Agile Multi-Source-Free Domain Adaptation

Xinyao Li[1], Jingjing Li[1,2*], Fengling Li[3], Lei Zhu[4], Ke Lu[1]

[1]University of Electronic Science and Technology of China (UESTC)
[2]Shenzhen Institute for Advanced Study, UESTC
[3]University of Technology Sydney
[4]School of Electronic and Information Engineering, Tongji University
xinyao326@outlook.com, lijin117@yeah.net, {fenglingli2023, leizhu0608}@gmail.com, kel@uestc.edu.cn

Abstract

Efficiently utilizing rich knowledge in pretrained models has become a critical topic in the era of large models. This work focuses on adaptively utilizing knowledge from multiple source-pretrained models to an unlabeled target domain without accessing the source data. Despite being a practically useful setting, existing methods require extensive parameter tuning over each source model, which is computationally expensive when facing abundant source domains or larger source models. To address this challenge, we propose a novel approach which is free of the parameter tuning over source backbones. Our technical contribution lies in the Bi-level ATtention ENsemble (Bi-ATEN) module, which learns both intra-domain weights and inter-domain ensemble weights to achieve a fine balance between instance specificity and domain consistency. By slightly tuning source bottlenecks, we achieve comparable or even superior performance on a challenging benchmark DomainNet **with less than 3% trained parameters and 8 times of throughput compared with SOTA method**. Furthermore, with minor modifications, the proposed module can be easily equipped to existing methods and gain more than 4% performance boost. Code is available at https://github.com/TL-UESTC/Bi-ATEN.

Introduction

Large-scale models have drawn significant attention for their remarkable performance across a spectrum of applications (Ramesh et al. 2022; Irwin et al. 2022; Lee et al. 2020). Considering that training large models from scratch requires tremendous computational costs, fine-tuning has become a predominant approach to transfer knowledge from large pretrained models to downstream tasks (Long et al. 2015; Guo et al. 2020). However, this paradigm heavily relies on labeled training data and suffers from significant performance decay when target data exhibits distribution shift from pretraining data (Ben-David et al. 2010). Moreover, we usually have multiple pretrained models trained on different sources or architectures on hand, e.g., medical diagnostic models trained on distinct regions or patient groups. Demands to maximally utilizing knowledge from multiple pretrained models are common in real world applications. To this end, Multi-Source-Free Domain Adaptation (MSFDA) (Ahmed et al. 2021; Dong et al. 2021) emerges as a promising technique to address these challenges by enabling holistic adaptation of multiple pretrained source models to an unlabeled target domain, while not accessing source training data.

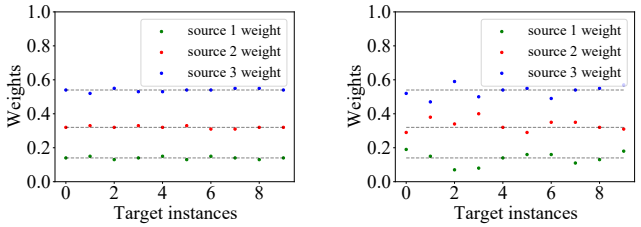
Method	Param.	Backbone	Acc.	Throughput
CAiDA	120.2M	ResNet50	46.8	91
PMTrans	447.4M	Swin	59.1	46
ATEN (ours)	4.9M	Swin	59.1	970
Bi-ATEN (ours)	10.6M	Swin	59.6	369

Table 1: Computation overhead and performance comparison between different methods on DomainNet.

Existing MSFDA methods (Ahmed et al. 2021; Dong et al. 2021; Han et al. 2023; Shen, Bu, and Wornell 2023) typically tackle the problem via a two-step framework, i.e., *(1)* Tune each source model thoroughly towards target domain, and *(2)* Learn source importance weights to assemble the source models. However, their overwhelming limitations in computational efficiency and scalability prevent their applications on large-scale problems. For *step (1)*, the number of models to tune increases linearly along with the number of source domains, which could become unacceptable for large-scale problems with abundant source domains. The necessity of tuning all parameters for each model also makes it infeasible to scale up these methods to larger models. In Table 1 we compare the performance and trainable parameters of CAiDA (Dong et al. 2021), PMTrans[1] (Zhu, Bai, and Wang 2023) and our methods on a challenging benchmark DomainNet (Peng et al. 2019) with 6 domains. As a typical MSFDA framework, CAiDA performs poorly due to limited performance of ResNet-50 (He et al. 2016) backbone. By equipping a stronger backbone SwinTransformer (Liu et al. 2021), a potential performance boost of +12.3% is achieved at a cost of four times of parameters to tune. On the other hand, we aim to achieve superior performance by equipping SwinTransformer while demanding significantly less training cost, presenting a more feasible and agile solution for MSFDA on large models. For *step (2)*, current MSFDA methods learn **domain-level** ensemble weights, ap-

*Corresponding author

[1]PMTrans is a single-source domain adaptation method and we evaluate it on MSFDA setting by taking its single-best results.



(a) without instance specificity (b) with instance specificity

Figure 1: Illustration of instance specificity and domain consistency. Dots are weights assigned to each target sample.

plying identical ensemble strategy across all target instances. Although the learned weights are intuitively interpretable in terms of domain transferablity, they unavoidably introduce misalignment and bias at instance-level. This controversy inherently introduces a trade-off between instance specificity and domain consistency of ensemble weights, which has not been well exploited by existing methods.

Recent success of model ensemble methods (Shu et al. 2021, 2022) suggests that it is effective to transfer knowledge by designing adaptive ensemble weights. While optimal strategies are hard to learn (Mohammed and Kora 2023), we resort to slight tuning of several domain-specific bottleneck layers, costing less than 0.1% of tuning the whole model. As stated above, the key to designing effective weights is to exploit both domain-level transferabilities and instance-level individual characteristics, as illustrated by Fig. 1. Existing MSFDA methods learn weights solely from feature representations, neglecting the potential transferability mismatch between features and outputs, i.e., transferable target features do not always lead to accurate predictions. To address this issue, we propose to introduce additional semantic information from classifiers for deriving weights. For each feature representation, we first learn *intra-domain weights* to mitigate transferability mismatch by finding the most compatible classifier that produces unbiased outputs. With unbiased outputs from the selected source classifier, we further learn *inter-domain ensemble weights* that combine source outputs into the final result. We propose a novel Bi-level ATtention ENsemble (Bi-ATEN) to effectively learn the two weights through attention mechanisms. Bi-ATEN is capable of tailoring its ensemble decisions to the particularities of each instance, while maintaining the broader transferability trends that are consistent across domains. This balance is essential for accurate domain adaptation, where a model needs to leverage domain-specific knowledge without losing the overarching patterns that drive adaptation.

The proposed Bi-ATEN can be simplified into inter-domain ATtention ENsemble (ATEN) and plugged into existing MSFDA methods by replacing their weight-learning module. Although leaning towards domain consistency in the specificity-consistency balance, ATEN still exhibits clear performance boost over baseline methods, proving the efficacy of our design. In a nutshell, we achieve adaptation primarily by assuring instance specificity and domain consistency along with slight tuning of bottlenecks. Table 1 provides comprehensive comparison between our methods and existing methods. Our contributions can be summarized as: (1) We propose a novel framework to agilely handle MSFDA by learning fine-grained domain adaptive ensemble strategies. (2) We design an effective module Bi-ATEN that learns both intra-domain weights and inter-domain ensemble weights. Its light version ATEN can be equipped to existing MSFDA methods to boost performance. (3) Our method significantly reduces computational costs while achieving state-of-the-art performance, making it feasible for real-life transfer applications with large source-trained models. (4) Extensive experiments on three challenging benchmarks and detailed analysis demonstrates the success of our design.

Related Work

Source-free domain adaptation (SFDA) assumes no labeled source data but a source-trained model is available for adaptation (Li et al. 2021a). SHOT (Liang, Hu, and Feng 2020) pioneers the problem by proposing a clustering algorithm for pseudo-labeling and utilizes information maximization loss. Several works (Li et al. 2020; Yang et al. 2021) follow the research line to improve or develop new clustering methods. Kundu et al. (2022) reveal insight on discriminability and transferability trade-offs and propose to mix-up original and corresponding translated generic samples to improve performance. Other relevant settings including source-free active domain adaptation (Li et al. 2022) and imbalanced SFDA (Li et al. 2021b) have also been explored.

Multi-source domain adaptation (MSDA) assumes that labeled source data from multiple domains are available, and tries to transfer simultaneously towards target domain with theoretical guarantees from pioneering works (Ben-David et al. 2010; Crammer, Kearns, and Wortman 2008). M^{3}SDA (Peng et al. 2019) provides theoretical insights that all source-target and source-source pairs should be aligned to achieve adaptation. DRT (Li et al. 2021c) proposes a dynamic module that adapts model parameters according to samples. ABMSDA (Zuo, Yao, and Xu 2021) proposes a Weighted Moment Distance to ensure higher attention among more related domains. STEM (Nguyen et al. 2021) generates a teacher-student framework to close the gap between source and target distributions.

Multi-source-free domain adaptation (MSFDA) combines SFDA and MSDA, aiming to learn optimal source model combinations that perform best on unlabeled target data. DECISION (Ahmed et al. 2021) first explores the problem and proposes to assemble source outputs with learnable weights while updating source models via weighted information maximization. CAiDA (Dong et al. 2021) proposes to use a similar framework but with a confident-anchor-induced pseudo label generator. Shen, Bu, and Wornell (2023) develop a generalization bound on MSFDA that reveals an inherent bias-variance trade-off. A hierarchical framework is further proposed to balance the trade-off. DATE (Han et al. 2023) evaluates source transferabilities via a Bayesian perspective before quantifying the similarity degree by a multi-layer perception. All forementioned methods learn domain-level importance regardless of instance characteristics, which unavoidably limits their performance.

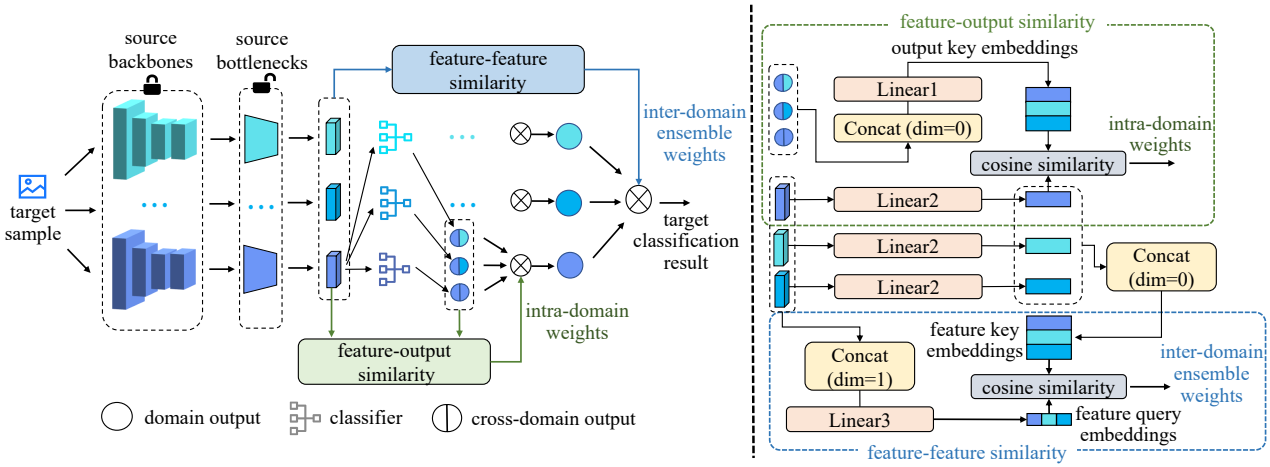


Figure 2: Framework of our method. Different colors represent different source domains. For cross-domain outputs, colors on the left semicircles represent domains of bottleneck features while that on the right semicircles represent domains of classifiers that generate the cross-domain output. Best viewed in color.

Method

Problem Definition

Assume we have n source-trained models $\{h_s^i\}_{i=1}^n$ for C-category classification task. Given an unlabeled target domain $\{X_t\}$ with identical categories, the goal is to optimize all n source models towards satisfactory performance on the target domain. Following (Tzeng et al. 2014), a bottleneck layer k_s with parameter θ_{k_s} is applied after the feature extractor f_s with parameter θ_{f_s}, and before the final fully-connected classifier g_s with parameter θ_{g_s}. Given a target sample x_t, we define its bottleneck feature with d_k dimensions produced by source model h_s^i as $\phi_t^i = (k_s^i \circ f_s^i)(x_t)$, and the output of source model h_s^i can be denoted as $y_t^i = g_s^i(\phi_t^i)$. Specifically, in this paper we consider cross-domain outputs obtained by forwarding ϕ_t^i through a classifier from another domain j, i.e., $y_t^{ij} = g_s^j(\phi_t^i)$. By learning intra-domain weights $\boldsymbol{\alpha^i}$, unbiased domain output for feature ϕ_t^i is denoted as $\tilde{y}_t^i = \sum_{j=1}^n \alpha_j^i y_t^{ij}$. Inter-domain ensemble weights $\boldsymbol{\beta}$ are further learned to obtain final output $\ddot{y}_t = \sum_{i=1}^n \beta_i \tilde{y}_t^i$. Our goal is to learn optimal $\{\boldsymbol{\alpha^i}\}_{i=1}^n$, $\boldsymbol{\beta}$ and bottleneck parameters θ_{k_s} that minimizes training loss.

Overview

Fig. 2 depicts our framework. A target sample is forwarded through the source models to extract the bottleneck features. Instead of directly generating outputs by specific source classifier, we compute all possible cross-domain outputs with respect to current feature by forwarding it through all source classifiers. Intra-domain weights $\{\boldsymbol{\alpha^i}\}_{i=1}^n$ are computed between the feature representation and all output vectors for obtaining unbiased outputs. Subsequently, inter-domain weights $\boldsymbol{\beta}$ are learned to assemble the unbiased domain outputs into the final classification result. Note that both source backbones and source classifiers remain frozen during the entire training process. Laying at the core of the framework is the Bi-ATEN module, as depicted on the right of Fig. 2. It simultaneously learns $\{\boldsymbol{\alpha^i}\}_{i=1}^n$ from feature-output similarities and $\boldsymbol{\beta}$ from feature-feature similarities. Next we elaborate on the detailed design of each module.

Bi-level Attention Ensemble

Intra-domain weights. All current MSFDA methods adopt an end-to-end training paradigm that treats each source model as a whole (Dong et al. 2021). However, the distribution shifts between target and source data can lead to mismatches within the source model components like bottlenecks and classifiers. Inspired by deep model reassembly methods (Yang et al. 2022), we propose to improve current MSFDA paradigms by performing a partial model reassembly. We explore compatible bottleneck-classifier pairs tailored towards target data characteristics, and obtain the reassembled result by summing over weighted cross-domain outputs of bottleneck-classifier pairs. Given bottleneck feature from the i_{th} source domain $\phi_t^i \in \mathbb{R}^{d_k}$, we first obtain its cross-domain outputs by:

$$O_t^i = Concat(\{\theta_{g_s}^j \phi_t^i\}_{j=1}^n, \text{ dim} = 0), \tag{1}$$

where $O_t^i \in \mathbb{R}^{n \times C}$ is cross-domain output matrix for the i_{th} feature. Since source classifier parameters are fixed, our aim to find the most compatible classifier can be converted to finding the most similar output vector after classification linear transformation θ_{g_s}. We adopt cosine similarity to eliminate norm mismatch between features and outputs:

$$Sim_t^i = Cosine(\phi_t^i W^F, O_t^i W^O), \tag{2}$$

where $Sim_t^i \in \mathbb{R}^n$ is similarity vector, $W^F \in \mathbb{R}^{d_k \times d_{emb}}$ (Linear2 in Fig. 2) and $W^O \in \mathbb{R}^{C \times d_{emb}}$ (Linear1 in Fig. 2) are linear transforms that transform feature and output into the same embedding dimension d_{emb}. Then, intra-domain weights are obtained by applying softmax operation over the similarity vector:

$$\boldsymbol{\alpha^i} = Softmax(Sim_t^i). \tag{3}$$

Finally, assembled output for domain i is obtained by:

$$\tilde{y}_t^i = \sum_{j=1}^{n} \alpha_j^i \theta_{g_s}^j \phi_t^i. \tag{4}$$

We regard output $\tilde{y}_t^i$ as unbiased if it is: (1) Confident. Ambiguous outputs imply multiple possible interpretations on the feature, increasing the risk of feature-output mismatch. (2) Diverse. Overly consistent classification results lead to mode collapse where certain classes are rarely considered. We apply IM loss (Liang, Hu, and Feng 2020), a base component shared by current MSFDA methods, to assure unbiased intra-domain ensemble:

$$\mathcal{L}_{intra} = \sum_{i=1}^{n} \mathcal{L}_{IM}(Softmax(\tilde{y}_t^i)), \tag{5}$$

where $\mathcal{L}_{IM}$ is defined as:

$$\mathcal{L}_{IM}(y) = \mathcal{L}_{ent}(y) - \mathcal{L}_{div}(y), \text{where} \tag{6}$$

$$\mathcal{L}_{ent}(y) = -\mathbb{E}_{x_t \in X_t} \left[\sum_{c=1}^{C} \delta_c(y) \log \delta_c(y) \right],$$

$$\mathcal{L}_{div}(y) = -\sum_{c=1}^{C} \bar{p}_c \, log \, \bar{p}_c,$$

where $\bar{p}_c = -\mathbb{E}_{x_t \in X_t} \delta_c(y)$ and $\delta_c(\cdot)$ takes the c_{th} logit.

Inter-domain weights. We derive ensemble weights from bottleneck features. Motivated by the success of attention mechanism (Vaswani et al. 2017), we obtain inter-domain weights by computing attention between different linear representations of bottleneck features. To allow intra-domain adjustments according to inter-domain weights, the transform matrix W^F is shared with that in Eq. (2):

$$\hat{\phi}_t^K = Concat(\{\phi_t^i W^F\}_{i=1}^n, \ \dim = 0). \tag{7}$$

For query embeddings, features are first concatenated before linearly transformed:

$$\hat{\phi}_t^Q = Concat(\{\phi_t^i\}_{i=1}^n, \ \dim = 1) W^{QF}, \tag{8}$$

where $W^{QF} \in \mathbb{R}^{(nd_k) \times d_{emb}}$ is the query transform matrix (Linear3 in Fig. 2). Similar to intra-domain weights, we compute inter-domain weights via:

$$\boldsymbol{\beta} = Softmax(Cosine(\hat{\phi}_t^Q, \hat{\phi}_t^K)). \tag{9}$$

Final ensemble result is then obtained by:

$$\ddot{y}_t = \sum_{i=1}^{n} \beta_i \tilde{y}_t^i. \tag{10}$$

Apart from being confident and diverse, the final ensemble result should more importantly be correct. Since no label is available, in this work we adopt a **dynamic-cluster-based strategy** to provide pseudo labels for classification. The dynamic is two-fold: dynamic feature combinations and dynamic centroids for each instance. We first compute centroid for class c generated by source model h_s^i by:

$$\mu_c^i = \frac{\sum_{x_t \in X_t} \delta_c(Softmax(\ddot{y}_t)) \phi_t^i}{\sum_{x_t \in X_t} \delta_c(Softmax(\ddot{y}_t))}, \tag{11}$$

where $\delta_c(\cdot)$ takes the c_{th} logit. Dynamic centroid for the m_{th} target sample x_t^m of class c is computed by assembling all centroids using instance-specific inter-domain weight $\boldsymbol{\beta^m}$:

$$\tilde{\mu}_c^m = \sum_{i=1}^{n} \beta_i^m \mu_c^i. \tag{12}$$

For target samples, their feature representations are dynamically obtained by assembling all source bottleneck features:

$$\tilde{\phi}_t^m = \sum_{i=1}^{n} \beta_i^m \phi_t^{mi}, \tag{13}$$

where ϕ_t^{mi} is bottleneck feature extracted by source model from domain i for sample x_t^m. Finally, we generate pseudo label for x_t^m by:

$$y_t = \arg\max_c \, Cosine(\tilde{\phi}_t^m, \ \tilde{\mu}_c^m). \tag{14}$$

Dynamic clustering greatly extends the diversity and flexibility of generated pseudo labels. As Bi-ATEN becomes more reliable, quality of pseudo labels is concurrently improved, which in turn helps the training of Bi-ATEN. With pseudo labels, objective for final output is formulated as:

$$\mathcal{L}_{inter} = \gamma CE(\ddot{y}_t, y_t) + \mathcal{L}_{IM}(Softmax(\ddot{y}_t)), \tag{15}$$

where γ is a hyperparameter and $CE(\cdot)$ is cross entropy loss with label smoothing (Szegedy et al. 2016). Overall objective is given as:

$$\mathcal{L} = \mathcal{L}_{inter} + \lambda \mathcal{L}_{intra}, \tag{16}$$

where λ is a trade-off hyperparameter. We train our model by solving the following optimization problem:

$$\boldsymbol{\alpha}, \ \boldsymbol{\beta}, \ \theta_{k_s} = \arg\min \ \mathcal{L}. \tag{17}$$

Attention Ensemble as a Pluggable Module

Consider an extreme situation where $\boldsymbol{\alpha^i}$ contains a single one at the i_{th} location and zeros elsewhere. It simplifies Bi-ATEN to ATEN with only inter-domain ensemble weights $\boldsymbol{\beta}$, which aligns with weight learning paradigm of existing MSFDA methods, and can therefore replace their weight learning module easily. Assume objective of the original MSFDA method as $\mathcal{L}_{origin}$, the optimization goal after equipping ATEN becomes:

$$\boldsymbol{\beta}, \ \theta_{k_s}, \ \theta_{f_s} = \arg\min \ \mathcal{L}_{origin}. \tag{18}$$

$\boldsymbol{\alpha^i}$s are fixed as one-hot vectors as described above, thus saving the training of W^O.

Training Process

We design an alternate training procedure for Bi-ATEN. We observe that for target domains with relatively smaller domain gap, the domain-specific source classifiers already show satisfactory performance, while for those with larger domain gap, intra-domain weights are vital for adaptive feature-classifier matching. Considering both cases, in certain epochs we manually set $\boldsymbol{\alpha^i}$ to one-hot vectors as in ATEN. Different from Eq. (18), we still update W^O via Eq. (5). Such alternate training utilizes the benefits of both strategies, striking a balance between intra-domain compatibility and domain-consistent adaptation.

Method	SF	Backbone	→clp	→inf	→pnt	→qdr	→rel	→skt	Avg.	Param.	Train time
M^3SDA	×	ResNet101	58.6	26.0	52.3	6.3	62.7	49.5	42.6	42.48M	/
LtC-MSDA	×		63.1	28.7	56.1	16.3	66.1	53.8	47.4	42.50M	/
STEM	×		72.0	28.2	61.5	25.7	72.6	60.2	53.4	43.78M	/
DRT	×		71.0	31.6	61.0	12.3	71.4	60.7	51.3	60.90M	/
DECISION	√	ResNet50	61.5	21.6	54.6	18.9	67.5	51.0	45.9	120.14M	2.9H
DATE	√		61.2	22.7	53.5	18.1	69.8	50.9	46.0	/	/
CAiDA	√		63.6	20.7	54.3	19.3	71.2	51.6	46.8	120.20M	3.0H
Surrogate	√		66.5	21.6	56.7	20.4	70.5	54.4	48.4	/	/
TransMDA	√		71.7	29.0	61.4	18.6	74.1	60.9	52.6	/	/
CDTrans-best	×	DeiT-base	69.0	31.0	61.5	27.2	72.6	58.1	53.2	428.23M	/
SSRT-best	×	ViT-base	70.6	37.1	66.0	21.7	75.8	59.8	55.2	442.74M	/
DRT	×	SwinTransformer	74.6	33.2	64.8	20.3	76.4	64.6	55.6	91.43M	/
AVG-ENS	√		74.1	35.3	66.1	15.0	81.6	62.9	55.8	/	/
PMTrans-best	×		74.1	35.3	**70.7**	**30.9**	79.8	63.7	59.1	447.43M	/
ATEN (ours)	√		76.6	37.2	68.6	24.0	83.5	64.6	59.1	**4.92M**	**0.6H**
Bi-ATEN (ours)	√		**77.0**	**38.5**	68.6	25.0	**83.6**	**64.9**	**59.6**	10.56M	1.2H

Table 2: Results on DomainNet. SF denotes whether the method follows source-free setting. Best results are in bold font.

Experiments

In this section we present main results and further analysis. Implementations are based on MindSpore and PyTorch.

Datasets and Baselines

Datasets. We evaluate our method on three MSFDA benchmarks Office-Home (Venkateswara et al. 2017), Office-Caltech (Gong et al. 2012) and DomainNet (Peng et al. 2019). Office-Home is divided into 65 categories with 4 domains Art, Clipart, Product and RealWorld. Office-Caltech is extended from Office31 (Saenko et al. 2010) by adding Caltech (Griffin, Holub, and Perona 2007) as a fourth domain. DomainNet is composed of 0.6 million samples from six distinct domains, each containing 345 categories.

Baselines. On Office-Home and Office-Caltech we validate boost obtained by equipping ATEN to existing MSFDA methods: DECISION (Ahmed et al. 2021), CAiDA (Dong et al. 2021), DATE (Han et al. 2023), and compare with other MSDA methods including M^3SDA (Peng et al. 2019), LtC-MSDA (Wang et al. 2020), MA (Li et al. 2020), NRC (Yang et al. 2021) and SHOT (Liang, Hu, and Feng 2020). Baseline results of forementioned methods are cited from DATE. On DomainNet we compare our ATEN and Bi-ATEN against various competing baselines implemented on various backbones. ResNet101 (He et al. 2016): M^3SDA, LtC-MSDA, STEM (Nguyen et al. 2021) and DRT (Li et al. 2021c). ResNet50 (He et al. 2016): DECISION, CAiDA, DATE, Surrogate (Shen, Bu, and Wornell 2023) and TransMDA (Li and Wu 2023). DeiT (Touvron et al. 2021): CDTrans (Xu et al. 2021). ViT (Dosovitskiy et al. 2020): SSRT (Sun et al. 2022). SwinTransformer (Liu et al. 2021): PMTrans (Zhu, Bai, and Wang 2023) and DRT implemented by ourselves.

Main Results

DomainNet. Table 2 illustrates classification accuracies on the DomainNet dataset. Note that methods end with -best are originally single-source domain adaptation approaches, and we select their single-best results on each target domain for fair comparison. AVG-ENS is a naive ensemble strategy by averaging over outputs from all source models, and is listed as a baseline. The results show that our Bi-ATEN achieves superior performance on most of the tasks, except for domains **pnt** and **qdr** we are behind PMTrans. This is because PMTrans has access to labeled source data, which helps to overcome the large domain gaps in DomainNet by distribution alignment. Bi-ATEN exhibits clear enhancements than ATEN, especially on the two most challenging tasks **inf** and **qdr**. Under such significant domain shift, bottleneck-classifier pairs learned by Bi-ATEN show better compatibility. The Train time column compares training time among available source-free methods on target **clp**. Our methods achieve higher accuracy in considerably less training time. The Param. column compares trainable parameters of existing open-source methods. Source-free methods train more parameter as they tune all source models. Larger transformer-based backbones also require heavy computation overheads. Our methods require significantly less trainable parameters to surpass all competing method while following source-free setting, demonstrating the efficacy and agility of our methods. Another key observation is that all existing MSFDA methods are implemented on ResNet50 backbone due to high computational complexities, largely limiting their performance. Our Bi-ATEN stands out as the first MSFDA method that introduces large models like SwinTransformer as backbone while maintaining a surprisingly low computation cost. Notably, our Bi-ATEN achieves a remarkable performance improvement of 7% over the current SOTA in MSFDA task, and additionally demonstrates comparable or even superior performance with source-available domain adaptation methods, strongly supporting the validity and efficiency of the proposed method.

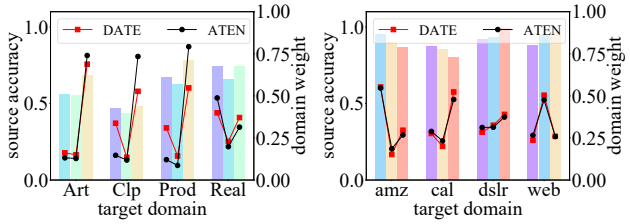
Office-Home and Office-Caltech. Table 3 gives performance improvements obtained by plugging ATEN to ex-

Method	SF	Office-home					Office-Caltech				
		→Art	→Clp	→Prod	→Real	Avg.	→amazon	→caltech	→dslr	→webcam	Avg.
M^{3}SDA	×	67.2	63.5	79.1	79.4	72.3	94.5	92.2	99.2	99.5	96.4
LtC-MSDA	×	67.4	**64.1**	79.2	80.1	72.7	93.7	95.1	99.7	99.4	97.0
MA	√	72.5	57.4	81.7	82.3	73.5	95.7	95.6	97.2	**99.8**	97.1
NRC	√	72.7	58.1	82.3	82.1	73.8	**95.9**	94.9	97.5	99.3	96.9
SHOT	√	72.2	59.3	82.9	82.8	74.3	95.7	95.8	96.8	99.6	97.0
DECISION	√	73.3	58.7	82.9	84.0	74.7	95.6	95.4	96.8	99.3	96.8
+ATEN (ours)	√	76.3	60.6	84.5	83.7	76.3	95.8	96.0	**100.0**	99.7	97.9
CAiDA	√	70.3	55.0	83.0	80.7	72.2	95.2	95.6	98.1	99.7	97.1
+ATEN (ours)	√	76.1	60.3	85.1	83.5	76.3	**95.9**	**96.3**	**100.0**	99.7	**98.0**
DATE	√	75.2	60.9	**85.2**	84.0	76.3	95.6	95.7	98.1	**99.8**	97.3
+ATEN (ours)	√	**76.7**	61.6	**85.2**	**84.7**	**77.1**	**95.9**	95.7	**100.0**	99.7	97.8

Table 3: Results on Office-Home and Office-Caltech. The '+ATEN' rows show improvements obtained by plugging ATEN into original methods. SF denotes whether the method follows source-free setting. Best results are in bold font.

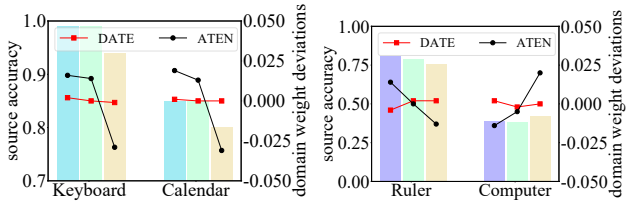
Method	→clp	→inf	→skt	Avg.
Bi-ATEN (ours)	**77.0**	38.5	**64.9**	**60.1**
ATEN (ours) (w/o intra-domain weights)	76.6	37.2	64.6	59.5
w/o alternate training	75.8	**38.6**	64.1	59.5
w/o $\mathcal{L}_{intra}$	76.1	35.8	63.4	58.4
w/o $\mathcal{L}_{IM}$	75.8	38.5	63.6	59.3

Table 4: Ablation study on three tasks from DomainNet. Best results are in bold font.



(a) Office-Home (b) Office-Caltech

Figure 3: Domain-level inter-domain weight comparison. Bars represent source-only accuracies of source models. Lines represent averaged weights assigned to each source.



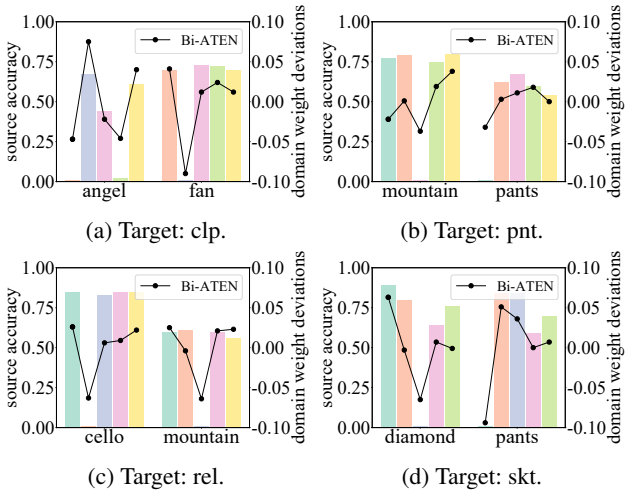
(a) Target: Art. (b) Target: Clp.

Figure 4: Class-level inter-domain weight comparison on Office-Home. Bars represent source accuracy. Lines represent weight deviations assigned to each source output.

isting MSFDA methods. Results show that computing ensemble weights by ATEN brings a maximal 4.1% overall accuracy boost and hardly any negative effects. The combination DATE+ATEN achieves the best accuracy on Office-Home with +0.8% improvement while more significant boost can be observed on baselines DECISION and CAiDA. On Office-Caltech, CAiDA+ATEN achieves the highest accuracy of 98%, approaching fully-supervised performance. We notice that accuracies obtained by plugging in ATEN tend to be similar within the same dataset despite the varying baseline performance. This phenomenon indicates that ATEN is able to learn stable ensemble strategies disregarding potential perturbations from origin method, which guarantees fair performance and steady improvements on various baselines. The experiment provides compelling evidence that ATEN is not only effective with fixed backbones but also offers promising enhancements when applied to existing MSFDA methods, suggesting that learning ensemble weights through our ATEN is beneficial.

Analytical Experiments

Ablation study. Table 4 presents ablation study by removing different modules in our framework, where w/o $\mathcal{L}_{IM}$ is to remove the IM loss in Eq. (15). It can be concluded that all modules contribute positively to our method, and the complete framework Bi-ATEN achieves the best overall accuracy. The alternate training procedure aims to balance the adaptation performance under both small and large distribution shift by focusing on domain specific bottleneck-classifier pairs in certain epochs. However, this procedure could harm the learning of intra-domain weights under significant domain shift as in task →**inf**. Therefore, removing alternate training can lead to slight accuracy increase in these challenging tasks. Removing $\mathcal{L}_{intra}$ brings the largest performance decay, suggesting that learning inappropriate intra-domain weights can harm final outcomes. The IM loss is more effective in easier tasks (→clp, →skt) where well-classified classes might mislead similar classes. On hard



(a) Target: clp. (b) Target: pnt.

(c) Target: rel. (d) Target: skt.

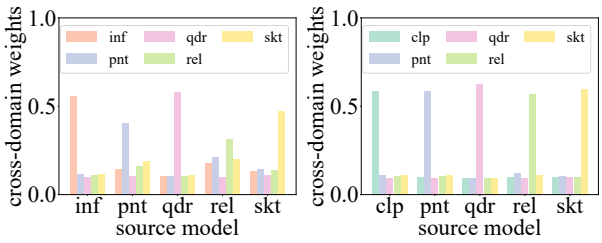
Figure 5: Class-level inter-domain weights on DomainNet. Bars represent source accuracies and lines represent domain weight deviations assigned to each source output.

tasks (→inf) where most samples are misclassified, mode collapse rarely occurs thus IM loss is less effective.

Weight analysis. We present a comprehensive analysis on the two types of weights learned in our framework. Fig. 3 shows that domain-level weights learned by ATEN aligns well with source model transferabilities and accuracies, and this similarity is comparable to that achieved by DATE. This demonstration emphasizes that ATEN effectively learns domain-consistent inter-domain ensemble weights.

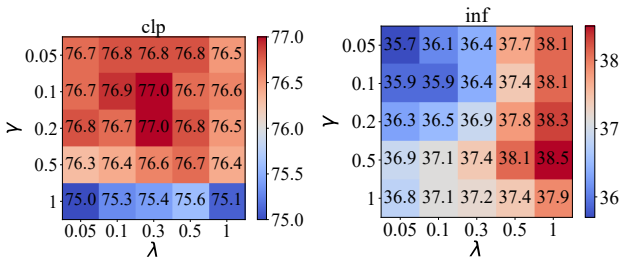
Limited flexibility of identical inter-domain ensemble weights prevent them from accommodating special instances with unique transfer characteristics, ultimately leading to a decline in performance. Our method addresses this by learning tailored inter-domain weights. We examine classes instead of instances for the sake of brevity. Fig. 4 represents how the class-level inter-domain weights deviates from domain-level weights, showcasing their ability to dynamically adapt to different classes that require distinct transferabilities. In contrast to DATE, which shows limited class-level adaptability, ATEN demonstrates its ability to learn individualized and effective strategies by striving to derive suitable weights customized for each class. However, without intra-class weights, this customization is limited, as the deviations are relatively subtle in Fig. 4. Fig. 5 provides the results on DomainNet of our full design. Under more significant transferability gap, Bi-ATEN is still able to adapt intelligently to source models with zero transferability by actively reducing their corresponding weights to prevent negative transfer. The tailored weights are deviated more significantly with the help of intra-domain weights. The collaborative evidence presented in Fig. 3, Fig. 4 and Fig. 5 strongly supports that our method indeed learns weights that are specific to instances and consistent on domains.

Intra-domain weights learned by Bi-ATEN are presented in Fig. 6. Each group of weights are corresponding intra-domain weights $\boldsymbol{\alpha}^i$ for the source bottleneck feature. It can



(a) Target: clp. (b) Target: inf.

Figure 6: Intra-domain weights on DomainNet. Bars represent intra-domain weights assigned to each source classifier.



(a) Target: clp. (b) Target: inf.

Figure 7: Hyperparameter analysis on DomainNet. Numbers represent overall accuracy obtained by each hyperparameter combination.

be seen that the classifiers from the same domain as bottleneck features receive the majority of attention. However, this attention can also dynamically match more compatible target domains, as exemplified in source **rel** of Fig. 6a.

Hyperparameter analysis. Fig. 7 gives accuracies under different hyperparameters in Eq. (15) and Eq. (16). Results show that a large γ harms performance, which suggests that overly relying on pseudo labels misguides the weight learning process. For target domains with larger domain gap (target **inf**), a larger λ is needed to constrain the intra-domain weights to avoid negative transfer, as stated in ablation study. Optimal parameter combinations might vary across different target data, but the overall performance is relatively stable.

Conclusion

This research aims to address the high computation costs associated with existing MSFDA methods. We present a novel framework that prioritizes the learning of instance-specific and domain-consistent ensemble weights, instead of extensively tuning each source model. We achieve this by designing a novel bi-level attention module that effectively learns intra-domain and inter-domain weights. Extensive experiments demonstrate that our methods significantly outperform state-of-the-art methods while requiring considerably lower computation costs. We believe that our work has the potential to encourage the exploration of more light-weight approaches to address the challenges posed by MSFDA.

Acknowledgements

We thank all reviewers for their hard work and thoughtful feedbacks. This work was supported in part by the National Natural Science Foundation of China under Grant 62250061, 62176042 and 62173066, and in part by Sichuan Science and Technology Program under Grant 2023NSFSC0483, and in part Sponsored by CAAI-Huawei MindSpore Open Fund.

References

Ahmed, S. M.; Raychaudhuri, D. S.; Paul, S.; Oymak, S.; and Roy-Chowdhury, A. K. 2021. Unsupervised multi-source domain adaptation without access to source data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10103–10112.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning*, 79: 151–175.

Crammer, K.; Kearns, M.; and Wortman, J. 2008. Learning from Multiple Sources. *Journal of Machine Learning Research*, 9(8).

Dong, J.; Fang, Z.; Liu, A.; Sun, G.; and Liu, T. 2021. Confident anchor-induced multi-source free domain adaptation. *Advances in Neural Information Processing Systems*, 34: 2848–2860.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, 2066–2073. IEEE.

Griffin, G.; Holub, A.; and Perona, P. 2007. Caltech-256 object category dataset.

Guo, Y.; Li, Y.; Wang, L.; and Rosing, T. 2020. Adafilter: Adaptive filter fine-tuning for deep transfer learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 4060–4066.

Han, Z.; Zhang, Z.; Wang, F.; He, R.; Su, W.; Xi, X.; and Yin, Y. 2023. Discriminability and Transferability Estimation: A Bayesian Source Importance Estimation Approach for Multi-Source-Free Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7811–7820.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Irwin, R.; Dimitriadis, S.; He, J.; and Bjerrum, E. J. 2022. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1): 015022.

Kundu, J. N.; Kulkarni, A. R.; Bhambri, S.; Mehta, D.; Kulkarni, S. A.; Jampani, V.; and Radhakrishnan, V. B. 2022. Balancing discriminability and transferability for source-free domain adaptation. In *International Conference on Machine Learning*, 11710–11728. PMLR.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Li, G.; and Wu, C. 2023. Transformer-Based Multi-Source Domain Adaptation Without Source Data. In *2023 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Li, J.; Du, Z.; Zhu, L.; Ding, Z.; Lu, K.; and Shen, H. T. 2021a. Divergence-agnostic unsupervised domain adaptation by adversarial attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 8196–8211.

Li, R.; Jiao, Q.; Cao, W.; Wong, H.-S.; and Wu, S. 2020. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9641–9650.

Li, X.; Du, Z.; Li, J.; Zhu, L.; and Lu, K. 2022. Source-free active domain adaptation via energy-based locality preserving transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5802–5810.

Li, X.; Li, J.; Zhu, L.; Wang, G.; and Huang, Z. 2021b. Imbalanced source-free domain adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3330–3339.

Li, Y.; Yuan, L.; Chen, Y.; Wang, P.; and Vasconcelos, N. 2021c. Dynamic transfer for multi-source domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10998–11007.

Liang, J.; Hu, D.; and Feng, J. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, 6028–6039. PMLR.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, 97–105. PMLR.

Mohammed, A.; and Kora, R. 2023. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University-Computer and Information Sciences*.

Nguyen, V.-A.; Nguyen, T.; Le, T.; Tran, Q. H.; and Phung, D. 2021. Stem: An approach to multi-source domain adaptation with guarantees. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9352–9363.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1406–1415.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, 213–226. Springer.

Shen, M.; Bu, Y.; and Wornell, G. W. 2023. On Balancing Bias and Variance in Unsupervised Multi-Source-Free Domain Adaptation. In *International Conference on Machine Learning*, 30976–30991. PMLR.

Shu, Y.; Cao, Z.; Zhang, Z.; Wang, J.; and Long, M. 2022. Hub-Pathway: Transfer Learning from A Hub of Pre-trained Models. *Advances in Neural Information Processing Systems*, 35: 32913–32927.

Shu, Y.; Kou, Z.; Cao, Z.; Wang, J.; and Long, M. 2021. Zoo-tuning: Adaptive transfer from a zoo of models. In *International Conference on Machine Learning*, 9626–9637. PMLR.

Sun, T.; Lu, C.; Zhang, T.; and Ling, H. 2022. Safe self-refinement for transformer-based domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7191–7200.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5018–5027.

Wang, H.; Xu, M.; Ni, B.; and Zhang, W. 2020. Learning to combine: Knowledge aggregation for multi-source domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, 727–744. Springer.

Xu, T.; Chen, W.; Pichao, W.; Wang, F.; Li, H.; and Jin, R. 2021. CDTrans: Cross-domain Transformer for Unsupervised Domain Adaptation. In *International Conference on Learning Representations*.

Yang, S.; van de Weijer, J.; Herranz, L.; Jui, S.; et al. 2021. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in neural information processing systems*, 34: 29393–29405.

Yang, X.; Zhou, D.; Liu, S.; Ye, J.; and Wang, X. 2022. Deep model reassembly. *Advances in neural information processing systems*, 35: 25739–25753.

Zhu, J.; Bai, H.; and Wang, L. 2023. Patch-Mix Transformer for Unsupervised Domain Adaptation: A Game Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3561–3571.

Zuo, Y.; Yao, H.; and Xu, C. 2021. Attention-based multi-source domain adaptation. *IEEE Transactions on Image Processing*, 30: 3793–3803.