2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 978-1-6654-9346-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/WACV56688.2023.00416

CoNMix for Source-free Single and Multi-target Domain Adaptation

Vikash Kumar* Rohit Lal* Himanshu Patil Anirban Chakraborty
Indian Institute of Science, Bengaluru, India
{vikashks,anirban}@iisc.ac.in, {take2rohit, hipatil1998}@gmail.com

Abstract

*This work introduces the novel task of **S**ource-**f**ree **M**ulti-**t**arget **D**omain **A**daptation and proposes adaptation framework comprising of **Co**nsistency with **N**uclear-Norm Maximization and **Mix**Up knowledge distillation (CoNMix) as a solution to this problem. The main motive of this work is to solve for Single and Multi target Domain Adaptation (SMTDA) for the source-free paradigm, which enforces a constraint where the labeled source data is not available during target adaptation due to various privacy-related restrictions on data sharing. The source-free approach leverages target pseudo labels, which can be noisy, to improve the target adaptation. We introduce consistency between label preserving augmentations and utilize pseudo label refinement methods to reduce noisy pseudo labels. Further, we propose novel MixUp Knowledge Distillation (MKD) for better generalization on multiple target domains using various source-free STDA models. We also show that the Vision Transformer (VT) backbone gives better feature representation with improved domain transferability and class discriminability. Our proposed framework achieves the state-of-the-art (SOTA) results in various paradigms of source-free STDA and MTDA settings on popular domain adaptation datasets like Office-Home, Office-Caltech, and DomainNet. Project Page: https://sites.google.com/view/conmix-vcl*

1. Introduction

The advent of Deep Learning has brought significant development in tasks like image classification, object detection, semantic segmentation, etc. However, the performance of the state-of-the-art methods trained with millions of labeled images suffers significantly in the environment where there is a mismatch between training and test distribution [39, 47], motivating researchers to design learning algorithms that are robust to shifts in data distribution. One such popular research direction is Unsupervised Domain Adaptation (UDA) for a labeled source domain to a unlabeled target domain adaptation. UDA with only one source and one target domain is termed Single Target Domain Adaptation (STDA) [52]. Multi-target Domain Adaptation (MTDA) consists of multiple unlabeled target domains against a single labeled source. STDA can be thought of as a special case of MTDA and is critical in solving a practical task such as adaptation from Synthetic data distribution to Real-World data distribution. In contrast, the MTDA framework is essential when we have multiple target domains with varying domain-shift.

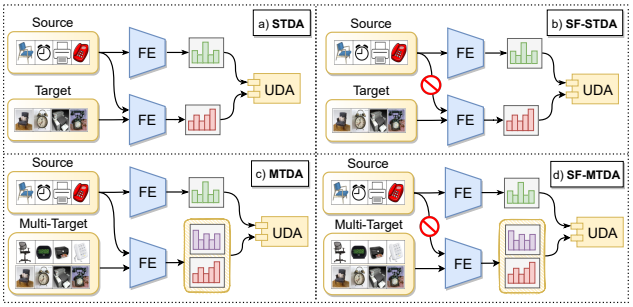

Figure 1: **1.(a)** represents vanilla STDA approach where along with unlabeled target data, labeled source data and source trained model are available during adaptation stage. **1.(b)** represents the adaptation strategy when only source trained model is available, but labeled source data is not available during adaptation stage. **1.(c)** is an extension of Fig1.(a), which shows an approach for UDA for the single source (always available) to multi-target domain adaptation (MTDA). **1.(d) (Ours)** is an extension to MTDA but without the access of labeled source data. Fig 1.(a), 1.(b), 1.(c) are already widely studied but 1.(d) remains unexplored.

Most of the existing Domain Adaptation (DA) methods assume availability of labeled source domain samples during adaptation which may not be possible for several use cases that mandate data privacy, such as biometrics, healthcare etc. Also, there can be situations where storing a large dataset is not feasible, for instance, training and deploying domain adaptation applications on an embedded system or on edge devices with limited memory. However, we can store source trained models because they are relatively

*Equal Contribution

2642-9381/23/$31.00 ©2023 IEEE
DOI 10.1109/WACV56688.2023.00416

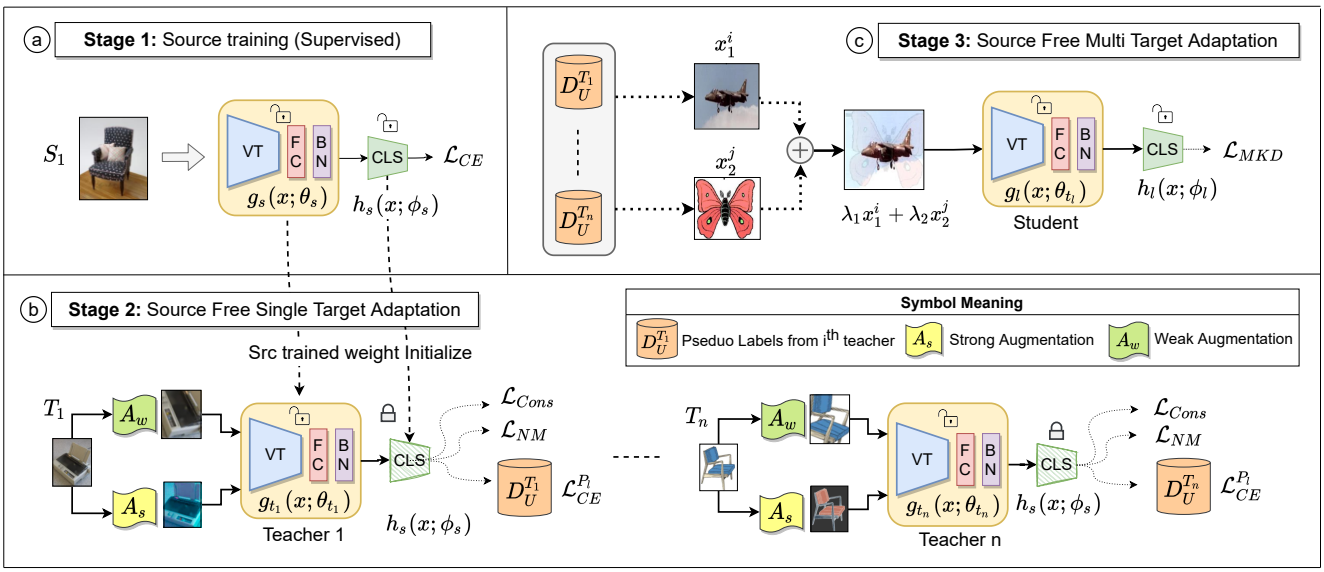

Figure 2: **Architecture of *CoNMix***. In Stage-1, we prepare a source trained model and get rid of source data. In Stage-2, we use previously source trained model and adapt to multiple target domains without having access to source. We store pseudo labels for all the target domains with the help of single domain adapted models (teachers) obtained in Stage-2. Finally, we perform knowledge distillation to a common student in Stage-3.

smaller in size compared to the dataset they are trained with. Therefore, traditional domain adaptation methods are not suitable in these situations and hence, we should look for *source-free* methods. Source-free UDA methods [20, 25, 27, 22, 54] aims to solve the adaptation problem without access to labeled or unlabeled source data during adaptation to a target domain. We show the high-level distinction of our proposed approach with respect to existing unsupervised domain adaptation tasks in Fig. 1.

Source-free STDA works well for specified target domain; however, it may fail for many practical use-cases where the test data can come from multiple target domains or some unseen/open domains which were not present during adaptation. One trivial extension would be to train specific source-free STDA models for each target domain. In that case, we need to have the domain information for selecting the appropriate model during inference. Still with this constraint, it may not generalize well for open domains. One such practical application is scene classification in autonomous driving built by taking data from sunny weather that should generalise well across different weather conditions such as winter, rainy, foggy, etc where target labels are not available. To alleviate the above problems, we propose a novel source-free MTDA setting. In SF-MTDA setting, there would be a single trained model that can generalize well across different domains. This would reduce storage costs and have a shorter inference time than having a specific model for each target domain. SF-MTDA poses the additional challenge of bridging the gap between multiple target domains by learning common representation. Existing state-of-the-art MTDA methods [41, 33] use a complex adversarial training strategy that needs access to both source and target datasets for learning domain invariant representation making them unsuitable for source-free tasks. We propose to leverage pseudo-label refinement [59] along with novel consistency constraint for mitigating the uncertainty associated with the pseudo-labels. To address SF-MTDA setting, we propose MixUp based Knowledge Distillation (MKD) to distill knowledge from multiple expert teachers (STDA models) to a single student. The overall novel framework for solving source-free domain adaptation tasks is dubbed as *CoNMix*. Further, we also investigate the role of different backbones on source-free adaptation tasks.

Almost all existing UDA methods use CNN based feature extractors [12, 4, 33, 28, 41] whose design includes strong human inductive bias such as local connectivity and pooling. Unlike CNN, VT has a global receptive field at every stage. Therefore, the learned representations are more meaningful for the downstream tasks. Self-attention in Vision Transformer is designed to assign more importance to salient objects of interest and lesser importance to less relevant information such as the background information. Therefore, it can mitigate the spurious correlation between prediction probability and domain dependent components such as lighting condition thereby making the feature representation more transferable, which is desirable in domain adaptation [54]. In fact, in our experiments too, we observe that the feature representation of VT has better domain-

transferability *(easy to transfer across different domains)* and class-discriminability *(ability to distinguish between classes)* compared to CNN based architecture (e.g.,ResNet). In summary, our main contributions are as follows:

- We propose a novel task of source-free multi-target domain adaptation and developed *CoNMix*, a novel framework for solving source-free single and multi-target domain adaptation tasks. We also provide empirical insights, backed by quantitative and qualitative results to substantiate the use of VT backbone for SF-SMTDA.
- We introduce a novel augmentation based consistency constraint and explore existing nuclear-norm maximization in our learning objective and pseudo label refinement strategy to mitigate the effect of noisy pseudo labels. Further, we judiciously combine these with Mixup knowledge distillation to propose the overall framework of CoNMix.
- We are among the first to extensively study this important SF-MTDA problem. We have advanced the SOTA for SF-STDA and SF-MTDA settings on popular benchmark datasets. We also provide a new baseline on the large-scale DomainNet dataset for source-free single and multi-target domain adaptation.

The insights we draw from our analysis constitute important contribution of this paper. Compared to previous methods, CoNMix also has various appealing aspects- (a) Safe: CoNMix is developed to maintain complete data privacy, as it keeps the data safe and avoids any leakages (b) Flexible: A single algorithm can be extended for both source-free single and multi target domain adaptation (SF-SMTDA) tasks.

2. Related Work

Single Target Domain Adaptation (STDA): One of the popular methods for STDA is to try learning domain invariant features by minimising domain discrepancy [57, 30, 46, 31, 19]. Methods such as [3, 49, 29, 10] leverage adversarial training for UDA. Generative modelling methods like [2, 17] try to minimise the gap between source and target images by transforming one feature space to another. Though these methods have been proven to be very effective for STDA, their dependency on source data during adaptation makes it undesirable for source-free approaches.

Source-free Domain Adaptation: Recently source-free methods [1, 24, 25, 20, 54] are getting a lot of attention for UDA tasks. In this setting, we only have access to source trained model and unlabeled target data. Liang *et al*. [25] uses information maximization and pseudo labeling to align the target domain to the source domain. 3C-GAN[24] improves the prediction through generated target-style data. Noisy pseudo label is one of the major problems in source-free adaptation tasks.

Multi Target Domain Adaptation: For MTDA, we need to generalise for multiple unlabeled target data distribution with the help of single labeled source data distribution [12, 4, 28, 35].Nguyen *et al*. [33] proposes training multiple adaptation networks and simultaneously distil knowledge from adapted models to small student network. The source-free MTDA is an important research direction, which has not been explored extensively yet to the best of our knowledge. Our proposed framework *CoNMix* attempt to address SF-STDA and SF-MTDA problems.

Vision Transformer (VT): Transformer achieved a lot of success in natural language processing since it was first introduced by Vaswani *et al*. [50]. Dosovitskiy *et al*. [8] represented image patch with position encoding as a sequence dataset and reported improved performance on ImageNet. Touvron *et al*. [48] uses smaller dataset for training compared to [8] utilising distillation token for learning the inductive bias. Kurmi *et al*. [21] proposes to get the weighted feature representation by multiplying backbone output with the attention map generated through the Bayesian discriminator. Due to the absence of source sample during adaptation, we can not use domain discriminator based architecture, therefore these methods can not be extended for source-free tasks. Yang *et al*. [54] uses the bigger variant of Vision Transformer (ViT-B) [8] along with student-teacher architecture for solving SF-STDA problem. ViT-B is overparameterized with 86M parameters, whereas existing methods use ResNet50 which has only 24M parameters.

3. Problem Setting and Proposed Approach

In this section we define the problem setting and our proposed approach towards solving this problem. We are trying to solve source-free single and multi-target domain adaptation, which involves solving adaptation task on single and multiple unlabeled target domain using source trained model without accessing the source dataset during adaptation. We introduce *CoNMix* (Fig. 2), a three-stage approach that utilizes Vision Transformer along with consistency constraint, nuclear norm maximization, pseudo label refinement and MixUp based knowledge distillation (MKD), designed at solving source-free single and multi-target domain adaptation problem.

Notation: We denote h as hypothesis or classifier. $\xi_S(h)$ and $\xi_T(h)$ are expected risk/error of hypothesis h for source domain and target domain respectively. $\mathcal{L}_{CE}$, $\mathcal{L}_{NM}$, $\mathcal{L}_{Cons}$, $\mathcal{L}_{CE}^{P_l}$, $\mathcal{L}_{MKD}$ represents cross-entropy loss, Nuclear-norm Maximization loss, Consistency loss, Pseudo label Cross-Entropy loss, and MixUp Knowledge Distillation loss respectively. $\mathcal{A}_W$ and $\mathcal{A}_S$ are weak and strong augmentation applied to input sample. We use $\mathcal{X}_\mathcal{S}$, $\mathcal{Y}_\mathcal{S}$, $\mathcal{X}_\mathcal{T}$ for representing source image, source label and target image respectively. $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$ are source and target distribution. $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T})$ is divergence between source and target domain distribution and $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{X}_\mathcal{S}, \mathcal{X}_\mathcal{T})$ is its empirical measure.

3.1. Backbone Selection

In this section, we demonstrate that the attention based backbone [8] provide tighter upper bound on $\xi_T(h)$ compared to popular ResNet [14] backbone. We provide empirical insight to show that $\xi_T^{VT}(h) < \xi_T^{RN50}(h)$, making Vision Transformer (VT) a more suitable candidate for solving domain adaptation tasks. Additional information and comparison pertaining to the choice of backbone in our proposed approach are provided in the following subsections.

3.1.1 Comparison of Backbone

A majority of current SOTA UDA techniques extract image features using a CNN based backbone, such as Resnet50. Given the recent success of VT [8, 48], we attempt to analyse the feature representation of VT based backbone for domain adaptation. We aim to show difference in learned representations using RN50 and DeiT backbones. In our experiments, we find that *VT features are more domain-transferable and class-discriminative compared to ResNet*. We corroborate above by explicitly measuring the $\mathcal{A}$-distance in Fig. 4 which is a popular way to measure the feature alignment in adversarial learning [10]. We found that $\mathcal{A}$-distance of VT feature representation is smaller compared to CNN based representation. This difference in $\mathcal{A}$-distance shown in Fig. 4 is significant and provides a direct evidence to why DeiT backbone leads to substatially better performance in UDA. Additionaly, we examined the t-SNE plot of the two representations and discovered that VT based representation are relatively better aligned (Fig. 3). We believe that the properties of VT (DeiT-S) such as having the global receptive field at every stage and self-attention help them learn more class-discriminative and domain-transferable feature representation than CNN (ResNet).

We also observe that *our suggested loss functions are better suited for VT than ResNet*. In supp. material, we perform an experiment to analyse the effect of loss functions on two backbones. We discovered that VT backbone results in significantly increased performance compared to its CNN counterpart. Based on this analysis, we can conclude that VT serves as a better feature extractor alternative than CNN for domain adaptation tasks. We also compare the effect of various VT models and better ImageNet models like EfficientNetV2-B3 and EfficientNetV2-S in suppl. material to further validate our analysis.

3.2. Source-Free Domain Adaptation

The *CoNMix* architecture attempt to solve two problems. Firstly, It can solve source-free Single-Target Domain Adaptation (SF-STDA) by utilizing Stage-1 and Stage-2 of the Architecture (Fig. 2). Secondly, by introducing Stage-3, *CoNMix* can solve source-free Multi-Target Domain Adaptation (SF-MTDA) effectively.

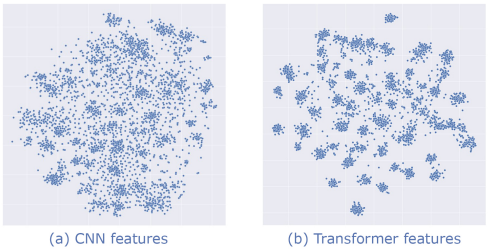

Figure 3: **t-SNE plots** (a) and (b) are t-SNE plots using features of Cl images obtained by passing through $Rw \rightarrow Cl$ adapted model for ResNet50 and VT respectively.

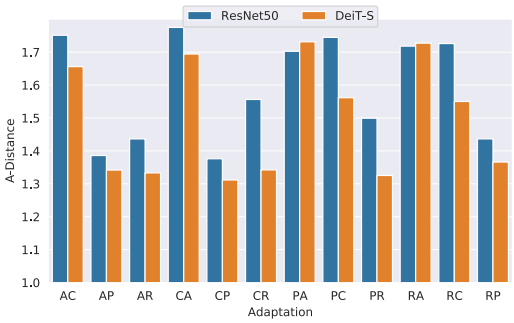

Figure 4: Plot compares $\mathcal{A}$-distance of VT and RN50 backbone for various Office-home splits (e.g., AC represents adaptation from Art to Clipart). Small **$\mathcal{A}$-distance** results in better feature alignment. VT features consistently shows smaller $\mathcal{A} - distance$.

3.2.1 Source Training

We aim to learn a model $f_s(\theta_s, \phi_s) : \mathcal{X}_\mathcal{S} \rightarrow \mathcal{Y}_\mathcal{S}$ where θ_s, ϕ_s are parameter for backbone network and classifier network respectively. Differing from traditional approaches, additional fully connected layer and batch norm layer was used after the backbone for better alignment of the projected features. We use VT backbone due to it's discussed benefits over ResNet. We sample $(\mathcal{X}_\mathcal{S}, \mathcal{Y}_\mathcal{S}) \sim \mathcal{D}_\mathcal{S}$ and train the complete network using cross-entropy $\mathcal{L}_{CE}$ loss as shown in Eq.1.

$$\mathcal{L}_{CE} = -\mathbb{E}_{(x_s,y_s)} \sum_{k=1}^{K} q_k \log \delta_k(f_s(x_s : \theta_s, \phi_s)) \quad (1)$$

where $\delta_k(a) = \frac{\exp(a_k)}{\sum_i \exp(a_i)}$ denotes the k-th element in the soft-max output of a K-dimensional vector a, and q is the one-of-K encoding of y_s where q_k is '1' for the correct class and '0' for the rest. In practice, we have used label smoothing [32] in place of one-hot encoding to avoid the network being overconfident. It helps to improve model generalization ability. q_k will be replaced with $q_k^{ls} = (1-\alpha)q_k + \alpha/K$

where q_k^{ls} is the smoothed label and α is the smoothing parameter set to 0.1. Refer Supp. material for more discussion on effect of smooth labels on adaptation. Post source training, we freeze the classifier network. Please note that the *labeled source sample will not be available for the next target adaptation stage* because we are trying to solve the source-free problem.

3.2.2 Source-free Single-target Domain Adaptation

We have access to sampled unlabeled target data $(\mathcal{X}_\mathcal{T}) \sim \mathcal{D}_\mathcal{T}$ and source trained model $f_s(\theta_s, \phi_s)$. Stage-2 aims to train T independent single source to single target domain adaptation networks while having no access to source data where T is a total number of target domains. Following [25], we propose to freeze the classifier parameter ϕ_s and update the backbone parameter θ_t which was initialized from source backbone parameter θ_s. We use $\mathcal{L}_{CE}^{P_l}$, $\mathcal{L}_{NM}$ and $\mathcal{L}_{Cons}$ for updating the backbone weights using back-propagation with SGD [23].

Nuclear-norm Maximization (NM): Many label insufficient situations such as Semi-supervised learning [60] or Unsupervised learning [13] suffers from higher data density near decision boundary, which results in poor class-discriminability. Directly minimizing the Shannon entropy [44] leads to uniformly smooth representation which improves discriminability by pushing samples to one of the class labels, however, it does not ensure diversity and may result in undesirable solution where all the minority class is pushed to the nearest majority class. Different variants such as Information maximization (IM) loss [16] address this issue with limited success. Nuclear-norm maximization (NM) [5] uses batch-statistics to achieve function-smoothing only in the required dimensions and to the required extent leading to superior representation therefore, it improves both class-discriminability as well as prediction diversity in a unified way making it desirable for SF-SMTDA tasks.

Class-discriminability in NM: We define $A \in \mathbb{R}^{B\times K}$ to be the classification-response matrix A, where B is the batch size, and K is the number of classes. Frobenius norm $\|A\|_F$ is defined in Eq.2.

$$||A||_F = \sqrt{\sum_{i=1}^{B}\sum_{j=1}^{K}|A_{ij}|^2} \tag{2}$$

where $0 \leq A_{ij} \leq 1$ and $\sum_{j=1}^{K}(A_{ij}) = 1$. We can obtain the upper bound of $||A||_F$ as shown in Eq.3

$$||A||_F \leq \sqrt{\sum_{i=1}^{B}(\sum_{j=1}^{K}A_{ij})(\sum_{j=1}^{K}A_{ij})} = \sqrt{\sum_{i=1}^{B}(1.1)} = \sqrt{B} \tag{3}$$

Upper bound in $||A||_F$ corresponds to the one-hot prediction for each sample in a batch. Therefore, maximizing $||A||_F$ leads to improved class-discriminability. Cui *et al*. in [7] proved that maximum value of $\|A\|_F$ comes where entropy achieves its minimum value.

Prediction-diversity in NM: If we define $\|A\|_*$ as nuclear norm and r as the rank of A then Recht *et al*. [38] provide relation between $||A||_F$ and $||A||_*$, which we show in Eq. 4. We provide proof in supp. material.

$$\|A\|_F \leq \|A\|_* \leq \sqrt{r}\|A\|_F \tag{4}$$

If $K < B$, then the r approximates the number of classes present in the batch by finding the linearly independent column vectors. Therefore, improving the rank of A is desirable. Our objective is to maximize the rank (r) of A which can be achieved by maximizing the nuclear-norm of it. Inequality shown in Eq.4 suggest that we can achieve the desired objective by maximizing the $||A||_F$ because it also provides the lower bound of $||A||_*$. Cui *et al*. shows that the approximation of NM using batch Frobenius norm improves the model performance and reduces the training time [6]. Hence, we define the Nuclear-Norm loss using its Frobenius approximation as shown in Eq.5.

$$\mathcal{L}_{NM} = -\|A\|_F = -\|(f_t(X_T^B;\theta_t,\phi_s)\|_F \tag{5}$$

where $\|(f_t(X_T^B;\theta_t,\phi_s)\|_F$ is Frobenius norm of classification-response matrix thereby minimizing $\mathcal{L}_{NM}$ improves class-discriminability as well as prediction-diversity.

Initial Pseudo label (PL): To improve model performance using self-training [45], we propose to use Pseudo label based cross-entropy loss ($\mathcal{L}_{CE}^{P_l}$ in Eq.10). Pseudo labels are inherently noisy, so directly computing target pseudo labels using source trained model is not desirable [40]. We use it along with nuclear norm maximization, which acts as soft regularization for self-training. We follow an iterative strategy similar to Liang *et al*. [25] to obtain pseudo labels. We get the initial class $c_k^{(init)}$ center using weighted k-means as shown in Eq.6

$$c_k^{(init)} = \frac{\sum_{x_t\in\mathcal{X}_\mathcal{T}}\delta_k(f_t(x_t;\theta_t,\phi_s))\; g_t(x_t;\theta_t)}{\sum_{x_t\in\mathcal{X}_\mathcal{T}}\delta_k(f_t(x_t;\theta_t,\phi_s))} \tag{6}$$

We can find the pseudo label $\forall x_t \in \mathcal{X}_\mathcal{T}$ based on their maximum cosine similarity with the initial class-center as shown in Eq.7.

$$\hat{y}_t^{init} = \arg\max_k \frac{\langle g_t(x_t;\theta_t), c_k^{(init)}\rangle}{\|g_t(x_t;\theta_t)\|\|c_k^{(init)}\|} \tag{7}$$

This allows us to assign each target sample to only one class. We can find the updated class center using the fraction of sample belonging to each class. We use Eq.6 and

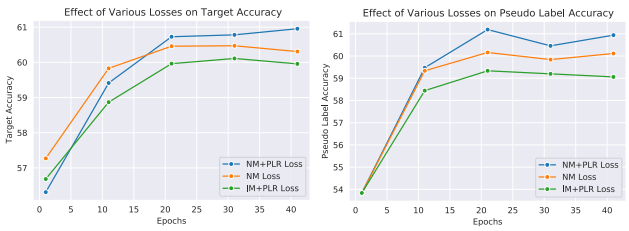

Figure 5: Plot for Nuclear-Norm Maximization (NM) vs. Information Maximization (IM) for Ar→Cl. *Left* and *Right* plot compares target accuracy and pseudo label accuracy for IM and NM with and without pseudo-label refinement (PLR) respectively. NM+PLR used in CoNMix provides the best performance.

Eq.7 in an iterative manner to find the updated pseudo label.

Pseudo Label Refinement (PLR): In order to reduce the noise in pseudo label, we refine pseudo labels using the temporal ensemble and consensus based weighting scheme [59]. Intuitively, If a pseudo label is consistent in two consecutive epochs then it should get more weight and vice-versa. Let $\tilde{y}_z^{n-1}$ and $\tilde{y}_z^n$ are pseudo label for z^{th} sample in epoch n and $n-1$ respectively. Let $W \in R^{K\times K}$ is cluster consensus matrix where, K is the total number of classes. If total number of samples in i^{th} class at n^{th} epoch is denoted as $I^n(i)$ then $W(i,j)$ is shown in Eq.8

$$W(i,j) = \frac{|I^{n-1}(i) \cap I^n(j)|}{|I^{n-1}(i) \cup I^n(j)|} \in [0,1] \tag{8}$$

Where $|.|$ is cardinality of a set. Row normalized $W(i,j)$ captures the similarity between i^{th} class and j^{th} class in epoch $n-1$ and n. Ideally, off-diagonal entries of matrix W should be close to zero. Finally, the updated pseudo label is shown in Eq.9.

$$\hat{y}_z^n = \alpha \tilde{y}_z^n + (1-\alpha) W^T \tilde{y}_z^{n-1} \tag{9}$$

where α is a hyper-parameter. We use refined pseudo label to calculate the $\mathcal{L}_{CE}^{P_l}$ loss as shown in Eq.10 where $\hat{y}$ is refined pseudo label and $\mathbb{1}_{[k=\hat{y}_t]}$ is indicator function. Fig. 5 shows the effectiveness of PLR when used with NM and IM loss.

$$\mathcal{L}_{CE}^{P_l} = -\mathbb{E}_{(x_t,\hat{y}_t)\in\mathcal{X}_\mathcal{T}\times\hat{\mathcal{Y}}_\mathcal{T}} \sum_{k=1}^{K} \mathbb{1}_{[k=\hat{y}_t]} \log \delta_k(f_t(x_t;\theta_t,\phi_s)) \tag{10}$$

Consistency Loss: For learning domain invariant representation, we propose weak and strong augmentation of the target image and seek consistent representation across the two label preserving augmentations as shown in Fig. 2(b). Let, X_{tw}^B, X_{ts}^B are weak and strong augmentation for target batch X_t^B and Y_{tw}^B, Y_{ts}^B are respective model softmax output i.e $Y_{tw}^B = \delta_k(h_t(g_t(X_{tw}^B))$ and $Y_{ts}^B = \delta_k(h_t(g_t(X_{ts}^B))$. We define an expectation ratio as $\mathcal{E}_{ratio} = \mathbb{E}[Y_{tw}^{all}]/\mathbb{E}[Y_{tw}^B]$. Y_{tw}^B is then normalized as $\hat{Y}_{tw}^B = \delta_k(Y_{tw}^B \mathcal{E}_{ratio})$ such that the row sum is 1. $\hat{Y}_{tw}^B$ acts as soft label ground truth for strong augmented output Y_{ts}^B and we minimize soft label based cross-entropy loss as shown below.

$$\mathcal{L}_{cons} = -\mathbb{E}_{(y_{ts})\in Y_{ts}^B} \sum_{k=1}^{K} \hat{y}_{tw}^k \log y_{ts}^k \tag{11}$$

$\mathcal{E}_{ratio}$ ensures that first order batch statistic matches with first order overall target data statistics. Overall loss for Stage-2 training is given by $\mathcal{L}_{total}$

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{NM} + \lambda_2 \mathcal{L}_{CE}^{P_l} + \lambda_3 \mathcal{L}_{cons} \tag{12}$$

where $\lambda_1, \lambda_2, \lambda_3$ are weights associated w.r.t three losses $\mathcal{L}_{NM}, \mathcal{L}_{CE}^{P_l}$ and $\mathcal{L}_{cons}$ respectively. In Fig. 5 we show that accuracy of pseudo label increases as training progresses which will result in improved adaptation. IM loss has been used in source-free domain adaptation task [25, 26] whereas NM loss is not explored for this task.

3.2.3 Source-free Multi-target Domain Adaptation

For extending *CoNMix* for SF-MTDA task, we propose a simple yet effective knowledge distillation based approach to transfer knowledge from all SF-STDA trained models (teachers) into a single student network (Stage 3 of Fig. 2). Seminal work in KD by Hinton *et al*. [15] showed that the high temperature distillation is equivalent to minimizing $\mathcal{L}_{KD} = 0.5(Z_t - Z_l)^2$ loss which pays significant attention in matching the logits from two networks. However, simply using Hinton Loss tends to overfit the teacher predictions. To avoid memorization and sensitivity to training examples, we propose MixUp Knowledge Distillation inspired from Zhang *et al*. work [58]. We first initialize a student model $g_l(x;\theta_l)$ with ImageNet trained weights. We store target image and it's corresponding pseudo label generated by each teacher network. An intermediate virtual domain image is generated by taking the convex combination of two randomly sampled images $\widetilde{x}_{ij} = \lambda x_i + (1-\lambda)x_j$ and $\widetilde{y}_{ij} = \lambda y_i + (1-\lambda)y_j$. Here (x_i, y_i) represents image and pseudo label pairs sampled from i^{th} domain. Here $\lambda \in [0,1]$. $(\widetilde{x}_{ij}, \widetilde{y}_{ij})$ represents a sample from an intermediate domain. We use all such pairs to train the student network using as a knowledge distillation loss as shown in Eq. 13. Derivation for Eq. 13 is shown in supplementary (see Section 2.4).

$$\begin{aligned} \mathcal{L}_{MKD} &= \mathcal{L}_{CE}^{P_l}(\widetilde{x}_{ij}, \widetilde{y}_{ij}) \\ \mathcal{L}_{MKD} &= \lambda \times \mathcal{L}_{CE}^{P_l}(\widetilde{x}_{ij}, y_i) + (1-\lambda) \times \mathcal{L}_{CE}^{P_l}(\widetilde{x}_{ij}, y_j) \end{aligned} \tag{13}$$

Intermediate domain acts as an implicit regularizer which helps to avoid over-fitting and generalize well on unlabeled target domains (Refer split domain test in supp. material). The proposed offline knowledge distillation allows us to distil knowledge from the best available STDA model because we aren't training teachers and students simultaneously. We can make inference with the final student model

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
Source train (RN50)	✓	45.1	67.5	74.7	52.4	61.7	65.7	52.6	39.7	71.8	64.4	44.5	77.4	59.8
G-SFDA (RN50) [55]	✓	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
CPGA (RN50) [37]	✓	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
SHOT (RN50) [25]	✓	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
SHOT++ (RN50) [26]	✓	57.9	79.7	82.5	68.5	79.6	79.3	68.5	57.0	83.0	73.7	60.7	84.9	73.0
CoNMix (RN50)	✓	57.6	77.2	82.2	68.4	78.8	78.3	67.1	54.7	81.5	74.0	60.2	85.3	72.1
Source train (DeiT-S)	✓	51.7	74.2	79.3	62.6	72.5	74.7	64.0	47.5	79.6	69.9	49.8	80.9	67.2
CDTrans (DeiT-S) [53]	✗	60.6	79.5	82.4	75.6	81.0	82.3	72.5	56.7	84.4	77.0	59.1	85.5	74.7
SHOT*(DeiT-S) [25]	✓	60.6	82.6	83.2	74.2	83.2	81.4	71.8	59.2	83.3	74.9	60.6	86.1	75.1
SHOT++*(DeiT-S) [26]	✓	62.6	83.4	83.9	**74.7**	83.3	**82.7**	72.2	59.0	83.7	74.7	60.6	**86.7**	75.6
CoNMix (DeiT-S)	✓	**63.4**	**83.5**	**84.6**	73.7	**83.3**	82.2	**73.4**	**59.9**	**84.4**	**75.6**	**62.3**	85.9	**76.0**

Table 1: Accuracy (%) on Office-Home for SF-STDA. Methods that uses DeiT-S are compared within shaded region. * represents experiments implemented by us. *CoNMix* (DeiT-S) achieves highest STDA average accuracy among all source-free methods.

without needing domain labels. Refer suppl. material for more direct analysis on how multi domain features align.

Method	SF	A→D	A→W	D→A	D→W	W→A	W→D	Avg
Source train (RN50)	✓	79.3	75.8	63.8	95.5	63.8	99.0	79.5
MA(RN50) [24]	✓	92.7	93.7	75.3	98.5	77.8	99.8	89.6
CPGA(RN50) [37]	✓	94.4	**94.1**	76.0	98.4	76.6	99.8	89.9
SHOT(RN50) [25]	✓	94.0	90.1	74.7	98.4	74.3	99.9	88.6
SHOT(RN50)++ [26]	✓	**94.3**	90.4	76.2	**98.7**	75.8	99.9	89.2
CoNMix (RN50)	✓	88.8	94.0	**77.3**	98.1	75.2	**100.0**	88.9
Source train (DeiT-S)	✓	79.9	82.3	70.3	96.6	71.2	99.8	83.3
CDTrans (Deit-S) [53]	✗	94.6	93.5	78.4	98.2	**78.0**	99.6	**90.4**
CoNMix (DeiT-S)	✓	90.6	**94.1**	77.2	98.1	77.0	99.6	89.4

Table 2: Accuracy (%) on Office-31 for STDA. Methods within shaded regions use DeiT backbone.

4. Experiments

We conducted experiments using four popular benchmarking datasets: Office-31 [42], Office-Home [51] and large-scale like DomainNet [34] and VisDA [36] dataset. After analyzing the benefits of VT over ResNet, we extended our analysis using VT as a backbone for *CoNMix*. For a fair comparison, we conducted experiments on Office-31 and Office-Home using a smaller VT network (DeiT-S [48]) with 22M parameter for both student's and teacher's backbone because DeiT-S is comparable to ResNet50 (25M parameter). For DomainNet, we used Hybrid ViT [8] for teacher models in Stage-1 and Stage-2. In Stage-3, ResNet101 is used as student model. Please refer to suppl. for additional training details.

4.1. Evaluation

Results for SF-STDA: We use Stage-2 of our proposed framework *CoNMix* for the SF-STDA. Table.1 and 2 illustrates results obtained for SF-STDA task for all combinations of domain pairs in Office-Home and Office-31. Our method outperforms existing source-free SOTA results with DeiT backbone in the case of the Office-Home dataset by a margin of **0.4%**. We have achieved significant improvement for STDA in DomainNet by **6.0%** (Refer supp. material table 1). Existing works [56, 41] only provide results for $Real$ and $Painting$ without comparing $Quickdraw$.

Method	SF	Office-Home				
		Ar	Cl	Pr	Rw	Avg
Source only (RN50)	✓	62.5	61.2	55.1	61.8	60.1
Source only (DeiT-S)	✓	68.4	71.2	63.6	66.4	67.4
Domain-Aggregation	✓	69.5	77.2	66.4	67.0	70.0
SHOT STDA in Stage-3	✓	73.1	77.7	69.2	72.4	73.1
CoNMix (ours)	✓	**75.6**	**81.4**	**71.4**	**73.4**	**75.4**

Table 3: SF-MTDA baselines. In *Domain-Aggregation*, we combines multiple target domains and treat it as a single domain. In *SHOT STDA*, we initialization for student network using SF-STDA SHOT weight. Highest performance for CoNMix highlights the importance of each design component in SF-MTDA.

We have compared against these works in Table.4. However, we also report STDA accuracy for all other possible splits in suppl. material. Even though *CoNMix* is source-free, we outperform non source-free method [53] by **0.3%** which showcases the efficacy of the proposed approach. We have included SF-STDA results for VisDA datasets in suppl. material (Table 4).

Results for SF-MTDA: There are no existing comprehensive studies related to SF-MTDA. Therefore, we formulated few baselines to evaluate SF-MTDA and reported the results in the Table. 3. We have considered source only training as a initial baseline, where we train on only on source dataset and evaluate its performance on all the target domains. For source train row in Table. 3 Art (Ar) represents training on Ar domain and testing on remaining domains. From Table. 3, we can observe that we achieve test accuracy of 60.1 %

Model	SF	DomainNet								
		R → S	R → C	R → I	R → P	P → S	P → R	P → C	P → I	**Avg** (%)
CDAN [29]	✗	40.7	51.9	22.5	49.0	39.6	57.9	44.6	18.4	40.6
HGAN [56]	✗	34.3	43.2	17.8	43.4	35.7	52.3	35.9	15.6	34.7
CDAN + DCL [41]	✗	45.2	58.0	23.7	54.0	45.0	61.5	50.7	20.3	44.8
D-CGCT [41]	✗	48.4	59.6	25.3	55.6	45.3	58.2	51.0	21.7	45.6
CoNMix (ours)	✓	**52.9**	**63.5**	**27.7**	**59.5**	**53.3**	**71.8**	**59.7**	**24.0**	**51.6**

Table 4: % Accuracy for SF-STDA on DomainNet Dataset. Our source-free method (Shaded region) outperforms the existing SOTA with significant margin even though they access source-dataset during target adaptation.

and 67.4% using ResNet and DeiT-S backbone respectively. Since, these results do not incorporate any adaptation, therefore, performing any adaptation using these models should lead to improvement in accuracy. Secondly, we considered aggregating all the target domain datasets together and train SF-STDA model. We can see its performance in Row-3 *(Domain-Aggregation)* of Table. 3 is better than source only but lesser than CoNMix. It shows that the proposed training strategy for CoNMix utilises domain information effectively. In another baseline (Row-4: *SHOT STDA in Stage-3*, we perform student training using SHOT SF-STDA weights in place of CoNMix SF-STDA weights. We can observe that its performance lies between domain aggregation and CoNMix. Therefore, the main components in SF-MTDA such as proposed teacher training and MixUp plays an important role in achieving the desired result.

Our SF-MTDA results on popular benchmark datasets will serve as a new baseline for research in this direction. Each cell in Table 5 and 6 reports classification accuracy of model which is adapted from *Source Domain* $\rightarrow$ *Rest of Target Domains*. SF represents whether the algorithm supports source-free method or not. We fine-tune the student network using MKD objective on all the target domains. Experiment with ResNet-101 provides initial baseline which consists of source training using ResNet-101 backbone and directly evaluating its performance on target dataset without performing any adaptation. Our source-free method achieves a significant improvement of **5.2%** over existing SOTA methods on the Office-Home dataset even though other methods access the labeled source data during adaptation. Our experiments with DomainNet dataset can be used to validate the scalability of our SF-MTDA (Table 6). We are the first to provide results for large-scale DomainNet dataset for both SF-STDA and SF-MTDA.

Model	SF	Office-31				Office-Home				
		A	D	W	**Avg**(%)	Ar	Cl	Pr	Rw	**Avg**(%)
RevGrad [11]	✗	78.2	72.2	69.8	73.4	58.4	58.1	52.9	62.1	57.9
CDAN [29]	✗	93.6	80.5	81.3	85.1	59.5	61.0	54.7	62.9	59.5
AMEAN [4]	✗	90.1	77.0	73.4	80.2	64.3	65.5	59.5	66.7	64.0
MT-MTDA [33]	✗	87.9	83.7	84.0	85.2	64.6	66.4	59.2	67.1	64.3
HGAN [56]	✗	88.0	84.4	84.9	85.8	-	-	-	-	-
CGCT [41]	✗	**93.9**	85.1	85.6	88.2	67.4	68.1	61.6	68.7	66.5
D-CGCT [41]	✗	93.4	**86.0**	**87.1**	**88.8**	70.5	71.6	66.0	71.2	69.8
Source train (RN50)	✓	76.3	68.7	67.0	70.7	62.5	61.2	55.1	61.8	60.1
Source train (DeiT-S)	✓	81.4	76.1	75.5	77.7	68.4	71.2	63.6	66.4	67.4
CoNMix (ours)	✓	92.4	81.8	80.4	84.9	**75.6**	**81.4**	**71.4**	**73.4**	**75.4**

Table 5: % Accuracy for Office-31 and Office-Home dataset for SF-MTDA. *CoNMix* outperforms SOTA in all possible splits of Office-Home.

Ablation for loss function: For better understanding of the effect of each loss function, we conducted an ablation study for different losses and show the result in Table 7. If we use $\mathcal{L}_{CE}^{P_l}$ or $\mathcal{L}_{Cons}$ individually, the performance is very poor. We observe that the $\mathcal{L}_{NM}$ is the most important loss component in overall optimization objective. Relative gains due to $\mathcal{L}_{CE}^{P_l}$ and $\mathcal{L}_{cons}$ may be smaller but we achieve best performance when all the components are present. The usage of both $\mathcal{L}_{CE}^{P_l}$ and $\mathcal{L}_{cons}$ together is not expected to handle noise present in pseudo label during training and it will deteriorate the model performance. These observations are consistent across various splits. We add additional analysis on pseudo label refinement and loss function in supplementary material (supp. Fig. 1 & 5).

Method	SF	Office-Caltech				
		A	C	D	W	Avg
ResNet-101 [14]	✗	90.5	94.3	88.7	82.5	89.0
SE [9]	✗	90.3	94.7	88.5	85.3	89.7
MCD [43]	✗	91.7	95.3	89.5	84.3	90.2
DANN [10]	✗	91.5	94.3	90.5	86.3	90.7
DADA [35]	✗	92.0	95.1	91.3	93.1	92.9
Source only	✓	90.7	96.1	90.2	90.9	92.0
SHOT	✓	96.2	97.3	96.3	96.2	96.5
CoNMix (ours)	✓	**96.4**	**97.4**	**96.9**	**96.8**	**96.9**

Method	SF	DomainNet						
		Cli.	Inf.	Pai.	Qui.	Rea.	Ske.	Avg
SE [9]	✗	21.3	8.5	14.5	13.8	16.0	19.7	15.6
MCD [43]	✗	25.1	19.1	27.0	10.4	20.2	22.5	20.7
DADA [35]	✗	26.1	20.0	26.5	12.9	20.7	22.8	21.5
CDAN [29]	✗	31.6	27.1	31.8	12.5	33.2	35.8	28.7
MCC [18]	✗	33.6	30.0	32.4	13.5	28.0	35.3	28.8
CGCT [41]	✗	36.1	33.3	35.0	10.0	39.6	39.7	32.3
D-CGCT [41]	✗	37.0	32.2	37.3	19.3	39.8	40.8	34.4
Source (RN101)	✓	25.6	16.8	25.8	9.2	20.6	22.3	20.1
CoNMix (ours)	✓	**41.8**	**29.2**	**39.9**	**17.5**	**32.7**	**41.2**	**33.7**

Table 6: Classification accuracy (%) on Office-Caltech and DomainNet for MTDA. Methods in shaded region are source-free.

$\mathcal{L}_{NM}$	$\mathcal{L}_{Cons}$	$\mathcal{L}_{CE}^{P_l}$	Ar→Cl	Cl→Ar	Pr→Cl	Cl→Pr	Rw→Cl	Cl→Rw
		✓	6.3	15.7	8.8	5.5	2.9	13.9
	✓		3.8	6.7	4.1	1.2	4.3	5.4
✓			59.5	71.1	55.9	78.6	58.3	78.7
✓	✓		60.3	72.7	57.6	80.4	59.8	78.6
✓	✓	✓	63.8	73.7	59.9	83.3	62.3	82.2

Table 7: Ablation for target accuracy when various losses are introduced sequentially across various splits of Office-Home.

5. Conclusion

In this work, we introduced a novel framework (*CoNMix*) for solving source-free Single and Multi-target domain adaptation. We achieved SOTA results for various datasets and in some cases CoNMix performed better than even non-source free methods. We provided baseline for source-free STDA and MTDA methods on DomainNet for the first time, which can help the domain adaptation research community further investigate this novel direction. Further, we provided empirical insights along with quantitative and qualitative results highlighting the benefit of VT and suggest that VT could be a potential choice for feature extractor in designing novel domain adaptation algorithms. We showed that our design choice, such as Nuclear-Norm Maximization, consistency constraint and label refinement mitigate uncertainty associated with noisy labels. *CoNMix* demonstrated effectiveness through various experimental findings across datasets, therefore we are keen to extend this further for more challenging source-free online adaptation where target domains are dynamic and continuously changing.

Acknowledgement: This work is supported by a Young Scientist Research Award (Sanction no. 59/20/11/2020-BRNS) from DAE-BRNS, India.

References

[1] Sk Miraj Ahmed, Dripta S Raychaudhuri, Sujoy Paul, Samet Oymak, and Amit K Roy-Chowdhury. Unsupervised multi-source domain adaptation without access to source data. *arXiv*, 2021.

[2] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

[3] Minghao Chen, Shuai Zhao, Haifeng Liu, and Deng Cai. Adversarial-learned loss for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3521–3528, 2020.

[4] Ziliang Chen, Jingyu Zhuang, Xiaodan Liang, and Liang Lin. Blending-target domain adaptation by adversarial meta-adaptation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2248–2257, 2019.

[5] 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.

[6] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Fast batch nuclear-norm maximization and minimization for robust domain adaptation. *arXiv preprint arXiv:2107.06154*, 2021.

[7] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *Proc. CVPR*, 2020.

[8] 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. *Proc. ICLR*, 2021.

[9] Geoff French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *Proc. ICLR*, 2018.

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[11] 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.

[12] Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing*, 29:3993–4002, 2020.

[13] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer, 2009.

[14] 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.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[16] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[17] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2018.

[18] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *Proc. ECCV*, 2020.

[19] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019.

[20] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4544–4553, 2020.

[21] Vinod Kumar Kurmi, Shanu Kumar, and Vinay P Namboodiri. Attending to discriminative certainty for domain adaptation. In *Proc. CVPR*, 2019.

[22] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 615–625, 2021.

[23] Quoc V Le, Jiquan Ngiam, Adam Coates, Ahbik Lahiri, Bobby Prochnow, and Andrew Y Ng. On optimization methods for deep learning. In *ICML*, 2011.

[24] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[25] 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.

[26] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[27] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021.

[28] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X Yu, and Boqing Gong. Open compound

domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2020.

[29] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1647–1657, 2018.

[30] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *arXiv preprint arXiv:1602.04433*, 2016.

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

[32] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *arXiv*, 2019.

[33] Le Thanh Nguyen-Meidine, Atif Belal, Madhu Kiran, Jose Dolz, Louis-Antoine Blais-Morin, and Eric Granger. Unsupervised multi-target domain adaptation through knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1339–1347, 2021.

[34] 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.

[35] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. *arXiv preprint arXiv:1904.12347*, 2019.

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

[37] 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. *CoRR*, abs/2106.15326, 2021.

[38] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[39] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *CoRR*, abs/1902.10811, 2019.

[40] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021.

[41] Subhankar Roy, Evgeny Krivosheev, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Curriculum graph co-teaching for multi-target domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5351–5360, 2021.

[42] 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.

[43] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018.

[44] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

[45] Inkyu Shin, Sanghyun Woo, Fei Pan, and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In *European conference on computer vision*, pages 532–548. Springer, 2020.

[46] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[47] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. When robustness doesn't promote robustness: Synthetic vs. natural distribution shifts on imagenet, 2020.

[48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

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

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv*, 2017.

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

[52] Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

[53] 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.

[54] Guanglei Yang, Hao Tang, Zhun Zhong, Mingli Ding, Ling Shao, Nicu Sebe, and Elisa Ricci. Transformer-based source-free domain adaptation. *arXiv preprint arXiv:2105.14138*, 2021.

[55] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. *CoRR*, abs/2108.01614, 2021.

[56] Xu Yang, Cheng Deng, Tongliang Liu, and Dacheng Tao. Heterogeneous graph attention network for unsupervised multiple-target domain adaptation. *TPAMI*, 2020.

[57] Ting Yao, Yingwei Pan, Chong-Wah Ngo, Houqiang Li, and Tao Mei. Semi-supervised domain adaptation with subspace learning for visual recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2142–2150, 2015.

[58] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. MixUp: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[59] Xiao Zhang, Yixiao Ge, Yu Qiao, and Hongsheng Li. Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3436–3445, 2021.

[60] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1), 2009.