Cooperative and Adversarial Learning: Co-Enhancing Discriminability and Transferability in Domain Adaptation

Hui Sun, Zheng Xie, Xin-Ye Li, Ming Li
Nanjing University,
Nanjing 210023, China
{xiez,sunh,lixy,lim}@lamda.nju.edu.cn

Abstract

Discriminability and transferability are two goals of feature learning for domain adaptation, as we aim to find the transferable features from the source domain that are helpful for discriminating the class label in the target domain. Modern domain adaptation approaches optimize discriminability and transferability by adopting two separate modules for the two goals upon a feature extractor, but lack fully exploiting their relationship. In this paper, we argue that by letting the discriminative module and transfer module help each other, better domain adaptation can be achieved. We propose Cooperative and Adversarial LEarning (CALE) to combine the optimization of discriminability and transferability into a whole, provide one solution for making the discriminative module and transfer module guide each other. Specifically, CALE generates cooperative (easy) examples and adversarial (hard) examples with both discriminative module and transfer module. While the easy examples that contain the module knowledge can be used to enhance each other, the hard ones are used to enhance the robustness of the corresponding goal. Experimental results show the effectiveness of CALE for unifying the learning of discriminability and transferability, as well as its superior performance.

1   Introduction

Transfer learning seeks to make the machine learning systems be able to perform well on new tasks by referring to the experience of some old tasks. Unsupervised domain adaptation (UDA), one of the most active subfields of transfer learning, shows a great ability to transfer knowledge across tasks [43]. It aims to build a model on an unlabeled target domain with the help of the source domain data, and has been applied to various tasks, including computer vision [34, 33], natural language processing [29, 62], information retrieval [30, 12, 25], reinforcement learning [10, 45], etc.

Due to the potent power of feature representation learning of the deep neural networks, most of the modern domain adaptation approaches are dedicated to designing elaborated mechanisms for learning domain invariant feature representations, such that the distributions of source and target domain are well aligned in the representation space [53, 15, 51]. Such features bridge the source domain and target domain, making it possible to transfer knowledge across domains.

A successful feature representation for domain adaptation requires two properties, *discriminability* and *transferability*, which form the fundamental goals of UDA feature learning [2, 9]. Discriminability refers to the effectiveness of the representation to be used for the discriminative task, e.g., classification. Transferability refers to the ability of the representation to capture the invariance across the domains so that the different domains can be well aligned under the learned representation.

(a) Before adaptation.　(b) Discriminability guides Transferability.　(c) Transferability guides discriminability.　(d) After adaptation.
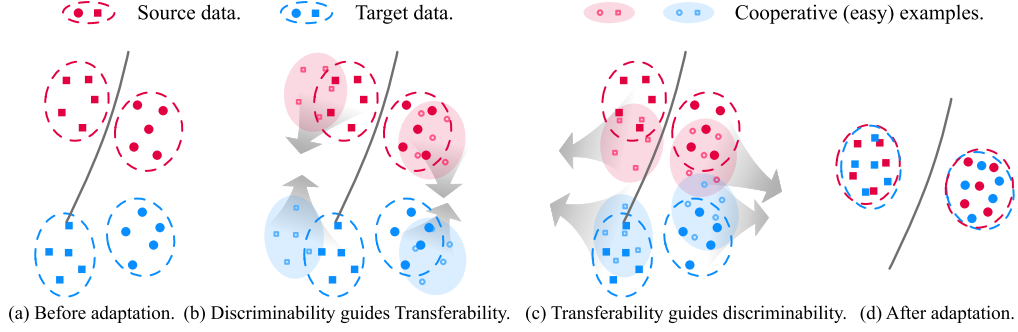
Figure 1: Demonstration of cooperative learning in CALE. The cooperative examples of discriminability and transferability help the feature learning when optimizing the other goal.

The modern domain adaptation approaches generally consist of a discriminative module (e.g., a classifier) and a transfer module (e.g., a domain discriminator based on Generative Adversarial Networks (GANs) [17] or a distribution distance function based on the statistical moment) upon the feature extractor. The two modules will be trained by the discriminability and transferability loss correspondingly.

However, existing literature treats the discriminability and transferability as two separate terms in the loss function, produced by two modules that are *optimized separately* [53, 15, 37, 60]. Since the discriminability and transferability can be partly conflicting, without considering optimizing them together, the two goals may adversely impact each other, leading to a degradation in the performance or even a negative transfer to some extent [59, 8, 26].
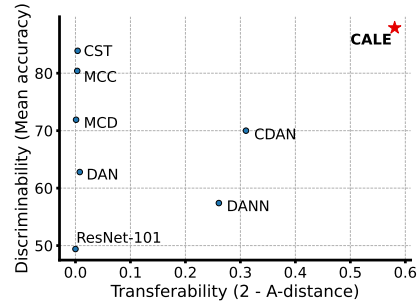

Figure 2: Discriminability and transferability.

In this paper, we argue that the discriminability and transferability should be considered jointly. By letting the discriminative module and transfer module help each other, better domain adaptation can be achieved. We propose Cooperative and Adversarial LEarning (CALE) for domain adaptation, which exploits the knowledge of optimizing one property to guide the another. Specifically, for cooperative learning, we generate easy examples in terms of discriminability loss and transferability loss, and then the discriminative and transfer module regularize their outputs to be consistent with the examples easy for their complementary modules. The easy examples contain the module knowledge of better feature representation, by exchanging the easy examples, the two modules take the guidance from each other, makes the feature extractor consider one property while optimizing another, and vice versa. Figure 1 demonstrates how the easy examples generated by the two modules can guide each other. For adversarial learning, we regularize the output consistency of the two modules on their adversarial hard example of their own loss. It further enhances the robustness of the feature learning in terms of discriminability and transferability. Figure 2 shows the discriminability and transferability of several recent domain adaptation approaches, in which we can observe that CALE not only achieve good discriminability, but also maintains great transferability. Detailed experiment protocol can be find in Section 4.2. The conceptual framework of CALE is shown in Figure 3. We emphasize our contributions in two aspects:

1. We argue that the discriminability and transferability in domain adaptation should be learned jointly instead of separately, and empirically validate our claim.

2. We propose a Cooperative and Adversarial Learning (CALE) framework to unify the learning of the discriminability and transferability. The CALE model not only achieves good performance but also keeps great transferability across domains.

2　Preliminaries

Unsupervised Domain Adaptation.　Let $\mathcal{X}$ be some feature space and $\mathcal{Y}$ be some label space. In this paper, we focus on the problem of unsupervised domain adaptation (UDA), where we want to

build a model for a target domain $\mathcal{T}$ defined over $\mathcal{X} \times \mathcal{Y}$ with the knowledge from a source domain $\mathcal{S}$. Formally, only a labeled set $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ with n_s labeled samples from the source domain $\mathcal{S}$ and an unlabeled set $\mathcal{D}_t = \{x_i^t\}_{i=1}^{n_t}$ with n_t unlabeled samples from the target domain $\mathcal{T}$ are available for building the model. As the two domains do not share the same probability distribution, the core problem of UDA is to handle the *domain shift* between $\mathcal{S}$ and $\mathcal{T}$ [43].

Revisit the Current State-of-the-Art.   The seminal work of Ben-David et al. [2] proposed the important theory for DA, on which most modern domain adaptation approaches are based:

Theorem 1. *Let $\mathcal{H}$ be the hypothesis space, and given a source domain $\mathcal{S}$ and a target domain $\mathcal{T}$. The upper bound of the expected error on the target domain is:*

$$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + C \tag{1}$$

where $\epsilon_{\mathcal{S}}(h)$ is the expected error on the source domain, $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ is a discrepancy metric of cross-domain distributions, and C is an ideal error defined as $C = \min_{h^* \in \mathcal{H}} [\epsilon_{\mathcal{S}}(h^*) + \epsilon_{\mathcal{T}}(h^*)]$.

This theorem shows that the expected error on the target domain is bounded by the sum of the expected error on source, the domain discrepancy, and a constant term related to the hypothesis class. From the perspective of feature representation learning, $\epsilon_{\mathcal{S}}(h)$ reflects the discriminability, and $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ reflects the transferability. Inspired by this, recent deep domain adaptation approaches generally consist of a discriminative module for enhancing the *discriminability*, and a transfer module for enhancing the *transferability*, upon a shared feature extractor. Concretely, in classification task, the discriminative module is usually a classifier that minimizes the discriminative loss on the domain data, e.g., cross-entropy on labeled source data and self-training loss on unlabeled target data. The transfer module guides the feature extractor to learn domain invariant feature representations through two mainstream technologies: moment matching [53, 35, 60] and adversarial confusing [15, 37].

Limitations of Current Approaches.   Current approaches mainly optimize discriminability and transferability separately, which has the following limitations. (1) When the discriminative module independently optimizes the discriminability, it may destroy the transferability of representations. For instance, to improve the discriminability on the source domain, the model extracts too many source-specific features, so the independent transfer module is difficult to align the target's feature representations with source-specific feature representations. (2) The transfer module optimizes transferability independently without caring about the discriminability of representations. This may destroy the discriminability of representations during the transfer process and makes the extracted features non-discriminable. For example, in image classification, if the backgrounds are similar across domains, the information of backgrounds can easily minimize $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ and enhance the transferability of representations. However, background in image classification always is non-discriminable. Useless background information may deteriorate the domain adaptation.

3   Cooperative and adversarial learning

In this section, we introduce Cooperative and Adversarial LEarning (CALE), which enhances the discriminability and transferability synchronously from an overall perspective.

CALE is a general regularization framework that combines the optimizations of discriminability and transferability into a whole. It works upon the common domain adaptation framework that consists of three parts: a feature extractor $z = F(x)$, a discriminative module $G(z)$ for enhancing the discriminability of the feature representations, and a transfer module $H(z)$ for enhancing the transferability of the feature representations [53, 35, 60, 15, 37, 5]. In domain adaptation for classification, the discriminative module is typically a category classifier $\hat{y} = G(z)$ trained with supervised (and sometimes with self-training) classification loss. The transfer module is usually implemented through moment matching or adversarial confusing. In moment matching, the transfer module minimizes the statistical moment distance to match the distributions across domains [53, 35, 60]. In adversarial confusing, the transfer module can be a domain discriminator $\hat{d} = G(z)$ based on GANs [17] that confuses the feature distributions in a min-max optimization [15, 37]. We denote the loss of $G(z)$ as discriminability loss $\mathcal{L}_{disc}$ and the loss of $H(z)$ as transferability loss $\mathcal{L}_{tran}$. Most existing domain adaptation approaches are generally trained by minimizing the overall loss in the
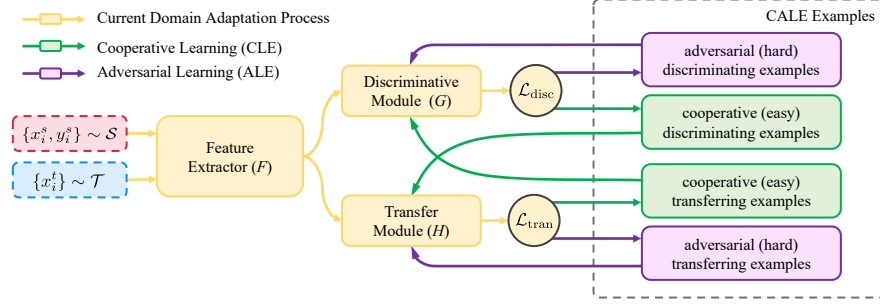
Figure 3: Cooperative and adversarial learning framework. The discriminative module G enhances the feature discriminability by minimizing a classification loss. The transfer module H enhances the feature transferability by minimizing domain distribution moments (moment matching) or maximizing an error of a domain discriminator (adversarial confusing).

following form:
$$\mathcal{L} = \mathcal{L}_{\mathrm{disc}} + \mathcal{L}_{\mathrm{tran}}. \tag{2}$$

Though discriminability loss and transferability loss are both minimized, the two modules $G(z)$ and $H(z)$ are independent, focusing on different goals during the optimization. As discussed before (Section 2), it leads to unsatisfactory feature learning. To tackle this problem, CALE (see Figure 3) provides a mechanism to bridge the discriminability module and the transferability module. Specifically, inspired by adversarial examples learning [18, 42], CALE generates cooperative (easy) examples and adversarial (hard) examples for both discriminability and transferability. Then, the easy examples are used to guide each other, and the hard ones are used to enhance their own robustness.

In the remainder of this section, we first introduce the cooperative learning (CLE) and the adversarial learning (ALE) of CALE, which together can be regarded as a general regularization framework for domain adaptation. Then, we provide an instantiated model based on the CALE framework.

3.1 Cooperative learning

CALE leverages a cooperative learning (CLE) mechanism between the discriminative module G and the transfer module H, as shown in Figure 3. To make the discriminative module and the transfer module guide each other, we try to let the discriminative module provide a clue of the feature learning direction for the transfer module to search transferable features in a more discriminable way, and vice versa. We achieve this by generating easy examples for training each other, as shown in Figure 1.

Discriminability guides Transferability. During the training, the easy examples for the discriminative module provide a clue of more discriminable directions for the transfer module. As shown in Figure 1b, CLE generates more discriminable examples by adding cooperative (in opposite to adversarial) perturbations to the original examples, and then minimizes the consistency loss of the transfer module on such discriminable examples:

$$\ell_{\mathrm{CLE\_tran}}(x) = \mathrm{Dist.}\left[H(F(x)) \| H(F(x_{\mathrm{disc}}))\right]; \quad x_{\mathrm{disc}} = \operatorname*{argmin}_{x';\|x-x'\|\le\epsilon} \ell_{\mathrm{disc}}(x'), \tag{3}$$

where ℓ_{disc} is a discriminability loss function corresponds to $\mathcal{L}_{\mathrm{disc}}$, and Dist. is a distance function for measuring the discrepancy between two distributions, e.g., KL-divergence or cross-entropy. It drives the transfer module to use more discriminable features to regularize (teach) the original features.

Transferability guides Discriminability. Similar to the previous part, CLE generates more transferable examples to provide guidance for the discriminative module, so that the discriminative module will be more likely to use transferable features for the discriminative task, i.e., Figure 1c. CLE generates transferable examples, and then minimizes the consistency loss of discriminative module on such transferable examples:

$$\ell_{\mathrm{CLE\_disc}}(x) = \mathrm{Dist.}\left[G(F(x)) \| G(F(x_{\mathrm{tran}}))\right]; \quad x_{\mathrm{tran}} = \operatorname*{argmin}_{x';\|x-x'\|\le\epsilon} \ell_{\mathrm{tran}}(x'), \tag{4}$$

where ℓ_{tran} is a transferability loss function corresponding to $\mathcal{L}_{\mathrm{tran}}$.

4

To calculate the cooperative examples, we approximate the minimizers above by move x towards its negative gradient of the loss, i.e., $x' = x - \nabla_x \ell(x) / \|\nabla_x \ell(x)\|_2$.

CLE bridges the discriminable module G and the transferable module H by swapping their easy examples, which makes the two modules learn each other's virtues. By learning from discriminative examples, the transfer module H pays more attention to the discriminative features when trying to align domains. Similarly, the discriminative module G focuses more on using domain invariant features for the discriminative task. Consequently, the two modules progress together to learn feature representations that are both transferable and discriminable.

3.2 Adversarial learning

In addition to exchanging the easy examples between the transfer module H and the discriminative module G, CALE further generates hard examples for themselves to enhance their own robustness of the feature learning in terms of discriminability and transferability, i.e., adversarial learning (ALE).

The adversarial regularization loss for discriminability can be formulated as:

$$\ell_{\text{ALE\_disc}}(x) = \text{Dist.}\left[G(F(x)) \| G(F(x_{\text{non-disc}}))\right]; \; x_{\text{non-disc}} = \operatorname*{argmax}_{x'; \|x-x'\| \leq \epsilon} \ell_{\text{disc}}(x'). \quad (5)$$

Similarly, the adversarial regularization loss for transferability can be formulated as:

$$\ell_{ALE\_\text{tran}}(x) = \text{Dist.}\left[H(F(x)) \| H(F(x_{\text{non-disc}}))\right]; \; x_{\text{non-tran}} = \operatorname*{argmax}_{x'; \|x-x'\| \leq \epsilon} \ell_{\text{tran}}(x'). \quad (6)$$

To calculate the adversarial examples, we move x towards its gradient of the losses, i.e., $x' = x + \nabla_x \ell(x) / \|\nabla_x \ell(x)\|_2$.

3.3 The CALE model

We provide an instantiated classification domain adaptation model based on our CALE framework.

Transfer Module. The transfer module has two mainstream technologies for domain alignment: moment matching and adversarial confusing. While CALE can be applied to both of them, in this paper, we focus on the adversarial confusing. Let $H(z)$ be a binary classification network for domain discrimination, the training of the transfer model is then a min-max game between feature extractor F and domain discriminator H: H aims to distinguish the domain label of feature representation while F aims to confuse H. Formally, the transferability loss can be written as:

$$\mathcal{L}_{\text{tran}}(\theta_F, \theta_H) = \mathbb{E}_{x_i^s \sim \mathcal{D}_s} \log\left[H(F(x_i^s))\right] + \mathbb{E}_{x_i^t \sim \mathcal{D}_t} \log\left[1 - H(F(x_i^t))\right]. \quad (7)$$

Discriminative Module. The discriminative module is a classifier that minimizes the classification error. On labeled data from the source domain, cross-entropy is adopted. In addition, the unlabeled data from the target domain are exploited through self-training by fitting their pseudo-labels, e.g., [33, 27, 6, 11]. The discriminability loss is then formulated as follows:

$$\mathcal{L}_{\text{disc}}(\theta_F, \theta_G) = \mathbb{E}_{(x_i^s, y_i^s) \sim \mathcal{D}_s} \ell_{ce}\left(G(F(x_i^s)), y_i^s\right) + \mathbb{E}_{x_i^t \sim \mathcal{D}_t} \mathbb{I}_{[\max(\hat{y}_i^t) \geq \tau]} \ell_{ce}\left(G(F(x_i^t)), \text{PL}(\hat{y}_i^t)\right)$$
(8)

Where $\ell_{ce}(p, q) = -\sum_i q_i \log(p_i)$ is the cross-entropy loss function, θ_G means the parameters of the category classifier G, and $\text{PL}(\hat{y}) = \text{onehot}(\operatorname{argmax}(\hat{y}))$ means the pseudo label of prediction.

Unified Cooperative and Adversarial Learning. Specifically, cooperative regularization losses for discriminability and transferability are calculated by:

$$\ell_{\text{CLE\_disc}}(x) = \text{Dist.}\left[G(F(x)) \| G(F(x_{tran}))\right]; \; x_{\text{tran}} = \operatorname*{argmin}_{x'; \|x-x'\| \leq \epsilon} \ell_{ce}(H(F(x')), H(F(x))),$$

$$\ell_{\text{CLE\_tran}}(x) = \text{Dist.}\left[H(F(x)) \| H(F(x_{disc}))\right]; \; x_{\text{disc}} = \operatorname*{argmin}_{x'; \|x-x'\| \leq \epsilon} \ell_{ce}(G(F(x')), G(F(x))).$$
(9)

Moreover, adversarial regularization losses are:

$$\ell_{\text{ALE\_disc}}(x) = \text{Dist.}\left[G(F(x)) \| G(F(x_{\text{non-disc}}))\right];$$
$$x_{\text{non-disc}} = \operatorname*{argmax}_{x'; \|x-x'\| \leq \epsilon} \ell_{ce}(G(F(x')), G(F(x))). \quad (10)$$

Algorithm 1 Cooperative and Adversarial LEarning (CALE)
___
Input: Source domain dataset $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ and target domain dataset $\mathcal{D}_t = \{x_i^t\}_{i=1}^{n_t}$.
Output: A model with parameters $\theta = (\theta_F, \theta_G, \theta_H)$
 1: Initialize parameters $\theta = (\theta_F, \theta_G, \theta_H)$ randomly.
 2: **for** $iter = 0$ to $MaxIteration$ **do**
 3:     Calculate the transferability loss $\mathcal{L}_{\text{tran}}$ and discriminability loss $\mathcal{L}_{\text{disc}}$ (Equ 7 and 8).
 4:     Calculate cooperative (easy) examples $\{x_{\text{disc}}\}$ and $\{x_{\text{tran}}\}$ (Equ 9).
 5:     Calculate adversarial (hard) examples $\{x_{\text{non-disc}}\}$ and $\{x_{\text{non-tran}}\}$ (Equ 10 and 11).
 6:     Calculate discriminability reg. $\mathcal{R}_{\text{disc}}(\theta_F, \theta_G)$ with $\{x_{\text{tran}}\}$ and $\{x_{\text{non-disc}}\}$ (Equ 12).
 7:     Calculate transferability reg. $\mathcal{R}_{\text{tran}}(\theta_F, \theta_H)$ with $\{x_{\text{disc}}\}$ and $\{x_{\text{non-tran}}\}$ (Equ 13).
 8:     Updating parameters by gradient backpropagation: $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}$ (Equ 14).
___

$$\ell_{\text{ALE\_tran}}(x) = \text{Dist.}\left[H(F(x)) \| H(F(x_{\text{non-tran}}))\right];$$
$$x_{\text{non-tran}} = \underset{x'; \|x-x'\| \le \epsilon}{\text{argmax}} \ \ell_{ce}(H(F(x')), H(F(x))). \tag{11}$$

Hence, the cooperative and adversarial regularization terms for discriminability are summarized as:

$$\mathcal{R}_{\text{disc}}(\theta_F, \theta_G) = \mathbb{E}_{x_i \sim \mathcal{D}_s \cup \mathcal{D}_t} \left[\ell_{\text{CLE\_disc}}(x_i; \theta_F, \theta_G) + \ell_{\text{ALE\_disc}}(x_i; \theta_F, \theta_G)\right], \tag{12}$$

and the cooperative and adversarial regularization terms for transferability are summarized as:

$$\mathcal{R}_{\text{tran}}(\theta_F, \theta_H) = \mathbb{E}_{x_i \sim \mathcal{D}_s \cup \mathcal{D}_t} \left[\ell_{\text{CLE\_tran}}(x_i; \theta_F, \theta_H) + \ell_{\text{ALE\_tran}}(x_i; \theta_F, \theta_H)\right]. \tag{13}$$

To sum up, the min-max paradigm can be written as:

$$\min_{\theta_F, \theta_G} \max_{\theta_H} \quad \mathcal{L}_{\text{disc}}(\theta_F, \theta_G) + \mathcal{L}_{\text{tran}}(\theta_F, \theta_H) + \lambda \left[\mathcal{R}_{\text{disc}}(\theta_F, \theta_G) + \mathcal{R}_{\text{tran}}(\theta_F, \theta_H)\right]. \tag{14}$$

4 Experiments

We conduct extensive experiments to show the effectiveness of the proposed CALE framework.

4.1 Experimental protocol

Datasets. We evaluate the CALE model and the baselines on three benchmark visual datasets: **Office-31** contains 4110 images from 31 categories of three distant domains, including Amazon (**A**), Webcam (**W**), and DSLR (**D**). **Office-Home**, a more challenging dataset, consists of 15588 images of 65 object classes in office and home environments, forming four extremely dissimilar domains: Artistic (**Ar**), Clip Art (**Cl**), Product (**Pr**), and Real World (**Rw**). **VisDA-2017**, a large dataset with 152397 **Synthetic** 3D rendered images and 55388 **Real-world** photos across 12 categories.

Baselines. We compare CALE with state-of-the-art deep DA approaches: Deep Adaptation Network (DAN) [35], Domain Adversarial Neural Network (DANN) [15], Conditional Domain Adversarial Network (CDAN) [37], Maximum Classifier Discrepancy (MCD) [47], Minimum Class Confusion (MCC) [24], Transferable Attention for Domain Adaptation (TADA) [58], Transferable Adversarial Training (TAT) [32], Batch Spectral Penalization (BSP) [9], and Cycle Self-Training (CST) [33]. Following the original implementation, all methods are with the ResNet [20] as their backbone. Due to the success of the Transformer [56, 14, 52] in various tasks, we further compare CALE with other baselines using the Data-efficient image Transformers (DeiT) [52] backbone, including CDTrans [61] and our implemented DeiT version of some representative baselines from above.

Implementation Details. We employ ResNet [20] and DeiT [52] pretrained on ImageNet1K [13] as the backbone network and attach a bottleneck layer with 256 units as the feature extractor F. We use a single fully-connected layer as the discriminative module G, and a three-layered conditional domain discriminator [37] as the transfer module H. The network architecture is followed the commonly used architecture of multiple recent domain adaptation approaches (e.g. [37, 9, 24, 23])

Table 1: Classification accuracy (%) on Office-Home with ResNet-50 (upper) and DeiT-Base (lower).

Method	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	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
DAN	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN	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	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
MCD*	51.6	72.7	77.6	62.5	68.6	70.4	62.7	52.1	78.2	74.4	57.9	82.2	67.6
MCC*	57.7	79.3	_82.8_	66.7	76.5	_77.8_	67.2	55.1	81.5	74.4	61.0	_85.9_	72.2
CST	_59.0_	**79.6**	**83.4**	_68.4_	_77.1_	76.7	_68.9_	_56.4_	83.0	_75.3_	_62.2_	85.1	_73.0_
TADA	53.1	72.3	77.2	59.1	71.2	72.1	59.7	53.1	78.4	72.4	60.0	82.9	67.6
TAT	51.6	69.5	75.4	59.4	69.5	68.6	59.5	50.5	76.8	70.9	56.6	81.6	65.8
BSP	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
CALE	**65.1**	_75.3_	80.8	**68.7**	**80.2**	**78.4**	**69.7**	**64.5**	**83.3**	**76.0**	**68.0**	**87.6**	**74.8**
DeiT-Base*	54.1	76.4	83.0	66.5	76.3	77.5	65.4	48.0	81.9	72.9	53.2	84.2	70.0
DAN*	56.7	76.0	83.2	68.4	75.7	78.6	66.3	50.6	81.3	74.8	56.4	84.7	71.1
DANN*	61.0	72.2	82.0	69.8	75.7	78.2	67.5	62.6	84.9	78.0	64.8	87.6	73.7
MCD*	60.2	77.8	83.9	72.4	73.2	75.5	68.6	59.2	82.8	80.7	62.3	86.4	73.6
MCC*	64.2	**85.8**	87.3	77.8	83.7	_85.6_	75.2	60.4	86.7	79.9	63.5	89.8	78.3
CST*	65.9	_85.3_	**88.0**	76.5	81.2	_85.6_	75.0	52.1	87.0	78.5	60.7	90.1	77.2
CDTrans	_68.8_	85.0	86.9	**81.5**	**87.1**	**87.3**	**79.6**	_63.3_	**88.2**	_82.0_	_66.0_	**90.6**	_80.5_
CALE	**71.5**	84.1	_87.6_	_78.4_	_86.3_	85.3	_79.2_	**70.7**	_87.7_	**82.4**	**74.2**	_90.4_	**81.5**

Table 2: Classification accuracy (%) on VisDA-2017 with ResNet-101 (upper) and DeiT-Base (lower).

Method	Plane	Bicycle	Bus	Car	Horse	Knife	Motor	Person	Plant	Ski	Train	Truck	Avg
ResNet-101	72.3	6.1	63.4	91.7	52.7	7.9	80.1	5.6	90.1	18.5	78.1	25.9	49.4
DAN	68.1	15.4	76.5	87.0	71.1	48.9	82.3	51.5	88.7	33.2	_88.9_	42.2	62.8
DANN	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
CDAN	93.6	82.3	66.2	_80.6_	92.7	10.7	87.1	70.0	94.6	38.4	76.6	47.2	70.0
MCD	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
MCC	94.5	80.8	78.4	65.3	90.6	79.4	_87.5_	_82.2_	_94.7_	81.0	86.0	44.6	80.4
CST*	_96.2_	**89.9**	73.3	**92.7**	_94.8_	**97.0**	81.5	79.4	**96.2**	_87.2_	83.3	_47.6_	_84.9_
CALE	**97.3**	_88.2_	**87.6**	74.7	**96.2**	_97.9_	**92.5**	**84.1**	94.5	**92.9**	**91.4**	**58.0**	**87.9**
DeiT-Base*	97.4	56.3	75.2	47.2	80.5	43.0	92.3	6.1	69.0	51.9	91.5	27.1	61.5
DAN*	97.2	50.3	_83.3_	46.2	93.5	77.7	_95.4_	18.1	87.2	69.8	_94.4_	26.6	70.0
DANN*	95.5	72.7	81.9	41.5	86.3	41.8	88.5	75.6	85.2	75.8	92.7	40.4	73.2
MCD*	94.7	82.6	66.4	77.9	86.1	97.0	92.5	73.6	95.0	27.0	89.2	50.0	77.7
MCC*	_98.1_	_90.7_	81.9	76.5	96.0	_97.8_	90.8	55.1	95.1	86.1	93.0	**63.9**	85.4
CST*	_98.1_	**93.5**	81.3	**90.5**	**97.6**	96.7	90.5	62.3	**98.0**	_93.8_	92.2	59.6	87.8
CDTrans	97.1	90.5	82.4	77.5	96.6	96.1	93.6	**88.6**	_97.9_	86.9	90.3	_62.8_	_88.4_
CALE	**98.7**	89.9	**88.0**	_87.8_	_97.4_	**98.5**	**96.0**	_86.5_	97.7	**95.5**	**95.4**	58.5	**90.8**

for a fair comparison. We adopt SGD with the learning rate annealed from 0.01 for training like Long et al. [37], Jiang et al. [23], and leverage FixMatch [48] for self-trining in $\mathcal{L}_{disc}$, which has been widely adopted in prior work (e.g., [33]). Throughout the experiments, the trade-off parameter of CALE regularization λ is set to 1, and the self-training threshold τ in Equation 8 is set to 0.95. The distance function Dist. is set as cross-entropy in *Office-31* and *Office-Home*, KL-divergence in *VisDA-2017*. More details about code and datasets can be found at `https://url` (removed for anonymous review).

4.2 Results and Analysis

Comparison with state-of-the-art. Experimental results on *Office-Home* and *VisDA-2017* are shown in Table 1, and Table 2, respectively. Due to space limitations, the results on the *Office-31*, the simplest one among three datasets, are reported in Table 4 (Appendix 7.1). The bolded and underlined numbers denote the best and the second best performance. The results marked with * are

Table 3: Results of ablation study on Office-Home (ResNet-50).

	$\mathcal{L}_{\mathrm{tran}}$	CALE		ST	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
1	✗	✗		✗	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
2	✗	✗		✓	51.0	66.4	75.9	48.0	65.0	64.5	55.3	46.2	77.1	70.9	55.3	79.5	62.9
3	✓	✗		✗	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
		$\mathcal{R}_{\mathrm{tran}}$	$\mathcal{R}_{\mathrm{disc}}$														
4	✓	✗	✓	✗	55.5	66.1	71.7	57.9	72.1	71.5	58.4	55.3	76.2	66.7	61.9	82.3	66.3
5	✓	✓	✗	✗	60.3	71.7	77.1	64.8	76.9	74.2	66.4	61.1	79.7	74.2	65.1	84.9	71.4
		CLE	ALE														
6	✓	✗	✓	✗	61.6	72.7	77.7	65.4	77.4	76.3	62.8	59.4	80.4	72.4	65.3	85.0	71.4
7	✓	✓	✗	✗	62.5	74.9	78.6	65.3	77.9	76.7	65.4	60.3	80.8	73.7	65.7	85.6	72.3
8	✓	✓		✗	63.2	75.9	79.9	68.1	79.3	77.3	68.4	64.0	82.1	74.8	67.1	86.7	73.9
9	✓	✓		✓	65.1	75.3	80.8	68.7	80.2	78.4	69.7	64.5	83.3	76.0	68.0	87.6	74.8

based on our reproduction, while all the other results are from their original paper. The results reveal several insightful observations: (1) With both backbones (ResNet and DeiT), CALE outperforms the compared SOTA approaches. Especially on the challenging datasets that have large domain shift: *OfficeHome* (ResNet-50: +1.8%, DeiT-Base: +1.0%) and *VisDA-2017* (ResNet-101:+3.0%, DeiT-Base:+2.4%), the enhancements are significant. (2) TADA and TAT try to let the transfer module asymmetrically guide the discriminative module. BSP attempts to penalize the transferability for enhancing discriminability heuristically. CALE outperforms TADA, TAT, and BSP significantly, which verifies that the proposed idea of bidirectional co-enhancing the discriminability and transferability is helpful for building a better domain adaptation model, compared to the unidirectional (TADA and TAT) or imbalanced (BSP) enhancement of discriminability or transferability.

Ablation Study. We conduct numerous ablation studies to validate the individual contributions of different components. The ablation results are reported in Table 3, where $\mathcal{L}_{\mathrm{tran}}$ denotes the transferability loss in Equation 7, *CALE* denotes the CALE regularization term, and *ST* denotes the self-training term in Equation 8. In rows 4 and 5, we split the CALE regularization into the transferability regularization $\mathcal{R}_{\mathrm{tran}}$ and the discriminability regularization $\mathcal{R}_{\mathrm{disc}}$ as shown in the Equation 14. Compare rows 4, 5, and 8, it shows that both $\mathcal{R}_{\mathrm{tran}}$ and $\mathcal{R}_{\mathrm{disc}}$ are necessary, especially enhancing transferability through $\mathcal{R}_{\mathrm{tran}}$ is critical. Moreover, in rows 5 and 6, we split the CALE regularization into cooperative learning (CLE) and adversarial learning (ALE) terms. The comparison of rows 5, 6, and 8 verifies that both the CLE and the ALE mechanisms are effective, and the enhancement from CLE is more remarkable. Last, it is worth noting that the result of row 8, our method is still the best on *Office-Home* even without the help of self-training. Other results also support our idea and the concrete CALE model.

Hyper-Parameter Sensitivity of Performance and Convergence. We study the sensitivity of the hyper-parameter λ in Equation 14, which determines the strength of the CALE regularization loss. The results of test accuracy on *VisDA-2017* are shown in Figure 4a, which reveal that global accuracy and class mean accuracy are stable with the various value of λ. Besides, we investigate the convergence of test global accuracy and the CALE regularization loss when setting different values of λ. The results on the convergence are shown in Figure 4b and Figure 4c.

Discriminability and Transferability. In Figure 2 we show the discriminability and transferability of the proposed CALE model and other domain adaptation models. We use classification accuracy to measure the discriminability. For transferability, we adopt $\mathcal{A}$-distance as a (negative) indicator. The $\mathcal{A}$-distance $= 2(1 - 2\epsilon)$ is first introduced by Ben-David et al. [1] to measure the discrepancy across domains, where ϵ means the test error of a binary classifier to distinguish the source domain from the target domain on the learned feature representation. As shown in Figure 2, CALE is the only Pareto optimal among compared approaches. The best discriminability on the target illustrates that CALE can learn more discriminative features for the target. Moreover, the best transferability between source and target indicates that CALE can transfer more knowledge from source to target.

(a) Test global accuracy and class mean accuracy w.r.t. λ.

(b) Convergence of the global accuracy with various λ.

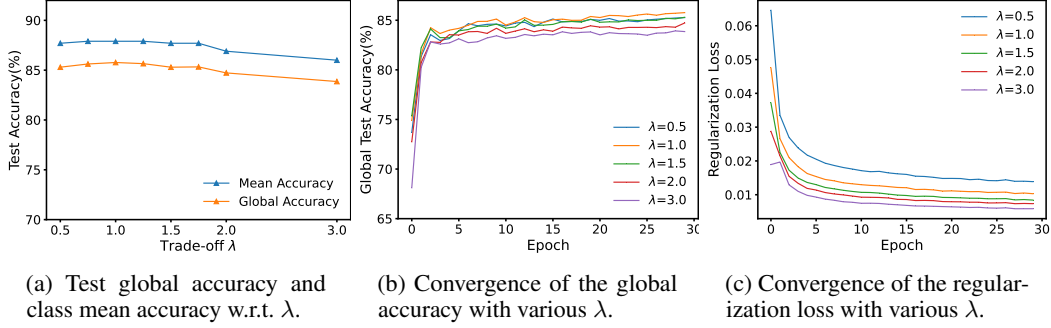(c) Convergence of the regularization loss with various λ.

Figure 4: Hyper-parameter sensitivity of classification performance and convergence.

5   Related works

Domain adaptation aims to overcome the *domain-shift* for transferring knowledge from a label-rich source domain to a label-poor target domain [43]. In the early shallow regime, Bickel et al. [3], Sugiyama et al. [49] re-weight source data for use in the target domain, Wang and Mahadevan [57], Blitzer et al. [4] learn representations invariant across domains, while Gao et al. [16] transfer model based on shared parameters and Mihalkova et al. [40] based on relational-knowledge.

The theoretical analysis of Ben-David et al. [2] and the success of deep neural networks inspire the modern deep domain adaptation. Two mainstream technologies rise: moment matching and adversarial confusing. The moment matching approaches design a distribution distance based on the statistical moment, and minimize the distance across domains. Sun and Saenko [50] matches the second-order statistical moment (covariance) in shared feature space, while Li et al. [31], Maria Carlucci et al. [39], Mancini et al. [38] match the first-order (mean) and second-order (variance) statistical moments in the Batch Normalization layer [22]. Tzeng et al. [53], Long et al. [35, 36] align the cross-domain means in Reproducing Kernel Hilbert Space (RKHS) based on Maximum Mean Discrepancy (MMD) [19] and its variants. In the other direction, motivated by GANs [17], the adversarial confusing approaches learn domain-invariant feature representations by training the feature extractor and a domain discriminator in an adversarial way. Ganin et al. [15] leverage Gradient Reverse Layer (GRL) to perform one-step optimization of adversarial min-max problem while Tzeng et al. [54] learns asymmetric feature extractors for source and targets. Long et al. [37] build a conditional domain discriminator based on Conditional-GANs [41]. Later, Pei et al. [44], Saito et al. [46, 47] propose various improvements for pursuing more fine-grained alignment. Most recently, Hoffman et al. [21], Chen et al. [7] adopt Cycle-GANs [63] to achieve impressive performances.

How to balance and consolidate the enhancement of discriminability and transferability in feature learning is getting more attention from researchers. Chen et al. [9] attempt to penalize the transferability for enhancing discriminability heuristically. They consider learning transferability too much may harm the discriminability. Wang et al. [58], Kurmi et al. [28], Liu et al. [32] try to let the transfer module asymmetrically guide the discriminative module. In this paper, we further argue that the guidance between the transfer module and the discriminative module can be reciprocal. We propose the CALE model, which further extends the limit of transferability and discriminability that the existing feature learning approaches can reach.

6   Conclusions

Discriminability and transferability are two goals of feature learning for domain adaptation, as we aim to find the transferable features from the source domain that are helpful for discriminating the class label in the target domain. We claim that the enhancement of discriminability and transferability should be considered jointly instead of separately. We propose Cooperative and Adversarial LEarning (CALE) for achieving the two goals simultaneously by letting the transfer module and the discriminative module help each other. Besides, we show that it is beneficial to use the information in one module to guide another. It might be possible to find different ways to unify the learning of discriminability and transferability, which could be an exciting direction in the future.

References

[1] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of Representations for Domain Adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2006.

[2] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning (ML)*, 79(1):151–175, 2010.

[3] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 81–88, 2007.

[4] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 120–128, 2006.

[5] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 343–351, 2016.

[6] T. Cai, R. Gao, J. Lee, and Q. Lei. A theory of label propagation for subpopulation shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1170–1182. PMLR, 2021.

[7] C. Chen, W. Xie, Y. Wen, Y. Huang, and X. Ding. Multiple-source domain adaptation with generative adversarial nets. *Knowledge-Based Systems (KBS)*, 199:105962, 2020.

[8] X. Chen, S. Wang, B. Fu, M. Long, and J. Wang. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[9] X. Chen, S. Wang, M. Long, and J. Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1081–1090. PMLR, 2019.

[10] X.-H. Chen, S. Jiang, F. Xu, Z. Zhang, and Y. Yu. Cross-modal domain adaptation for cost-efficient visual reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

[11] Y. Chen, C. Wei, A. Kumar, and T. Ma. Self-training avoids using spurious features under domain shift. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:21061–21071, 2020.

[12] Z. Chen, R. Xiao, C. Li, G. Ye, H. Sun, and H. Deng. Esam: Discriminative domain adaptation with non-displayed items to improve long-tail performance. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 579–588, 2020.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

[15] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research (JMLR)*, 17 (1):2096–2030, 2016.

[16] J. Gao, W. Fan, J. Jiang, and J. Han. Knowledge transfer via multiple model local structure mapping. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 283–291, 2008.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, 2014.

[18] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[19] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research (JMLR)*, 13:723–773, 2012.

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

[21] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1989–1998. PMLR, 2018.

[22] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 448–456. PMLR, 2015.

[23] J. Jiang, B. Chen, B. Fu, and M. Long. Transfer-learning-library. `https://github.com/thuml/Transfer-Learning-Library`, 2020.

[24] Y. Jin, X. Wang, M. Long, and J. Wang. Minimum Class Confusion for Versatile Domain Adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–480, 2019.

[25] H. Kanagawa, H. Kobayashi, N. Shimizu, Y. Tagami, and T. Suzuki. Cross-domain recommendation via deep domain adaptation. In *European Conference on Information Retrieval (ECIR)*, pages 20–29. Springer, 2019.

[26] R. Kontar, G. Raskutti, and S. Zhou. Minimizing negative transfer of knowledge in multivariate gaussian processes: A scalable and regularized approach. *IEEE Transactions on Pattern Analysis Machine Intelligence (TPAMI)*, 43(10):3508–3522, 2020.

[27] A. Kumar, T. Ma, and P. Liang. Understanding self-training for gradual domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5468–5479. PMLR, 2020.

[28] V. K. Kurmi, S. Kumar, and V. P. Namboodiri. Attending to discriminative certainty for domain adaptation. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition (CVPR)*, pages 491–500. IEEE, 2019.

[29] E. Lekhtman, Y. Ziser, and R. Reichart. Dilbert: Customized pre-training for domain adaptation with category shift, with an application to aspect extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 219–230, 2021.

[30] X. Li and D. Caragea. Domain adaptation with reconstruction for disaster tweet classification. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1561–1564, 2020.

[31] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou. Revisiting batch normalization for practical domain adaptation. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net, 2017.

[32] H. Liu, M. Long, J. Wang, and M. I. Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4013–4022. PMLR, 2019.

[33] H. Liu, J. Wang, and M. Long. Cycle Self-Training for Domain Adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[34] Y. Liu, W. Zhang, and J. Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition (CVPR)*, pages 1215–1224. IEEE, 2021.

[35] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 97–105. PMLR, 2015.

[36] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2208–2217. PMLR, 2017.

[37] M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1647–1657, 2018.

[38] M. Mancini, L. Porzi, S. R. Bulo, B. Caputo, and E. Ricci. Inferring latent domains for unsupervised deep domain adaptation. *IEEE Transactions on Pattern Analysis Machine Intelligence (TPAMI)*, 43(2):485–498, 2019.

[39] F. Maria Carlucci, L. Porzi, B. Caputo, E. Ricci, and S. Rota Bulo. Autodial: Automatic domain alignment layers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5077–5085. IEEE, 2017.

[40] L. Mihalkova, T. Huynh, and R. J. Mooney. Mapping and revising markov logic networks for transfer learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 608–614. IEEE, 2007.

[41] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[42] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training : A regularization method for supervised and. *IEEE Transactions on Pattern Analysis Machine Intelligence (TPAMI)*, 41(8):1979–1993, 2018.

[43] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(10):1345–1359, 2010.

[44] Z. Pei, Z. Cao, M. Long, and J. Wang. Multi-adversarial domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3934–3941. IEEE, 2018.

[45] K. Rao, C. Harris, A. Irpan, S. Levine, J. Ibarz, and M. Khansari. Rl-cyclegan: Reinforcement learning aware simulation-to-real. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition (CVPR)*, pages 11157–11166. IEEE, 2020.

[46] K. Saito, Y. Ushiku, T. Harada, and K. Saenko. Adversarial dropout regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net, 2018.

[47] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition (CVPR)*, pages 3723–3732. IEEE, 2018.

[48] K. Sohn, D. Berthelot, C. L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[49] M. Sugiyama, S. Nakajima, H. Kashima, P. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in Neural Information Processing Systems (NeurIPS)*, 20, 2007.

[50] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 443–450. Springer, 2016.

[51] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. A survey on deep transfer learning. In *Proceedings of the IEEE International Conference on Artificial Neural Networks (ICANN)*, pages 270–279. Springer, 2018.

[52] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 10347–10357. PMLR, 2021.

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

[54] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition (CVPR)*, pages 7167–7176. IEEE, 2017.

[55] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *The Journal of Machine Learning Research (JMLR)*, 9(11), 2008.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[57] C. Wang and S. Mahadevan. Manifold alignment using procrustes analysis. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1120–1127. PMLR, 2008.

[58] X. Wang, L. Li, W. Ye, M. Long, and J. Wang. Transferable attention for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 5345–5352. IEEE, 2019.

[59] Z. Wang, Z. Dai, B. Póczos, and J. Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE Conference Computer Vision Pattern Recognition (CVPR)*, pages 11293–11302. IEEE, 2019.

[60] S. Xie, Z. Zheng, L. Chen, and C. Chen. Learning semantic representations for unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5423–5432. PMLR, 2018.

[61] T. Xu, W. Chen, P. Wang, F. Wang, H. Li, and R. Jin. CDTrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021.

[62] H. Zhu, Z. Wang, H. Zhang, M. Liu, S. Zhao, and B. Qin. Less is more: Domain adaptation with lottery ticket for reading comprehension. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1102–1113, 2021.

[63] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232. IEEE, 2017.

7 Additional experiments

7.1 Comparison with state-of-the-art on Office-31

We here report the experimental results on *Office-31* in Table 4. It can be observed in the table that CALE achieves performance improvement with both ResNet and DeiT backbone.

Table 4: Classification accuracy (%) on Office-31 with ResNet-50 (upper) and DeiT-Base (lower).

Method	$A \rightarrow D$	$A \rightarrow W$	$D \rightarrow A$	$D \rightarrow W$	$W \rightarrow A$	$W \rightarrow D$	Avg
ResNet-50	68.9	68.4	62.5	96.7	60.7	99.3	76.1
DAN	78.6	80.5	63.6	97.1	62.8	99.6	80.4
DANN	79.7	82.0	68.2	96.9	67.4	99.1	82.2
CDAN	89.8	93.1	70.1	98.2	68.0	**100.0**	86.6
MCD*	91.4	89.2	70.0	98.2	68.5	**100.0**	86.2
MCC	**95.0**	**94.7**	73.0	98.6	<u>73.6</u>	**100.0**	<u>89.2</u>
CST*	<u>94.9</u>	85.0	<u>75.6</u>	<u>99.2</u>	73.3	<u>99.9</u>	88.0
TADA	91.6	<u>94.3</u>	72.9	98.7	73.0	99.8	88.4
TAT	93.2	92.5	73.1	**99.3**	72.1	**100.0**	88.4
BSP	93.0	93.3	73.6	98.2	72.6	**100.0**	88.5
CALE	92.8	91.6	**77.0**	98.6	**76.9**	**100.0**	**89.5**
DeiT-Base*	85.5	88.1	74.8	98.9	75.9	**100.0**	87.2
DAN*	89.2	91.8	77.4	98.9	75.9	**100.0**	88.9
DANN*	87.7	93.8	80.1	98.3	79.6	**100.0**	89.9
MCD*	94.7	95.1	72.9	98.6	73.8	<u>99.8</u>	89.2
MCC*	<u>97.0</u>	96.5	80.5	98.9	80.6	**100.0**	92.3
CST*	96.6	94.2	78.2	**99.2**	79.7	**100.0**	91.3
CDTrans	<u>97.0</u>	<u>96.7</u>	<u>81.1</u>	<u>99.0</u>	**81.9**	**100.0**	<u>92.6</u>
CALE	**97.6**	**98.0**	**82.4**	<u>99.0</u>	<u>80.7</u>	**100.0**	**93.0**

7.2 Feature Visualization.

To present the adaptation process intuitively, we leverage t-SNE [55] to visualize the feature representations of *VisDA-2017*. The results are shown in Figure 5, where different colors mean different categories, and circles and triangles denote source data and target data, respectively. Apparently, in Figure 5a, the feature of target before adaptation is non-discriminative and non-transferable. However, in Figure 5b, the target's feature adapted by CALE is discriminative (target features that come from different categories can be separated easily) and transferable (source feature and target feature are confused).
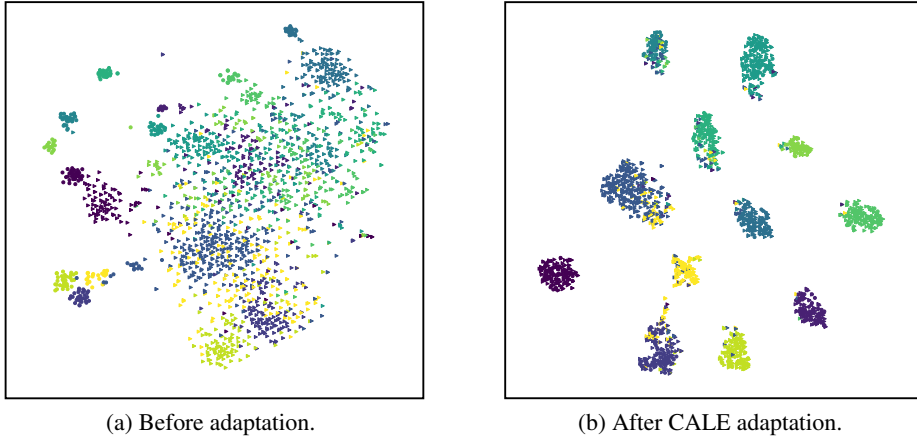

(a) Before adaptation.
(b) After CALE adaptation.

Figure 5: Feature visualizations.