Consistency Regularization for Generalizable Source-free Domain Adaptation

Longxiang Tang[1] Kai Li[2] Chunming He[1] Yulun Zhang[3] Xiu Li[1]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]NEC Laboratories America [3]ETH Zurich, Switzerland
{lloong.x, li.gml.kai, chunminghe19990224, yulun100}@gmail.com
li.xiu@sz.tsinghua.edu.cn

Abstract

Source-free domain adaptation (SFDA) aims to adapt a well-trained source model to an unlabelled target domain without accessing the source dataset, making it applicable in a variety of real-world scenarios. Existing SFDA methods ONLY assess their adapted models on the target training set, neglecting the data from unseen but identically distributed testing sets. This oversight leads to overfitting issues and constrains the model's generalization ability. In this paper, we propose a consistency regularization framework to develop a more generalizable SFDA method, which simultaneously boosts model performance on both target training and testing datasets. Our method leverages soft pseudo-labels generated from weakly augmented images to supervise strongly augmented images, facilitating the model training process and enhancing the generalization ability of the adapted model. To leverage more potentially useful supervision, we present a sampling-based pseudo-label selection strategy, taking samples with severer domain shift into consideration. Moreover, global-oriented calibration methods are introduced to exploit global class distribution and feature cluster information, further improving the adaptation process. Extensive experiments demonstrate our method achieves state-of-the-art performance on several SFDA benchmarks, and exhibits robustness on unseen testing datasets.

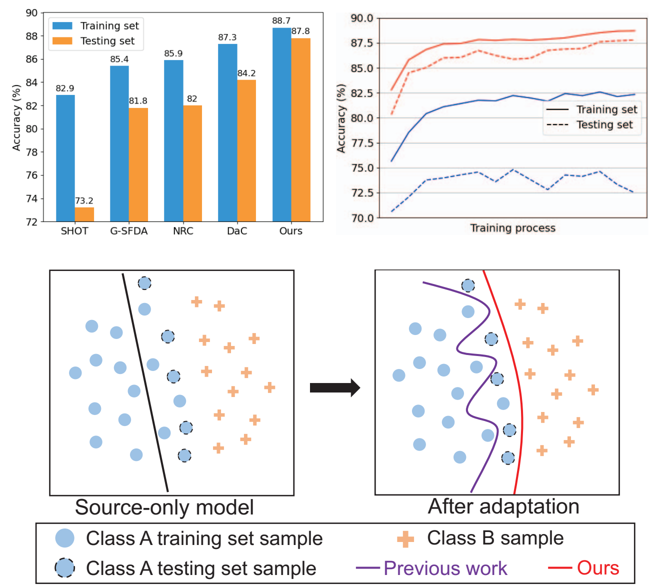


Figure 1. (**Top left**) Accuracy (%) comparison with existing methods [27, 51, 53] on target training and testing data of VisDA-2017 dataset [34]. (**Top right**) Training curve. Our results are in red, while SHOT [27] is in blue. (**Bottom**) Diagram of adaptation effect comparison. Existing SFDA methods only evaluate and improve their performance on target training data, which shows a performance degradation on unseen but identically distributed data. Our approaches construct a generalizable model, boosting the performance both on training and testing data.

1. Introduction

Deep neural networks achieve impressive success on various tasks [18, 9, 7, 25] while suffering from performance degradation when applied to data with a different distribution, which is called domain shift [45]. *Unsupervised Domain Adaptation* (UDA) approaches provide a promising solution to address this issue by learning an adaptive model jointly with labeled images (source domain) and unlabeled images with shifted distribution (target domain). Most existing approaches [35, 12, 40, 50] focus on inter-domain feature alignment by leveraging both domain data distribution, which implies an availability to source domain data when performing an adaptation strategy. However, considering real-world scenarios, source datasets storage costs and privacy issues may limit the accessibility to source data. Source datasets for pretraining are always on a large scale, which makes it infeasible to deploy traditional UDA methods to portable devices like mobile phones, though source model size is reasonable [27]. Meanwhile, source data might be privacy sensitive and related to copyright restrictions, like medical images and commercial data.

To this end, recently a new setting of UDA which alleviates the need to access source data, called *Source-free Domain Adaptation* (SFDA), has drawn increasing research interest. SHOT [27] performs a source hypothesis transferring by fixing the classifier and adopting information maximization loss, NRC [51] and G-SFDA [53] encourage similar prediction among nearest neighbors, and SFDA-DE [8] aligns surrogate features from estimated source distribution to target features. These existing paradigms have shown promising results under a transductive learning manner, where the models are trained and tested on the same dataset, which often leads to overfitting problems [19].

However, considering the real-world application, it's impractical to collect all target data for model training. So the primary object of SFDA is to perform better on unseen target domain data, instead of only being effective on the training set. Thus, it is essential to propose an overfitting-robust model with enhanced generalization ability.

Shown in the top row of Fig. 1, models trained with previous works are suffering from performance degradation when handling unseen but identically distributed images, i.e., the testing data. As illustrated in the bottom part, we argue that the reason for previous works who perform well on training set but fail to classify testing samples, is that they only consider existing target samples and construct an unsmooth model manifold [3]. In this case, overfitting issues easily occur during the transductive learning process, especially since there's no external supervision signal under the SFDA setting. Although neighborhood-based methods [51, 53] show some insights to constrain the local consistency by nearest neighbors, model manifold smoothness still cannot be guaranteed due to potentially spare feature space and limited target samples.

To address the above issues, in this paper, we propose a consistency regularization framework to realize a generalizable SFDA. (1) To avoid overfitting and promote the model's generalization ability, we enforce consistent predictions between target data and its perturbation form derived by image augmentation. It's implemented by generating soft pseudo-labels with model outputs of weakly augmented images, and supervising the strongly augmented ones. With this, we could construct a smooth model manifold which is beneficial to learn a transferable model and avoid overfitting issues. (2) We argue that pseudo-label selection is crucial for avoiding error accumulation because no explicit supervision is provided under the SFDA setting. Different from previous thresholding methods who may omit valuable intrinsic information with a rough threshold, we present a sampling-based pseudo-label selection method, filtering samples with probability associated with their prediction confidence metric, which leverages potentially useful samples. (3) We also introduce global information to calibrate the point-wise operations. We leverage class prediction distribution to apply class-wise weighting, preventing class imbalance issues without introducing any separate class balance loss. To identify the noisy outliers, feature cluster information is integrated by calculating class prototypes and selecting reliable training samples.

Our contributions can be summarized as follows:

- We propose the consistency regularization framework under the SFDA setting for the first time, generating a robust and generalizable model. It effectively avoids overfitting issues and boosts performance both on target training and testing sets.
- To advance the effect of consistency regularization, we propose the sampling-based pseudo-label selection strategy to leverage more potential samples and the global-oriented calibration techniques to constrain and enhance the adaptation process.
- Massive experiments show that our proposed method outperforms existing approaches on DA benchmarks, especially with a large margin on target testing set.

2. Related Work

Unsupervised Domain Adaptation. UDA attempts to learn an adaptive model jointly with labeled images (source domain) and unlabeled images (target domain) [45], which can erase the burden of labeling target domain data. These years UDA has achieved great success on different tasks [50, 6, 59]. Most common existing UDA approaches are trying to align feature distribution between source and target domain, for instance, Maximum Mean Discrepancy (MMD) [13] based methods [35, 28] align feature embedding in reproducing kernel Hilbert space, which can reduce the distance of representations from each domain. Holding the same point, DANN [12] achieve this alignment through an adversarial training approach, by fooling a newly added domain classifier for extracting domain-invariant feature representation. More recently, SRDC [40] proposes to directly uncover the intrinsic target discrimination via discriminative clustering of target data, FixBi [32] performs bidirectional matching and self-penalization on intermediate domains generated by fixed ratio-based mixup. The success of these paradigms depends on the assumption of simultaneous access to source and target data, which is hard to guarantee in real-world scenarios.

Source-free Domain Adaptation. Considering the privacy and copyright restrictions when adapting models across domains, a new setting of domain adaptation, SFDA, is proposed and eliminates the requirement to source data during the adaptation procedure [27, 51, 8, 41, 24]. SHOT [27] freezes the source model classifier module (hypothesis) and exploits both information maximization and self-supervised pseudo-labeling to achieve feature alignment. A^{2}Net [46] develops an adversarial network to seek a target-specific classifier that advances the target hard samples' recognition. Based on the observation that target data still form clear clusters with the source model, NRC [51] and G-SFDA [53] encourage label consistency towards nearest neighbors in feature space. BUFR [10] focuses on measurement shifts between domains and applies a bottom-up source feature restoration technique to boost target feature representation. DaC [58] divides the target data into source-like and target-specific samples, to learn global class clustering with former and learn intrinsic local structures with later. The methods mentioned above provide some insights to achieve SFDA, however, they only consider and adapt samples in the target domain in a transductive manner, which could lead to an unsmooth manifold and raise overfitting issues.

Consistency Regularization. Consistency regularization is extensively investigated in the semi-supervised learning area, and the key point is similar inputs should share consistent predictions, which can enhance the network's feature representation. Mean Teacher [42] averages model weights using EMA over training steps and enforce consistency between student and teacher models, UDA [47] proves the importance of data augmentation and provides a new effective way to noise unlabeled examples. FixMatch [39] proposes a new framework to achieve consistency regularization by generating pseudo-labels from weakly-augmented images and using them to supervise the outputs of strongly-augmented ones. These approaches are under SSL setting, which need labeled samples to guide the model training process and can't satisfy the SFDA scenarios.

3. Method

In this section, we formalize the definition of source-free domain adaptation and our methodology to tackle it. Firstly we introduce notations used in SFDA, then present our basic consistency regularization framework aiming to develop a more robust and generalizable model. For further boosting the performance, a sampling-based pseudo-label selection strategy is proposed and analyzed in detail. Lastly, we leverage global information to avoid degradation by performing global-oriented calibration methods.

3.1. Preliminaries

For the vanilla UDA classification task, we are given a source dataset $D_s = (x_i^s, y_i^s)_{i=1}^{n_s}$ with n_s labeled training samples and a target dataset $D_t = x_{t\,i=1}^{i\,n_t}$ with n_t unlabeled samples, where $x_s^i \in \mathcal{X}_s, x_t^i \in \mathcal{X}_t$ and $\mathcal{X}_s$ holds different underlying marginal distribution with $\mathcal{X}_t$ as domain shift exists. Under the close-set [36] setting, source and target domain share the same label set $\mathcal{C} = \{1, 2, \cdots, K\}$ under the K-way classification task, and $y \in \mathbb{R}^K$ is the one-hot label. SFDA considers the inaccessibility to source data problem in the adaptation phase, so only a source model $h = f \circ g$ pre-trained on D_s and a target dataset D_t without any label is provided when applying adaptation. Similar to existing works, the source model can be divided into two components: feature extractor f who generates features $z_i = f(x_i)$ from input images and classifier g who provides final prediction $p_i = p(x_i) = \delta(g(z_i))$ given z where δ is the softmax function.

3.2. Consistency Regularization

Existing paradigms consider SFDA as merely a transductive learning process, in which they train and test models on the same set of samples, ignoring the robustness and generalization ability. However, in real-world applications, it's almost impossible to collect every target domain sample for applying adaptation, so we attempt to develop a generalizable model which can achieve high performance on unseen but identically distributed target data.

Manifold smoothness assumption, which is mostly considered in the semi-supervised learning (SSL) area [21, 42, 31, 47, 39], describes that a model's output shouldn't change when realistic perturbations are occurred in data points [54], to construct a more reliable feature representation. We argue that SFDA, similar to SSL which mainly exploits potentially beneficial supervision signals from unlabeled data, should also satisfy this assumption to keep model consistency for a better understanding of target structure and robustness of unseen identically distributed data.

We introduce consistency regularization terms to generate a smooth manifold. To realize perturbations on data points, strong augmentations $\mathcal{A}$ are applied to input images, simulating the noisy form of the original image. Model prediction of a strong augmented image is $p_i' = p(x_i') = \delta\left(h\left(\mathcal{A}\left(x_i\right)\right)\right)$. As p_i' and p_i are derived from the same semantic information, it's obvious that they should share consistent results, which means $\mathrm{Dist}(p_i, p_i')$ is small.

Many SSL methods [21, 42, 47] attempt to minimize this distance with massive implementations, as an auxiliary task to labeled sample loss. Since no label information is available, these discrepancy minimization methods are insufficient to develop a target adaptive model under the SFDA setting. We hold that with less noise and transformation, p_i could provide supervision signals to p_i' as it has a more accurate prediction result. Formally, we generate soft pseudo-label $\hat{p}_k(x_i)$ from model predictions:

$$\hat{p}_k(x_i) = \frac{\exp(h_k(x_i)/T)}{\sum_{k=0}^{K} \exp(h_k(x_i)/T)} \tag{1}$$

where subscript k denotes the k^{th} class output, and $T < 1$ is Softmax temperature. This operation can be regarded as a sharpening [1] to the original class probability output, which makes prediction closer to one-hot and provides more clear supervision. We argue that soft pseudo-labels could be more robust for noisy samples than one-hot labels which

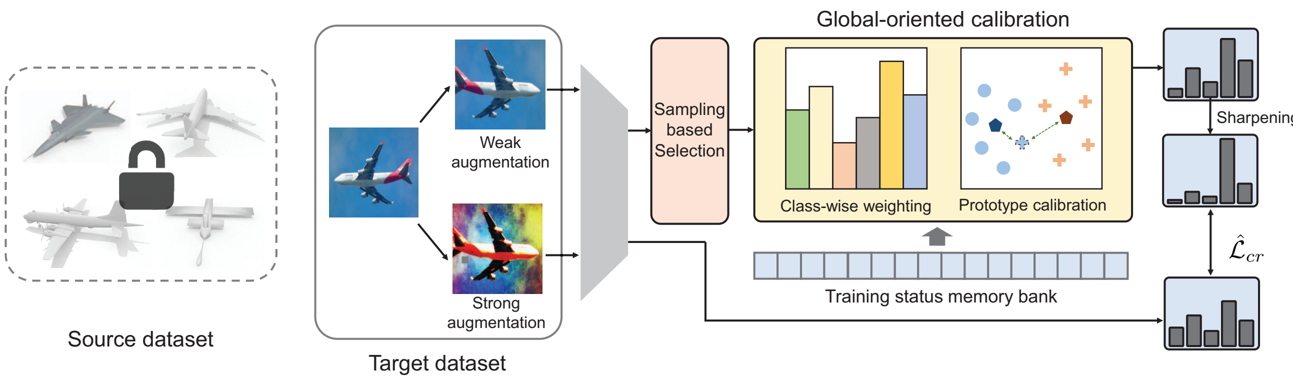


Figure 2. Framework of our proposed SFDA method. We leverage consistency regularization to boost model generalization ability, meanwhile, sampling-based pseudo-label selection strategy and global-oriented calibration module further strengthen the adaptation process.

may enlarge wrong learning signals, especially when there's no external constraint under the SFDA setting.

As mentioned above, aligning p'_i to $\hat{p}_i$ enhances the prediction consistency and provides supervision for domain adaptation. We apply cross-entropy loss here to accomplish alignment for sample x_i:

$$\mathcal{L}_{cr}^{(i)} = H(\hat{p}_i, p'_i) = -\sum_{k=1}^{K} \hat{p}_k(x_i) \log(p_k(\mathcal{A}(x_i))) \quad (2)$$

where $p_k(x_i)$ denotes the probability output of the k^{th} class. Besides construct consistency regularization, strong augmentation could also be regarded as an expansion of target data space, which is an effective approach to avoid overfitting issues and promotes the model's generalization ability. Notice that our implementation shares some similarities with FixMatch [39], however, 1) FixMatch applies consistency regularization to mine unlabeled samples aiming for setting up a new model, but we are trying to adapt the pretrained model to unlabeled data, 2) with labeled samples, Fixmatch generate vanilla one-hot pseudo-label, which is different from our approach, and 3) as discussed in Section 3.3, FixMatch needs labeled samples as anchors during the learning process, which is impractical for SFDA.

3.3. Sampling-based pseudo-label selection

One of the biggest challenges of SFDA is there's no explicit supervision signal when performing adaptation. It means error accumulation easily occur and causes degradation issue: if initially a sample was wrongly recognized as class k and continually involved in self-training, it will hardly be classified correctly thereafter, even if it's just close to class boundaries at the beginning due to domain shift. This fact implies that vanilla pseudo-labeling methods [22, 48, 55] who leverage all pseudo-labels are suboptimal. Meanwhile, some works [39, 56, 23, 17] introduce threshold to select "reliable" pseudo-labels. However, these methods need a carefully designed threshold, and too low takes no effect while too high omits useful information.

We provide insight into how to select pseudo-label and extract more potentially useful supervision by a sampling-based method. Consider an image input x_i and its corresponding prediction vector p_i, if p_i is close to one-hot, which means the model has high confidence to classify this sample, and the pseudo-label is more likely to be right. But at the same time, samples with relatively low confidence shouldn't be regarded wrong and abandoned, because they still carry some useful information. Only guided by "right" pseudo-labels, the model would tend to be biased and finally obtain sub-optimal results. We argue that conditionally adding relatively low-confidence samples to the self-training process could leverage more underlying distribution information and boost performance.

Formally, we sample pseudo-labels from the target dataset for training with a certain probability ξ related to its prediction confidence:

$$\xi_i = P(x_i \textit{ is selected}) = \mathcal{M}(\max_k p_k(x_i)) \quad (3)$$

where $k = 1, 2, \cdots, C$ and $\mathcal{M}$ denotes a mapping function with a range of $[0, 1]$. As $\mathcal{M}$ is a monotonically non-decreasing function, we can guarantee that high-confident samples are more considered during adaptation, and besides, low-confident samples aren't ignored. With the sampling strategy, hard samples caused by domain shift could also provide feature structure information for adaptation.

Implementing a probability sampling strategy often requires a relatively long training procedure and more samples to converge. But as an approximation, we can substitute sampling results with the expectation of each sample's consistency regularization loss, accelerating model convergence. Specifically, consistency regularization loss under sampling strategy with expectation approximation is:

$$\hat{\mathcal{L}}_{cr} = \mathbb{E}_{x_i \in D_t} \hat{\mathcal{L}}_{cr}^{(i)} = \frac{1}{n_t} \sum_{i=1}^{n_t} \xi_i \cdot H(\hat{p}_i, p'_i) \quad (4)$$

where ξ_i denotes the sampling probability in Eq. (3). Applying expectation instead of sampling results could helps the model perform better on small datasets because it introduces less random noise and makes the network less biased due to insufficient data points.

3.4. Global-oriented calibration

Since consistency regularization only leverages predictions of single input, global distribution information is under-considered, which is crucial to avoid degradation issues and form a compact feature representation.

Firstly, similar to previous works [27, 51, 53], SFDA paradigms easily suffer from degeneration problems where the model predicts all data to one or several specific classes, which greatly suppresses adaptation performance. Existing works [27, 51, 53] usually introduce an additional loss term, called diversity loss, to manually constrain class balance for tackling this local minimum issue, termed as $\mathcal{L}_{div} = \sum_k \mathrm{KL}(\overline{p}_k, \frac{1}{C})$ where $\overline{p}_k = \mathbb{E}_{x_i \in D_t}(p_k(x_i))$ is the mean probability output of class k. Although it shows some effect in avoiding trivial solution issues, it relies on a basic assumption that samples are uniformly distributed among different classes, which always fails on datasets [34, 43, 36, 16, 15] and real-world scenarios.

We present a class-wise weighting method to alleviate model degeneration without introducing any additional loss term, and can fit the class-imbalance situation better. Classes with fewer samples (minority) may encounter severer domain shifts, or just because they lack samples intrinsically. So instead of roughly constraining them to have the same amount of samples, we weigh their supervision signals with their current learning status. Formally, We denote the number of samples classified to class k as α_k:

$$\alpha_c = \sum_{i=1}^{n_t} \mathbb{1}(\max_k p_k(x_i) > \tau) \cdot \mathbb{1}(\arg\max_k p_k(x_i) = c) \tag{5}$$

where $\mathbb{1}$ is the indicator function and τ is the selection threshold. $[\alpha_c]$ shows the class-wise distribution in the current training stage. With it we can derive loss weight for samples from class c:

$$w_{div}(c) = \mathcal{T}\left(\frac{\alpha_c}{\max_c \alpha_c}\right) \tag{6}$$

where $\mathcal{T}$ is a monotonically decreasing mapping function. It's shown that minority class samples will obtain higher training weights and get promoted. Unlike previous works, our weighting method is adaptive for class imbalance: the majority of classes hold more samples but with low weight, producing a close scale of supervision signal to less but highly weighted samples. It can alleviate the negative impact of avoiding degeneration in previous works, which tends to compromise the accuracy of the majority classes.

Moreover, only considering single input is hard to recognize outliers, especially where domain shift exists. Clusters in feature space can provide distribution information inside each class and identify potentially wrongly classified samples [2, 57]. For all data instances classified to class c, we can compute the average feature representation of them, denoted as prototype η_c:

$$\eta_c = \frac{\sum_{x_i \in D_t} z_i \cdot \mathbb{1}(\arg\max p_i = c)}{\sum_{x_i \in D_t} \mathbb{1}(\arg\max p_i = c)} \tag{7}$$

If an image's feature is closer to η_A but classified as class B where $A \neq B$, we can indicate that it may be a latent outlier, and it's inadvisable to generate a supervision signal from it. Because once its initial prediction is slightly wrong, the error would accumulate during the self-training process and harm the final result. Based on this, we can leverage the inconsistency of feature prototype assignment and direct prediction, further modify the point-wise consistency regularization loss weight:

$$w_{proto}^{(i)} = \mathbb{1}(\arg\min_c \mathrm{dist}(z_i, \eta_c) = \arg\max p_i) \tag{8}$$

where $\mathrm{dist}(z_i, \eta_c)$ is the distance between sample feature and prototype. Eq. (8) can discard unreliable samples with disagreement of these two different semantics. With the prototype-based denoise adjustment, a more accurate set of samples is selected for the self-straining process. Note that this correction is performed in real-time, some initially abandoned samples may get rectified and be involved in adaptation, providing more supervision signals.

4. Experiments

Datasets We conduct experiments on three datasets: **Office-Home** [43] is an object recognition benchmark for domain adaptation, which consists of 12 adaptation tasks with four domains, sharing 65 classes and a total of around 15,500 images. **VisDA-2017** [34] is a synthetic-to-real object recognition dataset. It has 12 classes with two domains: the synthetic (source) domain contains 152K simulated images, and the real (target) domain consists of 55k real-world images. In addition, VisDA-2017 also provides 72k target domain test images, on which we evaluate model generalization ability. **DomainNet** [33] is a large-scale multi-source domain adaptation benchmark with 6 domains and 345 classes, which contains ~0.6 million images. Following [58, 37], we perform SFDA with selected four domains and 126 classes, and also evaluate testing set accuracy with their data split, shown in supplementary material.

Since consistency regularization with strong image augmentation adds noises to samples, it can be regarded as a more challenging learning setting and require more data to converge, so sample-insufficient datasets like Office-31 [36] are not chosen for evaluation.

Implementation details. To achieve fair comparisons with existing methods, we adopt ResNet-50 [18] for Office-Home, ResNet-101 for VisDA-2017, and ResNet-34 for

Table 1. Accuracy (%) on VisDA-2017 (ResNet101-based methods). SF means source-free. We highlight the best and second-best results.

Method	SF	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg.
MCC [20]	×	88.7	80.3	80.5	71.5	90.1	93.2	85.0	71.6	89.4	73.8	85.0	36.9	78.8
STAR [30]	×	95.0	84.0	84.6	73.0	91.6	91.8	85.9	78.4	94.4	84.7	87.0	42.2	82.7
RWOT [49]	×	95.1	80.3	83.7	90.0	92.4	68.0	92.5	82.2	87.9	78.4	90.4	68.2	84.0
SE [11]	×	95.9	87.4	85.2	58.6	96.2	95.7	90.6	80.0	94.8	90.8	88.4	47.9	84.3
SHOT [27]	✓	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
DIPE [44]	✓	95.2	87.6	78.8	55.9	93.9	95.0	84.1	81.7	92.1	88.9	85.4	58.0	83.1
A^2Net [46]	✓	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 [53]	✓	96.1	88.3	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
SFDA-DE [8]	✓	95.3	91.2	77.5	72.1	95.7	97.8	85.5	86.1	95.5	93.0	86.3	**61.6**	86.5
DaC [58]	✓	96.6	86.8	86.4	**78.4**	96.4	96.2	93.6	83.8	**96.8**	**95.1**	89.6	50.0	<u>87.3</u>
Ours	✓	**97.4**	**91.8**	**87.6**	78.1	96.6	**99.3**	**90.6**	**87.2**	95.6	94.6	88.9	57.3	**88.7**

DomainNet as the backbone. Consistent with SHOT [27] and other previous works [46, 51], we add two fully connected layers with weight normalization [38] after ResNet backbone. For model training, we adopt SGD with momentum 0.9, weight decay 1e-3, and learning rate scheduler $\gamma = \gamma_0 \cdot (1 + 10 \cdot p)^{-0.75}$ where p is the training progress changing from 0 to 1 and γ_0 is the initial learning rate. Our source model building-up process is identical to SHOT for a fair comparison. In the adaptation phase, we set $\gamma_0 =$ 4e-4, $\tau = 0.8$ and batch size of 64 for all experiments, while our strong image augmentation is from RandAugment [4]. For VisDA-2017, we set sharpen temperature T to 0.5 and epoch to 30, which are 0.1 and 60 for Office-Home and DomainNet due to the different number of classes. Considering simplification, we plainly set the sampling probability mapping function as $\mathcal{M}(x) = x$ and the class-wise weighting mapping function as $\mathcal{T}(x) = 1 - \log(x)$. Noting that $x > 0$ should be guaranteed for $\mathcal{T}$, we replace $\alpha_c = 0$ with the smallest non-zero $\alpha_{c'}$ in Eq. (6). Our final adaptation loss is formulated as $\mathcal{L} = \mathbb{E}_{x_i \in D_t}\left[w_{div} w_{proto} \hat{\mathcal{L}}_{cr}\right]$. For maintaining memory bank, we simply follow the common practice [51, 53] by saving and updating predictions and features. All experiments are run on NVIDIA 3090 GPU.

4.1. Results

We evaluate our proposed method and show comparison results with existing UDA and SFDA methods. Tables 1, 3 and 4 provide classification accuracy after adaptation on VisDA-2017, DomainNet and Office-Home datasets respectively. As shown in the tables, our method achieves state-of-the-art performance among existing methods, even including UDA paradigms which under the easier setting. Moreover, on large-scale datasets like VisDA-2017 and DomainNet, our results significantly surpass the second-best one and achieve the best class-wise accuracy in most classes, indicating a general improvement for SFDA task.

In Table 2, we evaluate the generalization ability of our adaptive model. We reproduce previous works with open-source code and test them on VisDA-2017 target domain testing data mentioned above, and evaluate the performance drop between the target training and testing data. Results demonstrate that our consistency regularization-based paradigm shows high accuracy both in training and testing data, avoiding a giant performance degradation on testing data like what appeared in previous works. It proves that we successfully construct a smooth manifold and avoid overfitting issues, which is essential for model generalization and real-world SFDA applications. Note that UDA methods compared in Table 2 are from VisDA2017 Classification Challenge [34] with no constraint for model backbone, so unfair comparisons may exist, like challenge winner team "GF_ColourLab_UEA" [11] (not included in the table) choosing ResNet-152 as a backbone.

Table 2. Accuracy (%) and its drop between target training and testing data on VisDA-2017 dataset. † means these results from *VisDA2017 Classification Challenge* [34] leaderboard, which are under **easier** vanilla UDA setting and may use **stronger** backbone causing **unfair** comparisons.

Method	SF	Target	Test	Drop
BUPT†	×	-	85.4	-
IISC_SML†	×	-	86.4	-
NLE†	×	-	<u>87.7</u>	-
SHOT [27]	✓	82.9	73.2	9.7(11.7%↓)
G-SFDA [53]	✓	85.4	81.8	3.6(4.2%↓)
NRC [51]	✓	85.9	82.0	3.9(4.5%↓)
DaC [58]	✓	87.3	84.2	3.1(3.6%↓)
Ours	✓	**88.7**	**87.8**	**0.9(1.0%↓)**

For the Office-Home dataset, as shown in Table 4, we achieve comparable results to previous state-of-the-art works. Consistent with what we discussed above, the use of strong image augmentations produces a harder optimization process, which calls for more training data to converge

Table 3. Accuracy (%) on DomainNet (ResNet34-based methods). SF means source-free. We highlight the best and second-best results.

Method	SF	Rw→Cl	Rw→Pt	Pt→Cl	Cl→Sk	Sk→Pt	Rw→Sk	Pt→Rw	Avg.
MME [37]	×	70.0	67.7	69.0	56.3	64.8	61.0	76.1	66.4
CDAN [29]	×	65.0	64.9	63.7	53.1	63.4	54.5	73.2	62.5
VDA [20]	×	63.5	65.7	62.6	52.7	53.6	62.0	74.9	62.1
GVB [5]	×	68.2	69.0	63.2	56.6	63.1	62.2	73.8	65.2
BAIT [52]	✓	64.7	65.4	62.1	57.1	61.8	56.7	73.2	63.0
SHOT [27]	✓	67.1	65.1	67.2	60.4	63.0	56.3	76.4	65.1
G-SFDA [53]	✓	63.4	67.5	62.5	55.3	60.8	58.3	75.2	63.3
NRC [51]	✓	67.5	68.0	67.8	57.6	59.3	58.7	74.3	64.7
DaC [58]	✓	70.0	68.8	70.9	62.4	**66.8**	60.3	**78.6**	68.3
Ours	✓	**72.2**	**69.4**	**72.1**	**63.2**	66.7	**63.4**	77.3	**69.2**

Table 4. Accuracy (%) on Office-Home (ResNet50-based methods). SF means source-free. We highlight the best and second-best results.

Method	SF	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.
GVB-GD [5]	×	57.0	74.7	79.8	64.6	74.1	74.6	65.2	55.1	81.0	74.6	59.7	84.3	70.4
RSDA [14]	×	53.2	77.7	81.3	66.4	74.0	76.5	67.9	53.0	82.0	75.8	57.8	85.4	70.9
TSA [26]	×	57.6	75.8	80.7	64.3	76.3	75.1	66.7	55.7	81.2	75.7	61.9	83.8	71.2
SRDC [40]	×	52.3	76.3	81.0	69.5	76.2	78.0	68.7	53.8	81.7	76.3	57.1	85.0	71.3
FixBi [32]	×	58.1	77.3	80.4	67.7	79.5	78.1	65.8	57.9	81.7	76.4	62.9	86.7	72.7
SHOT [27]	✓	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
G-SFDA [53]	✓	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
DIPE [44]	✓	56.5	79.2	80.7	**70.1**	**79.8**	78.8	67.9	55.1	**83.5**	**74.1**	59.3	84.8	72.5
A^{2}Net [46]	✓	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
SFDA-DE [8]	✓	**59.7**	79.5	82.4	69.7	78.6	**79.2**	66.1	**57.2**	82.6	73.9	60.8	85.5	**72.9**
DaC [58]	✓	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
Ours	✓	58.6	80.2	**82.9**	69.8	78.6	79.0	67.8	55.7	82.3	73.6	60.1	84.9	72.8

and boost model performance. With only 60 images per class on average, our method doesn't provide such significant improvement as that on large-scale datasets.

4.2. Analysis

Ablation study. We conduct ablation studies of our proposed modules in Table 5. Three class-wise accuracy shown here is carefully selected: class "car" owns the largest amount of samples while class "knife" and "truck" suffer from a more serious domain shift. Results in the first row are source-only model accuracy without adaptation. Row 2 and 3 show that a sampling-based pseudo-label selection strategy can significantly boost the consistency regularization learning paradigm because it identify noisy pseudo-labels and leverages potentially useful information. However, the accuracy of class "truck" decreases to 0% due to the degradation issues mentioned above. Class balancing constraint by class-wise weighting module solves this problem properly, making sure that classes with severe domain shifts are correctly recognized. Lastly, the prototype calibration module can further enhance the adaptation effect by recognizing and filtering outliers.

Consistency regularization for boosting generalization ability. Fig. 3 demonstrates the effect of our proposed consistency regularization on boosting model generalization

Table 5. Ablation study results of our proposed modules on VisDA-2017 dataset. Three typical class-wise and all classes' average accuracies are shown. (C.R.=Consistency Regularization, S.S.=Sampling-based pseudo-label Selection, C.W.=Class-wise Weighting, P.C.=Prototype Calibration)

C.R.	S.S.	C.W.	P.C.	car	knife	truck	Avg.
				75.7	4.4	7.9	46.6
✓				62.9	54.5	17.8	68.0
✓	✓			89.4	97.6	0	84.9
✓	✓		✓	90.0	98.2	0	85.5
✓	✓	✓		65.3	98.9	60.7	88.0
✓	✓	✓	✓	78.1	99.3	57.3	**88.7**

ability. As shown in Fig. 3a, removing the consistency regularization module and applying vanilla self-training, which has no strong data augmentation, will remarkably reduce the model's final accuracy, especially on the target testing set. This is because the self-training approach fails to construct a smooth manifold on target domain data representations. According to Fig. 3b, soft pseudo-labels are also essential to model generalization ability, because hard one-hot pseudo-labels may enlarge noise signals and cause error accumulation since there's no external supervision signal.

Sampling-based pseudo-label selection. We implement a

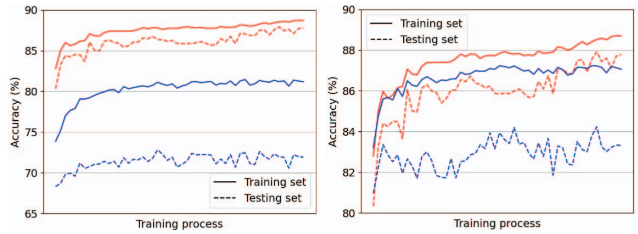
Table 6. Comparisons between our class-wise weighting method and vanilla class balancing loss, where β is the loss weight. Best results inside each method are in bold font. Our approach achieves higher accuracy and shows less sensitivity to hyper-parameter.

Method	VisDA	Ar→Cl	Ar→Rw
$\mathcal{L}_{div}$ ($\beta = 1$)	78.6	56.4	80.0
$\mathcal{L}_{div}$ ($\beta = 0.5$)	82.9	**57.7**	**81.8**
$\mathcal{L}_{div}$ ($\beta = 0.1$)	87.8	54.8	80.2
$\mathcal{L}_{div}$ ($\beta = 0.05$)	**88.1**	54.3	79.5
$\mathcal{L}_{div}$ ($\beta = 0.01$)	88.0	54.1	78.5
Ours ($\tau = 0.9$)	88.6	57.7	82.8
Ours ($\tau = 0.8$)	**88.7**	**58.6**	82.9
Ours ($\tau = 0.6$)	88.6	58.3	**83.0**
Ours ($\tau = 0.3$)	88.6	57.7	82.8
Ours ($\tau = 0.1$)	88.6	57.3	82.8

vanilla thresholding method to select pseudo-label with various thresholds and compare them with our non-parametric sampling-based approach in Fig. 4. While it's hard for thresholding to achieve high accuracy on all datasets at the same time, our proposed sampling-based method is free from adjusting any hyper-parameter and performs better for its consideration of more potentially useful samples. As shown in Fig. 4b, because limited samples are provided on the Office-Home dataset, our expectation approximation strategy further boosts the sampling-based selection effect. Note that on the VisDA-2017 dataset, the vanilla sampling way achieves the same results with expectation approximation due to efficient training samples.

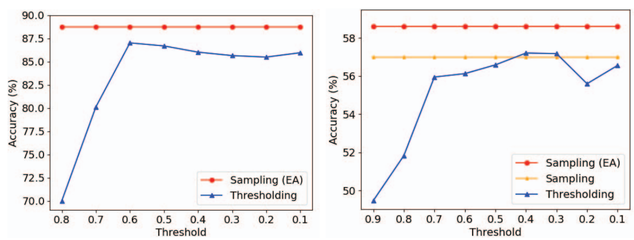
Class-wise weighting for class balancing. As what we discussed about Table 5, class-wise weighting can effectively balance class distribution and avoid trivial solution problems, and we further investigate the difference to widely used class balancing loss [27, 51, 53] $\mathcal{L}_{div}$ (see Section 3.4), shown in Table 6. Different from the high sensitivity to loss weight β settings of $\mathcal{L}_{div}$, our method is robust to different hyper-parameter settings and outperforms $\mathcal{L}_{div}$-based methods without introducing any additional loss. Meanwhile, our method could achieve the best results while τ keeps consistent across tasks, where the $\mathcal{L}_{div}$ approach needs to fine-tune their hyper-parameters in a large range.

Prototype calibration denoise effect. Fig. 5 demonstrates the effectiveness of prototypes in pseudo-label denoising. Fig. 5a shows that with prototype calibration, pseudo-labels could achieve relatively high accuracy at the beginning stage of adaptation. As what we discussed in Section 3.4, prototype calibration won't abandon potentially useful information due to its real-time property, and Fig. 5b shows how the ratio of selected pseudo-labels varies along the training stage. It illustrates that during the adaptation process, an increasing number of pseudo-labels are selected, which means sample predictions are getting more consistent with cluster relation results in the feature space.



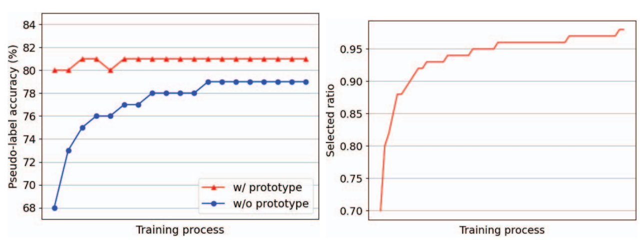
(a) (b)

Figure 3. Accuracy on target training (real line) and testing set (dashed line). Results of our proposed consistency regularization (CR) with soft pseudo-label are in Red, while blue lines indicate: (a) applying vanilla self-training instead of CR; (b) realizing CR with hard one-hot pseudo-labels.



(a) (b)

Figure 4. Comparisons between our sampling-based pseudo-label selection and thresholding way with different thresholds on (a) VisDA-2017 (b) Ar→Cl task of Office-Home. EA denotes the expectation approximation in Eq. (4).



(a) (b)

Figure 5. (a) Pseudo-label accuracy with and without prototype calibration on Ar→Pr. (b) Ratio of selected pseudo-labels along the adaptation process on Ar→Cl.

5. Conclusions

In this paper, based on the observation that previous SFDA works suffer from handling unseen but identically distributed data, we introduce a consistency regularization framework to achieve SFDA and boost model's generalization ability for the first time. By leveraging data augmentation and soft pseudo-label, our method can significantly improve the model performance both on target training and testing sets. In addition, sampling-based pseudo-label selection strategy and global-oriented calibration modules are proposed to further enhance the adaptation effect. Experiments demonstrate the effectiveness of our method.

Acknowledgment. This work was supported by Shenzhen Key Laboratory of next generation interactive media innovative technology (No.ZDSYS20210623092001004).

References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 3

[2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 5

[3] Lawrence Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005. 2

[4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 6

[5] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12455–12464, 2020. 7

[6] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[8] Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7212–7222, 2022. 2, 6, 7

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[10] Cian Eastwood, Ian Mason, Chris Williams, and Bernhard Schölkopf. Source-free adaptation to measurement shift via bottom-up feature restoration. In *International Conference on Learning Representations*, 2021. 3

[11] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017. 6

[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] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006. 2

[14] Xiang Gu, Jian Sun, and Zongben Xu. Spherical space domain adaptation with robust pseudo-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9101–9110, 2020. 7

[15] Chunming He, Kai Li, Guoxia Xu, Jiangpeng Yan, Longxiang Tang, Yulun Zhang, Xiu Li, and Yaowei Wang. Hqg-net: Unpaired medical image enhancement with high-quality guidance. *arXiv preprint arXiv:2307.07829*, 2023. 5

[16] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22046–22055, 2023. 5

[17] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *arXiv preprint arXiv:2305.11003*, 2023. 4

[18] 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*, pages 770–778, 2016. 1, 5

[19] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8988–8999, 2021. 2

[20] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *European Conference on Computer Vision*, pages 464–480. Springer, 2020. 6, 7

[21] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 3

[22] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 4

[23] Kai Li, Chang Liu, Handong Zhao, Yulun Zhang, and Yun Fu. Ecacl: A holistic framework for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8578–8587, 2021. 4

[24] Kai Li, Deep Patel, Erik Kruus, and Martin Renqiang Min. Source-free video domain adaptation with spatial-temporal-historical consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14643–14652, 2023. 2

[25] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. Adversarial feature hallucination networks for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13470–13479, 2020. 1

[26] Shuang Li, Mixue Xie, Kaixiong Gong, Chi Harold Liu, Yulin Wang, and Wei Li. Transferable semantic augmentation for domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11516–11525, 2021. 7

[27] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 1, 2, 5, 6, 7, 8

[28] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 2

[29] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *NeurIPS*, 2018. 7

[30] Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9111–9120, 2020. 6

[31] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 3

[32] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1094–1103, 2021. 2, 7

[33] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 5

[34] 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. 1, 5, 6

[35] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):801–814, 2018. 1, 2

[36] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010. 3, 5

[37] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8050–8058, 2019. 5, 7

[38] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 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. *Advances in neural information processing systems*, 33:596–608, 2020. 3, 4

[40] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8725–8735, 2020. 1, 2, 7

[41] Longxiang Tang, Kai Li, Chunming He, Yulun Zhang, and Xiu Li. Source-free domain adaptive fundus image segmentation with class-balanced mean teacher. *arXiv preprint arXiv:2307.09973*, 2023. 2

[42] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3

[43] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 5

[44] Fan Wang, Zhongyi Han, Yongshun Gong, and Yilong Yin. Exploring domain-invariant parameters for source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2022. 6, 7

[45] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 1, 2

[46] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9010–9019, 2021. 2, 6, 7

[47] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020. 3

[48] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 4

[49] Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4394–4403, 2020. 6

[50] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021. 1, 2

[51] Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 34:29393–29405, 2021. 1, 2, 3, 5, 6, 7, 8

[52] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Casting a bait for offline and online source-free domain adaptation. *arXiv preprint arXiv:2010.12427*, 2020. 7

[53] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8978–8987, 2021. 1, 2, 3, 5, 6, 7, 8

[54] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *arXiv preprint arXiv:2103.00550*, 2021. 3

[55] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019. 4

[56] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 4

[57] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021. 5

[58] 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 *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3, 5, 6, 7

[59] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 2