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.

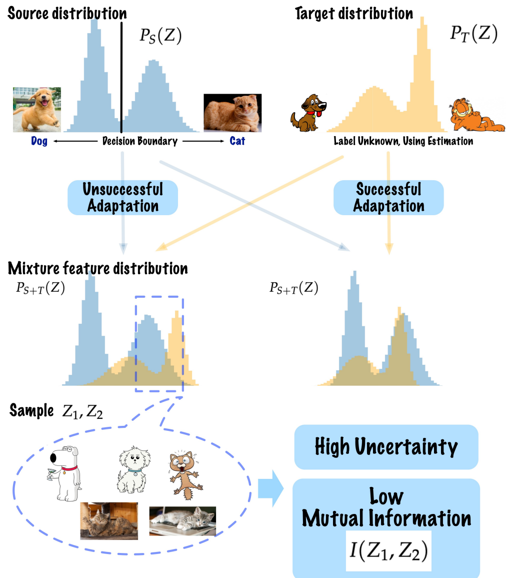

Figure 1: Our motivation is that higher intra-class feature mutual information implies better generalization ability across domain. As shown, unsuccessful domain adaptation implies lower intra-class feature mutual information. This is reflected in the fact that features from the same class do not represent each other well because of confusion with features from other classes.

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 under the premise that 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 $\{(x_S^i, y_S^i)\}_{i=1}^{n_S} \sim P_S(X,Y)$ and n_T unlabeled samples $\{(x_T^i)\}_{i=1}^{n_T} \sim P_T(X)$, and outputs a hypothesis composed of an encoder G and a classifier F. Let $\mathcal{Z}$ be the feature space. The encoder maps data to feature space, denoted by $G : \mathcal{X} \to \mathcal{Z}$. Then the classifier maps the feature to a corresponding class, $F : \mathcal{Z} \to \mathcal{Y}$.

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. $P^G(z,y) = P(x = G^{-1}(z), y)$.

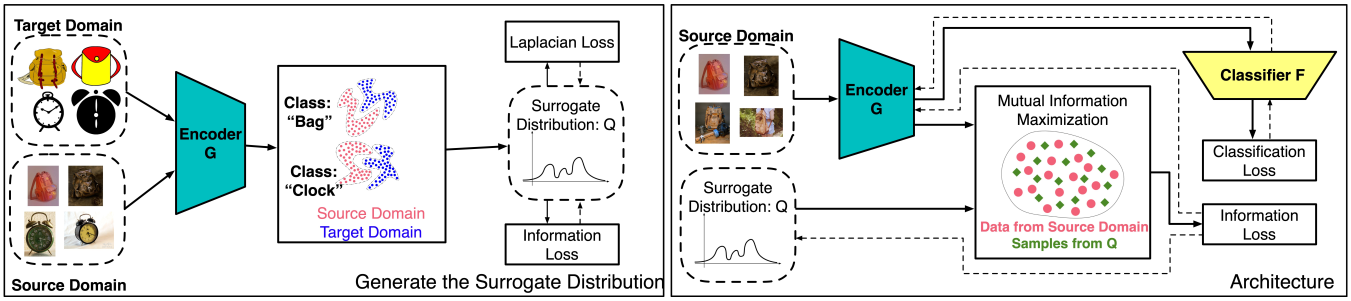

Figure 2: Overview of SIDA framework for training. Only encoder and classifier are involved in inference. The dashed arrow shows the path of the gradient backpropagation.

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

$$\epsilon_P(F) \triangleq \mathbb{E}_{P(z)}|\delta_{F(z)} - P(y|z)|_1, \tag{1}$$

where $\delta_{F(z)}(y)$ 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,

$$\min_{G,F} \epsilon_{P_T^G}(F). \tag{2}$$

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 $P_T^G(Z|Y)$ 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 [Hjelm *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

$$P_{S+T}(x, y) \triangleq \frac{1}{2}(P_S(x, y) + P_T(x, y)). \tag{3}$$

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

$$\begin{aligned} &P_{S+T}^G(z_1, z_2|y) \triangleq P_{S+T}^G(z_1|y)P_{S+T}^G(z_2|y), \\ &P_{S+T}^G(z_1, z_2) = \sum_y P_{S+T}^G(y)P_{S+T}^G(z_1, z_2|y). \end{aligned} \tag{4}$$

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:

$$\begin{aligned} &\arg\max_G I_{S+T}^G(Z_1; Z_2) \\ &= \int P_{S+T}^G(z_1, z_2) \log \frac{P_{S+T}^G(z_1, z_2)}{P_{S+T}^G(z_1)P_{S+T}^G(z_2)} dz_1 dz_2. \end{aligned} \tag{5}$$

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_{S+T}^G 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 $I_{S+Q}^G(Z_1; Z_2)$. 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 ($\mathcal{H}\Delta\mathcal{H}$-Divergence)**.** *Let* $F_1 \in \mathcal{H}, F_2 \in \mathcal{H}$ *be two hypotheses in hypothesis space* $\mathcal{H} : \mathcal{Z} \to \mathcal{Y}$.

Define $\epsilon_P(F_1, F_2)$ *as the disagreement between hypotheses* F_1, F_2 *w.r.t. distribution* P *on* $\mathcal{Z}$, $\epsilon_P(F_1, F_2) \triangleq \mathbb{E}_{z\sim P}\left|\delta_{F_1(z)} - \delta_{F_2(z)}\right|$. $\mathcal{H}\Delta\mathcal{H}$*-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* $d_{\mathcal{H}\Delta\mathcal{H}}(P_1, P_2) \triangleq 2\sup_{F_1,F_2\in\mathcal{H}} |\epsilon_{P_1}(F_1, F_2) - \epsilon_{P_2}(F_1, F_2)|$.

Theorem 1 (Bound of Target Domain expected risk)**.** *The expected risk on target domain can be upper-bounded by the negative MI between features, and* $\mathcal{H}\Delta\mathcal{H}$ *-divergence between features of two domains:*

$$\begin{aligned}&\epsilon_{P_T^G}(F) \le \epsilon_{P_S^G}(F) - 4I_{S+T}^G(Z_1; Z_2)\\&+\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(P_S^G(Z), P_T^G(Z)) + 4H(Y).\end{aligned} \tag{6}$$

The proof is in appendix. We give an explanation of the conditions for the upper bound to be equal. $I_{S+T}^G(Z_1; Z_2)$ is a lower bound of $I_{S+T}^G(Z; Y)$, and it measures how much uncertainty of Y is reduced by knowing the feature, and it's equal to $H(Y)$ if and only if $P_{S+T}^G(Y|Z)$ is deterministic, i.e., $P_{S+T}^G(Y|Z)$ is δ distribution, which means $P_S^G(Y|Z) = P_T^G(Y|Z) = \delta_{Y(Z)}$. Thus if the $\mathcal{H}\Delta\mathcal{H}$ -divergence is zero, i.e., $P_S^G(Z) = P_T^G(Z)$, then it's ensured that $P_S^G(Z, Y) = P_T^G(Z, Y)$, and $\epsilon_{P_T^G}(F) = \epsilon_{P_S^G}(F)$.

This upper bound decomposes the cross-domain generalization error into the divergence of feature marginal distribution and MI of features. It emphasizes that in addition to the divergence of the feature marginal distributions, only a MI term is enough for knowledge transfer across domains.

In this work, we minimize the expected risk on the source domain and maximize MI, for minimizing the upper bound of expected risk on target domain. Due to the lack of labels on target domain, we estimate MI based on surrogate distribution Q. The expected risk upper bound based on surrogate MI is further derived as follows.

Definition 2 (L_1-distance)**.** *Define* L_1*-distance of* P_1, P_2 *as* $d_1(P_1, P_2) \triangleq 2\sup_{B\in\boldsymbol{B}} |\Pr_{P_1}[B] - \Pr_{P_2}[B]|$ *where* $\boldsymbol{B}$ *is the set of measurable subsets under* P_1 *and* P_2.

Theorem 2 (Bound Estimation with Surrogate Distribution)**.** *Let* $B \triangleq d_1(P_T^G(Z), Q(Z)) + \epsilon_{P_T^G}(Q(Y|Z))$ *be the bias of surrogate distribution* Q *w.r.t target distribution. The expected risk on target domain can be upper-bounded by the negative surrogate MI between features,* $\mathcal{H}\Delta\mathcal{H}$ *-divergence between source and target domain, and additional bias of surrogate domain:*

$$\begin{aligned}&\epsilon_{P_T^G}(F) \le \epsilon_{P_S^G}(F) - 4I_{S+Q}^G(Z_1; Z_2) + B\\&+\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(P_S^G(Z), P_T^G(Z)) + 4H(Y).\end{aligned} \tag{7}$$

The proof is in appendix. This theorem supports the feasibility of domain adaptation via maximizing surrogate MI $I_{S+Q}^G(Z_1; Z_2)$. The bias of surrogate distribution is expressed in terms $d_1(P_T^G(Z), Q(Z)) + \epsilon_{P_T^G}(Q(Y|Z))$, 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.

Bias equal to zero if and only if surrogate feature distribution and conditional label distribution are the same as target distribution, i.e., $P_T^G = Q$, where surrogate distribution does not introduce errors.

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:

$$L_{model} = L_{Classify} + \alpha_1 L_{MI} + \alpha_2 L_{Auxiliary} + L_{Laplacian}. \tag{8}$$

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]:

$$\begin{aligned}I(Z_1; Z_2) \ge& \mathbb{E}_{P(z_1,z_2)}[f(z_1, z_2)]\\&- e^{-1}\mathbb{E}_{P(z_1)}[\mathbb{E}_{P(z_2)}\left[e^{f(z_1,z_2)}\right]],\end{aligned} \tag{9}$$

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

$$\begin{aligned}L_{MI} =& -\mathbb{E}_{P_{S+Q}^G(y)}\mathbb{E}_{P_{S+Q}^G(z_1|y)}\mathbb{E}_{P_{S+Q}^G(z_2|y)}[f(z_1, z_2)]\\&+ e^{-1}\mathbb{E}_{P_{S+Q}^G(z_1)}[\mathbb{E}_{P_{S+Q}^G(z_2)}\left[e^{f(z_1,z_2)}\right]],\end{aligned} \tag{10}$$

where $f(z_1, z_2)$ is constructed as $T_{m_1}^{m_2}(|z_1 - z_2|_2)$. $T_{m_1}^{m_2}$ is a threshold function, i.e., $T_{m_1}^{m_2}(a) = \max(m_1, \min(m_2, a))$.

Surrogate Distribution Construction

We decompose the surrogate distribution $Q(Z, Y)$ into two factors $Q(Z, Y) = Q(Y)Q(Z|Y)$, 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$\{G(x_T^i)\}_{i=1}^{n_T}$, and parameterized by $W \in R^{n_T \times n_Y}$, where n_Y is the number of labels:

$$Q(G(x_T^i)|y_j) = W_{ij}, \textbf{ s.t. } W_{ij} \in [0,1], \sum_i W_{ij} = 1, \forall j. \quad (11)$$

Compared with pseudo-labeling, our estimation method has the following advantages: (1) The surrogate marginal distribution of feature $Q(Z) = \sum_Y Q(Z|Y)P(Y)$ 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 $G(x_T^i)$ and $G(x_T^j)$ are, and $D = \mathrm{Diag}(A\mathbf{1})$ 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

$$\begin{aligned} L_{Laplacian} &= Tr(W^T L W) \\ &= \frac{1}{2}\sum_k \sum_{i,j} A_{ij}\left(\frac{W_{ik}}{D_{ii}} - \frac{W_{jk}}{D_{jj}}\right)^2, \end{aligned} \quad (12)$$

where L is the normalized Laplacian matrix $L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$. This regularization encourages W_{ik} and W_{jk} to be similar if feature $G(x_T^i)$ is similar to $G(x_T^j)$. 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.

$$L_{Classify} = -\frac{1}{n_Y}\sum_y E_{P_S(x|y)} \log P(F(G(x)) = y). \quad (13)$$

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.

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

Optimization of Surrogate Distribution
We optimize both $L_{\mathbf{Laplacian}}$ and $L_{\mathbf{MI}}$ 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. $\widetilde{W}_{i,j} = \mathbf{1}_{\mu_j \text{ nearest to } G(x_i)} \Delta \mathbf{1}_{d(G(x_i),\mu_j)<\theta}$, where μ_j is the j-th clustering center during clustering, and normalized as $W_{i,j} = \frac{\widetilde{W}_{i,j}}{\sum_i \widetilde{W}_{i,j}}$.

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 η_1 and η_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 ($\pm$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 only report available results in each baseline. We use NA, DA, CA, CT to note no adaptation method, domain alignment methods, conditional alignment methods and contrastive methods respectively.

Type	Methods	Plane	Bcycl	Bus	Car	Horse	Knife	Mcyle	Person	Plant	Sktbrd	Train	Truck	Avg
NA	ResNet-101	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DA	DAN	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	85.8	20.7	61.1
	DANN	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
CA	CDAN	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.9
	ALDA	93.8	74.1	82.4	69.4	90.6	87.2	89.0	67.6	93.4	76.1	87.7	22.2	77.8
CT	DRMEA	92.1	75.0	78.9	75.5	91.2	81.9	89.0	77.2	93.3	77.4	84.8	35.1	79.3
	DADA	92.9	74.2	82.5	65.0	90.9	93.8	87.2	74.2	89.9	71.5	86.5	48.7	79.8
	SAFN	93.6	61.3	84.1	70.6	94.1	79.0	91.8	79.6	89.9	55.6	89.0	24.4	76.1
Ours	SIDA	95.4	83.1	77.1	64.6	94.5	97.2	88.7	78.4	93.8	89.9	85.2	59.4	**84.0**

Table 1: Accuracy (%) on VisDA-2017

Type	Methods	A→W	D→W	W→D	A→D	D→A	W→A	avg
NA	ResNet-50	68.4±0.2	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
DA	DAN	80.5±0.4	97.1±0.2	99.6±0.1	78.6±0.2	63.6±0.3	62.8±0.2	80.4
	DANN	82.0±0.4	96.9±0.2	99.1±0.1	79.7±0.4	68.2±0.4	67.4±0.5	82.2
	JAN	85.4±0.3	97.4±0.2	99.8±0.2	84.7±0.3	68.6±0.3	70.0±0.4	84.3
CA	CDAN	94.1±0.1	98.6±0.1	100.0±0.0	92.9±0.2	71.0±0.3	69.3 ± 0.3	87.7
	DCAN	95.0	97.5	100.0	92.6	77.2	74.9	89.5
	ALDA	95.6±0.5	97.7±0.1	100.0±0.0	94.0±0.4	72.2±0.4	72.5±0.2	88.7
CT	ETD	92.1	100.0	100.0	88.0	71.0	67.8	86.2
	DADA	92.3±0.1	99.2±0.1	100.0±0.0	93.9±0.2	74.4±0.1	74.2±0.1	89.0
	SAFN	90.3	98.7	100.0	92.1	73.4	71.2	87.6
Ours	SIDA	94.5±0.6	99.2±0.1	100.0±0.0	95.7±0.3	76.6±0.6	76.2±0.4	**90.4**

Table 2: Accuracy(%) on Office-31

Type	Methods	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg
NA	ResNet-50	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DA	DAN	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
	DANN	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
	JAN	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CA	CDAN	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
	DCAN	54.5	75.7	81.2	67.4	74.0	76.3	67.4	52.7	80.6	74.1	59.1	83.5	70.5
	ALDA	53.7	70.1	76.4	60.2	72.6	71.5	56.8	51.9	77.1	70.2	56.3	82.1	66.6
CT	DRMEA	52.3	73.0	77.3	64.3	72.0	71.8	63.6	52.7	78.5	72.0	57.7	81.6	68.1
	ETD	51.3	71.9	85.7	57.6	69.2	73.7	57.8	51.2	79.3	70.2	57.5	82.1	67.3
	SAFN	54.4	73.3	77.9	65.2	71.5	73.2	63.6	52.6	78.2	72.3	58.0	82.1	68.5
Ours	SIDA	57.2	79.1	81.7	67.1	74.5	77.3	67.2	53.9	82.5	71.4	58.7	83.3	**71.2**

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 on 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.

MI	SD	A→W	A→D	D→A	W→A	Avg
×	×	90.25 ± 0.2	92.37 ± 0.1	74.21 ± 0.2	74.09 ± 0.1	82.7
×	√	92.08± 0.3	94.28±0.3	74.23±0.9	74.74± 0.8	83.8
√	×	94.03± 0.1	95.28± 0.1	75.86± 0.4	75.72 ± 0.5	85.2
√	√	94.52 ± 0.6	95.68 ± 0.1	76.62 ± 0.6	76.22 ± 0.4	85.8

Table 4: Ablation Study

5.5 Ablation Study

In this section, to 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 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 method. 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 framework. This problem is left for future work.

References

[Ben-David *et al.*, 2007] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 2007.

[Chen and Liu, 2020] Qingchao Chen and Yang Liu. Structure-aware feature fusion for unsupervised domain adaptation. In *Proc. of AAAI*, 2020.

[Chen *et al.*, 2019] Chaoqi Chen, Weiping Xie, Wenbing Huang, Yu Rong, Xinghao Ding, Yue Huang, Tingyang Xu, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. In *Proc. of CVPR*, 2019.

[Chen *et al.*, 2020a] Minghao Chen, Shuai Zhao, Haifeng Liu, and Deng Cai. Adversarial-learned loss for domain adaptation. In *Proc. of AAAI*, 2020.

[Chen *et al.*, 2020b] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 2020.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[Hjelm *et al.*, 2018] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *Proc. of ICLR*, 2018.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Proc. of NeurIPS*, 2020.

[Li *et al.*, 2020a] Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *Proc. of CVPR*, 2020.

[Li *et al.*, 2020b] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proc. of CVPR*, 2020.

[Li *et al.*, 2020c] Shuang Li, Chi Liu, Qiuxia Lin, Binhui Xie, Zhengming Ding, Gao Huang, and Jian Tang. Domain conditioned adaptation network. In *Proc. of AAAI*, 2020.

[Long *et al.*, 2015] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, 2015.

[Long *et al.*, 2017a] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *arXiv preprint arXiv:1705.10667*, 2017.

[Long *et al.*, 2017b] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, 2017.

[Long *et al.*, 2018] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018.

[Luo *et al.*, 2020] You-Wei Luo, Chuan-Xian Ren, Pengfei Ge, Ke-Kun Huang, and Yu-Feng Yu. Unsupervised domain adaptation via discriminative manifold embedding and alignment. In *Proc. of AAAI*, 2020.

[Nguyen *et al.*, 2010] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 2010.

[Park *et al.*, 2020] Changhwa Park, Jonghyun Lee, Jaeyoon Yoo, Minhoe Hur, and Sungroh Yoon. Joint contrastive learning for unsupervised domain adaptation. *arXiv preprint arXiv:2006.10297*, 2020.

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

[Poole *et al.*, 2019] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *Proc. of ICML*, 2019.

[Redko *et al.*, 2020] Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv e-prints*, 2020.

[Saenko *et al.*, 2010] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, 2010.

[Tang and Jia, 2020] Hui Tang and Kui Jia. Discriminative adversarial domain adaptation. In *Proc. of AAAI*, 2020.

[Thota and Leontidis, 2021] Mamatha Thota and Georgios Leontidis. Contrastive domain adaptation. In *Proc. of CVPR*, 2021.

[Venkateswara *et al.*, 2017] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proc. of CVPR*, 2017.

[Xie *et al.*, 2018] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International conference on machine learning*, 2018.

[Xu *et al.*, 2019] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proc. of ICCV*, 2019.