(C, C^*) - \varepsilon_t(C, C^*)\right|$.

Lemma 1. *Let $\nu_s, \nu_t \in \mathcal{P}(\mathcal{F})$ denote the probability measures of the source and target domain features, $\rho\left(f^s, f^t\right)$ be the cost of transporting a unit of material from location f^s satisfying $f^s \sim \nu_s$ to location f^t satisfying $f^t \sim \nu_t$, $W_1(\nu_s, \nu_t)$ represent the NWD, and K denote a Lipschitz constant. Given a family of classifiers $C \in \mathcal{H}_1$ and a ideal classifier $C^* \in \mathcal{H}_1$ satisfying the K-Lipschitz constraint, where $\mathcal{H}_1$ is a subspace of $\mathcal{H}$, the following holds for every $C, C^* \in \mathcal{H}_1$.*

$$\left|\varepsilon_s(C, C^*) - \varepsilon_t(C, C^*)\right| \leqslant 2KW_1(\nu_s, \nu_t). \tag{14}$$

Theorem 1. *Based on Lemma 1, for every $C \in \mathcal{H}_1$, the following holds*

$$\varepsilon_t(C) \leq \varepsilon_s(C) + 2KW_1(\nu_s, \nu_t) + \eta^*, \tag{15}$$

where $\eta^* = \varepsilon_s(C^*) + \varepsilon_t(C^*)$ is the risk of ideal joint hypothesis, which is a sufficiently small constant.

Table 1. Classification accuracy (%) on (a) Office-Home and (b) VisDA-2017 for unsupervised domain adaptation (using ResNet-50 and ResNet-101 as the backbone, respectively). † denotes that the results are reproduced using the publicly released code. The best accuracy is indicated in **bold red** and the second best accuracy is indicated in undelined blue. See **supplementary material** for more details.

(a) Office-Home.

Method	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg
ResNet-50 [16]	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
WDGRL†(18) [40]	44.1	63.8	74.0	47.3	57.1	61.7	51.8	39.1	72.1	64.9	45.9	76.5	58.2
MCD(18) [39]	48.9	68.3	74.6	61.3	67.6	68.8	57.0	47.1	75.1	69.1	52.2	79.6	64.1
BSP(19) [6]	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	59.3	81.9	66.3
BNM(20) [7]	52.3	73.9	80.0	63.3	72.9	74.9	61.7	49.5	79.7	70.5	53.6	82.2	67.9
GVB-GD(20) [9]	57.0	74.7	79.8	64.6	74.1	74.6	65.2	55.1	81.0	**74.6**	59.7	84.3	70.4
FGDA(21) [12]	52.3	77.0	78.2	64.6	75.5	73.7	64.0	49.5	80.7	70.1	52.3	81.6	68.3
TSA(21) [22]	53.6	75.1	78.3	64.4	73.7	72.5	62.3	49.4	77.5	72.2	58.8	82.1	68.3
CKB-MMD(21) [28]	54.2	74.1	77.5	64.6	72.2	71.0	64.5	53.4	78.7	72.6	58.4	82.8	68.7
SCDA(21) [23]	57.5	76.9	80.3	65.7	74.9	74.5	65.5	53.6	79.8	74.5	59.6	83.7	70.5
MetaAlign(21) [48]	**59.3**	76.0	80.2	65.7	74.7	75.1	65.7	**56.5**	81.6	74.1	61.1	85.2	71.3
DALN(Ours)	57.8	**79.9**	82.0	66.3	76.2	77.2	66.7	55.5	81.3	73.5	60.4	85.3	71.8
DANN(16) [11]	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
DANN+**NWD**	51.8	63.3	73.9	56.6	66.1	68.6	59.3	54.6	79.0	70.5	**61.5**	80.4	65.5
CDAN(18) [25]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
CDAN+**NWD**	54.8	70.7	77.9	60.5	69.6	71.8	61.2	55.0	80.9	**74.6**	59.4	83.4	68.3
MDD(19) [49]	54.9	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	60.2	82.3	68.1
MDD+**NWD**	55.8	76.1	79.1	64.3	73.3	73.2	63.6	55.0	80.2	73.8	61.1	84.0	70.0
MCC(20) [17]	55.1	75.2	79.5	63.3	73.2	75.8	66.1	52.1	76.9	73.8	58.4	83.6	69.4
MCC+**NWD**	58.1	79.6	**83.7**	**67.7**	**77.9**	**78.7**	**66.8**	56.0	**81.9**	73.9	60.9	**86.1**	**72.6**

(b) VisDA-2017.

Method	Avg
ResNet-101 [16]	52.4
WDGRL†(18) [40]	61.3
MCD(18) [39]	71.9
BSP(19) [6]	75.9
SWD(19) [19]	76.4
BNM(20) [7]	70.4
GVB-GD†(20) [9]	77.2
DADA(20) [42]	79.8
TSA(21) [22]	78.6
SCDA†(21) [23]	79.7
DALN(Ours)	80.6
DANN(16) [11]	57.4
DANN+**NWD**	80.0(22.6↑)
CDAN(18) [25]	73.9
CDAN+**NWD**	81.4(7.5↑)
MDD†(19) [49]	76.8
MDD+**NWD**	82.0(5.2↑)
MCC(20) [17]	78.8
MCC+**NWD**	**83.7(4.9↑)**

Table 2. Classification accuracy (%) on (a) Office-31 and (b) ImageCLEF-2014 for unsupervised domain adaptation (using ResNet-50 as the backbone). † denotes that the results are reproduced using the publicly released code. The best accuracy is indicated in **bold red** and the second best is indicated in undelined blue.

(a) Office-31.

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
ResNet-50 [16]	68.4	96.7	99.3	68.9	62.5	60.7	76.1
WDGRL†(18) [40]	72.6	97.1	99.2	79.5	63.7	59.5	78.6
MCD(18) [39]	88.6	98.5	**100.0**	92.2	69.5	69.7	86.5
SWD(19) [19]	90.4	98.7	**100.0**	94.7	70.3	70.5	87.4
BNM(20) [7]	91.5	98.5	**100.0**	90.3	70.9	71.6	87.1
DADA(20) [42]	92.3	**99.2**	**100.0**	93.9	74.4	74.2	89.0
GVB-GD(20) [9]	94.8	98.7	**100.0**	95.0	73.4	73.7	89.3
FGDA(21) [12]	93.3	99.1	**100.0**	93.2	73.2	72.7	88.6
MetaAlign(21) [48]	93.0	98.6	**100.0**	94.5	75.0	73.6	89.2
TSA(21) [22]	94.8	99.1	**100.0**	92.6	74.9	74.4	89.3
SCDA(21) [23]	94.2	98.7	99.8	95.2	75.7	76.2	90.0
DALN(Ours)	95.2	99.1	**100.0**	**95.4**	76.4	**76.5**	**90.4**
DANN(16) [11]	82.0	96.9	99.1	79.7	68.2	67.4	82.2
DANN+**NWD**	92.1	98.2	**100.0**	84.7	74.5	73.0	87.1
CDAN(18) [25]	94.1	98.6	**100.0**	92.9	71.0	69.3	87.7
CDAN+**NWD**	93.7	98.5	**100.0**	91.0	74.4	73.0	88.4
MDD(19) [49]	94.5	98.4	**100.0**	93.5	74.6	72.2	88.9
MDD+**NWD**	**95.5**	98.7	**100.0**	94.9	**76.6**	74.0	90.0
MCC(20) [17]	**95.5**	98.6	**100.0**	94.4	72.9	74.9	89.4
MCC+**NWD**	**95.5**	98.7	**100.0**	**95.4**	75.0	75.1	90.0

(b) ImageCLEF-2014.

Method	I→P	P→I	I→C	C→I	C→P	P→C	Avg
ResNet-50 [16]	74.8	83.9	91.5	78.0	65.5	91.2	80.7
WDGRL†(18) [40]	76.8	87.0	91.7	87.2	75.2	90.3	84.7
MCD(18) [39]	77.3	89.2	92.7	88.2	71.0	92.3	85.1
SWD(19) [19]	78.1	89.6	95.2	89.3	73.4	92.8	86.4
BNM(20) [7]	77.2	91.2	96.2	91.7	75.7	96.7	88.1
GVB-GD†(20) [9]	78.2	92.7	96.5	91.5	78.2	95.0	88.7
DADA†(20) [42]	78.7	92.3	97.2	91.6	78.5	95.3	88.9
CKB-MMD(21) [28]	**80.7**	92.2	96.5	92.2	79.9	96.7	89.7
SCDA†(21) [23]	78.7	91.8	96.7	92.8	78.5	95.2	89.0
TSA†(21) [22]	78.6	92.8	97.0	92.8	79.0	95.2	89.2
DALN(Ours)	80.5	93.8	97.5	92.8	78.3	95.0	89.7
DANN(16) [11]	75.0	86.0	96.2	87.0	74.3	91.5	85.0
DANN+**NWD**	78.0	89.2	97.3	93.3	78.5	92.0	88.1
CDAN(18) [25]	77.7	90.7	97.7	91.3	74.2	94.3	87.7
CDAN+**NWD**	78.6	92.5	97.2	91.7	79.3	94.6	89.0
MDD†(19) [49]	77.3	90.2	96.8	89.5	77.6	94.2	87.6
MDD+**NWD**	78.9	91.7	97.5	91.7	78.9	95.4	89.0
MCC†(20) [17]	78.3	**94.5**	97.3	92.3	77.3	96.3	89.3
MCC+**NWD**	79.8	**94.5**	**98.0**	**94.2**	**80.0**	**97.5**	**90.7**

Therefore, the risk of the target domain can be bounded by the risk of the source domain and the introduced NWD, providing theoretical guarantees for the proposed approach. Limited by space, **all the proofs and more details about the empirical measure of the target risk are provided in the supplementary materials.**

3.4. Regularizer to Existing UDA Methods

The proposed NWD can be easily integrated into existing methods to improve the prediction determinacy and diversity. Specifically, a gradient reverse layer is first added to the original task-specific classifier. Subsequently, the task-specific classifier with the introduced NWD can serve as a discriminator, which performs adversarial learning with the feature extractor. Formally, assuming the original loss $\mathcal{L}_{ori}$ of the model written as $\mathcal{L}_{ori} = \mathcal{L}_{cls} + \mathcal{L}_{spe}$, where $\mathcal{L}_{cls}$ is the standard supervised classification loss as that of the proposed method and $\mathcal{L}_{spe}$ is the special loss used in these methods, the reconstructed loss $\mathcal{L}_{rec}$ can be described as

$$\mathcal{L}_{rec} = \mathcal{L}_{cls} + \mathcal{L}_{spe} + \gamma \mathcal{L}_{nwd}, \tag{16}$$

where γ is the balance weight. For convenience, in our experiments, the values of γ for all other methods are set to 0.01. The results of taking the proposed discrepancy as a regularizer to benefit other UDA algorithms are presented in the experiments.

4. Experiments

In this section, we evaluate the proposed DALN and compare it with the SOTA methods for UDA classification. Additionally, we evaluate the effectiveness of NWD as a regularizer to benefit existing methods including DANN [11], CDAN [25], MDD [49], and MCC [17].

4.1. Setup

Dataset Description. We use four datasets including Office-Home [45], Office-31 [38], ImageCLEF [3], and VisDA-2017 [35] to perform the comparison experiments. **Office-Home** is a large-scale dataset that includes 15500 images and 65 categories. This dataset has four extremely different domains, i.e., Art (A), Clip Art (C), Product (P), and Real-World (R). **VisDA-2017** is a large-scale synthetic-to-real dataset that has two domains (synthetic (S) and real (R)). The dataset contains over 280K images across 12 classes. **Office-31** has a total number of 4110 images, which includes three domains, i.e., Amazon (A), Webcam (W), and DSLR (D); and each domain contains 31 categories. **ImageCLEF** consists of three domains derived from three public datasets: Caltech (C), ImageNet (I), and Pascal (P). Each domain contains 12 categories and each category has 50 images.

Implementation Details. The proposed method is implemented based on the PyTorch [33] framework running on a GPU (Tesla-V100). The SGD optimizer is used to train the model with a moment of 0.9, a weight decay of 1e-3, a batch size of 36, and a cropped image size of 224×224. The initial learning rate of classifier C is set to 5e-3, which is 10 times larger than that of feature extractor G. **Other details can be found in the supplementary material.**

4.2. Comparison Results

Results on Office-Home are shown in Table 1(a). Compared with SOTA methods, the proposed method achieves dramatic improvements in terms of classification accuracy. Particularly, in the case of domains suffering from large shifts and extremely unbalanced classes, e.g., A→R and C→R, the proposed method achieves 2.9% and 2.2% improvements compared to the existing SOTA methods. Moreover, by integrating the proposed NWD into the DANN, the average accuracy is improved by 7.9%. Combining the proposed NWD with MCC, it achieves SOAT performance of 72.6%, attaining 3.2% improvements. Dramatic improvements are obviously exhibited in P→R, C→A, and C→P tasks. These obtained gains come from the paradigm leveraging the predicted discriminative information and the introduced NWD encouraging the prediction determinacy and diversity.

Results on VisDA-2017 are displayed in Table 1(b). Despite the tremendous domain shift existing in synthetic and real data, the DALN achieves an average accuracy of 80.6%, outperforming the existing SOTA methods. Combining the proposed NWD with other methods, the performances of these methods are substantially improved by 22.6%, 7.5%, 5.2%, and 4.9% for the DANN, CDAN, MDD, and MCC, respectively. In particular, with the help of the proposed NWD, MCC achieves SOTA accuracy of 83.7%, demonstrating the effectiveness of the proposed method.

Results on Office-31 are presented in Table 2(a). The proposed DALN achieves SOTA performances in five adaptation sub-tasks, and attains the best performance on the average accuracy. Particularly, compared with WDGRL [40], which uses 1-Wasserstein distance with an additional discriminator, the proposed DALN dramatically improves the average accuracy by 11.8%. Additionally, taking the NWD as a regularizer, the typical methods can be improved by at least 0.6%, and the average accuracy of DANN is even improved by 4.9%. We note that the improvements are evidently achieved in the task of adapting a domain with a small number of samples (e.g., D and W) to a domain (e.g., A) containing large amounts of samples. These results occur because the proposed method encourages the prediction determinacy and diversity, which are highly important in such cases.

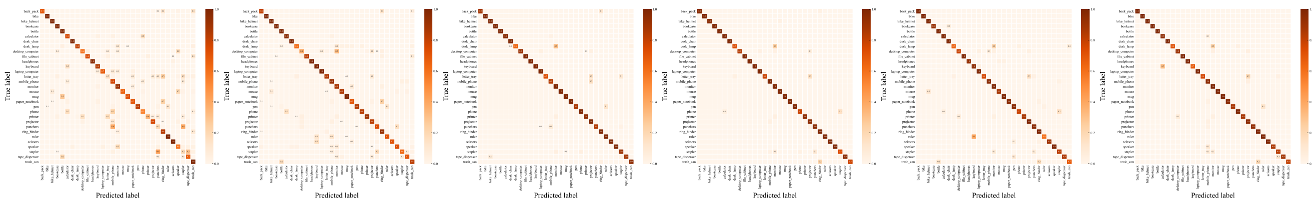
Results on ImageCLEF-2014 are provided in Table 2(b). Without bells and whistles, the proposed method achieves an average accuracy of 89.7%, which is same as the most competitive method CKB-MMD [28]. Moreover, the proposed method can be used as a regularizer, which is capable of improving the performance of the existing methods and thus contributes to SOTA performances. Specifically, the proposed method enables MCC to compare favorably against the SOTA methods. These results demonstrate that the proposed method is still highly effective in the case of all the domains containing the same samples and categories.

4.3. Insight Analysis

Here, we provide insight analyses for the proposed DALN and NWD. Limited by space, more detailed analyses regarding toy experiment, Proxy $\mathcal{A}$ distance, self-correlation matrix, convergence, and trade-off parameters (λ and γ) are provided in **supplementary material.**

Confusion Matrix. The comparison of the confusion matrices is shown in Fig. 4. The figure shows that the model trained on the source-only data suffers from severe class confusion. The DANN focuses on domain-level feature adaptation, but ignores feature discriminability, which results in misclassification for some categories (e.g., computers are misclassified as monitors and projectors). In contrast, benefiting from the introduced paradigm, DALN generates large values for the main diagonal elements of the confusion matrix. Moreover, by integrating the NWD into DANN and MDD, the off-diagonal elements of their confusion matrix are considerably decreased, demonstrating the effectiveness of NWD.

Determinacy. To evaluate the determinacy, we calculate

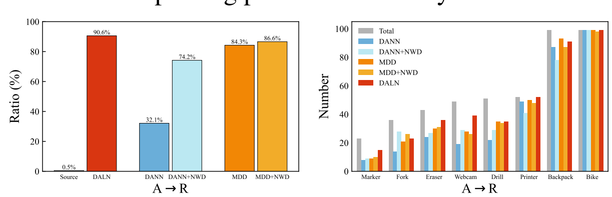


(a) Source only (b) DANN (c) MDD (d) DALN (e) DANN+NWD (f) MDD+NWD

Figure 4. The confusion matrices of different methods of the target domain on task A → W of Office-31. (Zoom in for a clear visualization.)

the ratio of the correctly classified samples that have high prediction certainty. Here, we consider task Ar→Rw of Office-Home. The prediction probability in the range of 0.9 to 1 is regarded as a high certainty prediction. As shown in Fig. 5(a), the model trained on the source-only data nearly cannot generate high certainty prediction. DANN and MDD considerably improve the ratio of high certainty prediction, but the improvements achieved by these methods cannot compete with those achieved by the proposed DALN. By taking the NWD as a regularizer, the ratios of high certainty prediction for DANN and MDD are improved, demonstrating the effectiveness of NWD in improving the determinacy.

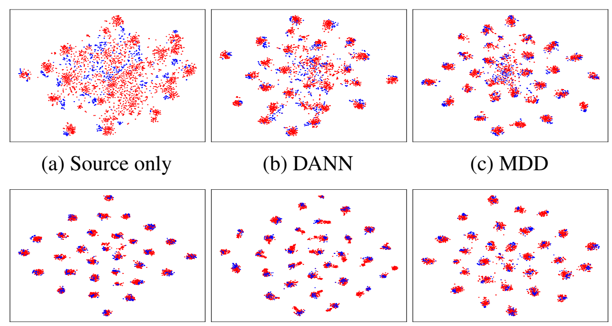
Diversity. As shown in Fig. 5(b), we compute the number of correctly classified samples for some typical categories that have a large or small number of samples. Compared with other methods, the proposed DALN correctly classifies more samples in the categories that have a small number of samples. Moreover, by adopting the NWD as a regularizer for the DANN and MDD, these methods achieve considerable improvements for categories that have a small number of samples. These results demonstrate the effectiveness of the NWD in improving prediction diversity.



(a) Determinacy (b) Diversity

Figure 5. Visualizations of (a) determinacy and (b) diversity on task A→R of Office-Home. The ratio in (a) is the proportion of the number of correctly classified samples, whose prediction probability is in the range of 0.9 to 1, to the total number of correctly classified samples in the target domain. The number of correctly classified samples in (b) is calculated for 8 typical categories that have a large or small number of samples.

t-SNE Visualization. The feature representations of the ResNet-50, DANN, MDD, DALN, DANN+NWD, and MDD+NWD are visualized in Fig. 6 using t-SNE [44]. Compared with the DANN and MDD, the proposed DALN not only confuses the feature representations, but also contributes to a more compact intra-class distribution and a more dispersed inter-class distribution, indicating that the features learned by the DALN are more discriminative. Combining the DANN and MDD with the proposed NWD, the intra-class features are pulled together while the inter-class features are pushed apart, demonstrating that the NWD can help them improve the discriminability.



(a) Source only (b) DANN (c) MDD

(d) DALN (e) DANN+NWD (f) MDD+NWD

Figure 6. t-SNE visualizations of feature distributions learned by different methods on task A→W of Office-31. Blue and red points represent source and target features, respectively.

5. Conclusions

In this work, we present a simple yet effective adversarial paradigm, i.e., reusing the task-specific classifier as a discriminator. To achieve this paradigm, we designed a new discrepancy NWD that has definite guidance meaning and correspondingly built a discriminator-free adversarial UDA model, i.e., DALN, which learns transferable and discriminative representations while promising prediction determinacy and diversity. Moreover, we demonstrated that the proposed NWD can be used as a plug-and-play regularizer to the existing methods, which helps these methods achieve more competitive performance. Extensive experiments on a variety of datasets demonstrate the effectiveness and superiority of the proposed method.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61727809, in part by the Special Fund for Key Program of Science and Technology of Anhui Province under Grant 201903c08020002, and in part by the National Key Research and Development Program of China under Grant 2019YFC0117800.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223, 2017. 2, 3, 4

[2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NeurIPS*, pages 137–144, 2007. 1, 5

[3] Barbara Caputo, Henning Müller, Jesus Martinez-Gomez, Mauricio Villegas, Burak Acar, Novi Patricia, Neda Marvasti, Suzan Üsküdarlı, Roberto Paredes, Miguel Cazorla, et al. Imageclef 2014: Overview and analysis of the results. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 192–211, 2014. 7

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 1

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017. 1

[6] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *ICML*, pages 1081–1090, 2019. 1, 6

[7] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *CVPR*, pages 3941–3950, 2020. 5, 6

[8] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Fast batch nuclear-norm maximization and minimization for robust domain adaptation. *arXiv:2107.06154*, 2021. 5

[9] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *CVPR*, pages 12455–12464, 2020. 1, 2, 6

[10] Zhekai Du, Jingjing Li, Hongzu Su, Lei Zhu, and Ke Lu. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In *CVPR*, pages 3937–3946, 2021. 2, 3

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, pages 2096–2030, 2016. 1, 2, 3, 5, 6, 7

[12] Zhiqiang Gao, Shufei Zhang, Kaizhu Huang, Qiufeng Wang, and Chaoliang Zhong. Gradient distribution alignment certificates better adversarial domain adaptation. In *ICCV*, pages 8937–8946, 2021. 1, 2, 3, 6

[13] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011. 1

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, volume 27, 2014. 2

[15] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *NeurIPS*, page 5769–5779, 2017. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 6

[17] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *ECCV*, pages 464–480, 2020. 4, 6, 7

[18] Vinod Kumar Kurmi and Vinay P Namboodiri. Looking back at labels: A class based domain adaptation technique. In *IJCNN*, pages 1–8, 2019. 2

[19] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *CVPR*, pages 10285–10295, 2019. 2, 3, 6

[20] Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *CVPR*, pages 13936–13944, 2020. 1

[21] Shuang Li, Fangrui Lv, Binhui Xie, Chi Harold Liu, Jian Liang, and Chen Qin. Bi-classifier determinacy maximization for unsupervised domain adaptation. In *AAAI*, 2021. 2

[22] Shuang Li, Mixue Xie, Kaixiong Gong, Chi Harold Liu, Yulin Wang, and Wei Li. Transferable semantic augmentation for domain adaptation. In *CVPR*, pages 11516–11525, 2021. 6

[23] Shuang Li, Mixue Xie, Fangrui Lv, Chi Harold Liu, Jian Liang, Chen Qin, and Wei Li. Semantic concentration for domain adaptation. In *ICCV*, pages 9102–9111, 2021. 2, 6

[24] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015. 1, 2

[25] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, volume 31, 2018. 1, 2, 3, 6, 7

[26] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, pages 2208–2217. PMLR, 2017. 2

[27] Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *CVPR*, pages 9111–9120, 2020. 2, 3

[28] You-Wei Luo and Chuan-Xian Ren. Conditional bures metric for domain adaptation. In *CVPR*, pages 13989–13998, 2021. 6, 7

[29] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014. 3

[30] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *CVPR*, pages 1094–1103, 2021. 1

[31] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *TNNLS*, 22(2):199–210, 2010. 1

[32] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *TNNLS*, 22(10):1345–1359, 2009. 1

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, pages 8026–8037, 2019. 7

[34] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *AAAI*, 2018. 3

[35] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv:1710.06924*, 2017. 7

[36] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010. 5

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

[38] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010. 7

[39] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, pages 3723–3732, 2018. 1, 2, 3, 6

[40] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *AAAI*, 2018. 2, 6, 7

[41] Nathan Srebro, Jason DM Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *NeurIPS*, volume 17, pages 1329–1336, 2004. 5

[42] Hui Tang and Kui Jia. Discriminative adversarial domain adaptation. In *AAAI*, volume 34, pages 5940–5947, 2020. 2, 3, 6

[43] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv:1412.3474*, 2014. 1, 2

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 8

[45] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, pages 5018–5027, 2017. 7

[46] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009. 4

[47] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 1

[48] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Metaalign: Coordinating domain alignment and classification for unsupervised domain adaptation. In *CVPR*, pages 16643–16653, 2021. 1, 6

[49] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *ICML*, pages 7404–7413. PMLR, 2019. 1, 2, 6, 7

[50] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *CVPR*, pages 5031–5040, 2019. 2, 3

[51] Sicheng Zhao, Xiangyu Yue, Shanghang Zhang, Bo Li, Han Zhao, Bichen Wu, Ravi Krishna, Joseph E Gonzalez, Alberto L Sangiovanni-Vincentelli, Sanjit A Seshia, et al. A review of single-source deep unsupervised visual domain adaptation. *TNNLS*, 2020. 1, 2