Augmented Self-Labeling for Source-Free Unsupervised Domain Adaptation

Anonymous ICCV submission

Paper ID ****

Abstract

Unsupervised domain adaptation aims to learn a model generalizing on target domain given labeled source data and unlabeled target data. However, source data sometimes may be unavailable when considering data privacy and decentralized learning architecture. In this paper, we address the source-free unsupervised domain adaptation problem where only the trained source model and unlabeled target data are given. To this end, we propose an Augmented Self-Labeling (ASL) method jointly optimizing model and labels for target data starting from the source model. This includes two alternating steps, where augmented self-labeling improves pseudo-labels via solving an optimal transport problem with Sinkhorn-Knopp algorithm, and model re-training trains the model with the supervision of improved pseudo-labels. We further introduce model regularization terms to improve the model re-training. Experiments show that our method can achieve comparable or better results than the state-of-the-art methods on the standard benchmarks.

1. Introduction

Deep learning has achieved great success in various applications and areas with the help of large amount of labeled data. However, annotating category labels, bounding boxes or even masks for different applications requires expensive labour cost and sometimes expert knowledge. To mitigate the reliance on manual annotations, unsupervised domain adaptation aims at adapting the model trained on a similar source domain to the unlabeled target domain.

Traditional unsupervised domain adaptation methods tackle the setting that labeled source data and unlabeled target data are available when adapting to target domain. Most methods seek to improve the model generalization ability on target domain by reducing the distribution discrepancy between domains according to the theoretical analysis in [1]. One prevailing paradigm is to learn domain-invariant representations by minimizing cross-domain feature discrepancy with certain metric. For example, Maximum Mean Discrepancy (MMD) [13] measures the feature discrepancy between two domains and lots of works [23, 25] align features by minimizing MMD in different layers. Inspired by the Generative Adversarial Networks (GAN) [11], many works [9, 36, 24, 15, 35] align domain distributions in different levels with the help of domain discriminators. Other works [8, 32, 40] utilize semi-supervised learning methods for model regularization or self-training with pseudo-labels.

However, source data might be inaccessible when adapting to target domain. This is possible in some privacy-sensitive applications. For example, federated learning collaboratively trains a model using decentralized data on mobile phones without fetching data into a centralized machine [2]. When adapting the model trained via federated learning, we have no access to the source data. This comes to the source-free unsupervised learning setting, where only the trained source model and unlabeled target data are given. Traditional unsupervised learning methods are not applicable to this setting because they usually seek to align distributions of source and target domains where samples from both domains are required. Few methods tackling this setting are published recently. For example, SHOT [22] alternately refine the pseudo-labels with a prototype classifier and fine-tunes the feature extractor together with a model regularization term maximizing mutual information between features and model outputs. 3C-GAN [21] collaboratively generates labeled target data using conditional GAN and fine-tunes the source model with the help of some model regularization terms.

In this paper, we propose a new Augmented Self-Labeling (ASL) method for the source-free unsupervised domain adaptation problem. This includes two alternating steps, where augmented self-labeling step aims to improve the pseudo-labels and model re-training step retrains the target model with the self-labeled target data. Firstly, we augment the self-labeling technique in [39] with data augmentation. Specifically, pseudo-labels obtained from the source model are noisy as the existence of cross-domain discrepancy. Thus, training target model with pseudo-labels may suffer error accumulation which decreases the performance of target model. We propose to use the ensemble of multiple predicted probabilities corresponding to different randomly augmented versions of the same sample for self-labeling. By minimizing a cross-entropy loss in addition to an entropy loss with respect to the labels, we can derive this problem to be an instance of optimal transport problem. In order to avoid the degenerate solution to this problem, we add the equi-partition constraint to the labels, which means each category contains similar number of samples. Thus this problem can be solved efficiently via a fast version of the Sinkhorn-Knopp algorithm [7].

Furthermore, we introduce several model regularization terms to improve the model re-training. Firstly, the conditional entropy minimization term is used to make the target features more discriminative and keep the decision boundaries far away from the data dense regions of the samples [32]. Secondly, virtual adversarial loss [26, 32] is added to guarantee the model's locally-Lipschitz which is important for empirical estimation of the conditional entropy. The adversarial perturbation introduced in virtual adversarial loss can also act as data augmentation which makes the model generalize better on the target domain. Thirdly, weight regularization term [21] is utilized to preserve knowledge from the source model and stable the target model training.

We apply the proposed ASL to the source-free unsupervised domain adaptation tasks. Experiments show that our method can achieve comparable or even better results than the state-of-the-art methods on the standard benchmarks. In addition, we conduct an ablation study to tease apart the contributions of each component in our method and perform hyper-parameter sensitivity analysis.

2. Related Work

Unsupervised Domain Adaptation. Most unsupervised domain adaptation methods seek to reduce the cross-domain discrepancy based on the theoretical guarantees in [1]. Related works can be divided into two categories, metric-based and adversarial training. Metric-based methods enforce the model to learn domain-invariant representations by minimizing feature discrepancy between domains with certain distance metric. Examples of these metrics include maximum mean discrepancy (MMD) [23, 25], second-order moment matching [33, 27], Wasserstein distance [31, 19], etc. Inspired by the Generative Adversarial Networks (GAN) [11], adversarial training has been utilized to align distribution between domains in different levels, including feature-level [9, 10, 36, 24, 5], input-level [15], output-level [35], etc.

Regularization terms from semi-supervised learning approaches can also be utilized to adapt the source model using unlabeled target data. Mean teacher [34] has been used in [8] to regularize the model predictions to be consistent across the student and teacher models. Entropy minimization [12] for unlabeled target data enforces the model's decision boundaries to be far away from data-dense regions [32, 6]. Virtual adversarial training [26] acts as a locally-Lipschitz constraint in [32] to guarantee the empirical approximation of conditional entropy when used together with the entropy minimization. Pseudo-labeling [20] has also inspired the self-training methods for unsupervised domain adaptation. [40] alternately select high-confident pseudo-labels using certain criteria and re-train the model with the pseudo-labeled target data.

Source-Free Unsupervised Domain Adaptation. In source-free unsupervised domain adaptation setting, labeled source data are unavailable which makes the problem more challenging. Traditional unsupervised domain adaptation methods are not applicable to this setting since both source and target data are required to align distributions in the previous methods. Few methods for the source-free unsupervised domain adaptation setting have been proposed recently. SHOT [22] refines the pseudo-labels by alternately computing the centroids for each class and performing weighted clustering in the target domain. PPDA [17] assigns pseudo-labels based on prototype classifier and a sample-level re-weighting scheme. 3C-GAN [21] and SDDA [18] utilize the conditional GAN to generate labeled target data through input-level adversarial training.

3. Preliminary

In this section, we briefly introduce the self-labeling technique as a preliminary for the proposed methodology. Self-labeling is proposed in [39] for the task of unsupervised representation learning.

In the unsupervised setting, we have only samples $\{x_i\}_{i=1}^N$ but have no access to the corresponding labels $\{y_i\}$. Self-labeling treats the labels as learnable variables and denote each of them as a one-hot vector $q_i = [q_{i1}, q_{i2}, \cdots, q_{iK}]$, and formulate the learning problem as a joint optimization over the model parameters θ and the labels $\{q_{iy}\}$ through a cross-entropy loss:

$$\begin{gathered}\min_{\theta, q} -\frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{K} q_{iy}\log p(y|x_i;\theta) \\ s.t.\ q_{iy}\in\{0,1\},\ \sum_{y=1}^{K} q_{iy}=1,\ \forall i,y.\end{gathered} \tag{1}$$

where K denotes the number of classes. However, this may lead to a degenerate solution such that the objective in Eq. (1) can be trivially minimized by assigning the same arbitrary label to all samples and making the model classify all samples to that class. To avoid this, one can constrain the label assignment such that each category contains similar number of samples, which is called equi-partition constraint in [39]. Moreover, to avoid the combinatorial optimization with respect to the binary labels q, one can relax the labels to be soft-labels, i.e. $q_{iy} \in [0,1]$. Thus, the learning problem becomes,

$$\min_{\theta,q} -\frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{K} q_{iy}\log p(y|x_i;\theta)$$
$$s.t.\ \ q_{iy}\in[0,1],\ \sum_{y=1}^{K} q_{iy}=1,\ \sum_{i=1}^{N} q_{iy}=\frac{N}{K}. \tag{2}$$

This problem is actually an instance of optimal transport problem [7]. By adding an additional entropy regularizer, it can be solved using a fast version of the Sinkhorn-Knopp algorithm [7].

4. Augmented Self-Labeling (ASL) for Source-Free Unsupervised Domain Adaptation

This paper tackles the source-free unsupervised domain adaptation problem where only source model and unlabeled target data are available. Specifically, for the traditional unsupervised domain adaptation (UDA), we have access to the labeled source data $(x_s, y_s) \in (\mathcal{X}_s, \mathcal{Y}_s)$ and unlabeled target data $x_t \in \mathcal{X}_t$. But in the source-free unsupervised domain adaptation setting, only the source model $f_s : \mathcal{X}_s \to \mathcal{Y}_s$ are given together with the unlabeled target data which we denote as x for simplification. The goal is to learn a model f generalizing well on the target data.

In this section, we present our proposed Augmented Self-Labeling (ASL) method for the source-free unsupervised domain adaptation problem. In the first part, we augment the self-labeling technique with data augmentation to obtain reliable pseudo-labels for the unlabeled target data, which can be used to train the target model. In the second part, several model regularization terms are introduced to further benefit the model re-training.

4.1. Augmented Self-Labeling

We initialize the target model f with the weights of the source model f_s. Given the unlabeled target data $\{x_i\}_{i=1}^{N}$, pseudo-labels can be obtained by choosing the high confident predictions from the model f and further used to finetune the model alternately.

However, as the existence of domain discrepancy, pseudo-labels for target data are noisy which may lead to error accumulation in the target model. On the other hand, data augmentation is a common regularization approach to enhance deep model's generalization ability. We thus propose an Augmented Self-Labeling method to optimize labels from the weighted average of multiple output predictions corresponding to samples with random data augmentations. Specifically, M different augmented version of samples $\{x_i^m\}_{m=1}^{M}$ can be obtained from the original sample x_i by applying random data augmentation M times, i.e.

$$x_i^1, x_i^2, \cdots, x_i^M = RandAugment(x_i), \tag{3}$$

where $RandAugment(\cdot)$ denotes a combination of multiple random data augmentations. The data augmentations we used include random resized crop, random auto-contrast and random color distortion [3].

In order to reduce the noise in the predicted probability, we take the ensemble of the $M+1$ probabilities corresponding to the M augmented version and the unaugmented version of sample x_i to get the average prediction indicating the probability of sample x_i belonging to class y,

$$p_{iy} = \frac{1}{2}p(y|x_i;\theta) + \frac{1}{2M}\sum_{m=1}^{M} p(y|x_i^m;\theta). \tag{4}$$

The reason half weight is assigned to the unaugmented version of predicted probability is that most target samples are still similar to the source samples and higher weight can make the obtained labels more stable and reliable.

We aim to optimize labels using the following objective,

$$\min_{\{q_{iy}\}} -\sum_{i=1}^{N}\sum_{y=1}^{K} q_{iy}\log p_{iy} + \lambda\sum_{i=1}^{N}\sum_{y=1}^{K}\log q_{iy}$$
$$s.t.\quad q_{iy}\in[0,1],\quad \sum_{y=1}^{K} q_{iy}=1,\quad \sum_{i=1}^{N} q_{iy}=\frac{N}{K}. \tag{5}$$

where K is the number of classes and we omit the coefficient $1/N$ in the cross-entropy term for the convenience of further derivation. In this objective, we relax the labels to be soft-labels to avoid the combinatorial optimization problem and after the optimization we will convert the soft-labels back to hard-labels. Secondly, the negative conditional entropy term is added to get smoothed soft-labels. The parameter λ controls the smoothness of the labels and higher λ leads to smoother soft-labels. Thirdly, the equi-partition constraint $\sum_i q_{iy} = N/K$ is added to avoid a degenerate solution where the same arbitrary label is assigned to all samples [39]. This constraint enforces that each category contains similar number of sample, which is reasonable in class-balanced dataset. But it is also rational in unbalanced dataset since it is actually maximizing the mutual information between the sample indices and labels according to [39].

The problem thus becomes an instance of optimal transport problem [7]. To make it more clear, we convert the notations in Eq. (5) to matrix form, where $[Q]_{iy} = q_{iy}$ is the label matrix with dimension of $N \times K$ and $[P]_{iy} = p_{iy}$ is the predicted probability matrix with dimension of $N \times K$. The objective in Eq. (5) can be derived to be:

$$\min_{Q\in U(r,c)} \langle Q, -\log P\rangle - \lambda H(Q) \tag{6}$$

where $\langle\cdot\rangle$ denotes Frobenius inner product, i.e. the sum of element-wise product between two matrices and log is applied in element-wise. The matrix Q is thus constrained to be an element of the transport polytope [7],

$$U(r,c) := \{Q \in \mathbb{R}_+^{N\times K} | Q\mathbf{1}_K = r, Q^\top \mathbf{1}_N = c\}. \quad (7)$$

where $r = \mathbf{1}_N, c = \frac{N}{K}\mathbf{1}_K$. This is equivalent to the constraints of labels in Eq. (5). This problem can be solved via the fast version of Sinkhorn-Knopp algorithm [7]. The closed-form solution to this problem is,

$$Q = \text{diag}(u) \cdot P^{1/\lambda} \cdot \text{diag}(v), \quad (8)$$

where u, v are vectors guaranteeing the constraints in Eq. (7) and can be computed with the Sinkhorn's fixed point iteration until convergence:

$$(u, v) \leftarrow (r./(P^{1/\lambda}v), c./([P^{1/\lambda}]^\top u)). \quad (9)$$

Specifically, we initialize v as normalized unit vector and then iteratively compute u and v until v converges. In practice, this iteration can converge in a few steps.

After the optimization, the obtained soft-labels are converted to hard-labels for model training. The target model $f(\cdot;\theta)$ can be optimized with the following cross-entropy loss using the augmented self-labeled target data,

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{K} q_{iy}\log p(y|x_i;\theta). \quad (10)$$

We can alternately perform augmented self-labeling and model re-training steps for multiple epochs such that the performance on target data can be gradually improved.

4.2. Model Regularization

Model regularization is also an important technique for deep model learning, especially for semi-supervised learning and unsupervised learning problem. Regularization terms usually constrain the model parameters or outputs based on empirical knowledge or characteristics of model and data.

In our source-free unsupervised domain adaptation problem, we believe that the target features shall be discriminative. Even though this can be achieved by the cross-entropy loss in Eq. (10), explicit regularization term can still benefit the training of target model. What's more, cluster assumption is reasonable in our setting, i.e. the target samples shall be in clusters and samples in the same cluster comes from the same class. If this assumption holds, the optimal decision boundaries shall be far away from the data dense regions of the samples [32]. To tackle these expectations, we add the conditional entropy minimization [12] term,

$$\mathcal{L}_{ent} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{K} p(y|x_i;\theta)\log p(y|x_i;\theta) \quad (11)$$

As shown above, the conditional entropy is empirically estimated using the available target data. According to [12, 32], this approximation holds only if the model is locally-Lipschitz. To this end, we add the virtual adversarial loss from [26] to guarantee the locally-Lipschitz constraint,

$$\mathcal{L}_{vat} = \mathbb{E}_x\left[\max_{\|r\|\le\epsilon} D_{\text{KL}}(f(x)\|f(x+r))\right], \quad (12)$$

where $D_{\text{KL}}(\cdot\|\cdot)$ is the Kullback-Leibler Divergence. According to [26], r is first initialized as Gaussian random noise with the same shape as the input batch samples to compute the KL divergence loss and then updated as the gradient of the loss w.r.t. r itself. After few iterations, the obtained r is treated as the perturbation that makes the model behaviours most differently and the corresponding KL loss is treated as the final virtual adversarial loss. By minimizing this loss, we are expecting the model can behaviours consistently within the norm-ball of each sample [32], which guarantees the locally-Lipschitz constraint. What's more, the perturbation added to the samples can be treated as a kind of data augmentation which makes the model generalize better on the target domain.

In our method, the source model $f_s(\cdot;\theta_s)$ is only used as an initialization of the target model $f(\cdot;\theta)$ so far. While fine-tuning on the source model with the self-labeled target data, the target model could possibly get far away from the source hypothesis. However according to the theoretical analysis in [1], the optimal classifier shall generalize well on both domains. Therefore, we add the weight regularization loss which computes the squared L2 distance between the source and target model parameters,

$$\mathcal{L}_{wr} = \|\theta - \theta_s\|_2^2. \quad (13)$$

On one hand, the weight regularization prevents the target hypothesis getting far away from the source, which helps preserve the source knowledge in the target model [21]. On the other hand, it stables the target model training since the obtained labels are noisy and updated every epoch.

Combining the cross-entropy loss in Eq. (10), the overall loss for model re-training is,

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_1(\mathcal{L}_{ent} + \mathcal{L}_{vat}) + \lambda_2\mathcal{L}_{wr}, \quad (14)$$

where λ_1 and λ_2 are trade-off parameters. The entropy loss and virtual adversarial loss empirically share the same trade-off parameter [32, 21].

Overall, the augmented self-labeling procedure assigns labels to the unlabeled target data according to Eq. (8), which is further used to re-train the target model by minimizing the loss in Eq. (14). The target model can be trained by alternating the two steps in each epoch. The algorithm for our proposed method is shown in Algorithm 1.

ICCV #****

Algorithm 1 Augmented Self-Labeling for Source-Free Unsupervised Domain Adaptation

Input: source model $f_s(\cdot;\theta_s)$, target model $f(\cdot;\theta)$ initialized with source model, unlabeled target data $\{x_i\}_{i=1}^N$, number of classes K, number of data augmentations M, self-labeling parameter λ, trade-off parameters λ_1, λ_2, learning rate η.

for $i = 1, \cdots, N\_epochs$ **do**
  $X = [x_1, x_2, \cdots, x_N]$ ▷ ASL starts
  $P^0 = f(X;\theta)$
  for $m = 1, \cdots, M$ **do**
    $\hat{X} = RandAugment(X)$
    $P^m = f(\hat{X};\theta)$
  end for
  $P = \frac{1}{2}P^0 + \frac{1}{2M}\sum_{m=1}^{M} P^m$
  $v = \mathbf{1}_K/K$ ▷ Sinkhorn's iteration
  $err = 1$
  while $err > 0.1$ **do**
    $u = 1./(P^{\lambda} v)$
    $v' = N./(K \cdot [P^{\lambda}]^{\top} u)$
    $err = \|v'/v - 1\|_1$
  end while
  $Q = \mathrm{diag}(u) \cdot P^{\lambda} \cdot \mathrm{diag}(v)$ ▷ get soft labels
  $Q = \arg\max Q$ ▷ ASL ends
  $\hat{X} = RandAugment(X)$ ▷ re-training starts
  $P = f(\hat{X};\theta)$
  $\mathcal{L} = \mathcal{L}_{ce}(P,Q) + \lambda_1(\mathcal{L}_{ent} + \mathcal{L}_{vat}) + \lambda_2\mathcal{L}_{wr}$
  $\theta = \theta - \eta\frac{\partial \mathcal{L}}{\partial \theta}$ ▷ re-training ends
end for

Output: target model $f(\cdot;\theta)$

5. Experiments

5.1. Setup

We evaluate the proposed Augmented Self-Labeling (ASL) method on the following standard benchmarks.

Office-31 [29] is a standard small-sized visual domain adaptation benchmark which contains images of 31 categories from three domains: Amazon (**A**), DSLR (**D**) and Webcam (**W**), each containing 2,817, 498 and 795 images respectively.

Office-Home [37] is a medium-sized dataset with images belonging to 65 categories from four distinct domains: Artistic images (**Ar**), Clip Art (**Cl**), Product images (**Pr**), and Real-World images (**Rw**), each including 2,427, 4,365, 4,439 and 4,357 images respectively.

VisDA-2017 [28] is a large-scale synthetic-to-real dataset with images in 12 categories from two domains, **Synthetic** and **Real**, each consists of 152,397 and 55,388 images respectively.

Baseline Methods. We compare our method ASL with the existing methods for source-free unsupervised domain adaptation setting, SHOT [22], PPDA [17], 3C-GAN [21] and SDDA [18]. As references, we also list results from recent state-of-the-art methods for standard unsupervised domain adaptation setting, including Domain Adversarial Neural Network (DANN) [10], Adversarial Discriminative Domain Adaptation (ADDA) [36], Maximum Classifier Discrepancy (MCD) [30], Conditional Domain Adversarial Network (CDAN) [24], BSP [4], TransNorm [38], SWD [19] and CAN [16].

5.2. Implementation Details

We use the same network architecture as the previous methods for fairness. For **Office-31** and **Office-Home** datasets, we use ResNet-50 [14] as the backbone network. Considering image quantity and for better performance, ResNet-101 [14] is utilized as the backbone module for **VisDA-2017** dataset. Following [9], the fully-connected (FC) layer in the ResNet network is replaced with a bottleneck and one FC layer, where the bottleneck layer is composed of one FC layer with 256 units and an one-dimensional Batch Normalization (BN) layer.

To get the trained source model, we randomly split each dataset into training set and validation set with the ratio 0.9/0.1. The ResNet model pretrained on ImageNet is used to initialize the backbone module and then the complete model is trained on the training set. We adopt mini-batch SGD with momentum 0.9 to optimize all networks. The batch size is set to be 64 considering GPU RAMs. Following [9], the learning rate is adjusted per batch iteration according to $\eta_i = \eta_0(1 + \gamma\frac{i}{n})^{-\beta}$, where $\gamma = 10, \beta = 0.75$, i is the iteration index and n is the total number of iterations. What's more, η_0 is the initial learning rate which is set to be 0.001 for the pretrained backbone module and 0.01 for the bottleneck and FC layers. The optimal model with best validation accuracy is saved as the source model.

When adapting to the target domain, we perform the self-labeling procedure once per epoch. The target model is first initialized with the weights of source model and then optimized using the same mini-batch SGD algorithm. The batch size is set to be 32 since the virtual adversarial loss costs more GPU RAMs. The learning rate is fixed to be 10^{-4} for the backbone and 10^{-3} for the bottleneck and FC layers such that all sample share the same weight in each iteration. The optimal trade-off parameters are $\lambda = 2, \lambda_1 = 1, \lambda_2 = 0.1, M = 4$ for **Office-31** and **Office-Home** datasets and $\lambda = 100, \lambda_1 = 1, \lambda_2 = 0.01, M = 1$ for **VisDA-2017** dataset.

5.3. Results

We evaluate our proposed method ASL on the three visual domain adaptation benchmarks including **Office-31**,

Table 1. Classification accuracy (%) on Office-31 (ResNet-50)

Methods	A → D	A → W	D → A	D → W	W → A	W → D	Avg.
ResNet-50 [14]	68.9	68.4	62.5	96.7	60.7	99.3	76.1
DANN [10]	79.7	82.0	68.2	96.9	67.4	99.1	82.2
ADDA [36]	77.8	86.2	69.5	96.2	68.9	98.4	82.9
CDAN+E [24]	92.9	94.1	71.0	98.6	69.3	100.	87.7
CDAN+BSP [4]	93.0	93.3	73.6	98.2	72.6	100.	88.5
CDAN+TransNorm [38]	94.0	95.7	73.4	98.7	74.2	100.	89.3
CAN [16]	95.0	94.5	78.0	99.1	77.0	99.8	90.6
SDDA [18]	85.3	82.5	66.4	**99.0**	67.7	99.8	83.5
SHOT [22]	93.1	90.9	74.5	98.8	74.8	**99.9**	88.7
3C-GAN [21]	92.7	93.7	75.3	98.5	**77.8**	99.8	**89.6**
ASL (Ours)	**93.4**	**94.1**	**76.0**	98.4	75.0	99.8	89.5

Office-Home and **VisDA-2017** under the source-free unsupervised domain adaptation setting.

Results on Office-31: Table 5.3 shows the performance of different methods on the six domain adaptation tasks of this small sized dataset, where the first part includes the source-only and unsupervised domain adaptation methods and the second part consists of the existing methods and our method under the source-free unsupervised domain adaptation setting. All methods use ResNet-50 [14] as the backbone network. Denoted as ResNet-50, source-only reports the performance of the target data evaluated directly using the source model. Comparing with the source-only results, our method improves the performance of all the six domain adaptation tasks and achieves an average 17.6% performance gain, which shows the effectiveness of our method. Comparing with the unsupervised domain adaptation methods, our method can outperform most previous methods even though the setting without source data is more challenging. What's more, our method achieves better performance than SHOT [22] and SDDA [18], illustrating the effectiveness of the augmented self-labeling procedure. We can also achieve similar performance as 3C-GAN [21] even though 3C-GAN uses generative model to generate lots of labeled target data which is time-costing and resource-costing. Especially, our method achieves the state-of-the-art performance on the first three tasks, i.e. A→D, A→W and D→A under the source-free unsupervised domain adaptation setting.

Results on Office-Home: Table 5.3 demonstrates the performance of different methods on the 12 domain adaptation tasks of this medium sized benchmark. All methods share the same ResNet-50 [14] backbone network. Comparing with the source-only, our method improves every task's performance and obtains an average 52% performance gain. What's more, our method outperforms all the unsupervised domain adaptation methods on this dataset given the more challenging source-free setting for our method. Comparing with existing source-free UDA methods, our method outperform PPDA [17] by a large margin. We can also obtain comparable results to the state-of-the-art method SHOT [22] even though SHOT raises the baseline (source-only) by using label smoothing and weight normalization when training source model.

Results on VisDA-2017: Table 5.3 illustrates the accuracy for each class and the average accuracy per class under different methods on this large scale benchmark. ResNet-101 [14] is used as the backbone network in all methods. Our method achieves the state-of-the-art performance under the source-free unsupervised domain adaptation setting. Comparing with the source-only, our method can improve the accuracy in every class and achieve 58% performance gain on average. We can also outperform most previous unsupervised domain adaptation methods even though under the more strict source-free constraint. Comparing with the methods under the same setting, our method outperforms PPDA [17], SHOT [22] and 3C-GAN [21] in most classes and on average we get the best result.

5.4. Ablation Study

We further perform an ablation study to tease apart the contributions of each component in our method and conduct hyper-parameter sensitivity analysis.

Contribution of each component: As shown in Table 4, we compare the performance of our method dropping different components with the naive pseudo-labeling method [20] which directly fine-tunes the source model with the pseudo-labeled target data. Firstly, we can see that both self-labeling and augmented self-labeling can easily outperform the naive PL method even without the model regularization terms. This demonstrates the superiority of self-labeling method in source-free unsupervised domain adaptation tasks. Secondly, the model regularization terms can also benefit the performance of both naive PL and our method. Thirdly, augmented self-labeling can promote the results by a large margin comparing with the self-labeling, which shows that data augmentation can truly benefit the self-labeling procedure. What's more, we can see that each model regularization term plays a positive role in achieving the final result.

Hyper-parameter sensitivity analysis: Table 5 shows the classification accuracy of our method on the task A → W of **Office-31** dataset under different times of random data augmentation used in augmented self-labeling step. We can see that multiple times of data augmentation do benefit the performance of target model but 4 times is enough for the task A→W. We can also see that the performance of augmented self-labeling method is not sensitive to the times of data augmentation used such that the proposed method under different parameter M all achieve similar and good performance comparing with other methods.

Table 5.4 shows the accuracy of our method on the same task under different parameter λ used in augmented self-labeling to control the smoothness of labels. Smaller λ can

Table 2. Classification accuracy (%) on Office-Home (ResNet-50)

Methods	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
ResNet-50 [14]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DANN [10]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
CDAN+E [24]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
CDAN+BSP [4]	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	59.3	81.9	66.3
CDAN+TransNorm [38]	50.2	71.4	77.4	59.3	72.7	73.1	61.0	53.1	79.5	71.9	59.0	82.9	67.6
PPDA [17]	48.5	71.3	75.6	63.9	69.0	72.1	62.4	43.5	76.0	70.4	50.1	76.1	64.9
SHOT [22]	**56.9**	**78.1**	**81.0**	**67.9**	**78.4**	**78.1**	**67.0**	54.6	**81.8**	**73.4**	58.1	**84.5**	**71.6**
ASL (Ours)	56.0	77.0	79.7	66.3	76.5	77.7	62.8	**54.9**	81.6	71.5	**58.4**	83.7	70.5

Table 3. Class-wise accuracy (%) on VisDA-2017 (ResNet-101)

Methods	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Per-class
ResNet-101 [14]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DANN [10]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
MCD [30]	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
CDAN [24]	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.9
CDAN+BSP [4]	92.4	61.0	81.0	57.5	89.0	80.6	90.1	77.0	84.2	77.9	82.1	38.4	75.9
SWD [19]	90.8	82.5	81.7	70.5	91.7	69.5	86.3	77.5	87.4	63.6	85.6	29.2	76.4
CAN [16]	97.0	87.2	82.5	74.3	97.8	96.2	90.8	80.7	96.6	96.3	87.5	59.9	87.2
PPDA [17]	81.5	79.4	80.3	61.8	92.3	91.9	84.5	82.7	86.5	58.4	74.2	43.5	76.4
SHOT [22]	92.6	81.1	80.1	58.5	89.7	86.1	81.5	77.8	89.5	**84.9**	84.3	49.3	79.6
3C-GAN [21]	94.8	73.4	68.8	**74.8**	93.1	**95.4**	88.6	**84.7**	89.1	84.7	83.5	48.1	81.6
ASL (Ours)	**97.3**	**85.3**	**86.9**	70.7	**96.4**	72.8	**93.0**	80.1	**95.5**	78.1	**87.7**	**50.3**	**82.8**

Table 4. Ablation study: accuracy (%) with each component.

Methods	Office-31
Source Only	76.1
Naive Pseudo-Labeling (PL) [20]	76.7
Naive PL + $\mathcal{L}_{ent} + \mathcal{L}_{vat} + \mathcal{L}_{wr}$	83.3
Self-Labeling (SL)	83.8
SL + $\mathcal{L}_{ent} + \mathcal{L}_{vat} + \mathcal{L}_{wr}$	86.7
Augmented Self-Labeling (ASL)	88.0
ASL + $\mathcal{L}_{ent} + \mathcal{L}_{vat}$	88.4
ASL + $\mathcal{L}_{ent} + \mathcal{L}_{vat} + \mathcal{L}_{wr}$	89.5

Table 5. Ablation study: accuracy (%) on task A→W under different times of random data augmentation (M)

M	1	4	7	10
A→W	93.0	94.1	92.5	92.5

Table 6. Ablation study: accuracy (%) on task A→W under different λ used in augmented self-labeling

λ	0.1	0.5	1	2	5	10
A→W	92.8	93.8	94.0	94.1	90.6	90.1

achieve better accuracy, which means less smoothed labels can benefit the model training on this task. We can also see that the accuracy is approximately concave related to the parameter λ and achieve the best performance in $\lambda = 2$. But our method still can get good performance under different value of λ, which means our method is not quite sensitive to the parameter λ.

6. Conclusion

In this paper, we propose a new Augmented Self-Labeling method for the source-free unsupervised domain adaptation, where only source model and unlabeled target data are available. We formulate this problem as a joint optimization over the labels and model. This can be divided into two alternating steps, where self-labeling improves the pseudo-labels with the help of equi-partition constraint and re-training trains the model with the self-labeled target data. We further exploit data augmentation to improve the self-labeling procedure by the ensemble of multiple probability matrices corresponding to augmented versions of samples. What's more, model regularization terms are introduced to further benefit the model re-training. Experiments on different sized benchmarks verify the effectiveness and superiority of our proposed method for the source-free unsupervised domain adaptation problem.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010. 1, 2, 4

[2] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, H Brendan McMahan, et al. Towards federated learning at scale: System design. *arXiv:1902.01046*, 2019. 1

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[4] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *ICML*, 2019. 5, 6, 7

[5] Shuhao Cui, Xuan Jin, Shuhui Wang, Yuan He, and Qingming Huang. Heuristic domain adaptation. In *NeurIPS*, 2020. 2

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

[7] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 2013. 2, 3, 4

[8] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *ICLR*, 2018. 1, 2

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 1, 2, 5

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2, 5, 6, 7

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1, 2

[12] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NeurIPS*, 2005. 2, 4

[13] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In *NeurIPS*, 2007. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6, 7

[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 *ICML*, 2018. 1, 2

[16] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, 2019. 5, 6, 7

[17] Youngeun Kim, Donghyeon Cho, and Sungeun Hong. Towards privacy-preserving domain adaptation. *IEEE Signal Processing Letters*, 27:1675–1679, 2020. 2, 5, 6, 7

[18] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *WACV*, 2021. 2, 5, 6

[19] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *CVPR*, 2019. 2, 5, 7

[20] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013. 2, 6, 7

[21] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *CVPR*, 2020. 1, 2, 4, 5, 6, 7

[22] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020. 1, 2, 5, 6, 7

[23] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 1, 2

[24] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018. 1, 2, 5, 6, 7

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

[26] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 2, 4

[27] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 2

[28] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv:1710.06924*, 2017. 5

[29] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010. 5

[30] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 5, 7

[31] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *AAAI*, 2018. 2

[32] Rui Shu, Hung Bui, Hirokazu Narui, and Stefano Ermon. A DIRT-T approach to unsupervised domain adaptation. In *ICLR*, 2018. 1, 2, 4

[33] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016. 2

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

[35] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 1, 2

[36] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 1, 2, 5, 6

[37] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017. 5

[38] Ximei Wang, Ying Jin, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable normalization: Towards improving transferability of deep neural networks. In *NeurIPS*, 2019. 5, 6, 7

[39] Asano YM., Rupprecht C., and Vedaldi A. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020. 1, 2, 3

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