Conditional Adversarial Domain Adaptation

Mingsheng Long†, Zhangjie Cao†, Jianmin Wang†, and Michael I. Jordan♯
†School of Software, Tsinghua University, China
†KLiss, MOE; BNRist; Research Center for Big Data, Tsinghua University, China
♯University of California, Berkeley, Berkeley, USA
{mingsheng,jimwang}@tsinghua.edu.cn caozhangjie14@gmail.com jordan@berkeley.edu

Abstract

Adversarial learning has been embedded into deep networks to learn disentangled and transferable representations for domain adaptation. Existing adversarial domain adaptation methods may struggle to align different domains of multimodal distributions that are native in classification problems. In this paper, we present conditional adversarial domain adaptation, a principled framework that conditions the adversarial adaptation models on discriminative information conveyed in the classifier predictions. Conditional domain adversarial networks (CDANs) are designed with two novel conditioning strategies: multilinear conditioning that captures the cross-covariance between feature representations and classifier predictions to improve the discriminability, and entropy conditioning that controls the uncertainty of classifier predictions to guarantee the transferability. Experiments testify that the proposed approach exceeds the state-of-the-art results on five benchmark datasets.

1 Introduction

Deep networks have significantly improved the state-of-the-arts for diverse machine learning problems and applications. When trained on large-scale datasets, deep networks learn representations which are generically useful across a variety of tasks [16, 11, 55]. However, deep networks can be weak at generalizing learned knowledge to new datasets or environments. Even a subtle change from the training domain can cause deep networks to make spurious predictions on the target domain [55, 52].

While in many real applications, there is the need to transfer a deep network from a source domain where sufficient training data is available to a target domain where only unlabeled data is available, such a transfer learning paradigm is hindered by the shift in data distributions across domains [39].

Learning a model that reduces the dataset shift between training and testing distributions is known as domain adaptation [38]. Previous domain adaptation methods in the shallow regime either bridge the source and target by learning invariant feature representations or estimating instance importances using labeled source data and unlabeled target data [24, 37, 15]. Recent advances of deep domain adaptation methods leverage deep networks to learn transferable representations by embedding adaptation modules in deep architectures, simultaneously disentangling the explanatory factors of variations behind data and matching feature distributions across domains [12, 13, 29, 53, 31, 30, 51].

Adversarial domain adaptation [12, 53, 51] integrates adversarial learning and domain adaptation in a two-player game similarly to Generative Adversarial Networks (GANs) [17]. A domain discriminator is learned by minimizing the classification error of distinguishing the source from the target domains, while a deep classification model learns transferable representations that are indistinguishable by the domain discriminator. On par with these feature-level approaches, generative pixel-level adaptation models perform distribution alignment in raw pixel space, by translating source data to the style of a target domain using Image to Image translation techniques [57, 28, 22, 43]. Another line of works align distributions of features and classes separately using different domain discriminators [23, 8, 50].

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.
Despite their general efficacy for various tasks ranging from classification \cite{12, 51, 28} to segmentation \cite{43, 50, 22}, these adversarial domain adaptation methods may still be constrained by two bottlenecks. First, when data distributions embody complex multimodal structures, adversarial adaptation methods may fail to capture such multimodal structures for a discriminative distribution alignment without mode mismatch. Such a risk comes from the equilibrium challenge of adversarial learning in that even if the discriminator is fully confused, we have no guarantee that two distributions are sufficiently similar \cite{3, 4}. Note that this risk cannot be tackled by aligning distributions of features and classes via separate domain discriminators as \cite{23, 8, 50}, since the multimodal structures can only be captured sufficiently by the cross-covariance dependency between the features and classes \cite{47, 44}. Second, it is risky to condition the domain discriminator on the discriminative information when it is uncertain.

In this paper, we tackle the two aforementioned challenges by formalizing a conditional adversarial domain adaptation framework. Recent advances in the Conditional Generative Adversarial Networks (CGANs) \cite{3, 4} disclose that the distributions of real and generated images can be made similar by conditioning the generator and discriminator on discriminative information. Motivated by the conditioning insight, this paper presents Conditional Domain Adversarial Networks (CDANs) to exploit discriminative information conveyed in the classifier predictions to assist adversarial adaptation. The key to the CDAN models is a novel conditional domain discriminator conditioned on the cross-covariance of domain-specific feature representations and classifier predictions. We further condition the domain discriminator on the uncertainty of classifier predictions, prioritizing the discriminator on easy-to-transfer examples. The overall system can be solved in linear-time through back-propagation. Based on the domain adaptation theory \cite{5}, we give a theoretical guarantee on the generalization error bound. Experiments show that our models exceed state-of-the-art results on five benchmark datasets.

## 2 Related Work

Domain adaptation \cite{38, 39} generalizes a learner across different domains of different distributions, by either matching the marginal distributions \cite{49, 37, 15} or the conditional distributions \cite{56, 10}. It finds wide applications in computer vision \cite{42, 18, 16, 21} and natural language processing \cite{9, 14}. Besides the aforementioned shallow architectures, recent studies reveal that deep networks learn more transferable representations that disentangle the explanatory factors of variations behind data \cite{6} and manifest invariant factors underlying different populations \cite{14, 56}. As deep representations can only reduce, but not remove, the cross-domain distribution discrepancy \cite{55}, recent research on deep domain adaptation further embeds adaptation modules in deep networks using two main technologies for distribution matching: moment matching \cite{52, 29, 31, 30} and adversarial training \cite{12, 53, 13, 51}.

Pioneered by the Generative Adversarial Networks (GANs) \cite{17}, the adversarial learning has been successfully explored for generative modeling. GANs constitute two networks in a two-player game: a generator that captures data distribution and a discriminator that distinguishes between generated samples and real data. The networks are trained in a minimax fashion such that the generator is learned to fool the discriminator while the discriminator struggles to be not fooled. Several difficulties of GANs have been addressed, e.g. improved training \cite{21} and mode collapse \cite{44, 7, 35}, but others still remain, e.g. failure in matching distributions \cite{41, 5}. Towards adversarial learning for domain adaptation, unconditional ones have been leveraged while conditional ones remain under explored.

Sharing some spirit of conditional GANs \cite{3, 4}, another line of works model the features and classes using separate domain discriminators. Hoffman et al. \cite{23} performs global domain alignment by learning features to deceive the domain discriminator, and category specific adaptation by minimizing a constrained multiple instance loss. In particular, the adversarial module for feature representation is not conditioned on the class-adaptation module for class information. Chen et al. \cite{8} performs class-wise alignment over the classifier layer; i.e., multiple domain discriminators take as inputs only the softmax probabilities of source classifier, rather than conditioned on the class information. Tsai et al. \cite{50} imposes two independent domain discriminators on the feature and class layers. These methods do not explore the dependency between the features and classes in a unified conditional domain discriminator, which is important to capture the multimodal structures underlying data distributions.

This paper extends the conditioning adversarial mechanism to enable discriminative and transferable domain adaptation, by defining the domain discriminator on the features while conditioning it on the class information. Two novel conditioning strategies are designed to capture the cross-covariance dependency between the feature representations and class predictions while controlling the uncertainty of classifier predictions. This is different from aligning the features and classes separately \cite{23, 8, 50}. 

2
3 Conditional Adversarial Domain Adaptation

In unsupervised domain adaptation, we are given a source domain \( D_s = \{ (x^i_s, y^i_s) \}_{i=1}^{n_s} \) of \( n_s \) labeled examples and a target domain \( D_t = \{ x^j_t \}_{j=1}^{n_t} \) of \( n_t \) unlabeled examples. The source domain and target domain are sampled from joint distributions \( P(x^s, y^s) \) and \( Q(x^t, y^t) \) respectively, while the IID assumption is violated as \( P \neq Q \). The goal of this paper is to design a deep network \( y = G(x) \) which formally reduces the shifts in the joint distributions across domains, such that the target risk \( \epsilon_t(G) = \mathbb{E}_{(x^t, y^t) \sim Q} [G(x^t) \neq y^t] \) can be bounded by the source risk \( \epsilon_s(G) = \mathbb{E}_{(x^s, y^s) \sim P} [G(x^s) \neq y^s] \) plus the distribution discrepancy \( D(P, Q) \) embodied by a novel conditional adversarial discriminator.

Adversarial learning, the key idea to enable Generative Adversarial Networks (GANs) \([12, 53]\), has been successfully explored to minimize the cross-domain discrepancy \([12, 51]\). Denote by \( f = F(x) \) the feature representation and by \( g = G(x) \) the classifier prediction generated from the deep network \( G \). Domain adversarial neural network (DANN) \([13]\) is a two-player game: the first player is the domain discriminator \( D \) trained to distinguish the source domain from the target domain and the second player is the feature representation \( F \) trained simultaneously to confuse the domain discriminator \( D \). The error function of the domain discriminator corresponds well to the discrepancy between feature distributions \( P(f) \) and \( Q(f) \) \([12]\), a key to bound the target risk in the domain adaptation theory \([5]\).

Conditional Discriminator We further improve existing adversarial domain adaptation methods \([12, 53, 51]\) in two directions. First, when the joint distributions of feature and class, i.e., \( P(x^s, y^s) \) and \( Q(x^t, y^t) \), are non-identical across domains, adapting only the feature representation \( f \) may be insufficient. Due to a quantitative study \([55]\), deep representations eventually transition from general to specific along deep networks, with transferability decreased remarkably in the domain-specific feature layer \( f \) and classifier layer \( g \). In other words, the joint distributions of feature representation \( f \) and classifier prediction \( g \) should still be non-identical in these domain adversarial networks. Second, when the feature distribution is multimodal, which is a real scenario due to the nature of multi-class classification, adapting only the feature representation may be challenging for adversarial networks. Recent work \([17, 2, 7, 1]\) reveals the high risk of failure in matching a only fraction of components underlying different distributions with adversarial networks. Namely, even if the discriminator is fully confused, we have no theoretical guarantee that two different distributions are made identical \([3, 4]\).

This paper tackles the two aforementioned challenges by formalizing a conditional adversarial domain adaptation framework. Recent advances in Conditional Generative Adversarial Networks (CGANs) \([34]\) disclose that different distributions can be matched better by conditioning the generator and discriminator on relevant information, such as associated labels and affiliated modality. Conditional GANs \([25, 35]\) generate globally coherent images from datasets with high variability and multimodal distributions. Motivated by conditional GANs, we observe that in adversarial domain adaptation, the classifier prediction \( g \) conveys discriminative information potentially revealing the multimodal structures, which can be conditioned on when adapting feature representation \( f \). By conditioning, domain variances in both feature representation \( f \) and classifier prediction \( g \) can be modeled simultaneously.

We formulate Conditional Domain Adversarial Network (CDAN) as a minimax optimization problem with two competitive error terms: (a) \( \min_D E_G \) on the source classifier \( g = G(x) \), which is minimized to guarantee lower source risk; (b) \( \min_D E_{D, G} \) on the source classifier \( g = G(x) \) and the domain discriminator \( D \) over the source and target domains, which is minimized over \( D \) but maximized over both \( f \) and \( G \):

\[
E_G = \frac{1}{n_s} \sum_{i=1}^{n_s} L(G(x^s_i), y^s_i),
\]

\[
E_{D, G} = -\frac{1}{n_s} \sum_{i=1}^{n_s} \log [D(f^s_i, g^s_i)] - \frac{1}{n_t} \sum_{j=1}^{n_t} \log [1 - D(f^t_j, g^t_j)],
\]

where \( L(\cdot, \cdot) \) is the cross-entropy loss, and \( h = (f, g) \) is the joint variable of feature representation \( f \) and classifier prediction \( g \). The minimax game of conditional domain adversarial network (CDAN) is

\[
\min_G \max_D E_G - \lambda E_{D, G}
\]

where \( \lambda \) is a hyper-parameter between the two objectives to tradeoff source risk and domain adversary.
We condition domain discriminator $D$ on the classifier prediction $g$ through joint variable $h = (f, g)$. This conditional domain discriminator can potentially tackle the two aforementioned challenges of adversarial domain adaptation. A simple conditioning of $D$ is $D(f \oplus g)$, where we concatenate the feature representation and classifier prediction in vector $f \oplus g$ and feed it to conditional domain discriminator $D$. This conditioning strategy is widely adopted by existing conditional GANs [34, 25, 35]. However, with the concatenation strategy, $f$ and $g$ are independent on each other, thus failing to fully capture multiplicative interactions between feature representation and classifier prediction, which are crucial to domain adaptation. As a result, the multimodal information conveyed in classifier prediction cannot be fully exploited to match the multimodal distributions of complex domains [47].

**Multilinear Conditioning** Multilinear map models the multiplicative interactions between different variables. The multilinear map of infinite-dimensional nonlinear feature maps has been successfully applied to embed joint distribution or conditional distribution into reproducing kernel Hilbert spaces [47, 44, 45, 39]. Given two random vectors $x$ and $y$, the joint distribution $P(x, y)$ can be modeled by the cross-covariance $E_{xy}[\phi(x) \circ \phi(y)]$, where $\phi$ is a feature map induced by some reproducing kernel. Such kernel embeddings enable easy manipulation of joint distributions, with the ability of modeling the cross-covariance, i.e., the multiplicative interactions across multiple random variables.

Besides the theoretical benefit of the multilinear map $x \circ y$ over the concatenation $x \oplus y$ [47, 46], we further explain its superiority intuitively. Assume linear map $\phi(x) = x$ and one-hot label vector $y$ in $C$ classes. As can be verified, mean map $E_{xy}[x \oplus y] = E_x[x] \oplus E_y[y]$ computes the means of $x$ and $y$ independently. In contrast, mean map $E_{xy}[x \circ y] = E_x[x|y=1] \circ \ldots \circ E_x[x|y=C]$ computes the means of each of the $C$ class-conditional distributions $P(x|y)$. Superior than concatenation, the multilinear map $x \circ y$ can fully capture the multimodal structures behind complex data distributions.

Taking the advantage of multilinear map, in this paper, we condition $D$ on $g$ with the multilinear map

$$T_\circ (f, g) = f \circ g,$$

where $T_\circ$ is a multilinear map and $D(f, g) = D(f \circ g)$. As such, the conditional domain discriminator successfully models the multimodal information and joint distributions of $f$ and $g$. Also, the multi-linearity can accommodate random vectors $f$ and $g$ with different cardinalities and magnitudes.

A disadvantage of the multilinear map is dimension explosion. Denoting by $d_f$ and $d_g$ the dimensions of vectors $f$ and $g$ respectively, the dimension of multilinear map $f \circ g$ is $d_f \times d_g$, often too high-dimensional to be embedded into deep networks without causing parameter explosion. This paper addresses the dimension explosion by randomized methods [40, 26]. Note that multilinear map holds

$$\langle T_\circ (f, g), T_\circ (f', g') \rangle = \langle f \circ g, f' \circ g' \rangle$$

$$= \langle f, f' \rangle \langle g, g' \rangle$$

$$\approx \langle T_\circ (f, g), T_\circ (f', g') \rangle,$$

where $T_\circ (f, g)$ is the explicit randomized multilinear map of dimension $d \ll d_f \times d_g$, and we define

$$T_\circ (f, g) = \frac{1}{\sqrt{d}} (R_f f) \odot (R_g g),$$
We condition the domain discriminator on the entropy-based certainty measure, which prioritizes the λ with both multilinear conditioning for discriminability and entropy conditioning for transferability is

\[
\begin{align*}
\text{minimize } & \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} L(G(x_i^s), y_i^s) \\
+ & \lambda \sum_{i=1}^{n_x} e^{-H(g_i^s)} \log [D(T(h_i^s))] + \lambda \sum_{j=1}^{n_t} e^{-H(g_j^t)} \log [1 - D(T(h_j^t))] \\
\text{maximize } & \sum_{i=1}^{n_s} e^{-H(g_i^s)} \log [D(T(h_i^s))] + \frac{1}{n_t} \sum_{j=1}^{n_t} e^{-H(g_j^t)} \log [1 - D(T(h_j^t))]
\end{align*}
\]

where \( \odot \) is element-wise product, \( R_f \) and \( R_g \) are random matrices sampled only once and fixed in training, and each element \( R_{ij} \) follows a symmetric distribution with unvariance, i.e. \( \mathbb{E}[R_{ij}] = 0, \mathbb{E}[R_{ij}^2] = 1 \). Applicable distributions include Gaussian distribution and Uniform distribution. As the inner-product on \( T_\odot \) can be accurately approximated by the inner-product on \( T_\odot \), we can directly adopt \( T_\odot(f, g) \) for computation efficiency. We guarantee such approximation quality by a theorem.

**Theorem 1.** The expectation and variance of randomized map \( T_\odot(f, g) \) for \( T_\odot(f, g) \) satisfy

\[
\text{var}[(T_\odot(f, g), T_\odot(f', g'))] = \sum_{i=1}^{d_f} \beta(R_i^f, f) \beta(R_i^g, g) + C,
\]

where \( \beta(R_i^f, f) = \frac{1}{d} \sum_{j=1}^{d_f} [f_j^2 f_j^2 \mathbb{E}[(R_{ij})^4] + C] \) and similarly for \( \beta(R_i^g, g), C, C' \) are constants.

**Proof.** The proof is given in the supplemental material.

Theorem 1 verifies that \( T_\odot \) is an unbiased estimate of \( T_\odot \), with estimation variance depending only on the fourth-order moments \( \mathbb{E}[(R_{ij})^4] \) and \( \mathbb{E}[(R_{ij})^4] \), which are constants for many symmetric distributions with unvariance, including Gaussian distribution and (centered) Uniform distribution. The bound reveals that we can further minimize the approximation error by normalizing the features.

For simplicity, we define the conditioning strategy used by the conditional domain discriminator \( D \) as

\[
T(h) = \begin{cases} 
T_\odot(f, g) & \text{if } d_f \times d_g \leq 4096 \\
T_\odot(f, g) & \text{otherwise},
\end{cases}
\]

where 4096 is the largest number of units in typical deep networks (e.g. AlexNet), and if dimension of the multilinear map \( T_\odot \) is larger than 4096, then we will opt to randomized multilinear map \( T_\odot \).

**Entropy Conditioning** The minimax problem for conditional domain discriminator may be problematic, since the domain discriminator imposes equal importance for different examples, while hard-to-transfer examples with uncertain predictions may deteriorate the adversarial adaptation procedure. Towards safe transfer, we quantify the uncertainty of classifier predictions by the entropy criterion \( H(g) = -\sum_{c=1}^{C} g_c \log g_c \), where \( C \) is the number of classes and \( g_c \) is the probability of predicting an example to class \( c \). The certainty of predictions can be computed by \( e^{-H(g_i)} \in [\frac{1}{e}, 1] \).

We condition the domain discriminator on the entropy-based certainty measure, which prioritizes the discriminator on those easy-to-transfer examples with certain predictions. The final discriminator with both multilinear conditioning for discriminability and entropy conditioning for transferability is

\[
E_{D, G} = -\frac{1}{n_s} \sum_{i=1}^{n_s} e^{-H(g_i^s)} \log [D(T(h_i^s))] - \frac{1}{n_t} \sum_{j=1}^{n_t} e^{-H(g_j^t)} \log [1 - D(T(h_j^t))].
\]

The discriminator encourages certain predictions, satisfying the entropy minimization principle [19].

**Conditional Domain Adversarial Network** We enable conditional adversarial domain adaptation over the domain-specific feature representation \( f \) and classifier prediction \( g \). We jointly minimize error (1) with respect to the source classifier \( G \), minimize error (10) with respect to domain discriminator \( D \), and maximize error (10) with respect to the feature extractor \( F \) and the source classifier \( G \). This leads to the minimax problem for the proposed Conditional Domain Adversarial Networks (CDANs):

\[
\begin{align*}
\min_G & \frac{1}{n_s} \sum_{i=1}^{n_s} L(G(x_i^s), y_i^s) \\
+ & \lambda \sum_{i=1}^{n_x} e^{-H(g_i^s)} \log [D(T(h_i^s))] + \lambda \sum_{j=1}^{n_t} e^{-H(g_j^t)} \log [1 - D(T(h_j^t))] \\
\max_D & \frac{1}{n_s} \sum_{i=1}^{n_s} e^{-H(g_i^s)} \log [D(T(h_i^s))] + \frac{1}{n_t} \sum_{j=1}^{n_t} e^{-H(g_j^t)} \log [1 - D(T(h_j^t))],
\end{align*}
\]

where \( \lambda \) is a hyper-parameter between source classifier and conditional domain discriminator, and note that \( h = (f, g) \) is the joint variable of domain-specific feature representation \( f \) and classifier prediction \( g \) for adversarial adaptation. As a rule of thumb, we can safely set \( f \) as the last feature layer representation and \( g \) as the classifier layer prediction. In cases where lower-layer features are not transferable as in pixel-level adaptation tasks [25, 22], we can change \( f \) to lower-layer representations.
Generalization Error Bound We provide an analysis of our method taking similar formalism of the domain adaptation theory [5]. We first consider fixed source and target over the representation space $f = F(x)$, and a family of source classifiers $G \in \mathcal{H}$ [13]. Denote by $\epsilon_p (G) = \mathbb{E}_{(f,y) \sim P} [G(f) \neq y]$ the risk of a hypothesis $G \in \mathcal{H}$ w.r.t. distribution $P$, and $\epsilon_p (G_1, G_2) = \mathbb{E}_{(f,y) \sim P} [G_1(f) \neq G_2(f)]$ the disagreement between hypotheses $G_1, G_2 \in \mathcal{H}$. Let $G^* = \arg \min_{G} \epsilon_p (G) + \epsilon_Q (G)$ be the ideal hypothesis that explicitly embodies the notion of adaptability. The probabilistic bound [5] of the target risk $\epsilon_Q (G)$ of hypothesis $G$ is given by the source risk $\epsilon_p (G)$ plus the distribution discrepancy:

$$\epsilon_Q (G) \leq \epsilon_p (G) + \epsilon_Q (G^*) + \left| \epsilon_p (G, G^*) - \epsilon_Q (G, G^*) \right|.$$  

(12)

The goal of domain adaptation is to reduce the distribution discrepancy $|\epsilon_p (G, G^*) - \epsilon_Q (G, G^*)|$. By definition, $\epsilon_p (G, G^*) = \mathbb{E}_{(f,y) \sim P} [G(f) \neq G^*(f)] = \mathbb{E}_{(f,g) \sim P_{G}} [g \neq G^*(f)] = \epsilon_{P_{G}} (G^*)$, and similarly, $\epsilon_Q (G, G^*) = \epsilon_{Q_{G}} (G^*)$. Define a (loss) difference hypothesis space $\Delta \triangleq \{\delta = |g - G^*(f)| | G^* \in \mathcal{H}\}$ over the joint variable $(f, g)$, where $\delta : (f, g) \mapsto \{0, 1\}$ outputs the loss of $G^* \in \mathcal{H}$. Based on the above difference hypothesis space $\Delta$, we define the $\Delta$-distance as

$$d_\Delta (P_G, Q_G) \triangleq \sup_{\delta \in \Delta} \left| \mathbb{E}_{(f,g) \sim P_G} \left[ \delta (f, g) \neq 0 \right] - \mathbb{E}_{(f,g) \sim Q_G} \left[ \delta (f, g) \neq 0 \right] \right|$$

$$\geq \mathbb{E}_{(f,g) \sim P_G} \left| g - G^* (f) \right| - \mathbb{E}_{(f,g) \sim Q_G} \left| g - G^* (f) \right| = \left| \epsilon_{P_G} (G^*) - \epsilon_{Q_G} (G^*) \right|.$$  

(13)

Hence, the domain discrepancy $|\epsilon_p (G, G^*) - \epsilon_Q (G, G^*)|$ can be upper-bounded by the $\Delta$-distance. Since the difference hypothesis space $\Delta$ is a continuous function class, assume the family of domain discriminators $\mathcal{H}_D$ is rich enough to contain $\Delta, \Delta \subset \mathcal{H}_D$. Such an assumption is not unrealistic as we have the freedom to choose $\mathcal{H}_D$, for example, a multilayer perceptrons that can fit any functions. Given the assumptions hold, we show that training domain discriminator $D$ is related to $d_\Delta (P_G, Q_G)$:

$$d_\Delta (P_G, Q_G) \leq \sup_{D \in \mathcal{H}_D} \left| \mathbb{E}_{(f,g) \sim P_G} \left[ D (f, g) \neq 0 \right] - \mathbb{E}_{(f,g) \sim Q_G} \left[ D (f, g) \neq 0 \right] \right|$$

$$\leq \sup_{D \in \mathcal{H}_D} \left| \mathbb{E}_{(f,g) \sim P_G} \left[ D (f, g) = 1 \right] + \mathbb{E}_{(f,g) \sim Q_G} \left[ D (f, g) = 0 \right] \right|.$$  

(14)

This is maximized by the optimal discriminator $D$ in CDANs, giving the upper bound of $d_\Delta (P_G, Q_G)$. Simultaneously, we learn representation $f$ to minimize $d_\Delta (P_G, Q_G)$, yielding better approximation of $\epsilon_Q (G)$ by $\epsilon_p (G)$. As verified above, the proposed CDAN models formally bound the target risk.

4 Experiments

We evaluate the proposed conditional domain adversarial networks with many state-of-the-art transfer learning and deep learning methods. Codes and datasets will be available at github.com/thuml

4.1 Setup

Office-31 [42] is the most widely used dataset for visual domain adaptation, with 4,652 images and 31 categories collected from three distinct domains: Amazon (A), Webcam (W) and DSLR (D). We evaluate all methods on six transfer tasks $A \rightarrow W, D \rightarrow W, W \rightarrow D, A \rightarrow D, D \rightarrow A,$ and $W \rightarrow A$.

ImageCLEF-DA⁷ is a dataset organized by selecting the 12 common classes shared by three public datasets (domains): Caltech-256 (C), ImageNet ILSVRC 2012 (I), and Pascal VOC 2012 (P). There are 50 images in each category and 600 images in each domain, while Office-31 has different domain sizes. We permute domains and build 6 transfer tasks: I $\rightarrow$ P, P $\rightarrow$ I, I $\rightarrow$ C, C $\rightarrow$ I, C $\rightarrow$ P, P $\rightarrow$ C.

Office-Home [53] is a better organized but more difficult dataset than Office-31, which consists of 15,500 images in 65 object classes in office and home settings, forming four extremely dissimilar domains: Artistic images (Ar), Clip Art (Cl), Product images (Pr), and Real-world images (Rw).

Digits We investigate three digits datasets: MNIST, USPS, and Street View House Numbers (SVHN). We adopt the evaluation protocol of CyCADA [22] with three transfer tasks: USPS to MNIST (U $\rightarrow$ M), MNIST to USPS (M $\rightarrow$ U), and SVHN to MNIST (S $\rightarrow$ M). We train our model using the training sets: MNIST (60,000), USPS (7,291), standard SVHN train (73,257). Evaluation is reported on the standard test sets: MNIST (10,000), USPS (2,007) (the numbers of images in the parentheses).

⁷http://imageclef.org/2014/adaptation
VisDA-2017 is a challenging simulation-to-real dataset, with two very distinct domains: Synthetic, renderings of 3D models from different angles and with different lightning conditions; Real, natural images. It contains over 280K images across 12 classes in the training, validation and test domains.

We compare Conditional Domain Adversarial Network (CDAN) with state-of-art domain adaptation methods: Deep Adaptation Network (DAN) [29], Residual Transfer Network (RTN) [31], Domain Adversarial Neural Network (DANN) [13], Adversarial Discriminative Domain Adaptation (ADDA) [31], Joint Adaptation Network (JAN) [30], Unsupervised Image-to-Image Translation (UNIT) [28], Generate to Adapt (GTA) [43], Cycle-Consistent Adversarial Domain Adaptation (CyCADA) [22].

We follow standard protocols for unsupervised domain adaptation [12, 30]. We use all labeled source examples and all unlabeled target examples for all datasets. We compare the average classification accuracy of each method on three random experiments, and report the standard error by all experiments of the same transfer task. We conduct the importance-weighted cross-validation [48] for all baseline methods and for our CDAN models to select hyper-parameter $\lambda$. As our CDAN models perform stably under different parameter configurations, we keep fixing $\lambda = 1$ throughout all experiments. For MMD-based methods (TCA, DAN, RTN, and JAN), we use the Gaussian kernel with bandwidth set to the median pairwise squared distances on the training data [29]. We examine the influence of deep architectures by exploring AlexNet [27] and ResNet-50 [20] as base architectures, respectively.

We implement all AlexNet-based methods by Caffe and all ResNet-based methods by PyTorch. We fine-tune from the AlexNet and ResNet models pre-trained on the ImageNet dataset [41]. We fine-tune all lower layers and train the classifier layer through back-propagation, where the classifier is trained from scratch with learning rate 10 times that of the lower layers. We adopt mini-batch SGD with momentum of 0.9 and the learning rate decay strategy of DANN [13]: the learning rate is adjusted from scratch with learning rate $10 \times$ that of the lower layers. We adopt mini-batch SGD with all lower layers and train the classifier layer through back-propagation, where the classifier is trained fine-tune from the AlexNet and ResNet models pre-trained on the ImageNet dataset [41]. We fine-tune deep architectures by exploring AlexNet [27] and ResNet-50 [20] as base architectures, respectively.

The results on Office-31 for unsupervised domain adaptation (AlexNet and ResNet) are reported in Table 1 with results of baselines directly reported from their original papers wherever available. The CDAN models significantly outperform all comparison methods on most transfer tasks, where CDAN-M is the top-performing variant and CDAN-RM performs slightly worse. It is desirable that CDAN promotes the classification accuracies substantially on hard transfer tasks, e.g. A $\rightarrow$ W and A $\rightarrow$ D, where the source and target domains are substantially different [42]. Note that, CDAN even outperforms generative pixel-level domain adaptation method GTA, which has a very complex design in both architecture and objectives.

The results on the ImageCLEF-DA dataset are reported in Table 2. Our CDAN models outperform the comparison methods on most transfer tasks, but with smaller rooms of improvement. This is

Table 1: Accuracy (%) on Office-31 for unsupervised domain adaptation (AlexNet and ResNet)

<table>
<thead>
<tr>
<th>Method</th>
<th>A $\rightarrow$ W</th>
<th>D $\rightarrow$ W</th>
<th>W $\rightarrow$ D</th>
<th>A $\rightarrow$ D</th>
<th>A $\rightarrow$ W</th>
<th>A $\rightarrow$ D</th>
<th>W $\rightarrow$ A</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>61.6±0.5</td>
<td>95.4±0.3</td>
<td>99.0±0.2</td>
<td>63.8±0.5</td>
<td>51.1±0.6</td>
<td>49.8±0.4</td>
<td>70.1</td>
<td></td>
</tr>
<tr>
<td>DAN</td>
<td>68.5±0.5</td>
<td>96.0±0.3</td>
<td>99.0±0.3</td>
<td>67.0±0.4</td>
<td>54.0±0.5</td>
<td>53.1±0.5</td>
<td>72.9</td>
<td></td>
</tr>
<tr>
<td>RTN</td>
<td>73.3±0.3</td>
<td>96.8±0.2</td>
<td>99.6±0.1</td>
<td>71.0±0.2</td>
<td>50.5±0.3</td>
<td>51.0±0.1</td>
<td>73.7</td>
<td></td>
</tr>
<tr>
<td>DANN</td>
<td>73.0±0.5</td>
<td>96.4±0.3</td>
<td>99.2±0.3</td>
<td>72.3±0.3</td>
<td>53.4±0.4</td>
<td>51.2±0.5</td>
<td>74.3</td>
<td></td>
</tr>
<tr>
<td>ADDA</td>
<td>73.5±0.6</td>
<td>96.2±0.4</td>
<td>98.8±0.4</td>
<td>71.6±0.4</td>
<td>54.6±0.5</td>
<td>53.5±0.6</td>
<td>74.7</td>
<td></td>
</tr>
<tr>
<td>JAN</td>
<td>74.9±0.3</td>
<td>96.6±0.2</td>
<td>99.5±0.2</td>
<td>71.8±0.2</td>
<td>58.3±0.3</td>
<td>55.0±0.4</td>
<td>76.0</td>
<td></td>
</tr>
<tr>
<td>CDAN-RM</td>
<td>77.9±0.3</td>
<td>96.9±0.2</td>
<td>100.0±0.0</td>
<td>75.1±0.2</td>
<td>54.5±0.3</td>
<td>57.5±0.4</td>
<td>77.0</td>
<td></td>
</tr>
<tr>
<td>CDAN-M</td>
<td>78.3±0.2</td>
<td>97.2±0.1</td>
<td>100.0±0.0</td>
<td>76.3±0.1</td>
<td>57.3±0.2</td>
<td>57.3±0.3</td>
<td>77.7</td>
<td></td>
</tr>
<tr>
<td>ResNet-50</td>
<td>68.4±0.2</td>
<td>96.7±0.1</td>
<td>99.3±0.1</td>
<td>68.9±0.2</td>
<td>62.5±0.3</td>
<td>60.7±0.3</td>
<td>76.1</td>
<td></td>
</tr>
<tr>
<td>DAN</td>
<td>80.5±0.4</td>
<td>97.1±0.2</td>
<td>99.6±0.1</td>
<td>78.6±0.2</td>
<td>63.6±0.3</td>
<td>62.8±0.2</td>
<td>80.4</td>
<td></td>
</tr>
<tr>
<td>RTN</td>
<td>84.5±0.2</td>
<td>96.8±0.1</td>
<td>99.4±0.1</td>
<td>77.5±0.3</td>
<td>66.2±0.2</td>
<td>64.8±0.3</td>
<td>81.6</td>
<td></td>
</tr>
<tr>
<td>DANN</td>
<td>82.0±0.4</td>
<td>96.9±0.2</td>
<td>99.1±0.1</td>
<td>79.7±0.4</td>
<td>68.2±0.4</td>
<td>67.4±0.5</td>
<td>82.2</td>
<td></td>
</tr>
<tr>
<td>ADDA</td>
<td>86.2±0.5</td>
<td>96.2±0.3</td>
<td>98.4±0.3</td>
<td>77.8±0.3</td>
<td>69.5±0.4</td>
<td>68.9±0.5</td>
<td>82.9</td>
<td></td>
</tr>
<tr>
<td>JAN</td>
<td>85.4±0.3</td>
<td>97.4±0.2</td>
<td>99.8±0.2</td>
<td>84.7±0.3</td>
<td>68.6±0.3</td>
<td>70.0±0.4</td>
<td>84.3</td>
<td></td>
</tr>
<tr>
<td>GTA</td>
<td>89.5±0.5</td>
<td>97.9±0.3</td>
<td>99.8±0.4</td>
<td>87.7±0.5</td>
<td>72.8±0.3</td>
<td>71.4±0.4</td>
<td>86.5</td>
<td></td>
</tr>
<tr>
<td>CDAN-RM</td>
<td>93.0±0.2</td>
<td>98.4±0.2</td>
<td>100.0±0.0</td>
<td>89.2±0.3</td>
<td>70.2±0.4</td>
<td>67.4±0.4</td>
<td>86.4</td>
<td></td>
</tr>
</tbody>
</table>
| CDAN-M     | 93.1±0.1           | 98.6±0.1           | 100.0±0.0          | 92.9±0.2           | 71.0±0.3           | 69.3±0.3           | 87.5               

4.2 Results

The results on Office-31 based on AlexNet and ResNet are reported in Table 1 with results of baselines directly reported from their original papers wherever available. The CDAN models significantly outperform all comparison methods on most transfer tasks, where CDAN-M is the top-performing variant and CDAN-RM performs slightly worse. It is desirable that CDAN promotes the classification accuracies substantially on hard transfer tasks, e.g. A $\rightarrow$ W and A $\rightarrow$ D, where the source and target domains are substantially different [42]. Note that, CDAN even outperforms generative pixel-level domain adaptation method GTA, which has a very complex design in both architecture and objectives.

The results on the ImageCLEF-DA dataset are reported in Table 2. Our CDAN models outperform the comparison methods on most transfer tasks, but with smaller rooms of improvement. This is
We study sampling strategies of the random matrices in Equation (6). We testify an Ablation Study Table 4. Note that the generative pixel-level adaptation methods UNIT, CyCADA, and GTA are specifically tailored to the digits and synthetic to real adaptation tasks. This explains why the previous approach that works reasonably well on all five datasets, and remains a simple discriminative model. It is desirable that CDAN is the only comparison methods on most transfer tasks, and with larger rooms of improvement. An interpretation is that the four domains in Office-Home are of equal size and balanced in each category, and are visually more similar than Office-31, making the former dataset easier for domain adaptation.

Strong results are also achieved on the digits datasets and synthetic to real datasets as reported in Table 3. Note that the generative pixel-level adaptation methods UNIT, CyCADA, and GTA are specifically tailored to the digits and synthetic to real adaptation tasks. This explains why the previous feature-level adaptation method JAN performs much worse. To our knowledge, CDAN is the only approach that works reasonably well on all five datasets, and remains a simple discriminative model.

### 4.3 Analysis

**Ablation Study** We study sampling strategies of the random matrices in Equation (6). We testify CDAN-RM (Gaussian) and CDAN-RM (Uniform) with their random matrices sampled only once from Gaussian and Uniform distributions, respectively. Table 5 shows that CDAN-RM (Uniform) performs better across the variants. We further investigate CDAN-M (w/o Entropy) and CDAN-M (w/ Entropy). Table 5 depicts that CDAN-M (w/ Entropy) outperforms CDAN-M (w/o Entropy), validating the efficacy of the entropy conditioning strategy for exploring easy-to-transfer examples.

**Conditioning Strategies** Besides multilinear conditioning, we investigate DANN-f and DANN-g with domain discriminator plugged in feature layer f and classifier layer g. DANN-[f,g] with domain discriminator imposed on the concatenation of f and g. Figure 2(a) shows the accuracies on A → W
Table 5: Accuracy (%) on Office-31 of CDAN variants for unsupervised domain adaptation (ResNet)

<table>
<thead>
<tr>
<th>Method</th>
<th>A → W</th>
<th>D → W</th>
<th>W → D</th>
<th>A → D</th>
<th>D → A</th>
<th>W → A</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDAN-RM (Gaussian)</td>
<td>91.8±0.1</td>
<td>97.4±0.1</td>
<td>99.7±0.1</td>
<td>87.4±0.2</td>
<td>70.6±0.3</td>
<td>69.0±0.3</td>
<td>86.0</td>
</tr>
<tr>
<td>CDAN-RM (Uniform)</td>
<td><strong>93.0±0.2</strong></td>
<td><strong>98.4±0.2</strong></td>
<td>100.0±0.0</td>
<td><strong>89.2±0.3</strong></td>
<td>70.2±0.4</td>
<td>67.4±0.4</td>
<td><strong>86.4</strong></td>
</tr>
<tr>
<td>CDAN-M (w/o Entropy)</td>
<td>91.7±0.2</td>
<td>98.3±0.1</td>
<td>100.0±0.0</td>
<td>92.5±0.2</td>
<td>70.0±0.2</td>
<td>67.8±0.2</td>
<td>86.8</td>
</tr>
<tr>
<td>CDAN-M (w/ Entropy)</td>
<td><strong>93.1±0.1</strong></td>
<td><strong>98.6±0.1</strong></td>
<td><strong>100.0±0.0</strong></td>
<td><strong>92.9±0.2</strong></td>
<td><strong>71.0±0.3</strong></td>
<td><strong>69.3±0.3</strong></td>
<td><strong>87.5</strong></td>
</tr>
</tbody>
</table>

Figure 2: Empirical analysis of conditioning strategies, distribution discrepancy, and convergence.

Figure 3: T-SNE of features by (a) ResNet, (b) DANN, (c) CDAN-f, (d) CDAN-fg (red: A; blue: W).

and $A \rightarrow D$ based on ResNet-50. The concatenation strategy is not successful, as it cannot capture the cross-covariance between features and classes, which are crucial to domain adaptation [10]. Figure 2(b) shows that the entropy weight $e^{-H(g)}$ corresponds well with the prediction correctness: entropy weight $\approx 1$ when the prediction is correct, and much smaller than 1 when prediction is incorrect (uncertain). This reveals the power of the entropy conditioning to guarantee example transferability.

Distribution Discrepancy The $A$-distance is widely used as a measure of distribution discrepancy [5] [33], defined as $d_A = 2(1 - 2c)$, where $c$ is the test error of a classifier trained to discriminate source and target. Figure 2(c) shows $d_A$ on tasks $A \rightarrow W$, $W \rightarrow D$ with features of ResNet, DANN, and CDAN-M. We observe that $d_A$ on CDAN-M features is smaller than $d_A$ on ResNet and DANN features, implying that CDAN-M features can reduce the domain gap more effectively. As domains $W$ and $D$ are similar, $d_A$ of task $W \rightarrow D$ is smaller than that of $A \rightarrow W$, implying higher accuracies.

Convergence We testify the convergence of ResNet, DANN, and CDANs, with the test errors on task $A \rightarrow W$ shown in Figure 2(d). CDAN enjoys faster convergence than DANN, while CDAN-M converges faster than CDAN-RM. Note that CDAN-M deals with high-dimensional multilinear map, thus each iteration costs slightly more than CDAN-RM, while CDAN-RM has similar cost as DANN.

Visualization We visualize by t-SNE [32] in Figures 3(a)[3(d)] the activations of task $A \rightarrow W$ (31 classes) by ResNet, DANN, CDAN-f, and CDAN-fg. The source and target are not aligned well with ResNet, better aligned with DANN but categories are not discriminated well. They are aligned better and categories are discriminated better by CDAN-f, while CDAN-fg is evidently better than CDAN-f. This shows the benefit of conditioning domain adversarial adaptation on discriminative predictions.

5 Conclusion

This paper presented conditional domain adversarial network (CDAN), a novel approach to domain adaptation with multimodal distributions. Unlike previous adversarial adaptation methods that solely match the feature representation across domains which is prone to under-matching, the proposed approach further conditions the adversarial domain adaptation on discriminative information to enable alignment of multimodal distributions. Experiments validated the efficacy of the proposed approach.

Acknowledgments

We thank Yuchen Zhang at Tsinghua University for insightful discussions. This work was supported by the National Key R&D Program of China (No. 2016YFB1000701), Natural Science Foundation of China (61772299, 61502265, 71690231) and DARPA Program on Lifelong Learning Machines.
References


