Understanding and Improving Source-free Domain Adaptation from a Theoretical Perspective

Yu Mitsuzumi[1,2] Akisato Kimura[1] Hisashi Kashima[2]
[1]NTT Corporation [2]Kyoto University
yu.mitsuzumi@ntt.com akisato@ieee.org kashima@i.kyoto-u.ac.jp

Abstract

Source-free Domain Adaptation (SFDA) is an emerging and challenging research area that addresses the problem of unsupervised domain adaptation (UDA) without source data. Though numerous successful methods have been proposed for SFDA, a theoretical understanding of why these methods work well is still absent. In this paper, we shed light on the theoretical perspective of existing SFDA methods. Specifically, we find that SFDA loss functions comprising discriminability and diversity losses work in the same way as the training objective in the theory of self-training based on the expansion assumption, which shows the existence of the target error bound. This finding brings two novel insights that enable us to build an improved SFDA method comprising 1) Model Training with Auto-Adjusting Diversity Constraint and 2) Augmentation Training with Teacher-Student Framework, yielding a better recognition performance. Extensive experiments on three benchmark datasets demonstrate the validity of the theoretical analysis and our method.

1. Introduction

Deep learning has suffered from the domain-shift problem where models perform well on domains seen in the training phase but struggle with unseen domains. Unsupervised domain adaptation (UDA) is a promising solution: it transfers knowledge learned from a labeled source domain to an unlabeled target domain. UDA methods show their effectiveness on various computer vision tasks such as classification [25, 26, 54], object detection [6, 37, 59], segmentation [15, 16, 50], *etc*.; however, they typically require both source and target domain data, which limits their applicability as this requirement poses privacy concerns about the source data and entails computational inefficiency. Recently, researchers have shifted focus to another direction of UDA called source-free domain adaptation (SFDA). SFDA bypasses the above issues by not using raw data from the source data. Instead, SFDA performs the training with a source-pre-trained model and unlabeled target data. Various SFDA methods have been proposed, including source estimation [11, 17, 33, 42, 49], pseudo-labeling [19, 34, 51], clustering [21–23], consistency [5, 52, 53, 57], and even without source data, they outperform UDA methods.

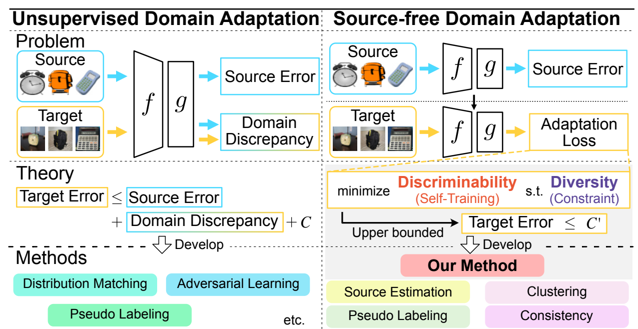

Figure 1. **Overview of our work.** Unsupervised domain adaptation has a theoretical background that has yielded a variety of methods. By contrast, the theoretical perspectives on source-free domain adaptation (SFDA) have not been well explored. Our research motivations are (1) to shed light on the theoretical perspectives of existing SFDA methods, and (2) to propose an improved method based on the theoretical insights.

Despite these promising achievements, a theoretical understanding of SFDA methods is still lacking. As shown in Fig. 1, UDA studies rely on the theoretical notion that the target error can be upper-bounded by the source error and the discrepancy between the two domains [3, 30, 55, 58], and develop various approaches such as distribution matching [26, 27, 44], adversarial learning [12, 28, 45, 54], and pseudo-labeling [20, 25, 60, 61]. By contrast, the theoretical analyses of SFDA are either absent or not general enough and cannot form the basis for the development of new methods. Moreover, the theory of UDA is not directly applicable to SFDA due inaccessibility of source data.

In this paper, we shed light on the theoretical perspective of existing SFDA methods through the theory of self-training based on the expansion assumption [48]. Self-training is an approach that utilizes the current model predictions of the unlabeled data for further training, and the expansion assumption states that the data distribution has good continuity within each class. The theory asserts that, under the expansion assumption, there is an upper bound on the target error when the model is trained on the objective with a self-training term encouraging prediction consistency among the augmented unlabeled samples and a constraint term ensuring prediction diversity. We reveal an interesting correspondence between this training objective and the SFDA training loss. Recent studies [9, 53] have discovered a feature common that most SFDA methods employ the combination of discriminability and diversity losses: the former improves the model discriminability to the unlabeled target samples while the latter ensures predictions for all classes. As illustrated in the middle right of Fig. 1, we find that the discriminability and diversity losses perform the same respective roles as the self-training term and the constraint term of the theory, which provides us the theoretical understanding of SFDA. In addition, our analysis brings the following new insights: 1) the trade-off between discriminability and diversity should be adjusted as training progresses, and 2) the upper bound of the target error depends on how we design the data augmentation.

Based on the above insights, we propose an improved SFDA method incorporating 1) Model Training with Auto-Adjusting Diversity Constraint and 2) Augmentation Training with Teacher-Student Framework. In the former training, we update the model on the basis of the discriminability and diversity losses while introducing a novel technique to automatically adjust the trade-off parameter between discriminability and diversity. In the latter training, we introduce a learnable data augmentation and update its parameters by using the predictions of the current model and the teacher model, yielding a tighter upper bound. Experimental results with three benchmarks (Office-31 [36], Office-Home [47], VisDA2017 [32]) show the validity of our theoretical analysis and the proposed method.

In summary, our contributions are: i) by using the theory of self-training based on the expansion assumption [48], we reveal that a model trained with discriminability and diversity losses will achieve a small target error; ii) we propose an improved SFDA method incorporating Model Training with Auto-Adjusting Diversity Constraint and Augmentation Training with Teacher-Student Framework.

2. Related Works

Unsupervised Domain Adaptation (UDA). On the basis of the theoretical foundation that the target error is upper bounded by the source error and the distributional discrepancy between the two domains [2, 3, 30, 55, 58], various UDA methods have been developed. Distribution matching approaches [26, 27, 44] directly minimize the measures of distribution discrepancy (*e.g.* maximum mean discrepancy (MMD)). Adversarial learning approaches [12, 28, 45, 54] reduce the discrepancy by learning domain-invariant representations using an additional domain classifier. Pseudo-labeling approaches [20, 25, 60, 61] not only minimize the domain discrepancy but also improve the feature discriminability by using the pseudo-labeled target samples.

Although the theory of UDA has yielded various methods, it is not applicable to SFDA due to the inaccessibility of the source data. In this study, we instead employed the theory of self-training based on the expansion assumption [48] as a way to understand SFDA methods.

Source-free Domain Adaptation (SFDA). With reference to [24], SFDA methods can be roughly categorized into four approaches. Source-estimation approaches [11, 17, 22, 33, 42, 49] generate pseudo-source data using a pre-trained model, which transforms the SFDA problem into a conventional UDA problem. Pseudo-labeling approaches [19, 34, 51] assign a class label to each unlabeled target sample using the current model and use them in a supervised manner. Based on the cluster assumption [46], clustering approaches [21–23] encourage minimizing the uncertainty of the model predictions or performing clustering over the target features. Inspired by the consistency regularization of semi-supervised learning [4, 39, 41], consistency approaches [5, 52, 53, 57] train the model to maximize the prediction consistency regardless of the perturbations on the input data or the model parameters.

Despite many successful SFDA methods, most do not have a theoretical foundation yet, except for the source estimation approaches that convert SFDA to conventional UDA. A few studies [57] have theoretically investigated their methods, but they are not applicable to the others. In this paper, we introduce the theory of self-training based on the expansion assumption [48] to give a theoretical perspective on a wide range of SFDA methods, and we propose an improved method based on our theoretical analysis.

3. Understanding SFDA from a Theoretical Perspective

In this section, we define the SFDA problem, and then, see the common feature of SFDA methods. Finally, we provide a theoretical analysis of SFDA through [48].

3.1. Problem Definition

The upper right of Fig. 1 illustrates the SFDA problem definition. We are given a model $F_S : \mathcal{X} \rightarrow \mathcal{Y}$ trained on a labeled source domain data D_S and an unlabeled target domain data $D_T = \{x_t^{(i)}\}_{i=1}^{n_t}$. Generally, the model F consists of a feature extractor f and a fully-connected layer based classifier g. The goal is to train the model F to obtain a target-adapted model $\hat{F}$ that has low target error without using source domain data D_S nor target labels $y_t^{(i)}$.

3.2. Common Feature of SFDA Methods

Whilst lacking a precise theoretical background, the following observation [9, 53] provides a key to understanding SFDA methods: the existing SFDA methods have a common feature that their training loss functions can be decomposed into discriminability and diversity losses wherein the discriminability loss enhances the model discriminability to the unlabeled target samples while the diversity loss ensures the model has predictions for diverse classes.

For example, SHOT-IM [23], a pioneering work on SFDA, trains the model on the basis of mutual information maximization, i.e., a training loss function comprising conditional entropy minimization (discriminability) and marginal entropy maximization (diversity):

$$\mathcal{L}_{\text{MIM}} = \underbrace{H(Y|X)}_{\text{discriminability}} - \lambda_{\text{div}} \underbrace{H(Y)}_{\text{diversity}}. \tag{1}$$

where λ_{div} represents the trade-off parameter of the loss.

Another example is AaD [53], which trains the model by maximizing the prediction similarities among the local neighborhoods in the feature space (discriminability) while minimizing those of the others (diversity) as follows:

$$\mathcal{L}_{\text{AaD}} = \frac{1}{|\mathbf{B}|} \sum_{i \in \mathbf{B}} \left\{ \underbrace{\frac{1}{|\mathbf{C}_i|} \sum_{\hat{\mathbf{p}} \in \mathbf{C}_i} -\mathbf{p}_i \cdot \hat{\mathbf{p}}}_{\text{discriminability}} + \lambda_{\text{div}} \underbrace{\sum_{j \in \mathbf{B} \setminus \{i\}} \mathbf{p}_i \cdot \mathbf{p}_j}_{\text{diversity}} \right\}, \tag{2}$$

where $\mathbf{B}$ is a mini-batch, $\mathbf{p}_i$ is the prediction of sample i, and $\mathbf{C}_i$ is a set of the predictions of K-nearest neighborhoods of sample i on the feature space.

This common feature is widely seen in other methods, such as those using pseudo-labeling or consistency regularization as their own discriminability loss with the above marginal entropy maximization (diversity) [34, 51, 52], and those employing contrastive learning, which maximizes the similarity of positive pairs (discriminability) and minimizes the similarity of negative pairs (diversity) [5, 19, 57].

Now that we have confirmed the discriminability and diversity losses to be the key to the success of SFDA methods, but *why are these losses crucial for the success of SFDA?*

3.3. Theoretical Understanding of SFDA

We will answer the above question by introducing the theory of self-training based on the expansion assumption [48]. This theory shows that, under certain assumptions, the model will have a low target error when it is self-trained based on prediction consistency while a constraint is imposed to ensure prediction diversity.

Notations. We let $\mathcal{A}$ denote the family of data augmentation and define an augmented sample set an input x as $\mathcal{B}(x) := \{x' \mid \exists A \in \mathcal{A} \text{ s.t. } \|x' - A(x)\| < r\}$, the neighborhoods of x as $\mathcal{N}(x) := \{x' \mid \mathcal{B}(x) \cap B(x') \neq \emptyset\}$, the neighborhoods of the set V as $\mathcal{N}(V) := \cup_{x \in V} \mathcal{N}(x)$, and prediction inconsistency for a C-class prediction model $F : \mathcal{X} \to [C]$ on a distribution P as $R_{\mathcal{B}}(F) := \mathbb{E}_P[\mathbf{1}(\exists x' \in \mathcal{B}(x) \text{ s.t. } F(x') \neq F(x))]$.

Assumptions. Before explaining the main theorem that is the key to our theoretical analysis, we must make two assumptions: *Expansion* and *Separation*[1].

Expansion assumes that the data of the same class are distributed in a continuous region and that any small region V has a neighborhood region of the same class larger than V. Concretely, $P_i(\mathcal{N}(V)) \geq dP_i(V)$ for $P_i(V) \leq 1/2$, $d > 1$, where $P_i(V)$ represents the proportion of subset V in the total class i data and d represents an expansion factor that corresponds to the strength of the data augmentation.

Separation assumes that the distribution of different classes is separated and the predictions of the ground-truth model $F^\star$ will not be altered by the data augmentation, *i.e.*, $R_{\mathcal{B}}(F^\star) < \mu$, where μ represents a negligible value.

Theory for Understanding SFDA. With the above assumptions, [48] derives the following theorem that is key to understanding why existing SFDA methods perform well.

Theorem 1. *Suppose that the above two assumptions hold for some d, μ such that $\min_{y \in [C]} P(\{x : F^\star(x) = y\}) > \max\{2/(d-1), 2\}\mu$. Then any minimizer $\hat{F}$ of*

$$\underbrace{\min_F R_{\mathcal{B}}(F)}_{\textit{self-training = discriminability}} \qquad \text{subject to}$$

$$\underbrace{\min_{y \in [C]} \mathbb{E}_P[\mathbf{1}(F(x) = y)] > \max\left\{\frac{2}{d-1}, 2\right\} R_{\mathcal{B}}(F)}_{\textit{constraint = diversity}} \tag{3}$$

satisfies

$$\text{Err}_{\text{U}}(\hat{F}) \leq \max\left\{\frac{d}{d-1}, 2\right\} \mu, \tag{4}$$

where $\text{Err}_{\text{U}}(\hat{F}) := \min_{\pi:[C] \to [C]} [\mathbf{1}(\pi(F(x)) \neq F^\star(x))]$, *and π represents a permutation.*

Theorem 1 shows that a model trained with the objective function (3) consisting of a self-training term that requires neighborhood predictions to be consistent and a constraint term that assigns a certain portion of predictions to all classes has an upper bound on the target error (4).

What we want to highlight is that the SFDA training loss with discriminability and diversity loss acts the same as the objective (3). Specifically, increasing the discriminability leads to a reduction in the prediction inconsistency of the self-training term while the diversity loss functions in the same way as the constraint term that ensures the minor class predictions. This indicates that we can apply Theorem 1 to the SFDA training loss, and thus, the SFDA-trained model also has an upper bound on the target error.

[1] Formal statements of the assumptions are given in Appendix A.

Method	Acc [%]
i) Source only	72.4
ii) Dis only	93.0
iii) Dis + Div	95.3
iv) Dis + Div w/ Decay	**95.7**

Table 1. Model accuracy w/ and w/o discriminability and diversity losses.

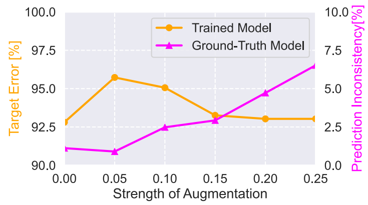

Figure 2. Model accuracy ($\mathrm{Err}_{\mathrm{U}}(\hat{F})$) and the prediction inconsistency of the ground-truth model (μ) against strength of augmentation (d).

Preliminary Experiment with Synthetic Data. We verified this correspondence through an experiment on a two-dimensional synthetic dataset. We employed a variant of the inter-twinning moons 2D dataset, where we simulated the domain shift by rotation. We used the training objective of SHOT-IM [23] and 2D-Gaussian perturbations as the data augmentation. In the experiment, we compared the accuracy of i) Source only, ii) Discriminability (Dis) only, and iii) Discriminability and Diversity (Dis + Div). Moreover, we measured the accuracy of the trained model $\mathrm{Err}_{\mathrm{U}}(\hat{F})$ and the prediction inconsistency of the ground-truth model corresponding to μ versus the strength of the augmentation corresponding to d. Other details on the experimental settings are described in Appendix B.

The results are summarized in Tab. 1 and Fig. 2. Tab. 1 shows that the accuracy is improved by incorporating the diversity loss in addition to the discriminability loss, which establishes the necessity of the diversity loss to function as the constraint of the objective (3). The prediction accuracy of the trained model in Fig 1 aligns with the upper bound (4); namely, as shown in Fig. 2, by keeping the prediction inconsistency of the ground-truth model μ low and increasing the strength of the data augmentation d, the model can achieve a low target error.

Theoretical Insights. Besides, Theorem 1 provides two insights for the further improvement of SFDA methods.

First, the weight of the diversity loss λ_{div} should be adjusted as training progresses, and this can be done by controlling the value of the discriminability loss. The right-hand side (RHS) of the constraint in the objective (3) includes $R_{\mathcal{B}}(F)$, which decreases during the training. This indicates that it is more reasonable to let the constraint decay along with $R_{\mathcal{B}}(F)$, *i.e.*, the discriminability loss. The results shown in Tab 1 also demonstrate that decaying λ_{div} (Dis + Div w/ Decay) brings a better result. However, most of the existing SFDA methods fix λ_{div}, which would be sub-optimal. Although AaD [53] exceptionally uses a manually designed scheduler, tuning it is laborious.

Second, the upper bound of the target error depends on the parameters d and μ which are relevant to the data augmentation properties. This indicates that how we design the data augmentation is a critical factor in training models with better accuracy. However, the prior studies on SFDA [5, 57] have paid less attention to it and have used pre-defined data augmentations [7, 8], which may not be optimal for SFDA.

4. Our Method: Improved SFDA based on Theoretical Insights

Fig. 3 shows the overview of our method, which has three major components: a prediction model $F = g \circ f$, a teacher model $F' = g' \circ f'$, and a learnable augmentation $\mathcal{A}$. On the basis of the above insights, we developed an improved SFDA method comprising 1) Model Training with Auto-Adjusting Diversity Constraint, and 2) Augmentation Training with Teacher-Student Framework.

4.1. Model Training with Auto-Adjusting Diversity Constraint

We update F upon the modified discriminability and diversity losses of AaD [53], coupled with a novel technique to automatically adjust the trade-off parameter between discriminability and diversity.

Discriminability and Diversity Losses. Considering that the prediction inconsistency $R_{\mathcal{B}}(F)$ in the self-training term is originally calculated among data-augmented samples, we modify the training loss (2) so as to calculate the prediction dissimilarity among data-augmented samples. The discriminability and diversity losses are formally defined as

$$\mathcal{L}_{\text{dis}} = \frac{1}{M|\mathbf{B}||\mathbf{C}_i|} \sum_{i \in \mathbf{B}} \sum_{\hat{\mathbf{p}} \in \mathbf{C}_i} \sum_{m=1}^{M} (1 - \mathbf{p}_i^m \cdot \hat{\mathbf{p}}), \tag{5}$$

$$\mathcal{L}_{\text{div}} = \frac{1}{M|\mathbf{B}|} \sum_{i \in \mathbf{B}} \sum_{j \in \mathbf{B} \setminus \{i\}} \sum_{m=1}^{M} \mathbf{p}_i^m \cdot \mathbf{p}_j, \tag{6}$$

where $\mathbf{p}_i^m = F(A_m(x_t^{(i)}))$, A_m is the m-th augmentation sampled from $\mathcal{A}$, M is the number of augmentations. To retrieve the K-nearest neighbors $\mathbf{C}_i$ efficiently, we build a memory bank that stores the feature vectors $\mathbf{z}_i = f(x_t^{(i)})$ and predictions $\mathbf{p}_i = F(x_t^{(i)})$ of all target samples in D_T.

Auto-Adjusting Diversity Constraint. As shown in Sec. 3.3, λ_{div} should be adjusted as the training progresses. Specifically, the RHS of the constraint of the objective (3) involves the prediction inconsistency $R_{\mathcal{B}}(F)$ which will get smaller as the training progresses. Moreover, the discriminability loss $\mathcal{L}_{\text{dis}}$ functions the same way as $R_{\mathcal{B}}(F)$. This means that we can easily control λ_{div} with $\mathcal{L}_{\text{dis}}$. Accordingly, the auto-adjusting λ_{div} can be simply expressed as

$$\lambda_{\text{div}} = \lambda_{\text{div}}^{\max} \mathcal{L}_{\text{dis}}, \tag{7}$$

where $\lambda_{\text{div}}^{\max}$ determines the maximum size of λ_{div}. Note that we apply the stop-gradient operation to λ_{div}.

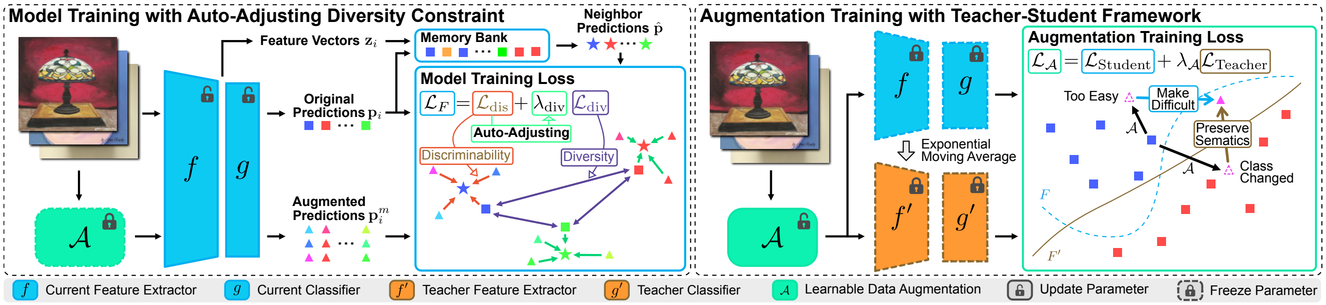

Figure 3. **Overview of our method.** In Model Training with Auto-Adjusting Diversity Constraint, we train the model by minimizing the discriminability loss $\mathcal{L}_{\text{dis}}$ and the diversity loss $\mathcal{L}_{\text{div}}$ while automatically adjusting the trade-off parameter between discriminability and diversity. In Augmentation Training with Teacher-Student Framework, we train a learnable data augmentation $\mathcal{A}$ to generate harder samples for the current model F while suppressing the prediction inconsistency of the teacher model F'.

Model Training. The training loss of the model using the learnable augmentation $\mathcal{A}$ is

$$\mathcal{L}_F = \mathcal{L}_{\text{dis}} + \lambda_{\text{div}} \mathcal{L}_{\text{div}}. \tag{8}$$

However, we find in an early study that the learnable data augmentation $\mathcal{A}$ (whose details will be described in Sec. 4.2) yields some heavy augmentations (*e.g.*, `rotation`, `invert`, *etc.*), which may impair the model training. Inspired by [1], we stabilize the training by incorporating a loss $\mathcal{L}'_F$ calculated among samples with weak augmentations (*e.g.*, `random_clip`, `random_flip`).

In summary, the total loss of the model training is

$$\min_F \mathcal{L}_F^{\text{Total}} = \lambda_F \mathcal{L}_F + (1 - \lambda_F) \mathcal{L}'_F. \tag{9}$$

where λ_F is a hyper-parameter to control the loss balance.

4.2. Augmentation Training with Teacher-Student Framework

As discussed in Sec. 3.3, the upper bound of the target error depends on how the data augmentation is designed. Motivated by [40], we update the learnable data augmentation $\mathcal{A}$ in the teacher-student framework to get a tighter bound.

Learnable Data Augmentation. $\mathcal{A}$ consists of L different augmentations $A^{(l)}$ $(l = 1, 2, \cdots, L)$. A single augmentation consists of N consecutive transformation operations $O_1^{(l)}, \cdots, O_N^{(l)}$. Each operation includes affine transformations (*e.g.* `shear_x`) and color enhancing operations (*e.g.* `contrast`), and it has a magnitude parameter $m_n^{(l)} \in [0, 1]$ to control the transformation strength and a probability parameter $p_n^{(l)} \in [0, 1]$ to control whether to apply the operation. To facilitate parameter optimization, we utilize Faster AutoAugment [13] to make these parameters differentiable and updatable by gradient descent.

Augmentation Training. We train $\mathcal{A}$ to make the bound tighter based on the observation that the upper bound of the target error (4) becomes tighter as 1) the strength of the augmentation d gets larger and 2) the prediction inconsistency of the ground-truth model μ gets smaller. Since we cannot actually access the ground-truth model $F^\star$, we use the teacher-student framework and assign the teacher model to be a proxy for $F^\star$; 1) we increase d by training $\mathcal{A}$ to augment samples that are harder for the current (student) model to predict, and 2) we decrease μ by training $\mathcal{A}$ to augment samples that are recognizable to the teacher model.

More specifically, 1) we train $\mathcal{A}$ to maximize the prediction entropy of the current model:

$$\mathcal{L}_{\text{Student}} = -\frac{1}{|\mathbf{B}|} \sum_{i \in \mathbf{B}} \sum_{c=1}^{C} p_i[c] \log(1 - p_i[c]), \tag{10}$$

where $p_i[c] = F(A(x_t^{(i)}))[c]$ is the prediction probability of sample i for class c, and A is a augmentation operation randomly sampled from $\mathcal{A}$. Following [40], we employ a non-saturating prediction entropy instead of the naive one.

Whereas, 1) we train $\mathcal{A}$ to minimize the prediction entropy of the teacher model:

$$\mathcal{L}_{\text{Teacher}} = -\frac{1}{|\mathbf{B}|} \sum_{i \in \mathbf{B}} \sum_{c=1}^{C} p'_i[c] \log p'_i[c], \tag{11}$$

where $p'_i[c] = F'(A(x_t^{(i)}))[c]$. The teacher model is updated using the exponential moving average strategy [41]:

$$F' \leftarrow (1 - \beta) F + \beta F', \tag{12}$$

where β is a momentum parameter.

The total loss of the augmentation training is

$$\min_{\mathcal{A}} \mathcal{L}_{\mathcal{A}} = \mathcal{L}_{\text{Student}} + \lambda_{\mathcal{A}} \mathcal{L}_{\text{Teacher}}, \tag{13}$$

where $\lambda_{\mathcal{A}}$ is a coefficient hyper-parameter.

Discussion. Our method is different from the teacher-student-based SFDA methods [5, 56] in that we use the teacher-student framework for training data augmentation. Our approach may appear similar to the prior work [43] but fundamentally differs in that ours treats the teacher model as a proxy of the ground truth model based on our theoretical insights.

Algorithm 1 Training Procedure of Our Method.

Input: Source-trained model F_s, Target domain data D_T
Output: Target-adapted model $\hat{F}$

```
for e in {1 ··· E} do
    for Mini-batch B in D_T do                       ▷ Model Training
        Calculate loss L_F^Total on B
        Update F to minimize L_F^Total
        Update F' based on (12)
    end for
    if e ≡ 0 mod ê then
        for Mini-batch B in D_T do          ▷ Augmentation Training
            Calculate loss L_A on B
            Update A to minimize L_A
        end for
    end if
end for
```

4.3. Training Procedure of Our Method

As shown Algorithm 1, our method alternately performs model training and data augmentation training. To control the speed of these pieces of training, we set a parameter $\hat{e}$ that determines the interval of the augmentation training.

5. Experiments

We experimentally compared the performance of our method and the existing SFDA methods on three benchmark datasets, and we examined the validity of our theoretical insights through further analyses.

5.1. Setups

Datasets. We used three benchmark datasets: *Office-31* [36] is a small-scale dataset, which consists of 31 categories and 3 domains (**A**mazon, **W**ebcam and **D**slr). *Office-Home* [47] is a moderate-scale dataset, which consists of 65 categories and 4 domains (**Ar**t, **Cl**ipart, **Pr**oduct and **Re**al world). *VisDA2017* [32] is a large synthetic-to-real adaptation benchmark dataset with 12 categories. For *Office-31* and *Office-Home*, we evaluated the accuracy on all source-target combinations, while we computed the average of per-class accuracies for *VisDA2017*.

Network Architectures & Augmentation Implementations. We use ResNet-50 [14] as the backbone network in the *Office-31* and *Office-Home* experiments, and ResNet-101 [14] in *VisDA2017*. All of the networks are pre-trained on Imagenet [10]. Following [23], we replaced the output layer of the backbone network with the following networks: a *fully-connected layer* $\rightarrow$ *batch normalization* [18] $\rightarrow$ *fully-connected layer with weight normalization* [38]. We implemented the learnable data augmentation $\mathcal{A}$ with a public library of differentiable data augmentation[2]. We set L to 25, and N to 2 in all of the experiments.

[2] https://github.com/moskomule/dda

Table 2. **Classification Accuracy (%) on Office-31 (ResNet-50).** The best and second best are highlighted in **bold** and with underline.

Method (Source $\rightarrow$ Target)	A $\rightarrow$ D	A $\rightarrow$ W	D $\rightarrow$ W	W $\rightarrow$ D	D $\rightarrow$ A	W $\rightarrow$ A	Avg.
3C-GAN [22]	92.7	93.7	98.5	99.8	75.3	77.8	89.6
SHOT [23]	94.0	90.1	98.4	$\underline{99.9}$	74.7	74.3	88.6
VDM-DA [42]	94.1	93.2	98.0	**100.0**	75.8	77.1	89.7
A^2Net [49]	94.5	94.0	**99.2**	**100.0**	$\underline{76.7}$	76.1	90.1
NRC [51]	96.0	90.8	99.0	**100.0**	75.3	75.0	89.4
CPGA [33]	94.4	94.1	98.4	99.8	76.0	76.6	89.9
CoWA-JMDS [21]	94.4	**95.2**	98.5	99.8	76.2	$\underline{77.6}$	$\underline{90.3}$
C-SFDA [19]	$\underline{96.2}$	93.9	98.8	99.7	**77.3**	**77.9**	**90.5**
AaD [53]	**96.4**	92.1	$\underline{99.1}$	**100.0**	75.0	76.5	89.9
Improved SFDA	95.3	$\underline{94.2}$	98.3	$\underline{99.9}$	76.4	77.5	$\underline{90.3\ (+0.4)}$

Source Training. We use Nesterov SGD with a mini-batch size of 64 as the optimization algorithm on all three datasets.

For *Office-31* and *Office-Home*, we set the learning rate η to $1\mathrm{e}{-2}$ for the last replaced layers and $1\mathrm{e}{-3}$ for the backbone layers, momentum to 0.9, and weight decay to $5\mathrm{e}{-4}$. We used a standard cross entropy loss with label smoothing for training. The label smoothing parameter was set to 0.1. We trained the model for 50 epochs.

For *VisDA2017*, we set the initial learning rate η_{init} to $1\mathrm{e}{-3}$ for the last replaced layers and $1\mathrm{e}{-4}$ for the backbone layers, momentum to 0.9, and weight decay to $1\mathrm{e}{-3}$. The learning rate η was scheduled as; $\eta = \eta_{\text{init}}(1 + 10p)^{-0.75}$, where p was linearly increased from 0.0 to 1.0 throughout the training. We used the training losses of *Office-31* and *Office-Home*. We trained the model for 10 epochs.

Target Training. We use the Nesterov SGD for training F and AdamW [29] for training $\mathcal{A}$ with mini-batch size 64.

For *Office-31* and *Office-Home*, we set Nesterov SGD parameters as follows: the learning rate η for the second last layer to $3\mathrm{e}{-3}$ and for the backbone layers to $3\mathrm{e}{-4}$, momentum to 0.9, and weight decay to $5\mathrm{e}{-4}$. We fixed the last layer during the training, which yielded better results. We set the parameters of AdamW to the Pytorch default values [31] except for the learning rate η^{aug} to $5\mathrm{e}{-4}$. The number of the training epoch E was set to 100, and the interval $\hat{e}$ was set to three. The other parameters were set as follows; $\lambda_{\text{div}}^{\max}$ to 0.4 for *Office-31* and 0.75 for *Office-Home*, λ_F to 0.5, $\lambda_{\mathcal{A}}$ to 1.0, K to two, M to three, and β to 0.99. We initialized $\mathcal{A}$ with AutoAugment [7] Imagenet Policies.

For *VisDA2017*, we set the Nesterov SGD parameters as follows: the initial learning rate η_{init} for the last two layers to $2.5\mathrm{e}{-3}$ and for the backbone layers to $2.5\mathrm{e}{-4}$, momentum to 0.9, and weight decay to $1\mathrm{e}{-4}$. We set the initial learning rate of AdamW $\eta_{\text{init}}^{\text{aug}}$ to $2\mathrm{e}{-3}$ and the other parameters are set as default. η and η^{aug} are scheduled as in the source training. The training epoch E was set to 100 and the interval $\hat{e}$ to three. The other parameters were set as follows; $\lambda_{\text{div}}^{\max}$ to 0.08, λ_F to 0.2, $\lambda_{\mathcal{A}}$ to 1.0, K to five, M to three, and β to 0.99. We randomly initialized $\mathcal{A}$.

Table 3. **Classification Accuracy (%) on Office-Home (ResNet-50).** The best and second best are highlighted in **bold** and with underline.

Method (Source → Target)	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
SHOT [23]	57.1	78.1	81.5	68.0	78.2	78.1	67.4	54.9	82.2	73.3	58.8	84.3	71.8
A^{2}Net [49]	58.4	79.0	82.4	67.5	79.3	78.9	**68.0**	56.2	82.9	74.1	60.5	85.0	72.8
G-SFDA [52]	57.9	78.6	81.0	66.7	77.2	77.2	65.6	56.0	82.2	72.0	57.8	83.4	71.3
NRC [51]	57.7	**80.3**	82.0	68.1	79.8	78.6	65.3	56.4	83.0	71.0	58.6	85.6	72.2
CPGA [33]	59.3	78.1	79.8	65.4	75.5	76.4	65.7	58.0	81.0	72.0	64.4	83.3	71.6
U-SFAN [35]	57.8	77.8	81.6	67.9	77.3	79.2	67.2	54.7	81.2	73.3	60.3	83.9	71.9
CoWA-JMDS [21]	56.9	78.4	81.0	69.1	80.0	**79.9**	67.7	57.2	82.4	72.8	60.5	84.5	72.5
DaC [57]	59.1	79.5	81.2	69.3	78.9	79.2	67.4	56.4	82.4	74.0	**61.4**	84.4	72.8
C-SFDA [19]	60.3	80.2	**82.9**	69.3	**80.1**	78.8	67.3	58.1	**83.4**	73.6	61.3	86.3	**73.5**
AaD [53]	59.3	79.3	82.1	68.9	79.8	79.5	67.2	57.4	83.1	72.1	58.5	85.4	72.7
Improved SFDA	**60.7**	78.9	82.0	**69.9**	79.5	79.7	67.1	**58.8**	82.3	**74.2**	61.3	**86.4**	73.4 (+0.7)

Table 4. **Classwise Accuracy (%) on VisDA2017 (ResNet-101).** The best and second best are highlighted in **bold** and with underline.

Method (Synthetic → Real)	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Per-class
3C-GAN [22]	94.8	73.4	68.8	74.8	93.1	95.4	88.6	84.7	89.1	84.7	83.5	48.1	81.6
SHOT [23]	94.3	88.5	80.1	57.3	93.1	94.9	80.7	80.3	91.5	89.1	86.3	58.2	82.9
VDM-DA [42]	96.9	89.1	79.1	66.5	95.7	96.8	85.4	83.3	96.0	86.6	89.5	56.3	85.1
A^{2}Net [49]	94.0	87.8	85.6	66.8	93.7	95.1	85.8	81.2	91.6	88.2	86.5	56.0	84.3
G-SFDA [52]	96.1	83.3	85.5	74.1	97.1	95.4	89.5	79.4	95.4	92.9	89.1	42.6	85.4
NRC [51]	96.8	91.3	82.4	62.4	96.2	95.9	86.1	80.6	94.8	94.1	90.4	59.7	85.9
CPGA [33]	95.6	89.0	75.4	64.9	91.7	**97.5**	89.7	83.8	93.9	93.4	87.7	**69.0**	86.0
U-SFAN [35]	-	-	-	-	-	-	-	-	-	-	-	-	82.7
AdaContrast [5]	97.0	84.7	84.0	77.3	96.7	93.8	91.9	84.8	94.3	93.1	**94.1**	47.9	86.8
CoWA-JMDS [21]	96.2	89.7	83.9	73.8	96.4	97.4	89.3	**86.8**	94.6	92.1	88.7	53.8	86.9
DaC [57]	96.6	86.8	86.4	78.4	96.4	96.2	**93.6**	83.8	**96.8**	**95.1**	89.6	50.0	87.3
C-SFDA [19]	**97.6**	88.8	86.1	72.2	97.2	94.4	92.1	84.7	93.0	90.7	93.1	63.5	87.8
AaD [53]	97.4	90.5	80.8	76.2	**97.3**	96.1	89.8	82.9	95.5	93.0	92.0	64.7	88.0
Improved SFDA	97.5	**91.4**	**87.9**	**79.4**	97.2	97.2	92.2	83.0	96.4	94.2	91.1	53.0	**88.4 (+0.4)**

Table 5. **Analysis of Augmentation Training.**

Init. Aug	Aug Training	Accuracy [%]	ΔAaD
Random		73.1	+ 0.4
	✓	**73.3**	+ **0.6**
AutoAugment [7]		73.3	+ 0.6
	✓	**73.4**	+ **0.7**

Table 6. **Analysis of Auto-Adjusting Diversity Constraint.**

	$\lambda_{\text{div}}^{\text{max}}$	Accuracy [%]
Fixed Diversity Constraint	0.1	69.3
	0.25	72.8
	0.5	72.6
	0.75	71.1
Auto-Adjusting Diversity Constraint	0.75	**73.4**

5.2. Main results

We evaluated the performance of our method by taking the average score of three different runs for all benchmarks.

Result on Office-31. The results are shown in Tab. 2. Our method improved accuracy by 0.4% on average compared with the baseline AaD [53]. Furthermore, ours was comparable in accuracy to the second best method, CoWA-JMDS [21] and only 0.2% off the best method C-SFDA [19].

Result on Office-Home. The results are shown in Tab. 3. Ours improved accuracy by 0.7% on average compared with AaD [53] and was second best. The accuracy difference from the best method, C-SFDA [19], was merely 0.1%.

Result on VisDA2017. The results are shown in Tab. 4. Ours improved accuracy by 0.4% compared with AaD [53]. The average of per-class accuracy reached 88.4%, which was the best among the compared methods.

5.3. Analysis

Ablation study. Using *Office-Home*, we analyzed the effectiveness of the proposed two components.

Augmentation Training with Teacher-Student Framework was validated by comparing its performance with that of a fixed augmentation variant. The results are shown in Tab. 5. When starting with the randomly initialized data augmentation, our augmentation training yielded a larger accuracy gain from the base method (ΔAaD), 1.5 times greater than without augmentation training. The accuracy gain is slightly smaller when the augmentation is initialized with AutoAugment policies, but ours still yielded better results. More detailed results are provided in Appendix C.1.

Model Training with Auto-Adjusting Diversity Constraint was validated by comparing our method with a variant that uses a fixed λ_{div}, *i.e.*, $\lambda_{\text{div}} = \lambda_{\text{div}}^{\text{max}}$. Considering that the optimal $\lambda_{\text{div}}^{\text{max}}$ for this "Fixed Diversity Constraint" may differ from that of our method, we performed experiments with several values of $\lambda_{\text{div}}^{\text{max}}$. The results are shown in

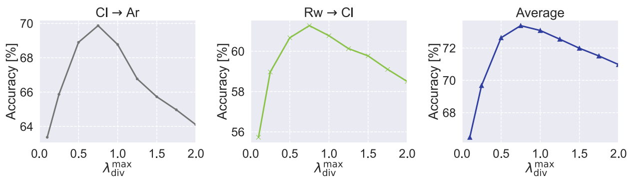

Figure 4. **Analysis of the coefficient parameter $\lambda_{\mathrm{div}}^{\max}$**

Table 7. **Analysis of Application to SHOT-IM.**

	Office-31	Office-Home	VisDA2017
SHOT-IM [23]	87.3	70.5	80.4
Improved SFDA	**88.6 (+1.3)**	**71.2 (+0.7)**	**83.3 (+2.9)**

Tab. 6. Our method outperformed all fixed λ_{div} variants, which indicates the validity of our proposed techniques. We also obtained the same results from the synthetic data experiment, which is described in Appendix B.2.

Applicability to other SFDA methods. Although we based on the implementation of our method on AaD, the theoretical insights and techniques we have made here should be applicable to many other SFDA methods. To confirm this, we empirically verified the applicability of our proposed techniques to another SFDA method, SHOT-IM [23]. Here, we used the same hyper-parameters as in Sec. 5.1, except for the following points; we set $\lambda_{\mathrm{div}}^{\max}$ to 0.7 and η to 2e$-$3 for *Office-31* and *Office-Home*, η_{init} to 1e$-$3 and $\lambda_{\mathrm{div}}^{\max}$ to 0.8 for *VisDA2017*.

The results are shown in Tab. 7. We can see a steady improvement in all the benchmarks, which demonstrate the effectiveness of our techniques for SHOT-IM. Moreover, since many of the existing SFDA methods are built upon SHOT-IM, this result implies that our method is valid for a wider range of SFDA methods.

Hyper-parameter analysis. We conducted experiments using *Office-Home* to analyze the effect of the parameter $\lambda_{\mathrm{div}}^{\max}$ in Model Training with Auto-Adjusting Diversity Constraint, and on the parameters $\hat{e}$ and $\lambda_{\mathcal{A}}$ in Augmentation Training with Teacher-Student Framework.

Analysis of $\lambda_{\mathrm{div}}^{\max}$ (Fig. 4) $\lambda_{\mathrm{div}}^{\max}$ controls the strength of the prediction diversity constraint. We varied $\lambda_{\mathrm{div}}^{\max}$ from 0.0 to 2.0 and evaluated the accuracy. $\lambda_{\mathrm{div}}^{\max}$ is optimal at 0.75, and the accuracy deteriorates if it is larger or smaller than the optimal value. In particular, the accuracy deteriorates more sharply when it takes a smaller value than a larger value. This result is in line with our theoretical analysis. Specifically, if $\lambda_{\mathrm{div}}^{\max}$ is too small, the diversity loss will not play the role of constraining the objective (3) sufficiently, and thus we can not obtain the upper bound for the target error. However, if it is too large, the diversity loss becomes an excessive constraint, and that reduces the accuracy.

Analysis of $\hat{e}$ (Fig. 5). $\hat{e}$ controls the frequency of the augmentation training. We analyzed the effect of $\hat{e}$ by varying it from one to six. The optimal value of $\hat{e}$ is two or three and

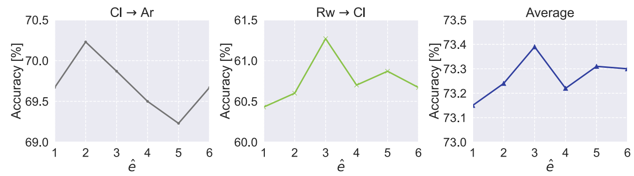

Figure 5. **Analysis of the interval parameter $\hat{e}$**

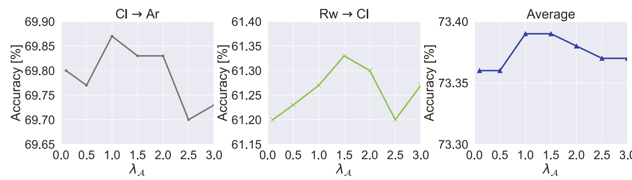

Figure 6. **Analysis of the coefficient parameter $\lambda_{\mathcal{A}}$**

the performance is lower if it is set to a larger or smaller value than this. Our augmentation training will not demonstrate its validity unless the current model and the teacher model are different to some extent. When $\hat{e}$ is small, the teacher and the current model are too close to exhibit the full potential of the augmentation training, while when $\hat{e}$ is large, the data augmentation is not trained well enough.

Analysis of $\lambda_{\mathcal{A}}$ (Fig. 6). $\lambda_{\mathcal{A}}$ controls the effect of $\mathcal{L}_{\mathrm{Student}}$ and $\mathcal{L}_{\mathrm{Teacher}}$, where $\mathcal{L}_{\mathrm{Student}}$ increases d of the bound (4) by encouraging the augmentation to generate more difficult samples while $\mathcal{L}_{\mathrm{Teacher}}$ decreases μ of the bound (4) by encouraging the augmentation to keep the semantics of the data. We analyzed how our method performed while varying $\lambda_{\mathcal{A}}$ from 0.0 to 3.0. The optimal value of $\lambda_{\mathcal{A}}$ is 1.0 to 1.5. If the balance is not proper, it will cause a decrease in d or an increase in μ, resulting in the prediction model having poor accuracy. The results thus demonstrate that our method is consistent with our theoretical insights.

6. Conclusion

We shed light on the theoretical perspective of existing SFDA methods through the theory of self-training based on the expansion assumption [48]. Our finding that the SFDA training loss with discriminability and diversity functions the same way as the training objective of the theory not only provided a way to understand existing SFDA methods but also yielded two novel techniques for improving the performance of SFDA methods. The experimental results and in-depth analysis justified the validity of our theoretical insights and proposed method. We expect that this work will encourage further development of SFDA research.

On the other hand, since this study aims to understand existing SFDA methods, we leave one limitation. That is, we cannot take into account how well the source model originally performs on the target domain, which is also not considered in existing SFDA research. One of our future works is to develop a method that overcomes this weakness and advances SFDA to be more practical.

References

[1] Yalong Bai, Yifan Yang, Wei Zhang, and Tao Mei. Directional self-supervised learning for heavy image augmentations. In *Proc. CVPR*, 2022. 5

[2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Proc. NeurIPS*, 2006. 2

[3] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010. 1, 2

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Proc. NeurIPS*, 2019. 2

[5] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proc. CVPR*, 2022. 1, 2, 3, 4, 5, 7

[6] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proc. CVPR*, 2018. 1

[7] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proc. CVPR*, 2019. 4, 6, 7

[8] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proc. NeurIPS*, 2020. 4

[9] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proc. CVPR*, 2020. 2, 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009. 6

[11] Yuntao Du, Haiyang Yang, Mingcai Chen, Juan Jiang, Hongtao Luo, and Chongjun Wang. Generation, augmentation, and alignment: A pseudo-source domain based method for source-free domain adaptation. *arXiv preprint arXiv:2109.04015*, 2021. 1, 2

[12] 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*, 17(1):2096–2030, 2016. 1, 2

[13] Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, and Hideki Nakayama. Faster autoaugment: Learning augmentation strategies using backpropagation. In *Proc. ECCV*, 2020. 5

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 6

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proc. ICML*, 2018. 1

[16] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proc. CVPR*, 2018. 1

[17] Yunzhong Hou and Liang Zheng. Source free domain adaptation with image translation. *arXiv preprint arXiv:2008.07514*, 2021. 1, 2

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. ICML*, 2015. 6

[19] Nazmul Karim, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-pang Chiu, Supun Samarasekera, and Nazanin Rahnavard. C-sfda: A curriculum learning aided self-training framework for efficient source free domain adaptation. In *Proc, CVPR*, 2023. 1, 2, 3, 6, 7

[20] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *Proc. ICML*, 2020. 1, 2

[21] Jonghyun Lee, Dahuin Jung, Junho Yim, and Sungroh Yoon. Confidence score for source-free unsupervised domain adaptation. In *Proc. ICML*, 2022. 1, 2, 6, 7

[22] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proc. CVPR*, 2020. 2, 6, 7

[23] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proc. ICML*, 2020. 1, 2, 3, 4, 6, 7, 8

[24] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *arXiv preprint arXiv:2303.15361*, 2023. 2

[25] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. In *Proc. NeurIPS*, 2021. 1, 2

[26] M Long, Y Cao, J Wang, and MI Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015. 1, 2

[27] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proc. ICML*, 2017. 1, 2

[28] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Proc. NeurIPS*, 2018. 1, 2

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. ICLR*, 2018. 6

[30] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proc. COLT*, 2009. 1, 2

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Proc. NeurIPS*, 2019. 6

[32] 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. 2, 6

[33] Zhen Qiu, Yifan Zhang, Hongbin Lin, Shuaicheng Niu, Yanxia Liu, Qing Du, and Mingkui Tan. Source-free domain adaptation via avatar prototype generation and adaptation. In *Proc. IJCAI*, 2021. 1, 2, 6, 7

[34] Sanqing Qu, Guang Chen, Jing Zhang, Zhijun Li, Wei He, and Dacheng Tao. Bmd: A general class-balanced multicentric dynamic prototype strategy for source-free domain adaptation. In *Proc. ECCV*, 2022. 1, 2, 3

[35] Subhankar Roy, Martin Trapp, Andrea Pilzer, Juho Kannala, Nicu Sebe, Elisa Ricci, and Arno Solin. Uncertainty-guided source-free domain adaptation. In *Proc. ECCV*, 2022. 7

[36] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proc. ECCV*, 2010. 2, 6

[37] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proc. CVPR*, 2019. 1

[38] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Proc. NeurIPS*, 2016. 6

[39] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proc. NeurIPS*, 2020. 2

[40] Teppei Suzuki. Teachaugment: Data augmentation optimization using teacher knowledge. In *Proc. CVPR*, 2022. 5

[41] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. NeurIPS*, 2017. 2, 5

[42] Jiayi Tian, Jing Zhang, Wen Li, and Dong Xu. Vdm-da: Virtual domain modeling for source data-free domain adaptation. *IEEE TCSVT*, 32(6):3749–3760, 2021. 1, 2, 6, 7

[43] Devavrat Tomar, Guillaume Vray, Behzad Bozorgtabar, and Jean-Philippe Thiran. Tesla: Test-time self-learning with automatic adversarial augmentation. In *Proc. CVPR*, 2023. 5

[44] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 1, 2

[45] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proc. CVPR*, 2017. 1, 2

[46] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020. 2

[47] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proc. CVPR*, 2017. 2, 6

[48] Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *Proc. ICLR*, 2020. 1, 2, 3, 8

[49] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *Proc. CVPR*, 2021. 1, 2, 6, 7

[50] Jinyu Yang, Chunyuan Li, Weizhi An, Hehuan Ma, Yuzhi Guo, Yu Rong, Peilin Zhao, and Junzhou Huang. Exploring robustness of unsupervised domain adaptation in semantic segmentation. In *Proc. ICCV*, 2021. 1

[51] Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In *Proc. NeurIPS*, 2021. 1, 2, 3, 6, 7

[52] Shiqi Yang, Yaxing Wang, Joost Van De Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proc. ICCV*, 2021. 1, 2, 3, 7

[53] Shiqi Yang, Yaxing Wang, Kai Wang, Shangling Jui, et al. Attracting and dispersing: A simple approach for source-free domain adaptation. In *Proc. NeurIPS*, 2022. 1, 2, 3, 4, 6, 7

[54] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proc. CVPR*, 2018. 1, 2

[55] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *Proc. ICML*, 2019. 1, 2

[56] Yixin Zhang, Zilei Wang, and Weinan He. Class relationship embedded learning for source-free unsupervised domain adaptation. In *Proc. CVPR*, 2023. 5

[57] Ziyi Zhang, Weikai Chen, Hui Cheng, Zhen Li, Siyuan Li, Liang Lin, and Guanbin Li. Divide and contrast: Source-free domain adaptation via adaptive contrastive learning. In *Proc. NeurIPS*, 2022. 1, 2, 3, 4, 7

[58] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *Proc. ICML*, 2019. 1, 2

[59] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proc. CVPR*, 2019. 1

[60] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proc. ECCV*, 2018. 1, 2

[61] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proc. ICCV*, 2019. 1, 2