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 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 $X$ be the data space and $Y$ be the label space. In UDA, there is a source distribution $P_S(X, Y)$ and a target distribution $P_T(X, Y)$ on $X \times Y$. 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 $Z$ be the feature space. The encoder maps data to feature space, denoted by $G : X \rightarrow Z$. Then the classifier maps the feature to a corresponding class, $F : Z \rightarrow Y$.

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

$$
\min_{G,F} \epsilon_{P^G_T}(F)
$$

## 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 $P_G^T(Z|Y)$ 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

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

Note that because $P_S(y) = P_T(y) = P_{S+T}(y)$, $P_{S+T}(x|y) = \frac{1}{2}P_S(x | y) + \frac{1}{2}P_T(x | y)$. Define the distribution of features from the same class as

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

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 below:

$$
\arg \max_{G} I_{S+T}^G(Z_1; Z_2)
$$

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

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 $P_{Q}^G$. 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 (H∆H-Divergence).** Let $F_1 \in \mathcal{H}, F_2 \in \mathcal{H}$ be two hypotheses in hypothesis space $\mathcal{H} : \mathcal{Z} \rightarrow \mathcal{Y}$.
Define $\epsilon_P(F_1, F_2)$ as the disagreement between hypotheses $F_1, F_2$ w.r.t. distribution $P$ on $Z$, $\epsilon_P(F_1, F_2) \triangleq \mathbb{E}_{z \sim P}[|\delta_{F_1}(z) - \delta_{F_2}(z)|]$. The $\Delta\mathcal{H}$-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 $d_{\Delta\mathcal{H}}(P_1, P_2) \triangleq 2\sup_{F_1, F_2 \in \mathcal{H}}[\epsilon_P(F_1, F_2) - \epsilon_P(F_1, F_2)]$.

**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 $\Delta\mathcal{H}$-divergence between features of two domains:

$$
\epsilon_{P_S}(F) \leq \epsilon_{P_S}(F) - 4T_{S+T}^{\mathcal{G}}(Z_1; Z_2) + \frac{1}{2}d_{\Delta\mathcal{H}}(P_S^{\mathcal{G}}(Z), P_S^{\mathcal{T}}(Z)) + 4H(Y).
$$

(6)

The proof is in appendix. We give an explanation of the conditions for the upper bound to be equal. $T_{S+T}^{\mathcal{G}}(Z_1; Z_2)$ is a lower bound of $T_{S+T}^{\mathcal{G}}(Z; Y)$, 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 $P_S^{\mathcal{T}}(Y | Z) = P_S^{\mathcal{G}}(Y | Z)$. Thus if the $\Delta\mathcal{H}$-divergence is zero, i.e., $P_S^{\mathcal{G}}(Z) = P_S^{\mathcal{T}}(Z)$, then it’s ensured that $P_S^{\mathcal{G}}(Z, Y) = P_S^{\mathcal{T}}(Z, Y)$, and $\epsilon_{P_S}(F) = \epsilon_{P_S}(F)$.

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 $d_1(P_1, P_2) \triangleq 2\sup_{B \in \mathcal{B}}[\mathbb{P}_{P_1}[B] - \mathbb{P}_{P_2}[B]]$ where $\mathcal{B}$ is the set of measurable subsets under $P_1$ and $P_2$.

**Theorem 2** (Bound Estimation with Surrogate Distribution). Let $B \triangleq d_1(P_S^{\mathcal{G}}(Z), Q(Z)) + \epsilon_{P_S}(Q(Y | Z))$ 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, $\Delta\mathcal{H}$-divergence between source and target domain, and additional bias of surrogate domain:

$$
\epsilon_{P_S}(F) \leq \epsilon_{P_S}(F) - 4T_{S+T}^{\mathcal{G}}(Z_1; Z_2) + B + \frac{1}{2}d_{\Delta\mathcal{H}}(P_S^{\mathcal{G}}(Z), P_S^{\mathcal{T}}(Z)) + 4H(Y).
$$

(7)

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

$$
L_{model} = L_{classification} + \alpha_1 L_{MI} + \alpha_2 L_{Auxiliary} + L_{Laplacian}.
$$

(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]:

$$
I(Z_1; Z_2) \geq E_{P(z_1, z_2)}[f(z_1, z_2)] - e^{-1}E_{P(z_1)}[E_{P(z_2)}[e^{f(z_1, z_2)}]],
$$

(9)

where $f$ is a score function in $Z \times Z \rightarrow R$. The equality holds when $\frac{e^{f(z_1, z_2)}}{E_{P(z_1)}[e^{f(z_1, z_2)}]} = \frac{P(z_1 | z_2)}{P(z_2)}$.\textsuperscript{1} The proof is in appendix. Therefore maximizing MI can be transformed into maximizing its lower bound, and the loss is:

$$
L_{MI} = -E_{P(G \perp \perp Q(y | z_1)}E_{P(G \perp \perp Q(z_1 | y)}E_{P(G \perp \perp Q(z_2 | y))}[f(z_1, z_2)]
$$

$$
+ e^{-1}E_{P(G \perp \perp Q)}[E_{P(G \perp \perp Q)}[e^{f(z_1, z_2)}]],
$$

(10)

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

**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, $P_T$ is similar to $P_S$, 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 $Q(Z | Y)$, 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 \( \{G(x^j_T)\}_{i=1}^{n_T} \), and parameterized by \( W \in \mathbb{R}^{n_T \times n_Y} \), where \( n_Y \) is the number of labels:

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

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 \( A \in \mathbb{R}^{n_T \times n_T} \) be the adjacent matrix of target features, where the entry \( A_{ij} \) measures how similar \( G(x^j_T) \) and \( G(x^l_T) \) 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

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

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 \( G(x^j_T) \) is similar to \( G(x^l_T) \). 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.

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

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., \( \bar{W}_{ij} = \mathbf{1}_{\mu_j \text{ nearest to } G(x_j)} I_{d(G(x_j), \mu_j) < \theta} \), where \( \mu_j \) is the \( j \)-th clustering center during clustering, and normalized as \( W_{ij} = \frac{\bar{W}_{ij}}{\sum_i \bar{W}_{ij}} \).

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/zhaohit/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.

### Table 1: Accuracy (%) on VisDA-2017

| Type      | Methods    | Plastic | Royal | Bus | Car | Horse | Knife | Make | Phone | Plant | Skis | Train | Tank | SVM   | MNIST | Avg |
|-----------|------------|--------|-------|----|----|-------|-------|------|-------|-------|------|-------|------|-------|-------|-----|
| NA        | ResNet-11  | 55.3   | 53.3  | 51.9 | 50.6| 53.7  | 52.6  | 51.9 | 50.6  | 53.7  | 52.6  | 51.9  | 50.6  | 53.7  | 52.6  | 51.9 |
| DS        | ResNet-101 | 81.6   | 80.1  | 78.8 | 77.2| 80.1  | 78.8  | 77.2 | 76.5  | 79.5  | 78.1  | 76.8  | 76.5  | 79.5  | 78.1  | 76.8 |
| ResNet-50 | 94.1      | 92.9   | 91.7  | 90.4| 92.9| 91.7  | 90.4  | 91.7 | 90.4  | 92.9  | 91.7  | 90.4  | 91.7  | 90.4  | 91.7  | 90.4 |
| ResNet-152| 95.4      | 94.2   | 93.0  | 91.8| 94.2| 93.0  | 91.8  | 93.0 | 91.8  | 94.2  | 93.0  | 91.8  | 93.0  | 94.2  | 93.0  | 91.8 |

### Table 2: Accuracy (%) on Office-31

| Type     | Methods     | A→W | A→D | D→A | W→A | Avg |
|----------|-------------|-----|-----|-----|-----|-----|
| NA       | ResNet-11   | 90.3| 85.8| 79.3| 79.2| 83.7|
| DS       | ResNet-101  | 94.5| 90.4| 84.7| 85.6| 89.0|
| ResNet-50| 95.6       | 91.7| 85.9| 85.6| 89.2| 91.7|

### Table 3: Accuracy (%) on Office-Home

| Type     | Methods     | A→W | A→D | D→A | W→A | Avg |
|----------|-------------|-----|-----|-----|-----|-----|
| NA       | ResNet-101  | 90.5| 89.5| 87.9| 87.6| 90.0|
| DS       | ResNet-101  | 94.2| 91.7| 85.2| 86.4| 90.4|
| ResNet-50| 95.3       | 91.7| 86.3| 87.5| 89.9| 91.7|

## 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 will 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 target, 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.

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.

### Table 4: Ablation Study

| Type     | Methods          | A→W | A→D | D→A | W→A | Avg |
|----------|------------------|-----|-----|-----|-----|-----|
| NA       | ResNet-101       | 90.5| 89.5| 87.9| 87.6| 90.0|
| DS       | ResNet-101       | 94.2| 91.7| 85.2| 86.4| 90.4|
| ResNet-50| 95.3             | 91.7| 86.3| 87.5| 89.9| 91.7|

## 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 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.

[Deng et al., 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, 2009.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In International conference on machine learning, 2015.

[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.

[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, Haosan 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, Qixia 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, Jonghun Lee, Jaeyoung Yoo, Minho 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 Pan- chanathan. 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.
7 Appendix

7.1 Proof for Theorem 4.1

Theorem 3 (Bound of Target Domain expected risk). The expected risk on target domain can be upper-bounded by the negative MI between features, and $\Delta H$-divergence between features of two domains:

$$
\epsilon_{p_S^G}(F) \leq \epsilon_{p_S^G}(F) - 4I_{S+T}^G(Z; Z_2) + \frac{1}{2}d_{\Delta H}(P_S^G(Z), P_T^G(Z)) + 4H(Y)
$$

Proof. The risk can be relaxed by triangle inequality

$$
\epsilon_{p_S^G}(F) \leq \epsilon_{p_S^G}(F') + \epsilon_{p_S^G}(F, F')
$$

\begin{align*}
&= \epsilon_{p_S^G}(F') + \epsilon_{p_S^G}(F, F') - \epsilon_{p_S^G}(F, F') \\
&\leq \epsilon_{p_S^G}(F') + \epsilon_{p_S^G}(F, F') + \epsilon_{p_S^G}(F) - \epsilon_{p_S^G}(F, F') \\
&\leq \epsilon_{p_S^G}(F) + \epsilon_{p_S^G}(F') + \epsilon_{p_S^G}(F') + \frac{1}{2}d_{\Delta H}(P_S^G(Z), P_T^G(Z))
\end{align*}

For the term $\epsilon_{p_S^G}(F') + \epsilon_{p_S^G}(F')$, we have

$$
\epsilon_{p_S^G}(F') + \epsilon_{p_S^G}(F') = \mathbb{E}_{p_S^G(z)}|P_S^G(y|z) - \delta_{F'(z)}|_1 + \mathbb{E}_{p_S^G(z)}|P_T^G(y|z) - \delta_{F'(z)}|_1 \\
= 2\mathbb{E}_{p_S^G(z)}(1 - P_S^G(F'(z)|z)) + 2\mathbb{E}_{p_S^G(z)}(1 - P_T^G(F'(z)|z)) \\
= 2(1 - \sum_z P_S^G(z)(P_S^G(F'(z)|z)) - \sum_z P_T^G(z)(P_T^G(F'(z)|z))) \\
= 2(2 - \sum_z P_S^G(F'(z), z) - \sum_z P_T^G(F'(z), z)) \\
= 2(2 - 2 \sum_z P_{S+T}^G(F'(z), z)) \\
= 4(1 - \sum_z P_{S+T}^G(z)(1 - P_{S+T}^G(F'(z)|z))) \\
= 4\mathbb{E}_{p_{S+T}^G(z)}(1 - P_{S+T}^G(F'(z)|z)) \\
\leq -4\mathbb{E}_{p_{S+T}^G(z)} \log P_{S+T}^G(F'(z)|z)
\end{align*}

While $F'$ can be any classifier, define $F'$ as the optimal classifier on $P_{S+T}^G$, i.e. $F'(z) = \arg\max_y P_{S+T}^G(y|z)$. Recall the definition of MI,

$$
I_{S+T}^G(Z; Y) = H_{S+T}(Y) - H_{S+T}(Y|Z) \\
= H_{S+T}(Y) + \mathbb{E}_{p_{S+T}^G(z)} \mathbb{E}_{p_{S+T}^G(y|z)} \log P_{S+T}^G(y|z) \\
\leq H_{S+T}(Y) + \mathbb{E}_{p_{S+T}^G(z)} \mathbb{E}_{p_{S+T}^G(y|z)} \log P_{S+T}^G(F'(z)|z) \\
= H_{S+T}(Y) + \mathbb{E}_{p_{S+T}^G(z)} \log P_{S+T}^G(F'(z)|z)
$$

Which means that

$$
\epsilon_{p_S^G}(F') + \epsilon_{p_S^G}(F') \leq -4\mathbb{E}_{p_{S+T}^G(z)} \log P_{S+T}^G(F'(z)|z) \\
\leq -4I_{S+T}^G(Z; Y) + 4H_{S+T}(Y)
$$

According to the MI chain rule, $I(Z_1; Y, Z_2) = I(Z_1; Y) + I(Z_1; Z_2|Y) = I(Z_1; Z_2) + I(Z_1; Y|Z_2)$. Since $Z_1$ and $Z_2$ are two samples from class $Y$, $Z_1$ and $Z_2$ are independent for a given $Y$, i.e., $I(Z_1; Z_2|Y) = 0$. So we can get $I(Z_1; Y) = I(Z_1; Z_2) + I(Z_1; Y|Z_2)$. Because $I(Z_1; Y|Z_2) \geq 0$, we finally get

$$
I(Z_1; Y) \geq I(Z_1; Z_2)
$$

Note that $I(Z_1; Y)$ is $I(Z; Y)$, because $Z_1$ and $Z_2$ both follow distribution $P(Z|Y)$. 
So now we get the conclusion by

\[
\epsilon_{p_g}(F') + \epsilon_{q_g}(F') \leq -4I_{S+T}^{G}(Z_1; Z_2) + 4H_{S+T}(Y)
\]  

(21)

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 if \(P_{S+T}^{G}(Y|Z)\) is deterministic, i.e., \(P_{S+T}^{G}(Y|Z) = \delta_{Y(Z)}\). Thus if the \(H\Delta H\)-divergence is zero, i.e., \(P_{S}^{G}(Y) = P_{T}^{G}(Y)\), then it’s ensured that \(P_{S}^{G}(Z, Y) = P_{T}^{G}(Z, Y)\), and \(\epsilon_{p_g}(F) = \epsilon_{q_g}(F)\).

\[
\square
\]

7.2 Proof for Theorem 4.2

**Theorem 4** (Bound Estimation with Surrogate Distribution). Let \(B = d_1(P_{S}^{G}(Z), Q(Z)) + \epsilon_{p_g}(Q(Y|Z))\) 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, \(H\Delta H\)-divergence between source and target domain, and additional bias of surrogate domain:

\[
\epsilon_{p_g}(F) \leq \epsilon_{p_g}(F) - 4I_{S+Q}^{G}(Z_1; Z_2) + B + \frac{1}{2}d_{H\Delta H}(P_{S}^{G}(Z), P_{T}^{G}(Z)) + 4H(Y)
\]

(22)

**Proof.** The expected expected risk can be relaxed by triangle inequality:

\[
\epsilon_{p_g}(F) \leq \epsilon_{p_g}(Q(Y|Z)) + \epsilon_{p_g}(F, Q(Y|Z))
\]

\[
\leq \epsilon_{p_g}(Q(Y|Z)) + \epsilon_{p_g}(F', Q(Y|Z)) + \epsilon_{p_g}(F', F) - \epsilon_{p_g}(F, F')
\]

\[
\leq \epsilon_{p_g}(Q(Y|Z)) + \epsilon_{p_g}(F', Q(Y|Z)) + \epsilon_{p_g}(F', F') + \epsilon_{p_g}(F') - \epsilon_{p_g}(F, F')
\]

\[
\leq \epsilon_{p_g}(Q(Y|Z)) + \epsilon_{p_g}(F') + \epsilon_{p_g}(F, Q(Z|Z)) + \epsilon_{p_g}(F', F) + \epsilon_{p_g}(F') + \epsilon_{p_g}(F') - \epsilon_{p_g}(F, F')
\]

\[
\leq \epsilon_{p_g}(Q(Y|Z)) + \epsilon_{p_g}(F') + \epsilon_{p_g}(F, F') + \frac{1}{2}d_{H\Delta H}(P_{S}^{G}(Z), P_{T}^{G}(Z))
\]

\[
\leq \epsilon_{p_g}(Q(Y|Z)) + \epsilon_{p_g}(F') + \frac{1}{2}d_{H\Delta H}(P_{S}^{G}(Z), P_{T}^{G}(Z))
\]

\[
\leq \epsilon_{p_g}(Q(Y|Z)) + \epsilon_{p_g}(F') + d_1(P_{S}^{G}(Z), Q(z)) + \epsilon_{p_g}(F') + \epsilon_{p_g}(F') + \frac{1}{2}d_{H\Delta H}(P_{S}^{G}(Z), P_{T}^{G}(Z))
\]

\[
= \epsilon_{p_g}(F') + B + \epsilon_{p_g}(F') + \frac{1}{2}d_{H\Delta H}(P_{S}^{G}(Z), P_{T}^{G}(Z))
\]

By the same method as the previous proof, terms \(\epsilon_{Q}(F') + \epsilon_{p_g}(F')\) can be deduced into MI \(-4I_{S+Q}^{G}(Z_1; Z_2) + 4H(Y)\).

\[
\square
\]

7.3 Proof for the Equality Condition of MI Estimation

**Proposition 5.** The following MI lower bound holds

\[
I(Z_1; Z_2) \geq E_{p(z_1, z_2)}[f(z_1, z_2)] - e^{-1}E_{p(z_1)}[E_{p(z_2)}[e^{f(z_1, z_2)}]]
\]

(24)

where \(f\) is arbitrary function in \(Z \times Z \to R\). The equality holds when \(E_{p(z_1, z_2)}[e^{f(z_1, z_2)}] = e\).

**Proof.** The proof is as follows:

\[
I(Z_1; Z_2) = E_{p(z_1, z_2)}[\log \frac{q(z_1 | z_2)}{P(z_1)}] + E_{p(z_2)}[KL(P(z_1 | z_2)||q(z_1 | z_2))] \geq E_{p(z_1, z_2)}[\log \frac{q(z_1 | z_2)}{P(z_1)}]
\]

(25)
where \( q \) is arbitrary variational distribution. Let \( q(z_1 \mid z_2) = \frac{p(z_1)}{Z(z_2)} e^{f(z_1,z_2)} \), where \( Z(z_2) = \mathbb{E}_{P(z_1)}[e^{f(z_1,z_2)}] \) is the normalization constant.

Then

\[
I(Z_1; Z_2) \geq \mathbb{E}_{P(z_1,z_2)}[f(z_1,z_2)] - \log \mathbb{E}_{P(z_2)}[Z(z_2)]
\]

By \( \log(x) \leq \frac{x}{a(x)} + \log(a(x)) - 1 \), which is tight when \( a(x) = x \),

\[
I(Z_1; Z_2) \geq \mathbb{E}_{P(z_1,z_2)}[f(z_1,z_2)] - \mathbb{E}_{P(z_2)} \left[ \frac{\mathbb{E}_{P(z_1)}[e^{f(z_1,z_2)}]}{a(z_2)} + \log(a(z_2)) - 1 \right]
\]

Let \( a(z_2) = e \), we get the final form of lower bound:

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

\[
\square
\]

### 7.4 Details for Surrogate Distribution Optimization

Let \( P \) be the conditional distribution matrix of source domain, i.e. \( P_{ij} = P_S(x_i^j | y_S^j) \), and \( W \) be the conditional distribution matrix of surrogate distribution \( Q \). Let \( M = \frac{1}{2} \begin{bmatrix} P & W \end{bmatrix} \) be the conditional distribution matrix of mixture distribution \( P_{S+T} \).

Let \( S \) is the score function matrix, i.e. \( S_{i,j} = f(G(x^i), G(x^j)), i, j = 1, \ldots, n_S + n_T \), where \( f \) is the score function of the MI.

For each transfer task, mean \((\mathbb{E}_x, \mathbb{E}_y)\) of surrogate distribution \( P \) is represented as follows:

\[
L_{MI} = -(\frac{1}{n_Y} \text{Tr}(M^T S M) - \frac{1}{en_Y} \mathbf{1}^T M^T e^S M \mathbf{1})
\]

The gradient w.r.t \( M \) is

\[
\nabla_M L_{MI} = -2(\frac{1}{n_Y} S M - \frac{1}{en_Y} e^S M \mathbf{1}^T)
\]

And thus the gradient w.r.t \( W \) is

\[
\nabla_W L_{MI} = -\left( \frac{1}{n_Y} (\mathbf{0}, I) S M - \frac{1}{en_Y} (\mathbf{0}, I) e^S M \mathbf{1}^T \right)
\]

Where \((\mathbf{0}, I) \in R_{n_x+n_S+n_T} \), \( I \) is identity matrix with size \( n_T \).

In practice, we find it harmful to minimize \( L_{MI} \) by \( \nabla_W L_{MI} \) directly, because it will encourage the distribution to concentrate on only a few samples rapidly. We thus adjust the descent direction of \( L_{MI} \) to update the distribution slowly. Let \( \nabla_W L_{Laplacian} \) be the entry-wise absolute value of \( \nabla_W L_{Laplacian} \). The descent directions are: \( d_1 = -\nabla_W L_{Laplacian} \) and \( d_2 = -|\nabla_W L_{Laplacian}| \odot (\nabla_W L_{MI}) \), where \( \odot \) is entry-wise multiplication. Due the property of \( \nabla_W L_{Laplacian} \), when \( d_2 \) is on the margin of each conditional distribution, this yield a diffusion-like update of \( W \), which prevent rapid collapse of surrogate distribution.

Therefore, the update rule of \( W \) is \( W^{k+1} = \Pi(W^k + \eta_1 d_1^k + \eta_2 d_2^k) \), where \( \Pi \) is the projection operator onto probability simplex for each column of \( W \). \( \eta_1 \) and \( \eta_2 \) are learning rate. The iteration is performed \( T \) times.

### 7.5 Implementation details

For each transfer task, mean (± std) over 5 runs of the test accuracy are reported. We use the ImageNet [Deng et al., 2009] 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. Following [kan, ], the final classifier layer of ResNet is replaced with the task-specific fully-connected layer to parameterize the classifier \( F \), and domain-specific batch normalization parameters are used. Code is attached in supplementary materials.

The model is trained in the finetune protocol, where the learning rate of the classifier layer is 10 times that of the encoder, by mini-batch stochastic gradient descent (SGD) algorithm with momentum of 0.9 and weight decay of 5e-4. The learning rate schedule follows [Long et al., 2017b; Long et al., 2015; Ganin and Lempitsky, 2015], where the learning rate \( \eta_p \) is adjusted following \( \eta_p = \frac{\eta_0}{(1+p)^\gamma} \), where \( p \) is the normalized training progress from 0 to 1. \( \eta_0 \) is the initial learning rate, i.e. 0.001 for the encoder layers and 0.01 for the classifier layer. For Office-31 and Office-Home, \( a = 10 \) and \( b = 0.75 \), while for VisDA-2017, \( a = 10 \) and \( b = 2.25 \). The coefficients of \( L_{MI} \) and \( L_{Auxiliary} \) are \( \alpha_1 = 0.3, \alpha_2 = 0.1 \) for Office-31, \( \alpha_1 = 1.3, \alpha_2 = 1.0 \) for Office-Home, and \( \alpha_1 = 3.0, \alpha_2 = 1.0 \) for VisDA-2017. The hyperparameters of surrogate distribution optimization include K-nearest Graph \( K = 3 \), number of iterations \( T = 3 \), learning rate \( \eta_1 = 0.5 \) and \( \eta_2 = 0.05 \).
Experiments are conducted with Python3 and Pytorch. The model is trained on single NVIDIA GeForce RTX 2080 Ti graphic card. For Office-31, each epoch takes 80 seconds and 10 seconds to perform inference. Code is attached in supplementary materials.

7.6 Visualization

Our visualization experiment is carried out on $D \rightarrow A$ and $W \rightarrow A$ tasks in the data set Office31, which are the two most difficult tasks in Office31. The baselines we chose were ResNet-50 pre-trained on ImageNet and CDAN[Long et al., 2018]. We chose CDAN because it is a typical conditional domain alignment method. Pre-trained ResNet-50 is fine-tuned on the source domain and then tested on the target domain. Results of CDAN are obtained by running the official code. We train all the models until convergence, then encode the data of source domain and target domain with the model, and take representation before the final linear classification layer as feature vectors. We use t-SNE to visualize the features, using the t-SNE function of scikit-learn with default parameters. The results are in the link.

Figure 3 shows the results of task $D \rightarrow A$. From top to bottom are the feature visualizations on task $D \rightarrow A$ of ResNet-50, CDAN, and SIDA, respectively. The left column is the feature comparison of the source and target domains. Red represents the source domain, and blue represents the target domain. The results show that SIDA emphasises discriminability of features. The right column shows the feature of different classes on the target domain. SIDA makes target features better distinguishable.

Figure 4 shows the results of task $W \rightarrow A$. The results on task $W \rightarrow A$ are similar to task $D \rightarrow A$.

The visualization results show that SIDA can make the features of different categories more distinguishable, a natural consequence of maximizing MI among features from the same category. Thus features can be easier for classification, as the visualization shows.
Figure 3: From top to bottom are the feature visualizations on task D→A of ResNet50, CDAN, and SIDA, respectively.
Figure 4: The result of task $W \rightarrow A$ is similar to task $D \rightarrow A$. 
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.

[Deng et al., 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 *IEEE conference on computer vision and pattern recognition*, 2009.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 2015.

[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.