Confident Anchor-Induced Multi-Source Free Domain Adaptation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Unsupervised domain adaptation has attracted appealing academic attentions by transferring knowledge from labeled source domain to unlabeled target domain. However, most existing methods assume the source data are drawn from a single domain, which cannot be successfully applied to explore complementary transferable knowledge from multiple source domains with large distribution discrepancy. Moreover, they require access to source data during training, which are inefficient and unpractical due to privacy preservation issues and storage concerns. To address these challenges, we develop a novel Confident Anchor-induced multi-source free Domain Adaptation (CAiDA) model, which is an earlier exploration of knowledge adaptation from multiple source domains to unlabeled target domain *without any source data, but with pre-trained source models*. Specifically, a source-specific transferability perception module is proposed to automatically quantify the transferability contributions of complementary transferable knowledge from multi-source domains to target domain. To generate pseudo labels for target domain without access to source data, we develop a confident anchor-induced pseudo label generator by constructing a confident anchor group and assigning each unconfident target sample with a semantic-nearest confident anchor. Furthermore, a class relationship-aware consistency loss is proposed to preserve consistent inter-class relationships by aligning soft confusion matrices across different domains. Theoretical analysis answers why multi-source domains are better than single source domain, and establishes a learning bound to show the effect of multi-source domains. Experiments on several representative datasets illustrate the superiority of our proposed model.

1 Introduction

Unsupervised Domain Adaptation (UDA) transfers the knowledge from labeled data in a source domain to classify unlabeled data in a target domain. UDA has achieved remarkable successes in many applications, *e.g.*, object detection [16], medical diagnose [7], sentiment analysis [27, 41], person re-identification [2], etc. Generally, most of existing methods employ adversarial learning [10] to encourage source and target features to be indistinguishable from each other [8,35,44,50], or minimize distribution discrepancy across domains by matching the distribution statistical moments [30, 33, 48].

However, above methods have a strong assumption that the source data are merely drawn from a single domain. Unfortunately, the source data are often collected under different deployment environments (*i.e.*, from multiple source domains with large distribution discrepancy) in real-world applications, which makes them difficult to explore complementary transferable knowledge from multiple source domains for target prediction. To achieve this, Multi-Source Domain Adaptation (MSDA) [25, 27, 53] is proposed to match the features across domains and then quantify the contributions of source domains by taking an average of source predictions [2, 30, 51]. Additionally, [23, 49, 55] aim to weight the source contributions by normalizing the distance similarities between source and target domains.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Unfortunately, recent MSDA methods [23, 27, 30, 32] require massive labeled source data when adapting to target domain. This could make them inefficient and unpractical in real-world applications with sensitive information (*e.g.*, medical diagnosis [7] and recommendation system [17]), due to privacy preservation issues, storage and security concerns [18]. To this end, a new challenging practical problem named *Multi-Source Free Domain Adaptation* (MSFDA) is researched, which explores transferable knowledge from multiple source domains to target domain *with only source models and without access to any source data*. The trivial solutions for tackling MSFDA via using existing single-source free domain adaptation methods [18, 24, 26, 47] are to adapt each source model individually and simply take an average prediction of source models. However, they cannot explore transferability contributions of complementary information from different source domains, due to lack of source data. Therefore, tackling MSFDA problem is a challenging but rarely-researched task.

To address MSFDA problem, we develop a novel Confident Anchor-induced multi-source free Domain Adaptation (CAiDA) model, which is an earlier exploration to capture transferable information from multiple source models to promote target prediction without access to source data. Specifically, a source-specific transferability perception module is designed to automatically capture the transferability contributions of complementary information from multiple source domains. We develop a confident anchor-induced pseudo label generator to mine pseudo labels for unlabeled target data without access to source data, by incorporating with the quantified source transferability contributions. We construct a confident anchor group to assign each target sample with a semantic-nearest confident anchor, and performs feature augmentation between them to generate confident target pseudo label. A class relationship-aware consistency loss is proposed to ensure the semantic consistency of underlying inter-class relationships across domains via soft confusion matrix alignment. Furthermore, based on mild assumptions, theoretical analysis guarantees that multiple source models could improve the possibility to generate more reliable pseudo labels. Our theory also provides a novel learning bound for MSFDA, which reveals that multiple source models improve the possibility to achieve a tighter generalization error bound for the target domain. Experiments on benchmark datasets verify the effectiveness of our model. In summary, the main contributions are as follows:

• We develop a novel Confident Anchor-induced multi-source free Domain Adaptation (CAiDA) method to explore transferable knowledge from multiple source domains to assist target prediction with only pre-trained source models and without access to source data. To our best knowledge, this paper is an earlier exploration of multi-source free domain adaptation task in transfer learning field.

• We propose a novel MSFDA theory, which shows that multiple source models could improve the possibility of obtaining more reliable pseudo labels under mild assumptions. Our theoretical analysis also provides a novel generalization bound for MSFDA to show the effect of multiple source models.

• A source-specific transferability perception module, and a class relationship-aware consistency loss are designed to quantify transferability contributions of multiple source domains and ensure the semantic consistency of underlying inter-class relationships across different domains, respectively.

• Based on the novel pseudo labeling strategy in our theoretical analysis, a confident anchor-induced pseudo label generator is proposed to generate pseudo labels for the target domain by establishing a confident anchor group and assigning each target sample with a semantic-nearest confident anchor.

2 Related Work

Unsupervised Domain Adaptation aims to borrow transferable knowledge from source domain to promote the prediction of the unlabeled target domain. After Hoffman *et al.* [15] introduce adversarial learning [10] into domain adaptation, enormous adversarial-based methods [8, 35, 44, 50] are proposed to perform feature-level or pixel-level distribution alignment. Besides, some moment matching-based methods [30, 33, 48] focus on matching the distribution statistical moments at different orders to minimize the distribution discrepancy across domains. Furthermore, some researches design the adversarial dropout [22], batch normalization [40] and auxiliary reconstruction tasks [4, 9] to narrow domain discrepancy. Unfortunately, the methods mentioned above assume massive labeled source data are available. This is unsafe and unpractical due to privacy and security concerns. To tackle this challenge, a self-supervised pseudo labeling strategy [26] and self-entropy criterion [18] are designed to mine confident pseudo labels without any labeled source data and with pre-trained source models. [24, 47] employ adversarial learning strategy to perform model adaptation with only pre-trained source models. However, they cannot be directly applied to multi-source free domain adaptation task, due to the large distribution discrepancy between different source domains.

Multi-Source Domain Adaptation is extended from vanilla domain adaptation setting [6–8] by exploring transferable knowledge from multiple source domains. To capture the relationship between different source domains and a given target domain, Guo *et al*. [12] design a mixture-of-experts model for unsupervised domain adaptation from multiple sources. [46] focuses on determining which source domain is best for target prediction by learning dynamic curriculum. [2] introduces the multi-source concept into person re-ID field by considering both domain-specific and domain-fusion views. Some discrepancy based methods aim to narrow distribution discrepancy across domains by minimizing different measures such as *Rényi*-divergence [14] and maximum mean discrepancy [12,52]. Moreover, some adversarial-based methods focus on optimizing $\mathcal{H}$-divergence [30, 51], generative adversarial loss [45,53] and Wasserstein distance [25,39] to make features from multiple sources indistinguishable for a shared discriminator. [27, 32] perform knowledge adaptation at the pixel-level by replaying multiple source domains (*i.e.*, domain translation). However, due to lack of source data, many strategies in MSDA may be invalid and unsuitable to address the challenging MSFDA problem.

3 Problem Setting

Let $\mathcal{X}$ and $\mathcal{Y} = [K]$ denote the feature space and label space. A domain is a joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$. There are n source domains $\{P^i_{XY}\}_{i=1}^n$. For any source domain P^i_{XY}, a corresponding neural network-based predictor (model) $\boldsymbol{h}^i : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|} = \mathbb{R}^K$ is given. Given a target domain P^t_{XY} and unlabeled target data $T = \{\mathbf{x}^j\}_{j=1}^m \sim P^t_X$, i.i.d., the aim of *multiple-source free domain adaptation* (MSFDA) is to classify the unlabeled target data by utilizing T and predictors $\{\boldsymbol{h}^i\}_{i=1}^n$.

Let ℓ denote a non-negative loss function defined over $\mathbb{R}^K \times \mathbb{R}^K$. Given a hypothesis space $\mathcal{H} \subset \{\boldsymbol{h} : \mathcal{X} \to \mathbb{R}^K\}$, we denote $\mathcal{L}^i_s(\boldsymbol{h}) = \mathbb{E}_{(\mathbf{x},y)\sim P^i_{XY}} \ell(\boldsymbol{h}^i(\mathbf{x}), \boldsymbol{\Phi}(y))$ and $\mathcal{L}_t(\boldsymbol{h}) = \mathbb{E}_{(\mathbf{x},y)\sim P^t_{XY}} \ell(\boldsymbol{h}^i(\mathbf{x}), \boldsymbol{\Phi}(y))$ as the risks with respect to i-th source domain and target domain, where $\boldsymbol{\Phi} : \mathcal{Y} \to \mathbb{R}^K$ maps any label y to a corresponding one-hot vector.

The neural network-based predictor $\boldsymbol{h}^i$ is a vector-valued function, *i.e.*, $\boldsymbol{h}^i(\mathbf{x}) = [h^i_1(\mathbf{x}), ..., h^i_K(\mathbf{x})]^T$, and consists of two basic components: feature extractor $\mathbf{f}^i : \mathcal{X} \to \mathbb{R}^d$ and classifier $\mathbf{c}^i : \mathbb{R}^d \to \mathbb{R}^K$, where d is the dimension of extracted features. Hence, $\boldsymbol{h}^i$ can be rewritten as $\mathbf{c}^i \circ \mathbf{f}^i$. After softmax function as the activation function used in the output layer, we have $h^i_k \geq 0$ and $\sum_{k=1}^K h^i_k = 1$. To ensure the each predictor $\boldsymbol{h}^i$ is a relatively accurate predictor for the domain P^i_{XY}, we assume that the predictor $\boldsymbol{h}^i$ is ϵ-accurate under ℓ_1 loss, *i.e.*, $\mathbb{E}_{(\mathbf{x},y)\sim P^i_{XY}} \|\boldsymbol{h}^i(\mathbf{x}) - \boldsymbol{\Phi}(y)\|_{\ell_1} < \epsilon$, for any $i \in [n]$.

4 Theoretical Analysis

Without any relations between source domains and target domain, MSFDA cannot be effectively addressed from the theoretical view. To bridge source and target domains, Mansour *et al*. [29], Wu *et al*. [42] and Miraj Ahmed *et al*. [1] assume that $P_{XY} = \sum_{i=1}^n \lambda_i P^i_{XY}$ and $P^t_{Y|X} = P^i_{Y|X} = P^j_{Y|X}$, for any $i, j \in [n]$, where $\lambda_i \geq 0$ and $\sum_{i=1}^n \lambda_i = 1$. This assumption is very strong and may be unrealistic in real-world applications. Motivated by the meta learning [20, 31] and domain generalization [3], some novel and mild assumptions are proposed in this paper.

Assumption 1 (Meta Assumption.) *$P^t_{XY}, P^1_{XY}, ..., P^n_{XY}$ are drawn (i.i.d.) from a meta distribution $\mathcal{P}$, which is defined over a joint distribution space $\mathscr{P}_{XY}$.*

Assumption 2 (Regular Domain.) *Let the joint distribution space $\mathscr{P}_{XY}$ is endowed with total variation distance $d_{\mathrm{TV}}(\cdot,\cdot)$. The target domain P^t_{XY} is a regular domain,* i.e.*, for any $\sigma > 0$, $\mathscr{P}(\mathcal{N}^\sigma_{P^t_{XY}}) > 0$, where $\mathcal{N}^\sigma_{P^t_{XY}} = \{P : d_{\mathrm{TV}}(P, P^t_{XY}) < \sigma\}$.*

It is easy to check, if the meta distribution $\mathcal{P}$ is discrete or continuous with continuous density function, then the target domain P^t_{XY} drawn by $\mathcal{P}$ is a regular domain with probability 1. In addition, to weaken the assumption $P^t_{Y|X} = P^i_{Y|X} = P^j_{Y|X}$, for any $i, j \in [n]$, our key strategy is to consider the anchor point assumption that is studied in label-noise learning [28, 43].

We say a given point $\mathbf{x}$ is a *τ-anchor point*, if there exists a predictor $\boldsymbol{h}^i$ with largest score $h^i_k(\mathbf{x})$, such that $h^i_k(\mathbf{x}) - h^i_c(\mathbf{x}) \geq \tau$, for any $c \in [K]$ and $c \neq k$. We say $\boldsymbol{h}^i$ is the *τ-anchor predictor* for $\mathbf{x}$.

Assumption 3 (τ-Anchor Point Assumption.) *Given a τ-anchor point $\mathbf{x}$, suppose that all τ-anchor predictors for $\mathbf{x}$ are $\boldsymbol{h}^{i_1}, ..., \boldsymbol{h}^{i_l}$, then the true label of the τ-anchor point $\mathbf{x}$ is k, if $h^j_k(\mathbf{x})$ is the largest score among scores $h^{i_1}_c, ..., h^{i_l}_c$, for any $c \in [K]$.*

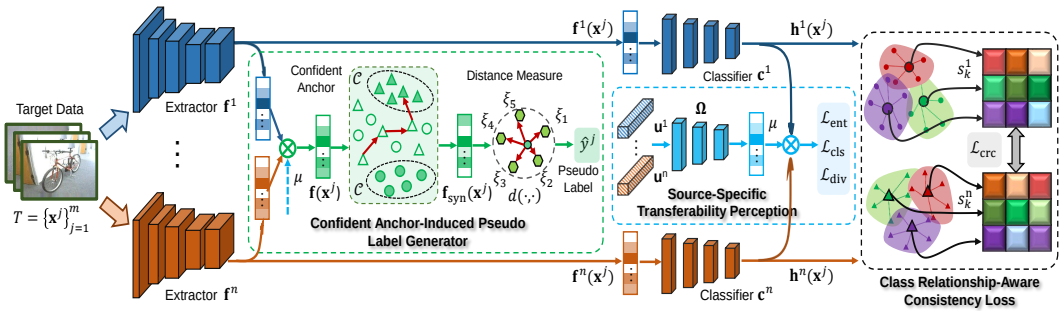


Figure 1: Overview of our CAiDA model, mainly including a *source-specific transferability perception* strategy to quantify transferability contributions of source domains, a *confident anchor-induced pseudo label generator* to generate confident pseudo labels for target domain, and a *class relationship-aware consistency loss* to ensure the semantic consistency of underlying inter-class relationships.

When the source predictors are accurate enough, Assumption 3 implies that the source and target conditional distributions are similar in the high confidence region. Hence, it is much weaker than the traditional assumption $P^t_{Y|X} = P^i_{Y|X} = P^j_{Y|X}$, for $i, j \in [n]$. τ is regarded as a threshold to distinguish which data has highly confident prediction. In general, τ is close to 1.

Theorem 1 *Assume Assumptions 1 and 2 hold and the conditional distribution $P^t_{Y|X}$ can be presented as a classifier ($P^t_{Y|X}(y|\mathbf{x})$ is 0 or 1). Given $\eta > 0$, if $m \geq n$ and $(1-\eta)(1-\tau) > \epsilon + 2\sigma + 2\sqrt{\log(2m/\delta)/2m}$, then with probability at least $1 - \delta - (1 - \mathscr{P}(\mathcal{N}^{\sigma}_{P_{X_t Y_t}}))^n > 0$, at least ηm target data are τ-anchor points, where ϵ is the upper bound of the accuracies of source predictors, and σ is introduced in Assumption 2.*

Theorem 1 shows multiple-source predictors improve the probability to obtain more τ-anchor points. To further understand the effect of multiple-source domains, we build a learning bound for MSFDA. Let A_τ be the set consisting of all τ-anchor points. Denote the empirical risk $\widehat{\mathcal{L}}^\tau_s(\boldsymbol{h})$ by $\frac{1}{|A_\tau \cap T|} \sum_{\mathbf{x} \in A_\tau \cap T} \ell(\boldsymbol{h}(\mathbf{x}), \boldsymbol{\Phi}(y))$, where y is the label of the anchor point $\mathbf{x}$.

Theorem 2 *Given Assumption 3 and same assumptions used in Theorem 1, and we assume the loss ℓ has upper bound $M > 0$ and hypothesis space $\mathcal{H}$ has finite Natarajan dimension, for $\eta > 0$, if $m \geq n$ and $(1-\eta)(1-\tau) > \epsilon + 2\sigma + 2\sqrt{\log(2m/\delta)/2m}$, then for any $\boldsymbol{h} \in \mathcal{H}$ and $b \in (0,1)$, there exists a constant $C(b, K)$ such that with the probability at least $1 - 2\delta - 2(1 - \mathscr{P}(\mathcal{N}^{\sigma}_{P^t_{XY}}))^n$:*

$$\left|\mathcal{L}_t(\boldsymbol{h}) - \widehat{\mathcal{L}}^\tau_s(\boldsymbol{h})\right| \leq MC(b, K)\sqrt{\frac{\log(2/\delta)}{\eta^{1-b} m^{1-b}}} + M\frac{2\sigma + \epsilon}{1 - \tau - 2\sigma - \epsilon}, \tag{1}$$

where ϵ is the upper bound of accuracies of source predictors, and σ is introduced in Assumption 2.

Theorem 2 shows that multiple source domains improve the probability to ensure a tighter generalization bound, *i.e.*, Eq. (1) holds. All detailed proofs are introduced in the supplementary material.

5 The Proposed CAiDA Model

The graphical illustration of our proposed CAiDA model is depicted in Figure 1. It mainly consists of three significant components: source-specific transferability perception, confident anchor-induced pseudo label generator and class relationship-aware consistency loss, which are elaborated as follows:

5.1 Source-Specific Transferability Perception

Generally, in multi-source domain adaptation (MSDA), different source domains have different contributions to improve the performance on target domain [14, 53]. To this end, many previous MSDA methods match the features between source and target domains, then quantify the contributions of source domains by taking an average of the trained source predictors [30, 51] or weight the trained source predictors by normalizing the distance similarities [23, 49, 55] between source and target domains. However, due to the lack of source data, these methods cannot employ source data to match features and cannot be successfully applied to multi-source free domain adaptation (MSFDA) tasks.

Therefore, a source-specific transferability perception module is developed to automatically quantify the transferability contributions of complementary knowledge from source domains, as shown in Figure 1. Specifically, the weight matrix $\mathbf{W}^i = [\mathbf{w}_1^i, \mathbf{w}_2^i, \cdots, \mathbf{w}_K^i] \in \mathbb{R}^{d' \times K}$ (d' denotes the feature dimension) of the last fully connected layer in classifier $\mathbf{c}^i$ could be considered as a distribution characterization of the i-th source domain, which is then employed to quantify the transferability contribution. Intuitively, $\{\mathbf{w}_k^i\}_{k=1}^K$ could be regarded as the k-th class prototype of the i-th source domain. We then take an average of all class prototypes to characterize the i-th source domain $\mathbf{u}^i = \frac{1}{K}\sum_{k=1}^K \mathbf{w}_k^i$, and concatenate all source domain characterizations together to obtain $\mathbf{U} = [\mathbf{u}^1, \cdots, \mathbf{u}^n]^\top \in \mathbb{R}^{n \times d'}$. $\mathbf{U}$ is then forwarded into a three-layer convolution network $\mathbf{\Omega}$ to automatically quantify the transferability contributions $\mu \in \mathbb{R}^n$ of n source domains: $\mu = \mathbf{\Omega}(\mathbf{U}) = \mathbf{\Omega}([\mathbf{u}^1, \mathbf{u}^2, \cdots, \mathbf{u}^n]^\top)$, such that $\sum_{i=1}^n \mu_i = 1$, where μ_i is the quantified contribution of the i-th source domain to target prediction.

When the source data are unavailable, it is difficult to narrow distribution discrepancy across domains, due to lack of any supervised information. Inspired by [26], we freeze the network parameters of classifiers $\{\mathbf{c}^i\}_{i=1}^n$ and solely perform distribution adaptation across domains on feature extractors $\{\mathbf{f}^i\}_{i=1}^n$ via information maximization [19, 34], since $\{\mathbf{c}^i\}_{i=1}^n$ contain class distribution information of source domains. However, Liang *et al*. [26] cannot be effectively applied to multi-source data free domain adaptation scenario, where different source domains have different transferability contributions on target adaptation. Therefore, $\mathcal{L}_{\text{ent}}$ is proposed to minimize conditional entropy of target outputs by incorporating with the source-specific transferability perception μ:

$$\mathcal{L}_{\text{ent}} = -\frac{1}{m}\sum_{j=1}^m [\sum_{k=1}^K h_k(\mathbf{x}^j)\log(h_k(\mathbf{x}^j))], \tag{2}$$

where $h_k(\mathbf{x}^j)$ is the k-th coordinate value of $\boldsymbol{h}(\mathbf{x}^j)$, here $\boldsymbol{h}(\mathbf{x}^j) = \sum_{i=1}^n \mu_i \boldsymbol{h}^i(\mathbf{x}_j)$ denotes the combination of source predictions. The larger value of μ_i indicates the larger transferability of the i-th source domain for target adaptation. Unfortunately, our proposed CAiDA model may suffer from the trivial solution of Eq. (2) by predicting all target data as a single class to minimize Eq. (2). To achieve this, $\mathcal{L}_{\text{div}}$ is designed to consider class prediction diversity by maximizing the entropy of empirical label distribution [5] predicted by different source domains:

$$\mathcal{L}_{\text{div}} = \sum\nolimits_{k=1}^K \hat{p}_k \log \hat{p}_k, \tag{3}$$

where $\hat{p}_k = \frac{1}{m}\sum_{j=1}^m h_k(\mathbf{x}^j)$ denotes the mean prediction probability over all target data.

5.2 Confident Anchor-Induced Pseudo Label Generator

Although the minimization of both $\mathcal{L}_{\text{ent}}$ and $\mathcal{L}_{\text{div}}$ promotes the class diversity and knowledge adaptation between target data and multiple unseen source data, it cannot circumvent erroneous label assignment due to the noisy target prediction brought by domain discrepancy. To alleviate this issue, based on Assumption 3, a confident anchor-induced pseudo label generator is developed to mine confident pseudo labels for target data, as shown in Figure 1. Specifically, based on the τ-anchor assumption, we select target data $\mathbf{x}^j$ as confident anchor when the maximum category prediction probability $h_{k^*}(\mathbf{x}^j)$ is larger than the rest of category prediction probabilities $h_k(\mathbf{x}^j)$ $(k \neq k^*)$ by a threshold τ_p. For each confident anchor, we integrate the features extracted from multiple source extractors together to obtain probability-based confident anchor group $\mathcal{C}_p$:

$$\mathcal{C}_p = \{\mathbf{f}(\mathbf{x}^j) | h_{k^*}(\mathbf{x}^j) - h_k(\mathbf{x}^j) \geq \tau_p, \forall k \neq k^*, j = 1, 2, \cdots m\}, \tag{4}$$

where $\mathbf{f}(\mathbf{x}^j) = \sum_{i=1}^n \mu_i \mathbf{f}^i(\mathbf{x}^j)$ denotes the feature of the j-th target data. τ_p is defined as the median value of the probability difference between the largest and the second largest probabilities over all target data. Furthermore, inspired by the proposed τ-anchor assumption, to circumvent the noisy anchor in $\mathcal{C}_p$, we construct a distance-based confident anchor group $\mathcal{C}_d$. The feature centroid of the k-th class induced by the i-th source domain for the whole target data is defined as $\xi_k^i = \sum_{j=1}^m h_k^i(\mathbf{x}^j)\mathbf{f}^i(\mathbf{x}^j) / \sum_{j=1}^m h_k^i(\mathbf{x}^j)$. $\{\xi_k^i\}_{i=1}^n$ are then weighted with source-specific transferability to obtain the feature centroid ξ_k of the k-th class over all source domains via $\xi_k = \sum_{i=1}^n \mu_i \xi_k^i$. The distance between the feature of $\mathbf{x}^j$ and the k-th feature centroid ξ_k is denoted as $d(\mathbf{f}(\mathbf{x}^j), \xi_k)$, where $d(\cdot,\cdot)$ is the distance measure function. When the minimum distance $d(\mathbf{f}(\mathbf{x}^j), \xi_{k^*})$ is shorter than the rest of distances $d(\mathbf{f}(\mathbf{x}^j), \xi_k)(k \neq k^*)$ by a threshold τ_d, we select the target data $\mathbf{x}^j$ into $\mathcal{C}_d$:

$$\mathcal{C}_d = \{\mathbf{f}(\mathbf{x}^j) | d(\mathbf{f}(\mathbf{x}^j), \xi_k) - d(\mathbf{f}(\mathbf{x}^j), \xi_{k^*}) \geq \tau_d, \forall k \neq k^*, j = 1, 2, \cdots m\}, \tag{5}$$

Algorithm 1 The Searching Process of Semantic-Nearest Confident Anchor $\mathbf{x}^{jc}$.

Input: $\mathbf{x}^j, \mathcal{C}, t = 1$;
Initialize: $\mathbf{x}^{j(t-1)} = \mathbf{x}^j, I = \varnothing$;
While $\mathbf{x}^{j(t-1)} \notin \mathcal{C}$ **do**
  Obtain $\mathbf{x}^{j(t)}$ via Eq. (6);
  Update I via adding $\mathbf{x}^{j(t)}$ into I;
  Update $\mathbf{x}^{j(t-1)}$ via $\mathbf{x}^{j(t-1)} \leftarrow \mathbf{x}^{j(t)}$;
End
Obtain $\mathbf{x}^{jc}$ via $\mathbf{x}^{jc} \leftarrow \mathbf{x}^{j(t-1)}$
Return $\mathbf{x}^{jc}$.

Algorithm 2 The Optimization of Our Model.

Input: $\{\boldsymbol{h}^i\}_{i=1}^n, T$, E epoches, B batches;
Initialize: κ_1, κ_2, the parameters of $\mathbf{\Omega}$;
For $e = 1, \cdots, E$ **do**
  Obtain pseudo labels via Eq. (7);
  For $b = 1, \cdots, B$ **do**
    Select a batch of target data;
    Update $\{\mathbf{f}^i\}_{i=1}^n$ via Eq.(9) while fixing $\mathbf{\Omega}$;
    Update $\mathbf{\Omega}$ via Eq. (9) while fixing $\{\mathbf{f}^i\}_{i=1}^n$;
  End
End
Return: $\{\boldsymbol{h}^i\}_{i=1}^n$ and $\mathbf{\Omega}$.

where we set τ_d as the median value of the distance difference between the shortest and the second shortest distances over all target data. Therefore, the final confident anchor group $\mathcal{C}$ is obtained by performing an intersection operation between $\mathcal{C}_p$ and $\mathcal{C}_d$, *i.e.*, $\mathcal{C} = \mathcal{C}_p \cap \mathcal{C}_d$.

To generate confident pseudo labels for unconfident target data, we select a semantic-nearest confident anchor from $\mathcal{C}$ for each target data via continual similarity searching. To be specific, given an unconfident target data $\mathbf{x}^j$, we search a serial of unconfident guiding data consecutively using the distance measure function $d(\cdot,\cdot)$ in the feature space, until the confident anchor from $\mathcal{C}$ is detected. During each searching process, the previous searched guiding data are not considered in the following iteration. Denote I as the set containing previously searched guiding data, $\mathbf{x}^{jc}$ as the searched confident anchor closest to $\mathbf{x}^j$. The guiding sample $\mathbf{x}^{j(t)}$ in the t-th search could be obtained by:

$$\mathbf{x}^{j(t)} = \arg\min_{\mathbf{x}^j \in T} d(\mathbf{f}(\mathbf{x}^{j(t-1)}), \mathbf{f}(\mathbf{x}^j)), \text{ subject to } \mathbf{x}^{j(t)} \neq \mathbf{x}^{j(t-1)}, \mathbf{x}^{j(t)} \notin I, \tag{6}$$

where $\mathbf{x}^{j(t-1)}$ is the guiding data in the previous search. The searching process of $\mathbf{x}^{jc}$ for the j-th target data is summarized in Algorithm 1. Moreover, we fuse the target data and their corresponding confident anchor in the feature space for feature augmentation, and the synthetic feature is denoted as $\mathbf{f}_{\text{syn}}(\mathbf{x}^j) = (1-\omega)\mathbf{f}(\mathbf{x}^j) + \omega\mathbf{f}(\mathbf{x}^{jc})$, where $\omega \in [0,1]$ is the random weight to determine the influence of confident anchor on $\mathbf{f}_{\text{syn}}(\mathbf{x}^j)$. Therefore, the confident anchor-induced pseudo label $\hat{y}^j$ of the j-th target data $\mathbf{x}^j$ and the classification loss $\mathcal{L}_{\text{cls}}$ of whole target data are formulated as:

$$\hat{y}^j = \arg\min_{k \in [K]} d(\mathbf{f}_{\text{syn}}(\mathbf{x}^j), \xi_k); \quad \mathcal{L}_{\text{cls}} = \frac{1}{m}\sum_{j=1}^{m}\sum_{k=1}^{K}[-\mathbb{1}_{\hat{y}^j=k}\log(h_k(\mathbf{x}^j))]. \tag{7}$$

5.3 Class Relationship-Aware Consistency Loss

The inherent relationships between different classes have semantic consistency across domains, regardless of distribution discrepancy. In light of this, aligning class relationships could promote more shared transferable knowledge from source domains towards target adaptation. To achieve this, as depicted in Figure 1, a class relationship-aware consistency loss $\mathcal{L}_{\text{crc}}$ is designed to encourage target data from the same class to be compactly clustered together while preserving the intrinsic inter-class relationships via soft confusion matrix alignment. To be specific, the soft label distribution s_k^i of the k-th class predicted via the i-the source predictor is formulated as $s_k^i = \frac{1}{m}\sum_{j=1}^{m}[\mathbb{1}_{\hat{y}^j=k}\mu_i h^i(\mathbf{x}^j)]$. The collection of soft label distributions $\{s_k^i\}_{k=1}^K$ represents a kind of soft confusion matrix associated with a particular domain, encoding the inter-class relationships learned by the i-th source predictor (*e.g.*, computers have more similar semantic relationship with desks than horses). Without access to the source data, $\mathcal{L}_{\text{crc}}$ aims to align soft confusion matrices between different source predictors:

$$\mathcal{L}_{\text{crc}} = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{k=1}^{K}\left(\text{KL}(s_k^i||s_k^{i'}) + \text{KL}(s_k^{i'}||s_k^i)\right), \tag{8}$$

where $\text{KL}(p||q) = \sum_r p_r \log\frac{p_r}{q_r}$ denotes the Kullback–Leibler (KL) divergence. The complexity of $\mathcal{L}_{\text{crc}}$ is not problematic in practice, due to the limited number of source domains.

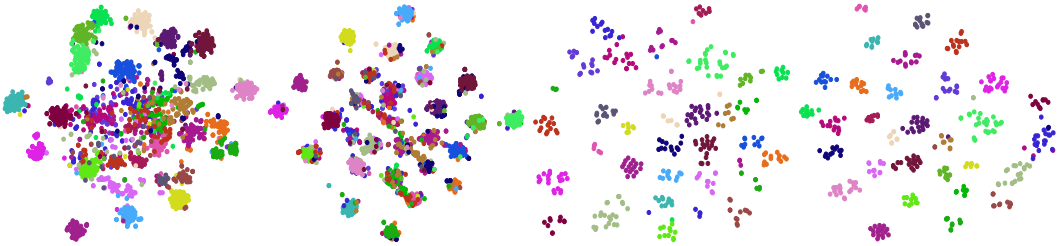
In summary, the overall optimization objective to optimize $\{\mathbf{f}^i\}_{i=1}^n$ and $\mathbf{\Omega}$ is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{div}} + \kappa_1\mathcal{L}_{\text{cls}} + \kappa_2\mathcal{L}_{\text{crc}}, \tag{9}$$

where κ_1, κ_2 are the balanced weights. The optimization of our model is presented in Algorithm 2.

Table 1: Comparisons between our model and other competing methods on Office-31 [15] dataset (the left block) and Office-Caltech [11] dataset (the right block).

Methods	Source Data	A, D→W	A, W→D	D, W→A	Avg.	A, D, C→W	A, C, W→D	C, D, W→A	A, D, W→C	Avg.
Source only [13]	✓	97.1	92.0	51.6	80.2	93.5	94.2	90.6	87.5	91.5
MDAN [51]	✓	99.2	95.4	55.2	83.3	99.4	98.7	93.5	91.6	95.8
DCTN [45]	✓	**99.6**	96.9	54.9	83.8	99.3	99.4	94.1	91.3	96.0
M^3SDA [30]	✓	99.4	96.2	55.4	83.7	99.5	99.2	94.5	92.2	96.4
MDDA [54]	✓	99.2	97.1	56.2	84.2	99.3	99.6	95.3	92.3	96.6
LtC-MSDA [38]	✓	**99.6**	97.2	56.9	84.6	99.4	99.7	93.7	95.1	97.0
Source model only	✗	95.4	97.5	60.2	84.4	98.0	99.5	96.3	92.1	96.5
BAIT [47]	✗	98.5	98.8	71.1	89.5	98.0	97.5	97.5	95.7	97.2
PrDA [18]	✗	93.8	96.7	73.2	87.9	97.6	97.1	97.3	94.6	96.7
SHOT [26]	✗	94.9	97.8	75.0	89.3	99.6	96.8	95.7	95.8	97.0
MA [24]	✗	96.1	97.3	75.2	89.5	**99.8**	97.2	95.7	95.6	97.1
DECISION [1]	✗	98.4	99.6	75.4	91.1	99.6	**100.0**	95.9	95.9	98.0
Ours-w/oEnt	✗	97.5	99.1	74.2	90.3	99.1	99.0	94.5	96.3	97.2
Ours-w/oDiv	✗	97.2	98.6	73.7	89.8	98.6	99.3	94.1	95.7	96.9
Ours-w/oCls	✗	96.7	98.4	73.0	89.4	97.3	98.4	93.6	95.2	96.1
Ours-w/oCrc	✗	98.6	99.5	75.4	91.2	99.6	**100.0**	95.3	96.5	97.9
Ours	✗	98.9	**99.8**	**75.8**	**91.6**	**99.8**	**100.0**	**96.8**	**97.1**	**98.4**



(a) Source model only (b) Ours (c) Source model only (d) Ours

Figure 2: t-SNE [36] visualizations on Office-31 [15] dataset when performing D, W→A (a)(b) and A, W→D (c)(d) adaptation tasks.

6 Experiments

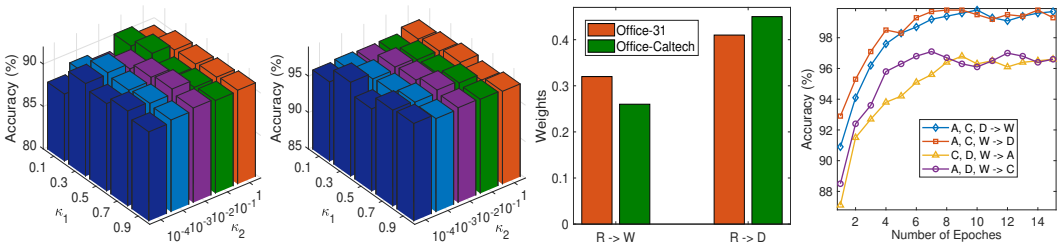
6.1 Datasets and Baseline Methods

Datasets: Office-31 [15] consists of three representative domains with 31 shared object categories in the office environment, *i.e.*, Amazon (A), DSLR (D), and Webcam (W). **Office-Caltech** [11] is an extension dataset of Office-31 [15] by adding an additional subset called Caltech-256 (C) on it and extracting 10 common object classes among them. **Office-Home** [37] is composed of four different domains including Product (Pr), Clipart (Cl), Art (Ar), and Realworld (Re). Each of these subsets consists of 65 shared object categories. **Digits-Five** [30] contains five digit recognition subsets including USPS (UP), SVHN (SV), MNIST-M (MM), MNIST (MT) and Synthetic Digits (SY).

Baseline Methods: To validate the effectiveness of our model, we conduct comparison experiments with a wide array of baseline methods. Specifically, MDAN [51], DCTN [45], M^3SDA [30], MDDA [54] and LtC-MSDA [38] are traditional representative multi-source domain adaptation methods with access to source data. BAIT [47], PrDA [18], SHOT [26] and MA [24] focuses on unsupervised single-source domain adaptation without access to source data. We compare against the multi-source extensions of [18, 24, 26, 47] by taking an average of target soft predictions from all adapted source models. DECISION [1] combines the source adaptation models with suitable weights automatically for multi-source free domain adaptation. Furthermore, Source only denotes the performance of evaluation on target data when combining the rest of source data for training, and Source model only represents the averaged performance over the predictions of all pre-trained source models. Accuracy (%) is considered as basic evaluation metric in our experiments. More details about experimental configuration and comparison results are presented in supplementary material.

6.2 Experiments on Office-31 and Office-Caltech Datasets

Performance Comparisons: The experimental comparisons between our proposed model and other state-of-the-art methods on Office-31 [15] and Office-Caltech [11] datasets are presented in Table 1. We have the following conclusions from the results in Table 1: 1) When compared with the multi-source domain adaptation methods [30, 38, 45, 51, 54] that employ source data for training, our proposed model without access to source data significantly outperforms them by a large margin of 1.4%∼8.3% in terms of mean accuracy. It verifies the superiority of our model to tackle multi-source



(a) Office-31 (b) Office-Caltech (c) Weights of domain A (d) Convergence

Figure 3: Qualitative analysis about parameters $\{\kappa_1, \kappa_2\}$ (a)(b), weights of domain A (c) and convergence on Office-Caltech (d), where R in (c) denotes the rest of domains except for the target.

Table 2: Comparisons between our model and other competing methods on Office-Home [37] dataset.

Methods	Source Data	Ar, Cl, Pr → Re	Ar, Cl, Re → Pr	Ar, Pr, Re → Cl	Cl, Pr, Re → Ar	Avg.
Source only [13]	✓	67.8	71.3	51.8	53.4	61.1
MDAN [51]	✓	77.3	77.6	62.2	65.4	70.6
DCTN [45]	✓	78.7	78.3	63.8	66.4	71.8
M^3SDA [30]	✓	79.4	79.1	63.5	67.2	72.3
MDDA [54]	✓	79.6	79.5	62.3	66.7	71.0
LtC-MSDA [38]	✓	80.1	79.2	64.1	67.4	72.7
Source model only	✗	76.3	78.8	50.1	50.9	64.0
BAIT [47]	✗	77.2	79.4	59.6	71.1	71.8
PrDA [18]	✗	76.8	79.1	57.5	69.3	70.7
SHOT [26]	✗	82.9	82.8	59.3	72.2	74.3
MA [24]	✗	81.7	82.3	57.4	72.5	73.5
DECISION [1]	✗	83.6	84.4	59.4	74.5	75.5
Ours-w/oEnt	✗	82.6	83.0	58.7	74.2	74.6
Ours-w/oDiv	✗	82.1	82.9	58.5	73.8	74.3
Ours-w/oCls	✗	81.4	82.7	57.9	73.1	73.8
Ours-w/oCrc	✗	83.5	84.4	59.7	74.9	75.6
Ours	✗	**84.2**	**84.7**	**60.5**	**75.2**	**76.2**

free domain adaptation. 2) Our model performs better than the multi-source extensions of single-source free domain adaptation methods [18,24,26,47], which validates the effectiveness of our pseudo label generation process. 3) The performance of our model is better than [1] for all evaluation tasks, since the confident anchor-induced pseudo label generator and class relationship-aware consistency loss promote the adaptation performance. Figure 2 shows that our model significantly narrows distribution discrepancy across domains on Office-31 [15] when compared with Source model only.

Ablation Studies: This subsection introduces the effectiveness of each component in our model via ablation studies on Office-31 [15] and Office-Caltech [11] datasets, as shown in Table 1. Ours-w/oEnt, Ours-w/oDiv, Ours-w/oCls and Ours-w/oCrc are the abbreviations of training our proposed model without $\mathcal{L}_{\text{ent}}$, $\mathcal{L}_{\text{div}}$, $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{crc}}$, respectively. When any component of our proposed model is abandoned, the performance degrades 0.4%∼2.3% in terms of average accuracy, which illustrates the rationality and effectiveness of all components to cooperate together. All modules play an indispensable role in improving performance, even though our model has no access to source data.

Parameter Investigation: This subsection investigates the effects of hyper-parameters κ_1 in a range of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and κ_2 in a range of $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ on Office-31 [15] and Office-Caltech [11] datasets, as shown in Figure 3 (a)(b). It validates that our proposed model achieves stable performance over a wide range of hyper-parameters selection. Moreover, the best performance of our proposed model on target domain is obtained when $\kappa_1 = 0.7$ and $\kappa_2 = 10^{-2}$.

Contribution Weights and Convergence Analysis: Figure 3 (c)(d) present the contribution weights of domain A and convergence curves of our proposed model on Office-31 [15] and Office-Caltech [11] datasets. The source-specific transferability perception module could quantify the transferability contributions of complementary information from multiple source domains to target prediction. Furthermore, the accuracy on Office-Caltech [11] dataset converges to a stable value after a few epoch, which demonstrates the convergence effectiveness of our proposed CAiDA model.

6.3 Experiments on Office-Home Dataset

Performance Comparisons: As presented in Table 2, we conduct extensive comparison experiments on Office-Home [37] dataset to validate the effectiveness of our model. We have the following observations from Table 2: 1) Without access to source data, our proposed model significantly outperforms the representative multi-source domain adaptation methods [30, 38, 45, 51, 54] by 3.5%∼5.6% in

Table 3: Comparisons between our model and other competing methods on Digits-Five [30] dataset, where R denotes the rest of four domains except for the single target domain.

Methods	Source Data	R → MM	R → MT	R → UP	R → SV	R → SY	Avg.
Source only [21]	✓	63.4	90.5	88.7	63.5	82.4	77.7
MDAN [51]	✓	69.5	98.0	92.4	69.2	87.4	83.3
DCTN [45]	✓	70.5	96.2	92.8	77.6	86.8	84.8
M^3SDA [30]	✓	72.8	98.4	96.1	81.3	89.6	87.7
MDDA [54]	✓	78.6	98.8	93.9	79.3	89.7	88.1
LtC-MSDA [38]	✓	85.6	99.0	98.3	83.2	93.0	91.8
Source model only	✓	25.2	90.0	93.3	42.8	77.8	65.8
BAIT [47]	✗	87.6	96.2	96.7	60.6	90.5	86.3
PrDA [18]	✗	86.2	95.4	95.8	57.4	84.8	83.9
SHOT [26]	✗	90.4	98.9	97.7	58.3	83.9	85.8
MA [24]	✗	90.8	98.4	98.0	59.1	84.5	86.2
DECISION [1]	✗	93.0	**99.2**	97.8	82.6	97.5	94.0
Ours-w/oEnt	✗	92.1	97.3	96.0	80.7	96.3	92.5
Ours-w/oDiv	✗	91.7	97.0	96.8	82.2	96.5	92.8
Ours-w/oCls	✗	91.3	96.6	96.4	80.5	95.8	92.1
Ours-w/oCrc	✗	92.8	98.2	98.1	82.8	97.7	93.9
Ours	✗	**93.7**	99.1	**98.6**	**83.3**	**98.1**	**94.6**

terms of average accuracy. 2) The confident pseudo label generator encourages our proposed model to perform better than [1], which verifies the superiority of our model for multi-source free adaptation. 3) The performance of our model improves a large margin of 1.9%∼5.5% in terms of mean accuracy, compared with [18, 24, 26, 47]. It validates the efficiency of source-specific transferability perception strategy and class relationship-aware consistency to narrow distribution discrepancy across domains.

Ablation Studies: As introduced in Table 2, we conduct ablation studies of our proposed model on Office-Home [37] dataset to illustrate the rationality of all designed modules. When compared with Ours, the performances of Ours-w/oEnt, Ours-w/oDiv, Ours-w/oCls and Ours-w/oCrc degrades 1.6%, 1.9%, 2.4% and 0.6% in terms of average accuracy, respectively. It validates that all designed modules without access to source data could cooperate well to achieve the best multi-source domain adaptation performance. Furthermore, a confident anchor-induced pseudo label generator could effectively reduce the distribution discrepancy by mining confident pseudo labels for the target domain.

6.4 Experiments on Digits-Five Dataset

This subsection presents the ablation studies and comparison experiments between our proposed model and other competing methods on Digits-Five [30] dataset, as introduced in Table 3. Some conclusions are drawn from the presented results in Table 3: 1) Our proposed model performs the best in terms of average accuracy when compared with multi-source domain adaptation methods [1, 30, 38, 45, 51, 54] and single-source free adaptation methods [18, 24, 26, 47]. The significant performance improvement demonstrates the effectiveness of our model, even without access to source data. 2) Our model outperforms [18,24,26,47] by 8.3%∼10.7% mean accuracy, since it automatically quantifies the contributions of different source domains to target adaptation. When we compare our model with [1, 30, 38, 45, 51, 54], the confident anchor-induced pseudo label generator and class relationship-aware consistency loss facilitate the performance improvement by mining confident pseudo labels and aligning soft confusion matrices across domains. 3) The performance degradation in ablation studies (*i.e.*, Ours-w/oEnt, Ours-w/oDiv, Ours-w/oCls and Ours-w/oCrc) validates that all proposed components are designed effectively and reasonably to explore transferable knowledge.

7 Conclusion

This paper proposes a novel confident anchor-induced multi-source free domain adaptation (CAiDA) model to capture transferable knowledge from multiple sources without access to source data. To be specific, a source-specific transferability perception module is developed to automatically weight the transferability contributions of multiple source domains. Meanwhile, we design a confident anchor-induced pseudo label generator to mine pseudo labels for the target domain by establishing a confident anchor group, and develop a class relationship-aware consistency loss to capture consistent inter-class relationships across domains. Theoretical analysis shows a novel generalization learning bound for MSFDA. Extensive experiments illustrate the superiority of our proposed CAiDA model.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[2] Zechen Bai, Zhigang Wang, Jian Wang, Di Hu, and Errui Ding. Unsupervised multi-source domain adaptation for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[3] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[4] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[5] John Bridle, Anthony Heading, and David MacKay. Unsupervised classifiers, mutual information and 'phantom targets. In J. Moody, S. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1992.

[6] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[7] Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4022–4031, June 2020.

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, Jan. 2016.

[9] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 597–613, Cham, 2016. Springer International Publishing.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[11] Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, page 2066–2073, USA, 2012. IEEE Computer Society.

[12] Jiang Guo, Darsh Shah, and Regina Barzilay. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[14] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998. PMLR, 10–15 Jul 2018.

[16] Han-Kai Hsu, Wei-Chih Hung, Hung-Yu Tseng, Chun-Han Yao, Yi-Hsuan Tsai, Maneesh Singh, and Ming-Hsuan Yang. Progressive domain adaptation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[17] Shatha Jaradat. Deep cross-domain fashion recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, page 407–410, New York, NY, USA, 2017. Association for Computing Machinery.

[18] Youngeun Kim, Donghyeon Cho, Priyadarshini Panda, and Sungeun Hong. Progressive domain adaptation from a source pre-trained model. *arXiv preprint arXiv:2007.01524*.

[19] Andreas Krause, Pietro Perona, and Ryan Gomes. Discriminative clustering by regularized information maximization. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[20] Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 942–950, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[22] Seungmin Lee, Dongwan Kim, Namil Kim, and Seong-Gyun Jeong. Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[23] Keqiuyin Li, Jie Lu, Hua Zuo, and Guangquan Zhang. Multi-source contribution learning for domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2021.

[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] Yitong Li, michael Murias, geraldine Dawson, and David E Carlson. Extracting relationships by multi-domain matching. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[26] 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 (ICML)*, pages 6028–6039, July 13–18 2020.

[27] Chuang Lin, Sicheng Zhao, Lei Meng, and Tat-Seng Chua. Multi-source domain adaptation for visual sentiment classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2661–2668, Apr. 2020.

[28] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2016.

[29] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the rényi divergence. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, pages 367–374. AUAI Press, 2009.

[30] 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 International Conference on Computer Vision*, pages 1406–1415, 2019.

[31] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[32] Paolo Russo, Tatiana Tommasi, and Barbara Caputo. Towards multi-source adaptive semantic segmentation. In Elisa Ricci, Samuel Rota Bulò, Cees Snoek, Oswald Lanz, Stefano Messelodi, and Nicu Sebe, editors, *Image Analysis and Processing – ICIAP 2019*, pages 292–301, Cham, 2019. Springer International Publishing.

[33] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2058–2065. AAAI Press, 2016.

[34] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. In *8th International Conference on Learning Representations (ICLR)*, Apr. 2020.

[35] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, 2017.

[36] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[37] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[38] Hang Wang, Minghao Xu, Bingbing Ni, and Wenjun Zhang. Learning to combine: Knowledge aggregation for multi-source domain adaptation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 727–744. Springer International Publishing, 2020.

[39] Haotian Wang, Wenjing Yang, Zhipeng Lin, and Yue Yu. Tmda: Task-specific multi-source domain adaptation via clustering embedded adversarial training. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1372–1377, 2019.

[40] Ximei Wang, Ying Jin, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Transferable normalization: Towards improving transferability of deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[41] Fangzhao Wu and Yongfeng Huang. Sentiment domain adaptation with multiple sources. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 301–310, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.

[42] Xi-Zhu Wu, Wenkai Xu, Song Liu, and Zhi-Hua Zhou. Model reuse with reduced kernel mean embedding specification. *CoRR*, abs/2001.07135, 2020.

[43] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *Advances in Neural Information Processing*

Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 6835–6846, 2019.

[44] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 6502–6509. AAAI Press, 2020.

[45] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[46] Luyu Yang, Yogesh Balaji, Ser-Nam Lim, and Abhinav Shrivastava. Curriculum manager for source selection in multi-source domain adaptation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 608–624, Cham, 2020. Springer International Publishing.

[47] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Unsupervised domain adaptation without source data by casting a bait. *arXiv preprint arXiv:2010.12427*, 2020.

[48] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (CMD) for domain-invariant representation learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[49] Jun Zhang, Weien Zhou, Xianqi Chen, Wen Yao, and Lu Cao. Multisource selective transfer framework in multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, 24(3):424–438, 2020.

[50] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5031–5040, 2019.

[51] Han Zhao, Shanghang Zhang, Guanhang Wu, José M. F. Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[52] Sicheng Zhao, Bo Li, Colorado Reed, Pengfei Xu, and Kurt Keutzer. Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169*, 2020.

[53] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[54] Sicheng Zhao, Guangzhi Wang, Shanghang Zhang, Yang Gu, Yaxian Li, Zhichao Song, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source distilling domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12975–12983, Apr. 2020.

[55] Yukun Zuo, Hantao Yao, and Changsheng Xu. Attention-based multi-source domain adaptation. *IEEE Transactions on Image Processing*, 30:3793–3803, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the limitations in Section 4. To make sure our problem is solvable from the mathematical perspective, some mild assumptions are necessary. Though we have some assumptions for our theory, those assumptions are commonly used in other important fields, such as meta learning, domain generalization and noisy learning. In addition, as we have mentioned, our assumptions are much weaker than previous existing works.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] One of the aims of our problem is to address the privacy preservation issues. Our work is positive for society. We do not think our work has negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 4.
   (b) Did you include complete proofs of all theoretical results? [Yes] See the supplementary material.
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We have introduced detailed instructions to reproduce the main experimental results in our paper and supplemental material, and will release the code after this paper has been accepted.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5 and the supplementary material A.1.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See the supplementary material A.1.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the supplementary material.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] See Section 6.

(b) Did you mention the license of the assets? [Yes] See Section 6.

(c) Did you include any new assets either in supplemental material or as a URL? [Yes]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 6.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 6.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]