CASUAL: Conditional Support Alignment for Domain Adaptation with Label Shift

Anh T Nguyen[1*], Lam Tran[2], Anh Tong[3], Tuan-Duy H. Nguyen[4], Toan Tran[2]
[1]University of Illinois Chicago
[2]VinAI Research
[3]Korea University
[4]National University of Singapore
anguy110@uic.edu, {lamtt12,toantm3}@vinai.io
duynht@u.nus.edu, anhtong12@korea.ac.kr

Abstract

Unsupervised domain adaptation (UDA) refers to a domain adaptation framework in which a learning model is trained based on the labeled samples on the source domain and unlabelled ones in the target domain. The dominant existing methods in the field that rely on the classical covariate shift assumption to learn domain-invariant feature representation have yielded suboptimal performance under label distribution shift. In this paper, we propose a novel **C**onditional **A**dversarial **SU**pport **AL**ignment (CASUAL) whose aim is to minimize the conditional symmetric support divergence between the source's and target domain's feature representation distributions, aiming at a more discriminative representation for the classification task. We also introduce a novel theoretical target risk bound, which justifies the merits of aligning the supports of conditional feature distributions compared to the existing marginal support alignment approach in the UDA settings. We then provide a complete training process for learning in which the objective optimization functions are precisely based on the proposed target risk bound. Our empirical results demonstrate that CASUAL outperforms other state-of-the-art methods on different UDA benchmark tasks under different label shift conditions.

Introduction

The remarkable success of modern deep learning models often relies on the assumption that training and test data are independent and identically distributed (i.i.d), contrasting the types of real-world problems that can be solved. The violation of that i.i.d. assumption leads to the data distribution shift, or out-of-distribution (OOD) issue, which negatively affects the generalization performance of the learning models (Torralba and Efros 2011; Li et al. 2017) and renders them impracticable. One of the most popular settings for the OOD problem is unsupervised domain adaptation (UDA) (Ganin and Lempitsky 2015; David et al. 2010) in which the training process is based on fully labeled samples from a source domain and completely unlabeled samples from a target domain.

While the *covariate shift* assumption has been extensively studied under the UDA problem setting, with reducing the feature distribution divergence between domains as the dominant approach (Ganin and Lempitsky 2015; Tzeng et al. 2017; Shen et al. 2018; Courty et al. 2017; Liu et al. 2019; Long et al. 2015, 2017, 2016, 2014), the *label shift* assumption (i.e., the marginal label distribution $p(y)$ varies between domains, while the conditional $p(x|y)$ is unchanged) remains vastly underexplored in comparison. Compared to the covariate shift assumption, the label shift assumption is often more reasonable in several real-world settings, e.g., the healthcare industry, where the distribution of diseases in medical diagnosis may change across hospitals, while the conditional distribution of symptoms given diseases remains unchanged.

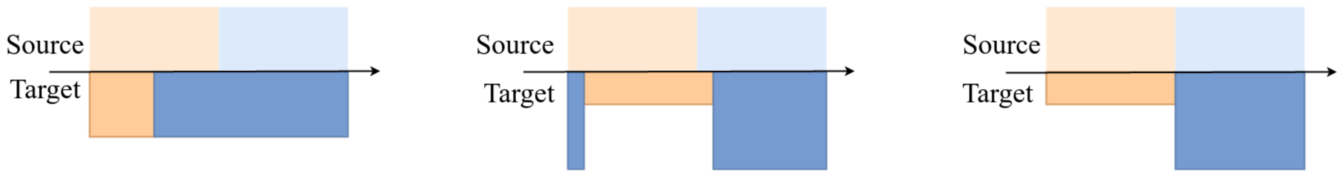
Several UDA methods that explicitly consider the label shift assumption often rely on estimating the importance weights of the source and target label distribution and strictly require the conditional distributions $p(x|y)$ or $p(z|y)$ to be domain-invariant (Lipton, Wang, and Smola 2018; Tachet des Combes et al. 2020; Azizzadenesheli et al. 2019). Another popular UDA under label shift framework is enforcing domain invariance of representation z w.r.t some *relaxed* divergences (Wu et al. 2019; Tong et al. 2022). Wu et al. (2019) proposed reducing β-admissible distribution divergence to prevent cross-label mapping in conventional domain-invariant approaches. However, choosing inappropriate values for β can critically reduce the performance of this method under extreme levels of label shift (Wu et al. 2019; Li et al. 2020).

This paper aims to develop a new theoretically sound, namely conditional adversarial support alignment (CASUAL) approach for UDA under the label shift. Our proposed method is relatively related to but different from the adversarial support alignment (ASA) (Tong et al. 2022) one, which utilizes the symmetric support divergence (SSD) to align the support of the marginal feature distribution of the source and the target domains. One of the critical drawbacks of the ASA method is that indiscriminately reducing the marginal support divergence may make the learned representation susceptible to conditional distribution misalignment. Our proposed CASUAL alleviates that issue by considering discriminative features when aligning the supports of two distributions. In particular, the conditional support alignment instead of the marginal case makes CASUAL less susceptible to misalignment of features between different classes than ASA, which is illustrated intuitively by Figure 1.

The main contributions of our paper are summarized as follows:

[*]Work done at VinAI Research.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(a) Distribution alignment (b) Support alignment (c) Conditional support alignment

Figure 1: Illustration of the learned latent space of different domain-invariant frameworks under label shift for a binary classification problem. It can be seen that the support alignment (b) can mitigate the high error rate induced by distribution alignment (a), whereas the conditional support alignment (c) can achieve the best representation by explicitly aligning the supports of class-conditioned latent distributions.

- We propose a novel conditional adversarial support alignment (CASUAL) to align the support of the conditional feature distributions on the source and target domains, aiming for a more label-informative representation for the classification task.
- We provide a new theoretical upper bound for the target risk for the learning process of our CASUAL. We then introduce a complete training scheme for our proposed CASUAL by minimizing that bound.
- We provide experimental results on several benchmark tasks in UDA, which consistently demonstrate our proposed method's empirical benefits compared to other existing UDA approaches.

Methodology

Problem Statement

Let us consider a classification framework where $\mathcal{X} \subset \mathbb{R}^d$ represents the input space and $\mathcal{Y} = \{y_1, y_2, \ldots, y_K\}$ denotes the output space. A domain is then defined by $P(x, y) \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$, where $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ is the set of joint probability distributions on $\mathcal{X} \times \mathcal{Y}$. The set of conditional distributions of x given y, $P(x|y)$, is denoted as $\mathcal{P}_{\mathcal{X}|\mathcal{Y}}$, and the set of probability marginal distributions on $\mathcal{X}$ and $\mathcal{Y}$ is denoted as $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{Y}}$, respectively. We also denote $P_{X|Y=y_k}$ as $P_{X|k}$, $k = 1, \ldots, K$ for convenience.

Consider an UDA framework in which the (labelled) source domain $D^S = \left\{\left(x_i^S, y_i^S\right)\right\}_{i=1}^{n_S}$, where $\left(x_i^S, y_i^S\right) \sim P^S(x, y)$, and the (unlabelled) target domain $D^T = \left\{x_j^T\right\}_{j=1}^{n_T}$, where $x_j^T \sim P_X^T \subset \mathcal{P}_{\mathcal{X}}$. Without loss of generality, we assume that both the source and target domains consist of K classes, i.e., $\mathcal{Y} = \{y_1, \ldots, y_K\}$. In this paper, we focus on the UDA setting with label shift, which assumes that $P_Y^S \neq P_Y^T$ while the conditional distributions $P_{X|Y}^S$ and $P_{X|Y}^T$ remain unchanged. Nevertheless, unlike relevant works such as Tachet des Combes et al. (2020); Lipton, Wang, and Smola (2018), we do not make a strong assumption about the invariance of $P_{X|Y}$ or $P_{Z|Y}$ between those domains, targeting more general UDA settings under label shift.

A classifier (or hypothesis) is defined by a function $h : \mathcal{X} \mapsto \Delta_K$, where $\Delta_K = \left\{\boldsymbol{\pi} \in \mathbb{R}^K : \|\boldsymbol{\pi}\|_1 = 1 \wedge \pi \geq \mathbf{0}\right\}$ is the K-simplex, and an induced scoring function $g : \mathcal{X} \mapsto \mathbb{R}^K$. Consider a loss function $\ell : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}_+$, satisfying $\ell(y, y) = 0,\ \forall y \in \mathcal{Y}$. Given a scoring function g, we define its associated classifier as h_g, i.e., $h_g(x) = \hat{y}$ with $\hat{y} \in \operatorname{argmin}_{y\in\mathcal{Y}} \ell(g(x), y)$. For conciseness, we consider any hypothesis h a scoring function g.

The ℓ-risk of a scoring function g over a distribution $P_{X\times Y}$ is then defined by $\mathcal{L}_P(g) := \mathbb{E}_{(x,y)\sim P}[\ell(g(x), y)]$, and the classification mismatch of g with a classifier h by $\mathcal{L}_P(g, h) := \mathbb{E}_{x\sim P_X}[\ell(g(x), h(x))]$. For convenience, we denote the source and target risk of scoring function or classifier g as $\mathcal{L}_S(g)$ and $\mathcal{L}_T(g)$, respectively.

A Support Misalignment-Based Target Risk Bound

Similar to Ganin et al. (2016), we assume that the hypothesis g can be decomposed as $g = c \circ f$, where $c : \mathcal{Z} \to \mathcal{Y}$ is the classifier, $f : \mathcal{X} \to \mathcal{Z}$ is the feature extractor, and $\mathcal{Z}$ represents the latent space. Let us denote the domain discriminator as $\phi : \mathcal{Z} \to \{0, 1\}$, and the marginal distribution of $Z \in \mathcal{Z}$ in source and target domains as P_Z^S and P_Z^T, respectively.

We next introduce several necessary definitions and notations for the theoretical development of the target error bound in our proposed CASUAL . Some are employed in constructing the IMD-based domain adaptation bound in (Dhouib and Maghsudi 2022).

Definition 1 (*Source-guided uncertainty* (Dhouib and Maghsudi 2022))**.** Let $\mathbb{H}$ be a hypothesis space, and let ℓ be a given loss function. The source-guided uncertainty of $g \in \mathbb{H}$ associated with ℓ is defined by:

$$\mathcal{C}_{\mathbb{H}}(g) = \inf_{h\in\mathbb{H}} \mathcal{L}_T(g, h) + \mathcal{L}_S(h), \tag{1}$$

where $\mathcal{L}_T(g, h)$ is the classification mismatch of g and h on P_X^T.

Remark 1. When ℓ is the cross-entropy loss, minimizing the conditional entropy of predictor on the target domain, along with $\mathcal{L}_S(h_g)$, effectively minimizes $\mathcal{C}_{\mathbb{H}}(g)$ (Dhouib and Maghsudi 2022).

Definition 2 (*Integral measure discrepancy* (Dhouib and Maghsudi 2022))**.** Let $\mathbb{F}$ be a family of nonnegative functions over $\mathbb{X}$, containing the null function, e.g., $\mathbb{F} = \{\ell(h, f_s); h \in \mathbb{H}\}$ with f_s as the source labeling function. The Integral Measure Discrepancy (IMD) associated to $\mathbb{F}$ between two distribution $\mathcal{Q}$ and $\mathcal{P}$ over $\mathbb{X}$ is

$$\mathrm{IMD}_{\mathbb{F}}(\mathcal{Q}, \mathcal{P}) := \sup_{f\in\mathbb{F}} \int f \, \mathrm{d}\mathcal{Q} - \int f \, \mathrm{d}\mathcal{P}. \tag{2}$$

Intuitively, this discrepancy aims to capture the distances between measures w.r.t. difference masses.

Definition 3 (*Symmetric support divergence* (Tong et al. 2022)). Assuming that d is a proper distance on the latent space $\mathcal{Z}$. The symmetric support divergence (SSD) between two probability distributions P_Z and Q_Z is then defined by:

$$\mathcal{D}_{\text{supp}}(P_Z, Q_Z) = \mathbb{E}_{z\sim P_Z}[d(z, \text{supp}(Q_Z))] + \mathbb{E}_{z\sim Q_Z}[d(z, \text{supp}(P_Z))], \quad (3)$$

where $\text{supp}(P_Z)$ and $\text{supp}(Q_Z)$ are the corresponding supports of P_Z and Q_Z, respectively.

Different from the work in Dhouib and Maghsudi (2022) that extends the bound with β-admissible distances, our proposed method goes beyond interpreting the bound solely in terms of support divergences (Tong et al. 2022). In particular, our method also incorporates the label structures of the domains, allowing a more comprehensive analysis of the underlying relationships.

A Novel Conditional SSD-Based Domain Adaptation Bound

In this work, we first introduce our definition for a novel *conditional symmetric support divergence* between the conditional distributions $P^S_{Z|Y}$ and $P^T_{Z|Y}$. For simplicity, we also denote d as a well-defined distance on the conditional $\mathcal{Z}|\mathcal{Y}$ space.

Definition 4 (*Conditional symmetric support divergence*). The conditional symmetric support divergence (CSSD) between the conditional distributions $P^S_{Z|Y}$ and $P^T_{Z|Y}$ is defined by

$$\mathcal{D}^c_{\text{supp}}(P^S_{Z|Y}, P^T_{Z|Y}) \quad (4)$$
$$= \sum_{y\in\mathcal{Y}} P^S(Y=y)\mathbb{E}_{z\sim P^S_{Z|Y=y}}[d(z, \text{supp}(P^T_{Z|Y=y}))] \quad (5)$$
$$+ P^T(Y=y)\mathbb{E}_{z\sim P^T_{Z|Y=y}}[d(z, \text{supp}(P^S_{Z|Y=y}))]. \quad (6)$$

We justify CSSD as a support divergence through the following result.

Proposition 1. *Assuming that $P^S(Y=y) > 0, P^T(Y=y) > 0$ for any $y \in \mathcal{Y}$, then $\mathcal{D}^c_{\text{supp}}(P^S_{Z|Y}, P^T_{Z|Y})$ defined in Eq.* (4) *is a support divergence.*

All proofs are provided in the Appendix. Compared to the SSD in Definition 3, our CSSD considers the class proportions in both source and target domains. As a result, the localized IMD considers per-class localized functions, which are defined as the $(\boldsymbol{\epsilon}, P^S_{Z|Y})$-localized nonnegative function denoted by $\mathbb{F}_{\boldsymbol{\epsilon}}$. Specifically, $\mathbb{F}_{\boldsymbol{\epsilon}} = \{f; f(z) \geq 0, \mathbb{E}_{P^S_{Z|k}}[f] \leq \epsilon_k, \quad k = 1 \ldots K\}$ with $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_K) \geq 0$. In the following lemma, we introduce upper bounds for the IMD using CSSD.

Lemma 1 (Upper bound IMD using CSSD). *Let $\mathbb{F}$ be a set of nonnegative and* 1*-Lipschitz functions, and let $\mathbb{F}_{\boldsymbol{\epsilon}}$ be a $(\boldsymbol{\epsilon}, P^S_{Z|Y})$-localized nonnegative function. Then, we can bound the IMD w.r.t the conditional support domain and CSSD, respectively, as*

$$\text{IMD}_{\mathbb{F}_{\boldsymbol{\epsilon}}}(P^T_Z, P^S_Z) \leq \sum_{k=1}^K q_k \mathbb{E}_{z\sim P^T_{Z|k}}[d(z, \text{supp}(P^S_{Z|k}))] + q_k\delta_k + p_k\epsilon_k; \quad (7)$$

$$\text{IMD}_{\mathbb{F}_{\boldsymbol{\epsilon}}}(P^T_Z, P^S_Z) \leq \mathcal{D}^c_{\text{supp}}(P^T_{Z|Y}, P^S_{Z|Y}) + \sum_{k=1}^K q_k\delta_k + p_k\gamma_k \quad (8)$$

where $\delta_k := \sup_{z\in\text{supp}\, P^S_{Z|k}, f\in\mathbb{F}_{\boldsymbol{\epsilon}}} f(z)$, $\gamma_k := \sup_{z\in\text{supp}\, P^T_{Z|k}, f\in\mathbb{F}_{\boldsymbol{\epsilon}}} f(z)$, $p_k = P^S(Y = y_k)$ *and* $q_k = P^T(Y = y_k)$.

We now provide a novel upper bound of the target risk based on CSSD in the following theorem, which is a straightforward result from Lemma 1.

Theorem 1 (Domain adaptation bound via CSSD). *Let $\mathbb{H}$ be a hypothesis space, g be a score function, and ℓ be a loss function satisfying the triangle inequality. Consider the localized hypothesis $\mathbb{H}^{\mathbf{r}} := \{h \in \mathbb{H}; \mathcal{L}_{S_k}(h) \leq r_k, k = 1 \ldots K\}$. Assume that $l(\mathbb{H}, \mathbb{H}) := \{l(h_1(\cdot), h_2(\cdot)); h_1, h_2 \in \mathbb{H}\} \subseteq \mathbb{F}$, and that all the assumptions for $\mathbb{F}_{\boldsymbol{\epsilon}}$ in Lemma 1 are fulfilled. Then, for any $\mathbf{r}^1 = (r^1_1, \ldots, r^1_K) \geq 0, \mathbf{r}^2 = (r^2_1, \ldots, r^2_K) \geq 0$ that satisfy $r^1_k + r^2_k = \epsilon_k$, we have:*

$$\mathcal{L}_T(g) \leq \mathcal{C}_{\mathbb{H}^{\mathbf{r}^1}}(g) + \mathcal{D}^c_{\text{supp}}(P^T_{Z|Y}, P^S_{Z|Y}) + \sum_{k=1}^K q_k\delta_k + p_k\gamma_k + \inf_{h\in\mathbb{H}^{\mathbf{r}^2}} \mathcal{L}_S(h) + \mathcal{L}_T(h). \quad (9)$$

Remark 2. In the case that we do not incorporate the label information, the IMD can be bounded as

$$\text{IMD}_{\mathbb{F}_\epsilon} \leq \mathbb{E}_{z\sim P^T_Z}[d(z, \text{supp}(P^S_Z))] + \delta + \epsilon, \quad (10)$$

where $\delta = \sup_{z\in\text{supp}(P^S_Z), f\in\mathbb{F}_\epsilon} f(z)$. Similar to the findings in Dhouib and Maghsudi (2022), the inequality in Equation 10 provides a justification for minimizing SSD proposed in Tong et al. (2022). Notably, this inequality extends to the case where $\epsilon \geq 0$, and thus recovers the bound with SSD in (Dhouib and Maghsudi 2022) as a special case. Note that in order to make a fair comparison between Eq. (7) and (10), we assume that $\sum_k \epsilon_k p_k = \epsilon$ making $\mathbb{F}_{\boldsymbol{\epsilon}} \subseteq \mathbb{F}_\epsilon$.

In comparison to the upper bound in Eq. (10), our expectation holds

$$\sum_k q_k \mathbb{E}_{z\sim P^T_{Z|k}} d(z, \text{supp}(P^S_{Z|k})) \geq \mathbb{E}_{z\sim P^T_Z} d(z, \text{supp}(P^S_Z))$$

due to the fact that $d(z, \text{supp}(P^S_{Z|k})) \geq d(z, \text{supp}(P^S_Z))$ and that Jensen's inequality holds.

However, this inequality does not imply that our bound is less tight than the bound using SSD. When considering the remaining part, we can observe that $\sum_k q_k\delta_k \leq \delta$ since $\text{supp}(P^S_{Z|k}) \subseteq \text{supp}(P^S_Z)$ for any $k = 1, \ldots, K$. In other words, there is a trade-off between the distance to support space and the uniform norm (sup norm) of function on the supports.

Remark 3. The proposed bound shares several similarities with other target error bounds in the UDA literature (Ben-David et al. 2010; Acuna et al. 2021). In particular, these bounds all upperbound the target risk with a source risk term, a domain divergence term, and an ideal joint risk term. The main difference is that we use the conditional symmetric support divergence instead of $\mathcal{H}\Delta\mathcal{H}$-divergence in Ben-David et al. (2010) and f-divergence in Acuna et al. (2021), making our bound more suitable for problems with large degrees of label shift, as a lower value of CSSD does not necessarily increase the target risk under large label shift, unlike $\mathcal{H}\Delta\mathcal{H}$-divergence and f-divergence (Tachet des Combes et al. 2020; Zhao et al. 2019). Furthermore, the localized hypothesis spaces $\mathbb{H}^{\mathbf{r}^1}$ and $\mathbb{H}^{\mathbf{r}^2}$ are reminiscent of the localized adaptation bound proposed in (Zhang et al. 2020). While lower values of $\mathbf{r}^1, \mathbf{r}^2$ can make the term $\sum_{k=1}^{K} q_k\delta_k + p_k\gamma_k$ smaller, the source-guided uncertainty term and ideal joint risk term can increase as a result. In our final optimization procedure, we assume the ideal joint risk term and $\sum_{k=1}^{K} q_k\delta_k + p_k\gamma_k$ values to be small and minimize the source-guided uncertainty (see section) and CSSD via another proxy (see section) to reduce the target domain risk.

For a more comprehensive comparison of the proposed error bound 1 to those of other generalized target shift methods (Tachet des Combes et al. 2020; Rakotomamonjy et al. 2022; Kirchmeyer et al. 2022; Gong et al. 2016), we refer readers to Appendix.

Training Scheme for CASUAL

So far, we have presented the main ideas of our CASUAL algorithm in a general manner. In the next section, we discuss the implementation details of our proposed framework.

Minimizing Source-Guided Uncertainty As stated in Remark 1, minimizing the source risk and the target conditional entropy reduces the source-guided uncertainty $\mathcal{C}_{\mathbb{H}^{\mathbf{r}_1}}(g)$, which is the second term in the target risk bound of Eq. (9). Under the assumption that the input distribution X contains well-separated same-class clusters, it is intuitive that minimizing the conditional entropy hedges the model towards high-confident, large margins classifiers, i.e., pushing the decision boundary away from high-density data regions. Minimizing the prediction entropy has also been extensively studied and resulted in effective UDA algorithms (Shu et al. 2018; Kirchmeyer et al. 2022; Liang et al. 2021). Hence, the total loss of CASUAL first includes the overall classification loss $\mathcal{L}_y(\cdot)$ on source samples and the conditional entropy loss on target samples $\mathcal{L}_{ce}(\cdot)$, defined as follows:

$$\mathcal{L}_y(g) = \frac{1}{n_S}\sum_{i=1}^{n_S} \ell(g(x_i^S), y_i^S); \quad (11)$$

$$\mathcal{L}_{ce}(g) = -\frac{1}{n_T}\sum_{i=1}^{n_T} g(x_i^T)^\top \ln g(x_i^T). \quad (12)$$

Enforcing Lipschitz Hypothesis The risk bound in Eq. (9) suggests regularizing the Lipschitz continuity of the classifier c. Inspired by the success of virtual adversarial training by Miyato et al. (2018) on domain adaptation tasks (Shu et al. 2018; Tong et al. 2022), we instead enforce the locally-Lipschitz constraint of the classifier, which is a relaxation of the global Lipschitz constraint, by enforcing consistency in the norm-ball w.r.t each representation sample z. This regularization ensures the learned classifier c is robust against changes within the norm-ball neighborhood of each sample z. In addition, we observe that enforcing the local Lipschitz constraint of $g = c \circ f$ instead of c leads to better performance in empirical experiments. Hence, we introduce the virtual adversarial loss term $\mathcal{L}_v(c, f)$ (Miyato et al. 2018), which enforces the classifier consistency within the ϵ-radius neighborhood of each sample x by penalizing the KL-divergence between predictions of nearby samples, and follow the approximation method in Miyato et al. (2018), which primarily relies on the power iteration method. In particular,

$$\mathcal{L}_v(c, f) = \frac{1}{n_S}\sum_{i=1}^{n_S} \max_{\|r\|<\epsilon} \mathrm{D}_{\mathrm{KL}}\left(g(x_i^S)\|g(x_i^S + r)\right) + \frac{1}{n_T}\sum_{i=1}^{n_T} \max_{\|r\|<\epsilon} \mathrm{D}_{\mathrm{KL}}\left(g(x_i^T)\|g(x_i^T + r)\right). \quad (13)$$

Minimizing Conditional Symmetric Support Divergence The next natural step for reducing the target risk bound in (9) is to minimize $\mathcal{D}^c_{\mathrm{supp}}(P^S_{Z|Y}, P^T_{Z|Y})$. However, it is challenging to directly optimize this term since in a UDA setting, we have no access to the labels of the target samples. Motivated by the use of pseudo labels to guide the training process in domain adaptation literature (French, Mackiewicz, and Fisher 2018; Chen et al. 2019; Long et al. 2018; Zhang et al. 2019a), we alternatively consider minimizing $\mathcal{D}^c_{\mathrm{supp}}(P^S_{Z|\widehat{Y}}, P^T_{Z|\widehat{Y}})$ as a proxy for minimizing $\mathcal{D}^c_{\mathrm{supp}}(P^S_{Z|Y}, P^T_{Z|Y})$, where $\widehat{Y}$ are pseudo labels. To mitigate the error accumulation issue of using pseudo labels under large domain shift (Zhang et al. 2019a; Liu, Wang, and Long 2021), we employ the entropy conditioning technique in Long et al. (2018) in our implementation of CASUAL . Nevertheless, given that the estimation of the conditional support alignment $\mathcal{D}^c_{\mathrm{supp}}(P^S_{Z|Y}, P^T_{Z|Y})$ is based on the approximation of P^T_Y, which is commonly time-consuming and error-prone, the following proposition motivates us to alternatively minimize the joint support divergence $\mathcal{D}_{\mathrm{supp}}(P^S_{Z,Y}, P^T_{Z,Y})$ to tighten the target error bound, without using any explicit estimate of the marginal label distribution shift as performed in Lipton, Wang, and Smola (2018); Tachet des Combes et al. (2020).

Proposition 2. *Assuming that* $P^S(\widehat{Y} = y) > 0$, $P^T(\widehat{Y} = y) > 0, \forall y \in \mathcal{Y}$, *and there exists a well-defined distance denoted by* d *in the space* $\mathcal{Z}\times\mathcal{Y}$. *Then* $\mathcal{D}^c_{\mathrm{supp}}(P^S_{Z|Y}, P^T_{Z|Y}) = 0$ *if and only if* $\mathcal{D}_{\mathrm{supp}}(P^S_{Z,Y}, P^T_{Z,Y}) = 0$.

Moreover, minimizing this joint support divergence can be performed efficiently in one-dimensional space. In particular, Tong et al. (2022) indicated that when considering the log-loss discriminator $\phi : \mathcal{X} \to [0, 1]$ trained to discriminate between two distributions P and Q with binary cross entropy loss function can be used to estimate $\mathcal{D}_{\mathrm{supp}}(P, Q)$. Instead of aligning the marginal distributions P^S_Z and P^T_Z, our approach focuses on matching the support of joint distributions, $P^S_{Z,\widehat{Y}}$

Algorithm	α					Average
	$\emptyset$	10	3	1	0.5	
No DA	73.9	73.8	73.5	73.9	73.8	73.8
DANN*	96.2	96.2	93.5	82.6	72.3	88.2
CDAN*	96.6	96.5	93.7	82.2	70.7	88.0
VADA	**98.1**	**98.1**	96.8	84.9	76.8	90.9
SDAT	97.8	97.5	94.1	83.3	68.8	88.3
MIC	98.0	97.6	95.2	82.9	70.8	88.9
DALN	97.2	96.3	91.2	83.4	72.5	88.2
IWDAN	97.5	97.1	90.4	81.3	73.3	87.9
IWCDAN	97.8	97.5	91.4	82.6	73.8	88.7
sDANN	87.4	90.7	92.1	89.4	85.2	89.0
ASA	94.1	93.7	94.1	90.8	84.7	91.5
PCT	97.4	97.2	94.3	82.3	71.8	88.6
SENTRY	97.5	91.5	91.4	84.7	82.3	89.5
OSTAR	97.7	97.6	95.7	91.3	84.7	93.4
BIWAA	97.5	97.5	91.6	82.7	73.3	88.6
MARS	96.8	92.6	88.5	85.3	83.5	89.4
CASUAL (Ours)	98.0	98.0	**97.2**	**96.7**	**88.3**	**95.6**

Table 1: Per-class average accuracies on USPS→MNIST.

Algorithm	α					Average
	$\emptyset$	10	3	1	0.5	
No DA	69.9	69.8	69.7	68.8	67.9	69.2
DANN*	75.9	74.9	74.1	71.7	69.5	73.2
CDAN*	75.6	74.3	73.7	71.8	69.7	73.0
VADA	**77.1**	75.5	73.8	71.3	68.0	73.1
SDAT	75.8	74.4	71.5	68.3	66.2	71.2
MIC	76.4	75.5	72.7	68.6	67.3	72.1
DALN	74.3	74.1	72.8	70.3	68.4	72.0
IWDAN	72.9	72.6	71.8	70.6	69.5	71.5
IWCDAN	72.1	72.0	71.5	71.9	69.9	71.5
sDANN	72.8	72.0	72.0	71.4	70.1	71.7
ASA	72.7	72.2	72.1	71.5	69.8	71.7
PCT	75.0	76.1	75.0	70.9	68.3	73.1
SENTRY	76.7	76.6	75.2	71.2	67.0	73.3
OSTAR	73.7	72.4	70.8	70.1	68.4	71.1
BIWAA	75.7	74.5	72.1	70.6	66.3	71.9
MARS	72.8	71.3	69.7	67.6	65.9	69.5
CASUAL (Ours)	76.9	**76.8**	**75.8**	**74.2**	**71.7**	**75.1**

Table 2: Per-class average accuracies on STL→CIFAR.

and $P^T_{Z,\widehat{Y}}$. We use a trained optimal discriminator ϕ^* to discriminate between these distributions, which are represented as the outer product $Z \otimes \widehat{Y}$ (Long et al. 2018). Consequently, our model incorporates the domain discriminator loss and support alignment loss to minimize $\mathcal{D}^c_{\text{supp}}(P^S_{Z|Y}, P^T_{Z|Y})$ in the error bound specified in Eq. (9).

$$\mathcal{L}_{dis}(\phi) = -\frac{1}{n_S}\sum_{i=1}^{n_S} \ln\left[G(x_i^S)\right] - \frac{1}{n_T}\sum_{i=1}^{n_T} \ln\left[1 - G(x_i^T)\right]; \tag{14}$$

$$\mathcal{L}_{align}(f) = \frac{1}{n_S}\sum_{i=1}^{n_S} d\left(G(x_i^S), \{G(x_j^T)\}_{j=1}^{n_T}\right) + \frac{1}{n_T}\sum_{i=1}^{n_T} d\left(G(x_i^T), \{G(x_j^S)\}_{j=1}^{n_S}\right), \tag{15}$$

where $G(x) = \phi(f(x) \otimes g(x))$ and $u \otimes v = uv^T$. Here, similar to (Tong et al. 2022), $d(\cdot,\cdot)$ is either the squared L2 or the L1 distance.

Overall, the training process of our proposed algorithm, CASUAL , can be formulated as an alternating optimization problem (see Algorithm 1),

$$\min_{f,c} \mathcal{L}_y(g) + \lambda_{align}\mathcal{L}_{align}(f) + \lambda_{ce}\mathcal{L}_{ce}(g) + \lambda_v\mathcal{L}_v(g); \tag{16}$$

$$\min_{\phi} \mathcal{L}_{dis}(\phi) \tag{17}$$

where $\lambda_{align}, \lambda_y, \lambda_{ce}, \lambda_v$ are the weight hyper-parameters associated with the respective loss terms.

Algorithm 1: Conditional Adversarial Support Alignment

Input: $D^S = \left\{\left(x_i^S, y_i^S\right)\right\}_{i=1}^{n_S}$, $D^T = \left\{x_j^T\right\}_{j=1}^{n_T}$

Output: Feature extractor f, classifier c, domain discriminator r

1: **for** number of training iterations **do**
2: Sample minibatch from source $\left\{\left(x_i^S, y_i^S\right)\right\}_{i=1}^{m}$ and target $\left\{x_i^T\right\}_{i=1}^{m}$
3: Update ϕ according to Eq. (17)
4: Update f, c according to Eq. (16)
5: **end for**

Experiments

Setup

Datasets. We focus on visual domain adaptation tasks and empirically evaluate our proposed algorithm CASUAL on benchmark UDA datasets **USPS** $\rightarrow$ **MNIST**, **STL** $\rightarrow$ **CIFAR** and **VisDA-2017**. We further conduct experiments on the **DomainNet** dataset and provide the results in Appendix due to the page limitation. For VisDA-2017 and DomainNet, instead of using the 100% unlabeled target data for both training and evaluation (Prabhu et al. 2021; Tanwisuth et al. 2021), we utilize 85% of the unlabeled target data for training, and the remaining 15% for evaluation, to mitigate potential overfitting to seen target samples during training(Garg et al. 2023).

Evaluation setting. To assess CASUAL 's robustness to label shift, we adopt the experimental protocol of Garg et al. (2023). We simulate label shift using the Dirichlet distribution, keeping the source label distribution unchanged and $P_Y^T(y) \sim Dir(\beta)$, with $\beta_y = \alpha.P_Y^{T0}(y)$, $P_Y^{T0}(y)$ as the original target marginal, and α values of $10, 3.0, 1.0, 0.5$. We also include a no-label shift setting, denoted as $\alpha =$ None, where both source and target label distributions are unchanged. For each method and label shift degree, we perform 5 runs with different random seeds and report average **per-class** accuracy on the target domain's test set, which can be a more informative evaluation metric than overall accuracy in (Garg et al. 2023).

Baselines. We assess the performance of CASUAL by comparing it with various existing UDA algorithms, including No DA (training using solely labeled source samples), DANN (Ganin et al. 2016), CDAN (Long et al. 2018), VADA (Shu et al. 2018), SDAT (Rangwani et al. 2022), MIC (Hoyer et al. 2023), DALN (Chen et al. 2022), IWDAN, IWCDAN (Tachet des Combes et al. 2020), sDANN (Wu et al. 2019), ASA (Tong et al. 2022), PCT (Tanwisuth et al. 2021), SENTRY (Prabhu et al. 2021), OSTAR (Kirchmeyer et al. 2022), MDD+IA (Jiang et al. 2020), BIWAA (Westfechtel et al. 2023) and MARS (Rakotomamonjy et al. 2022). Whereas IWDAN, IWCDAN, MARS and OSTAR rely on importance weighting methods, CDAN, IWCDAN, FixMatch and SENTRY employ target pseudolabels. Moreover, we apply the resampling and reweighting heuristics in Garg et al. (2023) to DANN, CDAN and FixMatch to make these baselines more robust to label shift, and denote them as DANN*, CDAN*, and FixMatch* in Table 1, 2 and 3. Further implementation details, including the hyperparameter values and network architectures, are provided in the Appendix.

Main Results

We report the results on USPS→MNIST, STL→CIFAR and VisDA-2017 in Table 1, 2 and 3 respectively. Methods focusing on distribution alignment such as DANN, CDAN, and VADA tend to achieve the highest accuracy scores under α = None. However, their performances degrade significantly as label shift becomes more severe. For instance, under $\alpha = 0.5$ in USPS→MNIST task, SDAT and MIC perform worse than No DA by 5.0% and 3.0%, respectively.

In contrast, CASUAL outperforms baseline methods on 10 out of 15 transfer tasks. It achieves the second-highest average accuracies when there is no label shift in the USPS→MNIST and STL→CIFAR tasks. This gap in performance of CASUAL and other baselines under no label shift could be because support alignment is an extreme relaxation of distribution alignment (Tong et al. 2022), which can still deliver competitive performance when the label shift level is mild (Zhao et al. 2019; Rangwani et al. 2022). Notably, under more severe label shifts, CASUAL outperforms the second-best methods by 3.6%, 1.8% and 0.7% under $\alpha = 0.5$ in the USPS→MNIST, STL→CIFAR and VisDA-2017 tasks, respectively. The robustness of CASUAL to different label shift levels is further demonstrated by its highest average accuracy results, surpassing the second-best methods by 2.2% and 1.9% on USPS→MNIST and STL→CIFAR, respectively. We, therefore, hypothesize that aligning the conditional feature support is more robust to label distribution shift than explicitly aligning the conditional feature distribution, which is performed in IWDAN, IWCDAN, MARS, and OSTAR.

Visualization and Additional Analysis

Analysis of individual loss terms. To study the impact of each loss term in Eq. (16), we provide additional experiment results, which consist of the average accuracy over 5 different random runs on USPS→MNIST in Table 4. It is evident that the conditional support alignment loss term $\mathcal{L}_{align}$, conditional entropy loss term $\mathcal{L}_{ce}$ and virtual adversarial loss term $\mathcal{L}_v$ all improve the model's performance across different levels of label shift. We further provide the hyperparameter values analysis and convergence analysis for each loss term in the Appendix.

Algorithm	α: ∅	10	3	1	0.5	Avg
No DA	55.6	56.0	55.5	55.2	55.1	55.5
DANN*	75.5	71.3	68.4	62.2	56.4	66.8
CDAN*	75.0	72.5	69.8	61.3	56.3	67.0
FixMatch*	71.6	67.5	65.6	60.1	58.7	64.7
VADA	75.2	72.3	69.6	59.2	52.6	65.8
SDAT	<u>75.4</u>	73.3	66.8	63.9	61.8	68.3
MIC	**75.6**	**74.5**	69.5	64.8	62.0	<u>69.3</u>
DALN	73.4	72.8	69.7	62.9	55.4	66.8
IWDAN	74.1	73.3	<u>71.4</u>	<u>65.3</u>	59.7	68.8
IWCDAN	73.5	72.5	69.6	62.9	57.2	67.1
sDANN	72.8	72.2	71.2	64.9	<u>62.5</u>	68.7
ASA	66.4	65.3	64.6	61.7	60.1	63.6
PCT	68.2	66.8	65.4	60.5	53.6	63.3
SENTRY	67.5	64.5	57.6	53.4	52.6	59.1
OSTAR	67.7	66.8	65.3	60.8	56.3	63.4
BIWAA	73.2	72.7	69.9	62.3	60.8	67.8
MARS	62.4	61.2	59.5	57.7	55.5	59.3
MDD+IA	72.9	72.4	70.2	64.6	60.3	68.1
CASUAL (Ours)	74.3	<u>73.4</u>	**71.8**	**66.3**	**63.2**	**69.8**

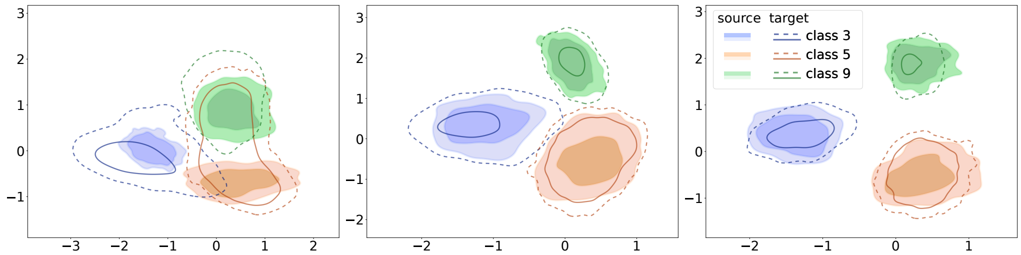
Table 3: Per-class accuracies on VisDA-2017.

$\mathcal{L}_{align}$	$\mathcal{L}_{ce}$	$\mathcal{L}_v$	α: ∅	10	3	1	0.5	Avg
✓			94.5	94.2	94.3	94.3	84.9	92.4
✓	✓		97.7	97.2	96.8	96.2	87.4	95.1
✓	✓	✓	**98.0**	**98.0**	**97.2**	**96.7**	**88.3**	**95.6**

Table 4: Ablation study of individual loss terms.

Visualization of learned feature embeddings under severe label shift. We first conduct an experiment to visualize the effectiveness of our proposed method. Samples from three classes (3, 5, and 9) are selected from USPS and MNIST datasets, following (Tong et al. 2022), to create source and target domains, respectively. The label probabilities are equal in the source domain, and $[22.9\%, 64.7\%, 12.4\%]$ in the target domain. We compare the per-class accuracy scores, Wasserstein distance $\mathcal{D}_W$, $\mathcal{D}^c_{\text{supp}}$, and 2D feature distribution of CDAN, ASA, and CASUAL . Fig. 2 shows that CASUAL achieves a higher target average accuracy, resulting in a clearer separation among classes and more distinct feature clusters compared to CDAN and ASA.

Conditional symmetric support divergence We analyze the impact of reducing the CSSD on the accuracy of different methods on the target domain. We follow the same data setting as the feature visualization experiment. Fig. 3 depicts the target accuracy and computed CSSD of 4 different methods during training. It is worth noting that as the proposed error bound 1 contains other unobservable terms, we do not claim that smaller CSSD always leads to better target domain's performance. However, we empirically observe in Fig. 3 that methods that more effectively reduce CSSD tend to achieve higher accuracy on the target domain. This result further validates the merits of aligning the supports of the conditional feature distribution, especially when label shift is severe.

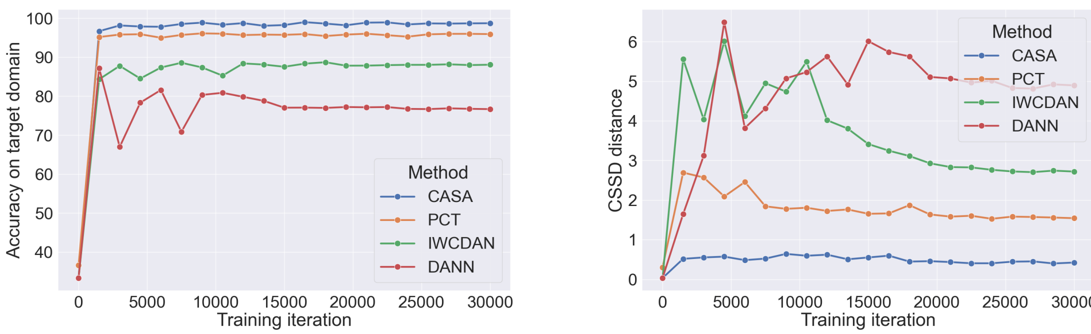


(a) CDAN (avg acc: 85%) $\mathcal{D}_W = 0.14$ $\mathcal{D}^c_{\text{supp}} = 0.13$

(b) ASA (avg acc: 93%) $\mathcal{D}_W = 0.63$ $\mathcal{D}^c_{\text{supp}} = 0.05$

(c) CASUAL (avg acc: 99%) $\mathcal{D}_W = 0.65$ $\mathcal{D}^c_{\text{supp}} = 0.02$

Figure 2: Visualization of support of feature representations for 3 classes in the USPS → MNIST task. Each plot illustrates the 2 level sets of kernel density estimates for both the source and target features.



(a) Target accuracy during training

(b) CSSD during training

Figure 3: *Left* Accuracy of various algorithms during training. *Right* Computed CSSD (Eq. 4) for learned feature.

Related Works

A dominant approach for tackling the UDA problem is learning domain-invariant feature representation, based on the theory of Ben-David et al. (2006), which suggests minimizing the $\mathcal{H}\Delta\mathcal{H}$-divergence between the two marginal distributions P_Z^S, P_Z^T. More general target risk bound than that of Ben-David et al. (2006) have been extended to multi-source domain adaptation (Zhao et al. 2018) or hypothesis-independent disparity discrepancy (Zhang et al. 2019b). Numerous methods have been proposed to align the distribution of source and target feature representation, using Wasserstein distance (Shen et al. 2018; Lee and Raginsky 2018), maximum mean discrepancy (Long et al. 2015, 2016), Jensen-Shannon divergence (Ganin and Lempitsky 2015; Tzeng et al. 2017), or first and second moment of the concerned distribution (Sun and Saenko 2016; Peng et al. 2019),.

However, recent works have pointed out the limits of enforcing invariant feature representation distribution, particularly when the marginal label distribution differs significantly between domains (Johansson, Sontag, and Ranganath 2019; Zhao et al. 2019; Wu et al. 2019; Tachet des Combes et al. 2020). Based on these theoretical results, different methods have been proposed to tackle UDA under label shift, often by minimizing β-relaxed Wasserstein distance (Tong et al. 2022; Wu et al. 2019), or estimating the importance weight of label distribution between source and target domains and aligning the conditional feature distribution instead (Lipton, Wang, and Smola 2018; Tachet des Combes et al. 2020; Azizzadenesheli et al. 2019; Kirchmeyer et al. 2022; Rakotomamonjy et al. 2022). Similarly, other methods tackle label shift by heuristics such as resampling during training and reweighing during inference (Jiang et al. 2020; Westfechtel et al. 2023; Garg et al. 2023). Different from these approaches, CASUAL aims to align the support of the conditional feature distribution, without explicitly estimating the marginal label distribution, which can be challenging under extreme label shift (Kirchmeyer et al. 2022; Rakotomamonjy et al. 2022).

Conclusion

In this paper, we propose a novel CASUAL framework to handle the UDA problem under label distribution shift. The key idea of our work is to learn a more discriminative and useful representation for the classification task by aligning the supports of the conditional distributions between the source and target domains. We next provide a novel theoretical error bound on the target domain and then introduce a complete training process for our proposed CASUAL . Our experimental results demonstrate that our CASUAL framework consistently outperforms other relevant UDA baselines on several benchmark tasks. In the future, we plan to extend our proposed method to more challenging problem settings, such as domain generalization and universal domain adaptation.

Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program(Korea University)).

References

Acuna, D.; Zhang, G.; Law, M. T.; and Fidler, S. 2021. f-domain adversarial learning: Theory and algorithms. In *International Conference on Machine Learning*, 66–75. PMLR.

Azizzadenesheli, K.; Liu, A.; Yang, F.; and Anandkumar, A. 2019. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning*, 79(1): 151–175.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2006. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19.

Chen, C.; Xie, W.; Huang, W.; Rong, Y.; Ding, X.; Huang, Y.; Xu, T.; and Huang, J. 2019. Progressive feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 627–636.

Chen, L.; Chen, H.; Wei, Z.; Jin, X.; Tan, X.; Jin, Y.; and Chen, E. 2022. Reusing the Task-specific Classifier as a Discriminator: Discriminator-free Adversarial Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7181–7190.

Courty, N.; Flamary, R.; Habrard, A.; and Rakotomamonjy, A. 2017. Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 30.

David, S. B.; Lu, T.; Luu, T.; and Pál, D. 2010. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 129–136. JMLR Workshop and Conference Proceedings.

Dhouib, S.; and Maghsudi, S. 2022. Connecting sufficient conditions for domain adaptation: source-guided uncertainty, relaxed divergences and discrepancy localization. *arXiv preprint arXiv:2203.05076*.

French, G.; Mackiewicz, M.; and Fisher, M. 2018. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1): 2096–2030.

Garg, S.; Erickson, N.; Sharpnack, J.; Smola, A.; Balakrishnan, S.; and Lipton, Z. C. 2023. Rlsbench: Domain adaptation under relaxed label shift. In *International Conference on Machine Learning*, 10879–10928. PMLR.

Gong, M.; Zhang, K.; Liu, T.; Tao, D.; Glymour, C.; and Schölkopf, B. 2016. Domain adaptation with conditional transferable components. In *International conference on machine learning*, 2839–2848. PMLR.

Hoyer, L.; Dai, D.; Wang, H.; and Van Gool, L. 2023. MIC: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11721–11732.

Jiang, X.; Lao, Q.; Matwin, S.; and Havaei, M. 2020. Implicit class-conditioned domain alignment for unsupervised domain adaptation. In *International Conference on Machine Learning*, 4816–4827. PMLR.

Johansson, F. D.; Sontag, D.; and Ranganath, R. 2019. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 527–536. PMLR.

Kirchmeyer, M.; Rakotomamonjy, A.; de Bezenac, E.; and Gallinari, P. 2022. Mapping conditional distributions for domain adaptation under generalized target shift. In *International Conference on Learning Representations*.

Lee, J.; and Raginsky, M. 2018. Minimax statistical learning with wasserstein distances. *Advances in Neural Information Processing Systems*, 31.

Li, B.; Wang, Y.; Che, T.; Zhang, S.; Zhao, S.; Xu, P.; Zhou, W.; Bengio, Y.; and Keutzer, K. 2020. Rethinking distributional matching based domain adaptation. *arXiv preprint arXiv:2006.13352*.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, 5542–5550.

Liang, J.; Hu, D.; Wang, Y.; He, R.; and Feng, J. 2021. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lipton, Z.; Wang, Y.-X.; and Smola, A. 2018. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, 3122–3130. PMLR.

Liu, H.; Long, M.; Wang, J.; and Jordan, M. 2019. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, 4013–4022. PMLR.

Liu, H.; Wang, J.; and Long, M. 2021. Cycle self-training for domain adaptation. *Advances in Neural Information Processing Systems*, 34: 22968–22981.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, 97–105. PMLR.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31.

Long, M.; Wang, J.; Ding, G.; Sun, J.; and Yu, P. S. 2014. Transfer joint matching for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1410–1417.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2016. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, 2208–2217. PMLR.

Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1979–1993.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1406–1415.

Prabhu, V.; Khare, S.; Kartik, D.; and Hoffman, J. 2021. Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8558–8567.

Rakotomamonjy, A.; Flamary, R.; Gasso, G.; Alaya, M. E.; Berar, M.; and Courty, N. 2022. Optimal transport for conditional domain matching and label shift. *Machine Learning*, 111(5): 1651–1670.

Rangwani, H.; Aithal, S. K.; Mishra, M.; Jain, A.; and Radhakrishnan, V. B. 2022. A closer look at smoothness in domain adversarial training. In *International Conference on Machine Learning*, 18378–18399. PMLR.

Shen, J.; Qu, Y.; Zhang, W.; and Yu, Y. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-second AAAI conference on artificial intelligence*.

Shu, R.; Bui, H. H.; Narui, H.; and Ermon, S. 2018. A DIRT-T approach to unsupervised domain adaptation. In *International Conference on Learning Representations*.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, 443–450. Springer.

Tachet des Combes, R.; Zhao, H.; Wang, Y.-X.; and Gordon, G. J. 2020. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33: 19276–19289.

Tanwisuth, K.; Fan, X.; Zheng, H.; Zhang, S.; Zhang, H.; Chen, B.; and Zhou, M. 2021. A prototype-oriented framework for unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 34: 17194–17208.

Tong, S.; Garipov, T.; Zhang, Y.; Chang, S.; and Jaakkola, T. S. 2022. Adversarial Support Alignment. In *International Conference on Learning Representations*.

Torralba, A.; and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR 2011*, 1521–1528. IEEE.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7167–7176.

Westfechtel, T.; Yeh, H.-W.; Meng, Q.; Mukuta, Y.; and Harada, T. 2023. Backprop Induced Feature Weighting for Adversarial Domain Adaptation with Iterative Label Distribution Alignment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 392–401.

Wu, Y.; Winston, E.; Kaushik, D.; and Lipton, Z. 2019. Domain adaptation with asymmetrically-relaxed distribution alignment. In *International conference on machine learning*, 6872–6881. PMLR.

Zhang, Q.; Zhang, J.; Liu, W.; and Tao, D. 2019a. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *Advances in Neural Information Processing Systems*, 32.

Zhang, Y.; Liu, T.; Long, M.; and Jordan, M. 2019b. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, 7404–7413. PMLR.

Zhang, Y.; Long, M.; Wang, J.; and Jordan, M. I. 2020. On localized discrepancy for domain adaptation. *arXiv preprint arXiv:2008.06242*.

Zhao, H.; Des Combes, R. T.; Zhang, K.; and Gordon, G. 2019. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, 7523–7532. PMLR.

Zhao, H.; Zhang, S.; Wu, G.; Moura, J. M.; Costeira, J. P.; and Gordon, G. J. 2018. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31.