Domain Adaptation via Maximizing Surrogate Mutual Information

Haiteng Zhao, Chang Ma, Qinyu Chen and Zhi-Hong Deng
Peking University
{zhaohaiteng,changma,chenqinyu,zhdeng}@pku.edu.cn

Abstract
Unsupervised domain adaptation (UDA) aims to predict unlabeled data from target domain with access to labeled data from the source domain. In this work, we propose a novel framework called SIDA (Surrogate Mutual Information Maximization Domain Adaptation) with strong theoretical guarantees. To be specific, SIDA implements adaptation by maximizing mutual information (MI) between features. In the framework, a surrogate joint distribution models the underlying joint distribution of the unlabeled target domain. Our theoretical analysis validates SIDA by bounding the expected risk on target domain with MI and surrogate distribution bias. Experiments show that our approach is comparable with state-of-the-art unsupervised adaptation methods on standard UDA tasks.

1 Introduction
Inspired by human beings’ ability to transfer knowledge across domains and tasks, transfer learning is proposed to leverage knowledge from source domain and task to improve performance on target domain and task. However, in practice, labeled data are often limited on target domains. To address such situation, unsupervised domain adaptation (UDA), a category of transfer learning methods [Long et al., 2015; Long et al., 2017b; Ganin et al., 2016], attempts to enhance knowledge transfer from labeled source domain to target domain by leveraging unlabeled target domain data.

Most previous work is based on the data shift assumption, i.e., the label space maintains the same across domains, but the data distribution conditioned on labels varies. Under this hypothesis, domain alignment and class-level method are used to improve generalization across source and target feature distributions. Domain alignment minimizes the discrepancy between the feature distributions of two domains [Long et al., 2015; Ganin et al., 2016; Long et al., 2017b], while class-level methods work on conditional distributions. Conditional alignment aligns conditional distributions and use pseudo-labels to estimate conditional distribution on target domain [Long et al., 2018; Li et al., 2020c; Chen et al., 2020a]. However, the conditional distributions from different categories tend to mix together, leading to performance drop.

Contrastive learning based methods resolve this issue by discriminating features from different classes [Luo et al., 2020], but still face the problem of pseudo-label precision. In addition, most of the class-level methods lack solid theoretical explanations for the relationship between cross domain generalization and their objectives. Some works [Chen et al., 2019; Xie et al., 2018] yield some intuition for conditional alignment and contrastive learning, but the relation between their training objectives and cross-domain error remains unclear.

In this work, we aim to address the generalization problem in domain adaptation from an information theory perspective. In failed case of domain adaptation, as shown in Figure 1, features from the same class do not represent each other well and this inspires us to use mutual information to reduce this confusion. Our motivation is to find more representative features for both domains by maximizing mutual information between features of the same class (on both source and target domains). Therefore, if our classifier can accurately predict features on source domain, then it would also function well on target domains where features share enough information with the source features.

Based on the above motivation, we propose Surrogate Information Domain Adaptation (SIDA), a general domain adaptation framework with strong theoretical guarantees. SIDA achieves adaptation by maximizing the mutual information (MI) between features within the same class, which improves the generalization of the model to the target domain. Furthermore, a surrogate distribution is constructed to approximate the unlabeled target distribution, which improves flexibility for selecting data and assists MI estimation. Also, our theoretical analyses directly establish a bound between MI of features and target expected risk, giving a proof that our model can improve generalization across domain.

Our novelties and contributions are summarized as follows:

- We propose a novel framework to achieve domain adaptation by maximizing surrogate MI.
- We establish an expected risk upper bound based on feature MI and surrogate distribution bias for UDA. This provides theoretical guarantee for our framework.
- Experiment results on three challenging benchmarks demonstrate that our method performs favorably against state-of-art class-level UDA models.
2 Related Work

Domain Adaptation Prior works are based on two major assumptions: (1) the label shift hypothesis, where the label distribution changes, and (2) a more common data shift hypothesis where we only study the shift in conditional distribution where the label distribution is fixed. Our work focuses on the data shift hypothesis, and previous work following this line can be divided into two major categories: domain alignment methods which align marginal distributions, and class-level methods addressing the alignment of conditional distributions.

Domain alignment methods minimize the difference between feature distributions of source and target domains with various metrics, e.g., maximum mean discrepancy (MMD) [Long et al., 2015], JS divergence [Ganin et al., 2016] estimated by adversarial discriminator, Wasserstein metric and others. Maximum Mean Discrepancy (MMD) is applied to measure the discrepancy in marginal distributions [Long et al., 2015; Long et al., 2017b]. Adversarial domain adaptation plays a mini-max game to learn domain-invariant features [Ganin et al., 2016; Li et al., 2020b].

Class-level methods align the conditional distribution based on pseudo-labels [Li et al., 2020c; Chen et al., 2020a; Luo et al., 2020; Li et al., 2020a; Tang and Jia, 2020; Xu et al., 2019]. Conditional alignment methods [Xie et al., 2018; Long et al., 2018] minimize the discrepancy between conditional distributions. In class-level methods, conditional distributions are assigned by pseudo-labels. The accuracy of pseudo-labels greatly influences performance and later works construct more accurate pseudo-labels [Chen et al., 2020a]. However, the major problem with this method is that error in conditional alignment leads to distribution overlap of features from different class, resulting in low discriminability on target domain. Contrastive learning addresses this problem by maximizing the discrepancy between different classes [Luo et al., 2020; Li et al., 2020a]. However, the performance of contrastive learning also relies on pseudo-labeling.

In addition, previous class-level works provide weak theoretical support for cross-domain generalization. Prior works mainly focus on domain alignment [Ben-David et al., 2007; Redko et al., 2020]. Some works [Chen et al., 2019; Xie et al., 2018] consider optimal classification on both domains, and yield some intuitive explanation for conditional alignment and contrastive learning, but the relation between their objective function and theoretical cross-domain error remains unclear.

Information Maximization Principle Recently, mutual information maximization (InfoMax) for representation learning has attracted lots of attention [Chen et al., 2020b; Hjelm et al., 2018; Khosla et al., 2020]. The intuition is that two features belonging to different classes should be discriminable while features of the same class should resemble each other. The InfoMax principle provides a general framework for learning informative representations, and provides consistent boosts in various downstream tasks.

We facilitate domain adaptation with MI maximization, i.e. maximizing the MI between features of the same class. Some works solve domain adaptation problem via information theoretical methods [Thota and Leontidis, 2021; Chen and Liu, 2020; Park et al., 2020], which maximize MI using InfoNCE estimation [Poole et al., 2019]. As far as we know, we are the first to provide theoretical guarantee for the target domain expected risk based on MI. Compared with InfoNCE, the variational lower bound of MI we use is tighter [Poole et al., 2019]. We also construct a surrogate distribution as a substitute for unlabeled target domain, which is more suitable for MI estimation.

3 Preliminaries

3.1 Notations and Problem Setting

Let \( \mathcal{X} \) be the data space and \( \mathcal{Y} \) be the label space. In UDA, there is a source distribution \( P_S(X,Y) \) and a target distribution \( P_T(X,Y) \) on \( \mathcal{X} \times \mathcal{Y} \). Note that distributions are also referred to as domains in UDA. Our work is based on the data shift hypothesis, which assumes \( P_S(X,Y) \) and \( P_T(X,Y) \) satisfy the following properties: \( P_T(Y) = P_S(Y) \) and \( P_T(X|Y) \neq P_S(X|Y) \).

In our work, we focus on classification tasks. Under this setting, an algorithm has access to \( n_S \) labeled samples \( \{(x_i^S,y_i^S)\}_{i=1}^{n_S} \sim P_S(X,Y) \) and \( n_T \) unlabeled samples \( \{(x_i^T)\}_{i=1}^{n_T} \sim P_T(X) \), and outputs a hypothesis composed of an encoder \( G \) and a classifier \( F \). Let \( Z \) be the feature space. The encoder maps data to feature space, denoted by \( G: \mathcal{X} \rightarrow Z \). Then the classifier maps the feature to a corresponding class, \( F: Z \rightarrow \mathcal{Y} \).

For brevity, given encoder \( G \) and data-label distribution \( P(X,Y) \), denote the distribution of \( G \)-encoded feature and label by \( P^G \), i.e. \( P^G(z,y) = P(x = G^{-1}(z), y) \).
Let $F$ be a hypothesis, and $P$ be the distribution of feature and label. The expected risk of a $F$ w.r.t. $P$ is denoted as

$$\epsilon_P(F) \triangleq \mathbb{E}_{P(z)}[\delta_F(z) - P(y|z)]_1,$$

where $\delta_F(z)$ equals to 1 if $y = F(z)$ and equals 0 in else cases. Our objective is to minimize the expected risk of $F$ on target feature distribution encoded by $G$,

$$\min_{G,F} \epsilon_{P_T^G}(F).$$

4 Methodology

4.1 Overview

In UDA task, the model needs to generalize across different domains with varying distributions; thus the encoder needs to extract appropriate features that are transferable across domains. The challenges of class-level adaptation are two folds: learning transferable features, and modeling $P_T$ without label information.

To solve the first problem, we use MI based methods. Following the InfoMax principle, we maximize the mutual information between features from the same class on the target and source mixture distribution. This encourages the features of the source domain to carry more information about the features of the same class in target domain, and thus provides opportunities for transferring classifier across domains.

As for the second challenge, we first revisit the data shift hypothesis. The distribution of labels $P(Y)$ remains independent of domains; therefore the key is to model the conditional distribution $P(Z|Y)$ on the target domain. However, modeling $P(Z|Y)$ is intractable, since labels on the target domain are inaccessible. To tackle this problem, we model a surrogate distribution $Q(Z|Y)$ instead.

We introduce the goal of maximizing MI in section 4.2, and theoretically explain how MI affects domain adaptation risk. In Section 4.3, we will introduce the model in detail, including the variational estimation of MI, the modeling of the surrogate distribution, and the optimization of the loss function of the model.

4.2 Mutual Information Maximization

MI measures the degree to which two variables can predict each other. Inspired by InfoMax principle [Hall et al., 2018], we maximize the MI between the features within the same class. It encourages features from different classes to be discriminable from each other.

We maximize MI between features on both source and target domain, regardless of which domain they come from. So we introduce mixture distribution $S + T$ of both domain, which is

$$P_{S+T}(x, y) \triangleq \frac{1}{2}(P_S(x, y) + P_T(x, y)).$$

Note that because $P_S(y) = P_T(y) = P_{S+T}(y)$,

$$P_{S+T}(x, y) = \frac{1}{2}P_S(x \mid y) + \frac{1}{2}P_T(x \mid y).$$

Define the distribution of features from the same class as

$$P_{S+T}^G(z_1, z_2 | y) = P_{S+T}^G(z_1 | y)P_{S+T}^G(z_2 | y),$$

$$P_{S+T}^G(z_1, z_2) = \sum_y P_{S+T}^G(y)P_{S+T}^G(z_1, z_2 | y).$$

which means the feature $z_1$ and $z_2$ are sampled independently from the conditional distribution of the same class, with equal probability from source domain and target domain.

MI between features is maximized within the mixture distribution, as formalized bellow:

$$\arg \max_G I^G_{S+T}(Z_1; Z_2) \quad \text{(5)}$$

$$= \int P_{S+T}^G(z_1, z_2) \log \frac{P_{S+T}^G(z_1, z_2)}{P_{S+T}^G(z_1)P_{S+T}^G(z_2)} \, dz_1, dz_2.$$ 

However, due to the lack of target domain labels, $P_{S+T}^G$ is hard to model and thereby it is infeasible to estimate $I^G_{S+T}$ directly. To address this problem, we propose a surrogate joint distribution $Q(Z, Y)$ as the substitute for target domain $P_{T}^G$. Then the mixture distribution becomes $P_{S+Q}^G = \frac{1}{2}(P_{S}^G + Q)$, and the objective becomes maximizing $I^G_{S+Q}(Z_1; Z_2)$. The construction and optimization of the surrogate joint distribution is explained in Section 4.3.

Theoretical Motivation for MI Maximization

We use theoretical bound to demonstrate the motivation for using MI maximization. Our theoretical results prove that minimizing the expected risk on the target domain can be naturally transformed into MI maximization and expected risk minimization on the source domain, which explains why MI maximization is pivotal to our framework. The proofs are in appendix.

Definition 1 (H\(\Delta\)H-Divergence). Let $F_1 \in H, F_2 \in H$ be two hypotheses in hypothesis space $H : Z \rightarrow Y$. 

Theorem 2

Define \( \epsilon_P(F_1, F_2) \) as the disagreement between hypotheses \( F_1, F_2 \) w.r.t. distribution \( P \) on \( Z \), 
\( \epsilon_P(F_1, F_2) \triangleq \frac{e}{\epsilon_{\Delta P_1 \Delta P_2}} \left[ \frac{e}{\epsilon_{\Delta P_1 \Delta P_2}} \right] \). \( \Delta \)-divergence, which is the discrepancy of two distributions \( P_1, P_2 \) w.r.t. any hypothesis \( F_1, F_2 \) where \( F_1, F_2 \in \mathcal{H} \), is defined as 
\( \mathcal{H}_F(H(Y)), \mathcal{H}_F(Y) \), and additional bias of surrogate distribution. The proof is in appendix. This theorem supports the feasibility of domain adaptation via maximizing surrogate MI \( I_{S+T}^2(Z;Z) \). The bias of surrogate distribution is expressed in terms 
\( d_F(P_{S+T}^2(Z), Q(Z)) + \epsilon_{\Delta P_1 \Delta P_2}^2(Q(Y|Z)) \), where the first term is the distance between the surrogate and target feature marginal distribution, and the second term is the risk of conditional label surrogate distribution. To minimize the upper bound, the bias of the surrogate distribution should be small.

4.3 SIDA Framework

We employ MI maximization and surrogate distribution in our SIDA framework, as shown in Figure 2. During training, a surrogate distribution is first built from target and source data via optimizing w.r.t. Laplacian and MI. Then a mixture data distribution is created by encoding source data to features and sampling target features from the surrogate distribution. The encoder is optimized by maximizing MI, and minimizing classification error. The overall loss is:

\[ L_{\text{model}} = L_{\text{classify}} + \alpha_1 L_{\text{MI}} + \alpha_2 L_{\text{Auxiliary}} + \lambda L_{\text{Laplacian}}. \] (8)

We elaborate each module in the following sections, and introduce the optimization of surrogate distribution in the last sections.

Mutual Information Estimation

Several MI estimation and optimization methods are proposed in deep learning [Poole et al., 2019]. In this work, we use the following variational lower bound of MI as proposed in [Nguyen et al., 2010]:

\[ I(Z_1;Z_2) \geq \mathbb{E}_{P(z_1, z_2)}[f(z_1, z_2)] - e^{-1} \mathbb{E}_{P(z_1)}[\mathbb{E}_{P(z_2)}[e^{f(z_1, z_2)}]]. \] (9)

where \( f \) is a score function in \( Z \times Z \rightarrow R \). The equality holds when 
\( \mathbb{E}_{P(z_1)}[e^{f(z_1, z_2)}]/\mathbb{E}_{P(z_2)}[e^{f(z_1, z_2)}] = e \). The proof is in appendix. Therefore maximizing MI can be transformed into maximizing its lower bound, and the loss is:

\[ L_{\text{MI}} = - \mathbb{E}_{P_{S+Q}(z)} \mathbb{E}_{P_{S+Q}(z_1|y)} \mathbb{E}_{P_{S+Q}(z_2|y)}[f(z_1, z_2)] + e^{-1} \mathbb{E}_{P_{S+Q}(z_1)}[\mathbb{E}_{P_{S+Q}(z_2)}[e^{f(z_1, z_2)}]]. \] (10)

Surrogate Distribution Construction

We decompose the surrogate distribution \( Q(Z, Y) \) into two factors \( Q(Z), Q(Y) \), and describe the construction of two factors individually.

According to the data shift assumption, \( P_T(Y) \) is similar to \( P_S(Y) \), thus \( Q(Y) \) should be similar to \( P_S(Y) \). However, source distribution may suffer from the class imbalance problem, which will harm the performance on classes with fewer data. A common solution to this problem is class-balanced sampling, which samples data on each class uniformly. In this work, for the balance across different classes, the marginal distribution \( P_S(Y) \) and \( Q(Y) \) are both considered as uniform distribution.

As for the second term, the conditional surrogate distribution \( Q(Z|Y) \) is constructed by weighted sampling method. We need to construct the \( Q(Z|Y) \) to calculate Eq. 10, which takes the form of expectation, and only needs samples from \( Q(Z|Y) \) to estimate. Instead of explicitly modeling \( Q(Y|Z) \), we use the ideas of importance sampling. For each class, the surrogate conditional distribution \( Q(Z|y_j) \) is constructed
by weighted sampling from target features. Thus \( Q(Z|Y) \) is a distribution on target features \( \{G(x_i^j)\}_{i=1}^{n_Y} \), and parameterized by \( W \in R^{n_T \times n_Y} \), where \( n_T \) is the number of labels:

\[
Q(G(x_i^j)|y_j) = W_{ij}, \text{ s.t. } W_{ij} \in [0,1], \sum_i W_{ij} = 1, \forall j. \tag{11}
\]

Compared with pseudo-labeling, our estimation method has the following advantages: (1) The surrogate marginal distribution of feature \( Q(Z) = \sum_Y Q(Z|Y) \hat{P}(Y) \) is not fixed, which enables us to select features more flexibly. (2) The construction process of the surrogate distribution makes MI estimation \( I(Z_1, Z_2) \) more convenient. Our surrogate distribution \( Q(Z|Y) \) provides weights so that weighted sampling can be performed directly.

The challenge is to optimize the sampling probability weights \( W_{ij} \) so as to minimize the bias of the surrogate distribution. We propose to optimize this distribution via Laplacian regularization as well as MI, which is explained in details in the following section.

**Surrogate Distribution Loss**

Inspired by semi-supervised learning, we expect that the surrogate distribution is consistent with the clustering structure of the feature distribution, based on the assumption that the feature is well-structured and clustered according to class, regardless of domains. We employ Laplacian regularization to capture the manifold clustering structure of feature distribution.

Let \( A \in R^{n_T \times n_T} \) be the adjacent matrix of target features, where the entry \( A_{ij} \) measures how similar \( G(x_i^j) \) and \( G(x_j^j) \) are, and \( D = \text{Diag}(A1) \) is the degree matrix, i.e. \( D_{ii} = \sum_j A_{ij} \) and \( D_{ij} = 0, \forall i \neq j \). We construct \( A \) as K-nearest graph on target features, and the Laplacian regularization of \( W \) is defined as

\[
L_{\text{Laplacian}} = \text{Tr}(W^T LW)
= \frac{1}{2} \sum_k \sum_{i,j} A_{ij} \frac{W_{ik}}{D_{ii}} - \frac{W_{jk}}{D_{jj}})^2, \tag{12}
\]

where \( L \) is the normalized Laplacian matrix \( L = I - D^{-1/2}AD^{-1/2} \). This regularization encourages \( W_{ik} \) and \( W_{jk} \) to be similar if feature \( G(x_i^j) \) is similar to \( G(x_j^j) \). It also enables the conditional surrogate distribution to spread uniformly on a connected region.

**Classification and Auxiliary Loss**

The model is optimized in supervised manner on the source domain. The classification loss is the standard cross-entropy loss via class-balanced sampling.

\[
L_{\text{Classify}} = -\frac{1}{n_Y} \sum_y E_{P_{\text{z}}(x|y)} \log P(F(G(x)) = y). \tag{13}
\]

And we use auxiliary classification loss on pseudo-labels from the surrogate distribution, as the classifier will benefit from label information of the surrogate distribution. We use mean square error (MSE) for pseudo-labels, which is more robust to noise than cross entropy loss.

\[
L_{\text{Auxiliary}} = \frac{1}{n_Y} \sum_y E_{Q(x|y)}(1 - P(F(G(x)) = y))^2. \tag{14}
\]

**Optimization of Surrogate Distribution**

We optimize both \( L_{\text{Laplacian}} \) and \( L_{\text{MI}} \) w.r.t. \( W \) for a structured and informative surrogate distribution. At the beginning of each epoch, \( W \) is initialized by K-means clustering and filtered by the distance to the clustering centers, i.e. \( W_{i,j} = 1_{\mu_j \text{ nearest to } G(x_i^j)} \delta(G(x_i^j), \mu_j) < \theta \), where \( \mu_j \) is the \( j \)-th clustering center during clustering, and normalized as \( W_{i,j} = \frac{W_{i,j}}{\sum_i W_{i,j}} \).

To minimize two losses w.r.t \( W \), the gradients are derived analytically. The derivation is in appendix.

Based on the gradient of these two losses, we perform T-step descent update of \( W \) with learning rate \( \eta_1 \) and \( \eta_2 \) respectively, and each step we project \( W \) back to the probability simplex. See appendix for details.

### 5 Experiments

In this section, We evaluate the proposed method on three public domain adaptation benchmarks, compared with recent state-of-the-art UDA methods. We conduct extensive ablation study to discuss our method.

#### 5.1 Datasets

VisDA-2017 [Peng et al., 2017] is a challenging benchmark for UDA with the domain shift from synthetic data to real imagery. It contains 152,397 training images and 55,388 validation images across 12 classes. Following the training and testing protocol in [Long et al., 2017a], the model is trained on labeled training and unlabeled validation set and tested on the validation set.

Office-31 [Saenko et al., 2010] is a commonly used dataset for UDA, where images are collected from three distinct domains: Amazon (A), Webcam (W) and DSLR (D). The dataset consists of 4,110 images belonging to 31 classes, and is imbalanced across domains, with 2,817 images in A domain, 795 images in W domain, and 498 images in D domain. Our method is evaluated on all six transfer tasks. We follow the standard protocol for UDA [Long et al., 2017b] to use all labeled source samples and all unlabeled target samples as the training data.

Office-Home [Venkateswara et al., 2017] is another classical dataset with 15,500 images of 65 categories in office and home settings, consisting of 4 domains including Artistic images (A), Clip Art images (C), Product images (P) and Real-World images (R). Following the common protocol, all 65 categories from the four domains are used for evaluation of UDA, forming 12 transfer tasks.

#### 5.2 Implementation details

For each transfer task, mean (±std) over 5 runs of the test accuracy are reported. We use the ImageNet pre-trained ResNet-50 [He et al., 2016] without final classifier layer as the encoder network \( G \) for Office-31 and Office-Home, and ResNet-101 for VisDA-2017. The details of experiments are in appendix. The code is available at https://github.com/zhaoht/SIDA.
5.3 Baselines

We compare our approach with the state of the arts. Domain alignment methods include DAN [Long et al., 2015], DANN [Ganin et al., 2016], JAN [Long et al., 2017b]. Class-level methods include conditional alignment methods (CDAN [Long et al., 2018], DCAN [Li et al., 2020c], ALDA [Chen et al., 2020a]), and contrastive methods (DRMEA [Luo et al., 2020], ETD [Li et al., 2020a], DADA [Tang and Jia, 2020], SAFN [Xu et al., 2019]). We report available results in each task. We use NA, DA, CA, CT to note no adaptation method, domain alignment methods, conditional alignment methods and contrastive methods respectively.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>BasicBay</td>
<td>90.25 ± 0.4</td>
<td>92.37 ± 0.1</td>
<td>74.21 ± 0.2</td>
<td>74.09 ± 0.1</td>
<td>84.7 ± 0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DA</td>
<td>DANN</td>
<td>92.08 ± 0.3</td>
<td>94.28 ± 0.3</td>
<td>74.23 ± 0.9</td>
<td>74.74 ± 0.8</td>
<td>83.8 ± 0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td>CDAN</td>
<td>94.03 ± 0.1</td>
<td>95.26 ± 0.1</td>
<td>75.86 ± 0.4</td>
<td>75.72 ± 0.5</td>
<td>85.2 ± 0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>DRMEA</td>
<td>94.92 ± 0.0</td>
<td>95.88 ± 0.1</td>
<td>76.2 ± 0.0</td>
<td>76.2 ± 0.4</td>
<td>85.8 ± 0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>SIDA</td>
<td>95.4 ± 0.1</td>
<td>96.9 ± 0.1</td>
<td>76.6 ± 0.6</td>
<td>76.0 ± 0.4</td>
<td>90.4 ± 0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Accuracy (%) on Office-31

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>BasicBay</td>
<td>90.25 ± 0.4</td>
<td>92.37 ± 0.1</td>
<td>74.21 ± 0.2</td>
<td>74.09 ± 0.1</td>
<td>84.7 ± 0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DA</td>
<td>DANN</td>
<td>92.08 ± 0.3</td>
<td>94.28 ± 0.3</td>
<td>74.23 ± 0.9</td>
<td>74.74 ± 0.8</td>
<td>83.8 ± 0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td>CDAN</td>
<td>94.03 ± 0.1</td>
<td>95.26 ± 0.1</td>
<td>75.86 ± 0.4</td>
<td>75.72 ± 0.5</td>
<td>85.2 ± 0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>DRMEA</td>
<td>94.92 ± 0.0</td>
<td>95.88 ± 0.1</td>
<td>76.2 ± 0.0</td>
<td>76.2 ± 0.4</td>
<td>85.8 ± 0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>SIDA</td>
<td>95.4 ± 0.1</td>
<td>96.9 ± 0.1</td>
<td>76.6 ± 0.6</td>
<td>76.0 ± 0.4</td>
<td>90.4 ± 0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Accuracy (%) on Office-Home

5.4 Results and Comparative Analysis

In this section we will present our results and compare with other methods for evaluation on three standard benchmarks mentioned earlier. We report average classification accuracies with standard deviations. Results of other methods are collected from original papers or the follow-up work. We provide visualizations of the features learned by the model in the appendix.

VisDA-2017 Table 1 summarizes our experimental results on the challenging VisDA-2017 dataset. For fair comparison, all methods listed here use ResNet-101 as the backbone network. Note that SIDA outperforms baseline models with an average accuracy of 84.0, surpassing the previous best result reported by +4%.

Office-31 The unsupervised adaptation results on six Office-31 transfer tasks based on ResNet-50 are reported in Table 2. As the data reveals, the average accuracy of SIDA is 90.4, the best among all compared methods. It is noteworthy that our proposed method substantially improves the classification accuracy on hard transfer tasks, e.g. W→A, A→D, and D→A, where source and target data are not similar. Our model also achieves comparable classification performance on easy transfer tasks, e.g. D→W, W→D, and A→W. Our improvements are mainly in hard settings.

Office-Home Results on Office-Home using ResNet-50 backbone are reported in Table 3. It can be observed that SIDA exceeds all compared methods on most transfer tasks with an average accuracy of 71.2. The performance reveals the importance of maximizing MI between feature in difficult domain-adaptation tasks which contain more categories.

In summary, our surrogate MI maximization approach achieves competitive performance compared to traditional alignment based methods and recent pseudo-label based methods for UDA. It underlines the validity of using information theory methods for UDA via MI maximization.

5.5 Ablation Study

In this section, we evaluate how different components of our work contribute to the final performance, we conduct ablation study for SIDA on Office-31. We mainly focus on harder transfer tasks, e.g. A→W, A→D, D→A and W→A. We investigate different combinations of two components: MI maximization and surrogate distribution (SD). Note that without surrogate distribution, we use pseudo label computed by the same method as surrogate distribution initialization to estimate MI. The average classification accuracy on four tasks are shown in Table 4.

From the results, we can observe that the model with MI maximization outperforms the base model without the two components by about 2.5% on average, which demonstrates the effectiveness of the maximization strategy. The surrogate distribution also improves the average performance by 1.1% compared to base model, confirming that the surrogate distribution improves the estimation quality of target domain compared to pseudo label. The combination of two components yields the highest improvement.

6 Conclusion and Future Work

In this work, we introduce a novel framework of unsupervised domain adaptation and provide theoretical analysis to validate our optimization objectives. Experiments show that our approach gives competitive results compared to state-of-the-art unsupervised adaptation methods on standard domain adaptation tasks. One unresolved problem is to integrate the domain discrepancy in target risk upper bound into mutual information formulation. This problem is left for future work.
References


